IMPUTATION-BASED ASSESSMENT OF NEXT GENERATION RARE EXOME VARIANT ARRAYS

ALICIA R. MARTIN[*]

Department of Genetics & Biomedical Informatics Training Program, Stanford University
Stanford, CA, 94305
Email: armartin@stanford.edu

GERARD TSE

Department of Computer Science, Stanford University
Stanford, CA, 94305
Email: gerardtse@gmail.com

CARLOS D. BUSTAMANTE

Department of Genetics, Stanford University
Stanford, CA, 94305
Email: cdbustam@stanford.edu

EIMEAR E. KENNY[*]

Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai
New York, NY 10029
Email: eimear.kenny@mssm.edu

A striking finding from recent large-scale sequencing efforts is that the vast majority of variants in the human genome are rare and found within single populations or lineages. These observations hold important implications for the design of the next round of disease variant discovery efforts—if genetic variants that influence disease risk follow the same trend, then we expect to see population-specific disease associations that require large samples sizes for detection. To address this challenge, and due to the still prohibitive cost of sequencing large cohorts, researchers have developed a new generation of low-cost genotyping arrays that assay rare variation previously identified from large exome sequencing studies. Genotyping approaches rely not only on directly observing variants, but also on phasing and imputation methods that use publicly available reference panels to infer unobserved variants in a study cohort. Rare variant exome arrays are intentionally enriched for variants likely to be disease causing, and here we assay the ability of the first commercially available rare exome variant array (the Illumina Infinium HumanExome BeadChip) to also tag other potentially damaging variants not molecularly assayed. Using full sequence data from chromosome 22 from the phase I 1000 Genomes Project, we evaluate three methods for imputation (BEAGLE, MaCH-Admix, and SHAPEIT2/IMPUTE2) with the rare exome variant array under varied study panel sizes, reference panel sizes, and LD structures via population differences. We find that imputation is more accurate across both the genome and exome for common variant arrays than the next generation array for all allele frequencies, including rare alleles. We also find that imputation is the least accurate in African populations, and accuracy is substantially improved for rare variants when the same population is included in the reference panel. Depending on the goals of GWAS researchers, our results will aid budget decisions by helping determine whether money is best spent sequencing the genomes of smaller sample sizes, genotyping larger sample sizes with rare and/or common variant arrays and imputing SNPs, or some combination of the two.

---

[*] Corresponding authors

1. Introduction

The ability to measure human genetic variation on a genome-scale reliably and inexpensively in research settings has fueled and shaped the movement toward personalized medicine in health care. A prominent strategy for discovering genetic variants underlying disease susceptibility is through genome-wide association studies (GWAS), in which a subset of genetic variation is observed or inferred via linkage disequilibrium (LD), and correlated with disease state. GWAS have been successful in identifying thousands of reproducible associations with complex disease, which have had some utility in clinical practice[1,2]. However, most variants identified in GWAS with genotyping arrays are of small effect and fail to explain a large portion of genetic variation, even when the disease is estimated to be highly heritable[3]. Population genetics and neutral theory suggest that common variation might be less important than rare variation in these cases because selective pressure has had more time to eliminate deleterious alleles. With the advent of next generation sequencing technology, large consortia seeking to identify nonsynonymous coding changes have emerged. A salient result of these large-scale projects is that the vast majority of genetic variation is rare and exhibits little sharing among diverged populations[4–6]. The sequencing costs for an exome still outweigh those of genotyping arrays, however, and large sample sizes are required to detect rare variants. This creates a budget dilemma for GWAS researchers trying to explain the genetic basis of disease regarding the number of individuals they can afford to study with sequencing versus genotyping methods.

As a consequence of these findings, researchers have designed a next generation genotyping array that enriches for nonsynonymous rare coding variants. More than 15 labs with exome sequencing data from ~12,000 individuals contributed to the ascertainment of SNPs to include in the first rare variant array. The current design of the first publicly available next generation array, the Illumina Infinium HumanExome BeadChip, consists of only ~250,000 variants, a fraction of the sites that most common variant arrays currently assay. The vast majority of sites are rare coding variants; the remaining sites include randomly selected synonymous single nucleotide polymorphisms (SNPs), Native American and African ancestry informative markers, GWAS tag SNPs, HLA tags, common scaffold SNPs, and ~2,000 variants from other functional classes. A potential way to bolster the number of sites is through statistical inference of variants not molecularly assayed on the genotyping array through phasing and imputation guided by publicly available reference panels[4,7,8]. Phasing and imputation methods rely on the correlated inheritance between neighboring alleles or linkage disequilibrium (LD) between assayed alleles. LD is substantially reduced between variants on the rare exome array overall, however, because the number of scaffold SNPs is substantially reduced compared to other GWAS arrays (5,286 SNPs total compared to hundreds of thousands on common variant arrays). Admixture mapping, an approach often used when ancestry confounds GWAS associations, also relies heavily on a dense scaffold of linked markers. For example, results from HapMix, a method for inferring local ancestry across chromosomes, indicated that accuracy is reduced with fewer than 50,000 scaffold markers even when admixture is recent[9].

In order to better understand the amenability of rare exome variant arrays to existing phasing and

imputation methods, we have performed evaluations of multiple LD-based methods as well as parameters that influence imputation accuracy, including sample size and population. We find that imputation with common variant arrays is more accurate across both the exomic and genomic regions of chromosome 22, highlighting the importance of contextual variants in imputation and suggesting that the Illumina Infinium HumanExome BeadChip is not ideal for imputation purposes.

2. Methods

2.1. *Evaluation overview*

We based all our evaluation on the data provided by the phase I 1000 Genomes project[10], wherein 1,092 individuals from 14 distinct populations were genome sequenced, exome sequenced, and genotyped to produce an integrated variant call set. These populations include three African populations, three East Asian populations, five European populations, as well as three populations from the Americas. We created a pipeline (Figure 1) to perform phasing and imputation using three methods: BEAGLE v3.3.2[11,12] for both phasing and imputation, MaCH-Admix[8] v2.0.198 for both phasing and imputation, and ShapeIt[13,14] v2.r644 for phasing followed by Impute2[15,16] v2.2.2 for imputation (process abbreviated as SHAPEIT2/IMPUTE2).

To fairly evaluate phasing and imputation performance we compared one rare and one common variant array of approximately the same SNP density (the Illumina Infinium HumanExome BeadChip and Illumina Infinium HumanHap 300v1 containing ~250K and ~300K SNPs, respectively). To evaluate performance versus cost trade-offs, we also included two higher-cost, higher-density common variant arrays, the Affymetrix Genome-Wide Human SNP Array 6.0 and Illumina Human Omni2.5 BeadChip containing 1M and 2.5M SNPs, respectively. To generate the phasing and imputation results for each array, we sampled individuals into a reference panel and a test set. The reference panel contained all of the sequence calls on chromosome 22, while the test set was further filtered to the markers on each of the corresponding arrays (Table 1). We generated a known truth set from the full phase I integrated call set and imputed set using the imputed sites not on each of the evaluated arrays for each run for accuracy evaluation.

Table 1 - Arrays evaluated in this study and number of sites across all of chromosome 22 versus exomic regions of chromosome 22. Exome sites were filtered using sites annotated with EXOME in the phase I 1000 Genomes integrated call set and are a subset of Genome sites. Minor allele frequency (MAF) distributions are as assessed in the 1000 Genomes phase I samples across all chromosome 22 sites and are drawn for each array from a frequency of 0 – 0.5. "Dark sites" are the sites that are on the array but not in the 1000 Genomes phase I reference panel.

Array	Genome	Exome	MAF distributions	Mean MAF	Dark sites (%)
Illumina HumanOmni2.5 BeadChip	33,188	1,631		0.173	6.99
Affymetrix Genome-Wide Human SNP Array 6.0	11,739	262		0.208	1.01
Illumina Infinium HumanHap 300v1	5,376	240		0.272	0.99
Illumina Infinium HumanExome BeadChip	3,442	3,009		0.050	69.81
Total reference panel sites	475,372	16,885			

Simulated data from each of the four arrays were run through the phasing and imputation pipeline. The reference panel for each run was used as an input to the pipeline to inform the phasing and imputation algorithms. The pipeline first phased the incomplete genotypes in the test set, then imputed markers up to the reference panel markers using the same test set markers as in the phasing step as a scaffold (Figure 1). In order to speed up computational run time, we split the reference panel sites into 5 Mb windows with 250 kb flanking on either ends that were removed in post-processing to reduce edge effects between windows. We ran separate instances of imputation for each chunk in parallel, enabling the pipeline to run with reasonable memory and in reasonable time. At the end of each run, we extracted the imputed genotypes and each algorithm's confidence score (R^2 in the cases of BEAGLE and MaCH-Admix and informative measure in the case of Impute2). We calculated diploid and haploid error for each imputed site from the known truth data.
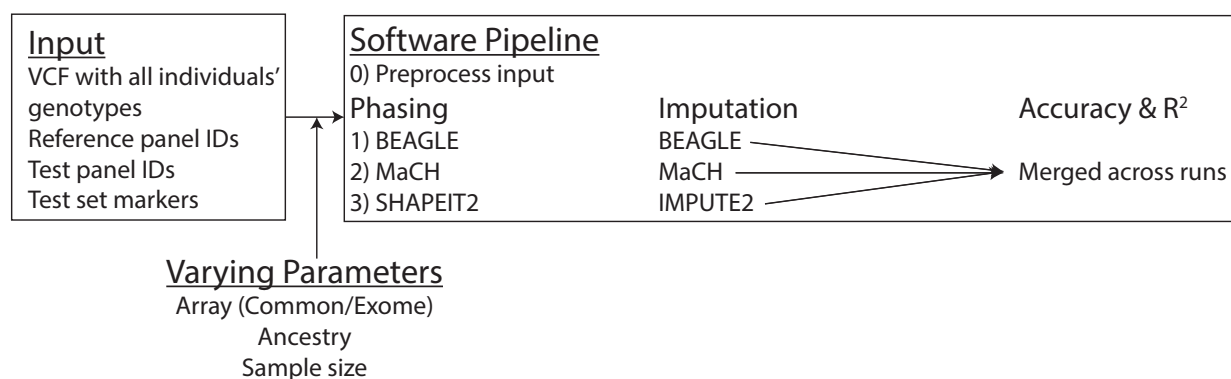

Figure 1 - Phasing and imputation pipeline. Inputs files are subsetted based on varying parameters specified, and for each set of parameters phasing and imputation was performed using three methods.

2.2. *Sampling strategy for test/reference size analyses*

Previous studies have assessed imputation accuracy on single chromosomes, including chromosomes 10 (~135 Mb), 20 (~62 Mb), and 22 (50 Mb), and have found highly consistent results[7,15,16], indicating that they are representative. As such, we used full sequence data from chromosome 22 for computational efficiency from all 1,092 individuals and sampled them randomly into two groups: A reference panel and a test set. To study the effect of different reference panels and GWAS study sizes on the accuracy of imputed haplotypes, we investigated 13 different configurations of test set and reference panel sizes: a test set of size 92 with varying reference panel sizes of 63, 125, 250, 500, and 1000; and test panel sizes of 300 and 500, each with reference panels of 62, 125, 250, and 500.

Using the reference panel to inform phasing and imputation, we ran the pipelines for each of the three common variant arrays and the rare exome array and collected the results. The results were compared to the true calls found in the unfiltered genotypes of individuals in the test set.

2.3. *Sampling strategy for population analyses*

We used full sequence data from all of the 1,092 individuals and separated them into 14 populations. Four different sampling strategies were employed to identify biases when different reference sets are used for each of the 14 populations, resulting in 56 sets of samplings, as follows. The first two samplings assessed imputation accuracy when a test population is not or is included in the reference panel, respectively. We created a test set with all individuals in each population and sampled 900 individuals from the rest of the genomes available in the 1000 Genomes project (strategy A, Figure 3). As a control for the presence of a population from the reference panel, we created another test set with half of all the individuals in each population and put the remaining half of the population in the reference panel, then added individuals from other populations randomly until the reference panel contained 900 individuals (strategy B).

The other two population samplings focused on the significance of having individuals from the same continent in the reference panel. We created a test set with 33 individuals in the population and sampled 148 from all other individuals from the same continental group (strategy C). These numbers were chosen for uniformity across populations in order to represent the smallest continental group in the data. We performed this evaluation for each population and considered four continental groups: Africans, Asians, Europeans, and Native Americans. As a control, we created another test set with 30 individuals in the population and sampled 148 from all other individuals regardless of origin (strategy D).

2.4. *Phasing and imputation summaries and analysis*

Using the reference panel to inform phasing and imputation, we ran the pipelines for each of the three common variant arrays and the rare exome array. The imputed genotypes were compared to the true calls in the unfiltered sequences of individuals in the test set. Data summaries for all three algorithms reported an informative metric (R^2), which were generated by the imputation algorithms. Because each algorithm calculates R^2 differently, we calculated diploid and haploid error, as well as minor allele frequency (MAF), in order to fairly compare the algorithms directly. We define the diploid error as any discordance between the most likely imputed and true calls, which is affected by MAF and therefore only used to compare method performances. In this scenario, if the true variant is homozygous reference, heterozygous or homozygous non-reference imputation dosages count equally toward the error. We also calculated haploid error, where in the previous scenario, a heterozygous call counts half as much toward the error as a homozygous non-reference call, which was highly correlated (>99%) with diploid error. We note that the diploid and haploid errors are critical to examine but that they are highly influenced by MAF. For example, at a site where a very rare variant exists in the reference panel, error is very low because the imputation algorithm frequently fills in the major allele, even in the absence of any surrounding variants. In contrast, when a common variant exists, the imputation algorithms require more neighboring information to correctly impute the variant. For these reasons, we assess imputation accuracy as R^2 as previously[15], except where

otherwise noted. In order to compare MAF versus imputation accuracy, we performed local regression weighted by least squares. Unless otherwise noted, the span was 0.75.

3. Results

We first compared the performance of three phasing and imputation algorithms, BEAGLE, MaCH-Admix, and SHAPEIT2/IMPUTE2 under multiple conditions. The informative measure metrics are defined slightly differently for each algorithm[7], and in all cases SHAPEIT2/IMPUTE2 reports the highest informative measures (data not shown). In order to determine which method was performing most accurately based on known truth data, we compared their performance via mean diploid error across all test panel sizes, reference panel sizes, and the four arrays we evaluated, as outlined in Methods. In each case, BEAGLE had the highest error, SHAPEIT2/IMPUTE2 performed comparably with MaCH-Admix, and MaCH-Admix resulted in the lowest error, which highlights the importance of using a directly comparable metric to assess method performance. Table 2 shows the average diploid error across chromosome 22 across all reference and test panel sizes using the Affymetrix Genome-Wide Human SNP Array 6.0 for each, which showed the same trends with other arrays (data not shown). Because MaCH-Admix resulted in the lowest imputation error, all following analyses show results using this method.

Table 2 - Diploid error across multiple sample sizes. Reported values are mean percentages across all variant sites in the phase I 1000 Genomes Project on chromosome 22 using sites on the Affymetrix Genome-Wide Human SNP Array 6.0 as test markers. Individuals in the test and reference panel are the same across methods for each comparison. Imputation R^2 values are shown for each algorithm, which are defined differently for each algorithm. Note that BEAGLE R^2 averages are calculated only for values that are not "NaN," which likely increases the R^2 reported with respect to other algorithms.

Test panel size	Reference panel size	BEAGLE (%)	MaCH-Admix (%)	Shapeit+Impute2 (%)	BEAGLE (R^2)	MaCH-Admix (R^2)	Shapeit+Impute2 (R^2)
500	500	6.36	4.21	4.35	.7349	**.3762**	**.5604**
500	250	6.37	4.27	4.38	.7329	.3333	.4735
500	125	6.63	4.41	4.56	.6820	.2959	.4048
500	62	6.77	4.63	4.74	.7403	.2464	.3175
300	500	6.31	4.16	4.32	.7387	.3724	.5348
300	250	6.60	4.39	4.56	.7392	.3279	.4567
300	125	6.57	4.37	4.53	.7344	.2954	.3928
300	62	6.87	4.66	4.79	.7331	.2513	.3191
92	1000	**6.36**	**4.13**	**4.30**	**.7653**	.3503	.4655
92	500	6.49	4.25	4.45	.7637	.3401	.4482
92	250	6.37	4.17	4.33	.7467	.3081	.3978
92	125	6.59	4.51	4.65	.7481	.2799	.3540
92	63	6.68	4.40	4.59	.7123	.2506	.3033

We next evaluated the impact of test and reference panel sizes on imputation accuracy, as assessed by R^2, for the four arrays described previously (Figure 2). We compared three test panel sizes (92, 300, and 500) and find that in all cases, larger test panels have greater imputation accuracy,

indicating that phasing and imputing a full study set together improves imputation accuracy. We also find that reference panel size has a greater impact on imputation accuracy than test panel size when the test panel contains greater than 92 individuals. These results indicate that large reference panels are necessary to accurately impute variants.
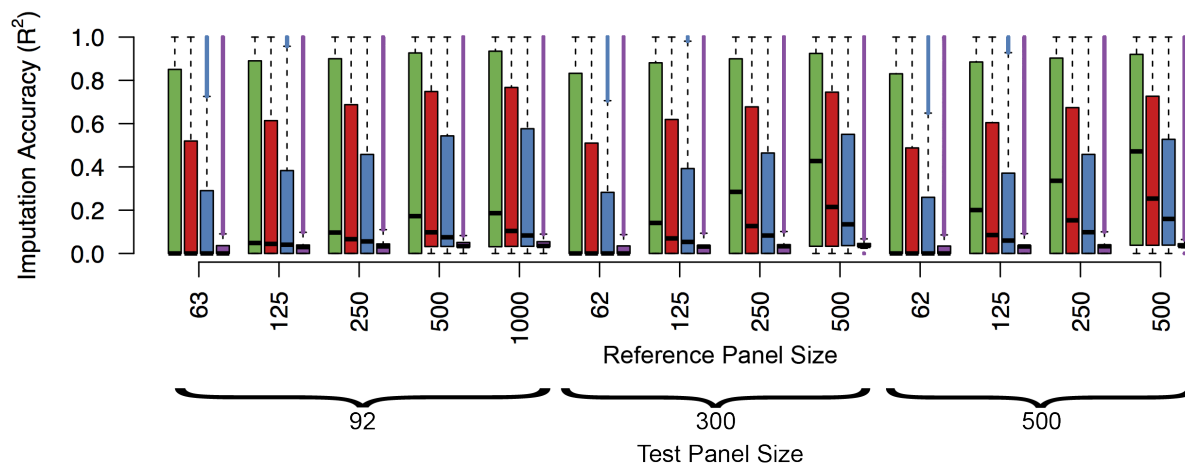

Figure 2 - Imputation accuracy across varying reference and test panel sizes. Phasing and imputation was performed using MaCH-Admix. Test panel markers were ascertained on chromosome 22 using sites from four arrays in the following colors: green – Illumina HumanOmni2.5 BeadChip, red – Affymetrix Genome-Wide Human SNP Array 6.0, blue – Illumina Infinium HumanHap 300v1, purple – Illumina Infinium HumanExome BeadChip. On the x-axis, the first number indicates the number of individuals included in the test panel, and the second number is the number of individuals included in the reference panel.

The effect of reference panel size on imputation accuracy is especially pronounced when fewer markers are assayed. For example, imputation accuracy is not substantially reduced for most common sites across chromosome 22 (MAF > 5%) when the reference panel size is reduced from 500 individuals to only 62 individuals using the dense Illumina HumanOmni2.5 BeadChip, and most common sites maintain an R^2 of ~0.9. In contrast, the accuracy drops considerably between a reference panel size of 500 versus 62 with the sparser Illumina Infinium HumanHap 300v1 (e.g. reduction of 13% from R^2=0.772 to 0.669 at MAF=0.3) and Illumina Infinium HumanExome BeadChip arrays (e.g. reduction of 26% from R^2=0.146 to 0.108 at MAF=0.3). We also find that accuracy plateaus as a function of minor allele frequency (MAF). Additionally, invariant reference panel SNPs likely drive the number of "dark sites" on each array (Table 1). Interestingly, the MAF at which accuracy peaks is array-specific. For example, the Illumina Infinium HumanHap 300v1 array has a similar number of sites on chromosome 22 as the Illumina Infinium HumanExome BeadChip (Table 1); however, accuracy peaks around MAF=0.3 on the Illumina 300k array and around MAF=0.5 on the exome array. Interestingly, imputed exome rare variant array sites from genome-wide arrays are imputed more accurately than across all chromosome 22 sites for varying allele frequencies (Figure 4A-C versus Figure 4I-K), likely because scaffold sites on genome-wide arrays are enriched near exonic regions, improving imputation accuracy.

Previous work has indicated that reference panels that share more haplotypes with the study panel improve imputation accuracy compared to a random panel[17]. We compared multiple population stratifications as described is Section 2.3 (Figure 3). In all scenarios, imputation performs the poorest in individuals of African descent. This is likely due to the reduced LD structure in African populations[18] and European ascertainment bias in genotyping arrays[19]. Imputation with both global reference panel strategies with a larger number of reference individuals, albeit from more distantly related populations overall (Figure 3A and Figure 3B), outperforms imputation with smaller continental reference panels (Figure 3C and Figure 3D). Low frequency alleles are imputed with greater accuracy when the reference panel includes individuals from the same population compared to when it does not (Figure 3B versus Figure 3A). This is especially true in European populations with the exception of TSI individuals, which likely arises from the greater genetic diversity and more complicated demographic history present in Italy compared to other European populations presented here[20,21].
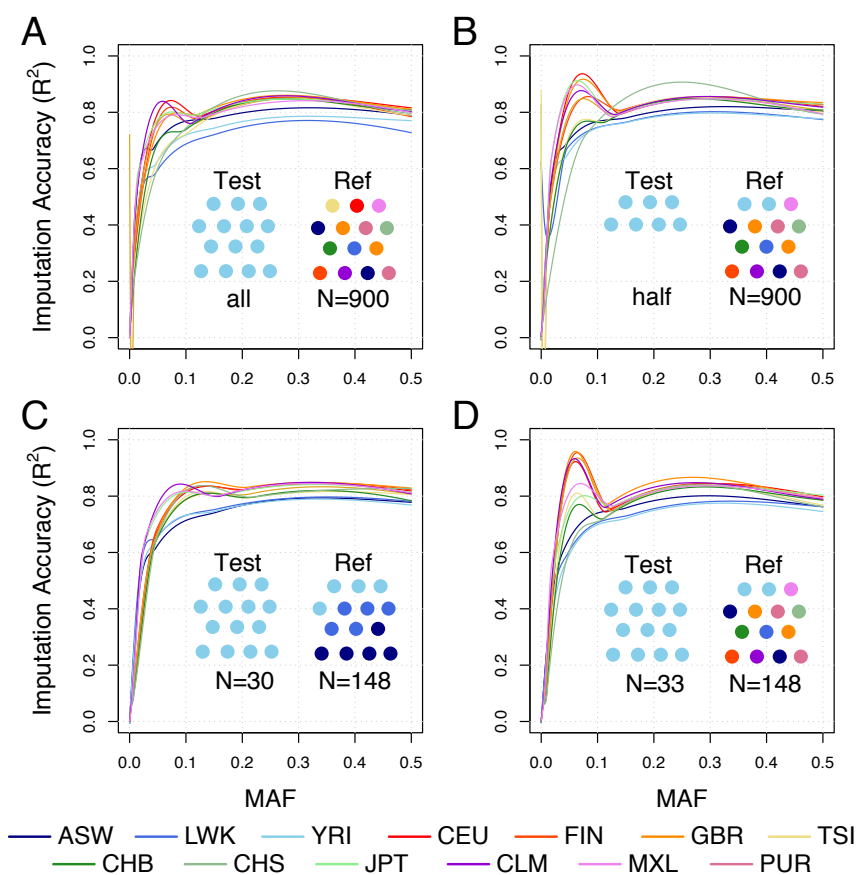

Figure 3 - Variability in imputation accuracy across populations. All simulations were performed using the Affymetrix Genome-Wide Human SNP Array 6.0 markers from chromosome 22 in the test set. Lines are local regression fits to the data, and local peaks near MAF=0 in A and B for the GBR and TSI, respectively, are simply due to smoothing edge

effects. A) Strategy A. B) Strategy B. C) Strategy C. D) Strategy D. Diagrams drawn under loess curves are cartoons of sampling strategies, as outlined in section 2.3. Abbreviations are as follows: ASW=HapMap African ancestry individuals from SW US, LWK=Luhya individuals, YRI=Yoruba individuals, CEU=CEPH individuals, FIN=HapMap Finnish individuals from Finland, GBR=British individuals from England and Scotland, TSI=Toscan individuals, CHB=Han Chinese in Beijing, CHS=Han Chinese South, JPT=Japanese individuals, CLM=Colombian in Medellin, Colombia, MXL=HapMap Mexican individuals from LA California, PUR=Puerto Rican in Puerto Rico.
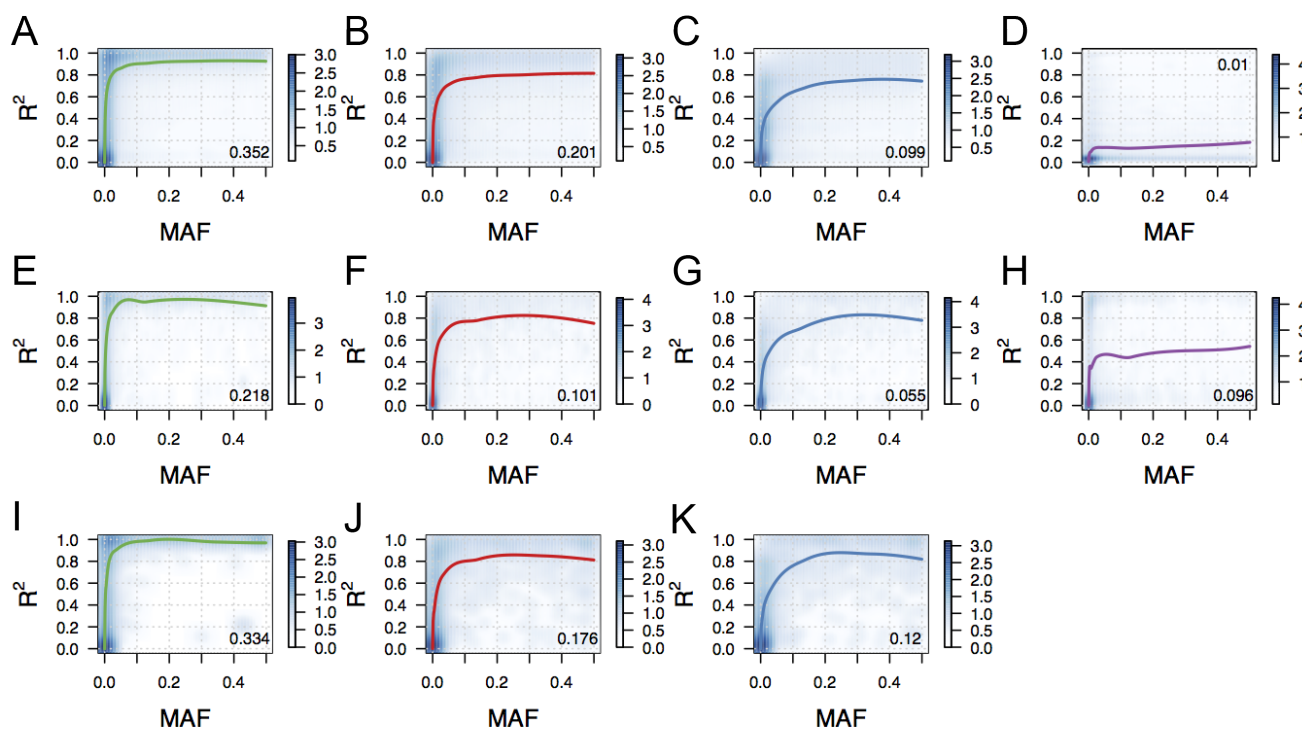
Figure 4 - Imputation accuracy across three common variant and one rare exome variant arrays in genomic, exomic, and imputable exome rare variant array regions of chromosome 22. Colors correspond with arrays, as in Figure 2. All subpanels show smoothened scatter plots with an overlaid local regression fit, and the proportion of sites imputed with $R^2 > 0.8$ is reported, which are consistent with previous results[22]. Span was adjusted to 0.5 in order to keep the fits within the bounds of the data. A-D) genomic regions of chromosome 22; E-H) exomic regions of chromosome 22; I-K) Imputation accuracy for imputable exome rare variant sites using each of the genome-wide common variant arrays.

We next compared imputation accuracy across three common variant and one rare exome variant genotyping array platforms. As expected, the common variant arrays impute sites across chromosome 22 more accurately than the Illumina Infinium HumanExome BeadChip. Surprisingly, all three common variant arrays also outperform the exome array in imputing the exome-only regions, though their accuracy is substantially reduced in the exome compared to the genome (Figure 4). Imputation accuracy is the poorest with the rare variant exome array, even though the Illumina 300k common variant array has slightly fewer assayed variants on chromosome 22 (Table 1). Aside from the exome array, accuracy improves with arrays tagging more variants, as expected. The accuracy in the rare variant exome array is increased in the exomic regions compared to all chromosome 22 variants

(Figure 4H and Figure 4D, respectively). As shown in Figure 4, the imputable exome variant sites are imputed with similar accuracy as all sites across chromosome 22 with common genome-wide arrays as a scaffold. While the "dark sites" on the exome chip will be missed, other imputable sites, which are enriched for biomedically relevant SNPs, are imputed with similar accuracy as any similar frequency SNP.

4. Discussion

We have evaluated multiple factors that influence imputation accuracy, including test and reference panel size, phasing and imputation methods, populations, and genotyping arrays. We find that both larger reference and test panels lead to greater imputation accuracy, and that reference panel size is more important than test panel size in most GWAS scenarios. Larger reference panels, regardless of population, aid imputation performance for common variants, while more closely related reference panels are critical for accurately imputing rare variants. Comparing three methods, our simulations revealed that BEAGLE was both the most computationally costly method (e.g. ~48 hours to run and 10.5G of memory for chromosome 22 with a reference size of 500 and test size of 500) and had the least accurate performance. SHAPEIT2/IMPUTE2 and MaCH-Admix were comparable in terms of computationally efficient (2 hours to run and 2G of memory versus 3.5 hours to run and 1G of memory with the same test and reference panel as in the BEAGLE case). These computational costs are consistent with previously reported values[8].

It is important to note that there is an obvious bias in imputation accuracy across populations, with the lowest accuracy in African populations. Greater accuracy in out-of-Africa groups is likely due to ascertainment bias as well as longer haplotypes from the serial founder effect during the peopling of the globe. We see improved imputation accuracy at the rare end of the allele frequency spectrum when the reference panel includes the same population as the test panel. These results suggest that nearby reference panels are especially important for large outbred groups.

Imputation with common variant arrays substantially outperforms imputation with the Illumina Infinium HumanExome BeadChip. This reduction in accuracy is apparent for all frequencies, including rare alleles, suggesting that covariance between rare and nearby alleles is low, and alleles are tagged poorly. This is likely in part due to the uneven distribution of variants on the exome array across the chromosome, reducing LD on the array. A scaffold of genomic variants will likely aid imputation accuracy in exome arrays. One potential way to assay a large number of rare variants accurately without losing important rare variant information is to combine arrays, coupling the exome array with one of the common arrays we evaluated, for example. The improved imputation accuracy by the exome array in exomic regions is likely due to denser markers and greater LD in this region. The reduction of imputation accuracy in exomic regions with the common variant arrays may be due to greater sequencing depth in the 1000 Genomes Project in the integrated call set, which contains, genotyping, genome-, and exome-sequencing data, leading to more low frequency calls passing variant filters.

Finally, alternative algorithms for phasing[23,24] that rely on identity-by-descent (IBD) structure preferentially rather than LD have recently been published. These methods take advantage of haplotypic structure and will likely aid imputation differentially depending on the degree of sharing within a population and the potential to improve phasing accuracy. A question for future work, for example, might compare phasing accuracy using LD-based and IBD-based methods in endogamous African populations where imputation with traditional arrays performs poorly but where cryptic relatedness is more likely to exist.

5. Conclusions

The next generation of genotyping arrays intends to capture rare, coding variation that is likely to contain more pathogenic variation than randomly ascertained SNPs. Here, we assess the ability of a commercially available rare variant exome array to adequately tag variation that has not been directly assayed, compared to common variant arrays. We assess multiple methods, sample sizes, and populations, and find that imputation accuracy is substantially reduced with the rare variant exome array compared to common variant arrays. This result is true both in genomic and exomic regions of chromosome 22, although the difference in imputation accuracy between common and exome arrays is reduced in exomic regions. We also find that the European ascertainment bias in common variant arrays is reflected in imputation accuracy across populations, with most European variants imputed more accurately than those of other continental groups. Additionally, closely related populations are critical in reference panels for low frequency variants. Finally, we compare three phasing and imputation methods and find that BEAGLE is the least accurate, and SHAPEIT2/IMPUTE2 performs slightly less accurately than MaCH-Admix for all reference and test panel sizes. This research provides guidelines for GWAS researchers to avoid the current design of exome rare variant arrays when imputing genotype data. We acknowledge, however, that these next generation arrays have potential utility when fine-mapping a variant that is suspected to be coding and not tagged by common variant genotyping arrays.

Acknowledgments

Appendix

All code written to run phasing and imputation simulations on a Sun Grid Engine can be downloaded here: https://github.com/armartin/compare_impute.

References

1. Hindorff, L. a *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America* **106,** 9362–7 (2009).
2. Manolio, T. a. Bringing genome-wide association findings into clinical use. *Nature reviews. Genetics* **14,** 549–58 (2013).
3. Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature reviews. Genetics* **11,** 415–25 (2010).
4. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
5. Gravel, S. *et al.* Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences of the United States of America* **108,** 11983–8 (2011).
6. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456,** 98–101 (2008).
7. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nature reviews. Genetics* **11,** 499–511 (2010).
8. Liu, E. Y., Li, M., Wang, W. & Li, Y. MaCH-admix: genotype imputation for admixed populations. *Genetic epidemiology* **37,** 25–37 (2013).
9. Price, A. L. *et al.* Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS genetics* **5,** e1000519 (2009).
10. Project, G., Asia, E., Africa, S., Figs, S. & Tables, S. An integrated map of genetic variation from 1,092 human genomes. *Nature* **135,** 0–9 (2012).
11. Browning, B. L. & Browning, S. R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American journal of human genetics* **84,** 210–23 (2009).
12. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American journal of human genetics* **81,** 1084–97 (2007).
13. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nature methods* **10,** 5–6 (2013).
14. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nature methods* **9,** 179–81 (2012).
15. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3 (Bethesda, Md.)* **1,** 457–70 (2011).
16. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics* **5,** e1000529 (2009).
17. Huang, L. *et al.* Haplotype variation and genotype imputation in African populations. *Genetic epidemiology* **35,** 766–80 (2011).
18. Henn, B. M. *et al.* Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proceedings of the National Academy of Sciences of the United States of America* **108,** 5154–62 (2011).
19. Albrechtsen, A., Nielsen, F. C. & Nielsen, R. Ascertainment biases in SNP chips affect measures of population divergence. *Molecular biology and evolution* **27,** 2534–47 (2010).
20. Esko, T. *et al.* Genetic characterization of northeastern Italian population isolates in the context of broader European genetic diversity. *European journal of human genetics : EJHG* **21,** 659–65 (2013).
21. Ralph, P. & Coop, G. The Geography of Recent Genetic Ancestry across Europe. *PLoS biology* **11,** e1001555 (2013).
22. Nelson, S. C. *et al.* Imputation-Based Genomic Coverage Assessments of Current Human Genotyping Arrays. *G3 (Bethesda, Md.)* (2013). doi:10.1534/g3.113.007161
23. Palin, K., Campbell, H., Wright, A. F., Wilson, J. F. & Durbin, R. Identity-by-descent-based phasing and imputation in founder populations using graphical models. *Genetic epidemiology* **35,** 853–60 (2011).
24. Williams, A. L., Patterson, N., Glessner, J., Hakonarson, H. & Reich, D. Phasing of many thousands of genotyped samples. *American journal of human genetics* **91,** 238–51 (2012).