RESEARCH ARTICLE

When Does Choice of Accuracy Measure Alter Imputation Accuracy Assessments?

Shelina Ramnarine[1], Juan Zhang[2], Li-Shiun Chen[3], Robert Culverhouse[4], Weimin Duan[1], Dana B. Hancock[5], Sarah M. Hartz[3], Eric O. Johnson[6], Emily Olfson[3], Tae-Hwi Schwantes-An[7], Nancy L. Saccone[1]*

1 Department of Genetics, Washington University, St. Louis, Missouri, United States of America, **2** Chinese Academy of Sciences, Key Laboratory of Brain Function and Disease, School of Life Sciences, University of Science and Technology of China, Hefei, Anhui, China, **3** Department of Psychiatry, Washington University, St. Louis, Missouri, United States of America, **4** Department of Medicine, Washington University, St. Louis, Missouri, United States of America, **5** Behavioral and Urban Health Program, Behavioral Health and Criminal Justice Division, Research Triangle Institute (RTI) International, Research Triangle Park, North Carolina, United States of America, **6** Fellow Program and Behavioral Health and Criminal Justice Division, RTI International, Research Triangle Park, North Carolina, United States of America, **7** Genometrics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland, United States of America

* nlims@genetics.wustl.edu

Abstract

Imputation, the process of inferring genotypes for untyped variants, is used to identify and refine genetic association findings. Inaccuracies in imputed data can distort the observed association between variants and a disease. Many statistics are used to assess accuracy; some compare imputed to genotyped data and others are calculated without reference to true genotypes. Prior work has shown that the Imputation Quality Score (IQS), which is based on Cohen's kappa statistic and compares imputed genotype probabilities to true genotypes, appropriately adjusts for chance agreement; however, it is not commonly used. To identify differences in accuracy assessment, we compared IQS with concordance rate, squared correlation, and accuracy measures built into imputation programs. Genotypes from the 1000 Genomes reference populations (AFR N = 246 and EUR N = 379) were masked to match the typed single nucleotide polymorphism (SNP) coverage of several SNP arrays and were imputed with BEAGLE 3.3.2 and IMPUTE2 in regions associated with smoking behaviors. Additional masking and imputation was conducted for sequenced subjects from the Collaborative Genetic Study of Nicotine Dependence and the Genetic Study of Nicotine Dependence in African Americans (N = 1,481 African Americans and N = 1,480 European Americans). Our results offer further evidence that concordance rate inflates accuracy estimates, particularly for rare and low frequency variants. For common variants, squared correlation, BEAGLE R^2, IMPUTE2 INFO, and IQS produce similar assessments of imputation accuracy. However, for rare and low frequency variants, compared to IQS, the other statistics tend to be more liberal in their assessment of accuracy. IQS is important to consider when evaluating imputation accuracy, particularly for rare and low frequency variants.

Citation: Ramnarine S, Zhang J, Chen L-S, Culverhouse R, Duan W, Hancock DB, et al. (2015) When Does Choice of Accuracy Measure Alter Imputation Accuracy Assessments? PLoS ONE 10(10): e0137601. doi:10.1371/journal.pone.0137601

Editor: Chuhsing Kate Hsiao, National Taiwan University, TAIWAN

Received: March 16, 2015

Accepted: August 19, 2015

Published: October 12, 2015

Data Availability Statement: Data used are from the 1000 Genomes Project (http://www.1000genomes.org/data). The reference panels used were obtained from the University of Michigan Center for Statistical Genetics (http://csg.sph.umich.edu/abecasis/MACH/download/). The nicotine dependence studies (Collaborative Genetic Study of Nicotine Dependence and the Genetic Study of Nicotine Dependence in African Americans) are available from NCBI dbGAP (accession number phs000813).

Funding: NLS, WD, JZ, SR, and RC were supported by R01DA026911 from the National Institute on Drug Abuse (NIDA). RC was also supported by

R21DA033827 from NIDA. NLS, WD, and RC were also supported by P01DA035825 from NIDA. DBH and EOJ were supported by R01DA035825 from NIDA. EOJ was also supported by R01DA025888 from NIDA. EO was supported by T32GM07200, UL1TR000448, F30AA023685, and TL1TR000449 from the National Institutes of Health. This work was also supported by P01CA089392 from NCI and HHSN268201100011I from NIH. LC was supported by K08DA030398 from NIH. SMH was supported by K08DA032680 from NIDA. TSA was supported by the Division of Intramural Research at National Human Genome Research Institute, National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors of this manuscript have the following competing interests: The spouse of NLS is listed as an inventor on Issued U.S. Patent 8,080,371,"Markers for Addiction" covering the use of certain SNPs in determining the diagnosis, prognosis, and treatment of addiction. This does not alter the authors' adherence to PLOS ONE policies on sharing data and materials. The other authors declare no conflicts of interest.

Introduction

In genomic analyses high-quality data are crucial to accurate statistical inferences. Data accuracy can typically be assessed by different methods and measures.

Genetic imputation provides an informative scenario for examining how the use of different accuracy measures can influence the assessment of accuracy. Genotype imputation is a valuable tool in association studies and meta-analyses. This process infers "in silico" genotypes for untyped variants in a study sample by matching genotyped variants in the study to corresponding haplotypes in a comprehensively genotyped reference panel [1–8]. Therefore, imputation accuracy is influenced by haplotype frequencies in the reference panel [9–10] and the typed single nucleotide polymorphism (SNP) coverage of the study sample [11–12]. Once untyped variants are inferred, statistics that measure imputation accuracy are calculated to identify poorly imputed SNPs.

Imputation accuracy statistics can be classified into two types: (1) statistics that compare imputed to genotyped data and (2) statistics produced without reference to true genotypes. Concordance rate, squared correlation, and Imputation Quality Score (IQS) [13] are examples of the first type. Because imputed SNPs usually do not have genotyped data for comparison, statistics of the second type are usually provided by imputation programs and are commonly relied upon in practice. However, a direct comparison of imputed and genotyped data can be made possible by masking a percentage of variants that were genotyped in the study sample [9, 14–15].

Lin et al (2010) introduced IQS, which is based on Cohen's kappa statistic for agreement [13]. Because of chance agreement, concordance rate, i.e. the proportion of agreement, can lead to incorrect assessments of accuracy for rare and low frequency variants. IQS adjusts for chance agreement [13]. Furthermore, Lin et al. (2010) used simulated data to show that requiring an IQS threshold > 0.9 removed all false positive association signals, while concordance rate > 0.99 still resulted in many false positives. Despite this evidence, IQS is not widely used in accuracy assessment.

This work builds upon previous studies by comparing IQS with commonly used accuracy measures—concordance rate, squared correlation, and built-in accuracy statistics—with the goal of identifying situations in which the choice of accuracy measure leads to differing assessments of accuracy. We compared imputed and genotyped data via masking, and used African-ancestry and European-ancestry populations to evaluate imputation accuracy in genomic regions associated with nicotine dependence and smoking behavior, some of which have also been implicated in lung cancer and chronic obstructive pulmonary disease (COPD).

Methods

We examined differences and similarities in accuracy assessment as measured by IQS, squared correlation, concordance rate and built-in accuracy statistics using: (1) 1000 Genomes as the sample and the reference, and (2) data from nicotine dependence studies as the sample and 1000 Genomes as the reference. Below we describe both approaches, beginning with analyses involving 1000 Genomes as the sample and the reference.

Masking and Imputation using 1000 Genomes Data

Because IQS adjusts for chance agreement [13], we used IQS as a benchmark for accuracy estimation. Calculating IQS, concordance rate, and squared correlation requires genotyped data for comparison with imputed data. We created a study sample for imputation by masking genotypes in the reference panel to mimic the typed SNP coverage of commercially available SNP arrays (Affymetrix—Affy 500 and Affy 6 as well as Illumina—Duo, Omni, and Quad matched by genomic position using Build 37.3/hg19). We used 1000 Genomes African (AFR) and European (EUR) continental reference panels with 246 and 379 individuals respectively

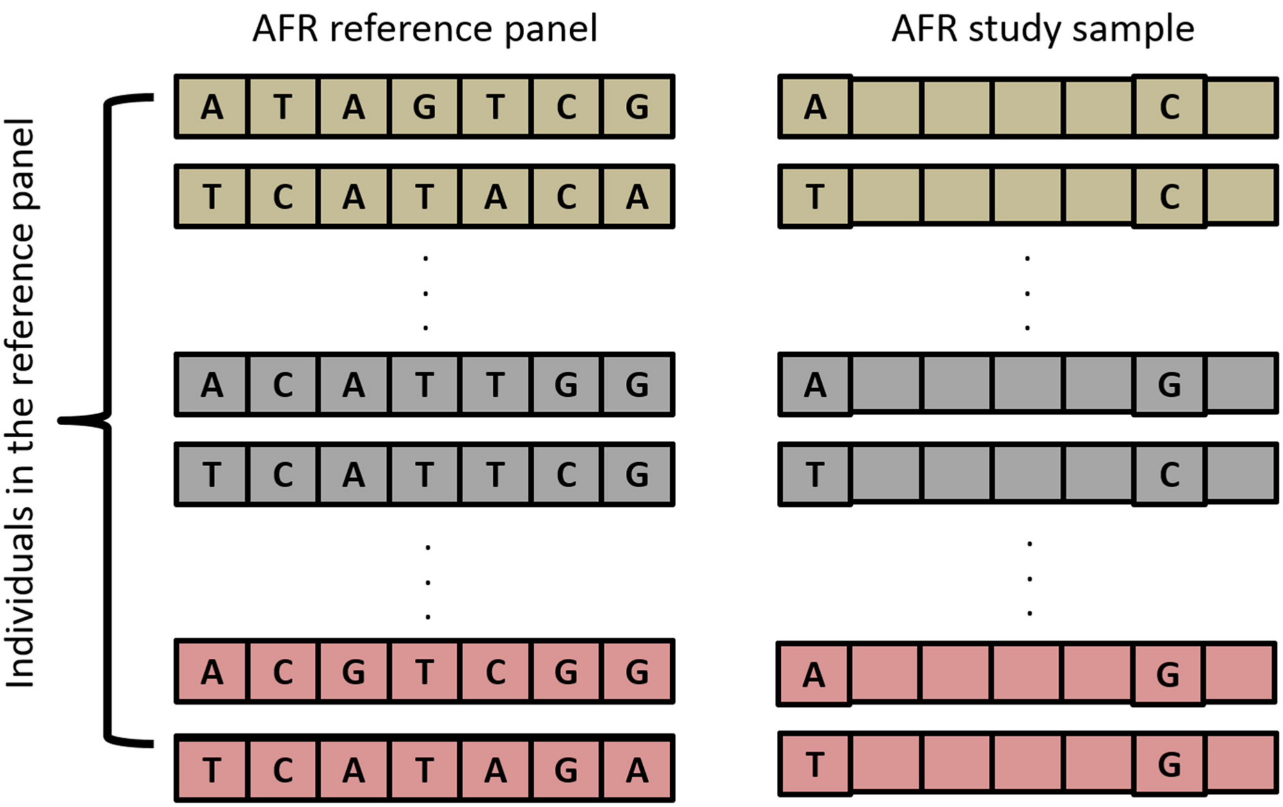

Fig 1. General process for creating the study sample for imputation. The reference panel was masked to mimic a commercial SNP array, resulting in a study sample which contains the same individuals as the reference panel.

doi:10.1371/journal.pone.0137601.g001

(S1 Table) [16]. All data analyzed here are de-identified, publicly available data from the 1000 Genomes (1000G) project, which provides these data as a resource for the scientific community. Participants provided informed consent to the 1000G Project for broad use and broad data release in databases [16–17]. We also have Washington University Human Research Protection Office approval for analyses of de-identified data.

The process of creating the study sample is described in Fig 1 and the numbers of typed variants are presented in S2 Table. Fig 1 illustrates several key characteristics of our masking approach. The reference panel individuals were the same as the study sample individuals. Our approach is expected to give an upper bound on accuracy because of the ideal match between the reference panel and study sample; the "correct" haplotype for each individual being imputed is present in the reference. Using population-specific reference panels (AFR and EUR) rather than a cosmopolitan reference panel maximizes the matching between the reference panel and study sample. Also, this design allowed us to compare accuracy estimates for variants not found on a SNP array. This sample data set was then imputed and the results were used to calculate accuracy statistics.

Imputation Programs

BEAGLE (version 3.3.2) [2, 8] and IMPUTE2 [1, 4–5] were used to obtain imputed genotype probabilities. We obtained the BEAGLE R^2 and IMPUTE2 INFO accuracy measures for each SNP; neither of these makes use of true genotypes. The BEAGLE R^2 and IMPUTE2 INFO accuracy measures are well established [3, 15]. BEAGLE R^2 approximates the squared correlation between the most likely genotype and the true unobserved allele dosage [2, 8]. IMPUTE2 INFO considers allele frequency as well as the observed and expected allele dosage [15]. We include their formulas for completeness, in Eqs 1 and 2, Here g_n represents the observed dosage, e_n represents the expected allele dosage, and $\hat{\theta}$ represents the sample allele frequency for sample n at a particular SNP, where n ranges from 1 to N, the total number of individuals and $0 < \hat{\theta} < 1$. Additionally, z_n represents the genotype with the highest posterior probability from imputation, i.e. 0, 1, or 2 corresponding to the number of copies of the coded allele. Finally, $f_n = p_{n1} + 4p_{n2}$ where p_{nk} represents the imputed probability of the genotypic class k (0, 1, and 2) corresponding to the nth sample.

$$BEAGLE\ \ R^2 = \frac{\left[\sum_{n=1}^{N} g_n e_n - \ (1/N)\left(\sum_{n=1}^{N} g_n \sum_{n=1}^{N} e_n\right)\right]^2}{\left[\sum_{n=1}^{N} f_n - (1/N)\left(\sum_{n=1}^{N} e_n\right)^2\right]\left[\sum_{n=1}^{N} z_n - \ (1/N)\left(\sum_{n=1}^{N} z_n\right)^2\right]} \quad (1)$$

$$IMPUTE2\ \ INFO = \ 1 - \frac{\sum_{n=1}^{N}(f_n - \ e_n^2)}{2N\hat{\theta}(1-\hat{\theta})} \quad (2)$$

Imputed probabilities produced by BEAGLE and the corresponding accuracy statistics showed variability, so we focus on these results. Analyses using IMPUTE2 were less informative in this matched sample-reference setting; this program appears to identify the matching individual in the reference and assign imputed data accordingly. The result was highly accurate imputation in this special context. Since we aim to compare concordance rate, squared correlation, and IQS in efforts to identify scenarios where these statistics produce similar or divergent conclusions regarding accuracy estimation, the variation produced by using BEAGLE for imputation allows us to address our question of interest.

Statistics that Compare Genotyped and Imputed Data

The imputed genotype probabilities produced by BEAGLE and IMPUTE2 were used to calculate concordance rate, squared correlation and IQS. These imputed genotype probabilities, one for each genotype class (e.g. AA, AB, or BB), are transformed to dosage values by multiplying by 0, 1 or 2 for each genotypic class. IQS is calculated from genotype probabilities while squared correlation uses dosage values. Note that a specific dosage value can correspond to multiple genotypic probabilities, but only one dosage value can result from a specific set of genotypic probabilities. Although the most likely (best guess) genotype for each variant can be used to calculate these statistics, it is not recommended because the discrete classification of each individual's genotype does not consider the probabilistic nature of imputation [18].

The incorporation of the genotypic classes into the IQS calculation is represented in Table 1, where each cell is the sum of the genotype probabilities for each genotyped and imputed genotypic class combination. The IQS calculation is demonstrated in Eq 3. IQS considers both the observed proportion of agreement (concordance rate or P_o shown in Eq 4) as well as chance agreement (P_c in Eq 5). Concordance rate (P_o) is the sum of probabilities for each matching genotypic class divided by the total sum of all genotype probabilities. Chance agreement is evaluated as the sum of the products of the marginal frequencies. An IQS score of one indicates that the data matched perfectly, while a negative IQS score indicates that the SNP was imputed worse than expected by chance [13]. Mathematically, the value of IQS will always be less than or equal to the value of concordance rate: $P_oP_c \leq P_c$, so $P_o – P_c \leq P_o - P_oP_c$, hence $(P_o-P_c)/(1-P_c) \leq$

Table 1. Calculating concordance (P_0) and IQS from imputed genotype probabilities and actual genotypes. The table was created by summing over probabilities for all N individuals (n = 1 to N) in each cell with p_{ij_n} representing the probability that the nth individual has the imputed genotype i and actual genotype j, where 1 corresponds to AA, 2 corresponds to AB, and 3 corresponds to BB. N_1 = number of individuals with AA actual genotype, N_2 = number of individuals with AB actual genotype, N_3 = number of individuals with BB actual genotype, and N = number of total individuals.

			Actual		
		AA	AB	BB	Total
	AA	$\sum_{n=1}^{N} p_{11\_n}$	$\sum_{n=1}^{N} p_{12\_n}$	$\sum_{n=1}^{N} p_{13\_n}$	$\sum_{j=1}^{3}\sum_{n=1}^{N} p_{1j\_n}$
	AB	$\sum_{n=1}^{N} p_{21\_n}$	$\sum_{n=1}^{N} p_{22\_n}$	$\sum_{n=1}^{N} p_{23\_n}$	$\sum_{j=1}^{3}\sum_{n=1}^{N} p_{2j\_n}$
Imputed	BB	$\sum_{n=1}^{N} p_{31\_n}$	$\sum_{n=1}^{N} p_{32\_n}$	$\sum_{n=1}^{N} p_{33\_n}$	$\sum_{j=1}^{3}\sum_{n=1}^{N} p_{3j\_n}$
	Total	$\sum_{i=1}^{3}\sum_{n=1}^{N} p_{i1\_n} = N1$	$\sum_{i=1}^{3}\sum_{n=1}^{N} p_{i2\_n} = N2$	$\sum_{i=1}^{3}\sum_{n=1}^{N} p_{i3\_n} = N3$	N

doi:10.1371/journal.pone.0137601.t001

(P_o-P_oP_c)/(1-P_c), which says that IQS ≤ P_o. Some statistics can be confounded with Hardy-Weinberg equilibrium (HWE) if they assume HWE to calculate "expected" genotype counts [19]. IQS avoids this concern since it uses imputed and experimentally determined genotypes.

$$\mathrm{IQS} \quad = \frac{Po - Pc}{1 - Pc} \tag{3}$$

$$Po = \frac{\sum_{n=1}^{N} \mathrm{p}_{11_\mathrm{n}} + \sum_{n=1}^{N} \mathrm{p}_{22_\mathrm{n}} + \quad \sum_{n=1}^{N} \mathrm{p}_{33_\mathrm{n}}}{\mathrm{N}} \tag{4}$$

$$Pc = \frac{\mathrm{N1} * \sum_{j=1}^{3} \sum_{n=1}^{N} \mathrm{p}_{1j\_\mathrm{n}} + \quad \mathrm{N2} * \sum_{j=1}^{3} \sum_{n=1}^{N} \mathrm{p}_{2j\_\mathrm{n}} + \mathrm{N3} * \sum_{j=1}^{3} \sum_{n=1}^{N} \mathrm{p}_{3j\_\mathrm{n}}}{\mathrm{N}^2} \tag{5}$$

Squared correlation is the square of the Pearson correlation coefficient between the imputed and genotyped dosage for each SNP. This is calculated using Eqs 6–11 where x_i and y_j are the imputed and genotyped dosage values for the nth sample respectively. It represents the proportion of the variability in the imputed data that can be explained by the least squared regression model.

$$\mathrm{R}^2 \quad = 1 - \frac{SSE}{SS_{yy}} \tag{6}$$

$$\mathrm{SS}_{\mathrm{yy}} = \sum_{n=1}^{N} (y_i - \bar{y})^2 \tag{7}$$

$$\mathrm{SSE} \quad = SS_{yy} - \widehat{\beta}_n (SS_{xy}) \tag{8}$$

$$\mathrm{SS}_{\mathrm{xy}} = \sum_{n=1}^{N} (y_i - \bar{y})(x_j - \bar{x}) \tag{9}$$

$$\widehat{\beta}_{\mathrm{n}} \quad = \frac{SS_{xy}}{SS_{xx}} \tag{10}$$

$$\mathrm{SS}_{\mathrm{xx}} = \sum_{n=1}^{N} (x_j - \bar{x})^2 \tag{11}$$

Evaluating Accuracy across MAF and LD

Imputation accuracy is influenced by a variant's minor allele frequency (MAF) and linkage disequilibrium (LD) with genotyped variants (measured by pairwise squared correlation r^2). We examined imputation accuracy in relation to these properties. The MAFs used here were based on the allele frequencies found in the genotyped data. We will use the terminology "rare" to denote variants with MAF $\leq$ 1%; and "low frequency" to refer to variants with 1% $<$ MAF $\leq$ 5%. For each imputed SNP, the genotyped SNP in the region with the highest LD was used to define the maximum r^2_{LD} with a genotyped SNP (denoted by max r^2_{LD}). PLINK was used to generate the LD values [20]. Bins for maximum r^2_{LD} and MAF were defined in 0.01 increments [13]. For each bin, the mean and one standard deviation of the values produced by each accuracy statistic were calculated.

Examining Regions Associated with Nicotine Dependence

We examined the imputation accuracy of two genomic regions known to be associated with nicotine dependence and smoking behavior. These regions were the nicotinic receptor subunit gene clusters on chromosome 15 (*CHRNA5-CHRNA3-CHRNB4*) and chromosome 8 (*CHRNB3-CHRNA6*) [21–26]. These signals were identified through genome-wide association studies (GWAS) and meta-analyses for smoking behavior, with the chromosome 15 region being the most significantly associated. We imputed 3Mb on each chromosome: 2Mb regions used for analysis plus two 500Kb flanking buffer regions according to Build 37.3/hg19. We focused our analyses on polymorphic variants with dbSNP identifiers in each 2MB region.

Masking and Imputation in a Real Data Application using a Nicotine Dependence Sample

A comparison of accuracy statistics was also conducted using nicotine dependence data as the study samples (N = 1,481 African Americans and N = 1,480 European Americans who were sequenced) and 1000 Genomes as the reference. The study sample was masked and imputed separately by race. This analysis provided a more conventional imputation scenario for comparison with the patterns found in the 1000 Genomes analyses.

The sequenced subjects in this applied analysis were from the Collaborative Genetic Study of Nicotine Dependence (COGEND) and the Genetic Study of Nicotine Dependence in African Americans (AAND). These studies are cross-sectional and contain extensive smoking behavior phenotypes in African Americans and European Americans [21]. These individuals were between the ages of 25–44 years old and were assessed for dependence as measured by the Fagerstrom Test for Nicotine Dependence (FTND) and cigarettes-per-day (CPD) [27]. The study protocol was approved by the appropriate Institutional Review Boards and written informed consent was obtained from all subjects.

Center for Inherited Disease Research (CIDR) performed next-generation targeted sequencing on genomic regions previously associated with smoking behaviors, using COGEND and AAND DNA samples derived from blood. Genotypic data that passed initial quality control at CIDR were released to the Quality Assurance/Quality Control analysis team at the University of Washington Genetics Coordinating Center. These data had mean on-target coverage of 180X with more than 96% of on-target bases containing a depth greater than 20X. A total of 1,481 African Americans and 1,480 European Americans were used in the analysis.

These sequencing data were masked to match the typed SNP coverage of the Omni 2.5 SNP array in a 500kb region on chromosome 15. The cosmopolitan reference panel, composed of individuals from a variety of ancestries, was used for imputation since it has been shown to produce the best accuracy estimates [9]. The imputation was performed using BEAGLE and IMPUTE2 to evaluate whether observed trends in accuracy were consistent across imputation programs. The imputed probabilities were compared to the masked sequencing data and accuracy statistics were calculated. We focused our analyses on polymorphic variants.

Results

We compared IQS with squared correlation, concordance rate, and BEAGLE R^2 to examine changes in accuracy assessment using 1000 Genomes as the study sample in Figs 2–5. IQS is our benchmark because it adjusts for chance agreement, in contrast to concordance rate which inflates assessments of accuracy [13]. We focus here on the results for the AFR reference population using Omni 2.5M typed coverage on chromosome 15 (13,442 imputed SNPs). We emphasize Omni 2.5 because it has the greatest genotype SNP coverage in the region (S2 Table).

Results for 1000 Genomes Imputation with Matching Reference

Results produced using BEAGLE and the AFR reference population are shown. Results for different chromosomal regions and populations were similar and are shown in S6–S8 Figs.

To help interpret results that are displayed by MAF and max r^2_{LD} bin, S1 Fig. shows the number of imputed variants in each MAF bin in panel A and max r^2_{LD} bin in panel B. This figure indicates that most of the imputed variants were rare and low frequency variants. There were 6,480 (48.21%) rare and low frequency rsID SNPs in the AFR population. The bins ranged in size from 7 variants ($0.49 \geq \text{MAF} < 0.50$) to 2,371 variants ($0.01 \geq \text{MAF} < 0.02$).

Concordance Rate and BEAGLE R^2 Inflate Assessments of Accuracy for Rare Variants

Results show that the choice of statistic is important when examining the imputation accuracy of rare and low frequency variants. Fig 2 displays the mean accuracy and one standard deviation in each MAF bin, after imputing from Omni 2.5M coverage. IQS (Panel A) and squared correlation (Panel B) produced similar means and standard deviations in each bin, though this does not necessarily represent similarity of values for particular SNPs. For rare and low frequency variants, both concordance rate (Panel C) and BEAGLE R^2 (Panel D) produce inflated assessments of accuracy. The higher concordance rate and BEAGLE R^2 values could mislead a researcher into assuming that these variants were imputed well, and that accuracy is best measured using concordance rate and BEAGLE R^2. IQS and squared correlation also show low accuracy for rare variants using other SNP array coverages (S2 Fig).

A MAF bin can have a wide range in accuracy values. Fig 2 shows variability within MAF bins across all MAF values. Standard deviations for IQS, squared correlation and BEAGLE R^2 can be sizeable for both rare and common variants (panels A, B and D); concordance rate does not reflect this as it classifies most variants as well imputed (panel C).

Rare and Low Frequency Variants can be Well Tagged but Poorly Imputed

We examined max r^2_{LD}, the maximum LD r^2 between imputed and genotyped SNPs, to understand the relationship between typed SNP coverage and imputation accuracy as measured by these accuracy statistics. Fig 3 displays the mean accuracy and one standard deviation in each max r^2_{LD} bin, after imputing from Omni 2.5M coverage, additional arrays are in S3 Fig. Mean accuracy tends to increase with increasing max r^2_{LD}, as expected. For low to moderate max r^2_{LD}, we observed substantial variability in IQS as well as squared correlation and BEAGLE R^2

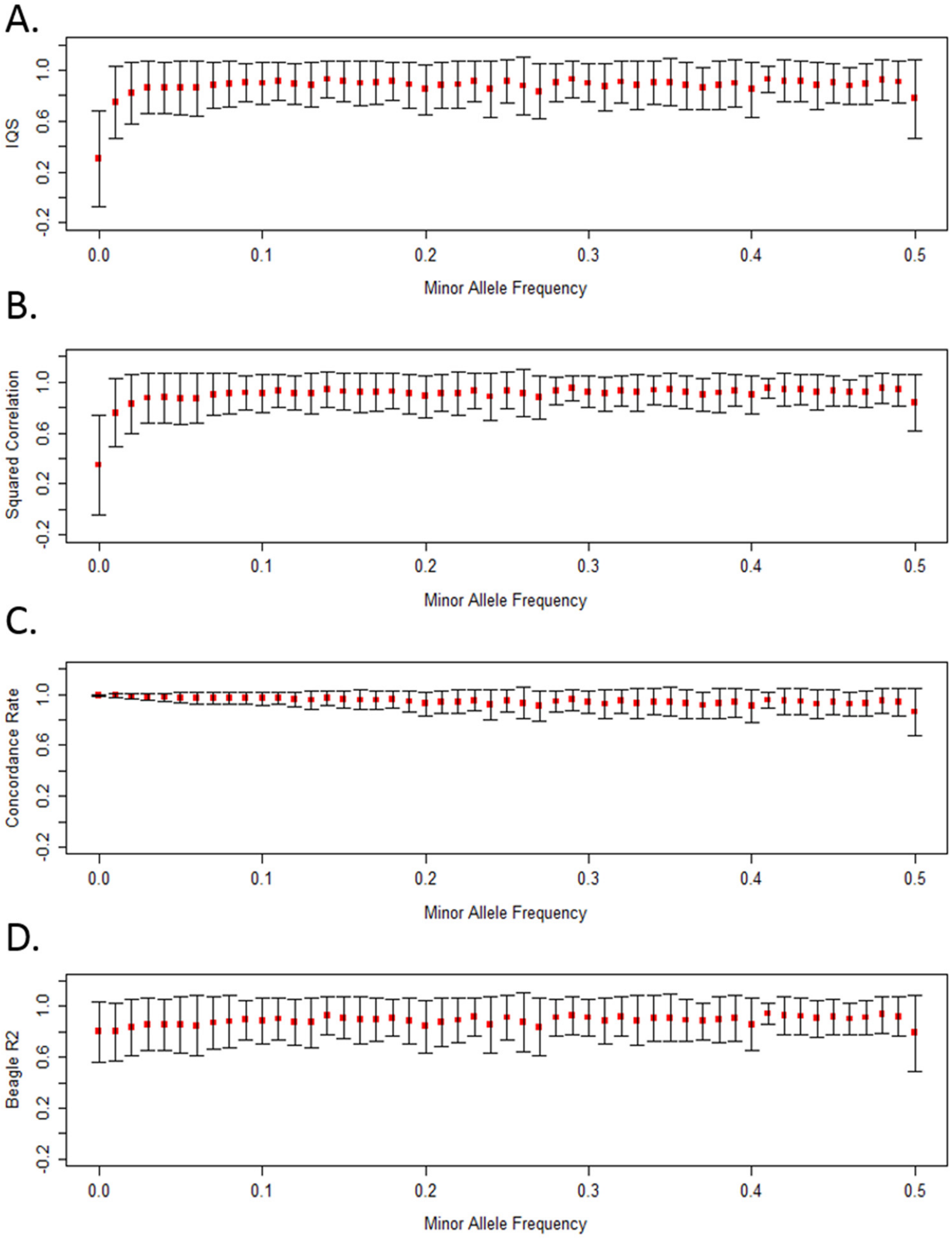

Fig 2. IQS, squared correlation, concordance rate, and BEAGLE R^2 are shown in MAF bins. Mean accuracy of SNPs in each MAF bin (defined by 0.01 increments with N = 13,442 variants total) is denoted by the red dots and the bars indicate one standard deviation (above and below the mean). These results are produced by using the 1000 Genomes AFR reference population as the study sample with Omni 2.5M typed coverage on chromosome 15.

doi:10.1371/journal.pone.0137601.g002

values; however, at high max r^2_{LD}, the variability decreases. IQS and squared correlation show a surprisingly wide standard deviation for variants in the highest max r^2_{LD} bin ($0.99 <$ max $r^2_{LD} \leq 1$) as well as the max r^2_{LD} bin $0.5 <$ max $r^2_{LD} \leq 0.51$. Upon investigation, we found that the variability was due to rare variants: after limiting to SNPs with MAF $> 5\%$, these standard deviations were comparable to those of the other bins, S4 Fig. This pattern suggests that even rare variants that are well tagged (as measured by max r^2_{LD}) can be poorly imputed.

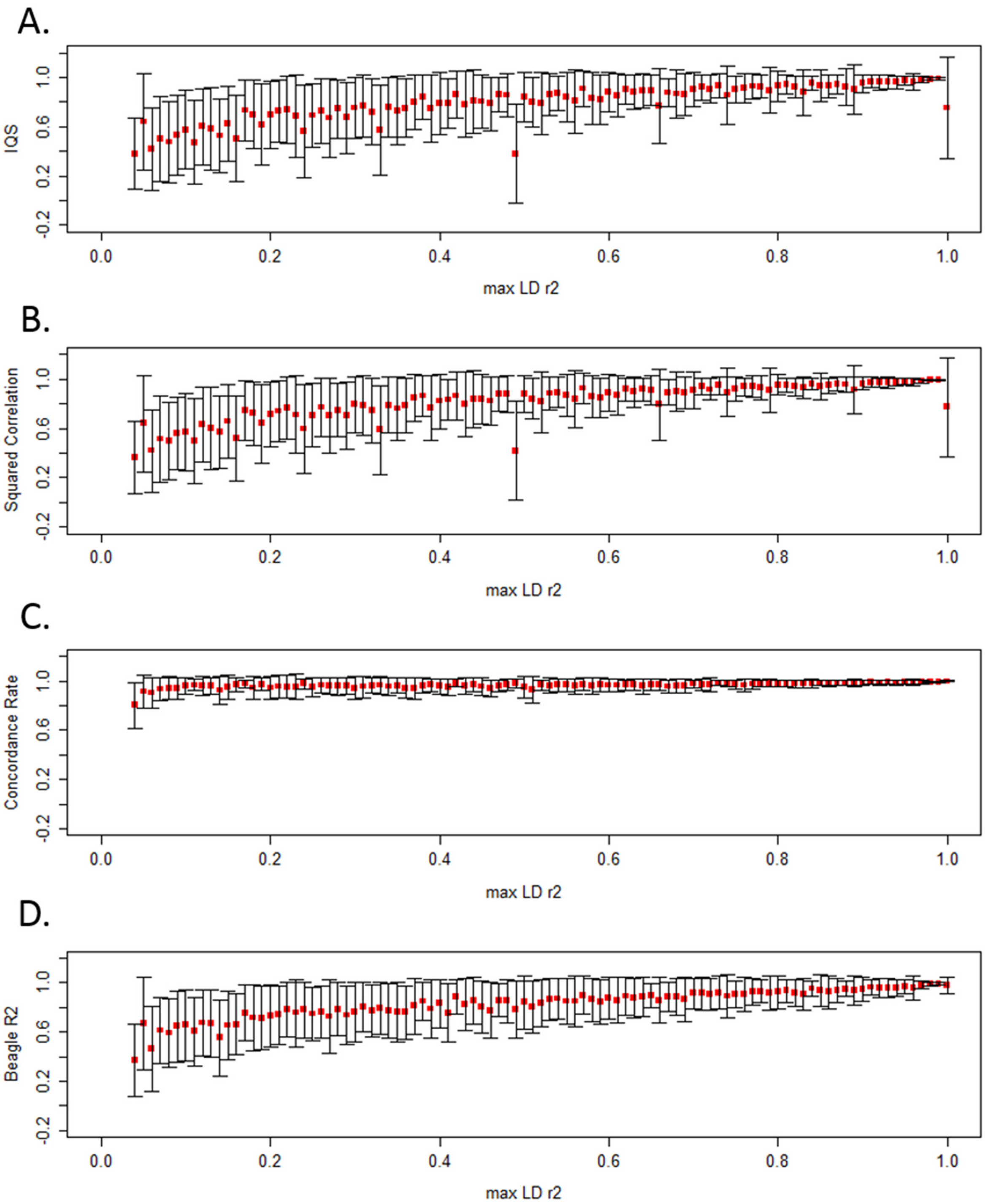

Fig 3. IQS, squared correlation, concordance rate, and BEAGLE R^2 are shown in max r^2_{LD} bins. Mean accuracy of SNPs in each MAF bin (defined by 0.01 increments with N = 13,442 variants total) is denoted by the red dots and the bars indicate one standard deviation (above and below the mean). These results were produced by using the 1000 Genomes AFR reference population as the study sample with Omni 2.5M typed coverage on chromosome 15.

doi:10.1371/journal.pone.0137601.g003

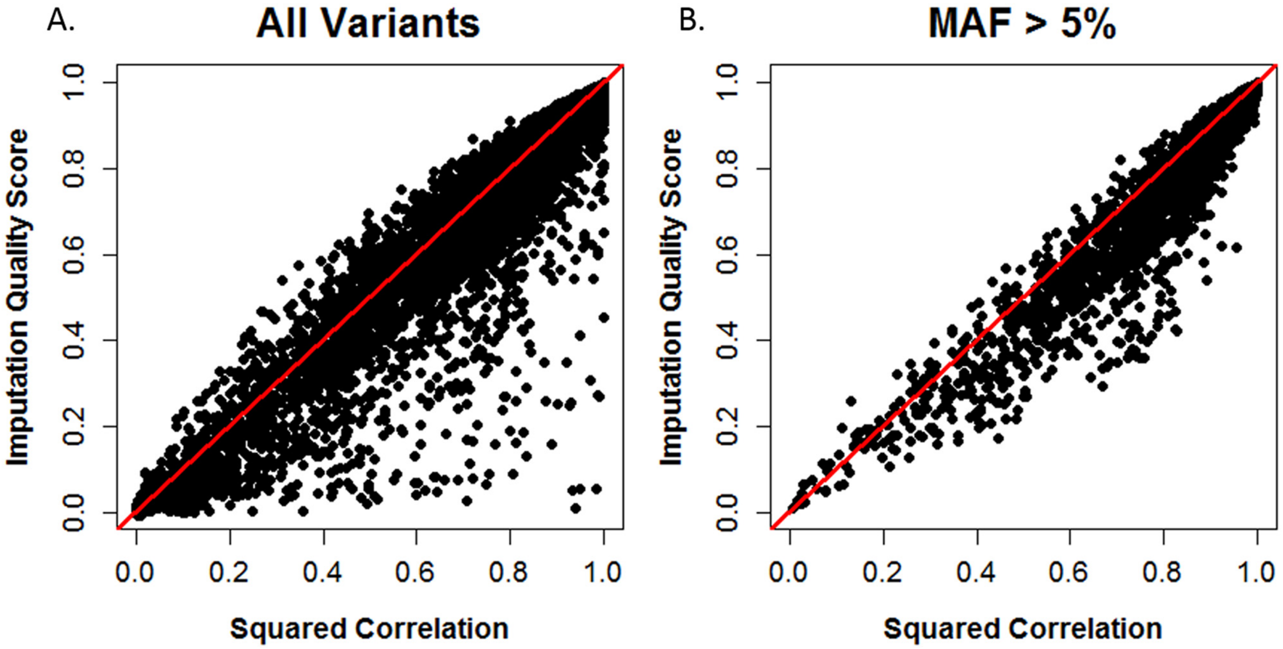

Fig 4. Scatterplots of squared correlation and IQS. Data for all 13,442 variants are displayed in panel A, while the results for variants with MAF>5% (N = 6,480) are found in panel B. The line y = x is denoted in red.

doi:10.1371/journal.pone.0137601.g004

Concordance Classifies Most Variants as Well Imputed

Concordance differs from IQS, squared correlation, and BEAGLE R^2 in that it indiscriminately classifies most variants as well imputed, across MAF (Fig 2) and r^2_{LD} bins (Fig 3). The results in Figs 2 and 3 support prior concerns regarding concordance rate [13] and led us to focus the rest of our evaluation on IQS, squared correlation, and BEAGLE R^2.

For Rare Variants, IQS and Squared Correlation Produce Different Assessments of Accuracy

Although squared correlation and IQS appeared similar overall in their assessment of imputation accuracy when examined using means and standard deviations by bin (Figs 2 and 3), further investigation showed that on an individual SNP level, these statistics produce divergent assessments of accuracy for rare and low frequency variants. We compared accuracy estimates produced by IQS and squared correlation in Fig 4 for each SNP. Panel A shows results for all variants, and panel B displays results for variants with MAF $>$ 5%. A comparison of these panels is useful to identify divergent trends for common variants versus rare and low-frequency variants. For most SNPs, IQS and squared correlation produced similar assessments of accuracy as seen by the many observations on and near the y = x line in panels A and B. This is consistent with the accuracy patterns observed for IQS and squared correlation in Figs 2 and 3. However, discrepancies in accuracy assessment do occur, with squared correlation generally being more liberal in assigning high accuracy compared to IQS. This is indicated by the sparseness of observations above the y = x line in panels A and B. The points below the y = x line indicate SNPs for which squared correlation values were higher than IQS. Panel B shows that widely discrepant values for IQS and squared correlation are attributable to rare and low frequency SNPs: filtering out SNPs with MAF $\leq$ 5% removes the widely discrepant observations.

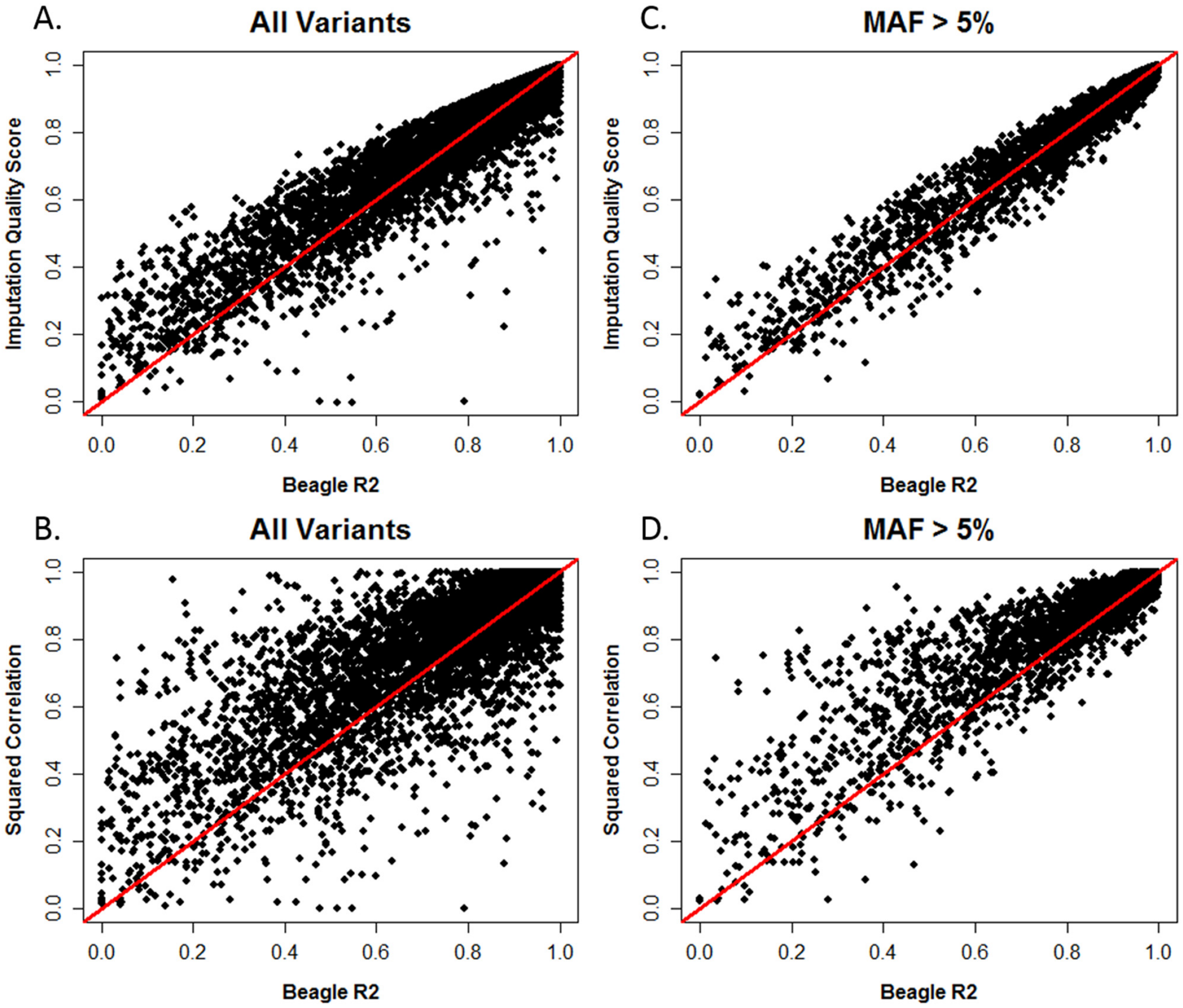

Fig 5. Scatterplots of IQS, squared correlation, and BEAGLE R^2. Panels A and B display all 13,442 variants, and panels C and D display variants with MAF>5% (N = 6,480). The line y = x is denoted in red.

doi:10.1371/journal.pone.0137601.g005

To further examine trends in the discrepancies between these statistics, we subtracted squared correlation from IQS for each variant and displayed this result across all MAF values in S5 Fig. Thus negative differences denote that squared correlation was greater than IQS (i.e. squared correlation more liberal) while positive differences indicate that IQS was greater than squared correlation. Large discrepancies occur over all MAF values with squared correlation tending to be higher than IQS, especially for SNPs with higher MAFs.

For Common Variants, IQS and BEAGLE R^2 Provide Similar Assessments of Accuracy

For common variants, BEAGLE R^2 produces a similar assessment of imputation accuracy as IQS, but BEAGLE R^2 can differ dramatically from squared correlation. In Fig 5, we compared BEAGLE R^2 to IQS (panels A and C) and squared correlation to BEAGLE R^2 (panels B and D).

For many variants, squared correlation and BEAGLE R^2 differ in accuracy assessment as seen by the variants above the y = x line in panel B. Although most of these variants are rare, there are still many common variants for which this trend is true (panel D). Large differences between IQS and BEAGLE R^2 occur mostly when rare variants are examined.

Results are Similar in Different Genomic Regions and Populations

Figs 2–5 displayed results for the AFR reference population and Omni 2.5M typed coverage in the chromosome 15 region. Results similar to those described above were also observed using the AFR reference on chromosome 8 (S6 Fig) as well as using the EUR reference panel for chromosomes 15 and 8 (S7 and S8 Figs respectively). In particular, low IQS values do occur for rare variants that have high squared correlation or high BEAGLE R^2. The number of variants for each imputation subset can be found in S3 Table.

Results are Consistent in Application to Nicotine Dependence Study Sample

Fig 6 shows results produced using African American individuals from the nicotine dependence data as the study sample and a 1000 Genomes cosmopolitan reference panel imputed using BEAGLE. These data show discrepancies in accuracy assessment between statistics. If IQS and squared correlation are compared, squared correlation tends to be similar or higher (i.e. more liberal) than IQS. In the applied scenario, we observed some variants with high IQS and low squared correlation (Fig 6, panel A, upper left quadrant), which was not observed for the upper bound values from the 1000 Genomes analysis (Fig 4, panel A); however, these discrepancies are few, and mostly among rare and low frequency variants (see Fig 6, panel D). When comparing IQS to Beagle R^2, the applied scenario showed IQS to be similar to or less than Beagle R^2 (Fig 6, panel B), which recapitulates patterns seen in 1000 Genomes (Fig 5, panel A).

In European Americans, from the nicotine dependence data, we also observed these same patterns as in African Americans, with squared correlation's more liberal assignment of accuracy as compared to IQS, S9 Fig. These results were also consistent using IMPUTE2 with African American and European American study samples, S10 and S11 Figs respectively. This confirms that these patterns are not limited to specific populations, chromosomes, or imputation programs.

Discussion

Genotype imputation is used to improve the density of genomic coverage and increase power by combining datasets [28], in efforts to identify and refine genetic variants associated with disease. We investigated how assessment of imputation accuracy changes when concordance rate, squared correlation and BEAGLE R^2 are compared to IQS, focusing on two genomic regions associated with smoking behavior.

Results showed that the choice of accuracy statistic matters for rare variants more than for common variants. This is important given that researchers are increasingly interested in imputing rare and low frequency variants [29–31]. While it has been recognized that rare variants are more difficult to impute accurately, our work here goes further by highlighting that choice of accuracy measure has an important role.

For common variants, squared correlation, IMPUTE2, and BEAGLE R^2 produce similar assessments of imputation accuracy as compared to IQS. For rare and low frequency variants, we observed varying assessments of accuracy compared to IQS. Our results also showed that discrepancies between IQS and squared correlation are most likely to occur at rare and low

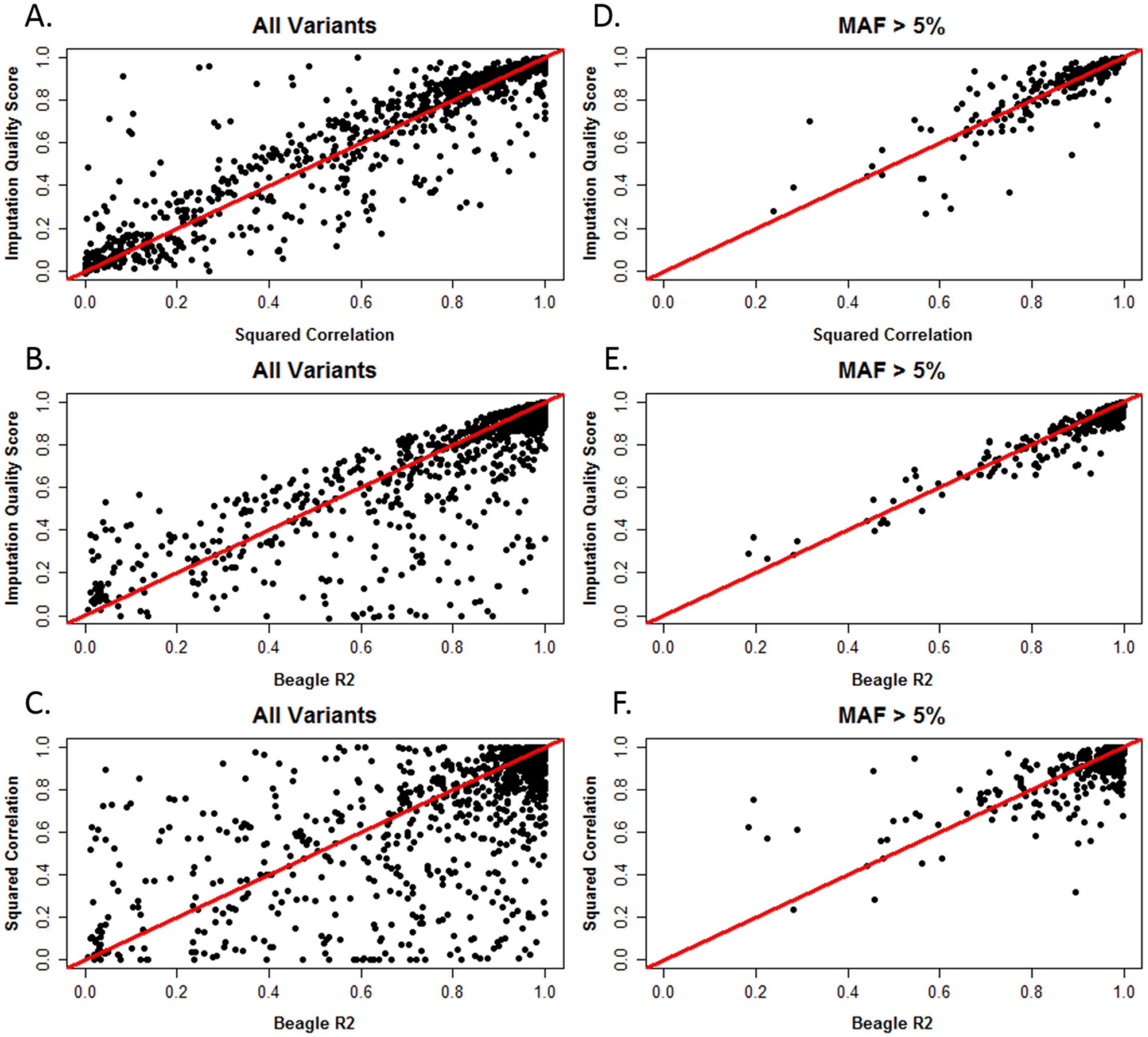

Fig 6. Scatterplots of IQS, squared correlation, and BEAGLE R^2 using the cosmopolitan reference panel and the African American nicotine dependence study sample for chromosome 15. Data for all 1,545 variants are displayed in panel A, B, and C while the results for variants with MAF>5% (N = 631) are found in panel D, E, and F. These results were generated using Omni SNP coverage. The line y = x is denoted in red.

doi:10.1371/journal.pone.0137601.g006

frequency variants, where squared correlation is more liberal in assigning higher accuracy as compared to IQS. An evaluation of nicotine dependence samples also showed discrepancies between IQS and squared correlation. We recommend calculating IQS to confirm imputation accuracy, especially for rare or low frequency variants.

The variability observed within a MAF or max r^2_{LD} bin is a reminder that not all variants that share the same MAF or max r^2_{LD} value can be imputed with the same level of accuracy. This is consistent with the expectation that the inference of untyped variants depends on haplotype block structure and not simply the pairwise relationships between the genotyped and untyped variants. For rare variants, high LD with a genotyped SNP may not guarantee high imputation accuracy. Still, overall, a high max r^2_{LD} usually implies high accuracy, as we observed increasing mean accuracy along with decreasing variability within max r^2_{LD} bins as max r^2_{LD} increases.

We applied this approach to genomic regions associated with our phenotype of interest, smoking behavior using an upper bound scenario and a nicotine dependence sample. Thus, one limitation is that rather than comprehensively examining the genome, we focused only on selected genomic regions. Furthermore we focused on certain populations (European and African ancestry). Nevertheless, different regions (on chromosome 8 and 15), different imputation programs, and different populations showed similar overall patterns, suggesting that our observations are relevant throughout the genome and across multiple populations.

In our masking process using only the 1000 Genomes reference data, the reference panel individuals were the same as the study sample individuals, and our masked SNPs are not limited to a SNP array, making our approach different from the two most common masking processes. One common masking method removes the genotypes for a portion of markers (e.g. 10%) found amongst the typed variants on a study sample SNP array. This method can provide accuracy comparisons only for SNPs on the array. Our approach is able to provide accuracy assessments for SNPs not on the array.

Another commonly used masking method is the "leave-one-out" masking of a comprehensively genotyped reference panel, in which one individual is imputed using the remaining reference panel members. Our study design differed from the leave-one-out method since all individuals in the reference panel and study sample were the same. Our approach was expected to give an upper bound on accuracy because of the ideal match between the reference and study sample; the "correct" genotype for each individual at each variant was present in the reference panel.

Our results provide further evidence that concordance rate inflates accuracy estimates particularly for rare and low frequency variants [13, 32]. These observations highlight a need to account for chance agreement not only when assessing imputation accuracy, but also more broadly in other situations for which concordance is traditionally used to assess accuracy, such as checking genotype agreement across duplicate samples [33–34]. Concordance rate will always produce a value greater than or equal to IQS due to their mathematical relationship (see Methods for proof).

IQS is important to consider, as it is designed to identify variants for which imputation accuracy is better than can be expected by chance; accordingly, other measures were generally more liberal in assigning high accuracy. Our analyses indicate that especially for rare and low frequency variants, IQS may be important to avoid overly liberal assessments of imputation quality. In practice, IQS can be computed by the leave-one-out method. Databases that provide per-SNP "imputability," such as that created by Duan et al. [35], would have increased usefulness if they included IQS values. As imputation methodology continues to develop and reference panels become more comprehensive, we expect that imputation will become increasingly accurate. However, it will be important to take chance agreement into account when assessing this accuracy, and IQS provides a means to do so.

Supporting Information

S1 Fig. Mean numbers of polymorphic variants in each MAF (panel A) and max r^2_{LD} (panel B) bin. These results are for the AFR population on chromosome 15 (13,442 imputed SNPs).
(TIF)

S2 Fig. Average accuracy of all SNPs according to 0.01 incremental MAF bins for each accuracy measure using several typed SNP array coverages. These results were produced by using

the 1000 Genomes AFR reference populations as the study samples for chromosome 15.
(TIF)

S3 Fig. Average accuracy of all SNPs in 0.01 incremental max r^2_{LD} bins for each accuracy measure using several typed SNP array coverages. These results were produced by using the 1000 Genomes AFR reference population as the study sample for chromosome 15.
(TIF)

S4 Fig. Accuracy scores produced by IQS, squared correlation, concordance rate and Beagle R^2 for SNPs with MAF > 5% (N = 6,480 SNPs) in max r^2_{LD} bins. Bins are defined by 0.01 increments. Mean accuracy is denoted by the red dots and the bars indicate one standard deviation (above and below the mean). These results were produced by using 1000 Genomes AFR reference population as the study sample with Omni 2.5M typed coverage on chromosome 15.
(TIF)

S5 Fig. Relationship between squared correlation and IQS by MAF. Squared correlation was subtracted from IQS for variants on chromosome 15 in the 1000 Genomes AFR reference population (N = 13,442 variants) as the study sample. Negative values indicate that the squared correlation score was higher while the positive values indicate that the IQS value was higher. The red line indicates the line y = 0.
(TIF)

S6 Fig. Scatterplots of IQS, squared correlation, and BEAGLE R^2 using the 1000 Genomes AFR reference panel as the study sample for chromosome 8. Data for all 10,937 variants are displayed in panel A, B, and C while the results for variants with MAF>5% (N = 4,533) are found in panel D, E, and F. These results were generated using Omni SNP coverage. The line y = x is denoted in red.
(TIF)

S7 Fig. Scatterplots of IQS, squared correlation, and BEAGLE R^2 using the 1000 Genomes EUR reference panel as the study sample for chromosome 15. Data for all 9,401 variants are displayed in panel A, B, and C while the results for variants with MAF>5% (N = 4,627) are found in panel D, E, and F. These results were produced by using Omni SNP coverage. The line y = x is denoted in red.
(TIF)

S8 Fig. Scatterplots of IQS, squared correlation, and BEAGLE R^2 using the 1000 Genomes EUR reference panel as the study sample for chromosome 8. Data for all 7,401 variants are displayed in panel A, B, and C while the results for variants with MAF>5% (N = 1,903) are found in panel D, E, and F. These results were produced by using Omni SNP coverage. The line y = x is denoted in red.
(TIF)

S9 Fig. Scatterplots of IQS, squared correlation, and BEAGLE R^2 using the cosmopolitan reference panel and the European American nicotine dependence study sample for chromosome 15. Data for all 1,170 variants are displayed in panel A, B, and C while the results for variants with MAF>5% (N = 387) are found in panel D, E, and F. These results were produced by using Omni SNP coverage. The line y = x is denoted in red.
(TIF)

S10 Fig. Scatterplots of IQS, squared correlation, and IMPUTE2 INFO using the cosmopolitan reference panel and the African American nicotine dependence study sample for chromosome 15. Data for all 1,878 variants are displayed in panel A, B, and C while the results for

variants with MAF>5% (N = 475) are found in panel D, E, and F. These results were generated using Omni SNP coverage. The line y = x is denoted in red.
(TIF)

S11 Fig. Scatterplots of IQS, squared correlation, and IMPUTE2 INFO using the cosmopolitan reference panel and the European American nicotine dependence study sample for chromosome 15. Data for all 1,253 variants are displayed in panel A, B, and C while the results for variants with MAF>5% (N = 259) are found in panel D, E, and F. These results were generated using Omni SNP coverage. The line y = x is denoted in red.
(TIF)

S1 Table. Sub-populations in the BEAGLE and IMPUTE2 AFR and EUR reference panels.
(PDF)

S2 Table. Numbers of SNPs in the 1000 Genomes study samples. Study sample variants were those found on each commercially available SNP array for the 2 MB chromosomal regions of interest. Only variants with dbSNP identifiers are listed in the number of variants in the reference panel column.
(PDF)

S3 Table. Polymorphic, imputed SNPs used in the comparison of accuracy measures. These variants were found in the 2 MB chromosomal regions of interest using 1000 Genomes as the study sample and were imputed using Omni 2.5 coverage.
(PDF)

Acknowledgments

The authors thank John Rice and Laura Bierut for helpful comments and discussion. The authors also thank Cynthia Helms for assistance in data analysis.

Author Contributions

Conceived and designed the experiments: SR JZ LC RC WD DBH SMH EOJ EO TS NLS. Performed the experiments: SR JZ WD TS. Analyzed the data: SR JZ. Contributed reagents/materials/analysis tools: SR JZ TS NLS. Wrote the paper: SR JZ LC RC WD DBH SMH EOJ EO TS NLS. Contributed to the interpretation of data: SR JZ LC RC WD DBH SMH EOJ EO TS NLS.

References

1. Howie BN, Donnelly P,Marchini J (2009) A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLoS Genet 5: e1000529. doi: 10.1371/journal.pgen.1000529 PMID: 19543373
2. Browning BL, Browning SR (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 84: 210–223. doi: 10.1016/j.ajhg.2009.01.005 PMID: 19200528
3. Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat Rev Genet 11: 499–511. doi: 10.1038/nrg2796 PMID: 20517342
4. Howie B, Marchini J, Stephens M (2011) Genotype Imputation with Thousands of Genomes. G3: Genes|Genomes|Genetics 1: 457–470. doi: 10.1534/g3.111.001198 PMID: 22384356
5. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet 44: 955–959. doi: 10.1038/ng.2354 PMID: 22820512
6. Liu EY, Li M, Wang W and Li Y (2013) MaCH-Admix: Genotype Imputation for Admixed Populations. Genetic epidemiology 37: 25–37. doi: 10.1002/gepi.21690 PMID: 23074066

7. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR (2010) MaCH: Using Sequence and Genotype Data to Estimate Haplotypes and Unobserved Genotypes. Genetic epidemiology 34: 816–834. doi: 10.1002/gepi.20533 PMID: 21058334

8. Browning SR (2006) Multilocus Association Mapping Using Variable-Length Markov Chains. American Journal of Human Genetics 78: 903–913. PMID: 16685642

9. Hancock DB, Levy JL, Gaddis NC, Bierut LJ, Saccone NL, Page GP, et al. (2012) Assessment of Genotype Imputation Performance Using 1000 Genomes in African American Studies. PLoS One 7: e50610. doi: 10.1371/journal.pone.0050610 PMID: 23226329

10. Sung YJ, Gu CC, Tiwari HK, Arnett DK, Broeckel U, Rao DC (2012) Genotype Imputation for African Americans Using Data From HapMap Phase II Versus 1000 Genomes Projects. Genetic epidemiology 36: 508–516. doi: 10.1002/gepi.21647 PMID: 22644746

11. Johnson EO, Hancock DB, Levy JL, Gaddis NC, Saccone NL, Bierut LJ, et al. (2013) Imputation across genotyping arrays for genome-wide association studies: assessment of bias and a correction strategy. Human genetics 132: 509–522. doi: 10.1007/s00439-013-1266-7 PMID: 23334152

12. Nelson SC, Doheny KF, Pugh EW, Romm JM, Ling H, Laurie CA, et al. (2013) Imputation-Based Genomic Coverage Assessments of Current Human Genotyping Arrays. G3: Genes|Genomes|Genetics 3: 1795–1807. doi: 10.1534/g3.113.007161 PMID: 23979933

13. Lin P, Hartz SM, Zhang Z, Saccone SF, Wang J, Tischfield JA, et al. (2010) A New Statistic to Evaluate Imputation Reliability. PLoS One 5: e9697. doi: 10.1371/journal.pone.0009697 PMID: 20300623

14. Shriner D, Adeyemo A, Chen G, Rotimi CN (2010) Practical Considerations for Imputation of Untyped Markers in Admixed Populations. Genetic epidemiology 34: 258–265. doi: 10.1002/gepi.20457 PMID: 19918757

15. Chanda P, Yuhki N, Li M, Bader JS, Hartz A, Boerwinkle E, et al. (2012) Comprehensive evaluation of imputation performance in African Americans. Journal of human genetics 57: 411–421. doi: 10.1038/jhg.2012.43 PMID: 22648186

16. The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65. doi: 10.1038/nature11632 PMID: 23128226

17. The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073. doi: 10.1038/nature09534 PMID: 20981092

18. Zheng J, Li Y, Abecasis GR, Scheet P (2011) A Comparison of Approaches to Account for Uncertainty in Analysis of Imputed Genotypes. Genetic epidemiology 35: 102–110. doi: 10.1002/gepi.20552 PMID: 21254217

19. Shriner D (2013) Impact of Hardy—Weinberg disequilibrium on post-imputation quality control. Human genetics 132: 1073–1075. doi: 10.1007/s00439-013-1336-x PMID: 23842951

20. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira Manuel AR, Bender D, et al. (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. American Journal of Human Genetics 81: 559–575. PMID: 17701901

21. Bierut LJ, Madden PAF, Breslau N, Johnson EO, Hatsukami D, Pomerleau OF, et al. (2007) Novel genes identified in a high-density genome wide association study for nicotine dependence. Human Molecular Genetics 16: 24–35. PMID: 17158188

22. Saccone SF, Hinrichs AL, Saccone NL, Chase GA, Konvicka K, Madden PA, et al. (2007) Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. Hum Mol Genet 16: 36–49. PMID: 17135278

23. Saccone NL, Culverhouse RC, Schwantes-An T-H, Cannon DS, Chen X, Cichon S, et al. (2010) Multiple Independent Loci at Chromosome 15q25.1 Affect Smoking Quantity: a Meta-Analysis and Comparison with Lung Cancer and COPD. PLoS Genetics 6: e1001053. doi: 10.1371/journal.pgen.1001053 PMID: 20700436

24. Liu JZ, Tozzi F, Waterworth DM, Pillai SG, Muglia P, Middleton L, et al. (2010) Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nat Genet 42: 436–440. doi: 10.1038/ng.572 PMID: 20418889

25. Tobacco and Genetics Consortium (2010) Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nat Genet 42: 441–447. doi: 10.1038/ng.571 PMID: 20418890

26. Thorgeirsson TE, Gudbjartsson DF, Surakka I, Vink JM, Amin N, Geller F, et al. (2010) Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. Nature genetics 42: 448–453. doi: 10.1038/ng.573 PMID: 20418888

27. Luo Z, Alvarado GF, Hatsukami DK, Johnson EO, Bierut LJ, Breslau N (2008) Race Differences in Nicotine Dependence in the Collaborative Genetic Study of Nicotine Dependence (COGEND). Nicotine & Tobacco Research 10: 1223–1230.

28. Winkler TW, Day FR, Croteau-Chonka DC, Wood AR, Locke AE, Mägi R, et al. (2014) Quality control and conduct of genome-wide association meta-analyses. Nat Protocols 9: 1192–1212. doi: 10.1038/nprot.2014.071 PMID: 24762786

29. Zheng H-F, Ladouceur M, Greenwood CMT, Richards JB (2012) Effect of Genome-Wide Genotyping and Reference Panels on Rare Variants Imputation. Journal of Genetics and Genomics 39: 545–550. doi: 10.1016/j.jgg.2012.07.002 PMID: 23089364

30. Zheng H-F, Rong J-J, Liu M, Han F, Zhang X-W, Richards JB, et al. (2015) Performance of Genotype Imputation for Low Frequency and Rare Variants from the 1000 Genomes. PLoS One 10: e0116487. doi: 10.1371/journal.pone.0116487 PMID: 25621886

31. Liu EY, Buyske S, Aragaki AK, Peters U, Boerwinkle E, Carlson C, et al. (2012) Genotype Imputation of MetabochipSNPs Using a Study-Specific Reference Panel of ∼4,000 Haplotypes in African Americans From the Women's Health Initiative. Genetic epidemiology 36: 107–117. doi: 10.1002/gepi.21603 PMID: 22851474

32. Asimit J, Zeggini E (2010) Rare Variant Association Analysis Methods for Complex Traits. Annual Review of Genetics 44: 293–308. doi: 10.1146/annurev-genet-102209-163421 PMID: 21047260

33. Truong L, Park H, Chang S, Ziogas A, Neuhausen S, Wang S, et al. (2015) Human Nail Clippings as a Source of DNA for Genetic Studies. Open Journal of Epidemiology: 41–50. PMID: 26180661

34. Rogers A, Beck A, Tintle NL (2014) Evaluating the concordance between sequencing, imputation and microarray genotype calls in the GAW18 data. BMC Proceedings 8: S22–S22. doi: 10.1186/1753-6561-8-S1-S22 PMID: 25519374

35. Duan Q, Liu EY, Croteau-Chonka DC, Mohlke KL, Li Y (2013) A comprehensive SNP and indel imputability database. Bioinformatics 29: 528–531. doi: 10.1093/bioinformatics/bts724 PMID: 23292738