

PROCEEDINGS

Open Access



Imputing rare variants in families using a two-stage approach

Samantha Lent^{*†}, Xuan Deng[†], L. Adrienne Cupples, Kathryn L. Lunetta, CT Liu and Yanhua Zhou

From Genetic Analysis Workshop 19
Vienna, Austria. 24-26 August 2014

Abstract

Background: Recent focus on studying rare variants makes imputation accuracy of rare variants an important issue. Many approaches have been proposed to increase imputation accuracy among rare variants, from reference panel selection to combinations of existing methods to multistage analyses. We aimed to bring the strengths of these new approaches together with our proposed two-stage imputation for family data.

Methods: Our imputation methods were tested on the region from 46.75Mb to 49.25Mb on chromosome 3. We did quality control based on the proportion of missing genotypes per variant and individual, leaving 495 individuals with 761 genome-wide association studies (GWAS) variants only, 45 with 14,077 sequence variants only, and 419 with both GWAS and sequencing data. All data were prephased using SHAPEIT2 with a duo hidden Markov model algorithm prior to performing imputation. Imputations were performed 100 times, each time masking the sequence data for 1 individual and imputing it from the GWAS data. We used well-imputed genotypes, defined as a probability of greater than 0.9, above 2 different minor allele frequency cutoffs—0.01 and 0.05—from Impute2 as input for Merlin, and compared these results to Impute2 and Merlin separately. The imputed results were evaluated using correlation measurement and the imputation quality score.

Results: Our method improved imputation accuracy, measured by imputation quality score, for variants with minor allele frequency between 0.01 and 0.40, but failed to improve accuracy for variants with minor allele frequency less than 0.01 when we used a minor allele frequency cutoff of 0.01 for the Impute2 results. In contrast, our 2-stage approach with a minor allele frequency cutoff of 0.05 performed the worst of all methods for variants with minor allele frequency between 0.01 and 0.40.

Conclusions: This method gave promising results, but may be further improved by changing the inclusion criteria of Impute2 variants. More analyses are needed on a larger region with different inclusion thresholds to assess the accuracy of this approach.

Background

Although existing population-based genotype imputation methods are very accurate for common variants, with overall best-guess error rates of 5 % to 7 % for the most common methods [1], they do not perform nearly as well with rare variants. Only 78 % of variants with a minor allele frequency (MAF) between 0.01 and 0.05 in the Illumina 550K panel and 57 % in the Affymetrix 500K panel can be well imputed ($r^2 > 0.7$) using BEAGLE [2].

Most efforts to improve rare variant imputation have focused on how the choice of reference panel affects imputation quality. However, recently Saad et al [3] and Kreiner-Møller et al [4] have proposed methods to improve imputation using multistep procedures. Saad et al proposed using 2 imputation methods independently, 1 population based (BEAGLE) and 1 family based (Genotype Imputation Given Inheritance [GIGI]), and choosing the imputed data from the method with the highest variance in genotype probabilities for each single nucleotide polymorphism (SNP). For instance, if the probabilities for genotypes AA, AB, and BB in an individual

* Correspondence: lent@bu.edu

†Equal contributors

Department of Biostatistics, Boston University, Boston, MA, USA



Table 1 Distribution of family size

Family size	3	4	5	6	7	8	9	10	11	12	13	14	15
No. of families	6	8	24	17	15	5	8	1	3	7	2	2	2

are 0, 0, and 1.0, respectively, for BEAGLE and 0, 0.5, and 0.5 for GIGI, Saad et al's method would choose BEAGLE for that variant, because the larger variance indicates more certainty in the call. Saad et al found that the combined method led to more accurate imputed genotypes than either method separately. Kreiner-Møller et al suggested a 2-step imputation using a local reference panel and the 1000 Genomes reference panel, implemented in MACH/Minimac [4, 5]. In the first step, they imputed the study sample to a densely genotyped local reference panel enriched for rare variants. Next, they used the best-guess genotypes from this imputation as well as the original genotypes to impute the study sample to the 1000 Genomes panel.

Our approach combined the strengths of Saad et al and Kreiner-Møller et al. We performed a 2-stage imputation, implementing Impute2 and Merlin sequentially, to test the hypothesis that increasing the density of genotypes in a sequenced reference panel using a population-based imputation before performing a family-based imputation would lead to higher imputation accuracy in a related genome-wide association studies (GWAS) study panel.

Methods

Quality control

Our sample consisted of 959 Mexican Americans from 20 families. All 959 subjects were genotyped on the Illumina platform, and 464 of these individuals were also sequenced. We removed all SNPs with more than 5 % missing data and all individuals with more than 5 % missing data ($N = 45$) from the GWAS samples, and limited our analysis to the 46.75 Mb to 49.25 Mb region on chromosome 3. This yielded 914 people with GWAS data and 761 Illumina variants. For the sequenced data, we removed any variant with more than 10 % missing data, leaving 14,077 sequenced variants. All sequenced individuals had less than 5 % missing data. Thus, all 959 individuals were included in the analyses: 495 with

GWAS only, 45 with sequencing only, and 419 with both GWAS and sequencing.

Phasing

All data were prephased using SHAPEIT2 prior to performing imputation [6]. We used the duo hidden Markov model (duoHMM) algorithm in SHAPEIT, which uses pedigree information from trios to improve phasing and eliminate Mendelian errors. GWAS and sequence data were phased in separate runs.

Imputation

We performed 100 imputations each with 3 different methods: population-based imputation with Impute2 2.3.1, family-based imputation with Merlin 1.1.2, and a combination of the two [7, 8]. For each of these 100 imputations, we masked the sequence data of 1 individual, using the individual's GWAS data instead, and imputed the sequenced variants not in the GWAS data. After the imputation, we compared this individual's imputed genotypes to his or her true sequenced genotypes. We chose which sequenced subjects to leave out by randomly ordering all 419 subject IDs—excluding the 45 participants with sequence data but no GWAS data—and choosing the first 100.

For the population-based imputation benchmark, we used Impute2 with the default settings. The reference panel included both a local reference panel of the sequenced study individuals and a cosmopolitan reference panel of all populations from the 1000 Genomes Project (1KGP) [5]. For the family-based imputation benchmark, we used Merlin, which combines sparse marker data and high-density genotype data on several individuals to infer unobserved high-density genotypes for related individuals [9]. In the Merlin-only imputation, only our population samples were used as the imputation backbone. Each Merlin imputation included the masked individual and their nuclear family, grandchildren, and grandparents. Table 1 shows the distribution of family size for 100 individuals. The maximum proportion of parents and spouses of the masked individuals with genotype data for sequence variants is 0.667 and the

Table 2 Summary statistics of correlation and IQScomparing the imputation with dense markers and sparse markers

Quality measurements		Minimum	Median	Mean	Maximum	SD
Correlation	Masked individuals with GWA	0.00049	0.6983	0.5766	1	0.3714
	Masked individuals with GWA in LE	0.00087	0.6798	0.5708	1	0.3729
	Impute with -cluster option	0.0000	0.6879	0.5748	1	0.3727
IQS	Masked individuals with GWA	-0.04793	0.4046	0.3758	0.9715	0.3007
	Masked individuals with GWA in LE	-0.04758	0.3883	0.3659	0.9682	0.2980
	Impute with -cluster option	-0.04887	0.3989	0.3712	0.9682	0.2995

GWA genome-wide association

Table 3 Tabulation of genotypes used for IQS calculation

Imputed Genotypes	True Genotypes			Total
	AA	AB	BB	
AA	n_{11}	n_{12}	n_{13}	$n_{1.}$
AB	n_{21}	n_{22}	n_{23}	$n_{2.}$
BB	n_{31}	n_{32}	n_{33}	$n_{3.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{..}$

minimum proportion is 0. The mean proportion is 0.3796 with a standard deviation of 0.35.

Because the algorithm used in Merlin depends on markers being in linkage equilibrium (LE), we also compared the family-based imputation qualities by using sparse markers, dense markers, or the haplotype-block approach [10] (with `-cluster` option in Merlin). To get sparse markers, we pruned the GWAS variants in the region (46.75Mb to 49.25Mb) on chromosome 3 by only keeping variants with pairwise r^2 less than 0.2 implemented in PLINK 1.9, which yielded 91 variants in approximate LE. The mean pairwise r^2 for the 91 variants was 0.0252 and the median was 0.0014. To get the clustered markers and haplotype frequencies, we searched for GWAS markers for which r^2 is larger than 0.2 and defined the clusters, including each identified pair and intervening markers, which were implemented in Merlin with the `-rsq` and `-cfreq` options. The imputations were conducted with all GWAS variants (dense markers), pruned GWAS variants in LE and dense markers with predefined haplotypes, separately. Table 2 presents the imputation quality measurements (correlation and imputation quality score [IQS]). Because of the slight differences between these 3 strategies as seen in Table 2 and the fraction of parents and spouses of the masked individuals having genotype data for sequence variants, we conclude that the linkage disequilibrium present in

the data is not affecting the Merlin imputation adversely in this study.

Finally, for the combined imputation method, we selected the best-guess genotypes for all SNPs with MAF greater than 2 different cutoffs—0.01 and 0.05—and posterior probability of the best-guess genotype greater than 0.9, and used these genotypes as well as the GWASSNPs as input for Merlin. Merlin automatically excluded from imputation any variant with Mendelian-inconsistent genotyping errors, but it is possible that Impute2 introduced Mendelian-consistent genotyping errors. However, the 2-stage and Merlin-only results were almost identical for variants with MAFs below the cutoff, which leads us to believe that these potential errors introduced by Impute2 did not negatively affect imputation quality in our sample.

Accuracy assessment

We used 2 different measures of accuracy: correlation between imputed dosage and true dosage and IQS, a measure developed by Lin et al in 2010 [11], inspired by Cohen’s Kappa statistic [12]. Cohen’s Kappa measures the agreement between 2 methods of classification, adjusting for chance agreement. To apply this to imputation results, we first tabulate the imputed best-guess genotypes and true genotypes, as shown in Table 3, where n_{ij} is the number of individuals with true genotype i and imputed genotype j . Cohen’s Kappa statistic is given by:

$$\kappa = \frac{\frac{\sum_i n_{ii}}{n_{..}} - \frac{\sum_i n_{i.} n_{.i}}{n_{..}^2}}{1 - \frac{\sum_i n_{i.} n_{.i}}{n_{..}^2}}$$

This statistic adjusts for agreement by chance by subtracting the expected cell counts along the diagonal

Table 4 Summary of Imputation Quality by MAF

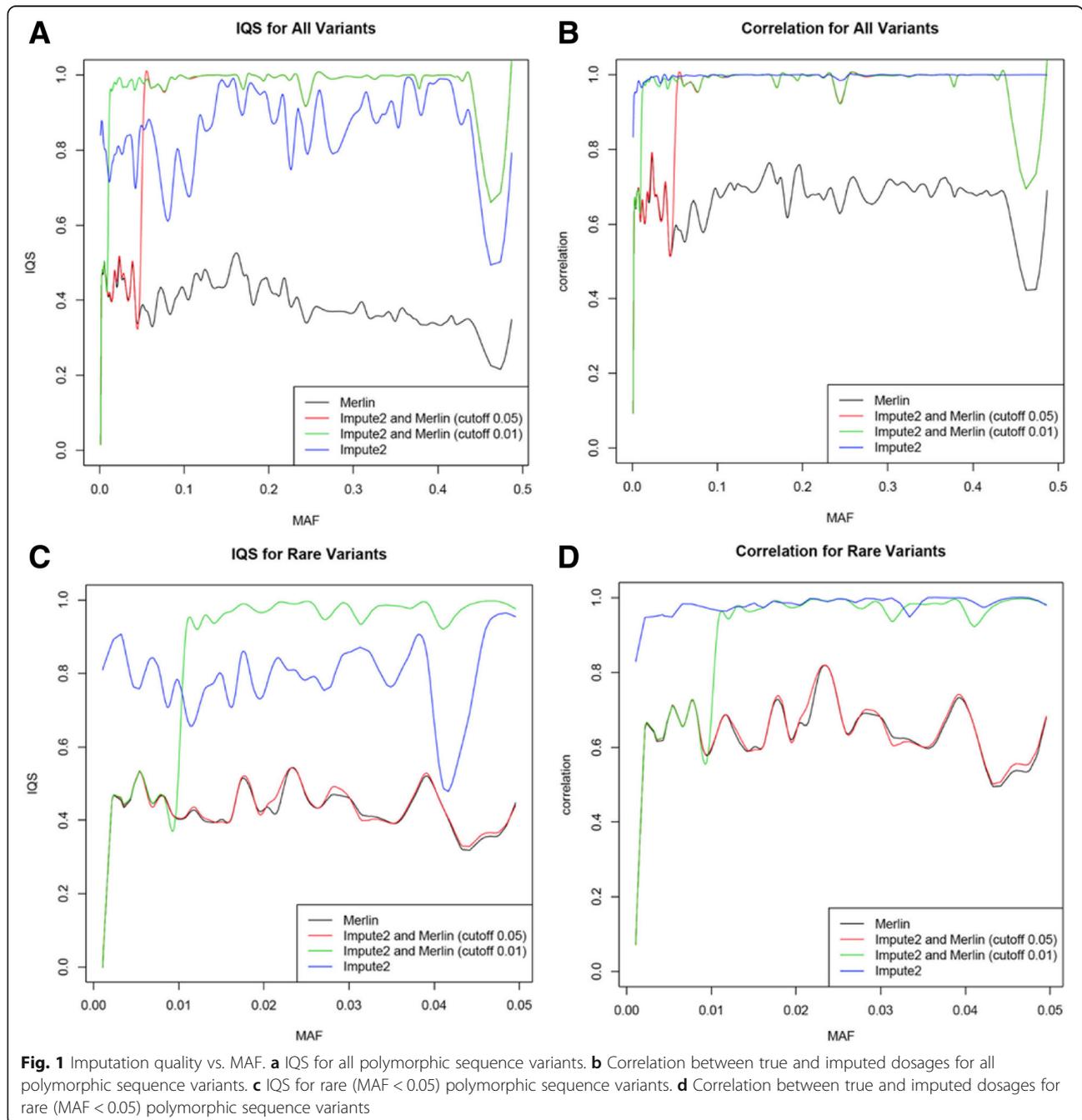
Imputation Approach	(0,0.01) 4028 SNPs			(0.01,0.05) 1416 SNPs			(0.05,0.4) 1142 SNPs		
	#SNP _p *	Mean	Var	#SNP _p *	Mean	Var	#SNP _p *	Mean	Var
	IQS								
Impute2	3023	0.840	0.097	1164	0.772	0.126	761	0.899	0.048
Merlin	4028	0.348	0.132	1416	0.437	0.041	1142	0.404	0.006
Combined (0.01) ^a	4028	0.350	0.133	1416	0.965	0.017	1142	0.992	0.004
Combined (0.05) ^a	4028	0.349	0.133	1416	0.443	0.041	1142	0.992	0.004
Correlation									
Impute2	3023	0.918	0.048	1164	0.981	0.006	761	0.999	0.00004
Merlin	4028	0.512	0.195	1416	0.663	0.064	1142	0.687	0.007
Combined (0.01) ^a	4028	0.514	0.196	1416	0.975	0.010	1142	0.994	0.003
Combined (0.05) ^a	4028	0.513	0.196	1416	0.669	0.063	1142	0.993	0.003

*#SNP_p is the number of SNPs with a MAF greater than 0 for both real and imputed genotypes (varies by method)

^aCombined (m) indicates the 2-stage imputation approach with MAF cutoff m

(which indicates agreement) from the observed proportion of agreement. In cases where the expected agreement is high, such as with variants with low MAFs, the second term in the numerator is higher, thus lowering the Kappa statistic. Lin et al extended this idea to incorporate the uncertainty of imputation by using the posterior probabilities of all 3 genotypes instead of the best-guess genotype, thus allowing the cells in Table 3 to have noninteger values. Cohen's Kappa and the IQS are equivalent when all cells in Table 3 are integers (ie, when

all posterior probabilities are 0 or 1), but differ when there is uncertainty in the imputation. Consequently, IQS is useful for rare variants because, unlike concordance, it accounts for allele frequency and adjusts for chance agreement. Furthermore, IQS can be computed using dosages, which gives more information about imputation quality than best-guess genotypes. Lin et al have compared the performance of IQS and concordance for population-based imputations implemented in Impute2. The authors show that concordance increases



with decreased MAF, whereas IQS drops as MAF decreases. The decreasing imputation quality with decreasing MAF is expected, as rare variants do not impute well [13], making IQS a better measure of imputation quality.

Results

Among 100 individuals that we selected, the number of imputed polymorphic sequence variants is 6726. The accuracy assessments with IQS and correlation were conducted within the 100 individuals and polymorphic variants. However, different imputation strategies yield different numbers of polymorphic variants with meaningful IQS or correlation (Table 4). This is because both imputed and true genotypes must be polymorphic to obtain a meaningful IQS or correlation, and the number of polymorphic imputed genotypes varied by method.

Generally, our proposed 2-step imputation method performed better than only using population-based imputation with Impute2 or only using family-based imputation with Merlin for the variants with a MAF larger than 0.1 and less than 0.4 (Figs. 1a and b). With decreasing the cutoff of MAF for selected imputed variants from population-based imputation using Impute2, the imputation of our method outperformed for most of rare variants with minor MAF between 0.01 and 0.05 (Figs. 1c and d). For common variants, the different cutoffs of the MAFs give similar imputations.

Discussion

Our combined method with a MAF cutoff of 0.01 performed better than either Merlin or Impute2 alone for variants with MAFs between 0.01 and 0.4, and our combined method with a MAF cutoff of 0.05 performed better than either Merlin or Impute2 alone for variants with MAFs >0.05. Because the performance suffers below our MAF cutoffs, this suggests that we should not filter Impute2 results by MAF at all, but filter only by best-guess genotype probability.

One potential limitation of this study is that families with more sequence data were more likely to be selected in our set of 100 individuals. We would expect higher imputation accuracy in these families, as there were more individuals included in the reference panels for imputation. More work needs to be done to determine exactly how much the number and relationships of sequenced family members available affect imputation quality. This was beyond the scope of our project, but may be useful in helping investigators choose which family members to sequence.

It is unclear from these results whether the sequential nature of the imputation increases accuracy. In the future, we should compare our method to a method combining independent results from Merlin and Impute2, both based on best-guess genotype probability and Saad

et al's proposed vote strategy [3]. Furthermore, future studies should be done on a larger region and larger sample size, and potentially include different probability thresholds for the Impute2 results.

Conclusions

Our 2-stage method with a MAF inclusion cutoff of 0.01 for Impute2 results achieved better IQSs than either Impute2 or Merlin alone, and similar correlation values, for variants with MAFs between 0.01 and 0.4. This method could be further improved by including all Impute2 imputed genotypes above a certain quality threshold regardless of MAF. Other probability thresholds should be tested, and this 2-stage method should be compared to results using Merlin and Impute2 independently to examine whether the sequential nature of the procedure increases accuracy above and beyond the increase obtained by combining population- and family-based methods.

Acknowledgements

The GAW19 whole genome sequence data were provided by the T2D-GENES Consortium, which is supported by NIH grants U01 DK085524, U01 DK085584, U01 DK085501, U01 DK085526, and U01 DK085545. The other genetic and phenotypic data for GAW18 were provided by the San Antonio Family Heart Study and San Antonio Family Diabetes/Gallbladder Study, which are supported by NIH grants P01 HL045222, R01 DK047482, and R01 DK053889. The Genetic Analysis Workshop is supported by NIH grant R01 GM031575. SL was supported by the National Institute of General Medicine grant T32 GM074905.

Declarations

This article has been published as part of *BMC Proceedings* Volume 10 Supplement 7, 2016: Genetic Analysis Workshop 19: Sequence, Blood Pressure and Expression Data. Summary articles. The full contents of the supplement are available online at <http://bmcpoc.biomedcentral.com/articles/supplements/volume-10-supplement-7>. Publication of the proceedings of Genetic Analysis Workshop 19 was supported by National Institutes of Health grant R01 GM031575.

Authors' contributions

All authors contributed to the design of the overall study. SL and XD conducted all analyses and drafted the manuscript. YZ, LAC, KLL, and CTL provided advice and critically revised the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare they have no competing interests.

Published: 18 October 2016

References

1. Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet.* 2010;11(7):499–511.
2. Li L, Li Y, Browning SR, Browning BL, Slater AJ, Kong X, et al. Performance of genotype imputation for rare variants identified in exons and flanking regions of genes. *PLoS Genet.* 2011;6(9):e24945.
3. Saad M, Wijsman E. Combining family- and population-based imputation data for association analysis of rare and common variants in large pedigrees. *Genet Epidemiol.* 2014;38(7):579–90.
4. Kreiner-Møller E, Medina-Gomez C, Uitterlinden A, Rivadeneira F, Estrada K. Improving accuracy of rare variant imputation with a two-step imputation approach. *Eur J Hum Genet.* 2015;23(3):395–400.
5. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature.* 2012;491(7422):56.

6. O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, et al. A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet.* 2014;10(4):e1004234.
7. 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(6):e1000529.
8. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet.* 2002;30(1):97–101.
9. Burdick JT, Chen WM, Abecasis GR, Cheung VG. In silico method for inferring genotypes in pedigrees. *Nat Genet.* 2006;38(9):1002–4.
10. Abecasis GR, Wigginton JE. Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. *Am J Hum Genet.* 2005;77(5):754–67.
11. Lin P, Hartz SM, Zhang Z, Saccone SF, Wang J, Tischfield JA, et al. A new statistic to evaluate imputation reliability. *PLoS One.* 2010;5(3):e9697.
12. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas.* 1960;20(1):37–46.
13. Asimit J, Zeggini E. Rare variant association analysis methods for complex traits. *Annu Rev Genet.* 2010;44:293–308.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

