Imputation of Rare Variants in Next-Generation Association Studies

Jennifer L. Asimit Eleftheria Zeggini

Wellcome Trust Sanger Institute, Hinxton, UK

Key Words

Association test · Low-frequency variant · Reference panel · Sequencing

Abstract

The role of rare variants has become a focus in the search for association with complex traits. Imputation is a powerful and cost-efficient tool to access variants that have not been directly typed, but there are several challenges when imputing rare variants, most notably reference panel selection. Extensions to rare variant association tests to incorporate genotype uncertainty from imputation are discussed, as well as the use of imputed low-frequency and rare variants in the study of population isolates.

Copyright © 2013 S. Karger AG, Basel

Introduction

Many disease associations with common variants have been detected via hundreds of genome-wide association studies (GWAS), yet these variants are only able to explain a small fraction of the heritable component of disease. Moreover, in the search for genetic susceptibility to disease, the analysis focus has shifted from common variants to the identification and association analysis of low-frequency and rare variants. Typically, common variants are defined to have minor allele frequency (MAF) >0.05,

while low-frequency variants have MAF 0.01–0.05, and rare variants are those with MAF <0.01.

In focusing on variants with low allele frequencies, the challenge of being able to accurately genotype them arises. The majority of genotype arrays were designed with common variants in mind, but with the changing focus to less common variants, high-density genotyping arrays (e.g. Illumina Omni 2.5M) have been designed enabling lower MAF variants to be genotyped. The most accurate way of accessing rare variants is through next-generation sequencing, which is not easily available for large samples due to its high cost. Imputation is a cost-effective alternative that is able to estimate genotypes that are not directly typed, and relies on the availability of a dense reference panel. Genetic correlations are then extrapolated from the reference panel to the study sample to approximate the unobserved genotypes, thereby increasing the number of variants available for association testing, and, consequently, providing a higher chance of detecting true associations [1–3]. Among the standard imputation tools, IMPUTE [4] and MACH [5] generally have better performance when estimating missing genotypes at rare polymorphisms, since they take into account all available markers and all available haplotypes, but this also makes them substantially more computationally intensive [6]. Computational time for imputation may be decreased by first pre-phasing the data, which has been demonstrated to be effective at low-frequency SNPs; for each individual within the GWAS sample, the haplotypes are first statisti-

Fig. 1. Flow chart of the steps taken in the use of imputation in a rare variant association analysis.

cally estimated and then missing genotypes are imputed into these estimated haplotypes [7].

There are several considerations when imputing low-frequency and rare variants, which are due to challenges that are not encountered when dealing with higher MAF SNPs. Low-frequency and rare variants are more difficult to tag than common SNPs, which often results in the poor performance of imputation, as illustrated by Pei et al. [8], Huang et al. [9] and Sung et al. [10]. However, there are ways of selecting a reference panel so as to maximize the number of low-frequency/rare variants imputed with high quality. Reference panel selection is discussed in the next section, followed by approaches to accessing non-common variants by making use of higher density data for a portion of the study sample in conjunction with imputation. Next, association analysis methods for imputed low-frequency and rare variants are discussed. The analysis of rare variants has difficulty in itself due to their low MAF, but the additional uncertainty from imputed genotypes adds to the challenge. There are several methods developed for rare variant association analysis that have been extended to handle imputed rare genotypes as well. The approaches to imputing rare variants in population isolates are subsequently discussed. Low haplotype diversity within population isolates makes it possible to use only a relatively small sequenced sample to impute variants in the remaining population. Studies in which rare variant imputation has been incorporated are then summarized, followed by a discussion on future directions in the imputation of rare variants. A flow chart of the steps taken in the use of imputation in a rare variant association analysis is provided in figure 1.

Reference Panel Selection

The objective of the International HapMap Project was to catalog common sequence variants and determine their frequencies as well as correlation patterns across the

genome [11]. HapMap is composed of samples taken from populations with African, Asian, and European ancestry as well as recently admixed populations from the Americas, and includes variants that are common in at least one of the study populations. There are currently 3 phases of HapMap: Phase 3 data are formed from genotyping 1.6 million common SNPs in 1,184 reference individuals from 11 global populations, and sequencing ten 100-kb regions in a subset of 692 individuals. In HapMap 3, it was confirmed that, even among closely related populations, lower-frequency variation is less shared than common variation [12].

In keeping with its design, the HapMap reference panel has good coverage down to MAF 0.05 and sparser coverage at lower MAFs [13]. This limitation may be surpassed by making use of the 1000 Genomes Project [14]. One of the key aims of the 1000 Genomes Project was to identify more than 95% of variants with MAF as low as 0.01 across the genome and from 0.001 to 0.005 in gene regions, by sequencing 2,500 individuals, thereby providing the scientific community with a catalogue of variation that is more extensive than the HapMap. Consequently, in comparison to HapMap-based imputations, many more variants, particularly rare and low-frequency variants, can be imputed by using the 1000 Genomes Project data reference panel.

A marker imputability database has been assembled (<http://www.unc.edu/~yunmli/imputability.html>) to provide imputation accuracy information for 1000 Genomes Project variants across four major continental groups and multiple genotyping platforms (pers. commun., Yun Li). A leave-one-out imputation procedure was applied to the 1000 Genomes Project samples, as a means of calculating the imputation accuracy of each variant, and markers that were absent from a given platform were treated as untyped.

With a focus on imputation performance in Europeans, the HapMap 2 and HapMap 3 reference panels were

compared by Jostins et al. [15], considering a variety of population combinations within each HapMap phase. The use of all populations in HapMap 3 as a reference panel was found to yield highly accurate imputation of low-frequency/rare variants close to the accuracy attained in common variant imputation. Further, if the ancestry-matched sample is not sufficiently large, then the imputation accuracy of low-frequency variants can increase with the addition of haplotypes from other populations. On the other hand, when there is a sufficiently large number of reference haplotypes from the relevant population, increasing the diversity of the reference panel may not improve imputation accuracy. This is also a consequence of a theoretical investigation of how reference panel properties affect imputation accuracy via a coalescent model [16]. The study showed that there is greater imputation accuracy from a modestly sized reference panel composed of individuals from the same population as the target haplotype than from a larger panel consisting of a different population.

Consistent with Jostins et al. [15], Sung et al. [10] showed that, for low-frequency variants, although the CEU panel from the 1000 Genomes Project (60 individuals, June 2010 release) and the EUR panel from the 1000 Genomes Project (283 individuals, August 2010 release) both achieved higher imputation performance than HapMap 2-based imputation, the larger and more diverse EUR 1000 Genomes Project panel was most successful. They found that 1000 Genomes Project-EUR-based imputations yielded 4 times as many successfully imputed low-frequency SNPs than those based on the HapMap-CEU panel, and the increase was 8-fold for rare variants.

Primarily, evaluation of imputation performance has focused on European populations. However, there is a growing interest in non-European populations, including African and those that are recently admixed, such as Hispanic. African populations are challenging to impute due to their lower levels of LD, and they were found to have the lowest imputation accuracy among 29 populations in the Human Genome Diversity Project [9]. African-American populations have the additional challenge of admixture, since it complicates the choice of an ideal reference panel.

Considering that the genomes of Hispanics, African-Americans, and other recently admixed populations may be partitioned into long genomic regions arising from one of a few ancestral populations, a method has been developed to construct a tailor-made reference population for each region of the genome [17]. Upon application of

this method to select a weighted haplotype reference panel based on a set of populations, any of the current imputation methods may then be employed. In assessing the use of weighted haplotype panels, the largest gain in accuracy was attained in regions of low LD; when used with IMPUTE v2 on a Hispanic dataset, the total number of SNPs with imputed correlation >0.6 increased from 76% (IMPUTE v2) to 83% (IMPUTE v2 with weighted haplotype panels). Moreover, with this same dataset, the error rate was found to be consistently lower across the range of MAFs (from <0.05 to common) when the tailor-made haplotype panels were used.

It was shown by Howie et al. [18] that cosmopolitan reference panels increase imputation accuracy at low MAF variants for a variety of populations, including African. In their comparison of imputing a Gambian population based on a Gambian-specific panel versus HapMap 3, which does not include Gambians, they found that the much smaller Gambian panel (~ 200 chromosomes) yielded higher accuracy than HapMap, which included around 800 chromosomes of African ancestry. However, the difference in imputation accuracy was not large, and smallest at low-frequency variants. Furthermore, for MAF 0.01–0.03, HapMap-based imputation had the highest accuracy, indicating that low-frequency alleles that are poorly represented in the Gambian panel could be captured by an ancestrally inclusive, non-specific reference panel. Upon using the HapMap 3 reference panel in addition to the Gambian panel, the imputation accuracy increased over that of the single Gambian reference panel. These results suggest that incorporating haplotypes from all available reference panels achieves the highest accuracy in African populations, among others.

In Sung et al. [19], three imputation strategies are evaluated for African-American populations, and a comparison is made between the HapMap 2 and 1000 Genomes Project as reference panels, in an attempt to overcome the issues in imputing such populations. The union strategy, which uses a panel consisting of SNPs polymorphic in either CEU or YRI, was found to have the best performance. The other strategies considered were intersection, which uses a reference panel composed of SNPs polymorphic in both CEU and YRI panels, and merge, which involves merging the results of two separate imputations, one YRI-based and one CEU-based, where each imputation uses SNPs polymorphic in the single panel. In line with the European imputation results of Sung et al. [10], they found that the 1000 Genomes Project-based imputations provided approximately 8 times as many well-imputed

rare and low-frequency variants as those that were HapMap-based. Furthermore, the accuracy attained in the 1000 Genomes Project-based imputations was similar to that of the HapMap-based imputations. It is anticipated that imputation performance for African-Americans will improve with versions of the 1000 Genomes Project data that include larger samples.

An alternative, but more costly option, to employing a publicly available reference panel is to sequence a subset of the study sample. Li et al. [3] sequenced (with a median depth of $27\times$) the exons and flanking regions of 202 genes that are current or prospective drug targets in over 14,000 samples including population-based and case collections for 12 diseases. Among the sequenced samples, 8,865 samples also had GWAS data available on either the Affymetrix 500K, Affymetrix 6.0, or Illumina 550K platforms. For each platform, the authors first partitioned these samples into a study set of size 270 and a reference set that ranged in size from 100 to 3,713, then used the reference data to impute the genotypes in the GWAS data. In order to assess imputation performance using the various reference sets and platforms, they compared the imputed genotypes with the genotypes derived from the high-depth sequence data, which enables evaluation for variants with $MAF < 0.01$, due to the high quality of the derived genotype calls. It was illustrated by Li et al. [3] that for very rare variants ($MAF < 0.005$) the imputation performance depends highly on the reference panel size, and such variants are not likely to be imputable if the reference panel does not contain at least 1,200 individuals. The authors suggest the routine use of imputation for common and moderately rare variants, based on their finding that for reference panel sizes of at least 1,200 subjects, approximately 40% of variants with $0.005 < MAF < 0.05$ and almost all common SNPs will be imputable. The results also indicate that Illumina 550K tended to result in higher imputation quality than Affymetrix 500K, particularly for low-frequency variants, where the proportion of variants that could be imputed well was 78 and 57% for Illumina 550K and Affymetrix 500K, respectively, when a reference panel of 3,713 individuals was used.

In simulation studies, Mägi et al. [20] compared the power of a regression-based collapsing method within four approaches to accessing rare variants: re-sequencing, genotyping the variants present in a reference panel, imputation, and genome-wide genotype data. They demonstrated that the power attained in analyzing genome-wide genotype data imputed up to a high-density reference panel is comparable to that of the much more costly re-sequencing gold standard.

A further approach is to impute based on both the 1000 Genomes Project data and a sequenced subset of the data. In Fridley et al. [21], an evaluation is made on the imputation performance of sequencing a subset of a sample to be used as a reference panel, as well as employing the 1000 Genomes Project data as a reference panel. They concluded that it is best to select the largest and most diverse reference panel for sequencing, and to maximize the number of markers that are genotyped for all subjects, based on their finding that imputation accuracy is influenced by both the number of markers genotyped and the reference sample size. The 1000 Genomes Project sequencing data may be used in conjunction with a reference panel based on a subset of the sample that is sequenced, in order to maximize the number of variants, particularly rare and low-frequency variants, that could be imputed. Fridley et al. [21] noted that a 1000 Genomes Project-based imputation without a sequenced subset from the sample may not give sufficient representation of the genetic diversity in affected individuals.

In the case that only low-frequency and common variants are of interest, low-depth sequencing may be used rather than high-depth. In a simulation study, Paşaniuc et al. [22] demonstrate that, for common and low-frequency variants, power similar to that attained by high-density SNP arrays could be achieved by extremely low-depth sequencing ($0.1\text{--}0.5\times$) with imputation based on the 1000 Genomes Project data. However, with this design the coverage is not sufficient to detect and test rare variants for association with a phenotype. Since high imputation accuracy at extremely low depth heavily depends on large reference panels, an increased accuracy is expected as the number of available haplotypes from the 1000 Genomes Project grows.

The ongoing UK10K Project (<http://www.uk10k.org/>) has the aim of identifying low-frequency and rare variants and their associations with disease and quantitative traits by sequencing 10,000 individuals primarily from the UK. Within the sample, the exomes of 6,000 cases with extreme obesity, rare disorders and neurodevelopmental disease are high-depth sequenced, while 4,000 individuals from population-based cohorts with deep phenotype data are genome-wide sequenced at low depth (average $6\times$). The low-depth whole-genome sequenced data will aid in imputation of previously genotyped cohorts and increase the power in such case-control studies. Moreover, the 6,000 high-depth sequenced exomes of extreme phenotypes will result in high power to detect rare variant associations with the diseases in-

cluded in the study and will provide the wider scientific community with control exomes (<http://www.uk10k.org/goals.html>).

Accessing Rare Variants

There are two general issues that need to be considered when testing for associations with rare variants. First, it needs to be decided how to access the rare variants. This may be done via sequencing, genotyping variants present in a reference panel, or imputation; the methods are listed in order of decreasing cost. Second, a method for detection of association needs to be chosen.

In accessing rare variants, an option is a two-platform design, where all individuals are genotyped, and a small subset is also genotyped on a more costly, denser platform or sequenced, as considered by Fridley et al. [21] and discussed in the previous section. This design was further explored by Sampson et al. [23], who imputed missing genotypes of individuals genotyped only on a low-density array (Illumina OmniExpress; >700,000 SNPs) by using a reference set that included both high-density genotype data (Illumina Omni2.5-Quad; approximately 2.5 million SNPs) in conjunction with public datasets. They demonstrated that even with high-density genotyping a subset as small as 100 individuals, it may be possible to observe more than 80% of the detectable associations from directly genotyping all individuals. This is 5–10% more than when imputation is based only on a public reference set and there is no high-density genotyping. However, the authors also illustrate that if rare variants have relative risks that are significantly larger than what has been observed for common variants, then the relative proportion would likely be lower. With a higher relative risk for variants with MAF <0.05, the total proportion of associations detected by the two-platform design increases, but the proportion relative to the number of detected associations when all individuals are genotyped decreases. This relative decrease is a consequence of low imputation accuracy for low MAF variants, so that genotyping only a subset of the individuals and then imputing performs poorer than directly genotyping everyone.

Along the lines of the two-platform design, there is the option of the more costly two-stage design in which a subset of the sample is sequenced to discover variants, and identified variants are then genotyped in the remaining sample, rather than genotyped on a higher density array. The portion chosen to be sequenced may be a balanced number of cases and controls, or only cases. Use of the

former strategy may detect protective variants for the disease phenotype with higher probability [24], while the latter approach may have higher power for detecting causal variant associations. In a comparison of the two study designs with varying sequenced subset size, Liu and Leal [25] showed that when relatively few samples are sequenced (e.g. 400 samples from a study of 3,000 with equal cases and controls), the cases-only strategy is more powerful, but this advantage grows smaller as the sequenced subset increases in size.

When portions of both cases and controls are selected for sequencing, the sequence and genotype data may be combined by meta-analysis methods. However, if only some cases are sequenced, integrating the genotyped and sequenced data directly, by comparing aggregated variant frequencies between cases and controls, leads to inflated type I errors in the test for rare variant associations [26]. Liu and Leal [25] discussed and compared various methods that have been developed to correct for this bias, and introduced a likelihood-based method, SEQCHIP, which makes an adjustment such that the variant genotypes found from sequencing only cases are corrected to follow the same distribution as the genotyped samples. These SEQCHIP data may then be used with any rare variant association test that incorporates genotype uncertainty. The authors illustrated that SEQCHIP-integrated data yielded well-controlled type I errors for rare variant association tests, as well as consistently higher power than other data integration methods.

The Analysis of Imputed Rare Variants

It is often difficult to detect associations with rare/low-frequency variants by applying single variant tests, due to their low frequency. However, imputation can lead to a boost in power at these variants [2]. The software SNPTEST v2 (https://mathgen.stats.ox.ac.uk/genetics_software/snpTest/old/snpTest_v2.3.0.html) implements single variant tests that account for uncertainty in imputed genotypes [4].

A computationally efficient method to test for single SNP associations in GWAS, when complete or imputed genotype data are available and also when including related individuals, is introduced by Zhou and Stephens [27] as Genome-Wide Efficient Mixed-Model Association (GEMMA). GEMMA efficiently performs exact calculations, rather than approximations, and is shown to provide numerically identical results to the widely used exact method Efficient Mixed-Model Association

(EMMA), which is not feasible for use on genome-wide data.

A gain in power over single variant tests may be achieved by combining information across the rare variants in a pre-specified region, such as a gene, and testing for an association with an accumulation of rare/low-frequency variants. Many locus-based tests have been developed to test for accumulations of rare variants, some of which have been extended to account for uncertainty in imputed variants. Reviews of approaches to rare variant association analysis are given by Asimit and Zeggini [28] and Bansal et al. [29], while comparisons of several representative methods in various simulation scenarios are provided by Basu and Pan [30] and Ladouceur et al. [31]. Here, we discuss some of the rare variant methods that have already been extended for the analysis of imputed data, as well as ways in which such an extension may be made.

An extension of a rare variant association test to incorporate imputation probabilities may be achieved by replacing each missing genotype by its expected value, which is referred to as the dosage. This is the approach used by Zawistowski et al. [32] to extend their Cumulative Minor-Allele Test (CMAT) so that it may handle probabilistic genotypes. CMAT is a pooling statistic for case-control data that separately, for cases and for controls, sums up the rare allele counts at predicted functionally relevant sites. Likewise, the software GRANVIL (<http://www.well.ox.ac.uk/GRANVIL>) implements the collapsing method test of Morris and Zeggini [33], and extends it in this manner to account for genotype uncertainty due to imputation [20]. This collapsing method models phenotype as a function of the proportion of rare variants that carry at least one minor allele within a generalized linear model framework.

A weighted imputation dosage test that may be used on GWAS data to detect associations with rare variants was introduced by Li et al. [34]. Unlike the burden (collapsing/aggregate) tests, it does not require the assumption of a common direction of effect among the causal rare variants. For each individual, a genetic score is computed as a weighted sum over the set of variants that contribute to disease risk, where this set is defined based on an examination of frequency difference between cases and controls.

AMELIA (Allele-Matching Extended and Integrated Locus-Wide Association test) is a further rare variant association test that does not require knowledge of the risk variants [35]. It is an allele-matching-based association test, where similarity scores between pairs of individuals

within cases and within controls are tested for a significant difference, while accounting for sequence quality scores, and is an extension of KBAT (Kernel-Based Association Test) [36]. The Allele Match (AM) similarity score is the count of common alleles between the genotypes of two individuals. We have extended this test to incorporate imputation probabilities by employing an expected allele count score for pairs of individuals, so that the imputed AM similarity score at a specific variant is given by

$$4 \sum_{g=0}^2 p_{ig} p_{jg} + 2[p_{i1}(1-p_{j1}) + p_{j1}(1-p_{i1})]$$

where p_{ig} is the probability that genotype g is observed at the specific variant for individual i . This extension has been implemented and is freely available from <http://www.sanger.ac.uk/resources/software/amelia/>.

A further test for association with a group of rare (and common) variants is SKAT (Sequence Kernel Association Test), which is a computationally efficient regression method that allows different directions and magnitudes of effect [37]. SKAT gives higher weights to lower MAF variants and its power is not adversely affected by the inclusion of non-causal variants. Upon applying SKAT to imputed data, by using imputed genotypes instead of true genotypes, Wu et al. [37] found that for 10% missing genotype data there was only a small reduction in power. The method SKAT-O has been developed as a unified approach that optimally combines SKAT with the burden test, so that it is powerful both when variants have different directions of effect and many are non-causal (as in SKAT) and when there is a common direction of effect for the majority of the variants (as in burden tests) [38].

Population Isolates and Rare Variant Imputation

Population isolates are of interest due to their lower phenotypic, environmental, and genetic heterogeneity, which allows easier detection of variants that are rare in frequency in the outbred populations from which the isolates diverged. Increases in genetic homogeneity and rare variant allele frequency tend to occur in populations resulting from the expansion of a relatively small number of founders who formed a population bottleneck, or from random genetic drift. Thus, variants that are rare elsewhere may have a higher frequency in isolated populations.

A long-range phasing approach was introduced by Kong et al. [39] to infer haplotype phase irrespective of

the availability of parental genotypes, provided that a sufficiently large sample from the isolated population has been genotyped. This method elicits latent information from such data by recognizing the usefulness of moderately to distantly related individuals, rather than focusing only on closely related pairs. The close genetic relationships in isolated founder populations are further taken advantage of in the genotype phasing algorithm developed by Palin et al. [40], referred to as systematic long-range phasing (SLRP). SLRP uses a model that explicitly considers identity-by-descent (IBD) segments and models multiple individuals sharing a haplotype.

High levels of relatedness in isolated founder populations are also capitalized in the imputation algorithm of Uricchio et al. [41] by using the IBD information in a small number of sequenced founder individuals for imputation of a much larger number of founders with only genome-wide SNP genotypes. The authors demonstrate that across the full range of allele frequencies, with high accuracy and low missing information, their method can phase and impute genotypes identified from sequencing in members of this founder population (South Dakota Hutterites).

With a focus on rare variant inference in the whole-genome analysis of a homogeneous isolate population, Gusev et al. [42] showed that low-pass sequencing together with multi-sample calling is able to yield a 2- to 3-fold gain in accurate variant detection over independent calling. The authors studied an isolated cohort from the Pacific island of Kosrae, Micronesia, and estimated that at least 60% of the cohort's 3,000 genomes could be inferred from a full panel of 40 sequenced individuals. However, they cautioned that this is a unique cohort, and in a more diverse isolated population it is likely that a lower proportion of the population genome could be inferred from sequencing a sample of size 40; in an isolated cohort of Ashkenazi Jewish origin only 23% of the genome could be inferred using 40 sequenced individuals.

Studies Employing Rare Variant Imputation

In this section, we summarize some representative examples of studies published to date that have searched for associations with low-frequency/rare variants by using imputation.

In a proof-of-principle for next-generation association study design [43], capitalizing on the unique characteristics of population isolates and combining whole-genome sequencing with imputation, Holm et al. [44] identified a

missense mutation associated with sick sinus syndrome in *MYH6* through GWAS and imputation of 1000 Genomes Project variants, followed by whole-genome sequencing for refinement and genotyping for validation purposes in an Icelandic cohort. The variant is not present in the HapMap or 1000 Genomes Project data and was not identified in any of the European-descent controls investigated.

In the study by Day-Williams et al. [45], a 1000 Genomes Project-based imputation into a GWAS dataset of osteoarthritis cases and controls from the UK was used to detect a previously unidentified risk locus on chromosome 13. Through large-scale replication, the authors established a robust association with SNPs in *MCF2L* (risk allele frequency of approximately 0.07).

A susceptibility locus for venous thrombosis on chromosome 11p11.2 was detected, and it was found that the signal was attributed to a rare functional variant, which is approximately 1 Mb away from the locus and has a MAF of approximately 0.03 [46]. This locus discovery arose by using the 1000 Genomes Project (August 2010 release) reference dataset for imputation of two GWAS. Previous imputation based on HapMap Phase 2 had missed the signal.

In an imputation analysis of *ADAMTSL3* as a candidate gene for schizophrenia, Dow et al. [47] identified several rare variants that may have a functional impact in terms of protein function, gene regulation and splicing, although verification is required. The authors sequenced 92 schizophrenia cases from a previous GWAS [48], and imputed the data using European-descent samples from both the HapMap Phase 3 and 1000 Genomes Project data as reference panels. In their sequencing study they did not detect any additional variants that were in strong LD with the previously identified GWAS SNPs, and the same previously identified variants were found to have the strongest association signals in the imputation analysis.

In the only example to use rare variant aggregate tests with imputed data to date, Mägi et al. [20] used GRANVIL to analyze 14,000 cases of 7 common complex diseases and 3,000 shared controls from the WTCCC [49], imputed using multiple populations from the 1000 Genomes Project Phase 1 reference panel (June 2011 interim release), to re-assess the evidence for association with rare variants. Genome-wide significant evidence of an association between an accumulation of rare variants and decreased risk of coronary artery disease was identified in *PRDM10*, while 10 genes (9 risk-increasing, 1 protective) within the MHC exhibited genome-wide significant rare variant associations with type 1 diabetes.

Discussion

The analysis of rare variants is a relatively new field that has grown exponentially with the development of analysis methods and access approaches, but to date there remain few published studies that employ rare variant association methods. However, it is anticipated that there will soon be an increase in the practical use of locus-based tests designed for rare variants, as a greater understanding of the methods and challenges are addressed. General difficulties in rare variant analysis include choice of a significance threshold for locus-based tests, as well as selection of regions for analysis. A few of the possible region choices are sliding windows of a fixed size, sliding windows of a fixed number of rare variants, or functional units.

Additional challenges include the functional annotation of non-coding rare variants, population stratification, and family-based analysis. The prediction of molecular function at non-coding variants is more difficult than at coding variants, and there are databases consisting of known non-coding variant that affect phenotype, but these are not of use for novel variants, as is often the case for rare variants [32]. There are existing methods that can adjust for confounding of population structure, but under certain conditions such methods have been demonstrated to be ineffective for rare variants [50]. Specifically, when the distribution of non-genetic risk is concentrated into a small, sharply defined region(s), rare variants were shown to have a tendency towards stronger stratification that is systematically different from that for common variants, so that the usual corrections used in the standard GWAS setting are not effective. The authors describe three ways in which such a scenario may occur:

(1) highly patchy localized environmental exposure; (2) systematic measurement bias at a recruitment centre, and (3) local variation in recruitment policy or misclassification rates.

Family-based designs are more likely to be enriched for rare variants, since if one parent possesses a rare minor allele, half of the children are expected to carry it as well [51]. Two adaptive weighting methods for rare variant association tests in family-based data have been recently proposed for sequenced data with quantitative traits by Fang et al. [52]. Through simulation studies, they showed that their methods are powerful, even when there are different directions of effect, and robust to population stratification. There is currently no extension to imputed data, but as with other rare variant methods, a quick extension is to replace genotype scores by dosages. However, it is unknown how this change to the genotype accuracy will affect the power of the test.

Methodology for testing rare variant associations continues to grow, and many challenges still remain. Imputation has been proven to be a useful tool in accessing non-common variants, but careful selection of a reference panel is essential to achieve highly accurate results. It has become more common for studies to take advantage of imputation to access low-frequency/rare variants for association testing, and as more studies combine such data with powerful rare variant analysis tools, additional pieces of the missing heritability problem are expected to be filled.

Acknowledgements

This work was supported by the Wellcome Trust (098051).

References

- 1 Howie BN, Donnelly P, Marchini J: A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 2009;5:e1000529.
- 2 Marchini J, Howie B: Genotype imputation for genome-wide association studies. *Nat Rev Genet* 2010;11:499–511.
- 3 Li L, Li Y, Browning SR, Browning BL, Slater AJ, et al: Performance of genotype imputation for rare variants identified in exons and flanking regions of genes. *PLoS One* 2011;6:e24945.
- 4 Marchini J, Howie B, Myers S, McVean G, Donnelly P: A new multipoint method for genome-wide association studies via imputation of genotypes. *Nat Genet* 2007;39:906–913.
- 5 Li Y, Ding J, Abecasis GR: Mach 1.0: rapid haplotype reconstruction and missing genotype inference. *Am J Hum Genet* 2006;79:S2290.
- 6 Li Y, Willer C, Sanna S, Abecasis G: Genotype Imputation. *Annu Rev Genomics Hum Genet* 2009;10:387–406.
- 7 Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR: Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* 2012;44:955–959.
- 8 Pei YF, Li J, Zhang L, Papasian CJ, Deng HW: Analyses and comparison of accuracy of different genotype imputation methods. *PLoS One* 2008;3:e3551.
- 9 Huang L, Li Y, Singleton AB, Hardy JA, Abecasis G, et al: Genotype imputation accuracy across worldwide human populations. *Am J Hum Genet* 2009;84:235–250.
- 10 Sung YJ, Wang L, Rankinen T, Bouchard C, Rao DC: Performance of genotype imputations using data from the 1000 Genomes Project. *Hum Hered* 2012;73:18–25.
- 11 International HapMap Consortium: The International HapMap Project. *Nature* 2003;426:789–796.
- 12 The International HapMap 3 Consortium: Integrating common and rare genetic variation in diverse human populations. *Nature* 2010;467:52–58.

- 13 Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al: Finding the missing heritability of complex diseases. *Nature* 2009; 461:747–753.
- 14 The 1000 Genomes Project Consortium: A map of human genome variation from population-scale sequencing. *Nature* 2010;467: 1061–1073.
- 15 Jostins L, Morley KI, Barrett JC: Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. *Eur J Hum Genet* 2011;19:662–666.
- 16 Jewett EM, Zawistowski M, Rosenberg NA, Zöllner S: A coalescent model for genotype imputation. *Genetics* 2012;191:1239–1255.
- 17 Paşaniuc B, Avinery R, Gur T, Skibola CF, Bracci PM, Halperin E: A generic coalescent-based framework for the selection of a reference panel for imputation. *Genet Epidemiol* 2010;34:773–782.
- 18 Howie B, Marchini J, Stephens M: Genotype imputation with thousands of genomes. *PLoS Genet* 2011;1:457–470.
- 19 Sung YJ, Gu CC, Tiwari HK, Arnett DK, Broeckel U, et al: Genotype imputation for African Americans using data from HapMap Phase II versus 1000 Genomes Projects. *Genet Epidemiol* 2012;36:508–516.
- 20 Mägi R, Asimit JL, Day-Williams AG, Zeggini E, Morris AP: Genome-wide association analysis of imputed rare variants: application to seven common complex diseases. *Genet Epidemiol* 2012, E-pub ahead of print. DOI: 10.1002/gepi.21675.
- 21 Fridley BL, Jenkins G, Deyo-Svendsen ME, Hebbiring S, Freimuth R: Utilizing genotype imputation for the augmentation of sequence data. *PLoS One* 2010;5:e11018.
- 22 Paşaniuc B, Rohland N, McLaren PJ, Garmella K, Zaitlen N, et al: Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat Genet* 2012;44:631–635.
- 23 Sampson JN, Jacobs K, Wang Z, Yeager M, Chanock S, et al: A two-platform design for next generation genome-wide association studies. *Genet Epidemiol* 2012;36:400–408.
- 24 Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, et al: Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet* 2011;43:1066–1073.
- 25 Liu DJ, Leal SM: SEQCHIP: a powerful method to integrate sequence and genotype data for the detection of rare variant associations. *Bioinformatics* 2012;28:1745–1751.
- 26 Li B, Leal SM: Discovery of rare variants via sequencing: implications for the design of complex trait association studies. *PLoS Genet* 2009;5:e1000481.
- 27 Zhou X, Stephens M: Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* 2012;44:821–824.
- 28 Asimit JL, Zeggini E: Rare variant association analysis methods for complex traits. *Ann Rev Genet* 2010;44:293–308.
- 29 Bansal V, Libiger O, Torkamani A, Schork NJ: Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet* 2010;11:773–785.
- 30 Basu S, Pan W: Comparison of statistical tests for disease association with rare variants. *Genet Epidemiol* 2011;35:606–619.
- 31 Ladouceur M, Dastani Z, Aulchenko YS, Greenwood CMT, Richards JB: The empirical power of rare variant association methods: results from sanger sequencing in 1,998 individuals. *PLoS Genet* 2012;8:e1002496.
- 32 Zawistowski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, Zöllner S: Extending rare variant testing strategies: analysis of noncoding sequence and imputed genotypes. *Am J Hum Genet* 2010;87:604–617.
- 33 Morris AP, Zeggini E: An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol* 2010;34:188–193.
- 34 Li Y, Byrnes AE, Li M: To identify associations with rare variants, just WHaIT: weighted haplotype and imputation-based tests. *Am J Hum Genet* 2010;87:728–735.
- 35 Asimit JL, Day-Williams AG, Morris AP, Zeggini E: ARIEL and AMELIA: testing for an accumulation of rare variants using next-generation sequencing data. *Hum Hered* 2012;73: 84–94.
- 36 Mukhopadhyay I, Feingold E, Weeks DE, Thalamuthu A: Association tests using kernel-based measures of multi-locus genotype similarity between individuals. *Genet Epidemiol* 2010;34:213–221.
- 37 Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X: Rare variant association testing for sequencing data using the Sequence Kernel Association Test. *Am J Hum Genet* 2011;89:82–93.
- 38 Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, NHLBI GO Exome Sequencing Project-ESP Lung Project Team, Christiani DC, Wurfel MM, Lin X: Optimal unified approach for rare variant association testing with application to small sample case-control whole-exome sequencing studies. *Am J Hum Genet* 2012;91:224–237.
- 39 Kong A, Masson G, Frigge ML, Gylfason A, Zusanovich P, et al: Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet* 2008;40:1068–1075.
- 40 Palin K, Campbell H, Wright AF, Wilson JF, Durbin R: Identity-by-descent phasing and imputation in founder populations using graphical models. *Genet Epidemiol* 2011;35: 853–860.
- 41 Uricchio LH, Chong JX, Ross KD, Ober C, Nicolae DL: Accurate imputation of rare and common variants in a founder population from a small number of sequenced individuals. *Genet Epidemiol* 2012;36:312–319.
- 42 Gusev A, Shah MJ, Kenny EE, Ramachandran A, et al: Low-pass genome-wide sequencing and variant inference using identity-by-descent in an isolated human population. *Genetics* 2012;190:679–689.
- 43 Zeggini E: Next-generation association studies for complex traits. *Nat Genet* 2011;43: 287–288.
- 44 Holm H, Gudbjartsson DF, Sulem P, Masson G, Helgadóttir HT, et al: A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nat Genet* 2011;43:316–323.
- 45 Day-Williams AG, Southam L, Panoutsopoulou K, Rayner NW, Esko T, et al: A variant in MCF2L is associated with osteoarthritis. *Am J Hum Genet* 2011;89:446–450.
- 46 Germain M, Saut N, Oudot-Mellakh T, Letenneur L, Dupuy AM, et al: Caution in interpreting results from imputation analysis when linkage disequilibrium extends over a large distance: a case study on venous thrombosis. *PLoS One* 2012;7:e38538.
- 47 Dow DJ, Huxley-Jones J, Hall JM, Francks C, Maycox PR, et al: ADAMTSL3 as a candidate gene for schizophrenia: gene sequencing and ultra-high density association analysis by imputation. *Schizophr Res* 2011;127:28–34.
- 48 Need AC, Ge D, Weale ME, Maia J, Feng S, et al: A genome-wide investigation of SNPs and CNVs in schizophrenia. *PLoS Genet* 2009; 5:e1000373.
- 49 The Wellcome Trust Case Control Consortium: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007;447:661–678.
- 50 Mathieson I, McVean G: Differential confounding of rare and common variants in spatially structured populations. *Nat Genet* 2012; 44:243–248.
- 51 Shi G, Rao D: Optimum designs for next-generation sequencing to discover rare variants for common complex disease. *Genet Epidemiol* 2011;35:572–579.
- 52 Fang S, Sha Q, Zhang S: Two adaptive weighting methods to test for rare variant associations in family-based designs. *Genet Epidemiol* 2012;36:499–507.