2014

Value of Mendelian Laws of Segregation in Families: Data Quality Control, Imputation, and Beyond

Elizabeth M. Blue

Lei Sun

Nathan L. Tintle
Dordt College, nathan.tintle@dordt.edu

Ellen M. Wijsman

Value of Mendelian Laws of Segregation in Families: Data Quality Control, Imputation, and Beyond

Abstract
When analyzing family data, we dream of perfectly informative data, even whole-genome sequences (WGSs) for all family members. Reality intervenes, and we find that next-generation sequencing (NGS) data have errors and are often too expensive or impossible to collect on everyone. The Genetic Analysis Workshop 18 working groups on quality control and dropping WGSs through families using a genome-wide association framework focused on finding, correcting, and using errors within the available sequence and family data, developing methods to infer and analyze missing sequence data among relatives, and testing for linkage and association with simulated blood pressure. We found that single-nucleotide polymorphisms, NGS data, and imputed data are generally concordant but that errors are particularly likely at rare variants, for homozygous genotypes, within regions with repeated sequences or structural variants, and within sequence data imputed from unrelated individuals. Admixture complicated identification of cryptic relatedness, but information from Mendelian transmission improved error detection and provided an estimate of the de novo mutation rate. Computationally, fast rule-based imputation was accurate but could not cover as many loci or subjects as more computationally demanding probability-based methods. Incorporating population-level data into pedigree-based imputation methods improved results. Observed data outperformed imputed data in association testing, but imputed data were also useful. We discuss the strengths and weaknesses of existing methods and suggest possible future directions, such as improving communication between data collectors and data analysts, establishing thresholds for and improving imputation quality, and incorporating error into imputation and analytical models.

Keywords
inference, type I error, power, next-generation sequence data, de novo mutation

Comments
- This is a pre-publication author manuscript of the final, published article.
- The definitive version is published by Wiley and available from Wiley Online Library wileyonlinelibrary.com). DOI 10.1002/gepi.21821
- Elizabeth E. Marchani listed as lead author on pre-publication version while the publisher's version lists her as Elizabeth M. Blue
- Title on pre-publication version is "On the Value of Mendelian Laws of Segregation in Familes: Data Quality Control, Imputation and Beyond" while final published version lists title as "Value of Mendelian Laws of Segregation in Families: Data Quality Control, Imputation, and Beyond"

TITLE: On the value of Mendelian laws of segregation in families: data quality control, imputation and beyond

Elizabeth E. Marchani[1*], Lei Sun[2,3], Nathan L. Tintle[4], Ellen M. Wijsman[1,5]

[1]Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA, USA

[2]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Canada

[3]Department of Statistics, University of Toronto, Canada

[4]Department of Mathematics, Statistics and Computer Science, Dordt College, Sioux Center, IA, USA

[5]Department of Biostatistics and Department of Genome Sciences, University of Washington, Seattle, WA, USA

*Correspondence to: Elizabeth Marchani, Division of Medical Genetics, University of Washington, BOX 357720, phone: (206) 685-4666.

Running title: Bad data begets bad data

ABSTRACT

When analyzing family data, we dream of perfectly informative data, even whole genome sequences for all family members. Reality intervenes, and we find next-generation sequence (NGS) data has error, and is often too expensive or impossible to collect on everyone. Genetic Analysis Workshop 18 groups "Quality Control" and "Dropping WGS through families using GWAS framework" focused on finding, correcting, and using errors within the available sequence and family data, developing methods to infer and analyze missing sequence data among relatives, and testing for linkage and association with simulated blood pressure.

We found that dense SNP, NGS, and imputed data are generally concordant, but that errors are particularly likely at rare variants, homozygous loci, within regions with repeated sequences or structural variants, and within sequence data imputed from non-relatives. Admixture complicated identification of cryptic relatedness, but information from Mendelian transmission improved error detection and provided an estimate of the *de novo* mutation rate. Both genotype and pedigree errors had an adverse effect on subsequent analyses. Computationally fast rules-based imputation was accurate, but could not cover as many loci or subjects as more computationally demanding probability-based methods. Incorporating population-level information into pedigree-based imputation methods improved results. Observed data outperformed imputed data in association testing, but imputed data were also useful.

We discuss the strengths and weaknesses of existing methods, and suggest possible future directions. Topics include improving communication between those performing data collection and analysis, establishing thresholds for and improving imputation quality, and incorporating error into imputation and analytical models.

Keywords: Inference, type 1 error, power, next-generation sequence data, *de novo* mutation

INTRODUCTION

Recent breakthroughs in next generation sequencing (NGS) technology are generating massive amounts of data on both rare and common variants. While the potential of this data deluge is staggering, so are the potential questions regarding analysis. To date, many methodological developments using NGS technologies either (a) assume that data are perfect and evaluate competing analytical techniques, or (b) focus entirely on data production and quality control, with little regard for the downstream implications regarding data processing.

At Genetic Analysis Workshop 18 (GAW18), two working groups considered data quality issues. The quality control (QC) group focused primarily on evaluating and developing ways to assess the quality of sequence and pedigree data, while discussing the potential implications of the data quality issues identified. The gene-dropping group explored how the pedigree structure of the data lent itself to novel approaches to imputation and statistical tests for genotype-phenotype relationships. By necessity, the gene-dropping group also discussed data quality and approaches to handling genotype and pedigree errors, as these errors can become particularly amplified by such approaches. After the workshop, the leaders of the groups decided it prudent to jointly summarize their findings to provide a more complete picture of approaches to assessing and solving data quality issues. We also evaluate the implications of these decisions regarding subsequent analyses where such mistakes can create potentially disastrous effects.

For over three decades, as new genotyping technologies have been introduced, the statistical genetics community has repeatedly wrestled with a host of issues related to data quality. Initial error rates from each technology were high (*e.g.*, for SNP microarrays see [Tintle et al., 2005]), with widespread implications including inflation of genetic map distances and biased estimates of the recombination fraction and linkage disequilibrium (LD) between loci [Buetow 1991;

Gordon and Finch 2005; Huang et al., 2004; Sobel et al., 2002]. Genotype errors can also inflate the type I error or reduce power of statistical analyses [Chang et al., 2006], depending on whether the errors are correlated with the phenotype [Gordon and Finch 2005]. Over time, data quality benefited from improvements in laboratory protocols, study design, genotype calling algorithms, and data screening approaches (*e.g.*, departure from Hardy-Weinberg equilibrium or high rates of missing data)[Laurie et al., 2010; Pluzhnikov et al., 2010], among other, older, methods to identify problematic genotypes [Ehm et al., 1996; Gordon and Finch 2005; O'Connell and Weeks 1998].

Such work was not long forgotten with the advent of genotype imputation, which uses inferred pedigree-based or population-based haplotypes to predict unmeasured genetic variants (*e.g.*, [Howie et al., 2009; Li et al., 2010]). The promise of "free" genotype data was and remains hugely appealing. However, genotype errors (this time produced exclusively *in silico*) are of concern in imputation, with the same issues of inflated type I errors and power loss [Beecham et al., 2010; Huang et al., 2009a; Huang et al., 2009b].

Now, with the advent of another sequencing technology, data quality issues are back in the spotlight with high error rates due to a variety of potential sources [Awadalla et al., 2010; Ilie et al., 2011; Nielsen et al., 2011]. Perhaps even more than for dense SNPs and imputation, the impact of these errors has the potential to be great. Rare variants are of profound biological importance, and are no longer ignored by data analysts. However, the high sequencing genotype error rate makes researchers ask, "Is that an error?" Preliminary assessment of gene-based tests of rare variant association show genotyping errors have strong effects on type I error and power [Garner 2009; Mayer-Jochimsen et al., 2013; Powers et al., 2011]. Renewed interest in family data, where rare variants may be easier to identify and associate with disease phenotypes, is

pushing the methodological and computational envelope. However, pedigree errors, as well as cryptic relatedness, often occur and also adversely affect downstream analyses. Application of imputation methods to sequence data in pedigrees, while potentially beneficial, can also dramatically magnify the adverse effects of initial sequencing errors.

Here, we discuss a variety of approaches to evaluate errors in NGS and pedigree data, with their implications. We begin by discussing approaches to QC for NGS data in pedigrees and for pedigree structures in an effort to inform best practice for the processing and imputation of genotypes. When multi-generation families are available, as in GAW18, we discuss a method that exploits apparent Mendelian inheritance errors to estimate the *de novo* mutation rate without additional genotyping for validation. We then explore a variety of approaches for genotype imputation in pedigrees, and the confidence we can have in the results, which rely heavily on data quality. Lastly, we briefly explore some implications of genotype and pedigree errors as well as joint use of population and pedigree data when testing genotype-phenotype association. We conclude with a discussion of open questions and our final conclusions.

ASSESSING DATA QUALITY

We begin by focusing on approaches taken by papers to assess data quality. QC papers tended to focus either on potential sample errors in the pedigree structures provided by GAW18, or on genotype quality. We structure the following sections accordingly.

Evaluating pedigree structure and cryptic relatedness

It is now well accepted that, despite the best practice in data collection, sample errors can occur within pedigrees (*e.g.*, sample swaps, non-paternity/maternity) or between pedigrees (*e.g.*, cryptic relatedness, population structure). Such errors can result in increased type I error or decreased power. The individuals in the GAW18 data were part of 20 distinct multi-generational

pedigrees (see GAW18 data description paper in this volume for more details). In summary, the pedigree information was reported to have been validated by estimation of kinship coefficients, principal components analysis, and investigation into apparent Mendelian errors. However, no other details were provided. Therefore, two papers evaluated potential sample errors remaining in the GAW18 data using the genotype data available for 959 individuals [Marchani et al., in press; Sun and Dimitromanolakis in press].

Sun and Dimitromanolaki [in press] used a likelihood-based method that assumes a homogeneous population [McPeek and Sun 2000] implemented in PREST-plus [Sun and Dimitromanolakis 2012] to estimate identity-by-descent (IBD), along with a formal hypothesis test for relationship errors. Among all possible pairs of individuals within families, strong evidence for misspecified relationships were found for 7 pairs, and plausible alternatives compatible with the observed genotypes were proposed. Sun and Dimitromanolaki [in press] also considered possible cryptic relatedness among the 147 purportedly unrelated individuals, and found four pairs with strong evidence for relatedness (half first-cousin to first-cousin).

Marchani et al. [in press] similarly evaluated cryptic relatedness, but used King Robust [Manichaikul et al., 2010] and REAP [Thornton et al., 2012] to accommodate population admixture, which was present in this sample. After analyzing all pedigrees, they found evidence of cryptic relatedness greater than second cousins for pairs of relatives belonging to a total of 7 families. These results were later confirmed by GAW18 organizers [John Blangero, personal communication]. When compared with the results of Sun and Dimitromanolaki [in press], the closest relationships were identified by both groups, but the two groups did not always identify the same cryptically related pairs. This illustrates the sensitivity of the conclusions to

incorporation of ancestry and admixture in the analysis model.

Evaluating genotype quality

Genotypes in the GAW18 data came from three distinct sources: direct NGS data, imputed NGS data using a novel population-based pipeline on 476 individuals (see data description paper in this volume), and GWAS data on approximately 550,000 common variants (population minor allele frequency, or MAF, >5%) on the entire sample. Genotype quality was evaluated by: (1) comparing genotypes of the same marker across platforms [Hinrichs et al., in press; Rogers et al., in press; Song et al., in press], (2) describing which errors are most common [Blackburn et al., in press; Pilipenko et al., in press; Rogers et al., in press] and (3) and evaluating when apparent genotyping errors are actually *de novo* mutations [Wang and Zhu in press].

Three groups evaluated marker consistency across platforms. Song et al. [in press] examined a single chromosome and removed families with high levels of missing data from the analysis. They found 10.43% of genotypes to be discordant, of which 71.26% were mismatched genotypes, while 28.74% were missing within a data set. Two other papers analyzed all available data, and found reasonably high average concordance between NGS (direct or imputed) and GWAS genotypes: 97.6% [Hinrichs et al., in press] and 97.37% [Rogers et al., in press]. Concordance increased when only called genotypes were considered: 99.74% [Hinrichs et al., in press] and 99.77% [Rogers et al., in press]. The discordant genotypes are generally found at NGS sites with higher rates of missing data, and at imputed NGS sites in particular.

Three papers further evaluated which types of variant sites were most likely to have discordant genotypes. Rogers et al. [in press] found that less common variants were more likely to be inconsistently genotyped. However, they only evaluated variants with MAF >5%, and so discordance rates among the truly rare variants are unknown. Rogers et al. [in press] also found

that heterozygote (GWAS) to non-reference allele homozygote (NGS) and reference allele homozygote (GWAS) to heterozygote (NGS) discrepancies were the most common, accounting for over 80% of all discordance among called genotypes. This suggests that imputation of minor alleles may be occurring more often than it should. Lastly, Pilipenko et al. [in press] used the pedigree structure and sequence data to examine the distribution of Mendelian Inheritance Errors (MIEs) across chromosome 3 and identify their characteristics. They found that MIEs tended to cluster near repetitive sequence locations. This corresponded to the findings of Blackburn et al. [in press] who found that inconsistencies between their own imputed genotypes (see below) and measured genotypes tended to occur in regions with structural variants (*e.g.*, copy number variants). Pilipenko et al. [in press] also concluded that, among the various QC parameters provided in the sequencing file, an SVM threshold >3.5 was the most effective at reducing MIEs.

While most QC work considered errors as a nuisance, Wang et al. [in press] took the view that not all errors are bad and developed a novel approach utilizing the apparent MIEs and the three-generation family data to accurately estimate the *de novo* mutation rate. In particular, MIEs identified in the first two generations were *de novo* mutation candidates, among which the true *de novo* mutations should be transmitted to the third generation following Mendelian laws. Using this approach, a *de novo* mutation rate of 1.64×10^{-8} per position was obtained. This is consistent with estimates obtained using more costly validation study designs requiring additional genotyping. The work of Wang et al. [in press] also showed that over 95% of the *de novo* mutation candidates were, in fact, genotyping errors.

IMPUTATION APPROACHES

Pedigree data provides powerful information for identifying genetic regions influencing traits, learning about *de novo* mutations, and detecting recombination hot spots. It is also an

investment, often taking years to collect samples for larger pedigrees. Researchers who have invested in genome-scan genotyping in their pedigrees may be tempted to couple that data with Mendel's laws of inheritance to impute NGS data collected on a subset of pedigree members into their relatives. Here, we summarize different strategies for imputing missing genotype data through pedigrees using strict rules (heuristic) or probabilistic methods. All methods used two sets of marker data: a sparser framework panel (often GWAS) observed in most relatives, and a rich, denser marker panel (often NGS) observed in a subset of those relatives. Pedigree relationships and framework genotypes, both assumed to be correct, are used to estimate IBD sharing among relatives, which are then used to impute rich marker genotypes in relatives with missing data.

Heuristic methods

Imputation of genotypes in large families, or those with distant relationships and missing data, is not well supported by existing tools, such as Merlin [Abecasis et al., 2002] and Mendel [Chen et al., 2012]. Two groups developed heuristic imputation methods, which perform these tasks accurately and expediently for a subset of markers and subjects. The heuristic methods only assign alleles where the inferred IBD information in a pedigree coupled with the observed marker data forces the allelic states at a marker and the phase between markers.

Blackburn et al. [in press] required true pedigree structure and that framework markers are observed for all relatives included in the analysis. They used Mendel's laws on trios in the pedigrees to phase founder parent haplotypes for the framework markers, and then knitted the trios together using a minimum recombination model [Qian and Beckmann 2002]. The rich marker panel was then phased using trio information, with the haploid genotypes mapped to each

of a founder's phased haplotypes. The inheritance patterns of founder framework haplotypes were then used to impute genotypes into relatives missing rich genotypes.

Song et al. [in press] applied PedIBD [Li and Li 2011], which also assumes true pedigree structure but tolerates missing data in relatives. This method considers pairs of individuals, with further constraints imposed when all sets of these pairs are considered. They used GWAS framework data to identify recombination break points, and then inferred phased haplotypes between them. Some individuals without framework marker data may still be assigned phased haplotypes because they are obligate carriers (*e.g.*, a parent with missing data and observed offspring), but as observed by Blackburn et al. [in press], allelic states for missing genotypes cannot be propagated indefinitely into parts of a pedigree lacking framework marker data.

To measure accuracy, Blackburn et al. [in press] divided the GWAS data into framework and a rich marker panels. Half of the genotypes in the rich panel were "masked", or treated as missing, and imputed. Average imputation accuracy was measured by the IQS statistic [Lin et al., 2010], which adjusts for chance concordance between masked and imputed genotypes, and MAF. The IQS statistic averaged 0.992 for the 211,736 markers with chance concordance <1. Although imputation was generally accurate, imputation of rare variants was less reliable (IQS = 0.972 at MAF $\leq$ 0.01) than for common variants (IQS = 0.994 at MAF >0.4).

Song et al. [in press] examined the accuracy of their imputation method by masking NGS data for 5 subjects from Family 21, each with multiple relatives with GWAS and NGS data. Accuracy ranged from 91.40% to 99.44%, with no clear relationship with the number of sequenced first- or second-degree relatives. Instead, NGS data quality in relatives appears to have a stronger influence on the accuracy of imputed data. Song et al. [in press] were able to

impute 82.9% of the NGS variants for 1,011 individuals (198 of whom without GWAS or NGS data), with nearly 90% of the uncalled variants causing Mendelian errors if called.

Although the heuristic imputation groups shared neither the same data nor the same statistics, a few common lessons are revealed. Heuristic methods did not incorporate population-level information when phasing haplotypes, such as MAF or LD. This may be considered an advantage when the pedigrees of interest do not clearly belong to a single reference population, such as in GAW18 where there is population stratification within a sample of admixed pedigrees. Both groups identified recombination points by looking for "switches" where an individual's haplotype changed between marker loci, and avoided imputing genotypes in such regions. Combined with the strict rules within their heuristic approaches, this means that their methods are very accurate, but some loci near recombination points will not be imputable, nor will information impute into individuals with ambiguous IBD information.

Probabilistic methods

Marchani et al. [in press] used a pedigree-based imputation approach that allows for imperfect IBD information. This probabilistic approach includes the rules behind the heuristic approaches, but incorporates additional information with computations that are consequently more demanding. The imputation program, GIGI [Cheung et al., 2013], uses Markov Chain Monte Carlo realizations of inheritance vectors [http://www.stat.washington.edu/thompson/Genepi/Pangaea.shtml], conditional on sparse genotype markers in linkage equilibrium, pedigree structure, allele frequencies, and meiotic marker map positions. These inheritance vectors are combined with the observed rich and framework marker data and allele frequencies to estimate genotype probabilities for each rich marker for each subject missing data. This provides the opportunity to estimate probable

genotypes in the face of greater uncertainty, such as near recombination break points or in unsampled individuals. Accuracy was measured as percent agreement between the imputed and observed GWAS data available in masked individuals. Marchani et al. [in press] compared and integrated GIGI with population-based imputation implemented in BEAGLE [Browning and Browning 2009]. BEAGLE uses a reference panel of genotypes from unrelated individuals and population-level LD to impute rich marker data for a set of unrelated individuals with framework marker data available.

Marchani et al. [in press] imputed chromosome 3 GWAS data for pedigree 10, using founders from several pedigrees as a reference panel for population-based imputation. Strategies to select whose rich data to observe and whose to impute were influential. Although GIGI (99.8% -98.1%) and BEAGLE (99.1%-97.8%) were both very accurate regardless of the amount of masked rich marker data, they were less likely to impute genotypes when less rich data was available. Accuracy at rare variants (MAF<0.05) was at least 99% for both GIGI and BEAGLE under both conditions. Interestingly, combining results from GIGI and BEAGLE using a few simple rules resulted in overall gains in call rates and accuracy comparable to those resulting from an increase in the amount of observed marker data.

Probabilistic imputation strategies offered promise, as they are able to impute data into more missing individuals across more loci than purely heuristic approaches [Blackburn et al., in press; Marchani et al., in press; Song et al., in press]. However, they come with a computational price: hours, instead of minutes, of computational time [Blackburn et al., in press; Marchani et al., in press], if inheritance vectors have not yet been sampled. Genotype imputation within pedigrees can improve with the incorporation of population-level data [Marchani et al., in press], but caution must be taken to ensure that Markov chains are reducible and the entire sample space of

inheritance vectors is represented. There was a reduction in call rate, and therefore the underlying probability of an imputed genotype, when pedigree members are absent from the population reference panel, and when less rich data is available within a pedigree [Marchani et al., in press]. The direct effect of this on accuracy relative to other pedigree-based imputation approaches is difficult to determine, as some papers chose not to impute uncertain calls, while others always imputed genotypes regardless of their probability. Rare alleles are often imputed with reduced accuracy and/or call rate by both heuristic and probabilistic approaches, although Marchani et al. [in press] found that rare alleles were more successfully imputed by a pedigree-based, rather than a population-based method.

GENOTYPE-PHENOTYPE ASSOCIATION TESTING

Many groups at GAW18 performed association testing, including family-based tests, and readers interested in this area should also examine the other group summary papers in this volume. However, association testing was also explored in the gene-dropping group, through the application of a new score test [Jiang et al., in press] and evaluation of the use of imputed data for analysis [Marchani et al.,et al., in press]. The implementation of the methods in both of these groups was based on the same MCMC-based approach and program to sample of inheritance vectors from complete pedigrees. Both groups also analyzed the simulated diastolic blood pressure values (SBP) from the first simulated data set, and analyzed only chromosome 3 GWAS data. Their quality control measures and choice of covariate adjustments varied, along with their choice of pedigrees for analysis, precluding direct comparison of the analysis results between the groups. However, both groups concluded that inclusion of pedigree data led to stronger p-values at the loci with most influence on SBP.

Marchani et al. [in press] compared mixed-model variance components association testing using a kinship matrix based on the known pedigree structure vs. estimated from all GWAS data in the absence of known pedigree structure. Although their analysis focused only on the top hits from another half-genome scan [Thornton et al., in press], results were consistent across loci: both forms of analysis identified significant associations around the simulated true loci, but the pedigree-based kinship matrix provided a stronger signal at the true causal locus than did the estimated kinship matrix. In light of the findings of Hainline et al. [in press], who found that pedigree-based kinship matrices resulted in an overly-conservative type I error relative to the estimated kinship matrices, pedigree-based kinship matrices may provide a stronger boost to power than initially suspected. However, Hainline et al. [in press] focused on rare variants, analyzed a binary phenotype, and used a different approach to estimate kinship matrices. Intuition suggests, at least for single-marker tests at common variants, if the true relationship is closer than what is specified by the given pedigree, then there may be an increased type I error, while if the true relationship is more distant than the given, then power may be reduced. When there is a mixture of both types of misspecification, effects are less predictable. This is an area that deserves further research.

Marchani et al. [in press] also compared association testing results using different amounts of imputed, rather than observed, GWAS data. Observed GWAS data provided a slightly stronger association signal than the imputed data, even when the imputed data was highly accurate. The ranked order of some variants also changed with the inclusion of imputed data. However, because the difference between the p-values was generally small, this does suggest that use of imputed data is useful when directly observed data are unavailable.

Jiang et al. [in press] introduced a novel gene-dropping score statistic that also uses sampled inheritance vectors from complete pedigrees, and also incorporates association information. They compared results with traditional family-based association testing FBAT [Rabinowitz and Laird 2000], which restricts the pedigree sizes, with an association analysis using only unrelated founders, and with their joint test. They found that conditioning their score test on inheritance vector information provided results comparable to the unconditional test. The most influential SBP variants on chromosome 3 were ranked as more significant by the score statistic with joint inheritance and association information than by either the association test alone or the FBAT test. There was very little correlation between p-values across the types of tests. This may be the result of the relatively weak contribution of the loci to the simulated trait: weaker associations are more vulnerable to noise in the data, and so their p-values may fluctuate more. Each association test compared also used different sources of data: association testing used only unrelated individuals, FBAT divided large pedigrees into smaller families, and the score test used the entire pedigree structure as well as both the linkage and association information. The results from this comparison suggest that joint testing for linkage and association between genetic and phenotypic variability maximizes the amount of data used.

Jiang et al. [in press] attempted to accommodate population stratification in this sample by including the first two principal components of genetic variation as covariates, but found only modest changes in p-values. This is consistent with the findings from the GAW18 Admixture group: the inclusion of the first few principal components as covariates are not sufficient for capturing the level of population stratification in this sample [Thornton et al., in press].

Both papers found minimal evidence for linkage to chromosome 3 for simulated blood pressure, using either parametric or variance components lod scores [Marchani et al., in press],

and the linkage component of the gene-dropping score test [Jiang et al., in press]. Personal communication with John Blangero at GAW18 revealed that phenotypic variation within the first simulated phenotype data set did not co-segregate well with genetic variation.

DISCUSSION

The analysis of rare variants, often the motivation for NGS data collection, is complicated by genotyping errors. Results from GAW18 raised several important points regarding the detection, and correction, of genotyping errors in this context. The discrepancy rate between GWAS SNPs and directly sequenced NGS SNPs was higher than an earlier comparison of genotype discrepancy rates between SNP-array platforms [Hinrichs and Suarez 2005]. The higher error rate of NGS data, and the difficulty finding the errors through standard pedigree-based methods [Hinrichs and Suarez 2005], suggests it may be useful to develop better tools to detect genotyping error during QC, or to incorporate errors into downstream analytical models, such as the detection of truly *de novo* variants. However, because the pattern of discrepancies between the NGS and GWAS genotype data is different than in earlier studies of SNP discrepancies, genotyping error models may need to be more sophisticated than the simple error models that have previously been used [Douglas et al., 2002; Epstein et al., 2000]. For example, error models may benefit from incorporation of observed differences in error rates as a function of genomic signatures, such as presence of structural variation.

There is also a need to jointly consider genotype and sample errors. Traditionally, the two categories of errors are evaluated separately, and each category is assessed with the assumption that the other type of error is absent. Intuitively, a small percentage of genotype errors should not alter the sample inference based on the whole genome, with the exception of monozygotic twins, which can be missed in the absence of a model that does not allow discordant genotypes [Epstein

et al., 2000]. Similarly, a small proportion of sample errors should not change genotype error conclusions based on the whole sample. However, it is not clear if sample QC followed by genotype QC leads to the same data for downstream analysis as does genotype QC followed by sample QC. Even within each category of error, there is a need to investigate the impact of different QC steps in a sequential approach and the utility of a joint analysis.

Error detection in families sampled from structured populations has its own challenges. Sun and Dimitromanolaki [in press] demonstrated that likelihood-based approach implemented in PREST-plus is more powerful than the method-of-moments implemented in PLINK [Purcell et al., 2007]. Marchani et al. [in press] showed that modeling admixture is important when the assumption that the individuals are all from the same homogeneous population might be violated. Ignoring population structure, such as variable amounts of admixture among individuals, can lead to spurious detection of weakly related individuals. Such individuals can simply share similar ancestry, not necessarily relatively close relatives. These differences in model assumptions probably account for why the two papers did not find the same pairs of cryptically-related families, other than one pair where the relationships across pedigrees were quite strong.

In addition to the errors caused by NGS data generation, the imputation of missing sequence data can introduce additional genotyping error. The combination of population-based and small-pedigree pedigree imputation methods used to provide the GAW18 imputed sequence yielded a disappointingly high inflation in the error rate relative to the directly sequence data. The large-pedigree based imputation methods proposed by several participants gave much more accurate results, illustrating that better methods exist that more appropriately make use of the existing data structures, and should be used for this purpose in the future. Future improvements to these methods could include more sophisticated combinations of population-level LD with pedigree-

based transmission. GAW18 results suggest that the information in these two sources is nearly independent for the purposes of genotype imputation, leading to potentially large gains in both quantity and quality of imputed genotypes that might be realized in the future.

GAW18 results, as well as earlier studies, show that use of imputed genotypes for downstream analysis is less desirable than use of directly measured genotypes. A small genotyping error rate can translate into a loss in power, or conversely, a reduced statistical signal, given a fixed sample size. However, because of sample availability and cost, imputation in studies of large pedigrees will still often be useful when NGS technologies are used. It would be useful to develop measures of the change in power, expected value of a test statistic, or sample size required as a function of imputation accuracy in pedigrees. It is possible that for association testing in a family-based setting, the relationship between the accuracy and required sample size may be similar to the familiar relationship known in the use of tag-SNPs for association testing: N/r^2 [Spencer et al., 2009], where r^2 is the squared correlation between the imputed and true haplotypes, and N is the required sample size in the absence of error. In other analytical settings, a different relationship may be more pertinent. It also would be useful to evaluate different approaches to capture and use the imputed genotypes for analysis, such as whether to use a single called most-likely genotype, an average "dose" across possible genotypes, or multiple imputations based on repeated analysis with a sampling of possible genotypes. Previously developed principles for the use of imputed data should also hold [Little 1992; Rubin 1996].

Although combining pedigree-based inheritance vectors and population LD information can be challenging, GAW18 participants found it also can be beneficial. Both of the papers that combined these sources of information reported gains: in the strength of a test of linkage and association [Jiang et al., in press], and in the fraction of genotypes that could be imputed at a

given accuracy level [Marchani et al., in press]. Both groups allowed the inheritance vectors and LD information to be obtained separately, and then combined. This is easier than constructing methods that accurately determine or sample inheritance vectors in the presence of tightly-linked markers in LD, and also provides some information about the relative importance of segregating variation in the pedigrees vs. population association in a given situation. It also allows each part of the analysis to be carried out using optimal marker spacing for that component, thereby increasing computational efficiency without losing power or accuracy: using moderately-spaced markers for estimation of inheritance vectors [Wilcox et al., 2005], and densely-space markers for capturing population-level association [Browning and Browning 2009].

It is not surprising that we found QC and subsequent analysis to be intertwined at GAW18. However, NGS data presents a new challenge, as "raw" NGS data has in fact undergone considerable pre-processing. Those protocols were not described here, and are often not sufficiently described to understand their potential sources of error. Such errors or biases incurred during data generation nevertheless will be transmitted within the data. Bioinformaticists, statistical geneticists, and others who implement these methods are already tackling multiple aspects of the challenges inherent to NGS data, sometimes in a redundant fashion. In order to further develop the QC and gene-dropping analyses summarized here, better communication across all stages of data analysis will be necessary.

ACKNOWLEDGEMENTS

Supported by NIH grants P50 AG005136, R01 AG039700, K99 AG040184, R15 HG006915, R15 HG004543, R37 GM042655, R01 MH092367, and R01 MH094293. The Genetic Analysis Workshop 18 was supported by NIH grant R01 GM0031575.

REFERENCES

Abecasis GR, Cherny SS, Cookson WO, Cardon LR. 2002. Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nature Genetics 30:97-101.

Awadalla P, Gauthier J, Myers RA, Casals F, Hamdan FF, Griffing AR, Cote M, Henrion E, Spiegelman D, Tarabeux J and others. 2010. Direct Measure of the De Novo Mutation Rate in Autism and Schizophrenia Cohorts. American Journal of Human Genetics 87(3):316-324.

Beecham GW, Martin ER, Gilbert JR, Haines JL, Pericak-Vance MA. 2010. APOE is not Associated with Alzheimer Disease: a Cautionary tale of Genotype Imputation. Annals of Human Genetics 74:189-194.

Blackburn AN, Dean AK, Lehman DM. in press. Imputation in families using a heuristic phasing approach. BMC Proceedings.

Browning BL, Browning SR. 2009. 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(2):210-223.

Buetow K. 1991. Influence of aberrant observations on high-resolution linkage analysis outcomes. American Journal of Human Genetics 49:985-994.

Chang YPC, Kim JDO, Schwander K, Rao DC, Miller MB, Weder AB, Cooper RS, Schork NJ, Province MA, Morrison AC and others. 2006. The impact of data quality on the identification of complex disease genes: experience from the Family Blood Pressure Program. European Journal of Human Genetics 14(4):469-477.

Chen GK, Wang K, Stram AH, Sobel EM, Lange K. 2012. Mendel-GPU: haplotyping and genotype imputation on graphics processing units. Bioinformatics 28(22):2979-2980.

Cheung CYK, Thompson EA, Wijsman EM. 2013. GIGI: An approach to effective imputation of dense genotypes on large pedigrees. American Journal of Human Genetics 92(4):504-516.

Douglas JA, Skol AD, Boehnke M. 2002. Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. American Journal of Human Genetics 70(2):487-495.

Ehm MG, Kimmel M, Cottingham RW. 1996. Error detection for genetic data, using likelihood methods. American Journal of Human Genetics 58(1):225-234.

Epstein MP, Duren WL, Boehnke M. 2000. Improved inference of relationship for pairs of individuals. American Journal of Human Genetics 67(5):1219-31.

Garner C. 2009. Confounded by sequencing depth in association studies of rare alleles. Genetic Epidemiology 35(3):261-268.

Gordon D, Finch SJ. 2005. Factors affecting statistical power in the detection of genetic association. Journal of Clinical Investigation 115(6):1408-1418.

Hinrichs AL, Culverhouse RC, Suarez BK. in press. Genotypic discrepancies arising from imputation. BMC Proceedings.

Hinrichs AL, Suarez BK. 2005. Genotyping errors, pedigree errors, and missing data. Genetic Epidemiology 29:S120-S124.

Howie BN, Donnelly P, Marchini J. 2009. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLoS Genetics 5(6).

Huang L, Li Y, Singleton AB, Hardy JA, Abecasis G, Rosenberg NA, Scheet P. 2009a. Genotype-Imputation Accuracy across Worldwide Human Populations. American Journal of Human Genetics 84(2):235-250.

Huang L, Wang CL, Rosenberg NA. 2009b. The Relationship between Imputation Error and Statistical Power in Genetic Association Studies in Diverse Populations. American Journal of Human Genetics 85(5):692-698.

Huang QQ, Shete S, Amos CI. 2004. Ignoring linkage disequilibrium among tightly linked markers induces false-positive evidence of linkage for affected sib pair analysis. American Journal of Human Genetics 75:1106-1112.

Ilie L, Fazayeli F, Ilie S. 2011. HiTEC: accurate error correction in high-throughput sequencing data. Bioinformatics 27(3):295-302.

Jiang Y, Emerson S, Wang L, Li L, Di Y. in press. Family-based association test using normal approximation to gene dropping null distribution. BMC Proceedings.

Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, Bhangale T, Boehm F, Caporaso NE, Cornelis MC, Edenberg HJ and others. 2010. Quality Control and Quality Assurance in Genotypic Data for Genome-Wide Association Studies. Genetic Epidemiology 34(6):591-602.

Li X, Li J. 2011. Haplotype Reconstruction in Large Pedigrees with Untyped Individuals through IBD Inference. Journal of Computational Biology 18(11):1411-1421.

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(8):816-834.

Lin P, Hartz SM, Zhang ZH, Saccone SF, Wang J, Tischfield JA, Edenberg HJ, Kramer JR, Goate AM, Bierut LJ and others. 2010. A New Statistic to Evaluate Imputation Reliability. PLoS One 5(3).

Little RJA. 1992. Regression with Missing Xs - a Review. Journal of the American Statistical Association 87(420):1227-1237.

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. 2010. Robust relationship inference in genome-wide association studies. Bioinformatics 26(22):2867-2873.

Marchani EE, Cheung CYK, Glazner CG, Conomos MP, Lewis SM, Sverdlov S, Thornton T, Wijsman EM. in press. Identity-by-descent graphs offer a flexible framework for imputation and both linkage and association analyses. BMC Proceedings.

Mayer-Jochimsen M, Fast S, Tintle NL. 2013. Assessing the impact of differential genotyping errors on rare variant tests of association. PLoS One 8(3):e56626.

McPeek MS, Sun L. 2000. Statistical tests for detection of misspecified relationships by use of genome-screen data. American Journal of Human Genetics 66(3):1076-1094.

Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011. Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics 12(6):443-451.

O'Connell JR, Weeks DE. 1998. PedCheck: A program for identification of genotype incompatibilities in linkage analysis. American Journal of Human Genetics 63(1):259-266.

Pilipenko V, He H, Kurowski B, Alexander ES, Zhang X, Ding L, Baye TM, Kottyan L, Fardo D, Martin LJ. in press. Using Mendelian inheritance errors as quality control criteria in whole genome sequencing dataset. BMC Proceedings.

Pluzhnikov A, Below JE, Konkashbaev A, Tikhomirov A, Kistner-Griffin E, Roe CA, Nicolae DL, Cox NJ. 2010. Spoiling the Whole Bunch: Quality Control Aimed at Preserving the Integrity of High-Throughput Genotyping. American Journal of Human Genetics 87(1):123-128.

Powers S, Gopalakrishnan S, Tintle N. 2011. Assessing the Impact of Non-Differential Genotyping Errors on Rare Variant Tests of Association. Human Heredity 72(3):153-160.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ and others. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81(3):559-575.

Qian DJ, Beckmann L. 2002. Minimum-recombinant haplotyping in pedigrees. American Journal of Human Genetics 70(6):1434-1445.

Rabinowitz D, Laird N. 2000. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Human Heredity 50(4):211-223.

Rogers A, Beck A, Tintle NL. in press. Evaluating the concordance between sequencing, imputation and microarray genotype calls in the GAW18 data. BMC Proceedings.

Rubin DB. 1996. Multiple imputation after 18+ years. Journal of the American Statistical Association 91(434):473-489.

Sobel E, Papp JC, Lange K. 2002. Detection and integration of genotyping errors in statistical genetics. American Journal of Human Genetics 70(2):496-508.

Song S, Shields RL, X., Li J. in press. Joint analysis of sequence data and SNP data using pedigree information for imputation and recombination inference. BMC Proceedings.

Spencer CCA, Su Z, Donnelly P, Marchini J. 2009. Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip. PLoS Genetics 5(5).

Sun L, Dimitromanolakis A. 2012. Identifying cryptic relationships. Methods Mol Biol 850:47-57.

Sun L, Dimitromanolakis A. in press. PREST-plus identifies pedigree errors and cryptic relatedness in the GAW18 sample using genome-wide SNP data. BMC Proceedings.

Thornton T, Conomos MP, Sverdlov S, Marchani EE, Cheung C, Glazner C, Lewis SM, Wijsman EM. in press. Estimating and adjusting for ancestry admixture in statistical methods for relatedness inference, heritability estimation, and association testing. BMC Proceedings.

Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, Risch N. 2012. Estimating kinship in admixed populations. American Journal of Human Genetics 91(1):122-138.

Tintle NL, Ahn K, Mendell NR, Gordon D, Finch SJ. 2005. Characteristics of replicated single-nucleotide polymorphism genotypes from COGA: Affymetrix and center for inherited disease research. BMC Genetics 6.

Wang H, Zhu X. in press. *De novo* mutations discovered in eight Mexican American families through whole-genome sequencing. BMC Proceedings.

Wilcox MA, Pugh EW, Zhang H, Zhong X, Levinson DF, Kennedy GC, Wijsman EM. 2005. Comparison of single-nucleotide polymorphisms and microsatellite markers for linkage analysis in the COGA and simulated data sets for Genetic Analysis Workshop 14: Presentation groups 1, 2, and 3. Genetic Epidemiology 29 (Suppl 1):S7-S28.