Impact of reference panel choice on imputation into genome-wide association studies of complex human traits

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by

Jack Flanagan

March 2022

Abstract

Thesis Title: Impact of reference panel choice on imputation into genome-wide association studies of complex human traits

Author: Jack Flanagan

Genome-wide association studies (GWAS) have been successful in identifying loci contributing genetic effects to complex human traits and diseases. These advances have been supported by "imputation", the process of predicting genotypes at single nucleotide polymorphisms (SNPs) that are not directly genotyped by GWAS arrays but are present on "reference panels" comprised of whole-genome sequence (WGS) data, such as the 1000 Genomes Project (1KG) and Haplotype Reference Consortium (HRC). The similarity of genetic ancestry between the GWAS and reference panel is an important factor in the imputation quality achieved. This thesis investigates variation in imputation quality into GWAS from different population groups and addresses the use of population specific WGS data to augment existing reference panels to improve imputation quality and enhance opportunities for GWAS locus discovery and fine-mapping.

Imputation quality across different population groups from a multi-ancestry GWAS in the Resource for Genetic Epidemiology Research on Adult Health and Aging (GERA)cohort was compared for different reference panels, including 1KG and HRC. Results indicated variation in imputation quality between population groups and highlighted limited imputation quality when imputing East Asian GWAS because of an underrepresentation of this ancestry group in commonly used reference panels, emphasising the need for greater population diversity in these resources.

Following on from these observations, the performance of imputation into GWAS from Biobank Japan was assessed for four Japanese population specific reference panels using WGS data of various sample sizes and sequencing depths to augment the 1KG. An augmented reference panel that included 7,472 WGS of mixed depth (1KG+7K) provided the greatest improvement in imputation quality relative to the 1KG panel alone, with ~2x the total number of imputed variants and ~3.8x the number of "well-imputed" variants (defined by $r^2 > 0.3$). Furthermore, most these improvements were observed for rare variants with a minor allele frequency (MAF) < 1%. To demonstrate the benefits of improved imputation in association analyses of complex traits, a GWAS of serum uric acid levels was performed in Biobank Japan after imputation up to the 1KG+7K panel. The analysis revealed eight potentially novel loci at genome-wide significance ($P<5x10^{-8}$) distinct from those previously reported in addition to the identification of multiple distinct signals at six loci. Improved imputation of rare variants supported a subsequent gene-based analysis, identifying two genes in which multiple rare coding variants were associated at exome-wide significance ($P<2.5x10^{-6}$) in genes with biological functions directly related to serum uric acid levels.

The work in this thesis highlights the improvements in imputation quality achieved by augmenting a publicly available reference panel with population specific WGS data, and the benefits for discovery and fine-mapping of complex trait loci This offers improved routes for clinical translation of GWAS through the development of more accurate polygenic risk scores, and the potential to aid the discovery of novel drug targets for the treatment and management of complex diseases.

ACKNOWLEDGEMENTS

I would like to thank my supervisors Professor Andrew Morris and Dr Momoko Horikoshi for giving me the opportunity to take on this project, for their constant support and guidance as well as the time they have invested in me throughout the past four years. I would also like to thank Dr Chikashi Terao and the members of the Laboratory for Statistical and Translational Genetics who helped guide and support my research in Japan, as well as Naoko Kuroki who organised and helped me with so many things during my stay in Japan. I am incredibly grateful to all who have supported me through this project and have helped me progress as a person and as a researcher.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

1KG	1000 Genomes Project Phase III
1KG+1K	1000 Genomes Project Phase III + 1,037 Japanese WGS
1KG+3K	1000 Genomes Project Phase III + 3,256 Japanese WGS
1KG+4K	1000 Genomes Project Phase III + 4,216 Japanese WGS
1KG+7K	1000 Genomes Project Phase III + 7,472 Japanese WGS
AMD	Age-related macular degeneration
AFR	African American ancestry
BBJ	Biobank Japan
CAAPA	Consortium on Asthma among African-ancestry populations in the Americas
CNV	Copy number variant
DNA	Deoxyribonucleic acid
EAS	East Asian ancestry
ERF	Erasmus Rucphen Family study
EUR	European ancestry
GAsP	Genome Asia panel
GERA	Resource for Genetic Epidemiology Research on Ageing
GoNL	Genome of The Netherlands
GWAS	Genome wide association study
HMM	Hidden Markov model
HRC	Haplotype reference consortium
HWE	Hardy-Weinberg equilibrium
IBD	Identity by descent
KB	Kilobase

KP RPGEH	Kaiser Permanente Research Program on Genes, Environment, and Health
LAT	Latino ancestry
LD	Linkage disequilibrium
MAC	Minor allele count
MAF	Minor allele frequency
MCTFR	Minnesota Center for Twin and Family Research
MSU	Monosodium urate
NGS	Next generation sequencing
PC	Principal component
PCA	Principal component analysis
PCR	Polymerase chain reaction
QC	Quality control
SFTP	Secure file transfer protocol
SKAT	Sequence Kernel Association Test
SNP	Single nucleotide polymorphism
VCDR	Vertical cup-disc ratio
WGS	Whole-genome sequence

CHAPTER 1

INTRODUCTION

1.1 Introduction

Genome-wide association studies (GWAS) present a comprehensive approach to utilising sequencing and genotype data with the primary aim of identifying associations between genetic variants and complex traits. Prior to the adoption of GWAS methods, linkage studies had proven to be a highly effective approach for the analysis of Mendelian traits (i.e. traits determined by a single gene) but remained limited in their ability to characterise the genetic contribution to complex traits (i.e. traits determined by multiple genes that potentially interact with the environment). This can be attributed to three key features of complex traits: variability in the age of onset and the severity of symptoms, the implication of multiple biological pathways, and the implication of numerous genes, each making modest contributions to the overall manifestation of the phenotype of interest (Tabor et al. 2002). Candidate-gene association studies offered a hypothesis-driven approach to investigating complex traits that implicate multiple sites of genetic variation of a modest effect size. Whilst these studies presented greater power to detect multiple variants of modest effect size, they remained limited in their effectiveness to fully understand genetic contribution to complex traits. This can be attributed to the need to pre-select candidate genes based on prior knowledge of the trait of interest, immediately limiting the scope of the study to a few select regions of the genome. GWAS however, represents a hypothesis free approach to association analysis, incorporating variation across the entire genome without limiting analysis to pre-selected genes. In its most simple terms, provided with the genotype information for a sample cohort and phenotype information to complement said sample group, we can test all individually genotyped sites for association with the phenotype of interest across the entire sample group.

Since their inception, GWAS and the associated methodologies have been rapidly evolving, supported by modern developments in genotyping platforms, whole genome sequencing (WGS) technologies, computing and software developments, and the advent of large repositories of genetic and medical information such as the UK Biobank. Often cited as the one of the first successful GWAS, a study on age-related macular degeneration (AMD) tested a total of 116,204 single nucleotide polymorphisms (SNPs) across a total cohort of 96 cases and 50 controls, identifying a single common variant of a large effect size (relative risk of 7.4 for individuals homozygous for the risk allele) (Klein et al. 2005). This landmark publication served to present the utility of GWAS in the discovery of trait associated loci (a genomic region associated with a trait, often defined by a single lead variant encompassing a region of +/- 500kb), but also highlights the drastic change in GWAS design in comparison to modern studies. One such study (Jansen et al. 2019) into the genetic component of insomnia amassed a total sample set of over 1.3 million individuals at ~39 million SNPs, performing GWAS to identify 202 loci, encompassing a total of 956 implicated genes. Whilst these examples do not capture the entirety of the developments made in GWAS methods, they do provide a clear example of the progress made, not only in terms of the total amount of information that can be incorporated into a modern GWAS, but also the level of detail at which complex traits can be described. The continued refinement and widespread adoption of GWAS methods has meant that as of 2021, more than 5,700 GWAS have been completed, encompassing over 3,330 traits (Uffelmann et al, 2021). However, even with the current wealth of information produced by GWAS, there is a substantial degree of heritability, meaning the proportion of trait variance that can be explained by genetics, for complex traits that remains unexplained. This has prompted further investigation into the adaptation of GWAS methods to capture sources of association that escape current studies. One such method of addressing this issue is through the use of imputation to increase the total number of variants that can be tested for association in a GWAS. Using established resources of WGS data, GWAS sample sets genotyped at a relatively small number of SNPs can be scaled up to encompass millions of variants through the statistical estimation of missing genotypes based on concepts discussed in sections 1.2 and 1.3.

This chapter will introduce the underlying concepts behind GWAS and provide context for the research described throughout this thesis. I will explore the topics of genetic variation, haplotypes and linkage disequilibrium to introduce the key genetic features that facilitate GWAS. Following this, I will explain the core components of standard GWAS study design, from sample set and phenotype selection, quality control procedures, to association analyses and an explanation of the software developments surrounding GWAS. Furthermore, descriptions of sequencing and genotyping methods and their modern developments will be provided, placing these technologies within the context of GWAS and their role in the evolution of the field. This chapter will also pay particular focus to the topics of imputation and the development of detailed reference panels of genetic variation, assessing the developments in imputation methods and the relevant software, and the construction and design of reference panels key to the imputation process.

1.2 Genome Wide Association Studies

1.2.1 Single nucleotide polymorphisms

Single nucleotide polymorphisms (SNPs) are the most common form of genetic variation and occur when, through a variety of mechanisms, a change in a single nucleotide base takes place. SNPs occur in both non-coding and protein coding regions. Those in protein coding regions are divided into two major categories, synonymous and non-synonymous. Non-synonymous SNPs produce a change of amino acid within a protein often leading to disruption with downstream protein-protein interactions, whilst synonymous SNPs do not alter the amino acids encoded by their specific gene. However, it is important to note that SNPs localised to non-coding regions and synonymous SNPs within coding regions still have the potential to impact biological functions through more indirect means. These SNPs have been implicated in altered messenger RNA splicing and stability and can exert influence on the regulation of gene expression (Hunt et al. 2009).

1.2.2 Haplotypes and Linkage disequilibrium

A haplotype is traditionally defined as a collection of alleles inherited together from a single parent (and therefore on the same chromosome). A low mutation rate in combination with knowledge that recombination events occur more frequently in specific regions of chromosomes implies that novel mutations can be grouped and inherited with specific sets of alleles. Novel haplotypes can be formed through either additional mutations or through recombination. However, these new haplotypes can be described as a mosaic of multiple known haplotypes, thus preserving a degree of knowledge in relation to the pattern of inherited groups of alleles (Pääbo, 2003).

Linkage disequilibrium (LD), the non-random association of alleles of different loci can be used to describe the correlation between alleles on the same haplotype and is particularly useful in GWAS. Measures of LD are based on the difference between the observed frequency of two alleles occurring together and the expected frequency of the two alleles under the assumption that they are independent of one another. One commonly used method of quantifying LD is through the measure of r^2.

$$r^2 = \frac{(\pi_{AB}\pi_{ab} - \pi_{Ab}\pi_{aB})^2}{\pi_A\pi_B\pi_a\pi_b}$$

Equation 1.2.1

This equation works on the assumption of two alleles (*A* and *a*) at one SNP and two alleles (*B* and *b*) at a second SNP. Therefore, π_{ab} represents the frequency of the haplotype containing both alleles *a* and *b*, with $\pi_a$denoting the frequency of allele *a* with $\pi_b$denoting the frequency of allele *b* (Bush and Moore, 2012). In summary, r^2 is reported on a range between 0 and 1, where r^2=1 corresponds to perfect LD between SNPs and where r^2=0 corresponds to linkage equilibrium (i.e. no correlation) between SNPs. The closer the r^2 value is to 1, the greater the correlations between SNPs. This facilitates the optimisation of genotyping within population groups via the selection of 'tag-SNPs' that can describe genetic variation in regions of high LD without the need for genotyping all SNPs within the region (Li et al. 2008)

1.2.3 Array based genotyping

Current methods for genotyping GWAS sample sets commonly use array-based genotyping platforms. Samples are genotyped at a pre-selected set of variants, acting as tag SNPs, with the goal to provide the greatest degree of genomic coverage whilst remaining cost efficient when genotyping increasingly large sample sets. Coverage is directly linked to the total number of genotyped variants and their potential to tag non-genotyped variants through LD, and is defined as the capacity for these variants to represent genetic variation across the genome.

Although the exact methods vary between manufacturers, the fundamentals behind the process of genotyping are shared. All genotyping platforms work on the basis that nucleotide bases bind to their complementary base (A to T and C to G). SNP arrays are comprised of hundreds of thousands of immobilised allele specific probes. Fluorescently labelled DNA fragments are hybridized to complementary probes on the array and the affinity to which the DNA fragments hybridize to the probes can be inferred from the signal intensities recorded via fluorescence (LaFramboise, 2009).

Companies such as Affymetrix and Illumina provide a variety of genotyping arrays that cater to different GWAS study designs and sample sets. This includes SNP selection to maximise coverage of sample sets comprised of singular ancestry groups, or alternatively to cater to sample sets of diverse ancestry groups. Furthermore, a tier system in terms of the number of SNPs included in the array allows for more control over the allocation of financial resources

1.2.4 Whole genome sequencing data

Whole genome sequencing (WGS) is the most comprehensive method of capturing the variation within an individual's genome, providing detailed information encompassing variant features such as single nucleotide variants, structural variants and indels. Single nucleotide variants generally fall into one of three categories; SNPs (section 1.2.1), single nucleotide insertions and single nucleotide deletions, which represent a change of nucleotide, insertion of an additional nucleotide or the removal

of an existing nucleotide, respectively. Structural variants can be defined as alterations to segments of DNA of 1 kilobase (kb) or more. This category of variation encompasses multiple types of structural variants including translocations (change in position of a chromosomal segment) and inversions (the reversal of a chromosome segment stemming from a break away from the chromosome followed by reintegration back into the chromosome) (Feuk et al, 2006). Indels are insertions or deletions of nucleotides in the genome up to lengths of 1kb. Indels that span 1kb are often classified as copy number variants (CNVs) and are grouped within the structural variant category (Ku et al, 2010).

At its inception, the process of sequencing the human genome was prohibitively expensive with estimates of the first 'draft' human genome sequence in the year 2000 reaching as high as $300 million. Additionally, the refinement process of this draft sequence undertaken by the Human Genome Project was estimated to cost an additional $150 million (National Human Genome Research Institute,2021). These costs can only remain as estimates due to the expansive nature of the research project and the difficulty in categorising the expenses involved. However, these estimates do serve as a reference point when discussing the rapid advancements in the field of WGS, commercialisation and the availability of WGS data.

The development of next generation sequencing (NGS) prompted the estimated cost of sequencing a human genome to fall from ~$20-25 million in 2006 to as low as ~$1500 by the end of 2015 by providing high throughput methods for the generation of large amounts of sequencing data in a timely manner (National Human Genome Research Institute, 2021). NGS covers a broad spectrum of sequencing methods that vary between different platforms and the respective companies that develop them. One of the most common methods for sequencing is on the platforms provided by Illumina, which share a general set of methods to accomplish template preparation, sequencing and imaging. DNA templates are generated by fragmentation of the DNA sample and through ligation, attaching adaptor sequences to said fragments. Adaptor sequences are short synthetic DNA sequences that facilitate the recognition of their attached DNA fragments by the sequencing platform used. These fragments are then

used to construct a sequencing library through amplification via polymerase chain reaction (PCR). Illumina NGS platforms specifically rely on a method of sequencing known as 'sequencing by synthesis'. This technique uses the library of DNA fragments as templates for the synthesis of DNA, recording the integration of nucleotides into the sequence as it is generated (Grada and Weinbrecht, 2013).

When considering the expenses involved in WGS based research, the general trend follows that of decreasing costs and increasing availability. However, as modern GWAS implicate sample sizes in the hundreds of thousands the generation of WGS for sample sets remains prohibitively expensive. Whilst the use of WGS techniques is not yet applicable to large sample sets, WGS data retains a key role in GWAS studies through the development of reference panels for imputation. The sample sets used in the development of reference panels are typically much smaller than those used in GWAS and so WGS is a more realistic prospect. This allows comprehensive descriptions of haplotypes and variants across the genomes of a sub-set of a select populations. Information concerning the haplotypes and LD structure within a population, alongside detailed variant information can be used to impute genotypes at untyped SNPs in GWAS sample sets that are only subject to genotyping at a relatively limited number of SNPs (see section 1.3).

1.2.5 GWAS sample cohort

The size and composition of the sample set used in a GWAS forms one of the key determinants of the statistical power of the study. Sample size remains one of the limiting factors in the discovery trait associated variants, as many causal variants contributing to complex traits are of small effect sizes (the impact of the variant on the phenotypic value) and therefore, the power to detect these variants is limited. This limitation in the power to detect is especially true of variants that occur at a low frequency within the population group tested. It is also important to note that whilst a particular variant may have only a modest impact on the phenotypic value, this does not translate to the biological and therapeutic relevance of the variant (Wray et al. 2018). Tam et al. 2019 assessed the number of loci identified as a function of GWAS sample size for three traits; body mass index, height, and waist to hip ratio

adjusted for BMI. Their results suggested that each trait had a threshold sample size at which the rate of locus discovery increased dramatically, observing no signs of plateau at a sample size of < 1 million. These results suggest that there are still gains to be made solely as a function of establishing larger sample sets.

With modern studies employing increasingly large sample sizes of hundreds of thousands and some touching upon millions of individual samples, many GWAS will utilise data from large population-based biobanks such as the UK Biobank and Biobank Japan. These large resources provide de-identified genetic and medical data, in addition to relevant personal information (e.g. sex, ethnicity, age), establishing a suitable resource for GWAS investigating complex traits/disease relevant to the selection criteria of the samples included within the resource.

The phenotype/trait selected is a major factor that impacts the availability of relevant samples to be included within a GWAS. Phenotypes are generally divided into two categories, binary (using a case control system reporting the presence or absence of the disease/trait) or quantitative/continuous traits that include measurements of variables such as serum uric acid levels. Depending on the phenotype used in GWAS, problems can arise from data availability. For example, particularly rare phenotypes will have a limited population from which samples can be taken, thereby limiting the power of any studies investigating these phenotypes. Furthermore, some traits/diseases can be difficult to measure in a quantitative manner, making it complex to establish a clear definition of the phenotype. This issue of imprecise phenotype definition can become particularly impactful when performing GWAS into binary traits, as incorrectly defining samples as either cases or controls will be detrimental to the studies power and can introduce bias in the effect estimates of causal variants.

Modern biobanks provide a great degree of support for GWAS, not only providing large collections of genotyping data, but also extensive medical records for the samples within the biobank. One example, Biobank Japan (Nagai et al, 2017), was established in 2003 with the aim of supporting genomics research into the Japanese

population. Participants were selected based on the diagnosis of one of 47 target diseases selected due to their clinical importance in Japan. 200,000 individuals were registered to the study over a five-year period and individuals enrolled in the study provided both DNA and serum samples for analysis. The information provided via biological sample was supported with a series of interviews documenting behavioural and environmental factors, in addition to further reviews of medical records. Furthermore, this information was supplemented with results from routine laboratory examinations as well as disease specific laboratory examinations. Patients enrolled in the programme were subject to follow up surveys to document any changes in an individual's behavioural, environmental and medical circumstances, including the development of any new diseases, and for 32/47 of the selected diseases, survival data was collected. Following the removal of individuals who withdrew consent, the conclusion of the initial five year period saw Biobank Japan amass a total of 199,982 individuals available for analysis, with a subset of 141,612 individuals for which survival time data was recorded. The final result was a comprehensive genomics resource to support research into 47 diseases within the Japanese population, providing detailed phenotype information alongside a plethora of recorded quantitative traits for a large sample set representing the Japanese population. Resources such as Biobank Japan establish the necessary phenotypic information to compliment the genotype information for each sample. With this information, samples for which the appropriate phenotype information is available can be included in the association analysis.

1.2.6 Sample quality control

Prior to performing association analysis, a fundamental step is to ensure the quality of the sample set to be used in said analyses. A standardised set of quality control (QC) protocols can aid in the detection of any errors that have occurred during sample collection, reporting and genotyping procedures, allowing the exclusion of any samples and SNPs impacted by errors to minimise bias and the potential for false positives to be reported in downstream results.

Genotype calling is the process of determining the genotype at a specific position and is based on the signal intensities of fluorescently labelled probes that correspond to the potential alleles at a given position. Given a site with three potential states (homozygous A, homozygous B and heterozygous), signal intensities of the allele specific probes can be clustered to represent said states through the use of genotype calling algorithms. In cases where signal intensity data is ambiguous and does not fit clearly into clusters, the genotype can be considered missing. High levels of missingness in samples are indicative of poor-quality DNA from those samples and so, applying minimum call rate thresholds is key to removing them from downstream analyses. Thresholds between 95% and 99% are commonly applied, although the exact thresholds are study dependent and are based on determining a balance between sample quality and maximising the total number of samples used in the analysis. Another indicator of poor sample quality is excessive heterozygosity on the autosomes. Software, such as PLINK (Purcell et al, 2007) can calculate observed and estimated homozygous genotype counts and report an estimated inbreeding coefficient *F*. The inbreeding coefficient (F) is a measure of the probability that two alleles are inherited from a common ancestor i.e. identical by descent (IBD). In instances where excessive heterozygosity (indicated by low *F* statistics) is observed, there is a likelihood of sample contamination and so the sample should be excluded from further analysis.

Inconsistencies between reported sex and the predicted sex based on genetic data can bring to light potential errors that have occurred through mislabelling and mishandling of samples. A common approach is to estimate genetic sex based on X chromosome heterozygosity rates under the assumption that females will display heterozygosity for calls made on the X chromosome whilst males will display homozygous calls for genotypes on the X chromosome. Males, expected to display extreme homozygosity on the X chromosome (because they only carry one X chromosome), will deviate from the expected heterozygosity and register *F* statistics close to the value of 1. Conversely, values close to 0 are expected in females (because they carry two X chromosomes). Filters can then be applied to remove individuals displaying discordance between the inferred genetic sex and their reported sex.

Duplicate and related samples also present the potential for the introduction of bias into downstream analyses. Filtering for related samples requires an initial step of LD pruning. This is the process of creating of subset of SNPs, filtering out those that are in high LD as the detection of population stratification is most effective under the assumption of no LD. Suggestions for the best approach to the pruning process have pointed towards the removal of exceptional regions of high LD followed by applying pairwise r^2 thresholds of 0.2 – 0.3 over windows of 2Mb (Weale, 2010). Related/duplicate samples can then be detected through IBD calculations. This process assesses the number of loci at which two samples share either zero, one or two alleles. Duplicate samples/monozygotic twins, sharing two alleles at all sites will return an IBD score close to 1, with more distant relationships and unrelated samples approaching values closer to 0. Filtering duplicated samples is a simple process however, determining the threshold of relatedness is more complicated. A minimum threshold for IBD of 0.1875 can be recommended to remove second/third degree relatives (Coleman et al, 2016) but determining more stringent thresholds is based on the individual study, to once again minimise bias whilst retaining the maximum sample size.

Population stratification is the presence of differing allele frequencies between population groups due to non-random mating. In GWAS, the presence of population stratification leads to the potential for the detection of false positive association signals, as population sub-groups within the broader sample set display differences in trait/disease frequency. Therefore, associations detected may reflect inter-population allele differences as opposed to genuine association with the trait of interest because of confounding. When accounting for population stratification it is important to consider the presence of potential 'outliers' within a GWAS sample set. Methods such as EIGENSTRAT (Price et al, 2006) use principal component analysis (PCA) to capture genetic variation in genotype data across sample sets, reducing the genome-wide genetic data down to a small number of dimensions that summarise this variation. We can apply these methods to a sample set and plot individuals along the principal components. Due to the nature of genetic structure derived from population stratification we can expect populations to cluster together based on

ancestry groups. Used in tandem with the principal components (PCs) of a reference data set we identify clusters and their corresponding ancestry group. Applying this information to the sample set will allow for the identification of sample mismatches/mislabelling where reported ancestry groups do not match the estimated group based on genetic data, as well as the identification of individual samples that map outside of the main clusters formed by the sample set data. Outliers identified through these methods may act as a potential source of bias and so, can be removed from the analyses.

1.2.7 SNP quality control

In addition to sample QC procedures, a further series of SNP-based QC protocols is key to ensuring high quality analyses. SNPs can be filtered based on call rate thresholds, this process assesses the call rate of individual SNPs across all samples included in the analysis. SNPs that have a consistently low call rate across the sample set should be removed and are likely due to poor assays for that specific SNP.

Hardy-Weinberg equilibrium (HWE) states that in the absence of evolutionary influences, the allele and genotype frequencies within a population will remain constant through different generations. Observing extreme deviations from the HWE can suggest the presence of genotyping errors, stemming from the calling process. As described above, genotype calling algorithms use probe signal intensities to cluster genotypes into the three potential states (homozygous A, homozygous B and heterozygous). Poor genotyping can result in poorly defined clusters which register as deviations from HWE through the observation of excessive hetero/homozygosity. Software such as PLINK can test each SNP for deviation from the HWE and provide summary statistics in the form of *P* values, allowing for an appropriate threshold to be applied for filtering purposes. However, deviation from the HWE is not exclusively due to genotyping errors and can be the result of mutations or population stratification. For this reason, it is important to check for deviation from the HWE independently within any potential sub-groups that exist within the main sample set (e.g., for different ancestry groups that exist in the sample set) (Turner et al, 2011).

Traditionally, QC protocols on SNPs involved a degree of filtering based of minor allele frequency (MAF) thresholds. SNPs with low MAF occur less frequently within the population and therefore there is limited information available for said SNPs, introducing a greater degree of uncertainty when completing processes such as genotype calling. Further, GWAS are generally underpowered to detect associations stemming from SNPs of modest effect size with exceptionally low MAF. This suggests that the inclusion of these SNPs would not aid discovery of genetic association with a trait (Weale, 2010). Whilst these remain important considerations, they are in part addressed by the increasing sample size of cohorts used for GWAS. In short, by increasing the size of the population tested we will observe more occurrences of the minor allele at SNPs with low MAF, improving the quality of genotype calling for these SNPs. Additionally, developments in GWAS study design and post-GWAS analyses provide the opportunity to utilize SNPs at lower MAF ranges to help understand the genetic component of complex traits that have not been explained through commonly occurring variation within population groups. So applying filters based on MAF should be completed on a study by study basis, largely based on the quantity of samples available for analysis.

1.2.8 Association analyses

A standard GWAS aims to quantify the degree of association between variants in the relevant sample set and the phenotype/trait of interest by individually testing each SNP from the final data set for association with the target phenotype. GWAS operate on the assumption of a specific genetic model and how specific alleles contribute to the overall phenotype. GWAS are most commonly performed under the assumption of an additive genetic model. Assuming two alleles, *A* and *a* with three potential genotypes of *A*/*A*, *A*/*a* and *a*/*a*, an additive model for allele *A* assumes a linear change in log-risk or mean trait value with each occurrence of the *A* allele (Bush et al, 2012).

The statistical model used is dependent on the format/type of trait studied. In cases where the phenotype is binary, using a case/control study design, a logistic regression model will be applied. Alternatively, where the trait is measured in a quantitative manner, a linear regression model is applied. The GWAS performed in the analyses

covered specifically in this thesis utilises the generalised linear model using the '--glm' option in PLINK2 (Chang et al, 2015). This model is summarised as shown below.

$$y_i = \beta_0 + G_i\beta_1 + e_i$$

Equation 1.2.2

This model is fitted sequentially to all SNPs included in the analyses. Here, *y* represents the vector for the phenotype, and *G* to denote the SNP genotype (counting the number of A alleles carried at the SNP). β_1 denotes the effect of the SNP on the phenotype *y* and β_0 is the intercept term. The error term *e* denotes the noise, or the component of phenotype *y* that cannot be explained by SNP *G*. One additional factor to consider is covariate information and its inclusion within the association analysis. Variables such as the age and sex of individual samples can impact the outcome of association analysis through their association with the trait, independent of the genetic factors being studied. Therefore, this information is a necessity for samples included in the GWAS as these factors will need to be accommodated for in the analysis. Covariate information can be incorporated into the regression model to limit confounding factors and often includes PCs of the sample set to account for population structure.

1.2.9 Genome-wide significance and multiple testing

When performing GWAS, each individual SNP included in the analysis is tested for association with the trait of interest. This introduces the issue of multiple testing and the associated increased risk of type I errors (false positives) at a given significance threshold. However, setting an overly stringent significance threshold can reduce power and result in type II errors (false negatives).

A common method to account for multiple testing is through the Bonferroni correction, although this works under the assumption that each test is independent. This is not true for testing SNPs because, as described in section 1.2.1, SNPs are inherited in blocks of LD and thus, each individual test is not truly independent. To account for this, an adjustment of the Bonferroni correction based on the assumption

of ~1 million blocks of LD (Pe'er et al, 2008) led to the standard threshold of $P < 5 \times 10^{-8}$ for genome-wide significance for common variants.

1.2.10 PLINK

PLINK (Purcell et al, 2007) is an open-source whole genome association analysis toolkit that offers an extensive library of options and capabilities to support genome wide association analyses. Designed to cater to large data sets PLINK allows for support throughout the entire pipeline of a standard GWAS, supporting data management and preparation, quality control procedures, analysis of input data and association analysis. This section will touch upon the main features of PLINK and supplementary features used throughout my research. This section serves as an introduction to the software and is not all encompassing with specific functionality documented in the relevant methods sections throughout.

Data management forms one of the core utilities of PLINK with key functionality stemming from the '--recode' option provided, facilitating the conversion of data sets between common formats such as ped formats and VCF formats. Furthermore, PLINK provides an extensive number of options for manipulating and sub-setting data through user inputted parameters.

In addition to basic data management features, PLINK provides options for detailed analysis and extraction of summary statistics of genetic data sets. This is especially important for applying QC protocols at both sample and SNP level. We can extract data for features such as missingness, allele frequencies, and deviation from HWE. Output from these analyses can be used in tandem with data management options to clean data sets via extraction of samples/SNPs based on such features listed.

PLINK can be used to run a variety of basic association analyses but for the purpose of GWAS throughout this project we focussed on the basic case/control association test through the '--glm option using a linear model (specifically using PLINK 2.0).

1.3 Haplotype Estimation and Imputation

1.3.1 Introduction

In the context of genetics, imputation is the statistical inference of unobserved genotypes. Imputation is a key component of any GWAS as sample sets for GWAS will only provide limited coverage of genomic variation across constituent samples. Imputation expands the genomic coverage and facilitates analysis of a greater number of variants. This process is heavily reliant on the use of WGS data to establish known haplotypes within populations that are used as a reference for patterns of genetic variation across the genome (Scheet et al, 2006).

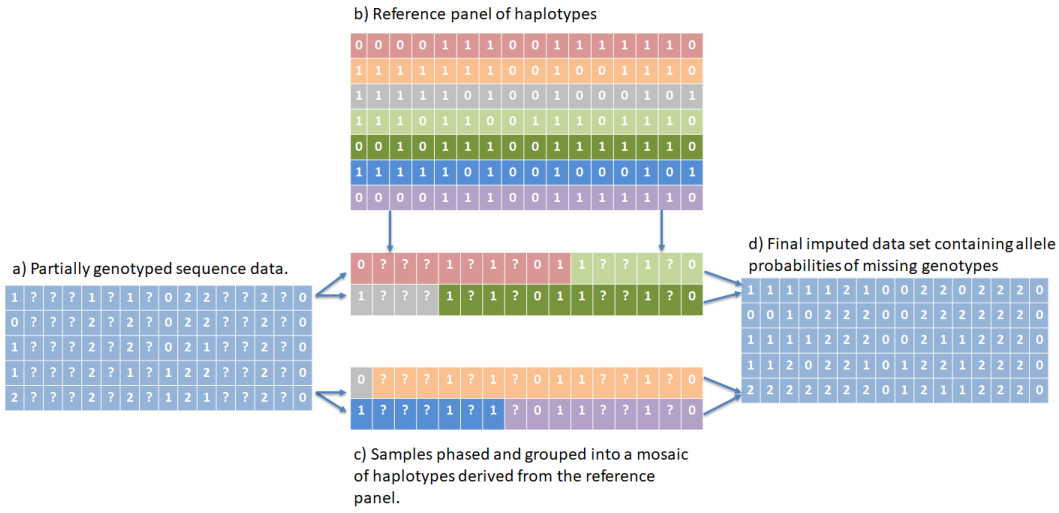

Figure 1.3.1: Simplified view of the haplotype estimation and imputation process adapted from Marchini & Howie 2010. a represents the initial GWAS sample set typed at a limited number of variants, where untyped variants are represented by '?'. b represents a reference panel of haplotypes typed at a much denser number of variants from which the mosaic of haplotypes in c is derived. d shows the imputation output with missing variants estimated based on the haplotype mosaic created in c.

Imputation methods identify the shared haplotypes between the sample data set and those found in a reference panel of haplotypes at a densely genotyped set of SNPs (figure 1.3.1b). Using the genotyped SNPs in the sample data set, imputation methods model the sample data set as a mosaic of haplogroups derived from the reference

panel (figure 1.3.1c). These combinations of haplogroups are then used to infer the untyped variants missing in the sample data set (figure 1.3.1d) (Marchini & Howie 2010).

1.3.2 Phasing

To facilitate the imputation process, GWAS sample sets first undergo the process of phasing. This requires appreciation of two key concepts, haplotypes and LD (see section 1.2.2) and how this impacts our understanding of patterns of inheritance. Phasing can be understood as the assignment of alleles to either paternal or maternal chromosomes i.e. the source from which the allele was inherited. In this context, as haplotypes are groups of alleles inherited together from a single parent, phasing refers to the process of haplotype estimation.

The methods used for phasing are dependent on the composition of the cohort to be phased, which is predominantly in reference to the relatedness between individuals in the cohort. As standard GWAS will typically involve large sample sets of unrelated individuals, this section will address the process of computationally inferring haplotypes in unrelated individuals. These methods work on the basis of pooling together information from across the sample set and estimating haplotype probabilities based on the observed genotypes within the sample set (Browning and Browning, 2011).

One of the most common methods of phasing unrelated samples is based on the use of approximate coalescent methods (to take into account recombination and mutation in the formation of new haplotypes) and hidden Markov models (HMMs). These are based on underlying “hidden states” (in this case the haplotypes that are being estimated) that cannot be directly observed and only inferred based on “known states” (the genotyped variants in the sample set). Using observed genotypes in the sample data set, the HMM can generate a pair of haplotypes compatible with the known typed variants. These estimates can be refined by repeating this process and generating a new pair of haplotypes with each iteration. After 20-100 iterations the

estimations can be compiled to construct a consensus haplotype (Browning and Browning, 2011).

HMM based methods of haplotype estimation have been implemented in a variety of software tools, such as SHAPEITv2. The underlying HMM used (as described in Li et al, 2010) to determine the joint probability of observed genotypes and the haplotype state is summarised below.

$$\boldsymbol{P(G,S) = P(S_1) \prod_{j=2}^{L} P(S_j \mid S_{j-1}) \prod_{j=1}^{L} P(G_j \mid S_j)}$$

Equation 1.3.1

In this application, unphased genotypes (G) are resolved into a sequence of mosaic template haplotypes (S). This model makes the following assumptions that *H* template haplotypes are genotyped at *L* loci and for each locus (*j*), H^2 possible states are assumed. $P(S_1)$ represents the probability of the prior mosaic state. $P(S_j \mid S_{j-1})$ represents the probability of transition between the mosaic states between SNPs *j* and *j-1*, which is dependent on the rate of recombination (recombination rates can be estimated from within the sample group or external estimates from resources such as the HapMap consortium can be used). $P(G_j \mid S_j)$ represents the probability of observed genotypes dependent on the current mosaic state (Li et al, 2010).

SHAPEIT (Delaneau et al, 2012) is a dedicated phasing/haplotype estimation software and provides an excellent option for the pre-phasing of GWAS sample sets prior to imputation. SHAPEIT requires an input of unphased genotypes split by chromosome prior to phasing as SHAPEIT requires chromosomes to be processed individually. Input files are automatically checked for missing data rates of 5% or greater and the presence of monomorphic SNPs. In addition, the user must submit a genetic map file containing information concerning the rate of recombination across different regions

of the genome, the HapMap phase II file can be downloaded directly from the SHAPEIT website.

SHAPEIT offers a variety of options to customise the algorithm and model parameters used by the software. The algorithm can be altered via the number of burn iterations use to identify the initial haplotype estimate, the iterations of the pruning stage and the total number of iterations during the main stage from which average haplotype estimates are determined. Further, a 'seed' option allows for the reproduction of a specific run, as without this option the results for identical input files may differ slightly due to the algorithm employed by SHAPEIT.

Model parameters can be adjusted by two key options, the number of conditioning states and the window size used in the run. These represent the number of states used when sampling pairs of compatible haplotypes underlying known genotypes, and the size of the region in which phasing is performed. As a default SHAPEIT uses 100 states per SNP to achieve a balance between accuracy and the computational time required to complete the run. The number of states can be increased to improve the accuracy of haplotype estimation, although the computational time required will increase in a linear fashion. Window size of the SHAPEIT model can also be adjusted based on the needs of the user but for the preparation of datasets to be used in GWAS a default window of 2 Mb is recommended.

Standard output is provided as the best haplotype estimations in the HAPS/SAMPLE format. These files provide the necessary pre-phasing information for imputation using IMPUTE2 software.

The EAGLE2 software (Loh et al, 2016) differs from other phasing software including SHAPEIT2 in its approach to the process. Traditionally, phasing software pools haplotype information from the sample set input provided by the user. This presents the issue of smaller sample sizes severely limiting the accuracy of the phasing process due to a lack of information from which haplotype estimations can be drawn. EAGLE2 differs in that instead of pooling information from the data set being phased, this

information is taken from a reference panel. By utilising a reference panel to phase data sets, the quality and speed of the phasing process is no longer dependent on the size of the sample being phased, leading to improvements when phasing smaller sample sizes (<100,000 samples).

Additionally, EAGLE2 further distances itself from previous HMM based methodology by using a combination of diploid genotypes in tandem with a set of conditioning haplotypes. A series of haplotype prefix trees are generated along the chromosome using the Burrows-Wheeler transform (Durbin, 2014). Diplotypes are then explored in computationally efficient manner, specifically targeting the diplotypes with the highest posterior probabilities. In comparison to previous methodologies, EAGLE2 offered a 20-fold increase in phasing speed (compared to SHAPEIT2) and a 10% increase in accuracy that was observed when phasing cohorts across a range of different ancestries (Loh et al, 2016).

1.3.4 Imputing genotypes at untyped SNPs

With the sample set phased, genotypes at untyped SNPs can then be derived from a reference panel. Reference panels are large collections of high quality WGS sequencing data that provide a resource of defined haplotypes and a collection of densely genotyped SNPs. The information that these panels provide can be integrated with GWAS data sets and used to identify recurring patterns of LD based on the higher density sequencing data provided in the reference panel.

Similarly to methods for phasing genotype data, imputation methods rely on extensions of HMMs. For example, if we consider the software IMPUTEv1 (Marchini et al, 2007) the fundamental process for the imputation of missing genotypes is based on the below model.

$$P(G_i|H,\theta,\rho) = \sum_{Z} P(\,G_i|Z,\theta),P(Z|H,\rho)$$

Equation 1.3.2

G_i is a vector for an individual's genotypes and is conditional on the reference set of haplotypes *H*. Assuming data for *L* SNPs in *K* individuals, we have both observed genotypes $G_{ik} \in \{0,1,2\}$, and missing genotypes. In this model *Z* denotes the hidden states (the pair of haplotypes copied from the reference panel), $Z = \{Z_1, \dots, Z_L\}$ with $Z_j = \{Z_{j1}, Z_{j2}\}$ and $Z_{jk} = \{1, \dots, N\}$. At SNP *j*, Z_j is the pair of haplotypes copied from the reference panel to the vector G_i. In this model, ρ denotes the fine scale recombination map (mapping the frequencies of recombination between sites across the genome) estimated from the HapMap phase II and θ represents a fixed mutation parameter. These parameters allow for the terms $P(Z|H, \rho)$ and $P(G_i|Z, \theta)$ to factor recombination and mutation events into the prediction of genotypes (Marchini et al, 2007, Marchini & Howie 2010).

Imputation software, such as IMPUTE2 (Howie et al, 2009) and MiniMac3 (Das et al, 2016), offer an assessment of imputation quality in addition to the imputed genotype data. This can be presented in the form of estimated squared correlation scores between the imputed and true genotypes in the case of MiniMac3 software or, in the case of IMPUTE2 this is presented with a scoring system unique to the software (IMPUTE2 provides an 'info score').

There are numerous factors that can impact imputation accuracy such as total sample size, quality of phasing and the choice of genotyping array, but the work in this thesis will place particular focus on two specific factors: the MAF of the imputed SNP and the choice of reference panel the sample set is imputed up to. Variants with lower MAFs are generally imputed with less confidence. The choice of reference panel is of particular importance. As the haplogroups and variants of the reference panel act as the source from which genotypes at the untyped SNPs will be imputed, factors such as the population groups incorporated into the reference panel and their relation to the sample data set will have a major impact on imputation quality. Additional factors such as the overall panel size, (i.e., the total number of haplotypes and variants included), and the degree of quality control applied to the GWAS sample set will all impact the quality of the imputation genotypes.

1.3.5 Software – IMPUTE2

IMPUTE2 is a freely available imputation and haplotype phasing software built upon the methodology described in Howie *et al*. 2009. IMPUTE2 distinguished itself in comparison to previous imputation methods by separating the process of phasing and imputation, process that had often been performed simultaneously. IMPUTE2 aims to estimate haplotypes at typed SNPs and then using this information, impute the alleles at non-genotyped SNPs.

IMPUTE2 offers two alternative approaches to the imputation process, deemed scenario A and B. Scenario A addresses instances where the entire study sample set is genotyped in a uniform manner, using a single genotyping platform. Scenario B however, applies to instances where the cohorts within a study differ in terms of the genotyping platform used. This involves the total sample set being broken down into various subsets. Consider a GWAS sample set comprised of group A and group B, where group A is subject to genotyping at a 1 million SNPs. Group B will be genotyped using an alternative genotyping platform, genotyping these samples at a subset of the SNPs in group A. In this scenario, both groups share a subset of SNPs for which they are both genotyped although group A are genotyped at a larger total number of SNPs.

In the following description of the IMPUTE2, process *T* refers to SNPs that are typed in both the reference panel and study sample set, whilst U refers to those SNPs that are present in the reference panel but absent in the sample set. Methodology for scenario A operates as follows; $H_R^{T,U}$ is defined as set of known haplotypes within the reference panel. H_R^T represents the known haplotypes in the reference panel for SNPs under the umbrella of T. Further, H_S^T represents this same set of SNPs (T) derived from the sample set and H_{Si}^T describing those for individual i. With this in mind, the preliminary step is for the software to make the first set of guesses for the set of haplotypes H_S^T. The secondary stage of the process is based upon the equation below.

$$\boldsymbol{Pr(H_{S,i}^{T} | G_{S,i}^{T}, H_{S(-i)}^{T}, H_{R}^{T}, \rho)}$$

Equation 1.3.3

$G_{S,i}^{T}$ is the individual's genotype at SNPs in *T*. $H_{S(-i)}^{T}$ is the current estimate for haplotypes for SNPs in *T* for all individuals excluding the individual noted by i. H_{R}^{T} as previously stated, represents the reference panel haplotypes for SNPs *T* and finally, ρ represents a population-scaled recombination map for the region. A two-step process of sampling new haplotype pairs for each individual based on SNPs in *T* through sampling from the conditional distribution described in equation 1.3.3, followed by imputing new alleles for SNPs in *U*, is repeated to complete the imputation process for all individuals. This conditional distribution is derived from the HMM common to imputation methods, and specifically building upon the HMM used in IMPUTEv1 (Marchini et al, 2007) and described in equation 1.3.2.

For imputation using IMPUTE2 the user is strongly recommended to prephase their sample genotypes using SHAPEIT2. In addition, the user must specify a reference panel to impute up to and provide a fine-scale recombination map, known haplotypes and legend file for the chosen reference panel.

1.3.6 Software – Minimac3

Minimac3 (Das et al, 2016) imputation software offers a computationally efficient method for imputing up to large reference panels whilst simultaneously retaining a high degree of accuracy. Similarly, to other available imputation software, Minimac3 is built upon the HMM, adapting their algorithm to maintain its performance in the presence of missing genotypes.

One method by which Minimac3 maximises its computational efficiency is through a concept of 'state space reduction'. This concept addresses the number of states of which the HMM will iterate over by using the similarities between haplotypes in a given space, reducing the number of iterations required of the HMM whilst outputting identical results where the state space reduction algorithm is not used.

Das et al (2016) assessed the computational efficiency of their methods in comparison to alternate imputation software including the previous iteration of minimac, minimac2 (Fuchsberger et al, 2014), IMPUTE2, and another HMM based imputation software, Beagle4.1 (Browning and Browning, 2016). They completed imputation into 100 individuals (European ancestry) up to several reference panels varying in total sample size, with a range of 1,091 to 32,390 samples. Across all panel sizes, minimac3 offered the shortest run times and lowest memory requirements, with the disparities in performance being exacerbated as the reference panel size increased. Imputation accuracy was also assessed across three MAF ranges (0.0001 – 0.5%, 0.5 – 5% and 5 – 50%). Across all panels and software choice, imputation quality was largely consistent, supporting minimac3 and its capacity to ensure high quality imputation at significantly lower computational cost.

1.3.7 Assessing Imputation Quality

The most commonly used method of assessing the quality of imputed variants is through the estimated squared correlation (r^2). The most effective way to assess the accuracy of imputation is to calculate r^2 between the imputed dosage and the true genotypes. However, as the true genotypes are typically unknown, imputation software will calculate an estimated r^2 between the imputed dosage observed and the estimated variance in alleles based on allele frequencies within the imputed data set. Imputation software such as Minimac3 (Das et al, 2016) use a model (as shown below) to estimate r^2 whilst considering the format of imputed variants that provide probabilities for genotypes.

$$\hat{r}^2 = \frac{\frac{1}{2n} \times \sum_{i=1}^{2n} (D_i - \hat{p})^2}{\hat{p}(1 - \hat{p})}$$

The above model details the estimated r^2 calculated by Minimac3. In this model n denotes the total number of samples being imputed, the alternate allele frequency is represented by $\hat{p}$, and D_i denotes imputed alternate allele probability at haplotype i

(Das et al, 2016). This model can be summarised as the variance of the imputed alleles over the variance of the alleles assuming perfect imputation. Overall, estimated r^2 metrics provide a good indication of imputation quality and r^2 thresholds are an effective method of filtering poorly imputed SNPs from an analysis. However, it is important to consider that these metrics are still estimated and therefore there will always be a degree of uncertainty especially when imputing particularly rare and low frequency variants (see section 6.2.3) (Pistis et al, 2015).

1.3.8 Analysis of imputed genotypes

The output from imputation is provided in the form of genotype dosage files. Imputed genotypes are not discrete variables and instead, imputation software outputs a probability distribution of possible genotype calls for each SNP in each individual. Therefore, rather than directly stating the genotype at any given point (e.g. common homozygote = 0, heterozygote = 1 and rare homozygote = 2), the imputation output will provide probabilities of the most likely genotype. When analysing imputed genotypes, simply considering the most likely genotype at each SNP based on the probabilities in the dosage data will introduce uncertainty into the analysis. Therefore, if we consider the regression model in equation 1.2.2, the genotypes G are replaced by dosages from the imputation output. These dosages provided the expected genotype call over the imputed probability distribution.

1.3.9 Fine mapping

Whilst basic association analysis can be used to identify trait associated loci, the SNP or variant that define these loci are not necessarily the causal variant (the variant of biological interest) and often show a strong association signal because they are in LD with the causal variant. Fine mapping is the process of exploring and refining these loci to identify the variants within the locus that exert direct influence on the associated phenotype/trait (rather than indirect association because of LD).

As the purpose of fine mapping studies is to determine the casual variant(s) at a locus, the level of SNP density at a locus will have a large influence on the quality of fine

mapping possible. Prior to applying fine mapping methods, it is important to ensure that all independent association signals within the locus have been fully described. This can be completed through a process of forward stepwise conditional analyses. This requires the lead SNP of a locus to be included in the regression model as a covariate and repeating the analysis of the region in question. If, after conditioning on the previous lead SNP of a locus, additional SNPs remain at genome-wide significance, the SNP with the lowest *p* value is added into the regression model. This process is repeated in sequential steps until no SNPs at the locus remain at genome-wide significance following the conditional analysis. This process identifies the independent signals of association within the locus and can provide evidence for the potential of multiple causal variants within a single region. Establishing these independent signals within a region is key to accurate fine-mapping as the process works on the assumption that each independent signal is driven by a single causal variant.

Standard SNP genotyping arrays used for GWAS will not provide a suitable level of SNP density to support extensive fine mapping, and alternatives such as WGS data generation are not a realistic approach for GWAS sample sets. Instead, imputing up to densely genotyped reference panels can provide a significant boost in SNP density to support fine mapping efforts. This can be through increasing the likelihood that all of the independent signals within a locus can be captured, and also increasing the likelihood of capturing the causal variants within the locus.

With knowledge of the independent associations within a locus, a common approach for fine-mapping casual variants is through the use of Bayesian methods to estimate probabilities that specific variants are causal. This approach, used in many fine mapping software tools determines the posterior inclusion probability of individual SNPs within the locus, outputting a ranking of SNPs most likely to be casual (Schaid et al, 2018)

1.3.10 Meta-analysis

One of the major benefits of improved imputation capabilities is the facilitation of meta-analyses. Meta-analyses aggregate information from multiple independent studies with the primary aim of improving the power to detect trait associated loci by increasing sample size, without the need for exchange of individual genotype and phenotype data. Naturally, different studies are likely to have genotyped their respective sample sets on different genotyping platforms. Thus, the different sample sets will be genotyped at different sets of SNPs. Imputation provides the methods by which these sample sets can be imputed up to the same reference panel to provide a uniform set of SNPs across which, all sample sets can be meta-analysed on. These analyses have been largely successful in further investigating established sample sets and in complex traits such as the development of colorectal cancer, they have enabled the detection of loci previously undiscovered in their composite studies (Al-Tassan et al, 2015).

1.4 Imputation reference panels

1.4.1 1000 Genomes Project

The development of the 1000 Genomes Project was undertaken with the intention of providing a resource to support efforts in understanding patterns of genetic variation across populations. Phase I of the 1000 Genomes project was released in 2012 and was comprised of 1,092 individual samples taken from 14 populations spanning Europe, East-Asia, sub-Saharan African and the Americas. Samples were subject to whole genome sequencing (low depth between 2-6x) and augmented with exome sequencing data performed at a much greater depth (50-100x). These efforts led to a dramatic increase in data availability in comparison to previous efforts such as the HapMap project, with a total of 38 million SNPs, 1.4 million indels and 14,000 large deletions included in the 1000 Genomes project pilot. The abundance of SNPs encompassed by the 1000 Genomes project provided much improved coverage at lower MAF ranges with ~50% of SNPs at 0.1% MAF from the UK10K WGS data included in the 1000 Genomes project. These data represented the pilot phase of the 1000 Genomes project with the final phase aiming to include a further 1,500

individuals sourced from an additional 12 new populations (The 1000 Genomes Project Consortium, 2012).

The completed 1000 Genomes project (phase III) included a total of 26 population groups, categorised into African ancestry (n=661), Americas (n=347), East Asian ancestry (n=504), European ancestry (n=503) and South Asian ancestry (n=489) with a final sample size of 2504 individuals. The updated cohort was subject to whole genome sequencing at a mean depth of 7.4x with targeted deep analysis of the exome reaching a mean depth of 65.7x.

In terms of the variants included in the final reference panel, the 1000 Genomes project was expanded to include multi-allelic SNPs (previously limited to bi-allelic SNPs) and a more comprehensive collection of structural variants (SVs). In summary 84,387,209 bi-allelic SNPs, 3,409,987 indels, 475,022 multi-allelic sites and 59,797 SVs for a total of 88,332,015 variants in the phase III iteration.

Haplotypes were constructed using a multi-stage approach, beginning with the creation of haplotype scaffolds using genotype array data from the Illumina Omni 2.5 and Affymetrix 6.0 microarrays included in the 1000 genomes project cohort, using the HMM based methods adapted by the SHAPEIT2 software (Delaneau et al, 2014). Bi-allelic variants and multi-allelic variants were processed separately with bi-allelic variants phased onto the scaffold via SHAPEIT2 and multi-allelic added onto the scaffold using the program MNVcall. The two scaffolds are then merged to produce the final realisation of the haplotype (The 1000 Genomes Project Consortium, 2015).

1.4.2 The Haplotype Reference Consortium (HRC)

The Haplotype Reference Consortium (HRC) (McCarthy et al, 2016) was established to create a conglomerate of as many whole genome and whole exome sequencing data sets as possible. The purpose of this was to develop the largest imputation reference panel to date, providing “a single centralized resource for human genetics researchers to carry out genotype imputation”. The HRC is comprised of 20

contributing cohorts supplying WGS data sets. The contributing cohorts are predominantly of European ancestry and low coverage WGS data. However, the inclusion of the 1000 Genomes project provides an exception with the most diverse collection of ancestries in terms of its constituent population groups.

Genotype calling was performed using a custom extension of the SNPTools (Wang et al, 2013) algorithm (GLPhase) to adapt the algorithm to work with extremely large sample sizes by incorporating the previously constructed haplotypes for each individual reference panel cohort in the process. Haplotype estimation was completed with SHAPEIT3 (O'Connell et al, 2016), built on the foundation of SHAPEIT2 but specialised to phase large scale data set such as those of biobanks.

Combining the 20 contributing cohorts, a total set of 95,855,206 variants was established. A series of filtering steps removed indels, SNPs with minor allele count (MAC) ≤ 5 and a QC protocol removed low quality variants to produce a final collection of 39,235,157 SNPs from 32,488 samples and 64,976 haplotypes.

1.4.3 Initial comparisons of imputation quality between publicly available reference panels

The 1000 Genomes Project, the Haplotype Reference Consortium, and their constituent phases and releases represent major steps in the development of publicly available whole genome sequence data sets. With each revision, the panels have displayed improved capability in terms of producing high quality imputation as they were augmented with additional sample and sequencing data across more diverse populations.

The incorporation of new technologies into panel composition and increased availability of data has allowed for large improvements in the number of haplotypes, variants, samples and population groups incorporated into each panel. These improvements in panel design have seemingly translated into better imputation quality, most notably at lower frequency variants.

The initial assessment of the 1000 Genomes Project's imputation capability was completed by excluding 9 – 10 individuals each from 6 different populations and imputing these individuals up to the reference panel. Imputation quality was assessed via squared correlation between the imputed and observed genotypes. Common variants displayed excellent imputation quality with an average squared correlation over 95%. However, the imputation quality suffered a drastic decrease when imputing at a MAF below 5% (The 1000 Genomes Project Consortium, 2015).

Internal assessments of the imputation capabilities of the HRC panel were completed using a GWAS sourced from the InCHIANTI study (Ferrucci et al, 2000). InCHIANTI consists of 1,154 individuals from Italy, and a subset of samples have been included as a component of the HRC. However, the imputation assessment made use of an additional 534 samples that were not included in the final HRC reference panel. The results of this initial assessment highlighted the improvements in imputation quality derived from the use of the HRC panel compared to 1000 Genomes Project panel, most apparent at lower frequency variants (McCarthy et al, 2016).

Iglesias et al (2007) completed a series of comparisons between the 1000 Genomes project (phase III) and the HRC panel. This included a comparison of meta-analyses of GWAS into vertical cup-disc ratio (VCDR) performed in the Rotterdam cohorts (Hofman et al, 2015) and The Erasmus Rucphen Family study (ERF)(Van Duijn and Zillikens, 2005). Initial comparisons within this study contrasted r^2 values across the different imputation outputs, concluding that mean r^2 values of imputed SNPs from the HRC imputation were higher than those from the 1000 Genomes project panel imputation (particularly at $0.001 < MAF < 0.01$ and $0.01 < MAF < 0.05$).

In addition to evaluating the imputation quality derived from each panel, it is important to consider how this impacts downstream processes, and whether we can observe an improvement in the results obtained from GWAS (i.e. identifying novel trait associated loci).

1.5 Thesis outline

The primary aim of this thesis is to assess how the choice of reference panel used for imputation impacts on imputation quality and downstream GWAS analyses. Primarily, this will investigate the relationship between the GWAS sample set and reference panel sample set in terms of the population groups from which samples are sourced.

Chapter two will introduce the relationship between the population groups individuals in the GWAS sample set are sourced from and the composition of the reference panel in terms of the population groups incorporated into the reference cohort, following with a description of how this relationship impacts the quality of the imputation. Major reference panels such as the HRC panel are predominantly composed of WGS data taken from individuals of European ancestry. Thus, it is important to investigate how the underrepresentation of non-European populations in commonly used reference panels has impacted the quality of imputation achieved for these groups, highlighting any potential effects this may have on GWAS performed on non-European populations. This chapter will begin with a description of current publicly available reference panels and their respective cohorts, followed by a review current literature that has addressed the topic of imputation quality when imputing from population groups underrepresented in reference panels. This section of the thesis will conclude with an assessment of four distinct population groups isolated from the Resource for Genetic Epidemiology Research on Aging (GERA) cohort (Kvale et al, 2015), imputing each group up to set of four different reference panels, and the imputation output derived from these imputations, contrasting the outcomes across population groups.

Chapter three introduces the concept of population specific reference panels and their application in GWAS. This section reviews the methods used in the design of population specific reference panels, the potential benefits for imputation quality and how these benefits translate to GWAS. The primary focus of this chapter is the design of 4 Japanese population specific reference panels, each using different amounts of WGS at varying depths to augment an established reference panel (1000

Genomes Project phase III). This is followed by an in-depth analysis of imputation output derived from each of these novel reference panels in comparison to the 1000 Genomes Project panel alone, concluding with recommendations of the novel reference panel that offers the greatest level of imputation quality and the reasoning behind this.

Building upon the conclusions made in chapter three, chapter four encompasses a GWAS of serum uric acid in the Japanese population. Utilising the finalised iteration of the Japanese population specific reference panel we investigated how the improvements in imputation quality observed from the output summary statistics translated into tangible benefits when performing GWAS. The association analysis was completed twice, once using the imputation output from the 1000 Genomes reference panel and again using the Japanese specific reference panel. We provide a comprehensive investigation into serum uric acid, comparing the capabilities of both reference panels in terms of redefining the lead SNP at known loci, the discovery of novel loci, and the capability for fine mapping at complex loci.

Chapter five focuses on the improvements of imputation at low MAF ranges offered by the Japanese specific reference panel. As most loci discovered through GWAS are defined by common variants and GWAS methods are typically underpowered to discover association signals driven by low frequency and rare variants, we used a gene-based approach to association analysis to identify novel sources of association. This serves to highlight the utility of the improved imputation of rare variants, facilitated by the Japanese specific reference panel. Finally, chapter six provides concluding remarks and a final summary of the research undertaken in this project.

CHAPTER 2

COMPARISONS OF IMPUTATION QUALITY ACROSS DIFFERENT POPULATION GROUPS WHEN IMPUTING UP TO WIDELY-USED REFERENCE PANELS

2.1 Introduction

2.1.1 Diversity of samples in reference panels

One of the major determining factors of imputation accuracy is the choice of reference panel and its composition in terms of its constituent samples and variants. Ensuring a degree of genetic similarity between the target population to be imputed and the population from which the reference panel is derived is one of the key factors in producing a high-quality imputation output. However, generation of WGS data has shown bias towards the sequencing of individuals from European ancestry population groups (Popejoy and Fullerton, 2016). Therefore, it is important to consider the impact this bias may have on publicly available reference panels and in turn, the effects on imputation accuracy when imputing from non-European cohorts.

If we consider three of the most widely-used genomics resources; The International HapMap Project, the 1000 Genomes Project, and the HRC as well as their respective phases and releases, we can observe general trends in improvements to reference panel design. As time has progressed, a particular focus in the design of these large resources has been on increasing sample sizes and the generation of high quality WGS to provide a detailed summary of genetic variation. Of the three resources, the HRC has been established as the largest reference panel with 32,470 samples in comparison to the 2,504 of the 1000 Genomes Project (phase 3). Assessments of imputation capabilities in comparison to the 1000 Genomes Project highlighted improved imputation accuracy (in a European ancestry cohort), particularly when imputing low frequency and rare variants (McCarthy et al, 2016). Further, these gains in imputation quality were shown to translate into novel associations when repeating

a GWAS of 93 circulating blood marker phenotypes (originally using the 1000 Genomes Project) utilising imputation output from the HRC reference panel. However, it is important to consider whether these improvements are consistent when imputing non-European ancestry cohorts and to quantify the difference in imputation quality when imputing up to smaller but more diverse reference panel (such as the 1000 Genomes project) and the larger but predominantly European ancestry dominated HRC.

2.1.2 Chapter aims

This Chapter will primarily focus on reference panel composition in terms of the populations from which samples are sourced, and the impact this has on imputation accuracy and quality when imputing from different population groups. I will introduce the Michigan Imputation Server as a platform for conducting imputation on remote servers up to several publicly available reference panels, as well as describing further publicly available reference panels integrated on the server in addition to the reference panels described in chapter one. This chapter will explore examples of currently published comparisons between major reference panels when imputing non-European ancestry cohorts, the difference in imputation quality/accuracy between these panels, and also contrast these results with comparisons imputing from European ancestry cohorts. Following this, the results in this chapter will present a comparison of imputation across four major ancestry groups (European, East Asian, African American and Latino populations) extracted from the Resource for Genetic Epidemiology Research on Ageing (GERA) cohort, imputing each cohort up to four separate reference panels. This comparison will primarily assess imputation up to two large reference panels (1000 Genomes Project and HRC) in addition to two further, smaller but population specific reference panels, the consortium on Asthma among African-ancestry populations in the Americas (CAAPA) and The Genome Asia (GAsPv1.0) reference panels. With these results I aim to contribute to further understanding of best imputation practices regarding reference panel choice in relation to the sample cohort to be imputed.

2.2 Methods

2.2.1 Resource for Genetic Epidemiology Research on Ageing (GERA)

The GERA cohort is comprised of 103,067 individuals participating in the 'Kaiser Permanente Research Program on Genes, Environment, and Health' (KP RPGEH). The cohort is representative of the population of Northern California, with individual members providing self-reported details for 23 unique categories of 'race, ethnicity, and nationality'. The 23 self-reported race/ethnicity categories were grouped into one of seven major groups: East Asian, Pacific Islander, Latino, African descent, White European, South Asian and Native American. This cohort provides a multi-ethnic genetic resource of a U.S. population with 80.8% of samples categorised as White-European and 19.2% of samples from the remaining six categories (Banda et al, 2015).

To complement the individuals to be sampled in this cohort and their respective ancestries, four distinct genotyping arrays were designed to maximise coverage of genetic variation from major population groups within the cohort. These arrays were developed for the European (EUR), East Asian (EAS), African (AFR) and Latino (LAT) groups previously described, and all genotyping arrays were designed to use the Affymetrix Axiom Genotyping Solution. The EUR genotyping array was developed first, and the design was guided with two key aims in mind; maximising the total number of SNPs included and prioritising the inclusion of SNPs with known trait associations. SNPs were categorised into levels of priority dependent on their known associations with traits/disease, in addition to their functionality as tag SNPs. Addressing the key aims required compromise, as SNPs deemed 'High priority SNPs' potentially require multiple probes or the presence of replicates to ensure that genotype calls are made with confidence. Thus, their inclusion reduced the total potential number of SNPs that could be included on the genotyping array. The final SNP count of the EUR genotyping array was approximately 674,518 SNPs, offering excellent coverage of targeted SNPs with more than 60% of SNPs with MAF ≥ 0.03 covered with r^2 of 0.8 (Hoffmann et al, 2011).

Following the initial design of a genotyping array to maximise coverage in specifically European ancestry populations, similar platforms were subsequently developed to genotype samples from East Asian, African American, and Latino population. Across the four arrays there is a substantial overlap of SNPs included in their designs. The EAS array was designed to incorporate coverage of EUR populations to account for individuals with mixed EAS/EUR ancestry. The AFR array primarily targeted SNPs relevant to West African ancestry as well as EUR ancestry, again to accommodate cases where there is shared ancestry. Finally, the LAT array was designed to account for a greater degree of admixture of ancestry groups, accounting for SNPs in EUR, AFR and Native American populations. Across the four arrays a total of 1,619,074 unique SNPs were included, of which 403,981 were present on two arrays, 156,270 on three arrays, and 254,438 SNPs representing common variation were present on all four genotyping arrays. These arrays all displayed a high degree of coverage of common variation in addition to more population-specific low frequency variation, with coverage of the AFR slightly lower than the EUR, EAS and LAT arrays (Hoffmann et al, 2011).

The preliminary cohort consisted of 110,266 individuals from which saliva samples were taken for DNA extraction. Following extraction and preparation of DNA samples (methods detailed in Kvale et al, 2015), samples were genotyped on the appropriate genotyping array based on self-reported ethnicity. A series of QC steps were undertaken to ensure high quality genotype calls across all arrays, with a total of 93.84% of all samples passing QC. The final post QC sample set consists of 103,067 individuals (Kvale et al, 2015).

2.2.2 Preparation of GERA cohort

Initial preparation of the GERA cohort entailed a preliminary series of QC steps. Samples were filtered based on a call rate threshold of <97% and SNPs using a call rate threshold of <95% (Cook and Morris, 2016). Further, SNPs that exhibited extreme deviation from the Hardy-Weinberg equilibrium (excluded SNPs with $p < 1 \times 10^{-5}$) were subsequently excluded. A series of LD pruned SNPs with $r^2 < 0.01$ overlapping the four arrays used to genotype GERA samples were used to estimate

IBD and in turn, for filtering of related individuals using a pi-hat > 0.2 threshold for exclusion. In each family set as determined by these IBD metrics the sample with the lowest call rate was subject to exclusion. Finally, the remaining SNPs were lifted to NCBI build GRCh37, removing any SNPs with unknown positions following this process. Additionally, SNPs with mismatched alleles in relation to the 1000 Genomes Project (phase 3) were removed from the final set. In preparation for imputation, a MAF filter of MAF <1% for each of the four genotyping arrays was applied and samples were pre-phased using SHAPEITv2.5.

Following QC, four random subsets (N=1000) were extracted from the cohort to establish a sample set for African American, Latino, East Asian and European population groups. Each subset was comprised of the same number of samples to ensure that we removed the effect of differing sample sizes when evaluating imputation quality. Using the '--PCA' option of PLINK v1.9, the top 10 principal components (PCs) were generated for the post-QC GERA cohort. Samples were plotted along PC1 vs PC2, PC2 vs PC3 and PC1 vs PC3, colour coded to indicate which of the four arrays the sample was genotyped on, based on their self-reported race/ethnicity data. Principal component analysis (PCA) was used to separate the GERA cohort based on each samples respective PCs and guide the estimation of clusters representative of the four major population groups. Sample sets of 1000 individuals were taken from clusters where self-reported and PCA estimated ethnic ancestry was concordant.

2.2.3 The Michigan Imputation Server

The Michigan imputation server (Das et al, 2016) is a freely available web-based imputation platform that provides researchers access to imputation up to several publicly available reference panels. Imputation of GWAS sample sets up to large reference panels is a computationally intensive process that requires the use of high-performance computing clusters alongside extensive storage capabilities. These resources can be costly, due to both initial set up costs and ongoing maintenance of the systems. In circumstances where computational resources are limited and not

suited to imputation, the Michigan imputation server provides researchers with an alternative solution to preforming imputation locally.

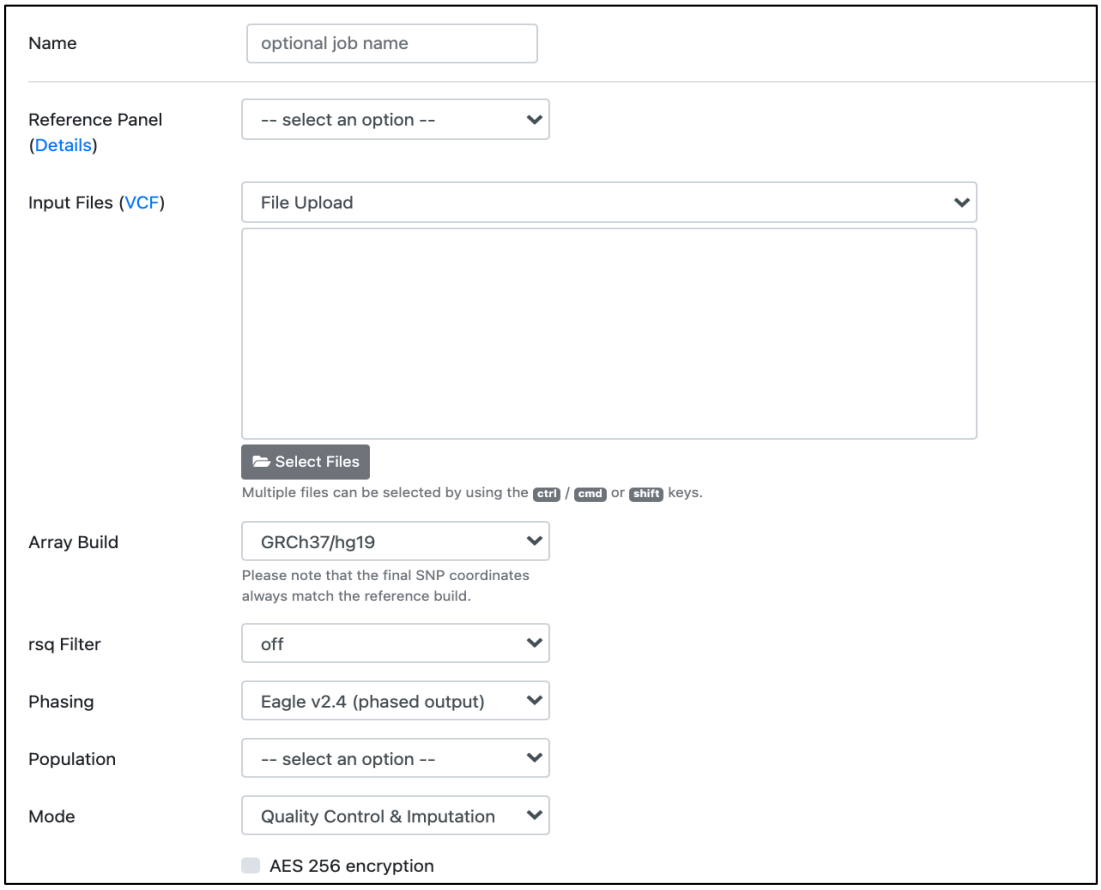

Figure 2.2.1: User interface for job submission to the Michigan imputation server.

The Michigan imputation server provides a simple user interface (Figure 2.1.1) that allows for a degree of user customisation when submitting jobs. Input files to be imputed are accepted in VCF format and can be uploaded to the server directly from local storage or through secure file transfer protocol (SFTP) via a remote server. A total of six reference panels are available for user selection (1000 Genomes Project phase I, 1000 Genomes Project phase III, CAAPA African American panel, Genome Asia Pilot, HapMap 2, and the HRC). Additionally, when imputing up to the 1000 Genomes Project or HRC reference panels, individual population groups within these panels can be specified to limit imputation up to the populations within the reference panel selected. Individual chromosomes must be prepared and uploaded separately, additionally the appropriate reference build (GRCh37/GRCh38) must be selected.

Further, users can determine the pipeline that best suits the data uploaded through the 'mode' option, which allows for an additional series of QC steps and phasing prior to imputation.

Once uploaded, chromosomes are split into 20Mb chunks and if the user has opted for the quality control option, will be processed through a series of QC checks prior to phasing or imputation. Initially each chunk is assessed for the number of variants that are also present in the chosen reference panel with a threshold of 50% required to process the chunk. Furthermore, sample call rate is assessed with a minimum of 50% of variants required to be called for each sample. Regarding variant QC, variants are excluded where invalid alleles or duplicates occur. Indels, monomorphic sites, SNPs with a call rate of less than 90%, and allele mismatches between the reference panel and study are also subject to exclusion. If unphased data is uploaded to the server phasing is performed prior to imputation via Eaglev2.4, processing the 20Mb chunks with a 5Mb overlap.

Imputation is carried out using Minimac4 (Das et al, 2016) the latest version of the Minimac imputation software and successor to Minimac3 (section 1.3.8). Minimac4 offers a greater degree of computational efficiency with imputation up to the 1000 genomes panel six times faster compared to Minimac3 and imputation up to the HRC two times faster compared to imputation with Minimac3. Imputation on the Michigan imputation server is completed using Minimac4 default parameters with the addition of the option 'allTypedSites'. This option includes sites that are genotyped in the sample set but absent from the reference panel in the final output. Following imputation, the output provided includes one dose.vcf.gz containing the dosage data for the imputed genotypes, and one info.gz file containing relevant summary statistics (e.g. SNP, reference and alternate alleles, r^2 score, MAF) for each chromosome submitted with the constituent chunks of each chromosome merged prior to the output being made available.

2.2.4 CAAPA reference panel

The Consortium on Asthma among African-ancestry populations in the Americas (CAAPA) was established to investigate the genetic contributions to asthma in populations of African ancestry within North, Central and South American, and Caribbean populations, addressing the issue of limited representation of these population groups within publicly available reference panels (Daya et al, 2019).

WGS data for 883 individuals (~30x depth) sourced from a total of 19 studies by CAAPA was generated on the Illumina HiSeq 2000 platform and used to establish a reference panel catered to population groups of African ancestry (Johnston et al, 2017). These data were made available on the Michigan imputation server, covering a total of 31,163,897 variants across the autosomes.

2.2.5 Genome Asia (GAsP) reference panel

The Genome Asia 100K consortium led further research to address the under representation of non-European ancestry groups in genomic data sets. Their pilot phase completed whole genome sequencing for 1,267 individuals (at an average depth of 36x) and incorporated an additional 596 whole genome sequences from publicly available data sets, for a total of 1,739 samples. These 1,739 samples, 80% of which originate from Asia, represented a total of 219 individual population groups, sourced from 64 countries. Whole genome sequencing for this cohort was completed following a standard Illumina protocol using the Illumina Hiseq 2500/4000 and X10 for sequencing. The pilot reference panel developed using this cohort entails a total of 1,654 of the 1,739 samples including 21,494,814 variants with multi-allelic site excluded from the reference panel.

An initial assessment of the reference panels utilised whole genome sequence data for samples originating from South, South East, and North East Asian population groups, limited to variants genotyped on the Illumina Global Screening Array v.1. These samples were imputed up to both the 1000 Genomes Project (phase III, all populations) and the novel GAsPv1.0 reference panels. Results from the imputation evaluation indicated that when imputing East Asian and South Asian populations up

to the 1000 Genomes Project panel, imputation accuracy was consistently and significantly below 90%. In contrast, imputing these same populations up to the GAsPv1.0 reference panel consistently achieved imputation accuracy in the range of 93 – 95% (GenomeAsia100K Consortium, 2019). The original GAsPv1.0 reference panel derived from the pilot study WGS data was made available on the Michigan imputation server for public use in 2019. However, as of 2021 an expanded reference panel has also been incorporated into the imputation server with an improved sample size of 6,461 individuals, establishing the GAsPv2.0 reference panel.

2.2.6 Imputation comparison

Each of the four sample sets (AFR, EUR, EAS and LAT) extracted from the GERA cohort were pre-phased separately using SHAPEITv2.5, and subsequently uploaded on to the Michigan imputation server. Each sample set was subject to four separate instances of imputation, this involved imputation up to the 1000 Genomes Project (Phase III), HRC, CAAPA, and GAsPv1.0 reference panels. Imputation output was retrieved from the Michigan imputation server and .info files were analysed to assess r^2 metrics for imputed SNPs (see section 1.3.6). Each imputation output was organised based on the MAF of the imputed SNPs in the dosage data. Eight MAF categories were established (MAF $\leq$ 0.01%, 0.01% < MAF $\leq$ 0.1%, 0.1% < MAF $\leq$ 0.5%, 0.5% < MAF $\leq$ 1%, 1% < MAF $\leq$ 2.5%, 2.5% < MAF $\leq$ 5%, 5% < MAF $\leq$ 10%, and 10% < MAF), placing imputed SNPs into one of eight categories within which their r^2 metrics were assessed. For each MAF category, the total number of SNPs was recorded, mean r^2 and mean MAF were calculated, and the number of SNPs with $r^2 > 0.3$ and SNPs with $r^2 > 0.8$ were recorded as raw totals and as a proportion of the total number of SNPs within the MAF category.

With these data, two separate comparisons of imputation quality were completed. The first comparison included the total output for each imputation, whilst a second comparison was limited to imputed SNPs present in all four reference panels. The second comparison allows for a direct comparison of imputation quality between all four reference panels across the same set of imputed SNPs. Comparisons using only the total imputation output may skew results towards larger reference panels (in

terms of total SNPs) based on the metrics we assessed. The performance of each reference panel was assessed within each population group to determine which reference panel supported the highest quality imputation in each of the four populations. Following this, the performance of each reference panel was compared across population groups to determine if/how imputation quality differed between population groups.

2.3 Results

2.3.1 PCA of GERA cohort

To capture sample sets (N=1000) from AFR, EAS, EUR and LAT population groups, the top 10 principal components were calculated, and individuals were plotted along their principal components to cluster individuals based on their genetic ancestry. Figure 2.3.1 shows the full GERA cohort distributed along PCs 1 and 2, 1 and 3, and 2 and 3. As per figure 2.3.1, isolating sample sets from EAS and EUR populations was relatively straightforward in comparison to isolating groups of LAT and AFR due to the increased admixture in these samples. EAS and EUR sample set were predominantly separated based on the observations of PCA plots of PCs 1 and 2.

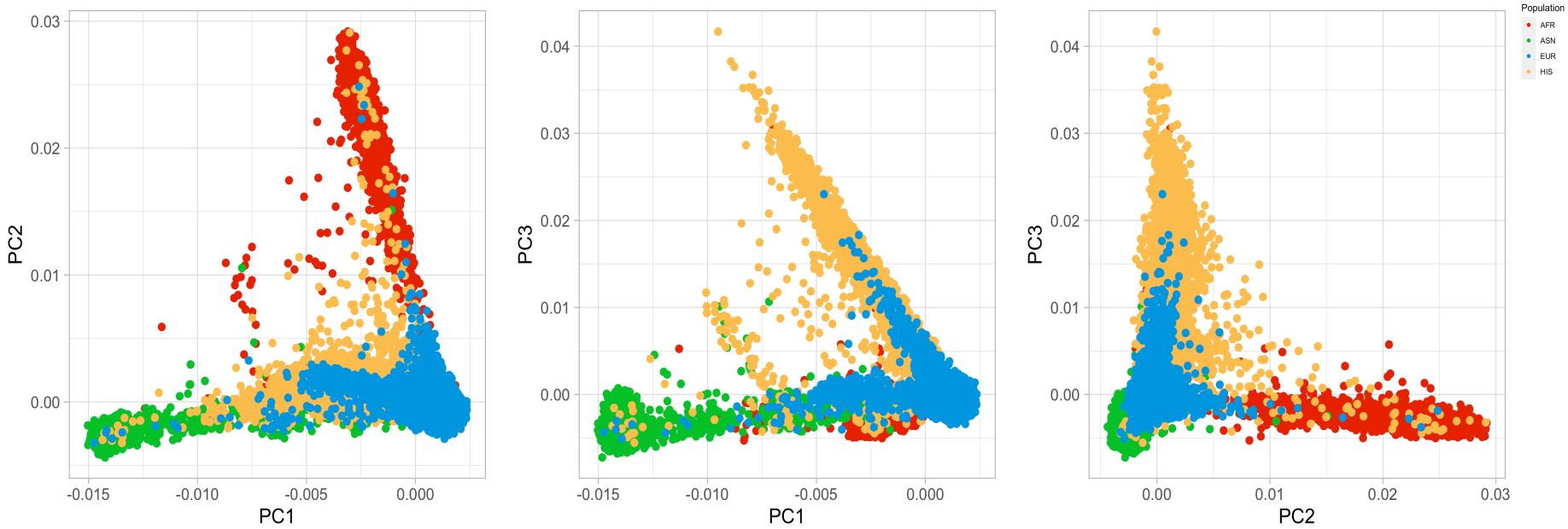

Figure 2.3.1: PCA plots for the complete GERA cohort, plotting PCS 1 vs 2, 1 vs 3, and 2 vs 3. Each data point represents a single individual and colour is indicative of the genotyping array used, determined by self-reported ethnicity. Yellow representing LAT, green for EAS, blue for EUR, and red of AFR respectively.

However, individuals attributed to LAT and AFR groups were relatively more complex with their distribution across the PCA plots. This is due to extensive admixture present in in AFR and LAT samples and so plots along PC1 vs PC3 in addition to PC2 vs PC3 were required to identify clear clusters for the AFR and LAT groups. Figure 2.3.2 highlights the overlap between the self-reported AFR and LAT members of the GERA cohort. Furthermore, whilst the GERA cohort is relatively diverse in regard to the population groups represented within it, the majority of samples were genotyped on the EUR array. Thus, the total number of AFR and LAT samples from which a subset of 1000 samples could be taken from was much smaller than that of the EUR category. Taking sample availability and admixture into account, the clusters from which the AFR and LAT subsets were created were less condensed in PCA plots in comparison to the EUR cluster (Figure 2.3.3).

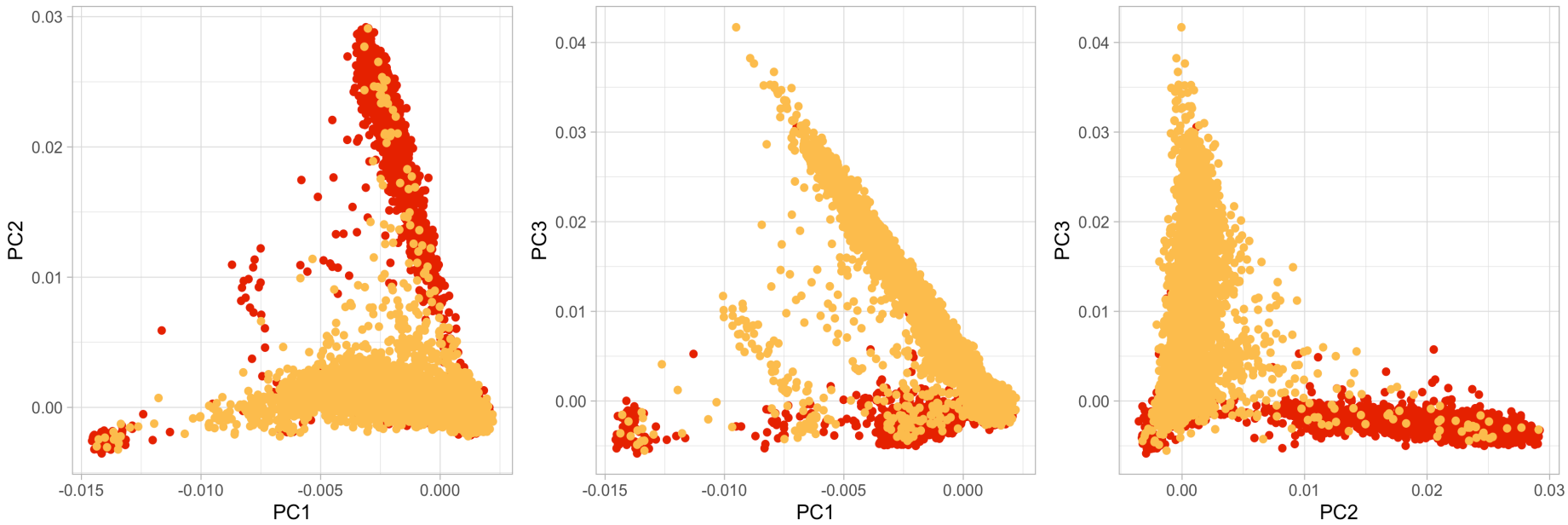

Figure 2.3.2: PCA plots for the self-reported AFR and LAT members of the GERA cohort, plotting PCS 1 vs 2, 1 vs 3, and 2 vs 3. Each data point represents a single individual and colour is indicative of the genotyping array used, determined by self-reported ethnicity. Yellow representing LAT, and red of AFR respectively.

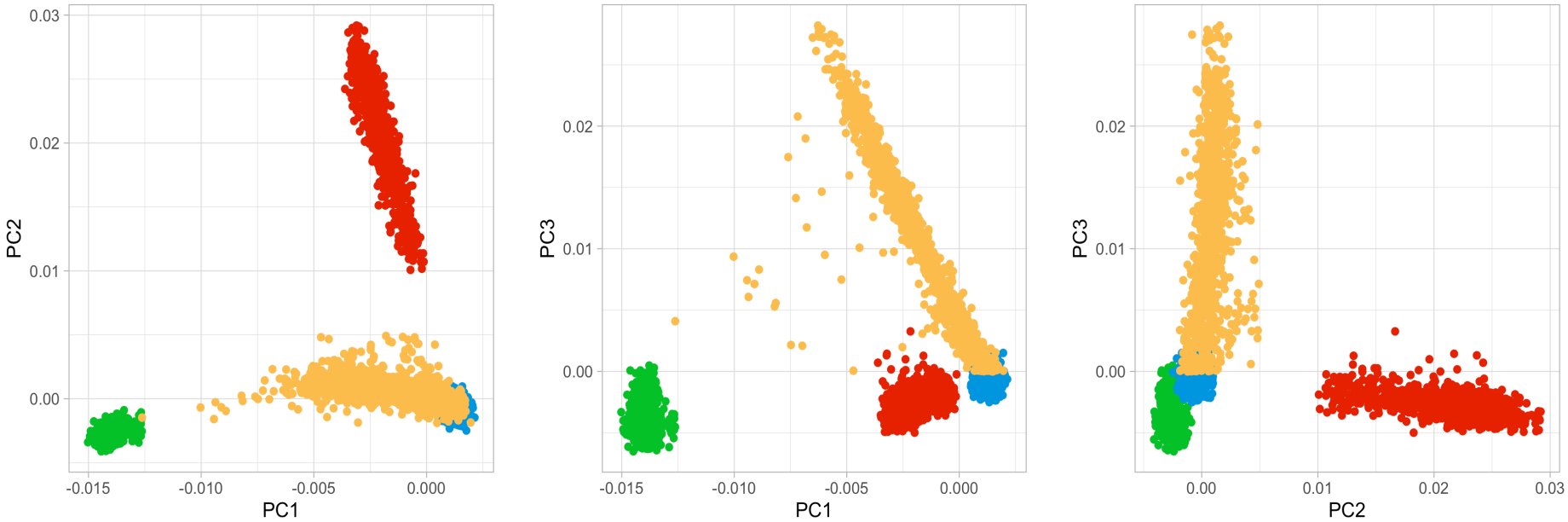

Figure 2.3.3: PCA plots 4 subsets of N=1000 extracted from the GERA cohort, plotting PCS 1 vs 2, 1 vs 3, and 2 vs 3. Each data point represents a single individual and colour is indicative of the genotyping array used, determined by self-reported ethnicity. Yellow representing LAT, and red of AFR respectively.

2.3.2 Imputation comparison for AFR subset

Each of the four reference panels used for imputation varied significantly in total sample size and the number of variants included in the reference panel (1000 Genomes N=2,504 and ~49 million variants, CAAPA N=883 and ~31 million variants, GAsPv1.0 N=1,654 and ~21 million variants, HRC N=32,470 and ~39 million variants). When imputing the AFR subset up to each of these reference panels and assessing the full imputation output we can see that the total number of SNPs imputed reaching a minimum imputation quality metric of $r^2 \geq 0.3$ (a common threshold to determine imputed SNPs viable for inclusion in results) is greatest in the 1000 Genomes Project output but remains comparable to that derived from the HRC (23,130,687 and 22,160,778 SNPs respectively). These totals are significantly larger in comparison to imputation up to the CAAPA (9,112,801 SNPs) and GAsPv1.0 (10,156,207 SNPs) panels (figure 2.3.4). However, if we are to look at well imputed SNPs meeting a more stringent quality threshold of $r^2 \geq 0.8$, there is a more significant disparity between the 1000 Genomes Project and HRC panels. In terms of high-quality imputation, 8,557,164 SNPs meet this threshold when imputing up to the HRC, a 55.6% increase relative to the 1000 Genomes Project output of 5,498,893 SNPs. Furthermore, imputation at lower MAF ranges (<1%) is consistently of higher quality (in terms of r^2 metric) when imputing up to the HRC. For example, in the MAF ranges of 0.01 – 0.1%, 0.1 – 0.5%, and 0.5 – 1%, imputing up to the HRC output 33%, 88% and 96% of SNPs, respectively, meet a minimum r^2 of 0.3. At these MAF ranges the imputation quality of the 1000 Genomes Project output was significantly lower with 11%, 57%, and 75% of the SNPs in these MAF bins attaining $r^2 \geq 0.3$. The CAAPA and GAsPv1.0 reference panels were comparable in terms of total numbers of SNPs meeting the r^2 thresholds of 0.3 and 0.8. Although, the CAAPA panel includes an additional ~10 million imputed SNPs relative to the GAsPv1.0 panel so the proportion of SNPs that meet these thresholds is overall lower when imputing up to the CAAPA reference panel.

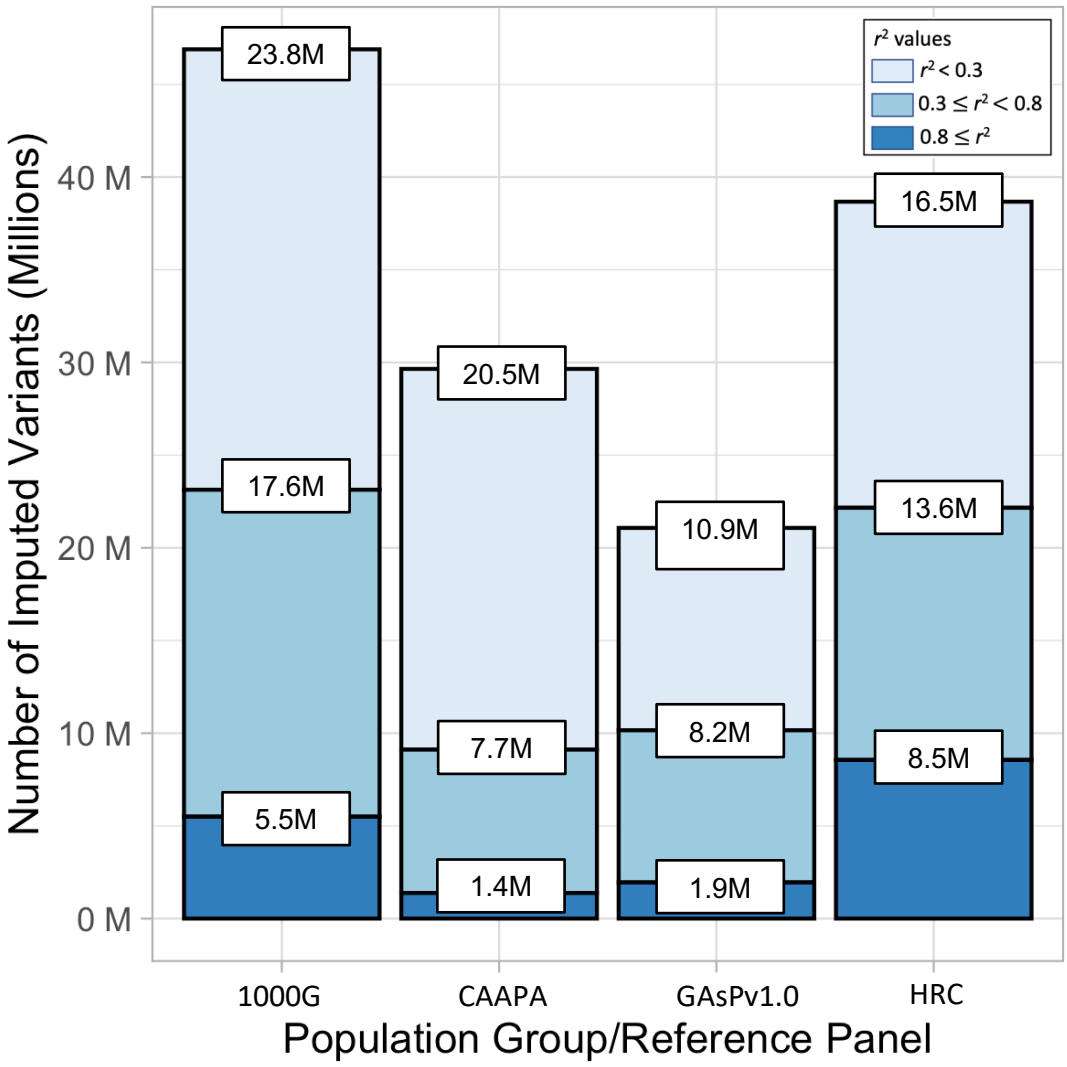

Figure 2.3.4: Total imputation output for the AFR subset imputed up to 1000 Genomes Project (1000G), CAAPA (African), GAsPv1.0 (Asian) and Haplotype Reference Consortium (HRC) reference panels. Output is categorised into r^2 ranges as per the key and total numbers of SNPs (in millions) are detailed for each category.

A comparison of imputation quality across a subset of 12.7 million SNPs present in the four imputation outputs provided a direct comparison of imputation quality across a shared set of SNPs. As per figure 2.3.5, imputation up to the HRC produces the highest mean r^2 values across the MAF spectrum, with the greatest difference in imputation quality at lower MAF ranges (especially at MAF <1%). Based on r^2 metrics the poorest quality imputation across this shared subset of SNPs stems from the

CAAPA reference panel. The output derived from this panel has significantly lower r^2 metrics across all MAF ranges in comparison to all other reference panel outputs.

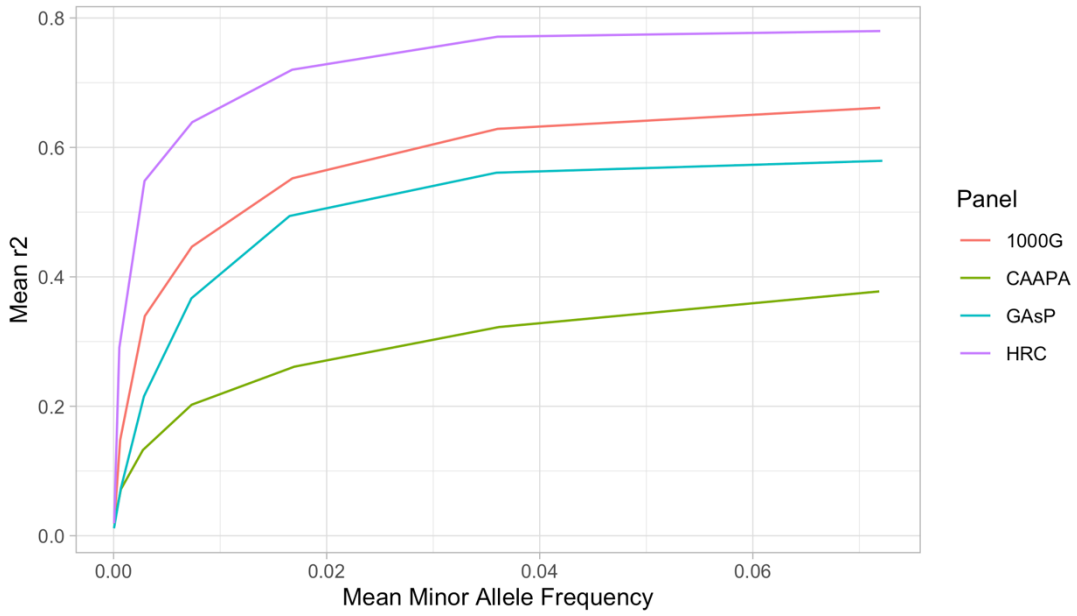

Figure 2.3.5: Mean MAF vs Mean r^2 plotted for a subset of 12.7 million SNPs shared across AFR imputation outputs from the 1000 Genomes Project (1000G), CAAPA, GAsPv1.0 and HRC reference panels. This plot places focus on rare and low frequency variants with an emphasis on those with MAF < 5% as this is where the majority of the imputed variants placed on the MAF spectrum and where the greatest disparities in imputation quality were observed between reference panels.

2.3.3 Imputation comparison for EAS subset

Figure 2.3.6 shows the imputation output from each EAS imputation up to the four reference panels and the quality (based on r^2 metrics) of their respective imputed SNPs. Imputation up to the HRC outputs the largest number of SNPs meeting the 0.3 (11,134,020 SNPs) and 0.8 (3,216,822) r^2 thresholds, slightly higher than the 1000 Genomes Project with 9,779,981 and 2,826,779 SNPs meeting these thresholds, despite the 1000 Genomes Project output containing ~8 million additional SNPs compared to the HRC panel. The difference in imputation quality is most apparent between these panels at a MAF < 1%. Across three MAF categories of 0.01 – 0.1%,

0.1 – 0.5%, and 0.5 – 1%, 15%, 52% and 69% of SNPs imputed in these categories have an $r^2 \geq 0.3$ when imputing up to the HRC. Conversely, at these same MAF categories imputing up to the 1000 Genomes Project, only 4%, 17% and 26% of imputed SNPs meet the minimum r^2 threshold of 0.3.

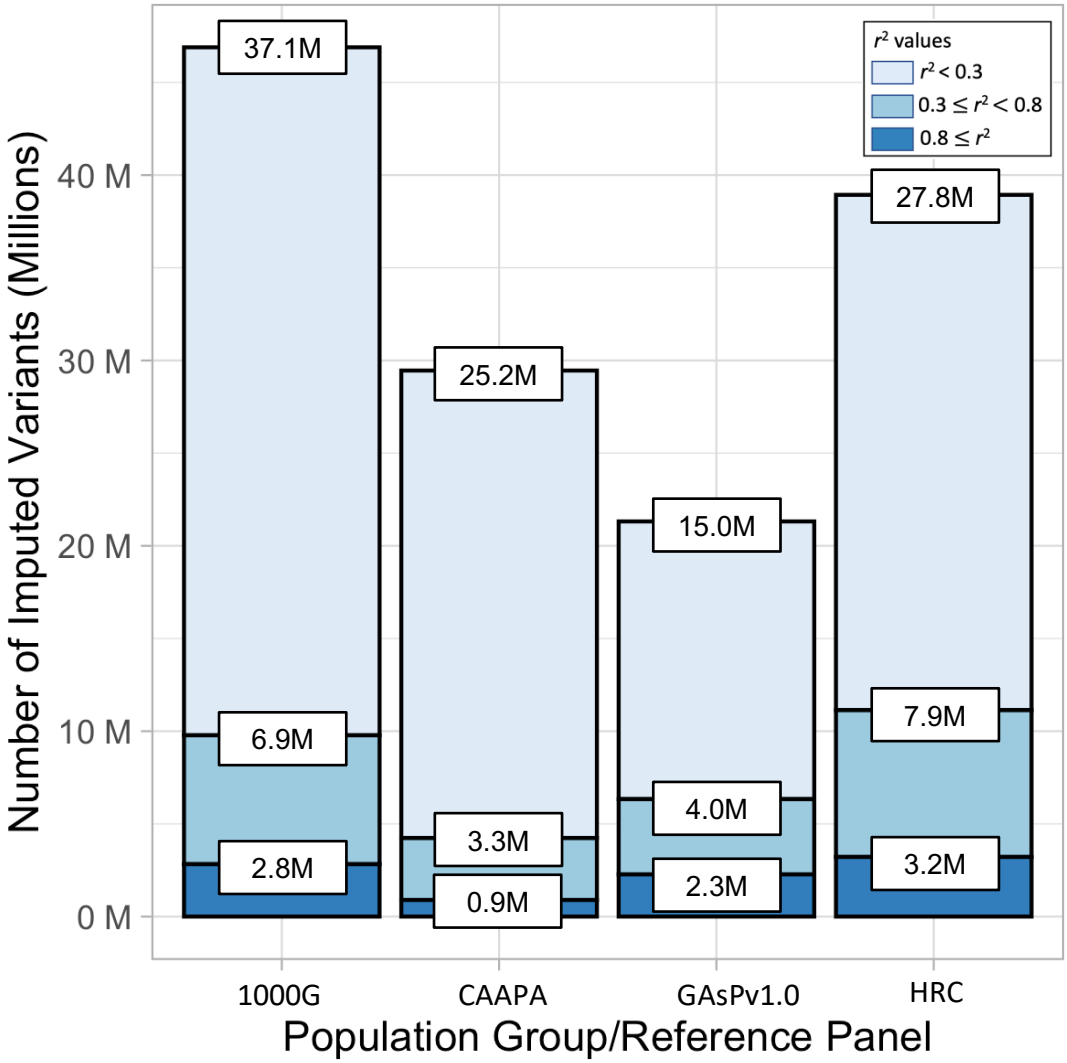

Figure 2.3.6: Total imputation output for the EAS subset imputed up to 1000 Genomes Project (1000G), CAAPA (African), GAsPv1.0 (Asian) and Haplotype Reference Consortium (HRC) reference panels. Output is categorised into r^2 ranges as per the key and total numbers of SNPs (in millions) are detailed for each category.

As made apparent in figure 2.3.6, the total number of imputed SNPs with $r^2 \geq 0.3$ is noticeably lower when imputing up to the CAAPA and GAsPv1.0 reference panels. When imputing from the EAS sample set, the CAAPA leads to the overall lowest

imputation quality with only 14.3% of all SNPs meeting a minimum of 0.3 r^2. In comparison, for imputation up to the HRC, 28.8% of SNPs met this threshold, 20.8% for the 1000 Genomes Project, and 29.7% for the GAsPv1.0 panel. So, whilst in terms of the overall number of SNPs with $r^2 \geq 0.3$ was relatively low for the GAsPv1.0 panel, as a proportion of the total SNPs included within the panel the GAsPv1.0 outperformed the HRC and 1000 Genomes Project by this metric.

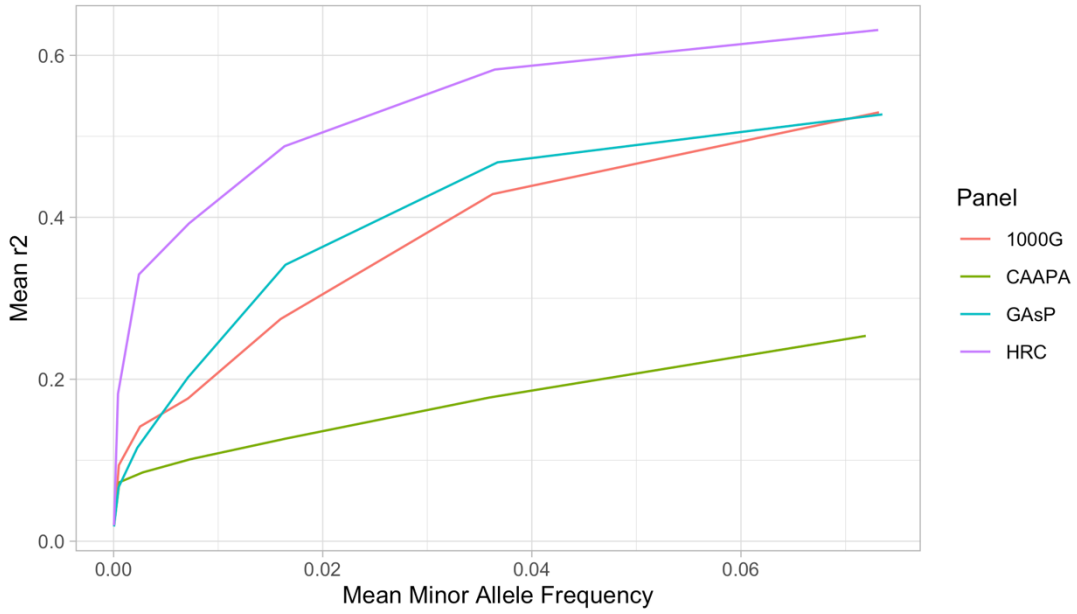

Figure 2.3.7: Mean MAF vs Mean r^2 plotted for a subset of 12.7 million SNPs shared across EAS imputation outputs from the 1000 Genomes Project (1000G), CAAPA, GAsPv1.0 and HRC reference panels. This plot places focus on rare and low frequency variants with an emphasis on those with MAF < 5% as this is where the majority of the imputed variants placed on the MAF spectrum and where the greatest disparities in imputation quality were observed between reference panels.

Across 12.7 million SNPs shared across the four imputation outputs, mean r^2 was consistently greatest for imputation up to the HRC. The most significant difference in mean r^2 generated across the four imputation outputs was at MAF < 1% where the HRC panels imputation quality is best relative to that of the remaining three reference panels. In the MAF range captured in figure 2.3.7 the GAsPv1.0 panel generally performs better than the 1000 Genomes Project in terms of mean r^2 values.

However, for more common variation above MAF of 5%, imputation up to the 1000 Genomes Project is of slightly higher quality based on mean r^2 values.

2.3.4 Imputation comparison for EUR subset

Similarly to both AFR and EAS sample sets, imputing into the EUR sample set shows a clear divide between the 1000 Genomes Project and HRC and the CAAPA and GAsPv1.0 reference panels when considering the total number of imputed SNPs with $r^2 \geq 0.3$. When imputing up to the 1000 Genomes Project, 26.2% (a total of 12,309,578 SNPs) of all SNPs meet the minimum r^2 threshold of 0.3, with 9.7% (total of 4,552,562) reaching the 0.8 threshold. Imputing up to the HRC however, results in a substantial improvement in imputation quality with 40.8% (15,897,924) of all SNPs with an $r^2 \geq 0.3$. Of these SNPs, 22.4% (8,730,659) meet the second threshold of $r^2 \geq 0.8$. Figure 2.3.8 highlights the contrast in imputation quality when imputing up to the CAAPA and GAsPv1.0 panels, both with a similarly poor capability when imputing in the EUR sample set. Imputation outputs from CAAPA (5,433,207 SNPs with $r^2 \geq 0.3$) and GAsPv1.0 (6,781,646 SNPs with $r^2 \geq 0.3$) provide considerably fewer well imputed SNPs, viable for use in further analyses such as GWAS.

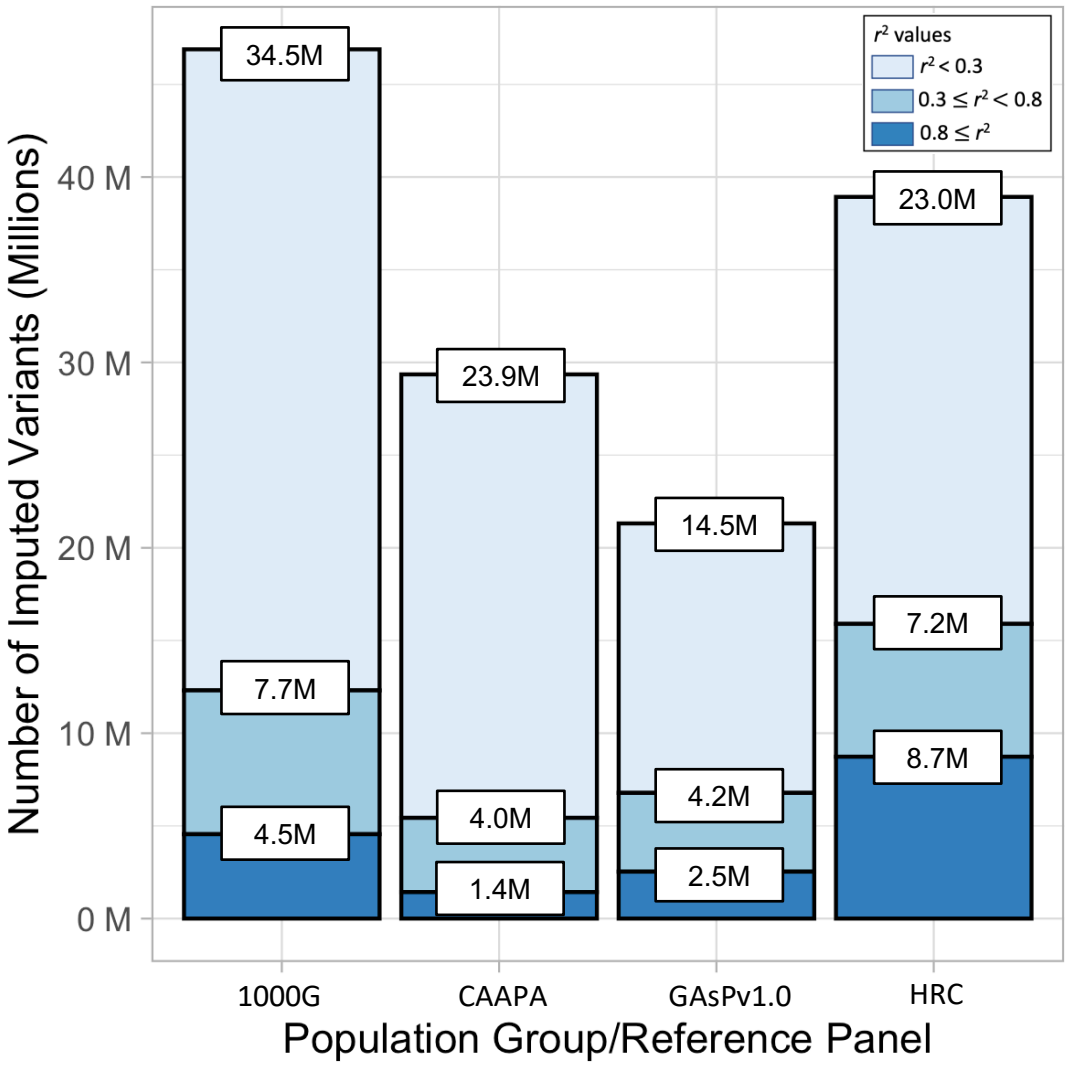

Figure 2.3.8: Total imputation output for the EUR subset imputed up to 1000 Genomes Project (1000G), CAAPA (African), GAsPv1.0 (Asian) and Haplotype Reference Consortium (HRC) reference panels. Output is categorised into r^2 ranges as per the key and total numbers of SNPs (in millions) are detailed for each category.

Figure 2.3.9 summarises the comparison across a shared subset of 12.7 million SNPs present in all four imputation outputs. There is an immediate distinction between the HRC imputation and the remaining three reference panels, with imputation at low MAF ranges markedly improved in the HRC imputation. For example, in the MAF range of 0.01 – 0.1%, 51.3% of SNPs are imputed with $r^2 \geq 0.3$ when using the HRC reference panel. In comparison, for the same MAF range, 12.3% of SNPs have an $r^2 \geq 0.3$ when imputing up to the 1000 Genomes Project, 6.5% for CAAPA, and 10.8% for

GAsPv1.0. Furthermore, when assessing the mean r2 for this subset of SNPs, imputing up to the HRC confers a mean r^2 of 0.56, higher than that of the 1000 Genomes Project imputation (0.43), CAAPA imputation (0.27), and GAsPv1.0 imputation (0.37).

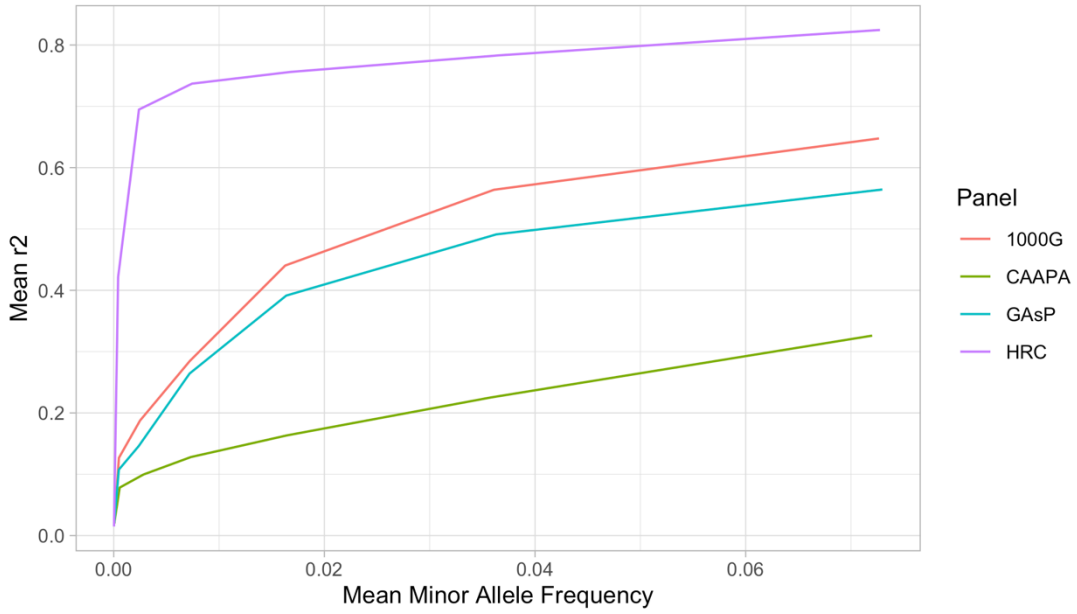

Figure 2.3.9: Mean MAF vs Mean r^2 plotted for a subset of 12.7 million SNPs shared across EUR imputation outputs from the 1000 Genomes Project (1000G), CAAPA, GAsPv1.0 and HRC reference panels. This plot places focus on rare and low frequency variants with an emphasis on those with MAF < 5% as this is where the majority of the imputed variants placed on the MAF spectrum and where the greatest disparities in imputation quality were observed between reference panels.

2.3.5 Imputation comparison for LAT subset

Imputing the LAT sample set up to all four reference panels produced similar results to that of the EUR sample set in terms of total imputation output and corresponding r^2 metrics. Imputation up to the HRC produced the best quality imputation in terms of total number and proportion of SNPs meeting the r^2 thresholds of 0.3 and 0.8, 19,489,770 SNPs (50.2%) and 8,622,804 (22.3%) respectively. Relative to imputation up to the 1000 Genomes Project this translates to a 16% increase in SNPs with $r^2 \geq 0.3$ and a 72% increase in SNPs with $r^2 \geq 0.8$. As with the remaining three population groups, there is a significant decrease in imputation quality across the total outputs

when imputing up to both the CAAPA and GAsPv1.0 reference panels. Imputation up to the CAAPA panel produces only 6,341,490 SNPs at r2 ≥ 0.3. Similarly, imputation up to the GasPv1.0 panel results in a total of 8,642,201 SNPs for the same r^2 threshold. Relative to the HRC panel, the CAAPA and GasPv1.0 outputs at $r^2 \geq 0.3$ are 32% and 44% of the total HRC output at this r^2 standard.

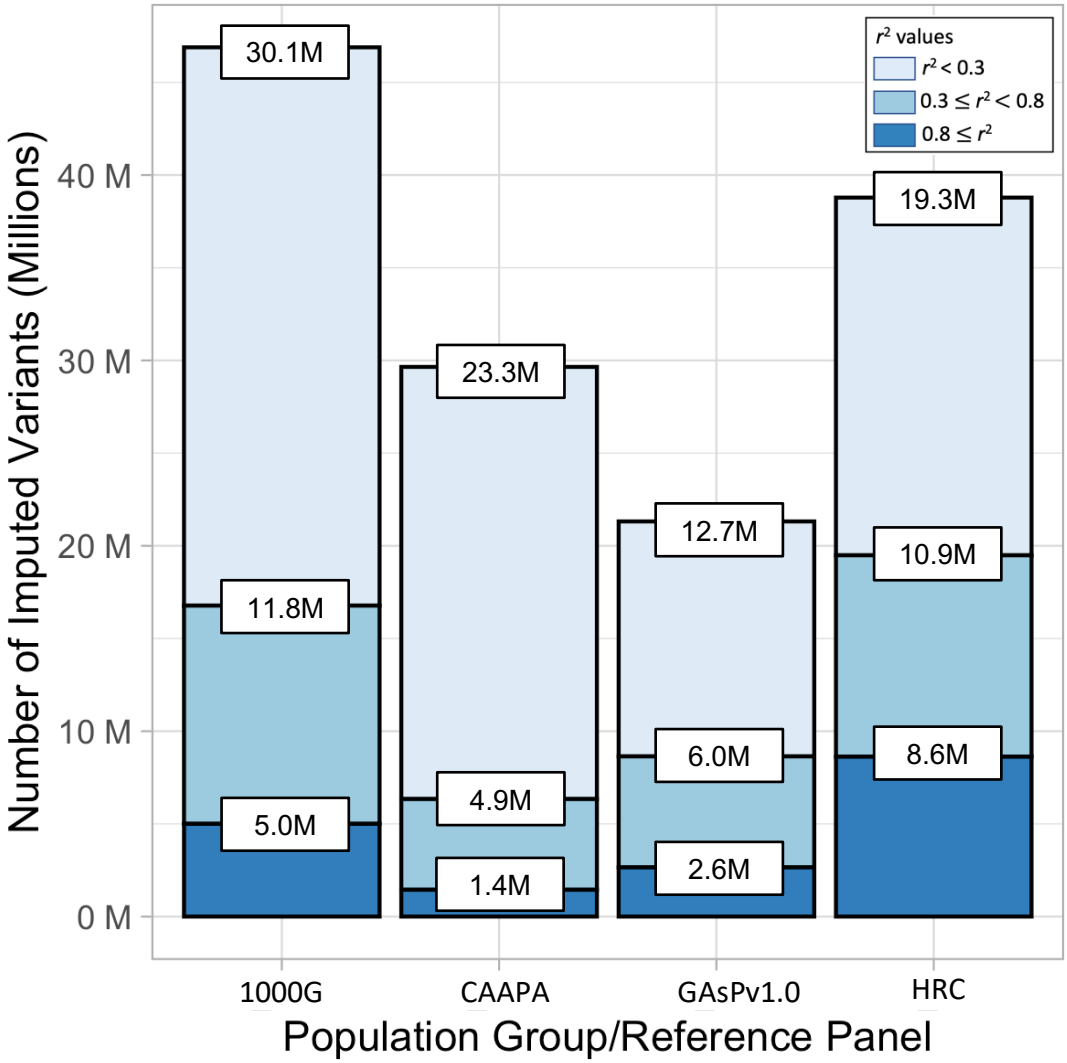

Figure 2.3.10: Total imputation output for the LAT subset imputed up to 1000 Genomes Project (1000G), CAAPA (African), GasPv1.0 (Asian) and Haplotype Reference Consortium (HRC) reference panels. Output is categorised into r^2 categories as per the key and total numbers of SNPs (in millions) are detailed for each category.

Figure 2.3.11 shows the mean r^2 values across the MAF spectrum for imputation of 12.7 million SNPs shared across all four reference panels. This difference in imputation quality again, is most apparent at low MAF ranges (MAF < 1%) and clear separation is apparent between reference panels regarding their imputation capabilities across this subset of SNPs. Imputation up to the HRC for the LAT sample set is consistently and significantly greater across the MAF spectrum with an overall mean r^2 of 0.567. In comparison the mean r^2 for imputation up to the 1000 Genomes Project panel is 0.435, with 0.279 when imputing up to the CAAPA panel and 0.373 when imputing up to the GasPv1.0 panel.

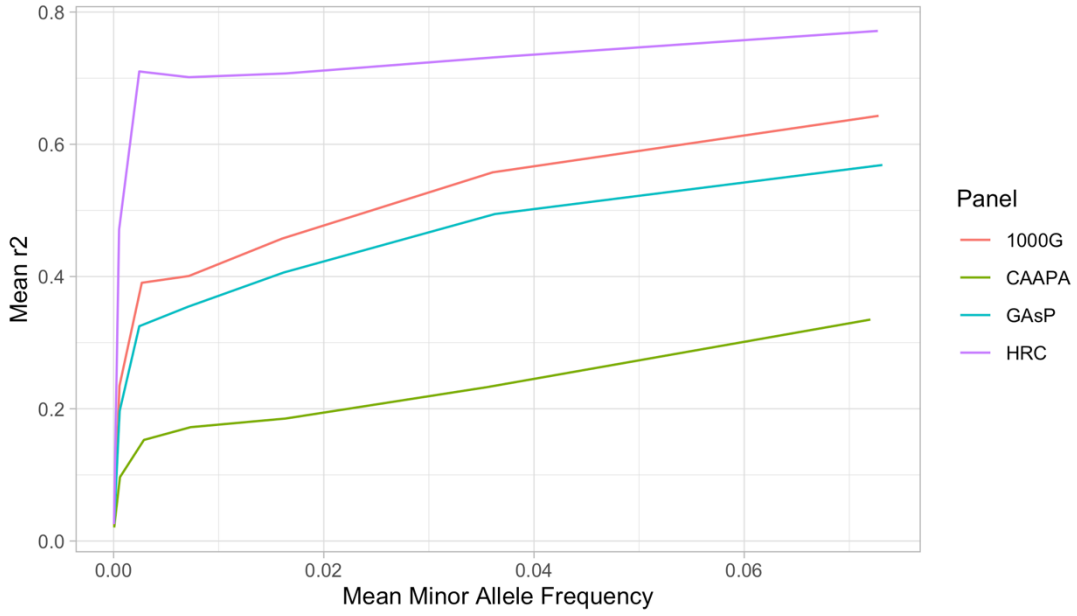

Figure 2.3.11: Mean MAF vs Mean r^2 plotted for a subset of 12.7 million SNPs shared across LAT imputation outputs from the 1000 Genomes Project (1000G), CAAPA, GasPv1.0 and HRC reference panels. This plot places focus on rare and low frequency variants with an emphasis on those with MAF < 5% as this is where the majority of the imputed variants placed on the MAF spectrum and where the greatest disparities in imputation quality were observed between reference panels.

2.3.6 Inter-population comparison across four reference panels

Initial comparisons considered the effectiveness of each reference panel to provide a high-quality imputation for the AFR, EAS, EUR, and LAT subsets. Following this initial endeavour, the comparison was expanded to assess how the imputation quality varied between the AFR, EAS, EUR, and LAT subsets. Figures 2.3.12 – 15 compare the mean r^2 values against mean MAF of each subset's imputation up to one of the four reference panels. These figures focus on imputation of rare and low frequency SNPs at low MAF ranges (MAF < 1.5%), as it is at these MAF ranges where the difference in imputation quality is most apparent between population groups.

Imputation up to the 1000 Genomes Panel was of the highest quality when imputing the AFR subset. Across all MAF bins, mean r^2 values were highest for the AFR subset, followed by the LAT, EUR and EAS subsets in that order. Mean r^2 values for the MAF range plotted in figure 2.3.12 were higher for the LAT subset in comparison to the EUR subset. However, for MAF > 2.5% mean r^2 values for the EUR subset were greater than those for the LAT subset. Imputation for EAS was consistently lower in comparison to the AFR, EUR, and LAT populations.

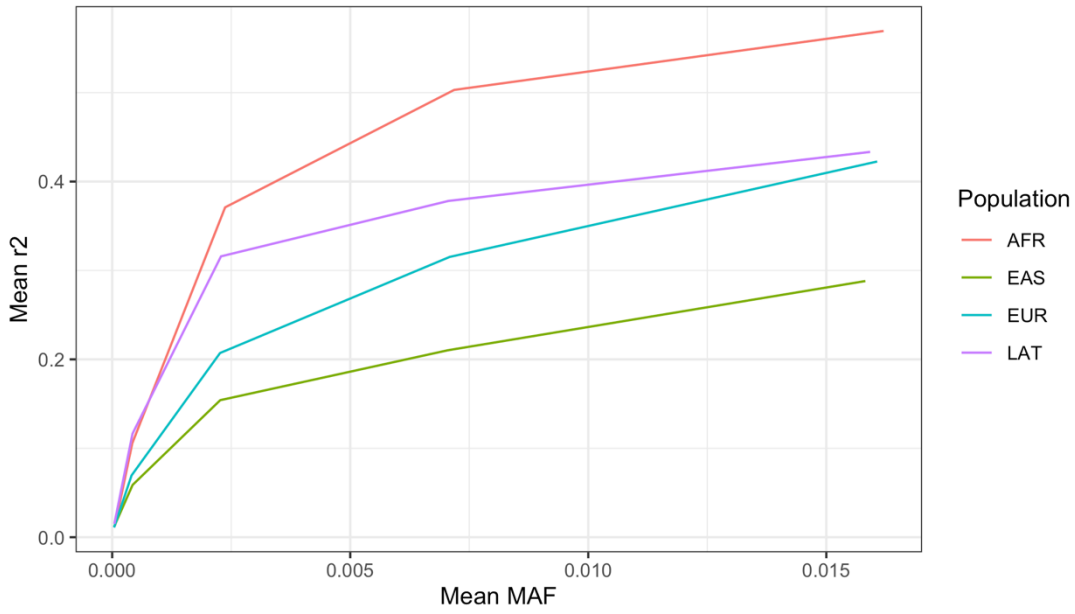

Figure 2.3.12: Mean MAF vs Mean r^2 across total outputs for imputation of each subset up to the 1000 Genomes Project reference panel. This plot places focus on rare and low frequency variants with an emphasis on those with MAF < 1.5% as this is where the greatest disparities in imputation quality were observed between population groups.

As shown in figure 2.3.13, the same trends in imputation quality observed in the 1000 Genomes Project imputation are present in imputation up to the CAAPA reference panel. Again, imputation for the AFR is of the highest quality, followed by the LAT, EUR and EAS subsets. However, imputation quality overall is significantly lower for all population groups in comparison to the 1000 Genomes Project imputation but in this case, there is a greater difference in overall mean r^2 between the AFR subset and the remaining three subsets.

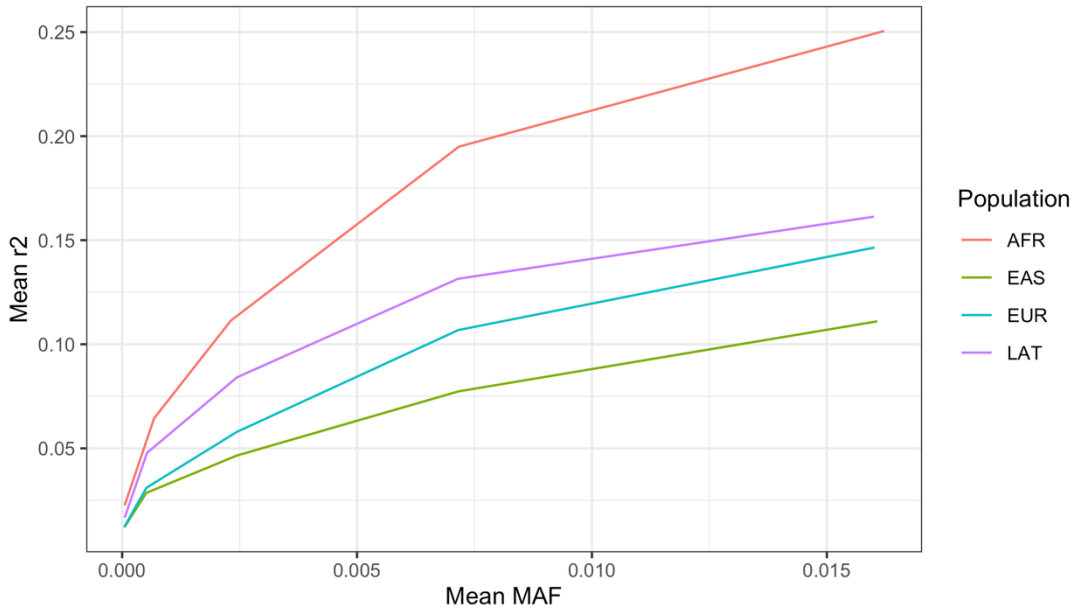

Figure 2.3.13: Mean MAF vs Mean r^2 across total outputs for imputation of each subset up to the CAAPA reference panel. This plot places focus on rare and low frequency variants with an emphasis on those with MAF < 1.5% as this is where the greatest disparities in imputation quality were observed between population groups.

Results from imputation up to the GAsPv1.0 panel differ from trends observed for the 1000 Genomes Project and CAAPA panels. At MAF of 0.5% and below, imputation quality is best when imputing the LAT subset. For example, in the MAF range of 0.1 – 0.5% mean r^2 for the LAT subset is 0.191, 0.130 for AFR, 0.098 for EUR, and 0.112 for EAS. However, at a MAF > 0.5% the best performing subset in terms of r^2 metrics, is the AFR subset. Whilst imputation quality for the EUR subset is lower in comparison to the LAT subset for MAF < 1.5%, mean r^2 values for variants with MAF > 1.5% are slightly higher in the EUR imputation output. Once again, imputation for the EAS subset is of the lowest quality based on mean r^2 but by a smaller margin in comparison to imputation up to the three alternative reference panels.

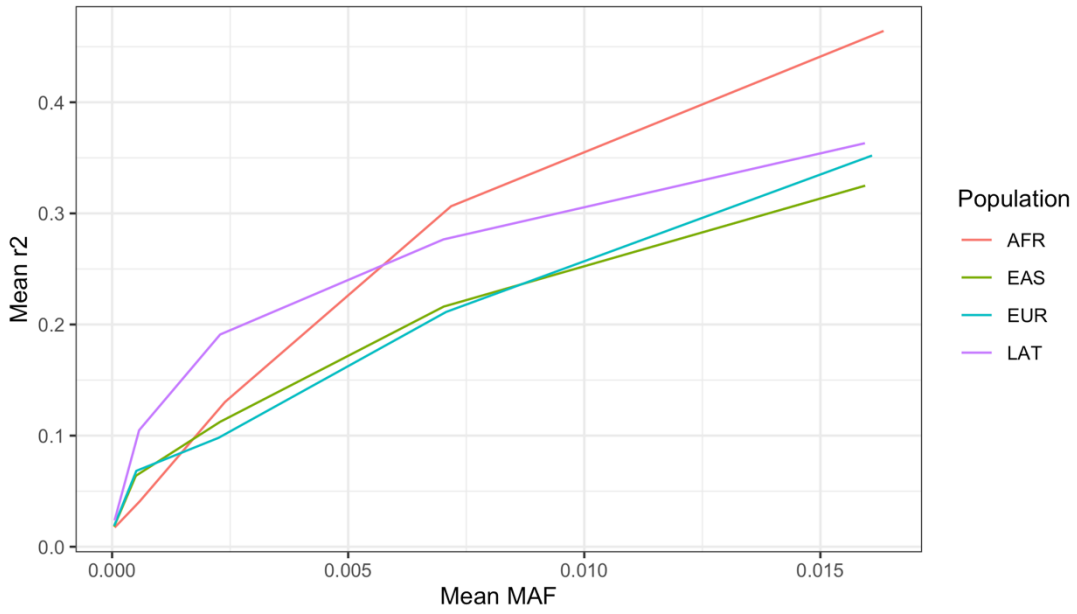

Figure 2.3.14: Mean MAF vs Mean r^2 across total outputs for imputation of each subset up to the GAsPv1.0 reference panel. This plot places focus on rare and low frequency variants with an emphasis on those with MAF < 1.5% as this is where the greatest disparities in imputation quality were observed between population groups.

Across all four subsets, imputation up to the HRC produced the highest quality imputation output relative to the 1000 Genomes Project, CAAPA, and GAsPv1.0 reference panels, based on the metrics described. However, imputation quality was not consistent across these population groups. Figure 2.3.15 shows the sharp increase of mean r^2 and mean MAF between the MAF range of 0 – 0.5%, and the contrast between the results for imputing AFR, EUR, and LAT subsets and the EAS subset. For variants with MAF between 0.1 – 0.5%, mean r^2 for the EUR subset is 0.661, almost double that of the EAS subset (0.338). Whilst imputation quality was best for the EUR subset acorss all MAF ranges, a similarly high quality imputaiton was achieved for the AFR and LAT subsets. The EAS imputation stands out in this respect as whilst imputing up to the HRC does provide an improvement in impuation quality compared to other reference panels assessed, the increase in mean r^2 values across the various MAF ranges is not comparable to that observed for the AFR, EUR and LAT subsets.

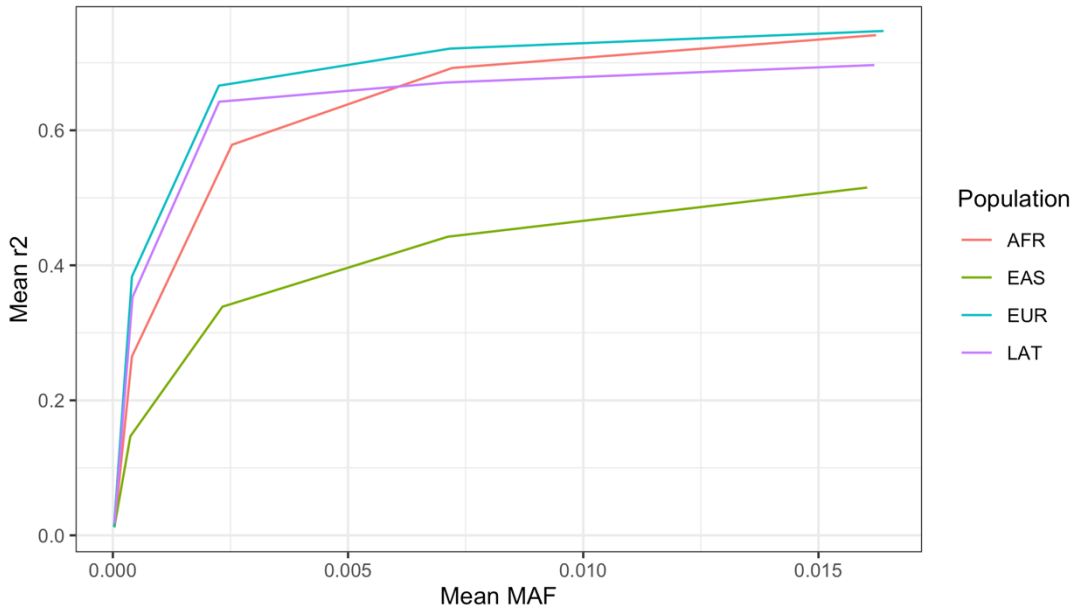

Figure 2.3.15: Mean MAF vs Mean r^2 across total outputs for imputation of each subset up to the HRC reference panel. This plot places focus on rare and low frequency variants with an emphasis on those with MAF < 1.5% as this is where the greatest disparities in imputation quality were observed between population groups.

2.4 Discussion

2.4.1 Imputation quality across population groups

Comparisons between reference panels shared a general trend across all four subsets in terms of imputation quality. Based on the metrics assessed and described in section 2.3, the HRC and 1000 Genomes Project reference panels produced the highest quality imputation output for all subsets. These reference panels represent the largest panels (in terms of both sample size and total number of variants) included in the comparison. Furthermore, whilst the HRC does display bias towards WGS data of European ancestry, these reference panels are not population specific in terms of their design and the samples they include. This contrasts with the smaller sample size of the CAAPA (West-African ancestry) and GAsPv1.0 (Asian population groups) reference panels, and the goal to source WGS data from specific regions/population groups.

Imputation quality derived from the CAAPA and GAsPv1.0 reference panels was significantly reduced in comparison to the two larger panels across all subsets. Improved imputation quality when imputing potentially complementary subsets up to these reference panels (AFR up to CAAPA and EAS up to GAsPv1.0) may be expected, but it appears that the imputation quality achieved may be severely limited by the size of these panels (CAAPA N=883 and GAsPv1.0 N=1,654). Furthermore, GAsPv1.0 reference panel sample set is encompassing population groups spanning the entire Asian continent (in addition to some European and African samples), including populations sourced from India, Pakistan and Russia, whereas the EAS subset is specifically East-Asian ancestry. Thus, the sample set may still be limited in terms of the representation of individuals of East-Asian ancestry.

Imputing up to the 1000 Genomes Project panel proved to be more effective than the CAAPA and GAsPv1.0 panels for all subsets. The 1000 Genomes Project (N=2,504) is significantly larger than both the CAAPA and GAsPv1.0 but also represents one of the most diverse collections of WGS data in regard to the population groups it encompasses, including 661 samples with African ancestry (including samples from African American populations) in addition to 504 samples with East-Asian ancestry (The 1000 Genomes Project Consortium, 2015). Thus, the 1000 Genomes Project sample set offers, as a proportion of its total sample set, substantial representation of samples with genetic ancestry similar to that of the AFR and EAS subsets. This may offer explanation for disparity in imputation quality between the 1000 Genomes Project and the CAAPA/GAsPv1.0 reference panels, as the 1000 Genomes Project not only offers a greater total number of reference haplotypes but also similar amount of data specific to the AFR and EAS subsets.

Overall, the HRC provided improved imputation quality for all four subsets and their respective imputations. The improvement in quality was greatest in the EUR subset but AFR and LAT subsets displayed a similar degree of improved imputation quality. This is not true for the EAS subset that only exhibited minor improvements in imputation quality in comparison to the 1000 Genomes Project imputation. As the HRC is a conglomerate of existing reference panels it contains the data from the 1000

Genomes Project in addition to a plethora of further sequencing data predominantly sourced from populations with European ancestry. Therefore, the large improvement in imputation quality for the EUR group is expected due to the much larger sample size and additional sequencing data. The key issue to address however, is the difference between the AFR, EAS and LAT subsets and why the EAS subset exhibits only minor improvement in imputation quality and does not benefit from the additional samples of the HRC to the same degree as AFR and LAT subsets.

One key feature that may influence the imputation quality attained for each sample set is the degree of admixture within the different populations involved in this comparison. Comparisons of genome-wide ancestry estimates within U.S. population groups (Bryc et al, 2015) gave estimates for African, Native American and European genetic ancestry within African American and Latino populations. In this study, they estimated proportions of genome-wide European genetic ancestry within African American populations (24.0%) and Latino populations (65.1%). Coupled with the observations made during PCA (figures 2.3.1 – 3) of a high degree of admixture in AFR and LAT groups, in comparison to the distinct clusters formed for EAS and EUR populations, there may be more overlap of genetic ancestry between the AFR, EUR and LAT subsets. This may offer explanation for the large improvements in imputation quality in the AFR and LAT subsets when imputed up to the HRC, as the additional samples with European ancestry may offer more utility to the imputation of the AFR and LAT subsets compared to the EAS subset.

2.4.2 Previous imputation comparisons between publicly available reference panels when imputing non-European ancestry cohorts

Vegara et al (2018) investigated the quality of imputation when imputing cohorts of African ancestry up to three publicly available reference panels; 1000 Genomes Project, HRC and a reference panel of 883 WGS sourced from individuals of African ancestry (CAAPA reference panel). 3,747 African Americans from the HCV Genetics Consortium and COPDGene cohorts, genotyped on Illumina Omni Quad and Express arrays were imputed up to the reference panels, assessing imputation accuracy

across the three panels. Mean r^2 was consistently higher for the HRC output across the MAF spectrum and lowest for the CAAPA output, with the greatest difference in mean r^2 values for variants with MAF < 1%. However, when assessing quality based on the proportion of SNPs meeting an imputation quality of $r^2 \geq 0.5$, 62.3% of 29,850,712 variants, 56.4% of 46,626,711 variants, and 52.9% of 39,019,012 variants meet this threshold for the CAAPA, 1000 Genomes Project and HRC respectively. In terms of a direct comparison between total numbers of SNPs meeting their quality threshold, the 1000 Genomes Project imputation displayed the strongest performance despite its smaller sample size, even when accounting for the inclusion of indels that are absent in the HRC. When assessing a subset of imputed variants overlapping the three reference panels, performance between the 1000 Genomes Project and HRC was comparable whilst that of the CAAPA lagged behind with ~3 million variants less meeting the $r^2 \geq 0.5$ threshold (Vegara et al 2018).

Whilst all reference panels facilitated high quality imputation, the 1000 Genomes Project and HRC panels appear to allow for higher quality imputation in comparison to the CAAPA panel. The results described do not seem to identify a clear improvement in imputation quality across the metrics measured when imputing up to the HRC rather than the 1000 Genomes Project. This suggests that the much larger sample size of the HRC does not have the expected impact on imputation quality on the cohorts imputed in the study completed by Vegara et al. However, the results reported in section 2.3 indicate a higher imputation quality for the AFR cohort when imputing up to the HRC in comparison to 1000 Genomes Project. This may be linked to the use of customised genotyping arrays designed to maximise coverage for African American samples in the GERA cohort.

The topic of imputation quality in relation to non-European cohorts was additionally reported by Lin et al (2018), completing comparisons between the 1000 Genomes Project and HRC panels when imputing from two cohorts taken from Han Chinese populations, the Anhui cohort of 1,132 Han Chinese samples and the Beijing cohort of 2,042 Han Chinese samples, in addition to the 1958 British Birth Cohort (58 BC) of 3,000 samples of European ancestry. As expected, their results showed that imputing

the 58 BC cohort up to the HRC led to greater imputation quality (measured by r^2 metric) across the MAF spectrum and again, significant improvement was observed at lower MAF ranges. However, in the case of imputing Han Chinese cohorts, mean r^2 across the MAF spectrum was consistently lower when imputing up to the HRC and, in fact for the Anhui and Beijing cohorts, imputation up to the 1000 Genomes Project panel led to observably higher mean r^2 values across all MAF ranges. Furthermore, mean r^2 values across MAF ranges were consistently lower for Han Chinese cohorts in comparison to the 58 BC. Summary statistics for the distribution of variants across r^2 bins showed greater imputation quality when imputing samples of European ancestry with 57.9% and 48.1% of imputed variants for the 58 BC cohort attaining r^2 between 0.8-1 when imputed up to the HRC and respectively. This contrasts with 41.5% and 44.0% of variants imputed up to the HRC and 1000 Genomes Project reaching this r^2 range for the Anhui cohort (Lin et al, 2018).

Studies into imputation quality/accuracy of Latino cohorts are relatively limited and do not address comparisons between the 1000 Genomes Project and the HRC as previous examples above. Latino population groups offer a greater insight into the capacity of these reference panels when imputing populations with a large degree of genetic admixture, as Latino populations specifically from the USA, display extensive admixture between European, West African and Native American ancestry groups (Bryc et al, 2015). Nelson et al (2016) however, addressed improvements in imputation quality when imputing Latino cohorts up to the 1000 Genomes Project phase I and III reference panels. A cohort 12,803 individuals sourced from Latino populations across four locations in the USA was imputed up to both phases of the 1000 Genomes Project and imputation quality/accuracy was assessed on chromosome 22 by comparing imputed and masked genotypes. Final results were as expected, and the increased sample size and diversity of the phase III panel led to improved r^2 metrics in imputed variants in comparison to phase I output (Nelson et al, 2016).

If we compare the results of these studies with those of our own analyses, the key difference appears to be in our assessment of imputation quality when imputing

African American and East Asian sample sets up to the HRC. In regard to the imputation of African American individuals, the improvement of imputation quality we observed when imputing up to the HRC as opposed to the 1000 Genomes Project, may be attributed to the difference in genotyping arrays used in these analyses. Vegara et al (2018) genotyped samples on the Illumina Omni Quad array and Illumina Omni Express array, whereas the African American sample sets derived from the GERA cohort were genotyped on a customised array, optimised for the African American population (Hoffmann et al, 2011). Similarly, whilst Lin et al (2018) reported a higher quality imputation when imputing up to the 1000 Genomes Project, our own analyses suggested that the HRC provided an improved imputation quality. This can potentially be linked to two factors, firstly, the sample sets used in Lin et al (2018) are specifically limited to Han Chinese samples whereas this analysis used a much broader category of 'East Asian', likely implicating numerous populations across the region leading to a more genetically diverse sample set. Secondly, the East Asian samples from the GERA cohort were again genotyped on a custom array, designed with the specific goal of maximising coverage when genotyping East Asian samples. The Han Chinese samples reported in Lin et al (2018) however, were genotyped on the Illumina Human 610-Quad array and the Affymetrix GeneChip Human Mapping 6.0 array.

2.4.3 Final Conclusions

This chapter assessed the imputation quality provided by two large reference panels using a diverse sample set derived from multiple population groups (the 1000 Genomes Project panel and the HRC), and two smaller reference panels using a more selective approach in their sampling methods to specifically capture variation in targeted population groups (the CAAPA panel and the GAsPv1.0 panel). Across all but one comparison in the four population groups assessed, the HRC and the 1000 Genomes project panels provided the highest and second highest levels of imputation quality respectively. Whilst the CAAPA and GAsPv1.0 reference panels were designed specifically for the imputation of African American and Asian population groups respectively, both the AFR and EAS cohorts achieved higher quality imputation when imputing up to the largest reference panel, the HRC. However, the

one exception to imputation quality trends observed was the performance of the 1000 Genomes project panel when imputing the EAS cohort. In this case, the GAsPv1.0 panel provided the second highest degree of imputation quality, most apparent at lower MAF ranges.

In addition to the four reference panels assessed in this comparison, the Michigan imputation server allows for imputation up to the more recently released TOPMed reference panel (Taliun et al, 2021). The TOPMed panel follows a similar structure to the HRC in that it is a conglomerate of existing WGS sourced from a variety of studies and cohorts to form an extremely large genomics resource, including 97,256 samples from diverse populations with a total of 308,107,085 variants across the autosomes and the X chromosome. Whilst this reference panel offers further developments in reference panel design, including greater representation of samples of non-European ancestry, the reference panel was not available during this imputation comparison. Thus, the work completed in this chapter was restricted to a comparison of imputation quality between the 1000 Genomes Project, HRC, CAAPA and GAsPv1.0 reference panels.

Based on the results observed in this series of imputation comparisons, the resources assessed exhibit a distinct decrease in their capacity to provide high quality imputation when imputing from the EAS subset in comparison to the other population groups included in the comparisons. Observations on genetic admixture in population groups were assessed, and composition of reference panel sample sets suggest that this could be linked to relatively poor representation of East Asian ancestry groups within current reference panels. Whilst imputation up to the HRC panel did exhibit an improvement in imputation quality in the EAS cohort, it was not to the degree observed in the remaining three populations. This suggests that the total number of samples in the reference panel does not appear to be the limiting factor in the imputation quality of the EAS cohort. Therefore, an additional focus on the representation East Asian populations within reference panels may support the improvement of imputation quality in East Asian GWAS cohorts.

CHAPTER 3

DESIGN AND DEVELOPMENT OF A JAPANESE POPULATION SPECIFIC REFERENCE PANEL USING WHOLE-GENOME SEQUENCE DATA

3.1 Introduction

3.1.1 Addressing disparities in imputation quality across population groups

The results described in Chapter 2 highlighted the disparities in imputation quality across different population groups and how this relates to the representation of said population groups within existing imputation reference panels. In particular, the imputation into individuals of East Asian ancestry produced the lowest quality imputation relative to AFR, EUR and LAT sample sets. This can be directly linked to the composition of these widely used reference panels and attributed to not only the relatively small sample size stemming from East Asian populations but also the limited diversity of the populations from which East Asian samples are taken. In comparison, samples of European ancestry in panels such as the HRC are sourced from a diverse collection of populations across the European continent whilst also representing the largest proportion of the total sample size of the reference panel. Population specific WGS data provides a unique opportunity to address this issue and this chapter will investigate the use of study specific WGS data alongside existing reference panels to improve imputation quality into underrepresented population groups.

3.1.2 Population specific reference panels

Large diverse publicly available reference panels are invaluable resources that have supported the adoption and development of association studies and their contributions to understanding the genetic component of complex traits. Imputation

up to these panels has proved successful in providing high quality imputation at common and to a degree, low frequency variants. Concerning imputation of rare variants, however, they are generally imputed with lower confidence and are thus limited in their utility in association analyses. This, as highlighted in section 2.3, is especially true for the imputation of cohorts comprised of population groups with limited representation within the reference panel of choice.

Even in regions, such as Europe, where the genetic differences between populations are relatively low, there exists distinct genetic structure linked to geographic origins of the population groups. Novembre et al (2008) analysed a subset of the Population Reference Sample (POPRES) cohort, comprised of 1,387 European individuals across 36 populations and genotyped on the Affymetrix 500K array (Novembre et al, 2008; Nelson et al, 2008). The POPRES subset was plotted along PCs1 and 2 to produce a 2-D map of genetic variation across the sample set. Their resulting map of genetic variation highlighted the correlation between genetic and geographic distances between the population groups, with the PCA plot showing a strong degree of similarity to a map of each sample's geographic origin (based on the central point of each populations corresponding country of origin). Their analysis of genetic variation among European populations showed that an individual's geographic origin can be estimated with a relatively high degree of accuracy. Using their multiple-regression-based assignment model, 50% of individuals were placed within 310km of their self-reported geographic origin and 90% within 700km of their self-reported geographic origin. With this in mind, and with the knowledge that one of the limiting factors of imputation quality is the genetic similarity between the sample set to be imputed and that of the reference panel, one approach to improve overall imputation quality and facilitate the confident imputation of rare variants has been to design novel reference panels specific to the population group of interest using WGS data. This approach considers that even though the imputation sample set, and reference panel samples may share ancestry to a degree (e.g., two European population groups), imputation quality can be further improved with addition of WGS sourced directly from target population group.

Early examples of such population specific reference panels were mostly based on European populations and include the Genome of The Netherlands (GoNL) (Deelen et al, 2014). The GoNL panel used WGS data for 769 individuals from the Netherlands at an average depth of 14x. Imputation quality was compared across three equal size (N=745) sample sets, representing Dutch, British and Italian (Milan, Rome and Naples) populations, with each individual sample genotyped on both the HumanHap550 and ImmunoChip platforms. The genotype data from the HumanHap550 array was imputed up to the 1000 Genomes project (Phase I, European ancestry samples, N=379), GoNL (A subset of N=379, randomly selected individuals), GoNL (full panel) and a merged panel comprised of both reference panels. Imputed variants were categorised by MAF in to common (MAF > 5%), low-frequency (0.5 – 5%) and rare (0.05 – 0.5%) groups. Imputation quality was based on the Pearson correlation r^2 between the imputed dosage and the observed genotypes sourced from the ImmunoChip genotyping data.

Mean r^2 values improved for all cohorts when imputing up to the GoNL panel compared to the 1000 Genomes Project, most noticeable when imputing rare variants. The most significant improvement in imputation quality was observed in the Dutch cohort (mean r^2 increasing from 0.61 to 0.71) and smaller improvements were observed in the British (mean r^2 0.58 rising to 0.65) and Italian (mean r^2 0.43 rising to 0.47) cohorts. These metrics were further improved by merging the GoNL WGS data with the 1000 Genomes Project and imputing up to this novel reference panel. A further comparison between the 1000 Genomes project and a subset of the GoNL panel using a matched sample size was used to ensure that the improvement in imputation quality was not attributed to the increased sample size of the full GoNL panel, with results showing a similar improvement in imputation quality when imputing up to the GoNL (N=379) panel. Whilst the highest imputation quality was achieved in the Dutch cohort, both the British and, to a lesser degree, Italian cohorts exhibited improved imputation quality when imputing up to the GoNL panel. Principal component analyses of the three imputation cohorts provided further insight regarding the different degrees by which imputation quality improved and showed substantial overlap between the Dutch and British cohorts with the GoNL. In

fact, the British cohort showed a greater degree of overlap with the GoNL panel in comparison to the 1000 Genomes Project European cohort. The clustering of the Italian cohort however, showed that even with the plethora of European samples included in the reference panels, the Italian cohort was not as well represented in any of the reference panels in comparison to the Dutch and British cohorts. This appears to manifest in the imputation results and the relatively lower imputation quality for rare variants in the Italian cohort across all panels. However, as with the Dutch and British cohorts, the Italian cohort does exhibit improved imputation quality with the GoNL panel. This may be attributed to two factors in the reference panel design; (i) Sequencing depth for the GoNL panel was at 14x whereas the 1000 Genomes Project data was comprised of low depth (4 – 6x) supplemented with targeted exome sequencing data and SNP genotyping data; (ii) Samples used in the GoNL reference panel were members of a set of 231 trios or 19 quartets, allowing for improved phasing and accurate haplotypes. Finally, it is important to note that this comparison of imputation quality was limited to SNPs found on the ImmunoChip genotyping array and therefore did not cover the entirety of the reference panels involved in the comparison (Deelen et al, 2014; The 1000 Genomes Project Consortium, 2012).

Similar research completed by Pistis et al (2014) also examined the benefits of augmenting established reference panels with study specific WGS data. This particular endeavour used WGS data to create two population specific reference panels with 2,120 Sardinian samples (average depth of ~4x) and 1,325 samples of European ancestry sourced from Minnesota (average depth of ~10x) (Pistis et al, 2014). These WGS data sets were used to create study specific reference panels for both the Sardinian and Minnesota cohorts. Further, to assess the imputation quality derived from combined reference panels, each WGS data set was merged with the 1000 Genomes Project (Phase I, All) and the 1000 Genomes Project (Phase I, European samples), creating two merged reference panels each for both the Sardinia and Minnesota WGS data sets. Imputation quality was assessed by imputing 6,602 samples from the SardiNIA cohort (genotyped on the HumanOmniExpress array) and 5,429 samples from the Minnesota Center for Twin and Family Research (MCTFR)

cohort (genotyped on the Illumina 660W-quad array) up both study specific and merged reference panels and assessing r^2 values, comparing the imputed dosage with true genotypes available for the same sample sets genotyped on the HumanExome array. When imputing the Sardinian sample set, imputation quality was higher at all MAF ranges when imputing up to the Sardinian study specific reference panel compared to the 1000 Genomes Project (both the total reference panel and the European subset) and the Minnesota study specific panel. The same pattern was observed for the Minnesota sample set with the study specific reference panel providing the highest imputation quality out of the non-merged reference panels.

When comparing the imputation quality derived from the Sardinian WGS + 1000 Genomes panels and the Minnesota WGS + 1000 Genomes reference panels, only a small improvement in imputation quality was observed in the Sardinian sample set (specifically at low MAF ranges). The Minnesota sample set benefitted from a larger improvement in imputation quality when merging the Minnesota study specific panel with the 1000 Genomes, again at low MAF ranges. The comparably minor improvements in imputation quality observed in the Sardinian sample set when imputing up to the Sardinian WGS + 1000 Genomes panels may be attributed to the lack of Sardinian samples included in the 1000 Genomes Project. Whilst this particular study limited imputation comparisons to only chromosome 20, they established the same final conclusions as other research addressing this topic and observed improved imputation when imputing up to population specific reference panels, particularly for low frequency and rare variants.

Following on from these initial designs of population specific reference panels, studies such as those completed by Mitt et al (2017) utilised high depth WGS data (30x depth) to again augment the 1000 Genomes Project panel. Additionally, the improvement in imputation quality afforded by population specific reference panels was compared to the improvements offered by increasing overall sample size, sourced from a diverse range of population groups as seen in reference panels such as the HRC panel. A subset of 2,244 Estonian individuals subject to whole-genome sequencing (30x depth) was used to create the novel EGCUT reference panel. This

WGS data set was also used to augment the 1000 Genomes Project by imputing up to both the EGCUT and 1000 Genomes Project reference panels using the 'imputation with two phased reference panels' option of the IMPUTE2 software. The methods used by Mitt et al (2017) to impute to a combined reference panel using both the 1000 Genomes Project in addition to study specific data, differ from those used in the previously discussed studies. Pistis et al (2014) and Deelen et al (2014) assessed the potential of using combined reference panels, however, both limited their final combined reference panel to variants present in both of the composite reference panels. The option to impute to two phased reference panels in IMPUTE2 allows for the selection of a primary and secondary reference panel. IMPUTE2 allows for imputation up to the genotypes of the primary reference panel whilst the secondary reference panel is used to provide additional haplotype data to support the imputation process. Thus, Mitt et al evaluated the effect of using the population specific WGS data in the roles of the primary and secondary reference panel. The final comparison was completed across the 1000 Genomes Project, the HRC, the EGCUT, the EGCUT (primary) + 1000 Genomes Project (secondary) and the 1000 Genomes Project (primary) + EGCUT (secondary). Once again, the most effective method to improve imputation quality was through the augmentation of an established reference panel using WGS data, leading to improved imputation of rare and low frequency variants. The use of multiple reference panels proved most effective when using the larger (in terms of number of variants) 1000 Genomes Project panel as the primary reference panel, with the population specific reference panel (EGCUT) supporting accurate imputation for this set of variants.

Mitt et al (2017) compared the capabilities of large reference panels comprised of diverse sample sets (i.e., the 1000 Genomes Project and HRC) to a much smaller but population matched panel. Their results indicated that the population matched reference panel (EGCUT panel) outperformed both larger panels in its capacity to output high quality imputed variants (this was assessed in terms of INFO score, an alternative to r^2 metrics and specific to the IMPUTE software (section 1.2.5)). However, the performance of the EGCUT panel was limited by the total number of variants included in the panel and so, the total imputed variants whilst of a generally

higher quality was less than that of the 1000 Genomes Project and the HRC (Mitt et al, 2017).

These three studies represent only a selection of the total publications that have discussed the development of population specific reference panels, as numerous reference panels have been developed since, for a multitude of specific population groups. Reference panels such as the ChinaMAP reference panel (Li et al, 2021) supporting imputation across a variety of Chinese populations, and the DV-GLx AFAM panel (O'Connell et al, 2021) for imputation of African American population groups, represent some of the most recent population specific panels published. The three studies discussed in detail, however, highlight key considerations in their design and relevance to imputation for association studies.

Across these studies, WGS data of 4x, 10x, 14x and 30x depth was used to augment existing reference panels, but comparisons between the efficacy of WGS data of different depths and combinations thereof was relatively limited. This has the potential to be a key area for consideration when designing population specific reference panels as quantifying the difference in improvements in imputation quality between WGS of varying depth could potentially aid in the guidance of panel design and the allocation of resources between the total number of WGS samples vs the depth at which each sample is sequenced.

3.1.3 Chapter aims

The primary aim of this chapter is to establish the most effective approach to improving imputation quality into Japanese GWAS utilising WGS data available from Biobank Japan. This chapter will document the design and development of multiple iterations of a Japanese population specific reference panel (using a collection of high and low depth WGS data) based on the augmentation of a widely used and publicly available, the 1000 Genomes Project (phase III) reference panel. With the development of multiple Japanese specific reference panels, we aim to compare the effects of incorporating high depth and low depth sequencing data into an established and diverse reference panel in addition to evaluating the performance of

a reference panel that incorporates a combination of both high and low depth sequencing data. Imputation will be performed using a GWAS cohort of 174,460 Japanese samples (non-overlapping with the WGS data) taken from the Biobank Japan and genotyped on the Illumina HumanOmniExpress and HumanExome arrays, imputing up to each iteration of the Japanese specific reference panels. The performance of these Japanese specific reference panels will be placed into context through comparison with the 1000 Genomes Project (phase III) alone. Imputation comparisons with additional widely used reference panels were limited as all imputation was performed locally on RIKEN servers due to data permissions regarding the GWAS cohort. This meant that data could not be uploaded to imputation servers to facilitate comparisons with the HRC and TOPMed reference panels.

This chapter will not be limited to only identifying the best performing reference panel based on the quantity and depth of WGS data used but will also place a focus on fully describing the gains in imputation quality derived from these novel reference panels. Specifically, we will assess reference panels based on the total imputed SNPs in their respective imputation outputs and the confidence with which these SNPs are imputed. Further, performance will be examined across the MAF spectrum to understand the MAF ranges in which the addition of WGS data is most impactful in its effect on imputation quality, and how this differs between the novel reference panels based on their composition in reference to WGS data. The chapter will conclude by determining the best performing novel reference panel based on the summary statistics extracted from each reference panel's imputation output.

3.2 Methods

3.2.1 Japanese reference panel design

WGS data of various depths were available for a subset of 7,517 Japanese individuals and sequencing details are as follows. 1,502 individuals were sequenced at 30x depth using one of three Illumina platforms: Illumina Hiseq 2500 Rapid, Illumina Hiseq 2500 V4, and Illumina Hiseq X Five. 1,786 individuals were sequenced at 15x depth on the

Illumina Hiseq X Five platform. The remaining 4,229 individuals were sequenced at 3x depth on the Illumina Hiseq X Five platform. The preparation of WGS data and subsequent reference panel design was completed by Naoko Miyagawa of the Laboratory for Statistical and Translational Genetics at the RIKEN Centre for Integrative Medical Sciences.

Prior to the development of the Japanese population specific reference panels, the 7,517 whole-genome sequences were subject to a series of quality control filters: (i) 5 samples were excluded due to genomic DNA failure; (ii) a further 12 samples were excluded due to read quality control including errors in sequencing index, fraction of uniformity coverage, GC bias, fraction of duplicate reads and low read depth; (iii) 19 duplicate samples removed based on identity by descent checks; (iv) 1 sample excluded following heterozygosity checks; (v) a further 2 samples filtered based on low concordance with genotyping array; (vi) a final 6 samples removed as a result of high contamination rate measured by verify BamID (thresholds of CHIPMIX ≥0.25 and FREEMIX ≥0.25). The final post QC whole-genome sequence data set was comprised of 7,472 samples.

A total of four Japanese population specific reference panels were designed using the cleaned whole-genome sequence data taken from BBJ, supplementing the 1000 Genomes Project Phase III (1KG) reference panel with varying amounts of WGS at 30x, 15x and 3x depths; (i) 1000 Genomes Project + 1,037 WGS at 30x depth (1KG+1K); (ii) 1000 Genomes Project + 3,256 WGS (1,491 at 30x depth and 1,756 at 15x depth) (1KG+3K); (iii) 1000 Genomes Project + 4,216 WGS at 3x depth (1KG+4K); (iv) 1000 Genomes Project + 7,472 WGS (1,491 at 30x depth, 1,756 at 15x depth and 4,216 WGS at 3x depth) (1KG+7K). Whilst a total of 1,491 WGS were available at 30x depth, the 1KG+1K panel was comprised of 1KG + 1,037 WGS. This was a decision taken by the RIKEN Laboratory for Statistical and Translational Genetics, to separate the WGS in sets of approximately 1000, 3000, 4000 and 7000 WGS. The total numbers of WGS used in each Japanese specific reference panel are slightly higher than their panels stated values to account for the potential loss of any samples during the QC protocols.

VCF files for the 1KG panel were downloaded from the 1000 Genomes Project and subsequently filtered to remove multi-allelic sites (via vcftools-0.1.16) before conversion to hap and legend file formats. Similarly, multi-allelic sites were removed from cleaned WGS data sets using vcftools-0.1.16. 1,037 WGS at 30x depth were phased (via SHAPEIT2), following this conversion from VCF format to hap and legend files was performed. The hap and legend files for the 1KG panel and 1,037 were then merged using the '--merge panel' option of IMPUTE2 and subject to a final QC stage, removing multi-allelic site, singletons and monomorphic sites to produce the final 1KG+1K reference panel. This process was repeated with the set of 3,256 WGS of 30x and 15x depth, and the set of 4,216 WGS at 3x depth to create the 1KG+3K and 1KG+4K reference panels. To establish the 1KG+7K panel, an additional step was required to properly integrate the WGS of mixed depth into the 1KG panel. The phased sets of 3,256 WGS (30x and 15x depth) and 4,216 WGS (3x depth) were merged using IMPUTE2, the merged set of 7,472 WGS was then merged again via IMPUTE2 with the hap and legend files of the 1KG reference panel.

3.2.2 GWAS cohort and quality control

A preliminary cohort of 181,927 Japanese individuals from BBJ (genotyped on the Illumina HumanOmniExpress and HumanExome Bead chips) were made available for imputation comparisons across the four novel Japanese reference panels. Of these initial 181,927 individuals, 212 withdrew consent and so, were excluded from any further analysis. PCA of the cohort was completed and a total of 93 individuals were removed after being identified as population outliers. These outliers were removed from the cohort to reduce the potential for false positives due to population structure. A further 7,162 individuals overlapping WGS data used in the Japanese reference panels were also excluded from the sample cohort. Following these initial filters, a further series of QC procedures were completed to prepare the cohort for imputation. QC performed on a sample level included: (i) removing individuals with call rate <98%; (ii) exclusion of individuals with discordant genotypes, gender-mismatch and removal of closely related samples. QC protocols for variants included: (i) exclusion of variants with call rate <99%; (ii) filtered based on Hardy-Weinberg equilibrium ($p<1x10^{-6}$); (iii) filtered palindromic variants and non-autosomal variants; (iv) filtered MAF <1%. The

final cleaned sample set for imputation was comprised of 174,460 individuals at 520,378 variants. This sample set was phased using EAGLE2v2.3 (no reference panel used for phasing) and imputed up to the 1KG+1K, 1KG+3K, 1KG+4K and 1KG+7K reference panels, in addition to the 1KG panel alone to act as a reference point for imputation quality. Imputation was completed using Minimac3v1.0.11 with default parameters engaged.

3.2.3 Comparison of imputation quality between reference panels

r^2 metrics sourced from info files output from Minimac3 were used as the primary indicator for imputation quality. For each of the five reference panels, separate info files for each chromosome were grouped together and filtered to remove any genotyped variants, producing a single file for each imputation output composed of imputed SNPs only. The imputed variants for each output file were grouped based on MAF in the imputed dosage data and divided into the following MAF bins: (i) MAF < 0.001%; (ii) 0.001% ≤ MAF < 0.01%; (iii) 0.01% ≤ MAF < 0.1%; (iv) 0.1% ≤ MAF < 0.5%; (v) 0.5% ≤ MAF < 1%; (vi) 1% ≤ MAF < 2.5%; (vii) 2.5% ≤ MAF < 5%; (viii) 5% ≤ MAF < 10%; (ix) MAF ≥ 10%. Summary statistics were collected for each MAF bin, documenting the total number of variants, the mean r^2 and mean MAF of variants within the bin. Further, to assess the total number of well imputed variants two thresholds of $r^2 > 0.3$ and $r^2 > 0.8$ were used. $r^2 > 0.3$ is a common threshold used in GWAS as a minimum quality standard required to use imputed variants in downstream analysis and results. $r^2 > 0.8$ represents a much more stringent threshold for imputation quality, important when analysing variants at low MAF ranges, where there is typically more uncertainty in the quality of their imputation.

The total output for each reference panel was assessed by examining the relationship between mean MAF and mean r^2 values across the nine MAF bins to capture imputation quality across the MAF spectrum. Following this, we documented the distribution of variants across the MAF bins and compared the total number of variants falling in the r^2 categories of $r^2 < 0.3$, $0.3 \leq r^2 < 0.8$, and $r^2 \geq 0.8$ for each panel's imputation output. Following the analysis of each reference panels total imputation output, we limited analysis to a subset of 39,973,082 SNPs shared across

all four Japanese specific reference panels in addition to the 1KG reference panel. With this set of shared SNPs, we repeated our analysis of imputation quality to achieve a direct comparison of each reference panels capability to impute the same set of SNPs.

3.3 Results

3.3.1 Comparison across total imputation output

Table 3.3.1: Number of variants imputed up to five reference panels using a cohort of 174,460 Japanese individuals from BBJ.

Reference panel	Number of variants				
	Total	$r^2 \geq 0.3$	% of total with $r^2 \geq 0.3$	$r^2 \geq 0.8$	% of total with $r^2 \geq 0.8$
1KG	43,122,360	11,339,966	26.30	6,421,806	14.89
1KG+1K	58,872,998	25,784,042	43.80	9,194,822	15.61
1KG+3K	71,898,691	33,361,661	46.40	11,466,614	15.95
1KG+4K	64,967,199	28,769,587	44.28	9,991,697	15.38
1KG+7K	86,203,220	42,996,951	49.88	12,402,161	14.39

Imputing up to the four Japanese specific reference panels in addition to the 1KG panel alone provided the preliminary results detailed in table 3.3.1. Imputation up to the 1KG+7K panel (using mixed depth WGS data) gave the largest number of imputed variants. A total of ~86 million imputed variants were included in this output, approximately twice the total output when imputing up to the 1KG panel alone. All reference panels incorporating WGS data produced significantly larger imputation outputs in comparison to the 1KG panel output. Furthermore, the number of additional imputed variants appears to be more dependent on the depth of the sequencing data as opposed to the number of WGS used to augment the 1KG panel. The 1KG+4K panel, using ~4000 low depth WGS to augment the 1KG panel, outputs a total of ~64 million imputed variants. However, using ~3000 high depth WGS to augment the 1KG panel (1KG+3K) results in a total output of ~71 million imputed variants, a 10.6% increase in imputed variants with ~1000 fewer sequences.

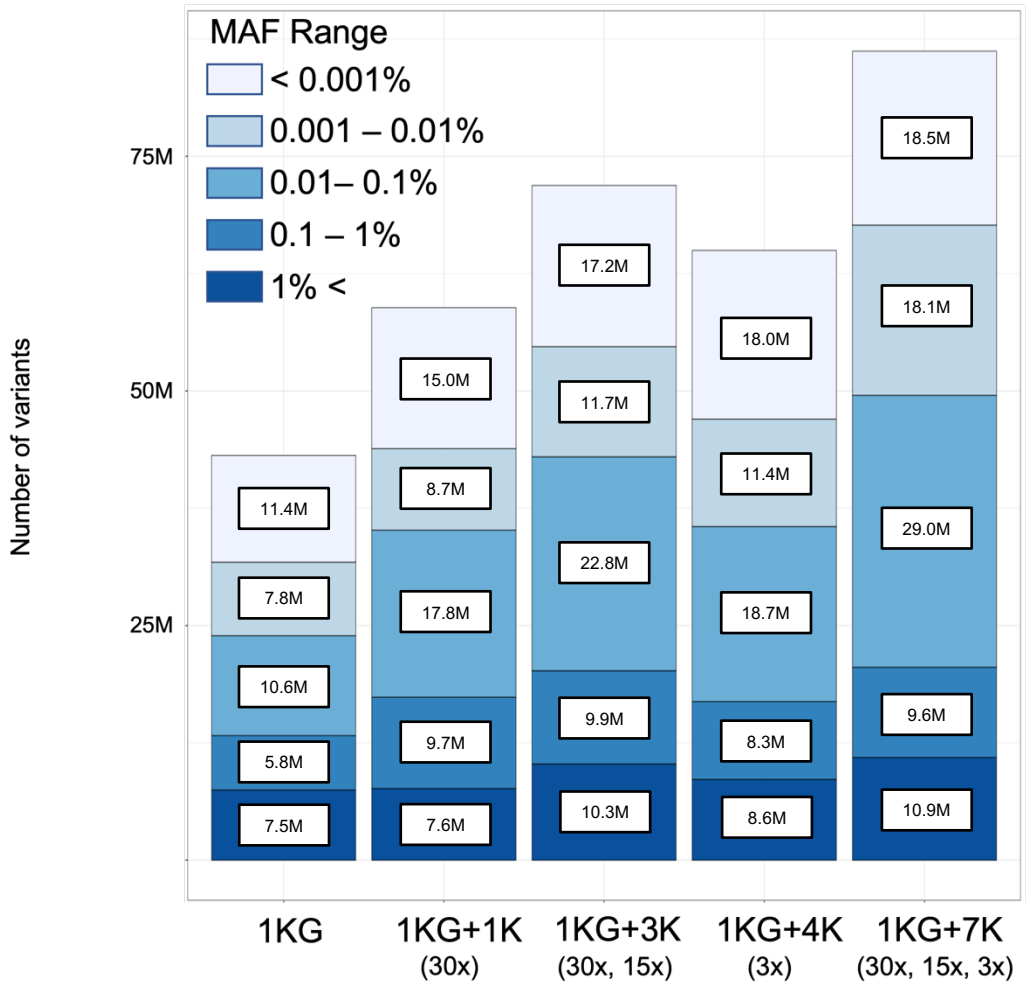

Figure 3.3.1: Number of variants imputed up to five reference panels using a cohort of 174,460 Japanese individuals from BBJ. Variants are categorised by MAF as described in the figure key and total numbers of SNPs (in millions) are detailed for each category.

Figure 3.3.1 details the distribution of imputed variants across various MAF categories used in the analysis. Looking at the distribution of each imputation output we observe the MAF ranges in which the additional imputed variants (sourced from WGS data) exist. As shown in figure 3.3.1, the majority of gains from incorporating WGS into the 1KG panel are low frequency and rare variants with MAF < 1%. The greatest difference in terms of imputed variants appears in the range 0.01% ≤ MAF < 0.1%. In this range, imputing up to the 1KG panel outputs 10,647,565 variants but the addition WGS data leads to significant increases in imputed variants; 17,783,071 in the 1KG+1K output, 22,790,959 in the 1KG+3K output, 18,664,321 in the 1KG+4K

output, and 28,969,767 in the 1KG+7K output (approximately 2.7x the output of the 1KG panel at this MAF range).

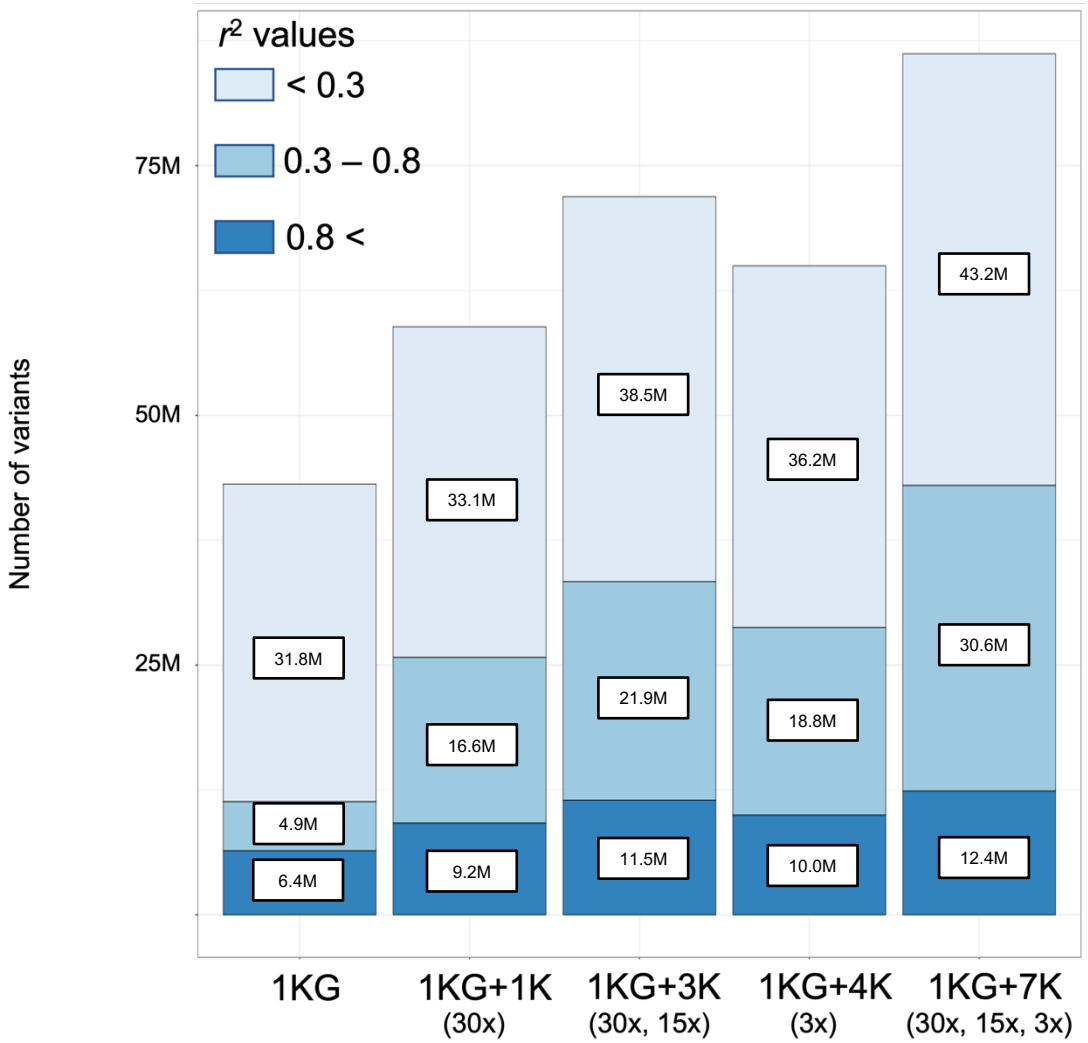

Figure 3.3.2: Number of variants imputed up to five reference panels using a cohort of 174,460 Japanese individuals from BBJ. Variants are categorised by r^2 as described in the figure key and total numbers of SNPs (in millions) are detailed for each category.

When looking at imputation quality across the totality of each of the five imputation outputs, figure 3.3.2 presents the number of variants imputed to different quality thresholds (based on r^2). As detailed in table 3.3.1, augmenting the 1KG panel with WGS provides a moderate increase in imputed variants meeting the highest r^2

threshold of $r^2 > 0.8$, with ~6 million imputed variants meeting this standard in the 1KG output and almost 2x this number (~12 million variants) in the 1KG+7K output. On this metric, the 1KG+1K and 1KG+4K panels performed similarly with 9.1 and 9.9 million variants meeting $r^2 > 0.8$ respectively. When considering a lower threshold of $r^2 > 0.3$, the increase in total imputed variants meeting this threshold increases dramatically with the introduction of WGS data. Once again, the 1KG+7K panel provides the largest number of imputed variants meeting this requirement with 42.9 million imputed variants with $r^2 > 0.3$, an almost 4-fold increase compared to the 1KG panel output. Imputation up to the two panels using high depth WGS data (1KG+1K and 1KG+3K) also showed a significant increase in variants with $r^2 > 0.3$, with increases of ~2x and ~3x of the 1KG panels output at this threshold respectively. The 1KG+4K panel, using only low depth WGS data displayed a small increase in output at this r^2 threshold in comparison to the 1KG+1K panel but did not match the performance of the 1KG+3K panel.

Table 3.3.2: Total imputation output for imputation up to the four Japanese specific reference panels and the 1KG. Imputed variants are categorised by MAF and again by r^2 value.

Reference panel	MAF category	Number of variants				
		Total	$r^2 \geq 0.3$	% of total with $r^2 \geq 0.3$	$r^2 \geq 0.8$	% of total with $r^2 \geq 0.8$
1KG	MAF < 0.001%	11,359,184	288,172	2.54	46,970	0.41
	0.001% ≤ MAF < 0.01%	7,840,949	289,252	3.69	32,871	0.42
	0.01% ≤ MAF < 0.1%	10,647,565	1,062,868	9.98	119,923	1.13
	0.1% ≤ MAF < 0.5%	4,562,977	1,854,633	40.65	290,374	6.36
	0.5% ≤ MAF < 1%	1,238,486	849,340	68.58	263,068	21.24
	1% ≤ MAF < 2.5%	1,357,934	1,045,885	77.02	512,060	37.71
	2.5% ≤ MAF < 5%	883,586	802,920	90.87	539,403	61.05
	5% ≤ MAF < 10%	999,195	971,467	97.22	783,925	78.46
	MAF ≥ 10%	4,232,484	4,175,429	98.65	3,833,212	90.57
1KG+1K	MAF < 0.001%	15,017,019	393,263	2.62	67,698	0.45
	0.001% ≤ MAF < 0.01%	8,694,697	678,676	7.81	67,874	0.78
	0.01% ≤ MAF < 0.1%	17,783,071	9,005,962	50.64	873,003	4.91
	0.1% ≤ MAF < 0.5%	8,139,960	6,948,019	85.36	1,259,108	15.47
	0.5% ≤ MAF < 1%	1,597,816	1,385,240	86.70	567,693	35.53
	1% ≤ MAF < 2.5%	1,479,464	1,300,659	87.91	787,894	53.26
	2.5% ≤ MAF < 5%	907,988	872,123	96.05	674,400	74.27
	5% ≤ MAF < 10%	1,016,202	1,000,525	98.46	886,318	87.22
	MAF ≥ 10%	4,236,781	4,199,575	99.12	4,010,834	94.67

Table 3.3.2 (cont.): Total imputation output for imputation up to the four Japanese specific reference panels and the 1KG. Imputed variants are categorised by MAF and again by r^2 value.

Reference panel	MAF category	Number of variants				
		Total	$r^2 \geq 0.3$	% of total with $r^2 \geq 0.3$	$r^2 \geq 0.8$	% of total with $r^2 \geq 0.8$
1KG+3K	MAF < 0.001%	17,180,457	537,765	3.13	94,132	0.55
	0.001% ≤ MAF < 0.01%	11,732,077	2,489,718	21.22	199,366	1.70
	0.01% ≤ MAF < 0.1%	22,790,959	15,531,304	68.15	2,292,908	10.06
	0.1% ≤ MAF < 0.5%	6,884,400	5,628,200	81.75	1,530,415	22.23
	0.5% ≤ MAF < 1%	3,059,240	1,251,006	40.89	615,063	20.11
	1% ≤ MAF < 2.5%	3,169,741	1,319,081	41.61	816,309	25.75
	2.5% ≤ MAF < 5%	1,245,465	934,929	75.07	692,804	55.63
	5% ≤ MAF < 10%	1,174,030	1,091,550	92.97	925,127	78.80
	MAF ≥ 10%	4,662,322	4,578,108	98.19	4,300,490	92.24
1KG+4K	MAF < 0.001%	17,973,198	515,897	2.87	92,551	0.51
	0.001% ≤ MAF < 0.01%	11,432,779	1,874,365	16.39	133,884	1.17
	0.01% ≤ MAF < 0.1%	18,664,321	11,087,843	59.41	1,194,996	6.40
	0.1% ≤ MAF < 0.5%	6,734,946	5,770,030	85.67	1,173,703	17.43
	0.5% ≤ MAF < 1%	1,552,826	1,346,552	86.72	549,464	35.38
	1% ≤ MAF < 2.5%	1,596,552	1,342,694	84.10	1,342,694	84.10
	2.5% ≤ MAF < 5%	1,022,188	943,434	92.30	703,094	68.78
	5% ≤ MAF < 10%	1,144,518	1,111,352	97.10	931,837	81.42
	MAF ≥ 10%	4,845,871	4,777,420	98.59	4,418,388	91.18
1KG+7K	MAF < 0.001%	18,529,367	772,782	4.17	124,055	0.67
	0.001% ≤ MAF < 0.01%	18,147,790	5,935,446	32.71	422,840	2.33
	0.01% ≤ MAF < 0.1%	28,969,767	21,460,803	74.08	2,872,672	9.92
	0.1% ≤ MAF < 0.5%	6,614,061	5,631,634	85.15	1,608,509	24.32
	0.5% ≤ MAF < 1%	2,999,756	1,254,867	41.83	636,866	21.23
	1% ≤ MAF < 2.5%	3,633,631	1,321,608	36.37	832,680	22.92
	2.5% ≤ MAF < 5%	1,427,662	949,517	66.51	693,502	48.58
	5% ≤ MAF < 10%	1,201,573	1,085,493	90.34	916,333	76.26
	MAF ≥ 10%	4,679,613	4,584,801	97.97	4,294,704	91.77

Table 3.3.2 offers insight into the additional well imputed variants afforded by the supplementation of the 1KG panel with WGS data. For three MAF categories representing the most common variants (MAF ≥ 10%, 5% ≤ MAF < 10%, and 1% ≤ MAF

< 2.5%), we observed similar number of imputed variants that attain r^2 values that meet the thresholds of $r^2 > 0.3$ and $r^2 > 0.8$ in each imputation output, albeit slightly higher numbers among the WGS augmented panels (1KG+4K panel seemingly producing the best imputation at these MAF ranges by a small margin). At the remaining MAF ranges (MAF < 1%) a general trend in the total number of well imputed variants emerges. All panels augmented with WGS impute noticeably greater number of variants at these MAF ranges with $r^2 > 0.3$ and $r^2 > 0.8$ in comparison to the 1KG panel alone. Across this series of MAF bins, the largest number of variants meeting the aforementioned r^2 thresholds is derived from imputing up to the 1KG+7K panel, followed by the 1KG+3K, 1KG+4K and 1KG+1K panels.

The majority of additional imputed variants from WGS data are located in the MAF < 0.001%, 0.001% ≤ MAF < 0.01%, and 0.01% ≤ MAF < 0.1% bins, and it is in these MAF ranges where the 1KG+7K panel displays the largest improvement in imputation quality. At the lowest MAF range, imputation up to the 1KG panel outputs 288,172 and 46,970 SNPs ($r^2 > 0.3$ and $r^2 > 0.8$), whereas imputation up to the 1KG+7K panel produces 772,782 and 124,055, an increase of more than 2.5x the total well imputed SNPs relative to the 1KG panel. However, even when imputing up to the best performing reference panel at this range (1KG+7K) imputation quality of SNPs with MAF < 0.001% remains extremely low with only 4% of imputed variants meeting the minimum r^2 of 0.3.

At the MAF range of 0.001% ≤ MAF < 0.01%, we observe a more substantial improvement in imputation quality when imputing up to the 1KG+7K both relative to the 1KG panel output and as a proportion of the total SNPs output at this range. A total of 5,935,446 and 422,840 SNPs with $r^2 > 0.3$ and $r^2 > 0.8$ are included in the 1KG+7K output with 32% of all SNPs (N=18,147,790) at this MAF range attaining $r^2 > 0.3$. This represents a marked improvement over the 1KG panel output were of the 7,840,949 total imputed SNPs, only 3.6% (N=289,252) of these SNPs met the minimum r^2 threshold of 0.3. The range of 0.01% ≤ MAF < 0.1% represents the range with the greatest level of improvement over 1KG alone in terms of imputation quality

when imputing up to the 1KG+7K panel. Imputation up to the 1KG+7K panel offers an almost 3x increase in the total number of imputed variants relative to the 1KG panel output. When imputing up to the 1KG+7K panel, this MAF bin is the first in which a majority of imputed SNPs meet the r^2 threshold of 0.3 (74% of 28,969,767 SNPs). Further, these 28,969,767 SNPs represent a 20-fold increase in the number of imputed variants with $r^2 > 0.3$ in comparison with the 1KG panel output.

3.3.2 Comparison across a subset of SNPs shared across all reference panels

Repeating the comparison of imputation quality across a subset of 39,973,082 SNPs polymorphic across all panels provided a direct comparison of imputation quality between the 1KG panel and the four WGS supplemented reference panels. Figure 3.3.3 summarises the mean r^2 and mean MAF stats for each imputation output (detailed in supplemental table A.9) and reinforces the initial observations made based on the comparison of total imputation outputs.

Across the subset of polymorphic SNPs, the mean r^2 across the MAF spectrum is higher for all WGS supplemented reference panel imputations in comparison to the 1KG imputation. Again, disparities in imputation quality are most apparent at lower MAF ranges, specifically MAF < 1.5%. At a MAF of 0 – 1% the highest quality imputation in terms of r^2 metrics is achieved by imputing up to the 1KG+7K panel. For imputed variants with MAF > 1%, imputing up to the 1KG+7K and 1KG+3K reference panels produces similar levels of imputation quality with the 1KG+3K producing the higher mean r^2 metrics at these MAF ranges. Whilst both the 1KG+1K and 1KG+4K panels produced a consistently higher quality imputation relative to the 1KG panel, both panels fail to achieve similarly high r^2 values to match the imputation quality derived from the 1KG+3K and 1KG+7K panels across all MAF ranges.

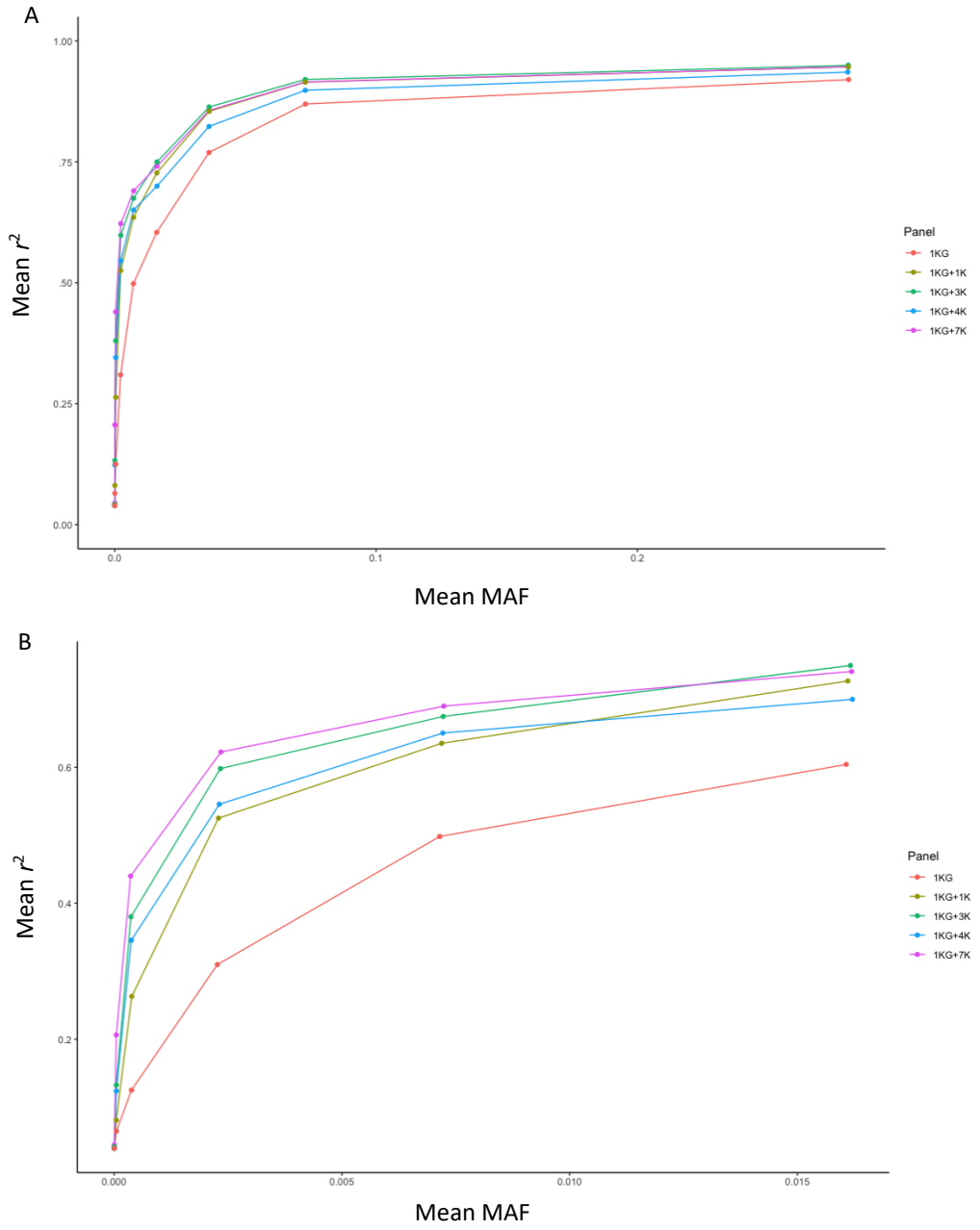

Figure 3.3.3: (A) Mean MAF vs mean r^2 for a subset of 39,973,082 SNPs polymorphic across all panels. (B) Zoomed in perspective of A, focussing on the MAF range of 0 – 1.5%.

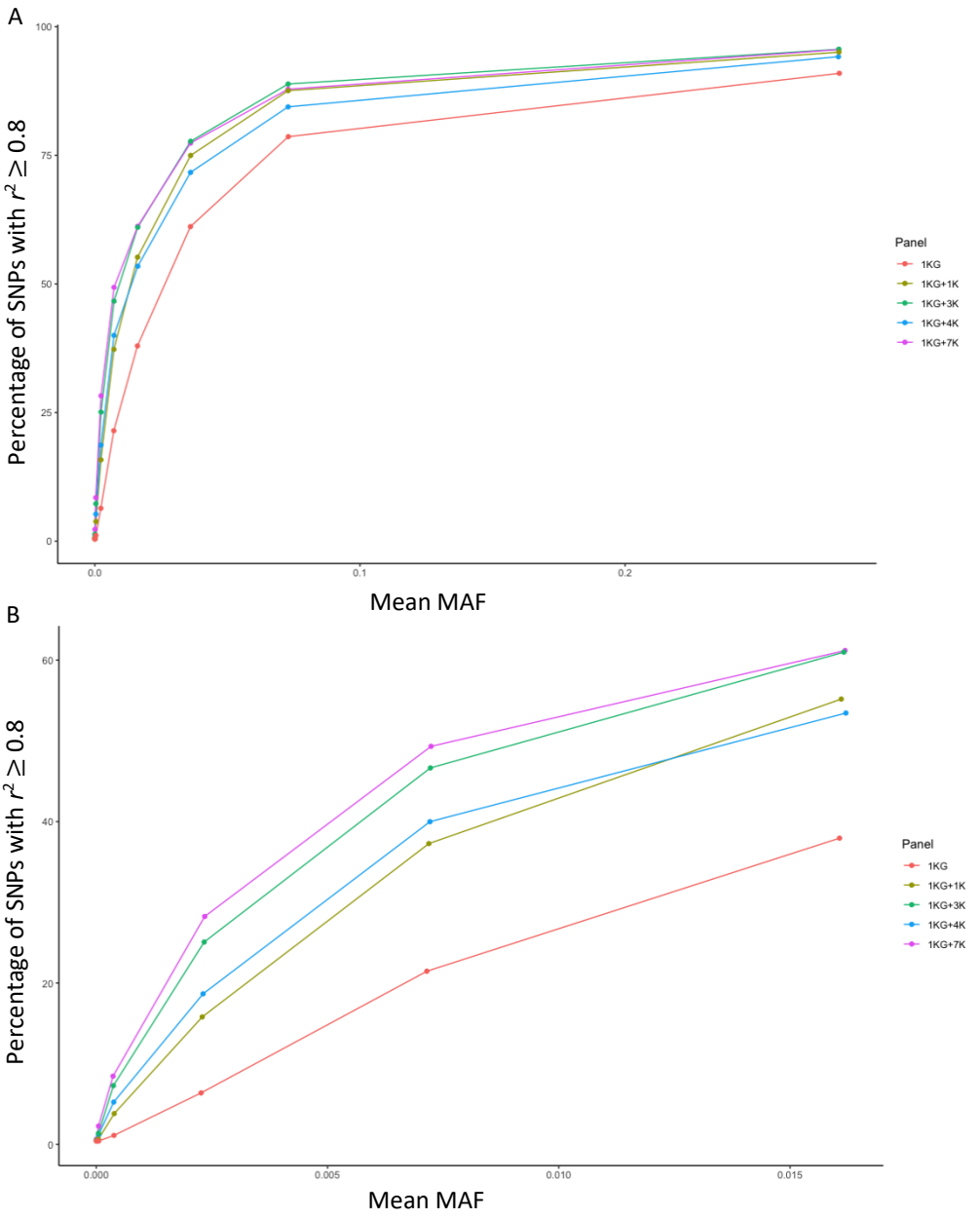

Figure 3.3.4: (A) Mean MAF for each MAF bin vs the percentage of SNPs with $r^2 \geq 0.8$ for a subset of 39,973,082 SNPs polymorphic across all panels. (B) Zoomed in perspective of A, focussing on the MAF range of 0 – 1.5%.

To further assess each imputation quality for each reference panel, we analysed each imputation output based on the proportion of imputed SNPs meeting the strict quality threshold of $r^2 > 0.8$. Observations based on this metric showed that all reference panels with additional WGS data performed to a much higher quality than

the 1KG panel alone. Similarly, to assessments made based on mean MAF vs mean r^2, we observed the 1KG+7K panel producing largest total of imputed SNPs with $r^2 > 0.8$ when imputing at MAF < 1.5%. At MAF >1.5% the 1KG+7K and 1KG+3K panels exhibit similar performance in terms of total imputed SNPs with $r^2 > 0.8$. Comparing the reference panels based on the percentage of SNPs with $r^2 > 0.8$ highlights the difference in imputation quality between the 1KG+7K and 1KG+3K panels and the 1KG+4K and 1KG+1K. The imputation quality of the 1KG+4K and 1KG+1K panels is noticeably lower than the remaining two Japanese specific panels and this is most apparent at lower MAF ranges (figure 3.3.4:B). In terms of overall output of high-quality imputed SNPs, imputation up to the 1KG+7K panel output 7,844,733 of the total 39,973,082 polymorphic SNPs with a minimum r^2 of 0.8 (19.6%). Imputation up to the 1KG+3K panel produced an almost identical output at this r^2 threshold with 19.0% (7,628,659) of SNPs with $r^2 > 0.8$. These two panels offered a small improvement at the highest threshold we assessed for imputation quality, in comparison to the 1KG+4K (17.7% of SNPs with $r^2 > 0.8$) and 1KG+1K (17.5% of SNPs with $r^2 > 0.8$) and a substantial improvement compared to the 1KG panel (14.5% of SNPs with $r^2 > 0.8$).

3.4 Discussion

3.4.1 Improvement in imputation quality with imputation up to population specific WGS data

The key metrics where we observed significant improvement over the 1000 Genomes project panel were the total number of imputed variants, total imputed variants with $r^2 > 0.3$ and total imputed variants with $r^2 > 0.8$, derived from each of the reference panels. For all three measurements of imputed variants the 1KG+7K panel, using all available WGS data of mixed depth produced the highest imputation quality. The 1KG+3K panel, using only high depth WGS data exhibited the second highest performance, with the 1KG+4K and 1KG+1K reference panels following in that order. It appears that the total number of sequences in the 1KG+4K panel, using only low depth WGS data (3x depth) was more effective (to a small degree) in comparison to the more limited 1KG+1K panel using only high depth (30x depth) WGS data. These

results do however indicate that to achieve imputation quality comparable to that of panel augmented with high depth WGS data, a significantly greater number of low depth WGS samples is required.

Next, we considered in what MAF ranges these additional imputed variants in terms of both total variants and well imputed variants. All reference panels augmented with WGS data followed the general trend regarding the distribution across the MAF spectrum, of additional SNPs sourced from WGS data. Regardless of the sample size or depth of the WGS used, the additional imputed SNPs were predominantly of low MAF with minimal increases in the number of common SNPs.

Repeating the comparisons of imputation quality limited to a subset of SNPs polymorphic across all reference panels facilitated a direct comparison of imputation quality across the five reference panels. Assessments of mean MAF vs mean r^2 and mean MAF vs total number of SNPs with $r^2 > 0.8$ provided important insight into how incorporating WGS into the 1KG panel impacts imputation quality of SNPs in different MAF bins. Imputation of rare and low frequency SNPs was greatly improved when utilising the WGS augmented reference panels. Imputation quality for common variants remained largely similar between the WGS augmented panels and the 1KG panel alone, with only small increases observed across the relevant metrics when imputing up to the WGS augmented panels. The top two performing panels based on these metrics were the 1KG+7K and 1KG+3K reference panels, whilst the 1KG+7K panel provided the best imputation quality at the lowest MAF ranges, at MAF of > 1% the imputation quality

3.4.2 WGS depth and reference panel design

Following our comparisons between all reference panels across both the total imputation output and the shared subset of SNPs, the initial observation is that augmenting the 1KG panel with WGS translates to improved imputation quality. The degree to which the imputation quality improves is dependent on the number of samples subject to WGS and the depth at which they are sequenced. Further, these

results highlighted the metrics by which imputation quality improved with the supplementation of WGS data.

The most effective way to improve imputation quality using a pre-existing reference panel augmented with WGS data is through the use of high depth WGS data. Overall, the addition of high depth WGS data leads to greater gains in imputation quality across all metrics tested in comparison to low depth WGS data. Low depth WGS however, was to a degree, effective at increasing imputation quality, although to a lesser extent compared to high depth WGS data. When considering the total number of imputed variants, supplementing the 1KG panel with a large amount of low depth WGS sequence data has the potential to offset the higher quality of the high depth WGS data to provide a larger total imputation output. When comparing imputation quality directly across a shared subset of SNPs the difference is not so clear cut. Imputation quality when imputing up to the 1KG+4K panel was greater than that of the 1KG+1K panel for rare SNPs but only by a small margin. When imputing SNPs with MAF > 1% however, the 1KG+1K panel provided a higher quality imputation. Generally, both the 1KG+4K and 1KG+1K panel provided comparable levels of imputation quality but due to the much larger sample size required of the low depth data.

Augmenting the 1KG panel with a combination of high depth WGS data (30x and 15x) proved to be a successful endeavour and provided a substantial improvement in imputation quality. In addition to large performance gains over the 1KG panel alone, the additional sequencing data led to further improvements in imputation quality when compared to the 1K+1K (30x depth). The additional WGS data (the majority of which was at 15x depth) supported improved r^2 metrics at rare and low frequency SNPs and also increased the total number of variants included in the reference panel. Thus, not only were r^2 metrics improved for SNPs polymorphic across all reference panels, the 1KG+3K output a greater number of well imputed SNPs when considering the total imputation output. These improvements in imputation quality, afforded by a larger sample size of high depth WGS data, are less apparent in the imputation of common SNPs and most pronounced in rare and low frequency SNPs. Common SNPs,

due to their high MAF, will have a high minor allele count (MAC) even in small sample sizes, and therefore will be captured effectively in smaller reference panels, providing ample information from which haplotypes are established. For SNPs with low MAF, the inverse is true, and much larger sample sizes are required to effectively capture enough occurrences of the mutation in a given population, to facilitate the accurate estimation of the haplotypes on which the SNP is present. As discussed in section 1.2, improving the accuracy of reference haplotypes leads to a higher degree of confidence in the imputed SNPs.

The 1KG+7K reference panel was designed to assess whether low depth sequencing data could be used to supplement additional high depth sequencing data and whether this would be to the benefit or detriment of the reference panels performance. The 1KG+7K panel used 30x, 15x and 3x depth WGS data, processing the high and low depth data separately prior to incorporation into the 1KG reference panel. Based on the series of comparisons completed between the WGS augmented reference panels, the 1KG+7K panel provided the highest quality imputation across the metrics we assessed. Imputation up to the 1KG+7K panel produced the largest total number of imputed SNPs in addition to the greatest number of imputed SNPs meeting the r^2 thresholds of 0.3 and 0.8. Furthermore, when assessing the subset of SNPs polymorphic across all imputation outputs, the highest imputation quality of rare and low frequency SNPs was obtained when imputing up to the 1KG+7K reference panel. Whilst imputation quality at a MAF > 1% was comparable with 1KG+3K panel (using only high depth data) the improved imputation of rare SNPs is significant enough to distinguish the 1KG+7K panel, as the primary goal of developing such population specific reference panels is to improve the imputation quality for population specific SNPs that are of typically low MAF and thus, generally imputed with less confidence.

The publications discussed in section 3.2.1 share the same general conclusions that the use of population specific WGS data can facilitate a higher quality imputation, an effect most pronounced at low MAF. However, these studies assessed the imputation quality achieved when imputing up to a standard publicly available reference panel,

a study specific reference panel and a combination of a publicly available panel and their own study specific WGS data. The results described in section 3.3 present a comparison between multiple iterations of a population specific reference panel, allowing for a greater understanding of how differing WGS sample set sizes and sequencing depths effect the improvements in imputation quality observed when supplementing a reference panel such as the 1KG with WGS data. As previous studies such as Deelen et al (2014) N=769, Pistis et al (2014) N=2,120 and N=1,325, and Mitt et al (2017) N=2,244, used relatively limited sample sets of WGS data, the results do not describe the effects of using much larger sample sizes when augmenting existing reference panels with WGS. The results in section 3.3 describe the effects of using up to 7,472 WGS to augment the 1KG reference panel and show large improvements in imputation quality from imputing up to the 1KG to the 1KG+1K/1KG+4K panels, and again from the 1KG+1K/1KG+4K to the 1KG+3K panel, followed by a much smaller improvement in imputation quality when moving to the 1KG+7K panel. This raises the possibility of diminishing returns when using WGS data to augment the 1KG panel. However, as the additional sequencing data used in the 1KG+7K panel was of low depth, the smaller improvements in imputation quality may be due sequencing depth and larger improvements in imputation quality may still be possible had the additional WGS data from the 1KG+3K to the 1KG+7K been of a similarly high depth.

These comparisons of imputation quality support the conclusion that incorporating WGS of mixed high and low depth can prove beneficial when designing population specific reference panels. Throughout these comparisons the mixed depth 1KG+7K provided a consistently high level of imputation quality, outperforming the competing reference panels using WGS of various quantities and depths. However, it important to consider that these assessments are only based on raw imputation output and metrics such as r^2 calculated by Minimac3v1.0.11. Therefore, it is necessary to further assess the utility of the 1KG+7K and whether or not the improvements in imputation quality metrics translate to tangible benefits when utilised in association studies.

CHAPTER 4

GWAS INTO SERUM URIC ACID FOLLOWING IMPUTATION UP TO THE JAPANESE POPULATION SPECIFIC REFERENCE PANEL

4.1 Introduction

4.1.1 Serum uric acid and gout

Gout is a form of inflammatory arthritis and estimated to have a worldwide prevalence between 0.1 – 10%, making it the most common form of inflammatory arthritis. The symptoms of gout develop as a direct consequence of the interaction between monosodium urate (MSU) crystals and various tissues in the body. MSU crystal formation within joints can trigger an acute inflammatory response initiated by the activation of vascular endothelial cells, facilitating leucocyte recruitment into the tissue in question. These acute inflammatory responses manifest as severe pain in the tissues implicated, with symptoms occurring over period of up to 10 days at a time (Dalbeth and Haskard, 2005). The development of gout is primarily caused by increased levels of serum uric acid, a condition known as hyperuricaemia that promotes the formation of MSU crystals. Risk factors for the development of gout include but are not limited to outside environmental factors, as a strong genetic contribution has been documented with the heritability of serum uric acid levels estimated to be in the range of 30 - 60% (Köttgen et al, 2013; Tin et al, 2019).

The prevalence of the disease varies significantly across different regions of the world, with developing countries reporting a general trend of increased incidence. However, even within individual countries, health care surveys have documented an increased predisposition to the development of gout within specific population groups (Kuo et al, 2015). For example, in New Zealand, where a relatively high prevalence of gout has been observed in the overall population, analysis of individual ethnic groups highlighted the disparities across population groups. Individuals aged ≥20 years of

European (3.24% prevalence), Asian (1.96% prevalence), Māori (6.06% prevalence) and Pacific (7.63% prevalence) ancestry recorded vastly different rates of gout across their respective demographics with Māori and Pacific populations exhibiting a much greater incidence of gout (Winnard et al, 2012).

One of the largest GWAS into serum uric acid levels in the Japanese populations was completed by Kanai et al (2018) where association analyses for 58 traits (including serum uric acid) were completed in a cohort of 162,255 Japanese individuals imputed up to the 1000 Genomes Project Phase I (East Asian reference haplotypes). In total 27 loci were reported based on their association with serum uric acid. Furthermore, more recent meta-analyses performed within the Japanese population have identified a total of 36 trait associated loci in samples imputed up to the 1000 Genomes Project Phase III (all ancestry haplotypes) and the 1000 Genomes Project Phase I (East Asian reference haplotypes) (Nakatochi et al, 2019). Additionally, a trans-ancestry meta-analyses of GWAS into serum uric acid reported by Tin et al (2019) provides a comprehensive report of trait associated loci identified in the UK biobank. GWAS of 457,690 individuals sourced from the UK biobank led to the identification of 183 loci exhibiting association with serum uric acid. These three publications will serve as a reference point for the work completed in this chapter as we perform further association analyses of serum uric acid in the Japanese population.

4.1.2 Chapter aims

This chapter builds on the results of the imputation comparison across the four Japanese specific reference panels and the 1000 Genomes Project. Following the comparison of imputation quality based on r^2 metrics across both total output and SNPs polymorphic across all panels, we determined that imputation up to the 1KG+7K panel produced the greatest level of imputation quality. This comparison, however, does not fully describe the utility of the 1KG+7K panel and the benefits of the improved quality and quantity of imputed SNPs sourced from the reference panel when applied to association studies. Therefore, this chapter will explore the potential

benefits of imputing up to the 1KG+7K reference panel in the context of GWAS of a complex human trait, serum uric acid levels.

The work in this chapter will cover the trait of serum uric acid levels in the Japanese population due to prior publications providing comprehensive documentation of trait associated loci, initially comparing GWAS results with those published Kanai et al (2018) but also expanding the comparison to those reported in the meta-analyses by Nakatochi et al (2019). With these publications as a reference point, we completed two separate GWAS in a Japanese cohort imputed up to the 1000 Genomes Project and 1KG+7K reference panels. With this we will address three objectives: 1) compare lead SNPs across known loci for serum uric acid when imputing up to the 1000 Genomes project (phase III) and the 1KG+7K panels and in turn, with previously published results; 2) assess the improvements in fine-mapping loci facilitated by imputation up to the 1KG+7K panel, identifying SNPs unique to the 1KG+7K panel and SNPs with significantly improved r^2 metrics in the 1KG+7K panel; 3) test the capacity of the 1KG+7K panel to support the identification of novel loci for serum uric acid in the Japanese population and determine whether these novel loci have been reported in other population groups.

4.2 Methods

4.2.1 Sample set and imputation

The sample set for GWAS into serum uric acid levels was sourced from the BBJ cohort previously described in section 3.2.2. Members of the BBJ cohort were subject to routine examinations and clinical data for a total of 47 diseases was collected. We extracted a total of 104,174 individuals from the post QC sample set for which serum uric acid levels had been recorded. Records of serum uric acid for each individual were adjusted for age, sex, 10 principal components of genetic ancestry and the status of the 47 diseases for which data was collected, and subsequently normalised to Z scores.

The subset of 104,174 individuals for which serum uric acid data was available was imputed up to both the 1000 Genomes Project (phase III) (1KG) and the 1KG+7K reference panels using Minimac3v1.0.11 with default parameters engaged following pre-phasing using EAGLE2v2.3 (no reference panel selected, default parameters engaged). The imputation output for each panel was filtered on the basis of an r^2 threshold of $r^2 > 0.3$.

4.2.2 Genome-wide association study of serum uric acid

We completed two separate GWAS of serum uric acid in 104,174 samples using the imputation outputs from the 1000 Genomes Project and 1KG+7K reference panels. All variants were tested for association in an additive model using the --glm option (linear regression model) of PLINKv2, analysing the imputed genotype dosages output by Minimac3v1.0.11. This initial GWAS placed primary consideration to a set of 27 loci previously reported in the Japanese population (Kanai et al, 2018). Each of the loci for serum uric acid documented in Kanai et al (2018) served as initial targets for comparison between the 1KG and 1KG+7K based GWAS, with summary stats including MAF, r^2 and p value for each of the original lead SNPs recorded for each GWAS output. Furthermore, where the lead SNP in the 1KG or 1KG+7K analyses differed from that originally published, we recorded summary statistics for both the previously reported and novel lead SNPs for said loci. Additionally, where the lead variant for a locus differed between the 1KG and 1KG+7K GWAS results we performed an additional step of conditional analyses. The lead variant identified in the 1KG output was incorporated into the regression model as a covariate in the association analyses for the 1KG+7K output to determine whether the SNPs from both the 1KG and 1KG+7K imputation both represented the same association signal.

In circumstances where SNPs exhibited exceptionally strong signals of association, summary statistics output by PLINKv2 included p values of $p = 0$. To address this, we used the Rmpfr package in R to determine the p values of SNPs where $p = 0$ based on the PLINKv2 summary statistics.

4.2.3 Conditional analyses

At the 27 trait associated loci we completed conditional analyses using a forward selection approach with the intention to identify regions comprised of multiple distinct signals. For each locus we used the --condition option of PLINK2 to incorporate the lead SNP into the regression model as a covariate. If, following conditional analysis, there are SNPs remaining at genome-wide significance ($p < 5 \times 10^{-8}$), the process was repeated. This meant, of the SNPs remaining at genome-wide significance, the SNP exhibiting the strongest association signal was included as an additional covariate in the same manner as the previous lead SNP. This iterative process was repeated until no SNPs in the locus achieved genome-wide significance, producing a final 'conditional set' of SNPs for said locus.

For loci implicating multiple SNPs in their respective 'conditional sets' we completed a further series of conditional analyses. These additional analyses were completed to identify the SNPs within the conditional sets that maintained association signals reaching genome-wide significance in a joint model. In the case that a SNP did not achieve genome-wide significance, the SNP in the conditional set with the weakest association signal was excluded from the set. This process was repeated until all SNPs within the conditional set attained genome-wide significance in the final joint model, and all signals of association sourced from SNPs outside of the conditional set were fully attenuated. Finally, to dissect the distinct association signals from each of the SNPs included in the final conditional sets, we excluded individual SNPs, one by one, from the conditional set whilst the remaining SNPs were incorporated into the regression model as covariates. The SNP with the strongest association signal (based on *p* value) when conditioned on all other SNPs in the conditional set, was labelled as the index variant for the signal.

4.2.4 Expanding the GWAS to identify novel associations

Following the comparison between the 1KG and 1KG+7K panel imputation outputs for 104,174 samples across 27 loci previously reported in the Japanese population, we expanded our analysis of the 1KG+7K imputation output to include any potential

trait associated loci that did not map to the initial 27. After applying a threshold of MAF < 0.1% to exclude exceptionally rare SNPs, we compared the lead SNPs identified in our expanded analyses with those reported in two additional publications of association analyses of serum uric acid: (i) a trans-ancestry genome-wide association study of serum uric acid in a total of 457,690 individuals (183 loci reported) (Tin et al, 2019); and (ii) a genome-wide meta-analysis of serum uric acid in 121,745 Japanese individuals (36 loci reported) (Nakatochi et al, 2019). In instances where lead SNPs from the 1KG+7K output were located >1Mb away from loci reported in these previous meta-analyses, the loci were reported as a novel association for serum uric acid. Lead SNPs from the 1KG+7K panel output that mapped within 1Mb of the loci reported in the meta-analyses were subject to additional analyses. In these cases, the lead SNP from the 1KG+7K output was subject to conditional analysis, using the lead SNP from the meta-analysis as a covariate in the regression model.

4.3 Results

4.3.1 GWAS results for 27 known loci

Our initial results contrast the lead SNPs across 27 known loci associated with serum uric acid levels in the Japanese population as reported in Kanai et al (2018). The lead SNP for each locus was compared across the those reported in the publication and those in our GWAS based on the 1KG and 1KG+7K imputation outputs. Imputation of 104,174 individuals from BBJ up to the 1KG panel led to an alternative lead SNP at 19/27 known loci. Extracting summary statistics for the alternative lead SNPs in our 1KG imputation showed the similarity in MAF and the strength of association (based on *p* values) between this set and those originally reported (1KG summary statistics reported in supplementary table A.10). As for the original lead SNPs, those reported in our 1KG imputation were common SNPs, most with MAF > 10%. The level of association observed between the original and new lead SNPs were similar for 18/19 lead SNPs. The exception however, being the *NRXN2-SLC22A12* locus originally defined by SNP rs57633992 (p = 7.3 x 10^{-845}). Imputation up to the 1KG panel indicated an alternative lead SNP, rs121907892 (p = 1.4377 x 10^{-1816}) with a stronger association signal. Imputation quality was high across all lead SNPs from the 1KG

imputation output with all SNPs meeting a minimum of $r^2 > 0.5$, although the majority of SNPs had extremely high r^2 metrics ($r^2 > 0.9$).

GWAS based on the 1KG+7K imputation output led to different lead SNPs in at a total of 20/27 (table 4.3.1) known loci in comparison to both the original lead SNPs and those reported in the 1KG imputation output. Similarly to the lead SNPs reported in the 1KG output, most of those reported in the 1KG+7K output were of similar MAF (all common SNPs) with high r^2 metric and exhibited similar levels of association in terms of *p* values. There were, however, two notable exceptions. The first exception being the *NRXN2-SLC22A12* locus, now defined by SNP rs121907892 (as reported in the 1KG output) with a stronger level of association compared to the previous lead SNP. The second exception was at the *LOC101927932* locus, previously defined by SNP rs6026578 in both Kanai et al (2018) and our own GWAS using the 1KG panel output. The alternative SNP based on the 1KG+7K output, rs202213319, is a common SNP (MAF = 16%), confidently imputed ($r^2 = 0.953$) and displays a stronger degree of association ($p = 3.9 \times 10^{-12}$) compared to rs6026578 (MAF = 28%, $r^2 = 0.975$, $p = 6.5 \times 10^{-10}$) than the previous lead SNP. Furthermore, rs202213319 is unique to the 1KG+7K panel and therefore absent from the 1KG panel.

Conditioning the lead SNPs from the GWAS based on the 1KG+7K imputation on those identified in the GWAS based on the 1KG imputation led to the attenuation of association signal at 16/27 loci (table 4.3.1). This suggests that at these loci, these pairs of SNPs both represented the same association signal. For 7/27 loci, the lead SNP in both the 1KG+7K and 1KG imputations was identical and the for the remaining four loci, the lead SNP from the 1KG imputation was not present in the 1KG+7K imputation and so no conditional analyses were performed at these loci.

In our comparison between the original set of 27 loci and those observed in the 1KG+7K output, we noted the absence of the original lead SNP for the *HNF4G* locus (rs1828911). This SNP represented a multi-allelic site, which had been removed from the 1KG+7K during its development. Further, when comparing the results from the 1KG with the 1KG+7K panel, we noted four additional sites where the lead SNP was

not present in the 1KG+7K panel: rs59578826 at the *MUC1* locus, rs150320174 at the *TP53INP1-NDUFAF6* locus, rs10886117 at the *EMX2-RAB11FIP2* locus, and rs6598541 at the *IGF1R* locus were all subject to filtering from the 1KG+7K panel due to being multi-allelic sites.

Table 4.3.1: Lead variants at 27 previously reported loci (Kanai et al. 2018) for serum uric acid in GWAS of 104,174 Japanese individuals from the Biobank Japan Project after imputation up to the 1KG+7K panel.

Locus	Chr	Lead variant from imputation up to 1KG+7K panel										
		rs ID	Position	Major allele	Minor allele	MAF	Beta	SE	*p*-value	r^2	Present in 1KG panel?	Conditional *p*-value on 1KG lead variant
NBPF10-NBPF20	1	rs34913946	145,724,937	AT	A	0.2156	-0.0566	0.0076	8.1E-14	0.476	Yes	Same variant
MUC1	1	rs2990223	155,184,975	G	A	0.1717	0.0484	0.0061	2.6E-15	0.869	Yes	rs59578826 not in 1KG+7K panel
GCKR	2	rs1260326	27,730,940	T	C	0.4422	-0.0346	0.0043	1.3E-15	1.001	Yes	Same variant
USP34	2	rs7570707	61,451,744	C	T	0.3683	0.0226	0.0045	4.5E-07	0.992	Yes	1.2E-01
LRP2	2	rs3821129	170,204,773	T	C	0.1940	0.0383	0.0055	2.9E-12	0.982	Yes	3.2E-01
SFMBT1-RFT1	3	rs2564934	53,024,580	G	A	0.4991	0.0234	0.0043	6.0E-08	0.988	Yes	9.5E-01
SLC2A9	4	rs3775948	9,995,182	C	G	0.4243	-0.1215	0.0043	1.5E-173	1.008	Yes	5.4E-07
PRDM8-FGF5	4	rs10857147	81,181,072	A	T	0.2965	-0.0299	0.0047	2.8E-10	0.992	Yes	2.9E-03
ABCG2	4	rs4148155	89,054,667	A	G	0.2891	0.1125	0.0047	2.5E-124	0.998	Yes	2.3E-02
SLC17A1	6	rs68094823	25,795,971	A	ACACACC	0.1623	-0.0529	0.0058	1.2E-19	0.999	Yes	4.1E-01
ZNF318	6	rs6930689	43,293,630	T	C	0.3860	0.0297	0.0045	3.7E-11	0.969	Yes	8.7E-02
UNCX-MICALL2	7	rs36022572	1,299,800	AC	A	0.4150	0.0269	0.0045	3.1E-09	0.927	Yes	2.8E-01
MLXIPL-VPS37D	7	rs13234378	73,026,151	A	T	0.1020	-0.0400	0.0071	1.8E-08	0.995	Yes	6.7E-02
HNF4G	8	rs2977945	76,475,125	G	A	0.4323	0.0372	0.0043	1.2E-17	0.998	Yes	8.8E-02
TP53INP1-NDUFAF6	8	rs56332377	95,980,168	C	CA	0.2460	-0.0306	0.0050	1.0E-09	0.995	Yes	rs150320174 not in 1KG+7K panel
BICC1	10	rs67713047	60,303,593	TA	T	0.4863	0.0356	0.0044	7.2E-16	0.950	Yes	1.4E-04
FAM35A	10	rs659494	88,953,694	T	A	0.3240	0.0321	0.0048	1.8E-11	0.926	Yes	1.1E-01

Table 4.3.1 (cont.): Lead variants at 27 previously reported loci (Kanai et al. 2018) for serum uric acid in GWAS of 104,174 Japanese individuals from the Biobank Japan Project after imputation up to the 1KG+7K panel.

Locus	Chr	Lead variant from imputation up to 1KG+7K panel										
		rs ID	Position	Major allele	Minor allele	MAF	Beta	SE	*p*-value	r^2	Present in 1KG panel?	Conditional *p*-value on 1KG lead variant
EMX2-RAB11FIP2	10	rs1886603	119,482,303	G	A	0.3740	0.0270	0.0044	1.2E-09	1.005	Yes	rs10886117 not in 1KG+7K panel
SBF2	11	rs2220970	9,857,749	G	A	0.3384	0.0231	0.0046	3.8E-07	0.997	Yes	Same variant
MPPED2-DCDC5	11	rs3925584	30,760,335	T	C	0.3200	-0.0241	0.0046	1.6E-07	1.005	Yes	1.1E-01
NRXN2-SLC22A12	11	rs121907892	64,361,219	G	A	0.0223	-1.2503	0.0141	7.8E-1713	0.995	Yes	Same variant
CUX2	12	rs149212747	111,836,771	A	AC	0.2738	-0.0873	0.0058	1.9E-50	0.681	Yes	7.9E-07
IGF1R	15	rs12898337	99,294,355	T	C	0.4950	-0.0347	0.0043	1.4E-15	0.980	Yes	rs6598541 not in 1KG+7K panel
CYB5B-MIR1538	16	rs8048032	69,599,803	G	A	0.1730	0.0328	0.0060	6.0E-08	0.886	Yes	Same variant
LINC01229-LOC102724084	16	rs11376510	79,745,672	G	GT	0.2910	-0.0356	0.0049	6.0E-13	0.916	Yes	Same variant
BCAS3	17	rs9895661	59,456,589	C	T	0.4597	0.0386	0.0043	2.8E-19	1.007	Yes	Same variant
LOC101927932	20	rs202213319	57,466,467	A	G	0.1600	0.0417	0.0060	3.9E-12	0.953	No	2.1E-04

4.3.2 Fine mapping complex loci

To assess the performance of the 1KG+7K panel and how its use impacts the potential to fine-map complex regions where multiple sources of association signals are present, we completed stepwise conditional analysis for all loci (as described fully in section 4.2.3). Out of the 27 known loci we identified six loci where we observed multiple distinct signals of association at genome-wide significance, independent of one another. Across these six loci, we identified a total of 35 index SNPs displaying independent association with serum uric acid and distributed across the loci as follows: *NBPF10-NBPF20* (three signals); *LRP2* (two signals); *SLC2A9* (six signals); *ABCG2* (three signals); *NRXN2-SLC22A12* (18 signals); and *LINC01229-LOC102724084* (three signals). Of the 35 signals we identified, 11 of the SNPs implicated were unique to the 1KG+7K panel output (table 4.3.2).

Table 4.3.2: Previously reported loci (Kanai et al. 2018) with multiple distinct signals of association at genome-wide significance ($p < 5 \times 10^{-8}$) for serum uric acid in GWAS of 104,174 Japanese individuals from the Biobank Japan

Locus	Chr	rs ID	Position	Major allele	Minor allele	MAF	Beta	SE	p-value	r^2	Present in 1KG panel?	Annotation
NBPF10-NBPF20	1	rs146397899	145,447,275	C	T	0.0007	-0.6790	0.1221	2.7E-08	0.429	Yes	
		rs782333127	145,607,993	C	T	0.0028	-0.3029	0.0503	1.7E-09	0.664	No	*POLR3C*: intron variant
		rs34913946	145,724,937	AT	A	0.2156	-0.0494	0.0077	1.4E-10	0.476	Yes	*PDZK1*: 2kb upstream variant
LRP2	2	rs34806803	170,012,192	G	GT	0.4581	-0.0262	0.0044	1.9E-09	0.981	Yes	*LRP2*: intron variant
		rs3821129	170,204,773	T	C	0.1944	0.0389	0.0055	1.3E-12	0.982	Yes	*LRP2*: intron variant
SLC2A9	4	rs1026406370	9,774,006	G	A	0.0004	-1.5322	0.1446	3.2E-26	0.570	No	*SLC2A9*: intron variant
		rs56050634	9,815,440	C	T	0.0861	-0.0576	0.0078	1.4E-13	0.999	Yes	*SLC2A9*: intron variant
		rs3775948	9,995,182	C	G	0.4243	-0.2126	0.0080	3.8E-155	1.008	Yes	*SLC2A9*: intron variant
		rs7657096	10,004,000	A	G	0.4824	0.0741	0.0083	4.9E-19	0.997	Yes	*SLC2A9*: intron variant
		rs34325511	10,028,867	G	A	0.2888	-0.0681	0.0059	1.6E-30	0.914	Yes	*SLC2A9*: intron variant
		rs192673734	10,340,489	C	T	0.0013	-0.4736	0.0611	9.0E-15	0.920	Yes	
ABCG2	4	rs72552713	89,052,957	G	A	0.0216	0.1623	0.0172	3.7E-21	0.751	Yes	*ABCG2*: stop gained
		rs4148155	89,054,667	A	G	0.2891	0.1126	0.0048	3.0E-121	0.998	Yes	*ABCG2*: intron variant
		rs66704028	89,093,036	C	T	0.1345	-0.0541	0.0064	2.4E-17	0.988	Yes	*ABCG2*: intron variant
NRXN2-SLC22A12	11	rs1324446966	63,928,771	C	T	0.0005	-0.8754	0.1245	2.1E-12	0.655	No	*MACROD1*: intron variant
		rs565795531	64,171,413	C	T	0.0005	-0.8948	0.1197	7.8E-14	0.643	Yes	
		N/A	64,317,578	A	C	0.0004	-1.1268	0.1404	1.0E-15	0.586	No	
		N/A	64,328,362	A	G	0.0001	-2.5255	0.4326	5.3E-09	0.409	No	

Table 4.3.2 (cont.): Previously reported loci (Kanai et al. 2018) with multiple distinct signals of association at genome-wide significance ($p < 5 \times 10^{-8}$) for serum uric acid in GWAS of 104,174 Japanese individuals from the Biobank Japan

Locus	Chr	rs ID	Position	Major allele	Minor allele	MAF	Beta	SE	p-value	r^2	Present in 1KG panel?	Annotation
NRXN2-SLC22A12	11	rs71581748	64,356,942	C	T	0.0340	0.0840	0.0115	3.3E-13	0.958	Yes	*SLC22A12*: 2kb upstream variant
		rs121907896	64,359,297	G	A	0.0026	-1.2958	0.0519	5.0E-137	0.670	No	*SLC22A12*: missense variant
		rs58174038	64,360,355	G	A	0.0006	-1.2316	0.0902	1.9E-42	0.801	Yes	*SLC22A12*: splice donor variant
		rs75786299	64,361,042	G	A	0.0062	0.3811	0.0285	7.1E-41	0.841	Yes	*SLC22A12*: synonymous variant
		rs201136391	64,361,124	G	A	0.0010	-0.3390	0.0573	3.4E-09	1.005	Yes	*SLC22A12*: missense variant
		rs121907892	64,361,219	G	A	0.0223	-1.2527	0.0140	5.8E-1754	0.995	Yes	*SLC22A12*: stop gained
		rs1291429571	64,446,967	G	A	0.0003	-1.0652	0.1485	7.3E-13	0.755	No	*NRXN2*: intron variant
		rs912882868	64,644,012	G	A	0.0002	-1.4673	0.1975	1.1E-13	0.715	No	*EHD1*: intron variant
		rs1253297381	64,674,066	C	T	0.0002	-1.4198	0.1375	5.4E-25	1.025	No	*ATG2A*: intron variant
		rs549413722	64,678,048	G	A	0.0002	1.3748	0.2058	2.4E-11	0.855	Yes (11:64678048:G:A)	*ATG2A*: intron variant
		rs540025991	64,726,351	C	T	0.0005	-1.1222	0.1242	1.7E-19	0.878	Yes (11:64726351:C:T)	*MAJIN*: intron variant
		rs540570840	64,742,946	G	A	0.0006	-1.1758	0.0943	1.1E-35	0.796	Yes (11:64742946:G:A)	
		N/A	64,744,403	G	A	0.0001	-2.5086	0.2715	2.5E-20	0.506	No	
		rs777218029	64,813,240	G	A	0.0003	-1.1220	0.1344	7.0E-17	0.867	No	*NAALADL1*: intron variant
LINC01229-LOC102724084	16	rs11376510	79,745,672	G	GT	0.2906	-0.0316	0.0050	2.2E-10	0.916	Yes	*LOC105371356*: intron variant
		rs4077450	79,931,595	T	G	0.4170	0.0280	0.0045	6.4E-10	0.929	Yes	
		rs374548262	80,123,137	ATG	A	0.0665	0.0557	0.0099	1.7E-08	0.781	Yes	*LOC105371357*: intron variant

At locus *NBPF10-NBPF20,* we identified a single rare SNP (MAF = 0.28%) unique to the 1KG+7K panel. This SNP, rs782333127 was imputed with a high degree of confidence (r^2 = 0.664) despite its low MAF. However, in terms of function, information was relatively limited. Similarly, at locus *SLC2A9,* we identified a single SNP (rs1026406370) with an extremely low MAF (MAF = 0.04%) unique to the 1KG+7K panel. Again, this SNP was well imputed (r^2 = 0.570) despite its low frequency in the population and annotated as an intron variant of the *SLC2A9* gene. The remaining nine signals unique to the 1KG+7K panel were all localised in the *NRXN2-SLC22A12* locus (figure 4.3.1) representing the most complex of the six loci where multiple signals were observed. These nine index SNPs were all of extremely low MAF (mean MAF = 0.052%) whilst still retaining a high level of confidence in their imputation (mean r^2 = 0.69). Most importantly however, is that this collection of nine SNPs included a functional variant, specifically, SNP rs121907896 (MAF = 0.26%, r^2 = 0.67, p = 4.96 x 10^{-137}). This SNP represents a missense variant (p.Arg90His) of the *SLC22A12* gene.

Outside of this collection of 11 signals unique to the 1KG+7K panel output, we observed multiple additional functional variants in our fine mapping efforts. At the *NRXN2-SLC22A12* locus we reported rs121907892 (a stop gained variant, p.Trp258Ter, MAF = 2.23%, r^2 = 0.995, p = 5.8 x 10^{-1754}); rs58174038 (a splice donor variant, c.506+1G>A, MAF = 0.06%, r^2 = 0.801, p= 1.9 x 10^{-42}); and rs201136391 (missense, p.Ala227Thr, MAF = 0.10%, r^2 = 1.00, p = 3.4 x 10^{-9}), all present in both the 1KG and 1KG+7K outputs. In addition to the functional variants in the *NRXN2-SLC22A12* locus, we reported an additional functional variant localised in the *ABCG2* locus: SNP rs72552713, a stop gained variant (p.Gln126Ter, MAF = 2.16%, r^2 = 0.751, p = 3.7 x 10^{-21}).

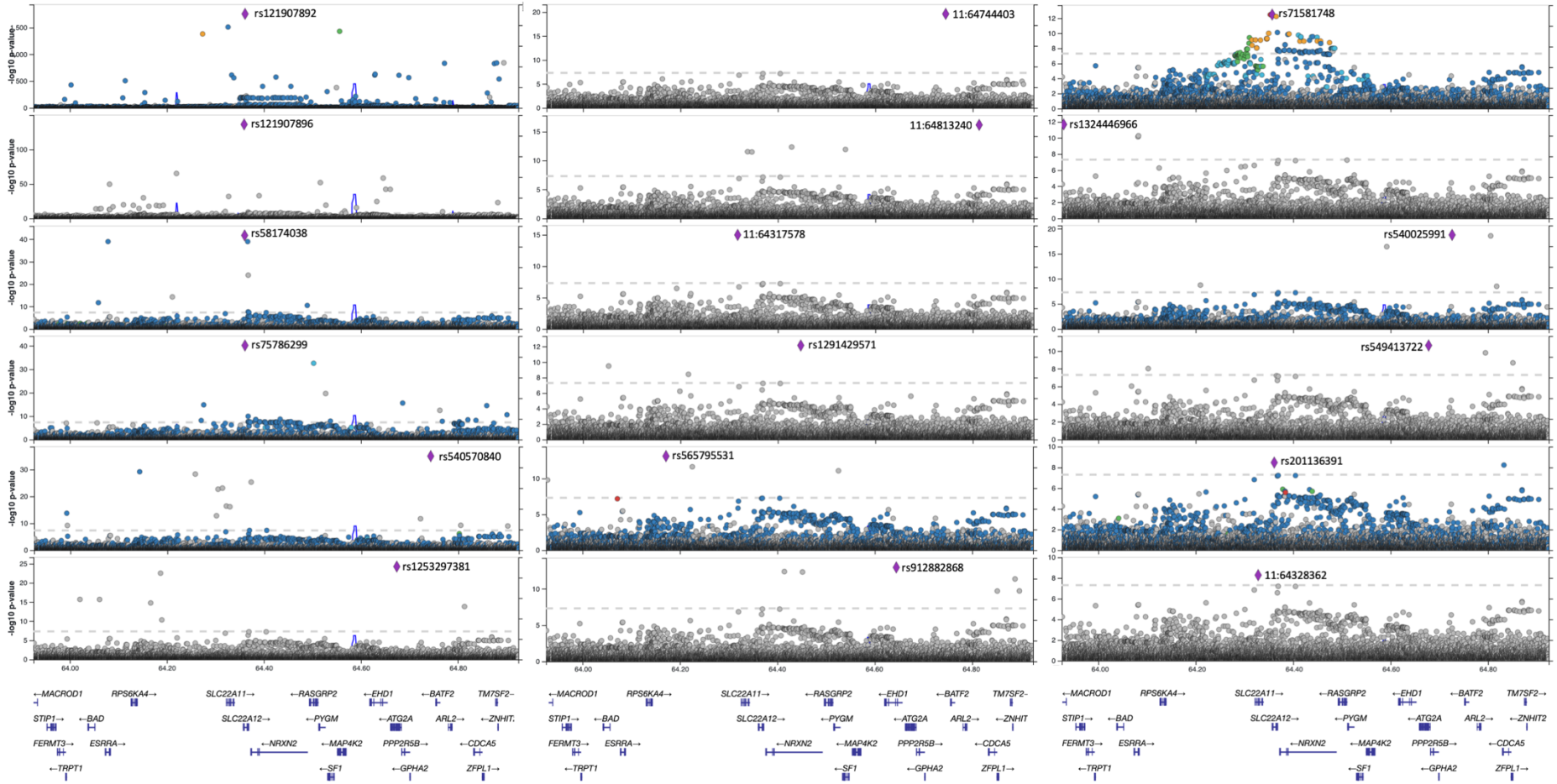
rs121907892
11:64744403
rs71581748
rs121907896
11:64813240
rs1324446966
rs58174038
11:64317578
rs540025991
rs75786299
rs1291429571
rs549413722
rs540570840
rs565795531
rs201136391
rs1253297381
rs912882868
11:64328362
-log10 p-value
64.00
64.20
64.40
64.60
64.80
←MACROD1
RPS6KA4→
STIP1→
←BAD
FERMT3→
ESRRA→
←TRPT1
SLC22A11→
←RASGRP2
←EHD1
←BATF2
TM7SF2-
SLC22A12→
←PYGM
←ATG2A
ARL2→
←ZNHIT
←NRXN2
←MAP4K2
PPP2R5B→
←CDCA5
←SF1
←GPHA2
ZFPL1→

Figure 4.3.1: Locus-Zoom plots for independent association signals for serum uric acid within the *NRXN2-SLC22A12* locus in a GWAS of 104,174 samples taken from BBJ, following imputation up to the 1KG+7K Japanese population specific reference panel. Each plot presents an independent association signal after conditioning on all other signals within the locus. The purple diamond denotes the index SNP for each signal whilst all other variants are colour coded according to LD (sourced from the 1KG East Asian subset) with the index variant: red $r^2 \geq 0.8$; gold $0.6 \leq r^2 < 0.8$; green $0.4 \leq r^2 < 0.6$; cyan $0.2 \leq r^2 < 0.4$; blue $r^2 < 0.2$; grey r^2 unknown. Gene annotations are taken from the University of California Santa Cruz genome browser.

4.3.3 Novel locus discovery

Following our analysis of the 27 known loci, we next assessed the evidence for further loci meeting the threshold for genome-wide significance ($p < 5 \times 10^{-8}$) and how imputing up to the 1KG+7K panel impacted the capacity to discover novel trait associated loci (summarised in figure 4.3.1). We identified 11 loci that had a lead SNP with MAF > 0.1% that were not previously reported in BBJ (Tables 4.3.3 and 4.3.4). Further, using the findings of the trans-ancestry meta-analysis of 457,690 individuals (Tin et al, 2019) and a meta-analysis of 121,745 Japanese individuals (Nakatochi et al, 2019) to exclude loci that did not represent novel associations, we confirmed that 5/11 loci did not map to those previously reported in either BBJ or these additional meta-analyses. Loci that did not map to known associations are as follows; *MEIS1* (rs553688146, MAF = 0.57%, $p = 4.7 \times 10^{-8}$); *OXSR1* (rs117297673, MAF = 4.54%, $p = 9.3 \times 10^{-11}$); *HCRTR2* (rs4715502, MAF = 7.31%, $p = 1.5 \times 10^{-11}$); *MAP7* (rs78302547, MAF = 0.57%, $p = 3.9 \times 10^{-9}$); and *MFSD12* (rs2240751, MAF = 39.52%, $p = 1.4 \times 10^{-8}$). Of these five loci, both *MEIS1* and *MAP7* are represented by lead SNPs that fall into the rare variant category: rs553688146 MAF = 0.57% and rs78302547 MAF = 0.57%. All lead SNPs for these loci were present in the 1KG output, but imputation quality for low frequency and rare variants was improved in the 1KG+7K output. The improvement in quality was most noticeable in the imputation of *MEIS1* (rs553688146), where imputation up to the 1KG panel output $r^2 = 0.168$ and failed to meet the minimum threshold of $r^2 > 0.3$. Imputation up to the 1KG+7K panel, however, led to a marked increase with $r^2 = 0.452$. Furthermore, the level of association (in terms of *p* values) observed in the lead SNPs of these loci was significantly improved when imputing up to the 1KG+7K panel. When imputing up to the 1KG panel 0/5 SNPs met the threshold for genome-wide significance.

The lead SNPs of the six remaining loci reaching genome-wide significance mapped within 1Mb of the lead SNP of a locus reported in the meta-analyses (Tin et al, 2019 & Nakatochi et al, 2019). In this case, the lead SNP for each locus was conditioned on the lead SNP for the previously reported locus. The *SESN2* locus with lead SNP rs74896528 was an exact match with that reported in Nakatochi et al (2019) and so was excluded without conditional analysis. The *MYL2* locus with lead SNP rs925368

was in close proximity to rs17550549 (Tin et al, 2019) and conditioning on said SNP resulted in a change in *p* value from $p = 8.4 \times 10^{-14}$ to $p = 2.0 \times 10^{-2}$ and was no longer genome-wide significant, indicating that these two SNPs represent the same association signal. In a similar manner, the *COMMD4* locus with the lead SNP rs140379576 was shown to be in extremely close proximity to rs73436803 (Tin et al, 2019) and conditional analyses on this SNP led to a loss of genome-wide significance; $p = 5.2 \times 10^{-9}$ to $p = 3.7 \times 10^{-3}$. The three remaining loci retained genome-wide significance following conditioning on the nearest previously reported lead SNP and so, *HAO2* (rs547500487, MAF = 0.20%, $p_{COND} = 4.8 \times 10^{-10}$); *HNF1A* (rs1169288, MAF = 49.70%, $p_{COND} = 4.2 \times 10^{-9}$); and *PDILT* (rs11646437, MAF = 27.74%, $p_{COND} = 3.3 \times 10^{-10}$) were deemed to be novel independent signals of association. Of these three loci, the *HAO2* locus and its lead SNP (rs547500487) is of particular interest as it is another example of a rare variant (MAF = 0.20%) with a large improvement in both imputation quality ($r^2 = 0.369$ to $r^2 = 0.974$) and to a lesser degree the level of association ($p = 9.0 \times 10^{-8}$ to $p = 4.8 \times 10^{-10}$).

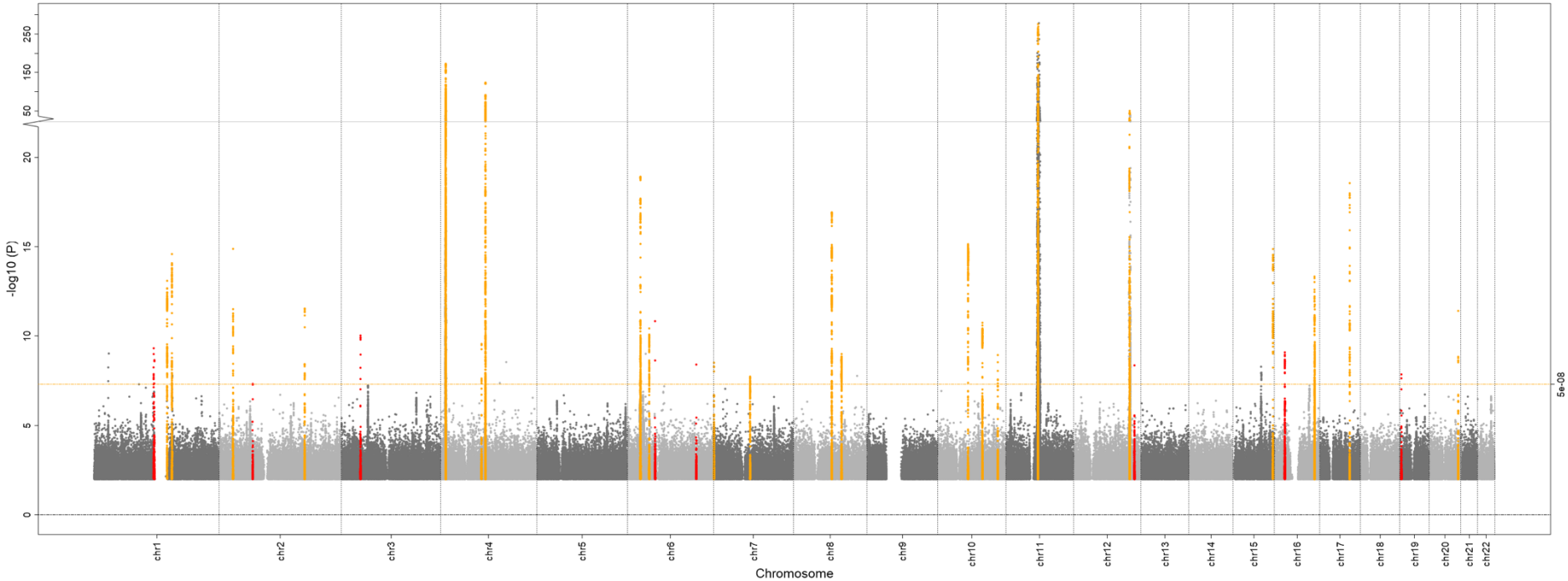

Figure 4.3.2: Manhattan plot of genome-wide association with serum uric acid in 104,174 Japanese individuals from the Biobank Japan Project. Each point represents a variant passing quality control ($r^2 > 0.3$ and MAF > 0.1%), plotted with their association *p*-value (on a -log10 scale) as a function of genomic position (NCBI build 37). Association signals attaining genome-wide significance ($p < 5x10^{-8}$, indicated by yellow horizontal line) are highlighted in yellow (previously reported loci) and red (novel loci).

Table 4.3.3: Lead variants mapping outside of 27 previously reported loci (Kanai et al. 2018) for serum uric acid in GWAS of 104,174 Japanese individuals from the Biobank Japan Project after imputation up to the 1KG panel.

Locus	Chr	rs ID	Position	Major allele	Minor allele	Imputation up to 1KG panel				
						MAF	Beta	SE	*p*-value	r^2
SESN2	1	rs74896528	28,598,287	C	T	0.0553	-0.0548	0.0094	5.6E-09	1.003
HAO2	1	rs547500487	119,081,763	A	G	0.0007	-0.7058	0.1320	9.0E-08	0.369
MEIS1	2	rs553688146	66,852,917	G	A	0.0011	0.1085	0.1582	4.9E-01	0.168
OXSR1	3	rs117297673	38,228,409	T	C	0.0207	-0.0809	0.0174	3.3E-06	0.755
HCRTR2	6	rs4715502	54,918,779	G	C	0.0819	0.0433	0.0092	2.5E-06	0.728
MAP7	6	rs78302547	136,860,353	A	C	0.0117	-0.1123	0.0249	6.7E-06	0.644
MYL2	12	rs925368	110,390,979	T	C	0.1100	-0.0616	0.0081	2.1E-14	0.727
HNF1A	12	rs1169288	121,416,650	A	C	0.4952	-0.0261	0.0043	1.1E-09	1.010
COMMD4	15	rs140379576	75,656,310	CCCAG	C	0.0746	-0.0489	0.0083	4.2E-09	0.969
PDILT	16	rs11646437	20,609,702	A	T	0.2741	0.0302	0.0050	1.2E-09	0.944
MFSD12	19	rs2240751	3,548,231	A	G	0.3381	-0.0315	0.0073	1.5E-05	0.393

Table 4.3.4: Lead variants mapping outside of 27 previously reported loci (Kanai et al. 2018) for serum uric acid in GWAS of 104,174 Japanese individuals from the Biobank Japan Project after imputation up to the 1KG+7K panel.

Locus	Chr	rs ID	Position	Major allele	Minor allele	Imputation up to 1KG+7K panel					Previously reported lead variant			
						MAF	Beta	SE	*p*-value	r^2	rs ID	Position	Conditional *p*-value	Reference
SESN2	1	rs74896528	28,598,287	C	T	0.0552	-0.0548	0.0094	5.7E-09	1.004	rs74896528	28,598,287	Same SNV	Nakatochi *et al*. 2019
HAO2	1	rs547500487	119,081,763	A	G	0.0020	-0.3023	0.0485	4.8E-10	0.974	rs141990161	119,943,525	4.8E-10	Tin *et al*. 2019
MEIS1	2	rs553688146	66,852,917	G	A	0.0057	0.2320	0.0425	4.7E-08	0.452	N/A	N/A	N/A	
OXSR1	3	rs117297673	38,228,409	T	C	0.0454	-0.0706	0.0109	9.3E-11	0.898	N/A	N/A	N/A	
HCRTR2	6	rs4715502	54,918,779	G	C	0.0731	0.0658	0.0097	1.5E-11	0.719	N/A	N/A	N/A	
MAP7	6	rs78302547	136,860,353	A	C	0.0057	-0.2099	0.0356	3.9E-09	0.642	N/A	N/A	N/A	
MYL2	12	rs925368	110,390,979	T	C	0.1060	-0.0731	0.0098	8.4E-14	0.509	rs17550549	111,357,471	2.0E-02	Tin *et al*. 2019
HNF1A	12	rs1169288	121,416,650	A	C	0.4970	-0.0251	0.0043	4.4E-09	1.008	rs1800574	121,416,864	4.2E-09	Tin *et al*. 2019
COMMD4	15	rs140379576	75,656,310	CCCAG	C	0.0842	-0.0504	0.0086	5.2E-09	0.805	rs73436803	75,619,201	3.7E-03	Tin *et al*. 2019
PDILT	16	rs11646437	20,609,702	A	T	0.2774	0.0302	0.0049	8.1E-10	0.952	rs77924615	20,392,332	3.3E-10	Tin *et al*. 2019
MFSD12	19	rs2240751	3,548,231	A	G	0.3952	-0.0330	0.0058	1.4E-08	0.571	N/A	N/A	N/A	

4.4 Discussion

4.4.1 GWAS results for 27 known loci using the 1KG+7K reference panel

A three-way comparison between the results of Kanai et al (2018), and our own GWAS using imputation outputs derived from the 1KG and 1KG+7K reference panels provided an insight into the performance of the 1KG+7K panel, concerning the identification of trait associated loci. For most loci, the identified lead SNPs were comparable in signal strength and MAF across all three sets of results. We did observe a single instance where a locus was defined by a SNP unique to the WGS data of the 1KG+7K reference panel. This particular SNP (rs202213319) is of relatively low MAF in the global population (MAF = 0.20%) but when considering East Asian population groups exclusively it is a common variant (MAF = 18%) (Phan et al, 2020). In both the results of Kanai et al (2018) and the imputation up to the 1KG panel, the lead SNP for this particular locus exhibit similar strengths of association signal and the novel SNP identified in the 1KG+7K panel does not provide any additional information of biological relevance.

As discussed in section 3.4.1, the key features of the 1KG+7K reference panel were the improved imputation quality of low frequency and rare variants, in addition to the increase in the total number of these variants in the reference panel. The majority of trait associated loci in this analysis were defined by common SNPs. As stated in section 3.4.1 there was no substantial improvement in the imputation quality of common SNPs and therefore, the performance of the 1KG and 1KG+7K panels was comparable across the 27 known loci, with their respective lead SNPs displaying similar levels of association.

One point to address was how the removal of multi-allelic sites may impact the ability of the 1KG+7K panel when defining trait associated loci. In the original Kanai et al (2018) results, a single multi-allelic variant was reported to be the lead SNP (rs1828911) for the *HNF4G* locus. Our own imputation up to the 1KG panel identified four loci where the lead SNP was multi-allelic: rs59578826 at the *MUC1* locus, rs150320174 at the *TP53INP1-NDUFAF6* locus, rs10886117 at the *EMX2-RAB11FIP2*

locus, and rs6598541 at the *IGF1R* locus. Whilst multi-allelic sites were removed from the 1KG+7K panel, as shown (Table 4.3.1), it did not impact the performance of the reference panel when identifying evidence of serum uric acid association across the 27 known loci, with equivalent performance between the 1KG and 1KG+7K panels.

4.4.2 Supporting fine mapping with the 1KG+7K reference panel

The fine mapping of complex loci provided the opportunity to explore the additional rare and low frequency SNPs afforded by imputation up to the 1KG+7K panel. A total of 35 signals driven by predominantly rare and low frequency SNPs were distributed across the *NBPF10-NBPF20*, *LRP2*, *SLC2A9*, *ABCG2*, *NRXN2-SLC22A12*, and *LINC01229-LOC102724084* loci. Fine-mapping of the *ABCG2*, *LRP2* and *LINC01229-LOC102724084* loci showed that both were comprised of multiple distinct signals. However, these signals were driven by common variants, thus did not provide ample opportunity to demonstrate the utility of additional information at low MAF ranges afforded by the 1KG+7K panel. The *NBPF10-NBPF20* and *SLC2A9* provided an example of the utility of the 1KG+7K panel because, within each locus, one signal was driven by a SNP present only in the 1KG+7K output, and both of these SNPs were particularly rare (both with MAF < 0.1%) but remained well imputed with both attaining $r^2 > 0.4$.

The *NRXN2-SLC22A12* locus proved to be the most complex locus (figure 4.3.1) we analysed and provided the best example of the benefits for fine-mapping of improved imputation at low MAF ranges. Following a series of conditional analyses on the locus we identified a total of 18 index SNPs linked to independent associations with serum uric acid of which nine were unique to the 1KG+7K. Not only where these nine SNPs absent from the 1KG panel, 8/9 were of extremely low MAF (MAF < 0.05%) yet remained well imputed (minimum $r^2 > 0.4$) due to the significant boost of imputation quality at low MAF ranges when imputing up to the 1KG+7K panel. The benefits of the 1KG+7K at this locus were not however limited to the additional SNPs not found in the 1KG panel. We also observed multiple SNPs with MAF < 0.05% that, whilst present in the 1KG panel, would typically be difficult to impute with a high degree of confidence. Imputation up to the 1KG+7K panel allowed for high quality imputation for these extremely rare variants including: rs565795531 (MAF = 0.05%, $r^2 = 0.643$),

rs549413722 (MAF = 0.02%, r^2 = 0.855), and rs540025991 (MAF = 0.05%, r^2 = 0.878) (full details included in table 4.3.2).

Overall, when performing fine mapping in loci with multiple independent signals the benefits of imputing up to the 1KG+7K panel were more apparent compared to the initial process of identifying known trait associated loci across the genome. At the six loci analysed, we observed improved imputation quality of the rare and low frequency variants implicated, as well as a greater level of detail attainable due to the larger total number of SNPs included in the 1KG=7K panel output.

4.4.3 Novel locus discovery using the 1KG+7K reference panel

After expanding our analysis to search for evidence of novel loci associated with serum uric acid, we were able to compare the capability of both the 1KG and 1KG+7K panels to detect potentially novel trait associated loci. The potentially novel associations we identified (see table 4.3.4) were polymorphic across the 1KG and 1KG+7K panels and so, were not unique to the WGS data of the 1KG+7K panel. The main benefit we observed from imputing up to the 1KG+7K panel was the improved strength of association across all of the 11 potentially novel loci. Furthermore, when imputing up to the 1KG panel, six of these loci did not meet the minimum threshold for genome-wide significance. This improved power to detect associations in addition to the improved imputation quality of the potential loci defined by low frequency SNPs, suggests that imputation up to the 1KG+7K panel provides additional support for the discovery of novel trait associated loci when compared to imputing up to the 1KG panel even when considering a trait that has been subject to thorough investigation.

4.4.4 ABCG2 locus

Disfunction within the *ABCG2* gene is a highly influential and well-known factor linked to increased serum uric acid levels. *ABCG2* encodes a high-capacity urate transporter with expression primarily observed in kidneys, intestines and liver, indicating its role in urate excretion is focussed within these regions (Nakayama et al, 2011). Whilst the

majority of urate excretion is localised in the kidneys, a substantial amount of urate excretions occurs within the intestines. Gene knockout studies in mice indicated that disfunction within the *ABCG2* gene appears to contribute to elevated serum uric acid levels due to a significant decrease in urate excretion within the intestines (potentially up to a 50% decrease) (Takada et al, 2014). Fine mapping of the *ABCG2* locus revealed the functional variant rs72552713, a common stop-gained mutation in the Japanese population. This specific mutation (also referred to as Q126X) was initially discovered through sequencing of the *ABCG2* gene in patients with hyperuricemia and has been linked to drastically reduced protein expression, in turn limiting urate excretion, promoting elevated serum uric (OR = 3.61) acid levels thus, increasing the risk of gout (Matsuo et al, 2009).

4.4.5 NRXN2-SLC22A12 locus

The *SLC22A12* gene encodes the protein URAT1, an organic anion transporter responsible for the regulation of urate levels in blood. *SLC22A12* is predominantly expressed in the epithelial cells of proximal tubules located in the renal cortex (Enomoto et al, 2002). Functional analysis of mutations within *SLC22A12* showed that in contrast with the *ABCG2* gene, mutations within the *SLC22A12* gene are linked to hypouricemia, low serum uric acid levels caused by the impaired reabsorption of urate in the kidneys (Ichida et al, 2004). Our fine mapping of the *NRXN2-SLC22A12* locus supported by imputation up to the 1KG+7K provided a detailed insight into the variants driving the association signal with serum uric acid, specifically in the *SLC22A12* gene where we identified four non-synonymous SNPs. These SNPs were imputed with a high degree of confidence across a range of MAFs (0.06 – 2.23%) and gave insight into the biological relevance of specific mutations within the *SLC22A12* gene.

The SNP rs121907896 (p.Arg90His) represents a rare (MAF = 0.26%) missense variant of the *SLC22A12* gene, initially identified through sequencing studies conducted on patients displaying hypo- and hyperuricemia within the Japanese population and linked to significantly lower serum uric acid levels (Ichida et al, 2004; Iwai et al, 2004). Furthermore, analysis of the p.Arg90His mutation in gout patients further suggest

that due to its promotion of hypouricemia, the mutation can act in a protective capacity against the development of hyperuricemia and gout (Sakiyama et al, 2016). rs58174038 is another rare (MAF = 0.06%), non-synonymous variant, specifically a splice-donor variant of the *SLC22A12* gene. This mutation has been identified in targeted sequencing of whole exon regions of *SLC22A12* in individuals exhibiting hypo- and hyperuricemia, and *in vivo* splicing assays suggest that the mutation causes aberrant splicing likely leading to the loss of function of URAT1 (Zhou et al, 2019; Misawa et al, 2020). The final rare, non-synonymous variant we identified in the *SLC22A12* gene was a missense variant, rs201136391 (MAF = 0.10%). Relatively limited information regarding the role/impact of this mutation is available. However, it has been implicated in the decreased transport function of URAT1 and thus, will likely promote hypouricemia (Toyoda et al, 2021). The final functional variant we identified was a stop gained mutation, rs121907892. This is a common mutation in the Japanese population (MAF = 2.23%) and is a well-known contributor to the development of renal-hypouricemia (Hamajima et al, 2011).

4.4.6 Summary of the performance of the 1KG+7K panel

Overall, through performing GWAS into serum uric acid, we were able to assess the benefits of imputation up to the 1KG+7K panel in a variety of scenarios, in all of which the 1KG+7K panel was successful in demonstrating its capacity to support GWAS.

Whilst minimal improvements were observed in our analysis of 27 known loci, this is expected as these loci are driven by common variant association signals and the 1KG+7K panel did not offer large improvements in the imputation quality of common variants. Furthermore, where the 1KG lead SNP was a multi-allelic variant (multi-allelic variants were excluded from the 1KG+7K panel), an alternative SNP present in the 1KG+7K panel was able to capture the multi-allelic SNP, and so the exclusion of multi-allelic SNPs in the 1KG+7K panel did not result in a loss of power. This resulted in the equivalent performance between the 1KG and 1KG+7K panels when analysing the 27 known loci. The detection of novel loci was seemingly improved when imputing up to the 1KG+7K panel and lead variants for these loci exhibited stronger association signals (based on p values) in comparison to the 1KG panel. Further,

signals driven by rare and low frequency SNPs were imputed with a higher degree of confidence and thus, more reliable.

The greatest benefits of imputing up to the 1KG+7K panel were observed when analysing rare and low frequency SNPs, and so, fine mapping complex loci was an effective way to demonstrate the benefits of improved imputation quality at these lower MAF ranges. The *NRXN2-SLC22A12* locus, in particular, provided an excellent example of the 1KG+7K panel's capabilities and supported the detailed fine mapping of a locus with 18 distinct signals (of which the majority were driven by rare variants). Of these 18 signals, we were able to identify multiple rare functional variants whose mutations were directly linked to hypo/hyperuricemia in previous publications.

The findings from this work suggest that whilst imputing up to the 1KG+7K panel was beneficial, even in GWAS of a thoroughly studied trait, the best application of the panel would appear to be in the analysis of rare and low frequency variants. Therefore, further association analyses with a focus on rare variants would provide the best opportunity to demonstrate the benefits of the improved imputation quality afforded by the 1KG+7K reference panel.

CHAPTER 5

GENE-BASED ASSOCIATION ANALYSIS OF SERUM URIC ACID IN THE JAPANESE POPULATION FOLLOWING IMPUTATION UP TO A JAPANESE POPULATION SPECIFIC REFERENCE PANEL

5.1 Introduction

5.1.1 SNP based association analyses and rare/low frequency variants

Whilst GWAS have been successful in the detection of trait associated loci, the majority of heritability for complex traits cannot be explained through the loci identified through GWAS. The concept of missing heritability overlooked by standard GWAS methodology has been a topic of much discussion in regard to reaching the full potential of these studies and their power to identify trait associated loci. More recent analyses of complex traits suggest that, through improvements in a multitude of factors including study design, sample size and expansion outside of a standard additive model, could lead to a much-improved proportion of heritability explained (with potential estimates as high as ~60%) (Tam et al, 2019). One of these areas for consideration is the presence of a multitude of rare (MAF < 1%) and low frequency ($1\% \leq MAF \leq 5\%$) variants with more modest effect sizes that are more difficult to detect. This problem is exacerbated by a high threshold for significance that is designed to limit false positives in GWAS but unfortunately acts as a barrier for the detection of said low frequency and rare variants (Manolio et al, 2009).

Improvements in reference panel design and the availability of WGS data have driven the progression from smaller, less diverse panels such as the HapMap project to larger reference panels such as the 1KG panel and the HRC panel. These novel panels can encompass a much greater number of variants from a larger and more diverse set of populations and have supported developments in imputing rare and low frequency variants. Additional developments in reference panel design, prompted

the inclusion of WGS data from target populations to supplement pre-existing and publicly available reference panels again, leading to better imputation outcomes for rare and low frequency variants. This matched our observations when imputing up to the novel 1KG+7K imputation panel (discussed in chapter three and further analysed in chapter four). The 1KG+7K panel offered a significant increase in not only the total number of imputed SNPs at low MAF ranges, but also a marked improvement in the imputation quality of these additional SNPs compared to the 1KG panel output. Yet, due to the limitations of standard GWAS much of this additional information may not have been fully utilised due to the typically smaller effect sizes of rare variants.

5.1.2 Gene-based association analysis

The limitations of SNP based association analyses to capture rare and low frequency variants with more modest effect sizes detracts from the plethora of additional data we have at lower MAF ranges as a product of the novel 1KG+7K panel. To capitalise on these additional low frequency and rare variants in our imputation output this chapter will shift the focus to gene-based testing. These analyses will aim to increase the power to detect rare and low frequency variants within genes that may have gone undetected in our previous SNP based analyses. Gene-based association analysis combats the limited detection of rare variants with modest effect sizes by analysing groups of these variants together. By assessing the accumulation of multiple rare variants within a pre-defined region/gene we facilitate the detection of more complex sources of association with a trait of interest (Morris and Zeggini, 2010) Further, this methodology has been demonstrably more effective in simulation studies detecting rare variants associated with the target trait of interest.

There are two main classes of gene-based testing methodologies, burden and dispersion tests. Burden tests work by collapsing the information for a set of variants in a pre-defined region, summarising the region and its association with the trait of interest. These tests are most effective in cases where the individual variants within the region share the same direction of effect. For example, in a region where a series of rare variants within a gene confer a protective effect against a target disease, a

burden test is an effective method by which to identify the association signal. This is of particular relevance to the results described in section 4.4.2 – 4.4.5. These results describe multiple complex loci (most notably the *NRXN2-SLC22A12* locus), comprised of numerous non-synonymous index SNPs with generally uniform effect directions.

In cases where a uniform direction of effect is not expected, dispersion tests present an alternative method for gene-based analysis. Dispersion tests such as the Sequence Kernel Association Test (SKAT) (Wu et al, 2011), take into account the individual variants within a region and their respective summary test statistics, rather than collapsing this information into a single set of summary statistics. This allows for a more accurate analysis of regions where multiple variants with different directions of effect exist. However, when considering regions of uniform effect direction, dispersion tests have a reduced power to detect association in comparison to burden tests.

5.1.3 Chapter aims

This chapter will cover the use of a burden test to explore the additional data produced at lower MAF ranges by imputing up to the 1KG+7K reference panel and testing for association with serum uric acid. The results from the previous SNP based GWAS and subsequent fine mapping of complex loci highlighted the additional imputed rare and low frequency variants afforded by the 1KG+7K imputation, however, these analyses remained limited by the lack of power to detect rare variants when performing standard GWAS. A gene-based testing approach, however, offers an alternative method to explore and identify new signals of association that the previous SNP based analyses may have lacked the power to detect. Furthermore, by incorporating annotation data into these analyses we can limit our findings to nonsynonymous variants to aid our search for biologically relevant mutations.

Association signals identified through gene-based testing were subject to further analyses in relation to the signals identified through GWAS to confirm their independence from the signals already identified in the previous analysis. Finally, this chapter will conclude with the exploration of the composite SNPs of each gene tested

within the burden analyses. By doing so, this chapter assesses the sources of association within the genes tested in our analysis and any potential functional mutations or biological relevancy linked to those variants identified in these analyses.

5.3 Methods

5.3.1 Gene based association testing

The methodology described in Morris & Zeggini (2010) forms the basis of the gene-based association analysis software GRANVIL. This approach models the phenotype in a linear regression framework, as a function of the proportion of rare variants in a gene transcript where a minor allele is present (within an individual). In this test, the phenotype (y_i) of the ith individual can be modelled as shown in equation 5.1.1, given the conditions of equations 5.1.2 and 5.1.3

$$y_i = \mathrm{E}[y_i] + \varepsilon_i$$

Equation 5.1.1

$$\varepsilon_i \sim N(0, \sigma_E)$$

Equation 5.1.2

$$\mathrm{E}[y_i] = \alpha + \lambda \frac{r_i}{n_i} + \beta x_i$$

Equation 5.1.3

Here, α denotes the intercept and σ_E the residual standard deviation. n_i represents the number of variants at which individual i has been genotyped and r_i denotes the number of these variants where individual i has at least one copy of the minor allele. The covariates for individual i are denoted by x_i , whilst β represents the corresponding regression coefficients. λ denotes the expected change in phenotype for a hypothetical individual who carries all minor alleles at the set of tested variants, compared to an individual carrying none.

$$f(y_i|a, \lambda, \beta, r_i, n_i, x_i) = \frac{1}{\sqrt{2\pi\sigma_E^2}} \exp\left[-\frac{(y_i - E[y_i])^2}{2\sigma_E^2}\right]$$

The model by which likelihood contribution of an individual is determined is shown above. These models form the basis of a test of association based on mutation load stemming from an accumulation of minor alleles at the variants tested. Additional simulation studies also confirmed the integrity of the methodology underpinning GRANVIL and showcased the high power to detect association signals driven by rare variants, even when presented with a substantial degree of missing genotype data (Mägi et al, 2011)

GRANVIL (v2.1.1) was used to test for association across a total of 70,663 gene transcripts spanning the autosomes. As with the previous GWAS, testing was completed on composite chunks of each chromosome and in cases where gene transcripts overlapped into more than one chunk, tests were repeated on a concatenation of the two chunks.

Association testing was performed using the default parameters of GRANVIL with a set of notable exceptions. Testing was completed in three separate groups, one for each set of SNPs extracted by the three linkage files denoted list 1, list 2 and list 3 (see table 5.3.1 for details of SNPs). Each list of extracted SNPs was tested with the following MAF and imputation quality (r^2) thresholds; MAF ≤ 5% and r^2 ≥ 0.3, MAF ≤ 5% and r^2 ≥ 0.8, MAF ≤ 0.5% and r^2 ≥ 0.3, MAF ≤ 0.5% and r^2 ≥ 0.8. When considering the threshold for exome-wide significance for this gene-based analysis, we used a Bonferroni correction under the assumption of ~70,000 gene transcripts used in the analysis, mapped to approximately 20,000 genes. The final *p* value threshold used in the analysis was set at 2.5×10^{-6}.

5.3.2 File preparation

The imputation output for 104,174 Japanese individuals imputed up to the 1KG+7K reference panel produced dosage files in the VCF file format. VCF files were converted into gen/sample file format via Plink (v2.0) using the --export oxford function allowing for compatibility with GRANVIL. Phenotype information (Z scores for serum uric acid)

used in the previous GWAS was incorporated into the sample file output by Plink (v2.0). The final output contains Family IDs (FIDS), Individual IDs (IIDs), Sex, Phenotype and covariate information and is referred to as the SAMPLE file. A 'GENELIST' file is also required, containing the following information for each gene transcript: Transcript ID, Chromosome number, Start position and end position (in bp).

Annotation files accompanying the 1KG+7K imputation output provide the basis for what are known as linkage files. Annotation files categorised SNPs in the imputation output into the following categories: frameshift deletion, frameshift insertion, nonframeshift deletion, nonframeshift insertion, nonsynonymous SNV, stopgain, stoploss, synonymous SNV, unknown. Three combinations of the aforementioned categories (see table 5.3.1 below) were used to extract SNPs from the annotation files to form three sperate linkage files in the format of Chromosome/SNP/Genetic_Distance/Position.

Table 5.3.1: Three combinations of SNP classifications used to extract SNPs to be tested in gene-based association testing.

List 1	List 2	List 3
Frame shift deletions	Frame shift deletions	Frame shift deletions
Frame shift insertions	Frame shift insertions	Frame shift insertions
Non-Frame shift deletions	Nonsynonymous	Stop gain
Non-Frame shift insertions	Stop gain	Stop loss
Nonsynonymous	Stop loss	-
Stop gain	-	-
Stop loss	-	-

5.3.3 Conditional analysis

In regions where multiple gene transcripts meet the threshold of exome-wide significance, we employed multiple rounds of gene based conditional analyses (using the --cond_gene option in GRANVIL) to determine which genes are independent of one another by including the burden of any overlapping gene transcripts as a covariate. Each gene within the region was conditioned on all other genes present and with each round of conditioning the gene with the weakest association was dropped from the analysis. This process was repeated until only the genes meeting genome wide significance remained.

5.3.4 Including GWAS index SNPs as covariates

Genes meeting the threshold for exome-wide significance were taken forward into a series of further conditional analyses. GRANVIL allows for conditioning on individual SNPs via the --cond_marker option and so, each gene was conditioned on the index SNP of the nearest loci identified in the serum uric acid GWAS. By incorporating these index SNPs into the regression model, gene transcripts where the association signal is driven either by an index SNP or a SNP in high LD with an index SNP, can be identified. This allows for the identification of gene-based association signals independent of the association signals identified in the serum uric acid GWAS.

5.3.5 Determining variants driving gene-based association signals

For each gene-based signal identified in this analysis, further investigation was completed to determine the variants within the gene that were driving the association signal. Log files included in the GRANVIL output detail the composite SNPs for each gene transcript tested. Using these lists we were able to sequentially test each gene transcript for association whilst excluding a single SNP from the analysis. After testing each individual SNP via exclusion, the SNP whose removal resulted in the greatest change in *p* value would be permanently excluded. This process is repeated with each subsequent iteration excluding an additional SNP until the association signal is fully attenuated and a list of SNPs contributing to said signal is identified.

5.4 Results

5.4.1 Identification of gene-based association signals

Approximately 70,000 gene transcripts were tested for association with serum uric acid in the Japanese population by performing an initial series of GRANVIL runs using default parameters (MAF ≤ 5% and r^2 ≥ 0.3) for the three subsets of functional variants defined in table 5.3.1. Analysis of the most inclusive functional variant subset 'List 1' (Frame shift deletions and insertions, non-frame shift deletions and insertions, non-synonymous, stop gain and stop loss functional variants) showed a total of 54 gene transcripts, encompassing 18 loci mapping to regions on chromosomes 1, 3, 4, 11 and 19, meeting the *p* value threshold of 2.5×10^{-6}. Subsequently, a subset of 52/54 gene transcripts also met exome-wide significance when filtering functional variants for those present in 'list 2' (table 5.3.1). Finally, when limiting analysis to functional variants present in 'list 3' (table 5.3.1) only six gene transcripts (mapping to chromosome 4) were identified. Results from this series of runs highlighted that, with the exception of a series of six gene transcripts on chromosome 4, all gene transcripts meeting exome-wide significance were comprised of SNPs falling within the 'non-synonymous' category defined in the annotation files. Finally, it is important to note that the functional variants included in 'list 2' and list 3' are subsets of those included in 'list 1'. Thus, all association signals captured when limiting analysis to variants in 'list 2' or 'list 3' are also captured when using the variants included in 'list 1'.

We next assessed the impact of more stringent thresholds in regard to MAF and imputation quality (r^2 values) using the subset of functional variants, 'List 1' for reference. Imputation up to the novel 1KG+7K reference panel provided a greater degree of confidence when imputing rare variants. This allowed for a significantly greater threshold for imputation confidence to be employed (r^2 ≥ 0.8 rather than the standard r^2 ≥ 0.3) without losing a large proportion of genes that reached genome wide significance (table 5.4.1). Applying a stricter threshold of MAF ≤ 0.5% appeared to be very restrictive relative to the default parameters of GRANVIL, as few variants appear to meet this MAF threshold. Excluding SNPs above this MAF threshold limited

power of the gene-based analysis as the total number of gene transcripts with $p \leq 2.5 \times 10^{-6}$ from 54 to 11 gene transcripts localised entirely on chromosome 11 (table 5.4.3). Furthermore, applying both thresholds of MAF ≤ 0.5% and $r^2 \geq 0.8$ severely reduces the number of SNPs tested and leaves only three gene transcripts reaching exome-wide significance.

Direct assessment of the results described in tables 5.4.1 and 5.4.2 show that even with the loss of numerous genes reaching exome wide significance when applying a filter of $r^2 \geq 0.8$, those that remain still cover the same regions on chromosomes 1, 3, 4 and 11. This does not necessarily lead to a loss of information regarding association signals as the genes reaching exome-wide significance are clustered within specific regions in their respective chromosomes. This, however, is not true for the signal on chromosome 19 that is lost when applying the $r^2 \geq 0.8$ threshold due to the variants driving the signal of association failing to meet the stricter r^2 threshold.

Table 5.4.1: Gene transcripts with $p \leq 2.5 \times 10^{-6}$, MAF ≤ 5% and $r^2 \geq 0.8$ when filtering for functional variants included in 'List 1'.

Chr	Gene	Start pos.	End pos.	No. of Variants	Average MAF	beta	se	p
1	uc009wrg.1	155,719,509	155,736,500	5	0.00690	0.33	0.06315	1.95E-07
1	uc001fmb.3	155,724,006	155,746,272	3	0.01030	0.23	0.04026	1.48E-08
3	uc003chz.3	38,307,297	38,319,806	4	0.01221	-0.25	0.04377	8.15E-09
3	uc011aym.1	38,307,297	38,317,886	3	0.01603	-0.19	0.03310	4.73E-09
3	uc011ayn.1	38,307,297	38,318,507	4	0.01221	-0.25	0.04377	8.15E-09
4	uc003gmc.2	9,827,847	10,023,114	3	0.00361	-0.40	0.05551	8.81E-13
4	uc003gmd.2	9,827,847	10,041,872	3	0.00361	-0.40	0.05551	8.81E-13
11	uc009ypo.1	64,107,689	64,121,642	9	0.00901	0.50	0.07490	2.84E-11
11	uc001nzy.2	64,107,689	64,125,006	10	0.00814	0.55	0.08280	4.37E-11
11	uc001nzz.1	64,110,202	64,118,167	5	0.01153	0.26	0.04829	4.28E-08
11	uc001oal.1	64,358,281	64,369,825	4	0.00599	-4.75	0.05569	5.77E-15
11	uc001oam.1	64,358,281	64,369,825	4	0.00599	-4.75	0.05569	5.77E-15
11	uc009yps.1	64,358,281	64,369,825	4	0.00599	-4.75	0.05569	5.77E-15
11	uc001oan.1	64,358,281	64,369,825	4	0.00599	-4.75	0.05569	5.77E-15
11	uc009ypr.1	64,358,281	64,366,395	3	0.00787	-3.61	0.04204	6.66E-15
11	uc009ypt.2	64,360,876	64,368,466	4	0.00599	-4.75	0.05569	5.77E-15
11	uc001obx.2	64,662,003	64,684,722	11	0.01528	-1.21	0.06575	1.52E-13
11	uc001ocm.2	64,808,375	64,812,300	4	0.00951	0.27	0.04942	7.56E-08
11	uc001ocn.2	64,812,294	64,826,009	10	0.00232	1.17	0.14644	7.77E-13
11	uc010rnw.1	64,812,294	64,826,014	10	0.00232	1.17	0.14644	7.77E-13
11	uc001ocq.1	64,851,693	64,855,874	4	0.01009	0.27	0.04517	1.45E-09
11	uc001oct.2	64,879,325	64,883,707	6	0.01004	-1.96	0.05756	4.44E-14

Table 5.4.1 (cont.): Gene transcripts with $p \leq 2.5 \times 10^{-6}$, MAF ≤ 5% and $r^2 \geq 0.8$ when filtering for functional variants included in 'List 1'.

Chr	Gene	Start pos.	End pos.	No. of variants	Average MAF	beta	se	p
11	uc010rny.1	64,879,325	64,883,707	6	0.01004	-1.96	0.05756	4.44E-14
11	uc001ocu.2	64,879,325	64,883,707	6	0.01004	-1.96	0.05756	4.44E-14
11	uc001ocv.2	64,879,373	64,883,707	6	0.01004	-1.96	0.05756	4.44E-14
11	uc010rnv.1	64,808,372	64,812,299	4	0.00951	0.27	0.04942	7.56E-08
11	uc009yqb.1	64,883,874	64,885,170	4	0.01657	0.26	0.03554	9.64E-13
11	uc001ocw.2	64,883,874	64,885,170	4	0.01657	0.26	0.03554	9.64E-13

5.4.2 Selection of genes of interest

The results detailed in table 5.4.1 summarise the gene transcripts with $p \leq 2.5 \times 10^{-6}$, MAF ≤ 5% and $r^2 \geq 0.8$ when analysing the functional variants included in 'List 1'. These gene transcripts are highlighted in the below Manhattan plot (figure 5.4.1), summarising our gene-based association analysis using GRANVIL. Where these genes transcripts overlapped, only the transcript with the minimum *p* value was considered for further downstream analyses.

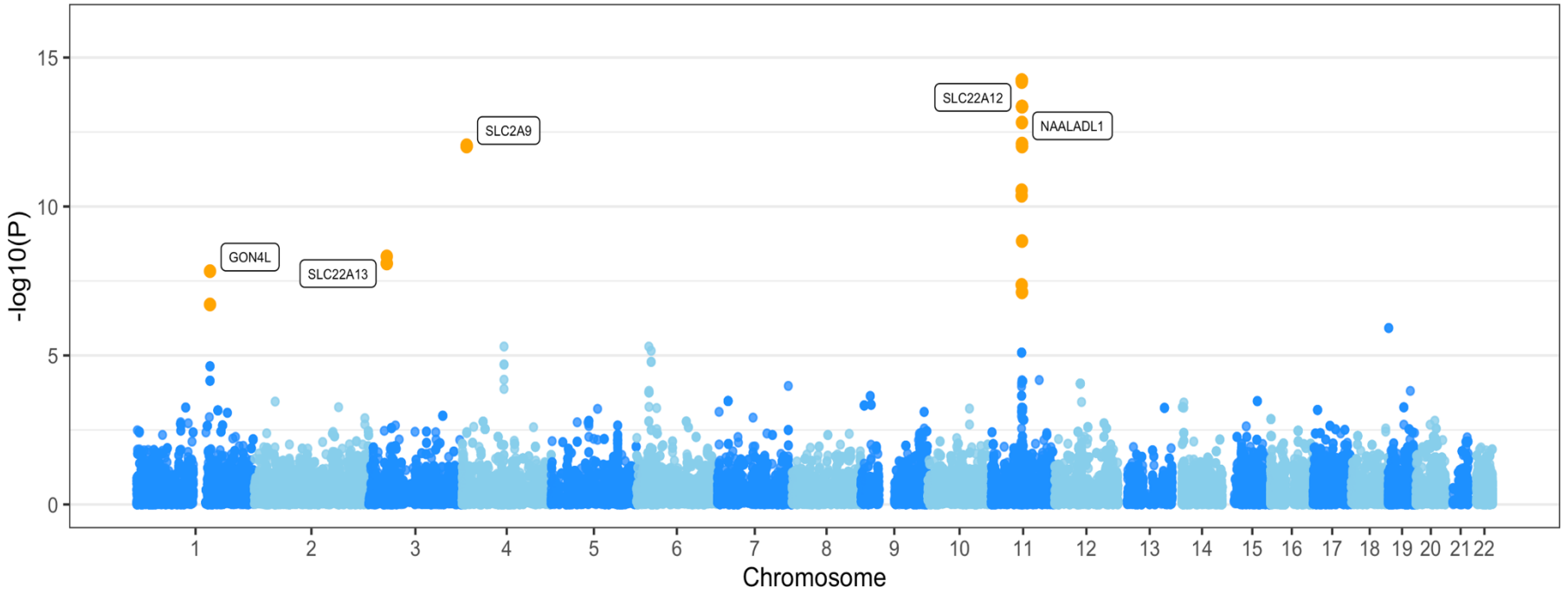

Figure 5.4.1: Manhattan plot of the List 1 subset of functional variants with MAF ≤ 5% and r^2 ≥ 0.8. Each point corresponds to a gene transcript plotted according to its starting position and yellow points represent gene transcripts reaching exome-wide significance ($p \leq 2.5 \times 10^{-6}$). In cases where multiple gene transcripts overlap, the transcript with the lowest p value was taken forwards and is labelled with its corresponding gene symbol.

Table 5.4.2: Finalised list of genes reaching genome wide significance to be subject to further downstream analyses filtering out overlapping genes, duplicates and genes comprised of single markers.

Chr	Transcript	Gene	No. of Variants	Average MAF	beta	se	*p*	Start	End	Closest GWAS position	GWAS Locus
1	uc001fmb.3	***GON4L***	3	0.01030	0.23	0.04026	1.48E-08	155,724,006	155,746,272	155,195,071	MUC1
3	uc011aym.1	**N/A**	3	0.01603	-0.19	0.03310	4.73E-09	38,307,297	38,317,886	38,228,409	OXSR1
4	uc003gmc.2	***SLC2A9***	3	0.00361	-0.40	0.05551	8.81E-13	9,827,847	10,023,114	9,985,376	SLC2A9
11	uc009ypo.1	***CCDC88B***	9	0.00901	0.50	0.07490	2.84E-11	64,107,689	64,121,642	64,361,219	NRXN2-SLC22A12
11	uc001oal.1	***SLC22A12***	4	0.00599	-4.75	0.05569	5.77E-15	64,358,281	64,369,825	64,361,219	NRXN2-SLC22A12
11	uc001obx.2	***ATG2A***	11	0.01528	-1.21	0.06575	1.52E-13	64,662,003	64,684,722	64,361,219	NRXN2-SLC22A12
11	uc001ocn.2	***NAALADL1***	10	0.00232	1.17	0.14644	7.77E-13	64,812,294	64,826,009	64,361,219	NRXN2-SLC22A12
11	uc001ocq.1	***ZFPL1***	4	0.01009	0.27	0.04517	1.45E-09	64,851,693	64,855,874	64,361,219	NRXN2-SLC22A12
11	uc001oct.2	***TM7SF2***	6	0.01004	-1.96	0.05756	4.44E-14	64,879,325	64,883,707	64,361,219	NRXN2-SLC22A12
11	uc009yqb.1	**N/A**	4	0.01657	0.26	0.03554	9.64E-13	64,883,874	64,885,170	64,361,219	NRXN2-SLC22A12

There are multiple gene transcripts with identical summary statistics that occupy similar and often identical regions. These transcripts are all comprised of the same set of SNPs and so only one transcript from each of these groups is taken forward for further analysis. With this in mind, the final list of variants for further analyses is documented in table 5.4.2 above.

5.4.3 Investigating the relationship between multiple exome-wide significant gene transcripts mapping to a shared locus

Table 5.4.1 shows that the association signals observed on chromosomes 1, 3 and 4 are all driven by single gene transcripts. However, the results indicate that multiple gene transcripts attaining exome-wide significance mapped to a shared locus on chromosome 11, specifically to a region spanning 64,358,281 bp – 64,885,170 bp.
To further investigate the relationship between these gene transcripts exhibiting association with serum uric acid we completed a series of conditional analyses (described in section 5.3.3) sequentially excluding individual gene transcripts from the analysis until only gene transcripts independent of others in the region remained (supplementary table A.11 details the step-by-step process below where each individual step involves the exclusion of the gene transcript with the weakest association signal following conditioning on all other overlapping gene transcripts).

Preliminary results from the stepwise condition analysis (supplementary table A.11) identified both uc001oal.1 (*SLC22A12*) and uc001ocn.2 (*NAALADL1*) as the independent signals within this cluster of gene transcripts and that these two genes can account for the association signals from within the region we defined. As an additional step to confirm these findings we completed further conditional analyses within this region. Each gene listed in table 5.4.6 was conditioned on both uc001oal.1 (*SLC22A12*) and uc001ocn.2 (*NAALADL1*) and contrasted with the original unconditioned summary statistics.

Table 5.4.3: Gene transcripts attaining exome-wide significance mapping to chromosome 11 (region: 64,358,281 - 64,885,170 (bp)). Summary statistics from the unconditioned gene-based association analyses are presented first, followed by beta, standard error and *p* value when incorporating gene transcripts uc001oal.1 (*SLC22A12*) and uc001ocn.2 (*NAALADL1*) into the regression model (these columns are denoted by 'Cond.').

Transcript	beta	se	p	Cond. beta	Cond. se	Cond. p
uc009ypo.1	0.50	0.07490	2.84E-11	0.22	0.07253	0.00231
uc001obx.2	-1.21	0.06575	1.52E-13	0.01	0.06535	0.92633
uc001ocq.1	0.27	0.04517	1.45E-09	0.11	0.04562	0.01224
uc001oct.2	-1.96	0.05756	4.44E-14	-0.19	0.06034	0.00139
uc009yqb.1	0.26	0.03554	9.64E-13	0.13	0.03519	0.00024

Table 5.4.3 demonstrates that conditioning on uc001oal.1 (*SLC22A12*) and uc001ocn.2 (*NAALADL1*) attenuates all signals of association in the region spanning 64,358,281 - 64,885,170 (bp) on chromosome 11 below the threshold of exome-wide significance.

5.4.4 Investigating relationships between index SNPs from GWAS analysis and gene-based association signals

Once we had established a list of genes of interest (uc001fmb.3, uc011aym.1, uc003gmc.2, uc001oal.1 and uc001ocn.2) we then sought to determine how loci defined in the previous serum uric acid GWAS and the SNPs driving the association signals within these loci related to the genes highlighted by the gene-based testing approach. Each gene of interest was conditioned on the index SNPs from the nearest loci defined in the serum uric acid GWAS. Therefore, in instances where a locus identified in the GWAS was represented by a single index SNP, the conditional analysis involved only this single SNP as a covariate. However, in complex loci such as *NRXN2-SLC22A12,* where multiple index SNPs were recorded, the gene-based signal was conditioned separately on all index SNPs from that locus.

Table 5.4.4: Gene transcript uc001fmb.3 (*GON4L*) conditioned on the index SNP (rs2990223) of the *MUC1* locus identified in the serum uric acid GWAS.

Condition	beta	se	*p*
None	0.23	0.04026	1.48E-08
rs2990223	0.12	0.04410	0.00766

Table 5.4.5: Gene transcript uc011aym.1 conditioned on the index SNP (rs117297673) of the *OXSR1* locus identified in the serum uric acid GWAS.

Condition	beta	se	*p*
none	-0.19	0.03310	4.73E-09
rs117297673	0.08	0.09998	0.43956

Table 5.4.6: Gene transcript uc003gmc.2 (*SLC2A9*) conditioned on the Index SNPs of the *SLC2A9* locus identified in the serum uric acid GWAS.

Condition	Beta	se	*p*
none	-0.40	0.05551	8.81E-13
rs1026406370	-0.40	0.05547	8.78E-13
rs56050634	-0.42	0.05550	9.06E-13
rs3775948	-0.41	0.05530	9.32E-13
rs7657096	-0.47	0.05556	6.98E-13
rs34325511	-0.41	0.05560	8.38E-13
rs192673734	-0.29	0.06919	2.14E-05

Table 5.4.7: Gene transcript uc001oal.1 (*SLC22A12*) conditioned on the index SNPs of the *NRXN2-SLC22A12* locus identified in the serum uric acid GWAS.

Condition	Beta	se	*p*
none	-4.75	0.05569	5.77E-15
rs1324446966	-4.75	0.05563	5.33E-15
rs565795531	-4.75	0.05568	6.66E-15
11:64317578:A:C	-4.75	0.05566	5.33E-15
11:64328362:A:G	-4.75	0.05568	4.22E-15
rs71581748	-4.74	0.05570	7.11E-15
rs121907896	-4.76	0.05547	5.55E-15
rs58174038	-4.75	0.05565	4.88E-15
rs75786299	-4.74	0.05565	5.77E-15
rs201136391	-4.94	0.05714	5.77E-15
rs121907892	-0.44	0.18847	1.84E-02
rs1291429571	-4.75	0.05568	6.88E-15
rs912882868	-4.75	0.05568	5.33E-15
rs1253297381	-4.75	0.05567	4.88E-15
rs549413722	-4.75	0.05569	6.22E-15
rs540025991	-4.75	0.05569	5.33E-15
rs540570840	-4.75	0.05565	6.66E-15
11:64744403:G:A	-4.75	0.05567	7.11E-15
rs777218029	-4.74	0.05568	5.55E-15

Table 5.4.8: Gene transcript uc001ocn.2 (*NAALADL1*) conditioned on the index SNPs of the *NRXN2-SLC22A12* locus identified in the serum uric acid GWAS.

Condition	Beta	se	p
none	1.17	0.14644	7.77E-13
rs1324446966	1.16	0.14629	8.15E-13
rs565795531	1.17	0.14641	8.07E-13
11:64317578:A:C	1.17	0.14636	8.02E-13
11:64328362:A:G	1.17	0.14642	8.08E-13
rs71581748	1.15	0.14640	7.93E-13
rs121907896	1.15	0.14591	8.29E-13
rs58174038	1.17	0.14633	7.10E-13
rs75786299	0.51	0.15683	1.12E-03
rs201136391	1.17	0.14643	8.36E-13
rs121907892	0.91	0.14125	1.42E-10
rs1291429571	1.17	0.14641	7.87E-13
rs912882868	1.17	0.14641	7.54E-13
rs1253297381	1.17	0.14638	7.11E-13
rs549413722	1.17	0.14644	6.97E-13
rs540025991	1.17	0.14642	8.02E-13
rs540570840	1.17	0.14635	7.21E-13
11:64744403:G:A	1.17	0.14638	7.24E-13
rs777218029	1.17	0.14639	8.12E-13

At the *MUC1* locus, the GWAS analysis identified a single association signal (lead SNP, rs2990223). After conditioning the gene-based association signal, gene transcript uc001fmb.3 (*GON4L*), on this lead SNP, the association signal (determined by p value) changed from $p = 1.48 \times 10^{-08}$ to $p = 0.00766$. Similarly, at the *OXSR1* locus, the GWAS identified a single association signal, attributed to the lead SNP, rs117297673. After conditioning the gene-based association signal (gene transcript uc011aym.1) on SNP rs117297673, the association signal of uc011aym.1 changed from 4.73×10^{-09} to 0.440 Association signals for both uc001fmb.3 (*GON4L*) and uc011aym.1, were attenuated following conditional analysis on the closest GWAS locus, suggesting that they represent the same association signal.

At locus *SLC2A9* identified in the GWAS, there are a total of six independent association signals represented by the index SNPs detailed in table 5.4.6 under the

column ‘Condition’. After conditioning the gene based association signal (gene transcript uc003gmc.2 (*SLC2A9*)) on each of the index SNPs, the results in table 5.4.6 show that conditioning on five of the six index SNPs had little effect on the association signal of transcript uc003gmc.2. However, conditioning on SNP rs192673734 led to a large decrease in the strength of association signal from (p = 8.81 x 10^{-13} to p = 2.14 x 10^{-05}, suggesting that they represent the same association signal.

The *NRXN2-SLC22A12* locus identified in the GWAS of serum uric acid is composed of 18 independent signals attributed to the index SNPs listed in table 5.4.7 under the column ‘Condition’. This locus represents the nearest GWAS locus in relation to both gene transcripts uc001oal.1 (*SLC22A12*) and uc001ocn.2 (*NAALADL1*), and both gene-based association signals were conditioned separately on the 18 index SNPs of the *NRXN2-SLC22A12* locus. Conditioning transcript uc001oal.1 (*SLC22A12*) on 17/18 of the index SNPs prompted minimal change in the association signal. However, conditioning on SNP rs121907892 led to a large decrease in the strength of association signal for transcript uc001oal.1 (*SLC22A12*), prompting a change in p value from p = 5.77 x 10^{-15} to p = 0.00112 and suggests that they represent the same association signal.

Gene transcript uc001ocn.2 (*NAALADL1*) was conditioned on the same set of index SNPs of the *NRXN2-SLC22A12* locus. Conditioning on 16/18 index SNPs led to minimal changes in the strength of association for transcript uc001ocn.2 (*NAALADL1*). Conditioning on SNP rs75786299 resulted in a decrease in association signal with a change in p value of p = 7.77 x 10^{-13} to p = 0.00112, and conditioning on rs121907892 led to a smaller decrease in the strength of association signal with a change of p value from p = 7.77 x 10^{-13} to p = 1.42 x 10^{-10}. This suggests that both the gene transcript uc001ocn.2 and SNP rs75786299 represent the same association signal.

5.4.5 Identifying variants driving gene-based association signals.

Following analysis of the five genes of interest as a sum of their respective SNPs, we proceeded with a series of GRANVIL runs isolating the individual contributions from each SNP within the gene. This analysis was completed to identify instances where gene-based association signals were driven by single SNPs within the gene transcripts identified. These signals, driven by single SNPs are not truly gene-based signals and the purpose of this gene-based analysis is to identify regions/transcripts where multiple SNPs contribute to the association signal. Shown below in table 5.4.9 are the gene transcripts analysed and their respective SNPs.

Table 5.4.9: Gene transcripts of interest and their constituent SNPs.

Transcript	SNPs
uc001fmb.3	rs140449886, rs556357817, rs61748905
uc011aym.1	rs1344774550, rs765934579, rs765934579
uc003gmc.2	rs121908323, 4:9922125:G:A, rs16890979
uc001ocn.2	rs377372068, rs373785662, rs200671505, rs182545269, rs200086647, rs140919277, rs1946989526, rs193248866, rs193248866, rs146841688, rs151287144
uc001oal.1	rs201136391, rs121907892, rs755277288, rs552267750

Table 5.4.10: Gene transcripts subject to sequential exclusion of individual SNPs from within the gene transcript. The original transcript summary statistics are recorded first, followed by the final list of excluded SNPs and the final summary statistics following their exclusion.

Transcript	Gene	Beta	se	*p*	SNP removed	Beta	se	*p*
uc001fmb.3	*GON4L*	0.23	0.04026	1.48E-08	rs140449886	0.00	0.19914	0.22546
uc011aym.1	NA	-0.19	0.03310	4.73E-09	rs765934579	0.13	0.14306	0.34895
uc003gmc.2	*SLC2A9*	-0.40	0.05551	8.81E-13	rs16890979, rs121908323	-0.05	0.08465	0.57719
uc001ocn.2	*NAALADL1*	1.17	0.14644	7.77E-13	rs182545269, rs146841688, rs193248866, rs373785662	0.47	0.28057	0.09732
uc001oal.1	*SLC22A12*	-4.75	0.05569	5.77E-15	rs121907892, rs201136391	-0.10	0.21985	0.65375

Excluding SNP rs140449886 (*GON4L*: Missense Variant) from transcript uc001fmb.3 (Gene: *GON4L*; Mean MAF: 0.01030) during association analysis removes all signs of association stemming from this gene transcript. Based on the complete attenuation of the association signal when excluding SNP rs140449886, the association signal for uc001fmb.3 appears to be driven by this single SNP. Similarly, the results described in table 5.4.10 show the exclusion of SNP rs765934579 (*SLC22A13*: Missense Variant) from uc011aym.1 (Mean MAF: 0.01603) fully attenuates the association signal from the gene.

The exclusion of SNP rs16890979 (*SLC2A9*: Missense Variant) from uc003gmc.2 (*SLC2A9*; Mean MAF: 0.00361) leads to the greatest change in *p* value (a change from $p = 8.81 \times 10^{-13}$ to $p = 9.80 \times 10^{-08}$) in the first series of analyses. The resulting *p* value still suggests that uc003gmc.2 retains exome wide significance even accounting for the exclusion of SNP rs16890979. The additional exclusion of SNP rs121908323 (*SLC2A9*: Missense Variant) further attenuates the association signal, leading to the final *p* value of 0.57719 and so, the association signal from uc003gmc.2 can be accounted for by SNPs rs16890979 and rs121908323.

At the uc001ocn.2 gene transcript, full attenuation of the association signal requires the exclusion of four separate SNPs (rs182545269: *NAALADL1*: Missense Variant, rs146841688: *NAALADL1*: Missense Variant, rs193248866: *NAALADL1*: Missense Variant, and rs373785662: *NAALADL1*: Missense Variant). Removing these four SNPs attenuates the association signal for gene transcript uc001ocn.2 (unconditioned *p* value = 7.77×10^{-13}) with a final *p* value of 0.0973 (table 5.4.10).

Finally, analysis of uc001oal.1 (Gene: *SLC22A12*; Mean MAF: 0.00599) highlighted SNP rs121907892 (*SLC22A12*: Stop Gained) as the major source of association from within the gene transcript. Excluding this specific SNP prompted a change in the *p* value of uc001oal.1 from $p = 7.77 \times 10^{-13}$ to $p = 1.18 \times 10^{-05}$. Further analyses identified SNP rs201136391 (*SLC22A12*: Missense Variant) as the source of the remaining degree of association and exclusion of this SNP led to full attenuation of the association signal stemming from uc001oal.1 with a final *p* value of 0.654. Cross

referencing the SNPs encompassed by gene transcript uc001oal.1 shows that the two SNPs that are the source of association signal (rs121907892 and rs201136391) are index SNPs of the *NRXN2-SLC22A12* locus identified in the GWAS of serum uric acid.

5.5 Discussion

5.5.1 Gene-based association analysis of serum uric acid

Throughout these analyses, we employed a significantly more stringent threshold of r^2 >0.8 in comparison to more traditional value of r^2 >0.3. Whilst this may limit findings in terms of the quantity of SNPs we were able to include within each gene transcript, the majority of genes involved in these analyses have an average MAF below 1%. Not only is it important to ensure that these analyses only incorporated high quality imputed SNPs due to the nature of working with rare and low frequency variants, but we were also able to demonstrate the proficiency of the 1KG+7K reference panel when imputing at the lowest MAF ranges. Although the majority of SNPs involved in these analyses were rare and low frequency variants, the addition of a MAF filter for MAF < 0.5% proved to be too strict in the initial series of association analyses performed and led to a substantial loss of potential SNPs incorporated in the analysis. Similarly, when assessing the three sets of functional variants (List 1, 2 and 3), limiting the analysis to subsets (list 2 and 3) of the functional variants included in list 1 was restrictive in practice and a large proportion of SNPs within gene transcripts were excluded in these analyses. Therefore, list 1 was used as the final set of functional variants used in the association analysis. The final parameters used for the association analysis used default parameters for GRANVIL v2.1.1, MAF < 5%, r^2 > 0.8 and limiting analysis to functional variants included in list 1 (table 5.3.1).

Whilst our previous SNP based investigation into serum uric acid levels was thorough and offered deep investigation into all the signals discovered, encompassing fine mapping across all complex loci, we were still able to detect new association signals unique to our gene-based approach. Following a series of conditional analyses to account for overlapping gene transcripts attaining exome-wide significance, the gene-based association analysis identified five transcripts with potential association

with serum uric acid: uc001fmb.3, mapping to the *MUC1* locus of chromosome 1; uc011aym.1 mapping to the *OXSR1* locus of chromosome 3; uc003gmc.2, mapping to the *SLC2A9* locus of chromosome 4; uc001ocn.2 and uc001oal.1, mapping to the *NRXN2-SLC22A12* locus of chromosome 11.

5.5.2 Non-overlapping gene transcripts at exome-wide significance

The gene transcript uc001fmb.3, mapping to the *MUC1* locus of chromosome 1, is composed of three individual SNPs (table 5.4.9) encompassing the region 155,724,006 – 155746272 (bp), implicating the gene *GON4L*. The closest locus identified in the GWAS of serum uric acid (index SNP rs2990223) is outside of the uc001fmb.3 transcript and does not represent one of the three SNPs of uc001fmb.3. Conditioning on rs2990223, as previously discussed removes the association for uc001fmb.3. Thus, this specific gene transcript and its association signal can be accounted for in the results of the previous SNP based testing approach. Furthermore, localising the sources of association within the genes of interest to specific SNPs highlighted the fact that the association signal from uc001fmb.3 could be explained by a single SNP (rs140449886). These results suggest that uc001fmb.3 is not a true hit in terms of gene-based testing as its association is not driven by a sum of its component SNPs and rather a single SNP in LD (r^2 = 0.298, sourced from the 1KG East Asian subset) with the lead SNP of the nearest GWAS identified locus.

The gene transcript uc011aym.1, mapping to the *OXSR1* locus of chromosome 3, marks the region spanning 38,307,297 – 38,317,886 (bp). The three composite SNPs are detailed in table 5.4.9**.** The nearest locus as identified in the previous GWAS is defined by the index SNP rs117297673, located outside of the uc011aym.1 region. Assessing the individual SNPs within uc011aym.1 showed that again, the association signal attributed to the gene was driven by a single SNP (rs117371763) that, when conditioned on index SNP rs117297673 loses exome-wide significance and the association signal is completely diminished. As with uc001fmb.3, the signal from uc011aym.1 is not a true gene-based association signal and instead is fully explained by the index SNP of the *OXSR1* locus described in the previous GWAS results (r^2 =

0.817 sourced from the 1KG East Asian subset). Irrespective of its eligibility as a gene-based hit, the SNP underpinning the gene-based association signal (rs117371763) does appear to show a high degree of biological relevance to the phenotype in question, serum uric acid. The SNP rs117371763 is a missense variant in the *SLC22A13* gene, which encodes a urate transporter expressed predominantly in the kidneys (Bahn et al, 2008). By using a combination of association analysis performed on exon sequencing data of *SLC22A13* proceeded by a series of *in vitro* cell based experiments, Higashino *et al* (2020) determined that rs117371763 represents a dysfunctional missense variant linked to a decreased levels of serum uric acid and in turn a lower risk of gout.

The gene transcript uc003gmc.2, mapping to the *SLC2A9* locus of chromosome 4, spans the region 9,827,847 – 10,023,114 (bp). The closest locus as defined in the previous GWAS is represented by a total of six index SNPs and defined by the lead SNP rs3775948. Although this locus is located within the boundaries of the uc003gmc.2 gene transcript, there is no overlap with the SNPs representing uc003gmc.2 in this analysis. Whilst conditioning on SNP rs192673734 leads to a reduction in the association signal of uc003gmc.2, the signal is not completely diminished suggesting that the relationship with the locus defined by rs3775948 does not fully explain the association signal from uc003gmc.2.

Assessing the individual SNPs within uc003gmc.2 confirmed that this signal was not derived from a single SNP within the gene transcript, and instead two separate SNPs contributed to the overall association signal. The uc003gmc.2 transcript implicates the gene *SLC2A9*, a fructose transporter that has been observed to play a role in the modulation of serum uric acid levels via renal urate reabsorption (Le MT, et al 2008). The SNP rs16890979 is a missense variant we identified as one of the two SNPs contributing to the association signal of uc003gmc.2, with a MAF of 0.00810. However, in terms of global frequency derived from the 1000 genomes dataset an MAF of 0.264 is documented, making this a much more common mutation outside of the Japanese population. Meta analyses of 11 studies (sum of 1,472 cases and 3,269 controls) from European and Asian populations identified rs16890979 as a protective

mutation against the development of gout due its association with lower levels of serum uric acid (Lee et al, 2017). Further, studies specifically in East Asian cohorts documented the significance of rs16890979 as a low frequency variant with strong associations with hypouricemia (low serum uric acid levels) (Cho et al, 2020).

The SNP rs121908323 is the second signal we observed from uc003gmc.2, with a MAF of 0.00202. This is a particularly rare variant that has been identified as another missense variant within the *SL2A9* gene. In current literature the mutation derived from rs121908323 is referred to as the p.Pro412Arg variant of Glut9 (Anzai et al, 2008). Information regarding this mutation is somewhat ambiguous with studies describing a decrease in activity as a urate transporter and others observing no significant changes regarding this function (Matsuo et al, 2008; Ruiz et al, 2018). Ruiz et al, also made the concession that although they observed no change in function in reference to the p.Pro412Arg variant, the studies referenced were performed *in vitro* on xenopus oocytes and the impact of the p.Pro412Arg variant remains inconclusive.

The gene transcript uc001ocn.2 is the first of two independent signals identified in the gene-based association analysis mapping to the *NRXN2-SLC22A12* locus of chromosome 11. The transcript uc001ocn.2 spans the region 64,812,294 – 64,826,009 (bp) and 10 SNPs were included in our analysis. In comparison to the other four genes involved in the analyses, uc001ocn.2 was by far the most complex, with 4 of the 10 SNPs contributing to the association signal. Conditioning uc001ocn.2 on the *NRXN2-SLC22A12* locus defined by SNP rs121907892 led to some reduction in the strength of association signal but it was not fully attenuated. The transcript uc001ocn.2 implicates the gene *NAALADL1*, encoding an ileal aminopeptidase predominantly expressed in the small intestine (Tykvart et al, 2015). However, no direct link between mutations within *NAALADL1* and serum uric acid levels has been reported to date. The SNPs rs182545269, rs373785662, rs193248866 and rs146841688 are all missense variants within the *NAALADL1* gene with MAF in the Japanese population of 0.00978, 0.00652, 0.00301 and 0.00152 respectively. None of these variants have been previously referenced in any current publication. Thus, we are unable to make any conclusive statements concerning the validity of these

variants and any relation they may have to serum uric acid levels in Japanese populations.

The second signal on chromosome 11 is uc001oal.1, a transcript spanning 64,358,281 – 64,369,825 (bp) encompassing 4 SNPs of which two contribute to the association signal. The transcript uc001oal.1 overlaps the *NRXN2-SLC22A12* locus defined by rs121907892. This locus occupies the same physical location on chromosome 11 as the uc001oal.1 gene transcript and our analyses showed that the two SNPs within uc001oal.1 that drive the gene-based association signal are rs121907892 and rs201136391. These are both index SNPs that were previously identified in the serum uric acid GWAS and represent distinct association signals at the *NRXN2-SLC22A12* locus. Therefore, uc001oal.1 does not represent results derived from a gene-based testing approach as the significant components of this gene transcript had previously been identified in the GWAS of chapter four.

The additional information afforded by our gene-based analyses shows the utility of this approach and its ability to supplement and support traditional SNP based GWAS. Through this we have highlighted additional signals of potential relevance to serum uric acid levels and have done so in a way that took advantage of improved imputation quality at lower MAF ranges offered by the 1KG+7K panel.

CHAPTER 6

DISCUSSION AND CONCLUSION

6.1 Summary

The work within this thesis has explored the utility of reference panels for imputation into GWAS from different population groups. Chapter two provided a detailed comparison of imputation quality attainable from imputation up to four publicly available reference panels for a total of four sample sets of different ancestry groups (African American, East Asian, European and Latino). The results in this chapter highlighted that the developments in reference panel design, predominantly manifesting as much larger reference sample sets (such as those found in the HRC), supported improved imputation quality across all population groups relative to imputation up to older and smaller reference panels (such as the 1000 Genomes project). However, the key observation from this comparison was the disparities in imputation quality between population groups, most noticeable when imputing the sample set comprised of individuals of East Asian ancestry. Whilst many factors can impact the imputation quality (such as the use of the customised genotyping array for EAS samples in this cohort), one of the key considerations is the relationship between the reference sample set and the GWAS sample set, and how poor representation of samples with East Asian ancestry translates to a lower level of imputation quality.

These results are followed by discussion of the utility of population specific reference panels, using subsets of WGS data from GWAS sample sets to supplement existing reference panels to support improved imputation quality. Chapter three presented a comprehensive comparison of imputation quality across four Japanese specific reference panels, each using different combinations of WGS data at different depths and sample sizes. Results indicated that the highest level of imputation quality was derived from a combination of the 1000 Genomes Project (Phase III) and 7000 WGS of mixed depth. Further analysis of the reference panel's capabilities was completed

in chapter four, where a GWAS into serum uric acid in the Japanese population showed that the improved imputation quality supported the discovery of novel trait associated loci and improved fine-mapping of complex loci. However, as the majority of improvements in imputation quality were observed at low frequency and rare variants, chapter five adopted a gene-based approach to association analysis to assess how the improvements in imputation quality supported analysis of low frequency and rare variants. Here gene-based analysis identified multiple gene-based association signals, implicating numerous well imputed rare variants with confirmed biological relevance to serum uric acid level, further supporting the adoption of population specific reference panels and providing examples of improved imputation quality translating to tangible benefits in association analyses.

6.2 Discussion

6.2.1 Further comparisons between reference panels

During the comparisons of imputation quality between reference panels completed in chapter three, the imputation quality derived from four Japanese population specific reference panels (1000 Genomes project and WGS data) was compared to imputation up to the 1000 Genomes project (Phase III) alone. These comparisons are somewhat limited as since the release of the 1000 Genomes Project reference panel (2015) there have been major advancements in the developments of large publicly available reference panels. These include the HRC (2016) and the TOPMed reference panel (2021) which represent conglomerates of established WGS data sets used to create large, diverse reference panels.

The primary limitation that arises when considering multiple reference panels and GWAS sample data sets is the permissions relating to local storage, sharing and uploading to remote servers. As panels such as the HRC and TOPMed panels rely on WGS data collected from numerous sources and studies, management of the final data set and therefore reference panel, in regard to data permissions and the relevant consent is a complicated issue. Whilst both the HRC and TOPMed reference panels are accessible for researchers for the purpose of imputation, they require the

use of remote servers as access to individual level data within the reference panels is not available. Thus, imputation up to these reference panels can only be performed using the Michigan imputation server and the TOPMed imputation server. This method allows for broader access to the reference panels for imputation without giving out access to the reference panel data itself (Taliun et al, 2021).

Establishing remote imputation servers offers additional benefits outside of the regulation and management of access to the WGS data. Servers such as the Michigan imputation server (see section 2.2.3) offer user friendly interfaces that simplify the imputation process, in addition to removing the need for high performance computing clusters, making the imputation of genotyping data more accessible overall. This is particularly important when considering the computational challenges presented when imputing up to large reference panels such as the TOPMed panel. The use of remote servers provides a solution to this as not all researchers will have access to the computational resources required to perform imputation up to such a large reference panel. However, these design choices come with certain drawbacks and limitations including limited control over the imputation process as options and software is limited to those supported by the imputation server. Furthermore, the use of these remote servers is limited in situations where GWAS sample sets cannot be uploaded to said servers due to data sharing restrictions. However, it is important to note that servers such as the Michigan imputation server implement measures to protect data security. These include the use of Hypertext Transfer Protocol Secure (HTTPS) to ensure secure interactions between the user and the server interface, input data for imputation is deleted following its use, encryption of imputation results and time-limited storage of encrypted results (Das et al, 2016).

For reference, the 1000 Genomes Project encompasses 2,504 samples and includes 88 million variants. The release of the HRC provided a drastic increase in total sample size in comparison to the 1000 Genomes Project and is comprised of 32,488 samples of mostly European ancestry, including 39 million variants. TOPMed however, represents a further dramatic increase in sample size, number of variants and the diversity of populations included in the data set with a total of 97,256 samples

including 308 million variants. Since the release of the 1000 Genomes Project the amount of information included in more recently released publicly available reference panels has increased and led to higher quality imputation. Studies employing the use of the TOPMed panel have reported large improvements in imputation quality, especially at low MAF variants. For example, a meta-analysis of type 2 diabetes in Latino populations highlighted the improved imputation quality of the TOPMed panel, noting 5 million imputed variants (MAF 0.1% – 0.05%) with r^2 > 0.8 compared to 1.6 million variants when imputing up to the 1000 Genomes Project panel (Huerta et al, 2021).

The genotyping data used in the imputation comparison and association analyses detailed in this thesis was sourced from Biobank Japan and required a degree of user restrictions. All analyses completed in this thesis were performed locally on the servers in RIKEN, Yokohama, and data was restricted to these servers alone. Due to this, the imputation comparison completed in chapter three was limited to reference panels that were stored locally and so, the HRC and TOPMed reference panels could not be incorporated into these comparisons. Whilst this thesis demonstrates the high quality imputation facilitated by the Japanese population specific reference panels (most notably the 1KG+7K panel) it would be beneficial to place these results in the context of modern publicly available reference panels and compare how smaller reference panels designed to cater to one specific population group compare to reference panels with diverse sample sets almost 10x larger in comparison.

6.2.2 Rare variant analysis

Chapter five explored the use of rare variant analysis to assess the benefits of improved imputation quality at low MAF ranges. The software GRANVIL was used to complete gene-based analysis using a burden test, grouping together rare variants based on a list of available gene transcripts and providing summary statistics for each collection of variants. Observations made in table 4.3.2 highlight the uniform effect direction of the majority of rare variants within each locus. Therefore, performing gene-based association analysis using a burden test would have an increased power to detect associations based on these prior observations.

To further investigate the contribution of rare variants to serum uric acid levels in the Japanese population, performing additional rare variant analyses using alternative tests may be beneficial. In regions where multiple variants with differing effect directions exist, burden tests have a reduced power to detect association (Lee et al, 2014). Dispersion tests offer an alternative method of analysis for regions with multiple variants of differing effect directions that does not suffer the same reduction in power. Repeating the analysis completed in chapter five using methods based on a dispersion test such a SKAT, may offer an additional interpretation of the data potentially highlighting novel associations for further investigation. Using a combination of different tests, given an understanding of their strengths and limitations, may support a more in-depth analysis of rare variants and their contribution to the trait in question.

6.2.3 Imputation quality thresholds

The work throughout this thesis has placed a strong emphasis on the degree of imputation quality achieved through imputation up to a variety of reference panels. For the GWAS completed in chapter four, a minimum threshold of $r^2 > 0.3$ was applied to excluded poorly imputed SNPs. However, the results reported in this chapter did bring attention to the confidence with which each reported index SNP was imputed and the relationship between imputation quality and MAF when imputing up to both the 1KG and 1KG+7K reference panels. The rare variant analysis completed in chapter five applied a considerably stricter threshold for imputation quality, only considering SNPs with $r^2 > 0.8$ due to the additional uncertainty when imputing SNPs with low MAF. Whilst minimum thresholds for imputation quality of $r^2 > 0.3$ are common in GWAS, it is important to consider how the use of population specific reference panels and the analysis of rare variants impacts the effectiveness of standard r^2 thresholds to filter out poorly imputed SNPs.

Pistis et al (2015) completed an additional analysis of imputation quality metrics to supplement their design and subsequent assessment of two population specific reference panels. A subset of SNPs on chromosome 20 were categorised as ‘good

quality' (r^2 > 0.5) and 'bad quality' (r^2 < 0.2) based on r^2 metrics between imputed dosages and true masked genotypes. Following this, the r^2 estimates output from MACH for this subset of SNPs were used to assess the effectiveness of MACH r^2 values in filtering 'bad quality' imputed SNPs whilst minimising the exclusion of 'good quality' imputed SNPs. When considering variants with MAF ≥ 1%, using the standard imputation quality threshold of r^2 > 0.3 was effective at filtering out 'bad quality' imputed SNPs whilst minimising the exclusion of 'good quality' imputed SNPs. In the context of GWAS, this suggests that a preliminary filter of r^2 > 0.3 remains suitable as the majority of loci discovered through GWAS will defined by SNPs with MAF ≥ 1%. This is largely due to the fact that GWAS methodology is generally underpowered to detect association stemming from rare variants. However, for variants with MAF < 1%, their results suggested that a threshold of r^2 > 0.3 is less effective at filtering out poorly imputed SNPs and that a higher threshold (suggested r^2 > 0.6) may be required. This was due to the observation of over-estimated MACH r^2 values for variants at low MAF compared to the 'true' r^2 metrics from the use of the masked genotypes, making it harder to exclude poorly imputed SNPs from analysis using a standard r^2 > 0.3 threshold (Pistis et al, 2015).

The requirements for more stringent filters for imputation quality further highlights the benefits of imputation up to the 1KG+7K reference panel. Not only does this reference panel include a multitude of additional rare variants, but imputation quality was also largely improved for low frequency and rare variants. So much so, that applying a high r^2 threshold of r^2 > 0.8 was feasible in the gene-based analysis of chapter five without severely limiting the detection of association signals.

6.3 Future Work

6.3.1 Adoption of the 1KG+7K reference panel for Biobank Japan data

The GWAS cohort used for the imputation comparison and association analyses completed in this thesis was sourced from Biobank Japan (Nagai et al, 2017). BBJ samples were genotyped on the Illumina HumanOmniExpress and HumanExome arrays and were selected based on the presence of at least one of 47 diseases.

Comprehensive medical records including data from routine check-ups provided phenotype information relevant to the 47 diseases. These data were subject to detailed analysis reported in Kanai et al (2018), were GWAS was performed using a subset of 162,255 Japanese individuals from BBJ imputed up to the 1000 Genomes Project (Phase I, version 3) East-Asian reference haplotypes. They tested for association with a total of 58 quantitative traits, reporting a final total of 1,407 trait associated loci (679 novel).

Chapter four details GWAS into serum uric acid (one of the 58 quantitative traits covered in Kanai et al, 2018) using a subset of 104,174 samples from BBJ imputed up to a combined reference panel of the 1000 Genomes project (phase III) augmented with 7,472 WGS of mixed depth. Kanai et al (2018) reported a total of 27 loci, but the improved imputation capabilities of the 1KG+7K reference panel led to the identification of an additional 11 loci, of which 8 represent potentially novel associations. Furthermore, factoring in the additional information available when exploring complex loci (with several index SNPs unique to the 1KG+7K panel), it becomes clear that the 1KG+7K reference panel supports a more detailed analysis of the genetic component for these traits. The results from the GWAS and subsequent gene-based analysis of serum uric acid suggest that expanding these analyses to encompass the additional 57 quantitative traits for which phenotype information exists, would provide a wealth of additional information and further insight into the 47 diseases included in BBJ.

6.3.2 Investigating the limits of WGS supplementation

Relative to other population specific reference panels discussed in this thesis, including the GoNL panel (N = 769), The Sardianain panel (N = 2,120), The Minnesota panel (N = 1,325), and the EGCUT panel (N = 2,244), the 1KG+7K panel represents a particularly large collection of WGS data sourced from a single population group (Deelen et al, 2014; Pistis et al, 2014; Mitt et al 2017). Among the four Japanese specific reference panels created, the 1KG+7K panel produced the highest quality imputation. However, it is unclear to what extent additional WGS would further improve the imputation quality achievable in comparison to the 1KG+7K panel.

With the development of multiple Japanese specific reference panels using the total WGS data set, we were able to compare the effects of multiple sets of WGS data on imputation quality when augmenting the 1000 Genomes project reference panel. The results in chapter three show that the difference in imputation quality between the 1KG+3K and 1KG+7K panels was considerably smaller than the difference between the 1KG+1K and 1KG+3K panels. If additional WGS data were to become available for the BBJ cohort, the construction of additional reference panels, building on those already created could provide insight into efficient use of WGS data to design population specific reference panels. Additional high depth WGS data may help in identifying the point at which gains in imputation quality are minimal relative to the costs of sourcing WGS data.

Larger population specific reference panels have been designed, such as the ChinaMAP panel, an exceptionally large reference panel (N = 10,155) using high depth (40.80x) WGS data sourced from 14 population groups across China (Li et al, 2021). Whilst producing excellent imputation quality when imputing cohorts of Chinese samples, without producing multiple iterations of the reference panel with progressively increasing amounts of WGS data, the results do not provide insight into the point at which we observe diminishing returns in imputation quality.

6.3.3 Additional assessments of imputation quality

The assessment of imputation quality across the four Japanese population specific reference panels was completed using r^2 metrics calculated by minimac 3 during the imputation process. The results as described in chapter three are thorough and clearly indicate that the 1KG+7K panel produces the highest quality imputation. However, given additional WGS data for a subset of samples within the BBJ GWAS cohort, additional assessments of imputation could be completed to reinforce the results reported.

Some studies, such as Li et al (2021), have completed a direct comparison between imputed and true genotypes. A subset of the GWAS cohort that has been subject to

WGS data can be used to establish a collection of 'true' genotypes. From this subset of additional WGS data, the genotypes found on the genotyping array used for the GWAs cohort (in this case the HumanOmniExpress) can be extracted and subsequently imputed up to the four Japanese specific reference panels. This allows for the direct comparison between 'masked' genotypes in the subset of WGS data with imputed genotype dosages derived from imputing the array subset of genotypes for the same set of samples, with higher quality imputations resulting in a greater degree of concordance between the 'true' and imputed genotypes.

6.4 Concluding remarks

The work completed in this thesis has investigated the disparities in imputation quality between major population groups when imputing up to publicly available reference panels, specifically the poor imputation quality when imputing East Asian cohorts relative to European ancestry samples. This has highlighted the need for additional resources for the imputation of GWAS cohorts from populations underrepresented in currently available reference panels. The development of the 1KG+7K Japanese specific reference panel has provided an excellent resource to support association analyses of the genotype data accumulated in Biobank Japan. Direct comparisons to previous GWAS completed using BBJ GWAS cohorts imputed up to the 1000 Genomes Project have shown that the improved imputation quality has allowed for enhanced discovery of trait associated loci. Large improvements in the imputation quality of low frequency and rare variants has opened up additional avenues for investigating the genetic component of complex traits and allowed for the analysis of high quality imputed SNPs at low MAF ranges and as demonstrated in chapter five this provides additional information regarding trait associated loci. Overall, the 1KG+7K reference panel has proven to be an excellent resource that will facilitate improved analysis of the 47 diseases recorded in Biobank Japan. The enhanced imputation quality supports the identification of additional trait associated loci as well as improved fine-mapping capabilities. These factors allow for two routes to clinical translation: 1) the improved localisation of causal genes, some of which may present the potential for novel drug targets offering new routes to treatment,

and 2) improved opportunities for risk prediction through the development of polygenic risk scores to identify at risk individuals.

BIBLIOGRAPHY

Al-Tassan, N. A., Whiffin, N., Hosking, F. J., Palles, C., Farrington, S. M., Dobbins, S. E., ... & Houlston, R. S. (2015). A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer. *Scientific reports*, *5*(1), 1-11.

Anzai, N., Ichida, K., Jutabha, P., Kimura, T., Babu, E., Jin, C. J., ... & Sakurai, H. (2008). Plasma urate level is directly regulated by a voltage-driven urate efflux transporter URATv1 (SLC2A9) in humans. *Journal of biological chemistry*, *283*(40), 26834-26838.

Bahn, A., Hagos, Y., Reuter, S., Balen, D., Brzica, H., Krick, W., ... & Burckhardt, G. (2008). Identification of a new urate and high affinity nicotinate transporter, hOAT10 (SLC22A13). *Journal of Biological Chemistry*, *283*(24), 16332-16341.

Banda, Y., Kvale, M. N., Hoffmann, T. J., Hesselson, S. E., Ranatunga, D., Tang, H., ... & Risch, N. (2015). Characterizing race/ethnicity and genetic ancestry for 100,000 subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort. *Genetics*, *200*(4), 1285-1295.

Browning, B. L., & Browning, S. R. (2016). Genotype imputation with millions of reference samples. *The American Journal of Human Genetics*, *98*(1), 116-126.

Browning, S. R., & Browning, B. L. (2011). Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, *12*(10), 703-714.

Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D., & Mountain, J. L. (2015). The genetic ancestry of african americans, latinos, and european Americans across the United States. *The American Journal of Human Genetics*, *96*(1), 37-53.

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS computational biology*, *8*(12), e1002822.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, *4*(1), s13742-015.

Cho, S. K., Kim, B., Myung, W., Chang, Y., Ryu, S., Kim, H. N., ... & Won, H. H. (2020). Polygenic analysis of the effect of common and low-frequency genetic variants on serum uric acid levels in Korean individuals. *Scientific reports*, *10*(1), 1-10.

Coleman, J. R., Euesden, J., Patel, H., Folarin, A. A., Newhouse, S., & Breen, G. (2016). Quality control, imputation and analysis of genome-wide genotyping data from the Illumina HumanCoreExome microarray. *Briefings in functional genomics*, *15*(4), 298-304.

Cook, J. P., & Morris, A. P. (2016). Multi-ethnic genome-wide association study identifies novel locus for type 2 diabetes susceptibility. *European Journal of Human Genetics*, *24*(8), 1175-1180.

Dalbeth, N., & Haskard, D. O. (2005). Mechanisms of inflammation in gout. *Rheumatology*, *44*(9), 1090-1096.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., ... & Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature genetics*, *48*(10), 1284-1287.

Daya, M., Rafaels, N., Brunetti, T. M., Chavan, S., Levin, A. M., Shetty, A., ... & Barnes, K. C. (2019). Association study in African-admixed populations across the Americas recapitulates asthma risk loci in non-African populations. *Nature communications*, *10*(1), 1-13.

Deelen, P., Menelaou, A., Van Leeuwen, E. M., Kanterakis, A., Van Dijk, F., Medina-Gomez, C., ... & Swertz, M. A. (2014). Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'. *European Journal of Human Genetics*, *22*(11), 1321-1326.

Delaneau, O., & Marchini, J. (2014). Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nature communications*, *5*(1), 1-9.

Delaneau, O., Marchini, J., & Zagury, J. F. (2012). A linear complexity phasing method for thousands of genomes. *Nature methods*, *9*(2), 179-181.

Durbin, R. (2014). Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). *Bioinformatics*, *30*(9), 1266-1272.

Enomoto, A., Kimura, H., Chairoungdua, A., Shigeta, Y., Jutabha, P., Ho Cha, S., ... & Endou, H. (2002). Molecular identification of a renal urate–anion exchanger that regulates blood urate levels. *Nature*, *417*(6887), 447-452.

Fadista, J., Manning, A. K., Florez, J. C., & Groop, L. (2016). The (in) famous GWAS P-value threshold revisited and updated for low-frequency variants. *European Journal of Human Genetics*, *24*(8), 1202-1205.

Ferrucci, L., Bandinelli, S., Benvenuti, E., Di Iorio, A., Macchi, C., Harris, T. B., & Guralnik, J. M. (2000). Subsystems contributing to the decline in ability to walk: bridging the gap between epidemiology and geriatric practice in the InCHIANTI study. *Journal of the American Geriatrics Society*, *48*(12), 1618-1625.

Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. *Nature Reviews Genetics*, *7*(2), 85-97.

Fuchsberger, C., Abecasis, G. R., & Hinds, D. A. (2015). minimac2: faster genotype imputation. *Bioinformatics*, *31*(5), 782-784.

GenomeAsia100K Consortium. (2019). The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature*, *576*(7785), 106.

Grada, A., & Weinbrecht, K. (2013). Next-generation sequencing: methodology and application. *The Journal of investigative dermatology*, *133*(8), e11.

Hamajima, N., Naito, M., Hishida, A., Okada, R., Asai, Y., & Wakai, K. (2011). Serum uric acid distribution according to SLC22A12 W258X genotype in a cross-sectional study of a general Japanese population. *BMC medical genetics*, *12*(1), 1-6.

Higashino, T., Morimoto, K., Nakaoka, H., Toyoda, Y., Kawamura, Y., Shimizu, S., ... & Matsuo, H. (2020). Dysfunctional missense variant of OAT10/SLC22A13 decreases gout risk and serum uric acid levels. *Annals of the rheumatic diseases*, *79*(1), 164-166.

Hoffmann, T. J., Kvale, M. N., Hesselson, S. E., Zhan, Y., Aquino, C., Cao, Y., ... & Risch, N. (2011). Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. *Genomics*, *98*(2), 79-89.

Hoffmann, T. J., & Witte, J. S. (2015). Strategies for imputing and analyzing rare variants in association studies. *Trends in Genetics*, *31*(10), 556-563.

Hoffmann, T. J., Zhan, Y., Kvale, M. N., Hesselson, S. E., Gollub, J., Iribarren, C., ... & Risch, N. (2011). Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. *Genomics*, *98*(6), 422-430.

Hofman, A., Brusselle, G. G., Murad, S. D., van Duijn, C. M., Franco, O. H., Goedegebure, A., ... & Vernooij, M. W. (2015). The Rotterdam Study: 2016 objectives and design update. *European journal of epidemiology*, *30*(8), 661-708.

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics*, *5*(6), e1000529.

Huerta, A., Cole, J. B., Dornbos, P., Dicorpo, D. A., Leong, A., Florez, J. C., ... & MX, B. (2021). 245-OR: Comprehensive Genome-Wide Association Study (GWAS) Meta-analysis Using TOPMed Imputation in Latinos Identifies Rare Variation Associated with Type 2 Diabetes (T2D). *Diabetes*, *70*(Supplement_1).

Hunt, R., Sauna, Z. E., Ambudkar, S. V., Gottesman, M. M., & Kimchi-Sarfaty, C. (2009). Silent (synonymous) SNPs: should we care about them?. *Single nucleotide polymorphisms*, 23-39.

Ichida, K., Hosoyamada, M., Hisatome, I., Enomoto, A., Hikita, M., Endou, H., & Hosoya, T. (2004). Clinical and molecular analysis of patients with renal hypouricemia in Japan-influence of URAT1 gene on urinary urate excretion. *Journal of the American Society of Nephrology*, *15*(1), 164-173.

Iglesias, A. I., Van Der Lee, S. J., Bonnemaijer, P. W., Höhn, R., Nag, A., Gharahkhani, P., ... & van Duijn, C. M. (2017). Haplotype reference consortium panel: Practical implications of imputations with large reference panels. *Human mutation*, *38*(8), 1025-1032.

Iwai, N., Mino, Y., Hosoyamada, M., Tago, N., Kokubo, Y., & Endou, H. (2004). A high prevalence of renal hypouricemia caused by inactive SLC22A12 in Japanese. *Kidney international*, *66*(3), 935-944.

Jansen, P. R., Watanabe, K., Stringer, S., Skene, N., Bryois, J., Hammerschlag, A. R., ... & Posthuma, D. (2019). Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. *Nature genetics*, *51*(3), 394-403.

Johnston, H. R., Hu, Y. J., Gao, J., O'Connor, T. D., Abecasis, G. R., Wojcik, G. L., ... & Qin, Z. S. (2017). Identifying tagging SNPs for African specific genetic variation from the African Diaspora Genome. *Scientific reports*, *7*(1), 1-9.

Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., ... & Kamatani, Y. (2018). Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nature genetics*, *50*(3), 390-400.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., ... & Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science*, *308*(5720), 385-389.

Köttgen, A., Albrecht, E., Teumer, A., Vitart, V., Krumsiek, J., Hundertmark, C., ... & Ernst, F. (2013). Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nature genetics*, *45*(2), 145-154.

Ku, C. S., Loy, E. Y., Salim, A., Pawitan, Y., & Chia, K. S. (2010). The discovery of human genetic variations and their use as disease markers: past, present and future. *Journal of human genetics*, *55*(7), 403-415.

Kuo, C. F., Grainge, M. J., Zhang, W., & Doherty, M. (2015). Global epidemiology of gout: prevalence, incidence and risk factors. *Nature reviews rheumatology*, *11*(11), 649-662.

Kvale, M. N., Hesselson, S., Hoffmann, T. J., Cao, Y., Chan, D., Connell, S., ... & Risch, N. (2015). Genotyping informatics and quality control for 100,000 subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort. *Genetics*, *200*(4), 1051-1060.

LaFramboise, T. (2009). Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic acids research*, *37*(13), 4181-4193.

Le, M. T., Shafiu, M., Mu, W., & Johnson, R. J. (2008). SLC2A9—a fructose transporter identified as a novel uric acid transporter. *Nephrology Dialysis Transplantation*, *23*(9), 2746-2749.

Lee, S., Abecasis, G. R., Boehnke, M., & Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics*, *95*(1), 5-23.

Lee, Y. H., Seo, Y. H., Kim, J. H., Choi, S. J., Ji, J. D., & Song, G. G. (2017). Associations between SLC2A9 polymorphisms and gout susceptibility. *Zeitschrift für Rheumatologie*, *76*(1), 64-70.

Li, L., Huang, P., Sun, X., Wang, S., Xu, M., Liu, S., ... & Wang, W. (2021). The ChinaMAP reference panel for the accurate genotype imputation in Chinese populations. *Cell Research*, *31*(12), 1308-1310.

Li, M., Li, C., & Guan, W. (2008). Evaluation of coverage variation of SNP chips for genome-wide association studies. *European journal of human genetics*, *16*(5), 635-643.

Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic epidemiology*, *34*(8), 816-834.

Lin, Y., Liu, L., Yang, S., Li, Y., Lin, D., Zhang, X., & Yin, X. (2018). Genotype imputation for Han Chinese population using Haplotype Reference Consortium as reference. *Human Genetics*, *137*(6), 431-436.

Loh, P. R., Danecek, P., Palamara, P. F., Fuchsberger, C., Reshef, Y. A., Finucane, H. K., ... & Price, A. L. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nature genetics*, *48*(11), 1443-1448.

Mägi, R., Kumar, A., & Morris, A. P. (2011, December). Assessing the impact of missing genotype data in rare variant association analysis. In *BMC proceedings* (Vol. 5, No. 9, pp. 1-6). BioMed Central.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., ... & Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747-753.

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, *11*(7), 499-511.

Marchini, J., Howie, B., Myers, S., McVean, G., & Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics*, *39*(7), 906-913.

Matsuo, H., Takada, T., Ichida, K., Nakamura, T., Nakayama, A., Ikebuchi, Y., ... & Shinomiya, N. (2009). Common defects of ABCG2, a high-capacity urate exporter, cause gout: a function-based genetic analysis in a Japanese population. *Science translational medicine*, *1*(5), 5ra11-5ra11.

Matsuo, H., Chiba, T., Nagamori, S., Nakayama, A., Domoto, H., Phetdee, K., ... & Shinomiya, N. (2008). Mutations in glucose transporter 9 gene SLC2A9 cause renal hypouricemia. *The American Journal of Human Genetics*, *83*(6), 744-751.

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., ... & Haplotype Reference Consortium. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics*, *48*(10), 1279.

Misawa, K., Hasegawa, T., Mishima, E., Jutabha, P., Ouchi, M., Kojima, K., ... & Nagasaki, M. (2020). Contribution of rare variants of the SLC22A12 gene to the missing heritability of serum urate levels. *Genetics*, *214*(4), 1079-1090.

Mitt, M., Kals, M., Pärn, K., Gabriel, S. B., Lander, E. S., Palotie, A., ... & Palta, P. (2017). Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *European Journal of Human Genetics*, *25*(7), 869-876.

Morris, A. P., & Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic epidemiology*, *34*(2), 188-193.

Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., ... & Kubo, M. (2017). Overview of the BioBank Japan Project: study design and profile. *Journal of epidemiology*, *27*(Supplement_III), S2-S8.

Nakatochi, M., Kanai, M., Nakayama, A., Hishida, A., Kawamura, Y., Ichihara, S., ... & Matsuo, H. (2019). Genome-wide meta-analysis identifies multiple novel loci associated with serum uric acid levels in Japanese individuals. *Communications biology*, *2*(1), 1-10.

Nakayama, A., Matsuo, H., Takada, T., Ichida, K., Nakamura, T., Ikebuchi, Y., ... & Shinomiya, N. (2011). ABCG2 is a high-capacity urate transporter and its genetic impairment increases serum uric acid levels in humans. *Nucleosides, Nucleotides and Nucleic Acids*, *30*(12), 1091-1097.

National Human Genome Research Institute (2021, November, 1). DNA Sequencing Costs: Data. National Human Genome Research Institute.

https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

Neale, B. M., & Sham, P. C. (2004). The future of association studies: gene-based analysis and replication. *The American Journal of Human Genetics*, *75*(3), 353-362.

Nelson, M. R., Bryc, K., King, K. S., Indap, A., Boyko, A. R., Novembre, J., ... & Lai, E. H. (2008). The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. *The American Journal of Human Genetics*, *83*(3), 347-358.

Nelson, S. C., Stilp, A. M., Papanicolaou, G. J., Taylor, K. D., Rotter, J. I., Thornton, T. A., & Laurie, C. C. (2016). Improved imputation accuracy in Hispanic/Latino populations with larger and more diverse reference panels: applications in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). *Human molecular genetics*, *25*(15), 3245-3254.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., ... & Bustamante, C. D. (2008). Genes mirror geography within Europe. *Nature*, *456*(7218), 98-101.

O'Connell, J., Sharp, K., Shrine, N., Wain, L., Hall, I., Tobin, M., ... & Marchini, J. (2016). Haplotype estimation for biobank-scale data sets. *Nature genetics*, *48*(7), 817-820.

O'Connell, J., Yun, T., Moreno, M., Li, H., Litterman, N., Kolesnikov, A., ... & McLean, C. Y. (2021). A population-specific reference panel for improved genotype imputation in African Americans. *Communications biology*, *4*(1), 1-9.
Pääbo, S. (2003). The mosaic that is our genome. *Nature*, *421*(6921), 409-412.

Pe'er, I., Yelensky, R., Altshuler, D., & Daly, M. J. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, *32*(4), 381-385.

Phan, L., Jin, Y., Zhang, H., Qiang, E., Shekhtman, E., Shao, D., Reyoe, D., Villamarin, R., Ivanchenko, E., Kimura, M., Wang, Z.Y., Hao, L., Sharapova, N., Bihan, M., Sturcke, A., Lee, M., Popova, N., Wu, W., Bastiani, C., Ward, M.,…& Kattman, B.L. (2020). ALFA: Allele Frequency Aggregator. National Center for Biotechnology Information, U.S. National Library of Medicine. www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa/.

Pistis, G., Porcu, E., Vrieze, S. I., Sidore, C., Steri, M., Danjou, F., ... & Sanna, S. (2015). Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. *European Journal of Human Genetics*, *23*(7), 975-983.

Popejoy, A. B., & Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature*, *538*(7624), 161-164.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, *38*(8), 904-909.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, *81*(3), 559-575.

Ruiz, A., Gautschi, I., Schild, L., & Bonny, O. (2018). Human mutations in SLC2A9 (Glut9) affect transport capacity for urate. *Frontiers in physiology*, *9*, 476.

Sakiyama, M., Matsuo, H., Shimizu, S., Nakashima, H., Nakamura, T., Nakayama, A., ... & Shinomiya, N. (2016). The effects of URAT1/SLC22A12 nonfunctional variants, R90H and W258X, on serum uric acid levels and gout/hyperuricemia progression. *Scientific reports*, *6*(1), 1-6.

Schaid, D. J., Chen, W., & Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, *19*(8), 491-504.

Scheet, P., & Stephens, M. (2006). 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.

Tabor, H. K., Risch, N. J., & Myers, R. M. (2002). Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nature Reviews Genetics*, *3*(5), 391-397.

Takada, T., Ichida, K., Matsuo, H., Nakayama, A., Murakami, K., Yamanashi, Y., ... & Suzuki, H. (2014). ABCG2 dysfunction increases serum uric acid by decreased intestinal urate excretion. *Nucleosides, Nucleotides and Nucleic Acids*, *33*(4-6), 275-281.

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., ... & Stilp, A. M. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, *590*(7845), 290-299.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, *20*(8), 467-484.

1000 Genomes Project Consortium. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56-65.

1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68.

Tin, A., Marten, J., Halperin Kuhns, V. L., Li, Y., Wuttke, M., Kirsten, H., ... & Pistis, G. (2019). Target genes, variants, tissues and transcriptional pathways influencing human serum urate levels. *Nature genetics*, *51*(10), 1459-1474.

Toyoda, Y., Kawamura, Y., Nakayama, A., Nakaoka, H., Higashino, T., Shimizu, S., ... & Matsuo, H. (2021). Substantial anti-gout effect conferred by common and rare dysfunctional variants of URAT1/SLC22A12. *Rheumatology*, *60*(11), 5224-5232.

Turner, S., Armstrong, L. L., Bradford, Y., Carlson, C. S., Crawford, D. C., Crenshaw, A. T., ... & Ritchie, M. D. (2011). Quality control procedures for genome-wide association studies. *Current protocols in human genetics*, *68*(1), 1-19.

Tykvart, J., Bařinka, C., Svoboda, M., Navrátil, V., Souček, R., Hubálek, M., ... & Konvalinka, J. (2015). Structural and biochemical characterization of a novel aminopeptidase from human intestine. *Journal of Biological Chemistry*, *290*(18), 11321-11336.

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., ... & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1), 1-21.

van Duijn, CM, Zillikens MC (2005, February) Erasmus Rucphen Family study. ErasmusMC. http://www.gefos.org/?q=content/erasmus-rucphen-family-study-0

Vergara, C., Parker, M. M., Franco, L., Cho, M. H., Valencia-Duarte, A. V., Beaty, T. H., & Duggal, P. (2018). Genotype imputation performance of three reference panels using African ancestry individuals. *Human genetics*, *137*(4), 281-292.

Wang, Y., Lu, J., Yu, J., Gibbs, R. A., & Yu, F. (2013). An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data. *Genome research*, *23*(5), 833-842.

Weale, M. E. (2010). Quality control for genome-wide association studies. *Genetic variation*, 341-372.

Winnard, D., Wright, C., Taylor, W. J., Jackson, G., Te Karu, L., Gow, P. J., ... & Dalbeth, N. (2012). National prevalence of gout derived from administrative health data in Aotearoa New Zealand. *Rheumatology*, *51*(5), 901-909.

Wray, N. R., Wijmenga, C., Sullivan, P. F., Yang, J., & Visscher, P. M. (2018). Common disease is more complex than implied by the core gene omnigenic model. *Cell*, *173*(7), 1573-1580.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, *89*(1), 82-93.

Zhou, Z., Wang, K., Zhou, J., Wang, C., Li, X., Cui, L., ... & Shi, Y. (2019). Amplicon targeted resequencing for SLC2A9 and SLC22A12 identified novel mutations in hypouricemia subjects. *Molecular genetics & genomic medicine*, *7*(7), e00722.

APPENDIX

Imputation comparisons across four reference panels

Table A.1: AFR cohort imputed up to four reference panels available on the Michigan imputation server with imputed variants categorised by r^2 thresholds.

Reference panel	MAF category	Number of variants		
		Total	$r^2 \geq 0.3$	$r^2 \geq 0.8$
1000 Genomes	MAF < 0.01%	3,230,607	0	0
	0.01% ≤ MAF < 0.1%	15,518,355	1,794,179	144,463
	0.1% ≤ MAF < 0.5%	9,542,011	5,444,253	583,224
	0.5% ≤ MAF < 1%	3,291,280	2,493,596	418,358
	1% ≤ MAF < 2.5%	3,864,814	3,215,073	719,055
	2.5% ≤ MAF < 5%	2,588,430	2,257,582	634,488
	5% ≤ MAF < 10%	2,383,465	2,106,008	707,231
	MAF ≥ 10%	6,471,392	5,819,996	2,292,074
CAAPA	MAF < 0.01%	66,212	0	0
	0.01% ≤ MAF < 0.1%	3,425,688	175,697	8,358
	0.1% ≤ MAF < 0.5%	9,598,355	1,249,843	52,769
	0.5% ≤ MAF < 1%	3,069,352	801,697	49,784
	1% ≤ MAF < 2.5%	3,596,355	1,262,159	105,620
	2.5% ≤ MAF < 5%	2,381,112	1,052,174	129,111
	5% ≤ MAF < 10%	2,171,199	1,122,591	188,604
	MAF ≥ 10%	5,343,560	3,448,640	847,206
GAsPv1.0	MAF < 0.01%	116,525	0	0
	0.01% ≤ MAF < 0.1%	3,428,804	83,297	2,784
	0.1% ≤ MAF < 0.5%	5,503,031	823,038	18,766
	0.5% ≤ MAF < 1%	1,872,488	875,528	41,883
	1% ≤ MAF < 2.5%	2,358,061	1,702,379	199,868
	2.5% ≤ MAF < 5%	1,649,488	1,364,842	250,799
	5% ≤ MAF < 10%	1,579,311	1,326,847	293,639
	MAF ≥ 10%	4,567,296	3,980,276	1,142,825
HRC	MAF < 0.01%	10,686,253	0	0
	0.01% ≤ MAF < 0.1%	7,279,073	2,464,059	461,563
	0.1% ≤ MAF < 0.5%	5,724,890	5,086,895	840,988
	0.5% ≤ MAF < 1%	2,637,017	2,543,762	823,270
	1% ≤ MAF < 2.5%	3,283,492	3,215,337	1,408,402
	2.5% ≤ MAF < 5%	2,185,427	2,147,093	1,134,410
	5% ≤ MAF < 10%	1,953,202	1,903,334	1,097,931
	MAF ≥ 10%	4,921,791	4,800,298	2,790,600

Table A.2: EAS cohort imputed up to four reference panels available on the Michigan imputation server with imputed variants categorised by r^2 thresholds.

Reference panel	MAF category	Number of variants		
		Total	$r^2 \geq 0.3$	$r^2 \geq 0.8$
1000 Genomes	MAF < 0.01%	8,136,097	0	0
	0.01% ≤ MAF < 0.1%	17,116,272	721,326	47,138
	0.1% ≤ MAF < 0.5%	9,665,721	1,705,024	81,685
	0.5% ≤ MAF < 1%	2,260,896	595,523	45,483
	1% ≤ MAF < 2.5%	1,891,372	731,553	117,287
	2.5% ≤ MAF < 5%	1,142,076	637,553	184,322
	5% ≤ MAF < 10%	1,323,794	882,483	341,645
	MAF ≥ 10%	5,354,878	4,506,519	2,009,219
CAAPA	MAF < 0.01%	1,221,863	0	0
	0.01% ≤ MAF < 0.1%	5,826,983	82,365	6,131
	0.1% ≤ MAF < 0.5%	8,060,924	196,931	11,455
	0.5% ≤ MAF < 1%	2,915,712	139,478	7,351
	1% ≤ MAF < 2.5%	3,184,115	290,445	24,141
	2.5% ≤ MAF < 5%	1,916,464	342,264	45,496
	5% ≤ MAF < 10%	1,669,945	498,871	95,087
	MAF ≥ 10%	4,659,849	2,683,341	704,390
GAsPv1.0	MAF < 0.01%	1,519,547	0	0
	0.01% ≤ MAF < 0.1%	5,498,721	281,122	39,487
	0.1% ≤ MAF < 0.5%	6,205,702	765,914	63,321
	0.5% ≤ MAF < 1%	1,350,473	379,530	40,151
	1% ≤ MAF < 2.5%	1,125,533	497,040	104,878
	2.5% ≤ MAF < 5%	761,801	461,906	162,975
	5% ≤ MAF < 10%	967,414	657,327	287,007
	MAF ≥ 10%	3,887,619	3,293,900	1,578,020
HRC	MAF < 0.01%	15,842,552	0	0
	0.01% ≤ MAF < 0.1%	10,231,932	1,582,137	120,194
	0.1% ≤ MAF < 0.5%	4,354,222	2,281,439	107,403
	0.5% ≤ MAF < 1%	1,281,338	891,562	78,048
	1% ≤ MAF < 2.5%	1,290,970	1,012,335	175,195
	2.5% ≤ MAF < 5%	869,085	715,368	240,059
	5% ≤ MAF < 10%	1,038,515	850,452	402,809
	MAF ≥ 10%	4,026,316	3,800,727	2,093,114

Table A.3: EUR cohort imputed up to four reference panels available on the Michigan imputation server with imputed variants categorised by r^2 thresholds.

Reference panel	MAF category	Number of variants		
		Total	$r^2 \geq 0.3$	$r^2 \geq 0.8$
1000 Genomes	MAF < 0.01%	9,769,076	0	0
	0.01% ≤ MAF < 0.1%	16,901,730	1,027,742	137,322
	0.1% ≤ MAF < 0.5%	8,034,349	2,235,810	207,890
	0.5% ≤ MAF < 1%	1,965,771	865,708	137,589
	1% ≤ MAF < 2.5%	1,913,816	1,135,996	274,919
	2.5% ≤ MAF < 5%	1,298,617	929,263	348,474
	5% ≤ MAF < 10%	1,416,373	1,118,329	554,661
	MAF ≥ 10%	5,590,616	4,996,730	2,891,707
CAAPA	MAF < 0.01%	1,224,999	0	0
	0.01% ≤ MAF < 0.1%	5,864,540	101,178	8,170
	0.1% ≤ MAF < 0.5%	8,139,040	338,924	13,423
	0.5% ≤ MAF < 1%	2,924,594	291,066	16,924
	1% ≤ MAF < 2.5%	3,078,222	482,804	48,443
	2.5% ≤ MAF < 5%	1,788,013	458,770	80,940
	5% ≤ MAF < 10%	1,570,817	624,388	169,260
	MAF ≥ 10%	4,765,573	3,136,077	1,090,791
GAsPv1.0	MAF < 0.01%	1,174,432	0	0
	0.01% ≤ MAF < 0.1%	5,695,832	335,403	55,659
	0.1% ≤ MAF < 0.5%	5,829,240	556,748	33,230
	0.5% ≤ MAF < 1%	1,318,118	358,338	30,291
	1% ≤ MAF < 2.5%	1,271,454	634,589	105,133
	2.5% ≤ MAF < 5%	902,343	592,566	178,409
	5% ≤ MAF < 10%	1,039,886	754,958	327,100
	MAF ≥ 10%	4,084,784	3,549,044	1,802,450
HRC	MAF < 0.01%	17,803,821	0	0
	0.01% ≤ MAF < 0.1%	8,769,025	4,251,865	1,537,767
	0.1% ≤ MAF < 0.5%	3,671,996	3,360,495	1,210,799
	0.5% ≤ MAF < 1%	1,065,928	1,010,366	448,981
	1% ≤ MAF < 2.5%	1,310,392	1,218,305	701,416
	2.5% ≤ MAF < 5%	1,009,415	925,305	642,592
	5% ≤ MAF < 10%	1,117,962	1,054,412	799,122
	MAF ≥ 10%	4,183,660	4,077,176	3,389,982

Table A.4: LAT cohort imputed up to four reference panels available on the Michigan imputation server with imputed variants categorised by r^2 thresholds.

Reference panel	MAF category	Number of variants		
		Total	$r^2 \geq 0.3$	$r^2 \geq 0.8$
1000 Genomes	MAF < 0.01%	6,525,365	0	0
	0.01% ≤ MAF < 0.1%	17,559,838	2,247,230	396,963
	0.1% ≤ MAF < 0.5%	9,798,235	4,591,361	563,579
	0.5% ≤ MAF < 1%	2,405,524	1,375,550	163,523
	1% ≤ MAF < 2.5%	2,119,052	1,354,999	249,017
	2.5% ≤ MAF < 5%	1,338,850	973,341	315,760
	5% ≤ MAF < 10%	1,456,194	1,154,259	547,309
	MAF ≥ 10%	5,687,291	5,078,062	2,768,769
CAAPA	MAF < 0.01%	724,509	0	0
	0.01% ≤ MAF < 0.1%	5,753,960	221,238	21,362
	0.1% ≤ MAF < 0.5%	8,571,680	724,561	35,741
	0.5% ≤ MAF < 1%	3,051,424	433,561	20,303
	1% ≤ MAF < 2.5%	3,205,910	588,136	45,411
	2.5% ≤ MAF < 5%	1,851,853	500,576	75,209
	5% ≤ MAF < 10%	1,634,058	669,272	178,647
	MAF ≥ 10%	4,858,439	3,204,146	1,071,660
GAsPv1.0	MAF < 0.01%	372,923	0	0
	0.01% ≤ MAF < 0.1%	4,691,228	523,737	60,510
	0.1% ≤ MAF < 0.5%	7,223,273	1,818,744	209,443
	0.5% ≤ MAF < 1%	1,596,882	620,970	59,941
	1% ≤ MAF < 2.5%	1,328,181	700,338	100,684
	2.5% ≤ MAF < 5%	894,647	599,665	162,784
	5% ≤ MAF < 10%	1,064,772	782,226	332,123
	MAF ≥ 10%	4,144,183	3,596,521	1,728,551
HRC	MAF < 0.01%	12,028,116	0	0
	0.01% ≤ MAF < 0.1%	11,300,737	5,092,733	1,525,277
	0.1% ≤ MAF < 0.5%	6,188,433	5,599,351	1,840,545
	0.5% ≤ MAF < 1%	1,427,648	1,345,373	443,513
	1% ≤ MAF < 2.5%	1,394,430	1,318,018	517,579
	2.5% ≤ MAF < 5%	1,034,723	943,620	517,424
	5% ≤ MAF < 10%	1,154,636	1,061,594	712,610
	MAF ≥ 10%	4,258,216	4,129,081	3,065,856

Table A.5: AFR cohort imputed up to four reference panels available on the Michigan imputation server with imputed variants categorised by r^2 thresholds limited to SNPs polymorphic across all imputation outputs.

Reference panel	MAF category	Total number of variants	Mean r^2	r^2≥0.3		r^2≥0.8	
				Number of variants	% of variants	Number of variants	% of variants
1000 Genomes	MAF < 0.01%	6,928	0.018	0	0.0	0	0.0
	0.01% ≤ MAF < 0.1%	207,245	0.148	35,004	16.9	2,300	1.1
	0.1% ≤ MAF < 0.5%	1,305,969	0.339	687,626	52.7	45,997	3.5
	0.5% ≤ MAF < 1%	1,123,876	0.447	780,283	69.4	88,406	7.9
	1% ≤ MAF < 2.5%	2,085,903	0.552	1,713,393	82.1	334,646	16.0
	2.5% ≤ MAF < 5%	1,842,408	0.629	1,643,694	89.2	466,359	25.3
	5% ≤ MAF < 10%	1,737,323	0.661	1,581,135	91.0	550,819	31.7
	MAF ≥ 10%	4,441,611	0.689	4,093,795	92.2	1,725,206	38.8
CAAPA	MAF < 0.01%	4,563	0.020	0	0.0	0	0.0
	0.01% ≤ MAF < 0.1%	229,675	0.072	13,721	6.0	1,078	0.5
	0.1% ≤ MAF < 0.5%	1,376,262	0.132	216,852	15.8	9,801	0.7
	0.5% ≤ MAF < 1%	986,339	0.202	267,011	27.1	16,258	1.6
	1% ≤ MAF < 2.5%	1,930,097	0.261	709,592	36.8	62,635	3.2
	2.5% ≤ MAF < 5%	1,878,410	0.322	866,538	46.1	110,811	5.9
	5% ≤ MAF < 10%	1,830,445	0.378	991,135	54.1	172,044	9.4
	MAF ≥ 10%	4,515,472	0.472	3,020,197	66.9	765,457	17.0

Table A.5(cont.): AFR cohort imputed up to four reference panels available on the Michigan imputation server with imputed variants categorised by r^2 thresholds limited to SNPs polymorphic across all imputation outputs.

Reference panel	MAF category	Total number of variants	Mean r^2	r^2≥0.3		r^2≥0.8	
				Number of variants	% of variants	Number of variants	% of variants
GAsPv1.0	MAF < 0.01%	4,035	0.011	0	0.0	0	0.0
	0.01% ≤ MAF < 0.1%	247,209	0.073	15,390	6.2	765	0.3
	0.1% ≤ MAF < 0.5%	1,731,871	0.215	513,830	29.7	14,676	0.8
	0.5% ≤ MAF < 1%	1,297,967	0.367	760,593	58.6	39,848	3.1
	1% ≤ MAF < 2.5%	2,073,245	0.494	1,607,017	77.5	193,888	9.4
	2.5% ≤ MAF < 5%	1,560,836	0.561	1,316,905	84.4	245,087	15.7
	5% ≤ MAF < 10%	1,503,282	0.579	1,279,720	85.1	284,889	19.0
	MAF ≥ 10%	4,332,818	0.613	3,793,564	87.6	1,092,649	25.2
HRC	MAF < 0.01%	61,216	0.019	0	0.0	0	0.0
	0.01% ≤ MAF < 0.1%	333,523	0.291	131,532	39.4	13,581	4.1
	0.1% ≤ MAF < 0.5%	1,182,983	0.548	1,032,287	87.3	121,256	10.3
	0.5% ≤ MAF < 1%	1,099,350	0.639	1,040,436	94.6	221,548	20.2
	1% ≤ MAF < 2.5%	2,110,357	0.720	2,057,534	97.5	776,296	36.8
	2.5% ≤ MAF < 5%	1,824,370	0.771	1,794,808	98.4	938,675	51.5
	5% ≤ MAF < 10%	1,717,031	0.780	1,677,779	97.7	975,240	56.8
	MAF ≥ 10%	4,422,433	0.780	4,319,380	97.7	2,532,080	57.3

Table A.6: EAS cohort imputed up to four reference panels available on the Michigan imputation server with imputed variants categorised by r^2 thresholds limited to SNPs polymorphic across all imputation outputs.

Reference panel	MAF category	Total number of variants	Mean r^2	r^2≥0.3		r^2≥0.8	
				Number of variants	% of variants	Number of variants	% of variants
1000 Genomes	MAF < 0.01%	425,405	0.021	0	0.0	0	0.0
	0.01% ≤ MAF < 0.1%	1,934,610	0.094	134,158	6.9	14,525	0.8
	0.1% ≤ MAF < 0.5%	2,957,838	0.142	389,501	13.2	30,356	1.0
	0.5% ≤ MAF < 1%	1,117,055	0.176	207,060	18.5	21,143	1.9
	1% ≤ MAF < 2.5%	1,066,229	0.274	366,876	34.4	75,879	7.1
	2.5% ≤ MAF < 5%	721,251	0.429	414,170	57.4	133,390	18.5
	5% ≤ MAF < 10%	895,919	0.530	629,443	70.3	255,016	28.5
	MAF ≥ 10%	3,633,664	0.666	3,171,567	87.3	1,509,412	41.5
CAAPA	MAF < 0.01%	86,907	0.021	0	0.0	0	0.0
	0.01% ≤ MAF < 0.1%	588,229	0.073	34,678	5.9	4,749	0.8
	0.1% ≤ MAF < 0.5%	1,829,009	0.085	114,061	6.2	10,257	0.6
	0.5% ≤ MAF < 1%	1,319,383	0.101	96,586	7.3	6,572	0.5
	1% ≤ MAF < 2.5%	2,068,947	0.127	228,946	11.1	21,915	1.1
	2.5% ≤ MAF < 5%	1,530,579	0.177	295,034	19.3	41,530	2.7
	5% ≤ MAF < 10%	1,391,227	0.254	440,629	31.7	86,784	6.2
	MAF ≥ 10%	3,937,690	0.434	2,355,913	59.8	636,048	16.2

Table A.6(cont.): EAS cohort imputed up to four reference panels available on the Michigan imputation server with imputed variants categorised by r^2 thresholds limited to SNPs polymorphic across all imputation outputs.

Reference panel	MAF category	Total number of variants	Mean r^2	r^2≥0.3		r^2≥0.8	
				Number of variants	% of variants	Number of variants	% of variants
GAsPv1.0	MAF < 0.01%	1,044,176	0.018	0	0.0	0	0.0
	0.01% ≤ MAF < 0.1%	3,041,951	0.066	161,495	5.3	24,657	0.8
	0.1% ≤ MAF < 0.5%	2,315,318	0.116	289,837	12.5	37,845	1.6
	0.5% ≤ MAF < 1%	578,006	0.202	140,062	24.2	25,165	4.4
	1% ≤ MAF < 2.5%	661,338	0.341	294,456	44.5	85,237	12.9
	2.5% ≤ MAF < 5%	607,151	0.468	374,856	61.7	145,508	24.0
	5% ≤ MAF < 10%	867,092	0.527	597,445	68.9	265,936	30.7
	MAF ≥ 10%	3,636,939	0.653	3,104,838	85.4	1,496,023	41.1
HRC	MAF < 0.01%	1,779,787	0.020	0	0.0	0	0.0
	0.01% ≤ MAF < 0.1%	2,555,859	0.182	499,464	19.5	41,889	1.6
	0.1% ≤ MAF < 0.5%	1,687,572	0.329	855,970	50.7	45,010	2.7
	0.5% ≤ MAF < 1%	673,202	0.392	417,260	62.0	34,946	5.2
	1% ≤ MAF < 2.5%	865,483	0.488	644,287	74.4	118,759	13.7
	2.5% ≤ MAF < 5%	697,564	0.582	567,575	81.4	197,203	28.3
	5% ≤ MAF < 10%	897,596	0.631	737,310	82.1	353,278	39.4
	MAF ≥ 10%	3,594,908	0.745	3,409,969	94.9	1,899,033	52.8

Table A.7: EUR cohort imputed up to four reference panels available on the Michigan imputation server with imputed variants categorised by r^2 thresholds limited to SNPs polymorphic across all imputation outputs.

Reference panel	MAF category	Total number of variants	Mean r^2	r^2≥0.3		r^2≥0.8	
				Number of variants	% of variants	Number of variants	% of variants
1000 Genomes	MAF < 0.01%	473,751	0.020	0	0.0	0	0.0
	0.01% ≤ MAF < 0.1%	1,857,393	0.127	229,349	12.3	39,891	2.1
	0.1% ≤ MAF < 0.5%	2,398,219	0.188	540,885	22.6	52,452	2.2
	0.5% ≤ MAF < 1%	1,003,577	0.285	380,162	37.9	62,110	6.2
	1% ≤ MAF < 2.5%	1,281,023	0.440	789,634	61.6	200,014	15.6
	2.5% ≤ MAF < 5%	927,497	0.564	702,193	75.7	272,175	29.3
	5% ≤ MAF < 10%	1,003,461	0.648	829,924	82.7	428,342	42.7
	MAF ≥ 10%	3,806,342	0.743	3,493,557	91.8	2,153,783	56.6
CAAPA	MAF < 0.01%	65,025	0.020	0	0.0	0	0.0
	0.01% ≤ MAF < 0.1%	513,516	0.078	33,729	6.6	3,846	0.7
	0.1% ≤ MAF < 0.5%	1,822,895	0.100	155,480	8.5	7,934	0.4
	0.5% ≤ MAF < 1%	1,386,899	0.128	166,275	12.0	11,317	0.8
	1% ≤ MAF < 2.5%	2,121,434	0.163	376,477	17.7	41,149	1.9
	2.5% ≤ MAF < 5%	1,473,137	0.225	403,516	27.4	74,004	5.0
	5% ≤ MAF < 10%	1,320,104	0.326	553,176	41.9	155,133	11.8
	MAF ≥ 10%	4,048,253	0.506	2,752,772	68.0	988,601	24.4

Table A.7(cont.): EUR cohort imputed up to four reference panels available on the Michigan imputation server with imputed variants categorised by r^2 thresholds limited to SNPs polymorphic across all imputation outputs.

Reference panel	MAF category	Total number of variants	Mean r^2	r^2≥0.3		r^2≥0.8	
				Number of variants	% of variants	Number of variants	% of variants
GAsPv1.0	MAF < 0.01%	648,915	0.020	0	0.0	0	0.0
	0.01% ≤ MAF < 0.1%	2,357,782	0.108	255,921	10.9	50,420	2.1
	0.1% ≤ MAF < 0.5%	2,282,758	0.146	383,911	16.8	27,979	1.2
	0.5% ≤ MAF < 1%	792,456	0.264	288,194	36.4	26,127	3.3
	1% ≤ MAF < 2.5%	1,039,439	0.391	587,852	56.6	99,853	9.6
	2.5% ≤ MAF < 5%	830,318	0.491	565,496	68.1	171,161	20.6
	5% ≤ MAF < 10%	971,894	0.564	719,432	74.0	312,736	32.2
	MAF ≥ 10%	3,827,701	0.678	3,348,634	87.5	1,710,108	44.7
HRC	MAF < 0.01%	3,073,011	0.015	0	0.0	0	0.0
	0.01% ≤ MAF < 0.1%	1,747,537	0.422	897,785	51.4	413,267	23.6
	0.1% ≤ MAF < 0.5%	786,028	0.695	709,401	90.3	347,040	44.2
	0.5% ≤ MAF < 1%	484,347	0.737	455,778	94.1	234,853	48.5
	1% ≤ MAF < 2.5%	1,018,941	0.756	946,902	92.9	572,464	56.2
	2.5% ≤ MAF < 5%	895,385	0.783	826,778	92.3	582,740	65.1
	5% ≤ MAF < 10%	998,316	0.825	947,118	94.9	725,965	72.7
	MAF ≥ 10%	3,747,698	0.869	3,664,147	97.8	3,075,176	82.1

Table A.8: LAT cohort imputed up to four reference panels available on the Michigan imputation server with imputed variants categorised by r^2 thresholds limited to SNPs polymorphic across all imputation outputs.

Reference panel	MAF category	Total number of variants	Mean r^2	r^2≥0.3		r^2≥0.8	
				Number of variants	% of variants	Number of variants	% of variants
1000 Genomes	MAF < 0.01%	122,007	0.023	0	0.0	0	0.0
	0.01% ≤ MAF < 0.1%	979,718	0.235	290,395	29.6	48,675	5.0
	0.1% ≤ MAF < 0.5%	2,889,028	0.391	1,705,274	59.0	227,769	7.9
	0.5% ≤ MAF < 1%	1,415,904	0.401	860,929	60.8	105,749	7.5
	1% ≤ MAF < 2.5%	1,478,947	0.457	999,458	67.6	190,498	12.9
	2.5% ≤ MAF < 5%	958,785	0.558	736,614	76.8	247,523	25.8
	5% ≤ MAF < 10%	1,032,637	0.643	856,585	83.0	423,399	41.0
	MAF ≥ 10%	3,874,237	0.732	3,554,598	91.7	2,069,113	53.4
CAAPA	MAF < 0.01%	21,937	0.020	0	0.0	0	0.0
	0.01% ≤ MAF < 0.1%	349,596	0.097	34,820	10.0	3,689	1.1
	0.1% ≤ MAF < 0.5%	1,760,389	0.153	325,405	18.5	20,586	1.2
	0.5% ≤ MAF < 1%	1,440,911	0.172	291,707	20.2	15,617	1.1
	1% ≤ MAF < 2.5%	2,214,942	0.185	483,854	21.8	40,038	1.8
	2.5% ≤ MAF < 5%	1,514,146	0.233	439,634	29.0	68,315	4.5
	5% ≤ MAF < 10%	1,361,455	0.335	589,747	43.3	162,753	12.0
	MAF ≥ 10%	4,087,887	0.505	2,794,884	68.4	966,465	23.6

Table A.8(cont.): LAT cohort imputed up to four reference panels available on the Michigan imputation server with imputed variants categorised by r^2 thresholds limited to SNPs polymorphic across all imputation outputs.

Reference panel	MAF category	Total number of variants	Mean r^2	r^2≥0.3		r^2≥0.8	
				Number of variants	% of variants	Number of variants	% of variants
GAsPv1.0	MAF < 0.01%	114,447	0.027	0	0.0	0	0.0
	0.01% ≤ MAF < 0.1%	1,575,506	0.197	387,785	24.6	51,497	3.3
	0.1% ≤ MAF < 0.5%	3,238,195	0.325	1,537,079	47.5	198,024	6.1
	0.5% ≤ MAF < 1%	1,037,665	0.354	547,809	52.8	56,480	5.4
	1% ≤ MAF < 2.5%	1,084,544	0.406	654,291	60.3	95,753	8.8
	2.5% ≤ MAF < 5%	821,490	0.495	571,510	69.6	155,790	19.0
	5% ≤ MAF < 10%	995,652	0.569	745,724	74.9	318,076	31.9
	MAF ≥ 10%	3,883,764	0.669	3,394,135	87.4	1,639,572	42.2
HRC	MAF < 0.01%	650,368	0.025	0	0.0	0	0.0
	0.01% ≤ MAF < 0.1%	1,642,781	0.472	1,018,902	62.0	328,992	20.0
	0.1% ≤ MAF < 0.5%	2,605,530	0.710	2,485,940	95.4	1,030,362	39.5
	0.5% ≤ MAF < 1%	908,606	0.701	868,206	95.6	339,749	37.4
	1% ≤ MAF < 2.5%	1,161,751	0.707	1,104,546	95.1	454,281	39.1
	2.5% ≤ MAF < 5%	920,925	0.732	845,792	91.8	469,902	51.0
	5% ≤ MAF < 10%	1,034,896	0.771	957,245	92.5	648,647	62.7
	MAF ≥ 10%	3,826,406	0.826	3,718,887	97.2	2,785,983	72.8

Imputation comparison across four Japanese population specific reference panels

Table A.9: Quality of imputation up to five reference panels into GWAS of 174,460 Japanese individuals at a subset of variants polymorphic across all panels.

Reference panel	MAF category	Total number of variants	Mean r^2	$r^2 \geq 0.3$		$r^2 \geq 0.8$	
				Number of variants	% of variants	Number of variants	% of variants
1KG	MAF < 0.001%	10,670,101	0.039	267,482	2.5	43,449	0.4
	0.001% ≤ MAF < 0.01%	7,328,299	0.065	268,508	3.7	30,243	0.4
	0.01% ≤ MAF < 0.1%	9,898,426	0.125	990,326	10.0	110,381	1.1
	0.1% ≤ MAF < 0.5%	4,208,273	0.310	1,714,013	40.7	268,540	6.4
	0.5% ≤ MAF < 1%	1,133,044	0.498	778,380	68.7	243,097	21.5
	1% ≤ MAF < 2.5%	1,237,165	0.604	955,744	77.3	469,399	37.9
	2.5% ≤ MAF < 5%	801,317	0.769	729,383	91.0	489,896	61.1
	5% ≤ MAF < 10%	901,517	0.870	877,015	97.3	708,753	78.6
	MAF ≥ 10%	3,794,940	0.920	3,747,067	98.7	3,450,764	90.9
1KG+1K	MAF < 0.001%	14,134,201	0.040	360,976	2.6	62,250	0.4
	0.001% ≤ MAF < 0.01%	7,379,838	0.081	412,211	5.6	48,577	0.7
	0.01% ≤ MAF < 0.1%	7,225,971	0.263	2,570,398	35.6	275,849	3.8
	0.1% ≤ MAF < 0.5%	3,435,708	0.525	2,722,449	79.2	542,680	15.8
	0.5% ≤ MAF < 1%	1,052,189	0.635	878,636	83.5	392,160	37.3
	1% ≤ MAF < 2.5%	1,217,217	0.727	1,064,393	87.4	671,691	55.2
	2.5% ≤ MAF < 5%	812,773	0.854	782,930	96.3	609,316	75.0
	5% ≤ MAF < 10%	912,811	0.915	899,836	98.6	799,275	87.6
	MAF ≥ 10%	3,802,374	0.946	3,771,083	99.2	3,614,579	95.1

Table A.9(cont.): Quality of imputation up to five reference panels into GWAS of 174,460 Japanese individuals at a subset of variants polymorphic across all panels.

Reference panel	MAF category	Total number of variants	Mean r^2	$r^2 \geq 0.3$		$r^2 \geq 0.8$	
				Number of variants	% of variants	Number of variants	% of variants
1KG+3K	MAF < 0.001%	16,240,618	0.042	472,126	2.9	87,115	0.5
	0.001% ≤ MAF < 0.01%	6,435,627	0.133	874,068	13.6	87,775	1.4
	0.01% ≤ MAF < 0.1%	6,852,252	0.380	3,898,627	56.9	498,364	7.3
	0.1% ≤ MAF < 0.5%	2,744,044	0.598	2,375,851	86.6	688,077	25.1
	0.5% ≤ MAF < 1%	969,128	0.675	819,645	84.6	452,084	46.6
	1% ≤ MAF < 2.5%	1,192,280	0.750	1,047,581	87.9	727,288	61.0
	2.5% ≤ MAF < 5%	816,770	0.863	784,234	96.0	634,710	77.7
	5% ≤ MAF < 10%	916,696	0.920	902,250	98.4	814,553	88.9
	MAF ≥ 10%	3,805,667	0.950	3,773,451	99.2	3,638,693	95.6
1KG+4K	MAF < 0.001%	15,948,088	0.039	428,408	2.7	80,780	0.5
	0.001% ≤ MAF < 0.01%	6,359,102	0.124	799,578	12.6	74,021	1.2
	0.01% ≤ MAF < 0.1%	6,906,401	0.346	3,592,508	52.0	362,163	5.2
	0.1% ≤ MAF < 0.5%	3,005,697	0.546	2,491,779	82.9	560,843	18.7
	0.5% ≤ MAF < 1%	972,824	0.650	828,094	85.1	389,026	40.0
	1% ≤ MAF < 2.5%	1,226,891	0.700	1,031,279	84.1	655,777	53.5
	2.5% ≤ MAF < 5%	831,714	0.823	773,684	93.0	596,127	71.7
	5% ≤ MAF < 10%	921,905	0.898	899,641	97.6	778,179	84.4
	MAF ≥ 10%	3,800,460	0.936	3,759,115	98.9	3,578,363	94.2

Table A.9(cont.2): Quality of imputation up to five reference panels into GWAS of 174,460 Japanese individuals at a subset of variants polymorphic across all panels.

Reference panel	MAF category	Total number of variants	Mean r^2	$r^2 \geq 0.3$		$r^2 \geq 0.8$	
				Number of variants	% of variants	Number of variants	% of variants
1KG+7K	MAF < 0.001%	16,523,507	0.045	568,530	3.4	108,375	0.7
	0.001% ≤ MAF < 0.01%	6,508,121	0.206	1,701,646	26.1	147,709	2.3
	0.01% ≤ MAF < 0.1%	6,612,577	0.440	4,558,677	68.9	559,016	8.5
	0.1% ≤ MAF < 0.5%	2,618,854	0.623	2,360,451	90.1	739,574	28.2
	0.5% ≤ MAF < 1%	952,475	0.690	818,939	86.0	469,774	49.3
	1% ≤ MAF < 2.5%	1,209,469	0.741	1,044,215	86.3	740,187	61.2
	2.5% ≤ MAF < 5%	822,950	0.856	782,349	95.1	636,720	77.4
	5% ≤ MAF < 10%	921,553	0.915	904,589	98.2	809,526	87.8
	MAF ≥ 10%	3,803,576	0.947	3,770,461	99.1	3,633,852	95.5

GWAS into serum uric acid: 27 previously reported loci (Kanai et al. 2018)

Table A.10: Lead variants at 27 previously-reported loci (Kanai et al. 2018) for serum uric acid in GWAS of 104,174 Japanese individuals from the Biobank Japan Project after imputation up to the 1KG reference panel.

Locus	Chr	Lead variant from imputation up to 1KG panel								
		rs ID	Position	Major allele	Minor allele	MAF	Beta	SE	*p*-value	r^2
NBPF10-NBPF20	1	rs34913946	145,724,937	AT	A	0.1714	-0.0620	0.0081	1.5E-14	0.501
MUC1	1	rs59578826	155,195,071	ATTAT	A	0.1541	0.0487	0.0061	2.0E-15	0.941
GCKR	2	rs1260326	27,730,940	T	C	0.4408	-0.0346	0.0043	1.6E-15	0.997
USP34	2	rs66667063	61,709,169	T	C	0.3940	0.0233	0.0046	3.6E-07	0.923
LRP2	2	rs16856823	170,200,452	A	T	0.1793	0.0385	0.0058	4.0E-11	0.923
SFMBT1-RFT1	3	rs2581824	53,022,408	A	C	0.4989	-0.0232	0.0043	7.8E-08	0.992
SLC2A9	4	rs7679724	9,985,376	T	G	0.4226	-0.1237	0.0044	2.6E-174	0.978
PRDM8-FGF5	4	rs11437847	81,191,519	C	CT	0.3246	-0.0333	0.0050	4.0E-11	0.830
ABCG2	4	rs2231142	89,052,323	G	T	0.2895	0.1124	0.0047	1.8E-124	1.001
SLC17A1	6	rs1177442	25,809,069	G	A	0.1617	-0.0529	0.0058	1.2E-19	1.001
ZNF318	6	rs34453672	43,340,010	AT	A	0.3545	0.0293	0.0045	8.3E-11	0.990
UNCX-MICALL2	7	rs4724828	1,298,448	C	T	0.4286	0.0274	0.0047	7.9E-09	0.839
MLXIPL-VPS37D	7	rs7800944	73,035,857	T	C	0.1183	-0.0423	0.0070	1.2E-09	0.915
HNF4G	8	rs11351186	76,510,328	TA	T	0.4179	0.0392	0.0045	2.1E-18	0.949
TP53INP1-NDUFAF6	8	rs150320174	95,974,852	TG	T	0.3079	-0.0294	0.0048	6.7E-10	0.960
BICC1	10	rs34066134	60,372,359	TA	T	0.4808	0.0336	0.0045	9.9E-14	0.907
FAM35A	10	rs9420446	88,880,689	T	C	0.3692	0.0314	0.0048	5.7E-11	0.863

Table A.10(cont.): Lead variants at 27 previously-reported loci (Kanai et al. 2018) for serum uric acid in GWAS of 104,174 Japanese individuals from the Biobank Japan Project after imputation up to the 1KG reference panel.

Locus	Chr	Lead variant from imputation up to 1KG panel								
		rs ID	Position	Major allele	Minor allele	MAF	Beta	SE	*p*-value	r^2
EMX2-RAB11FIP2	10	rs10886117	119,480,578	G	A	0.3825	0.0274	0.0045	8.3E-10	0.985
SBF2	11	rs2220970	9,857,749	G	A	0.3426	0.0238	0.0045	1.6E-07	1.000
MPPED2-DCDC5	11	rs3600200	30,759,781	C	CT	0.2429	-0.0315	0.0057	2.6E-08	0.785
NRXN2-SLC22A12	11	rs121907892	64,361,219	G	A	0.0221	-1.2556	0.0141	7.8E-1713	0.995
CUX2	12	rs11066001	112,119,171	T	C	0.2267	-0.0974	0.0064	2.8E-52	0.643
IGF1R	15	rs6598541	99,271,135	A	G	0.4986	-0.0343	0.0044	5.5E-15	0.962
CYB5B-MIR1538	16	rs8048032	69,599,803	G	A	0.1565	0.0314	0.0059	1.1E-07	1.000
LINC01229-LOC102724084	16	rs11376510	79,745,672	G	GT	0.2853	-0.0349	0.0050	2.0E-12	0.922
BCAS3	17	rs9895661	59,456,589	C	T	0.4597	0.0390	0.0043	1.2E-19	1.008
LOC101927932	20	rs6026578	57,463,472	G	C	0.2831	0.0299	0.0048	6.5E-10	0.975

Locus-zoom plots for GWAS into serum uric acid

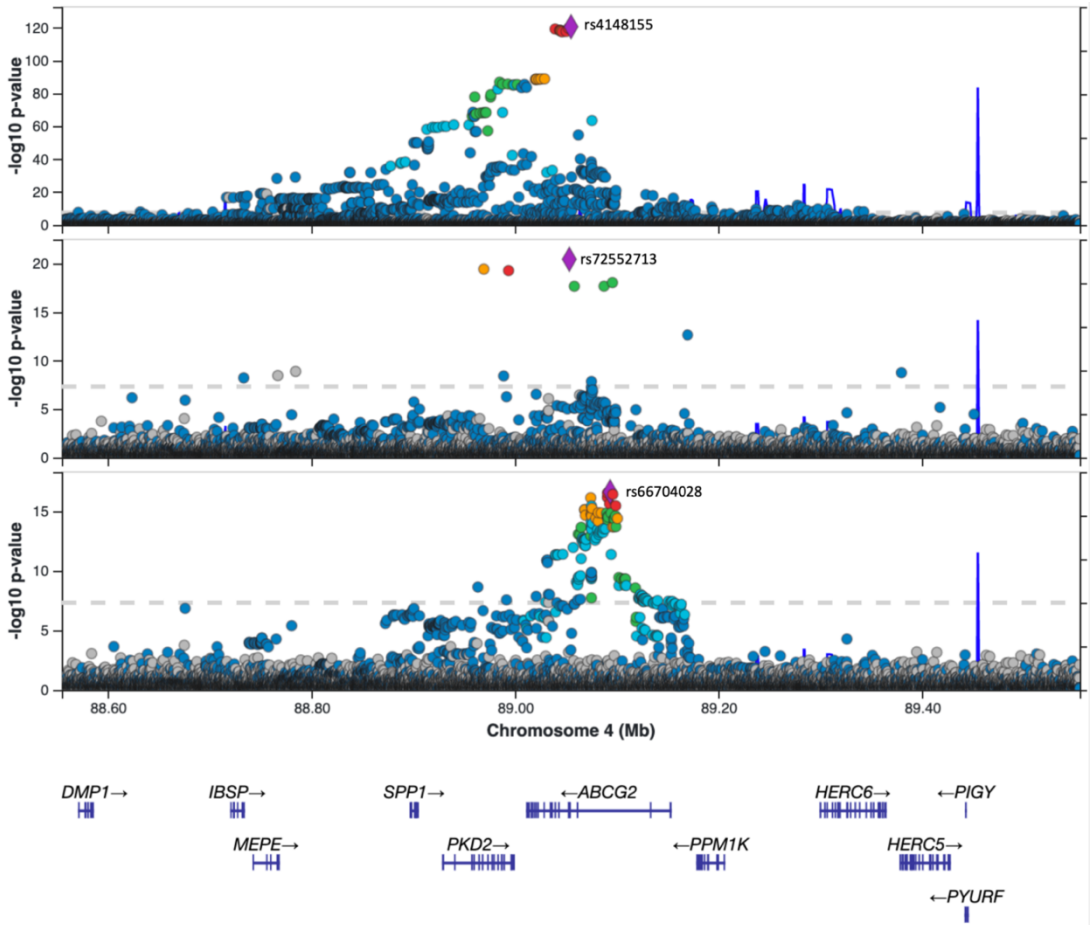

Figure A.1: Signal plots of distinct associations with uric acid at the ABCG2 locus in GWAS of 104,174 Japanese individuals from the Biobank Japan Project. The GWAS was imputed up to a merged reference panel of 1KG and Japanese population-specific WGS from 7,472 individuals. Each panel represents a distinct signal of association after conditioning on all other signals at the locus. In each panel, the index variant is highlighted by the purple diamond. All other variants are coloured according to their LD (1KG East Asian) with the index variant: red $r^2 \geq 0.8$; gold $0.6 \leq r^2 < 0.8$; green $0.4 \leq r^2 < 0.6$; cyan $0.2 \leq r^2 < 0.4$; blue $r^2 < 0.2$; grey r^2 unknown. Gene annotations are taken from the University of California Santa Cruz genome browser.

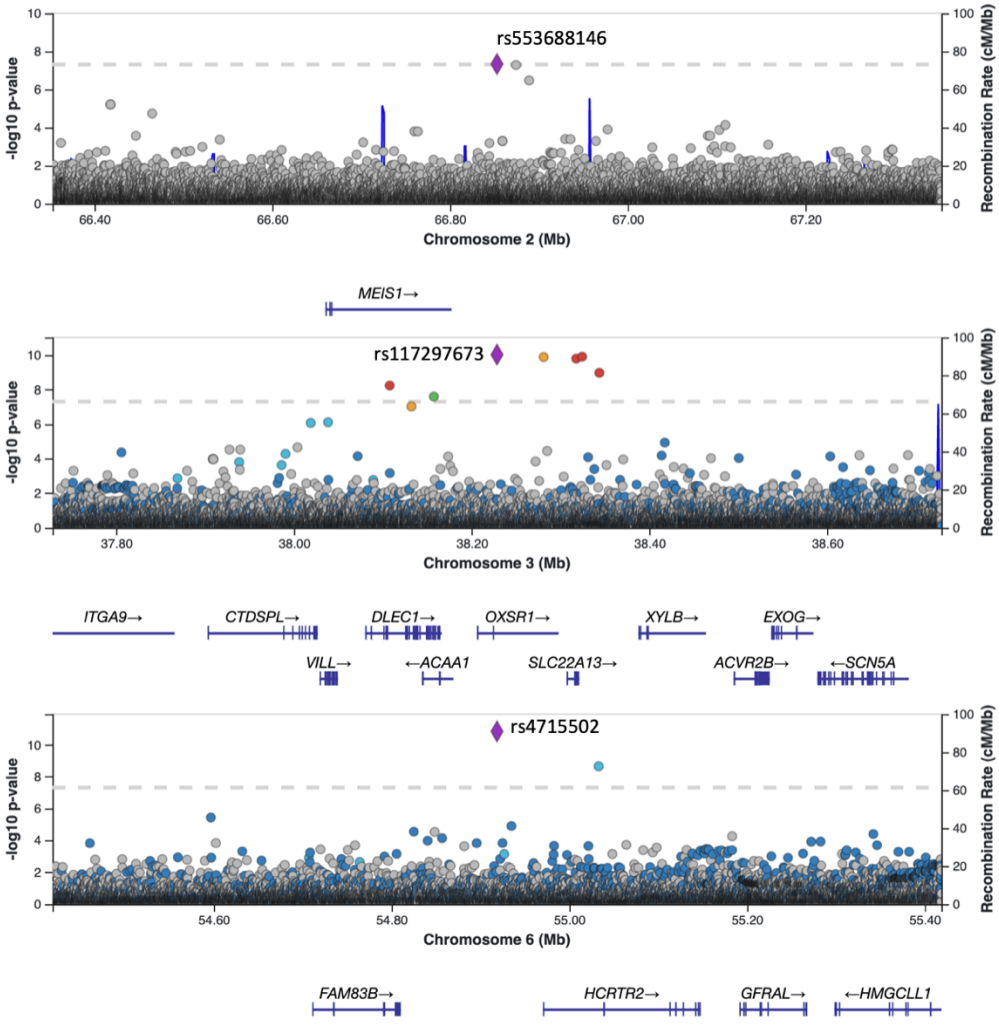

Figure A.2: Signal plots of associations with uric acid at novel loci in GWAS of 104,174 Japanese individuals from the Biobank Japan Project. The GWAS was imputed up to a merged reference panel of 1KG and Japanese population-specific WGS from 7,472 individuals. In each panel, the lead variant is highlighted by the purple diamond. All other variants are coloured according to their LD (1KG East Asian) with the index variant: red $r^2 \geq 0.8$; gold $0.6 \leq r^2 < 0.8$; green $0.4 \leq r^2 < 0.6$; cyan $0.2 \leq r^2 < 0.4$; blue $r^2 < 0.2$; grey r^2 unknown. Gene annotations are taken from the University of California Santa Cruz genome browser.

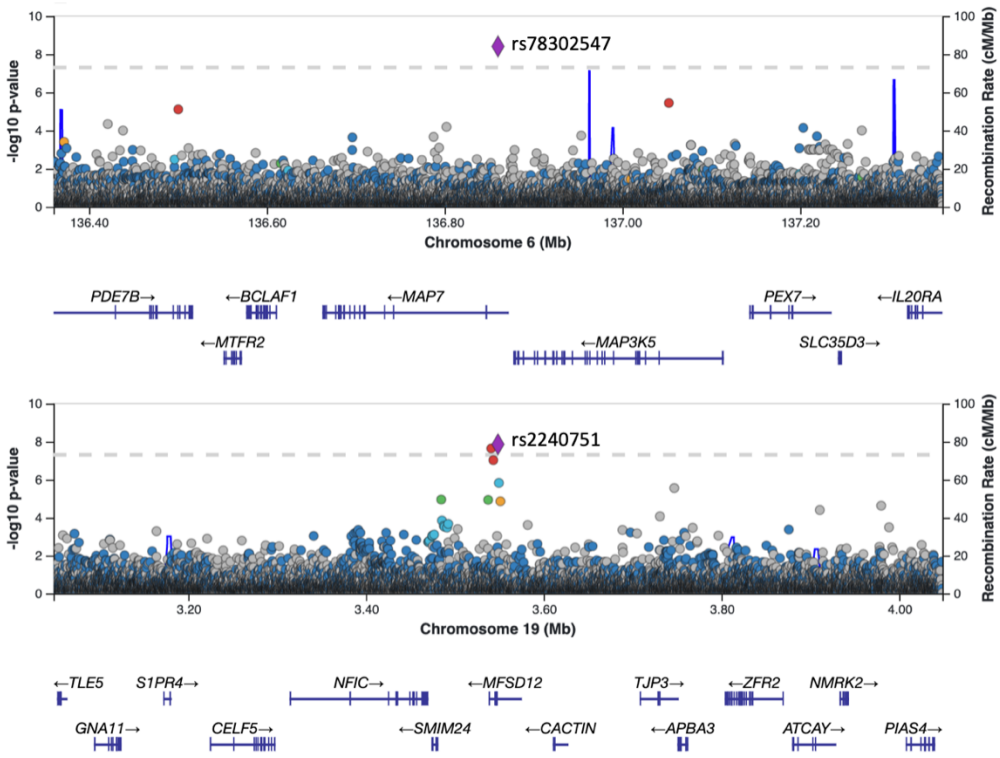

Figure A.3: Signal plots of associations with uric acid at novel loci in GWAS of 104,174 Japanese individuals from the Biobank Japan Project. The GWAS was imputed up to a merged reference panel of 1KG and Japanese population-specific WGS from 7,472 individuals. In each panel, the lead variant is highlighted by the purple diamond. All other variants are coloured according to their LD (1KG East Asian) with the index variant: red $r^2 \geq 0.8$; gold $0.6 \leq r^2 < 0.8$; green $0.4 \leq r^2 < 0.6$; cyan $0.2 \leq r^2 < 0.4$; blue $r^2 < 0.2$; grey r^2 unknown. Gene annotations are taken from the University of California Santa Cruz genome browser.

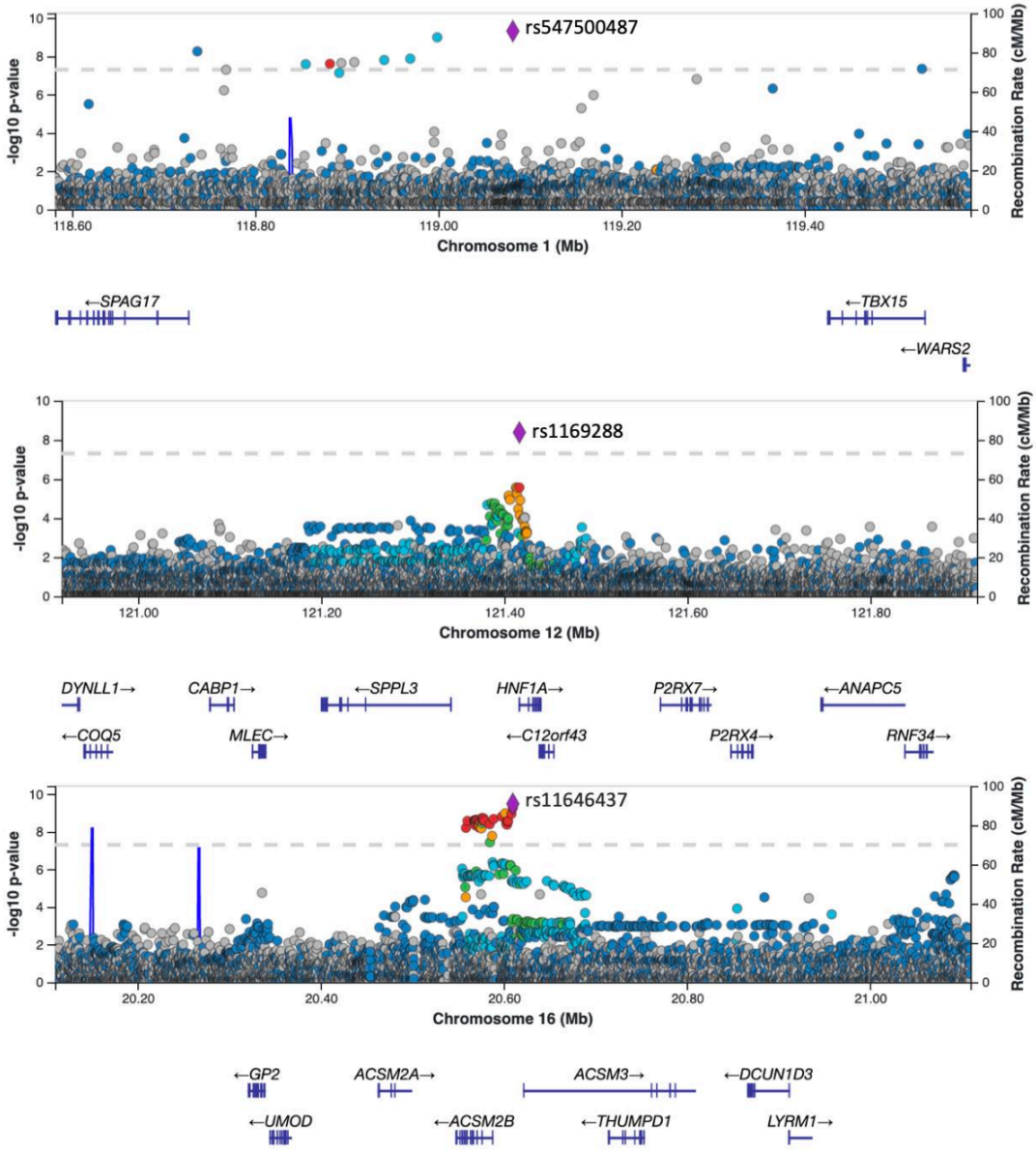

Figure A.4: Signal plots of associations with uric acid at novel loci in GWAS of 104,174 Japanese individuals from the Biobank Japan Project. The GWAS was imputed up to a merged reference panel of 1KG and Japanese population-specific WGS from 7,472 individuals. In each panel, the lead variant is highlighted by the purple diamond. All other variants are coloured according to their LD (1KG East Asian) with the index variant: red $r^2 \geq 0.8$; gold $0.6 \leq r^2 < 0.8$; green $0.4 \leq r^2 < 0.6$; cyan $0.2 \leq r^2 < 0.4$; blue $r^2 < 0.2$; grey r^2 unknown. Gene annotations are taken from the University of California Santa Cruz genome browser.

Stepwise conditional analysis of gene transcripts

Table A.11: Stepwise conditional analysis of gene transcripts at the NRXN2-SLC22A12 locus.

Step 1	Gene transcript	No. of variants	Average MAF	beta	se	p
	uc009ypo.1	9	0.00901	0.205	0.073	0.00495
	uc001oal.1	4	0.00599	-4.663	0.061	6.66E-15
	uc001obx.2	11	0.01528	0.017	0.066	0.79326
	uc001ocn.2	10	0.00232	0.800	0.148	6.65E-08
	uc001ocq.1	4	0.01009	0.010	0.064	0.88050
	uc001oct.2	6	0.01004	-0.182	0.061	0.00269
	uc009yqb.1	4	0.01657	0.111	0.050	0.02605
Step 2	c009ypo.1	9	0.00901	0.204	0.073	0.00500
	uc001oal.1	4	0.00599	-4.663	0.061	7.77E-15
	uc001obx.2	11	0.01528	0.016	0.066	0.80477
	uc001ocn.2	10	0.00232	0.804	0.145	3.19E-08
	uc001oct.2	6	0.01004	-0.182	0.061	0.00269
	uc009yqb.1	4	0.01657	0.116	0.035	0.00102
Step 3	uc009ypo.1	9	0.00901	0.204	0.073	0.00505
	uc001oal.1	4	0.00599	-4.660	0.060	6.88E-15
	uc001ocn.2	10	0.00232	0.802	0.145	3.29E-08
	uc001oct.2	6	0.01004	-0.181	0.060	0.00278
	uc009yqb.1	4	0.01657	0.117	0.035	0.00096
Step 4	uc001oal.1	4	0.00599	-4.667	0.060	6.22E-15
	uc001ocn.2	10	0.00232	0.784	0.145	6.52E-08
	uc001oct.2	6	0.01004	-0.181	0.060	0.00275
	uc009yqb.1	4	0.01657	0.123	0.035	0.00047
Step 5	uc001oal.1	4	0.00599	-4.734	0.056	6.22E-15
	uc001ocn.2	10	0.00232	0.793	0.145	4.47E-08
	uc009yqb.1	4	0.01657	0.129	0.035	0.00024
Step 6	uc001oal.1	4	0.00599	-4.740	0.056	4.22E-15
	uc001ocn.2	10	0.00232	0.907	0.142	1.55E-10