Genotype Imputation Enables Powerful Combined Analyses of Genome-Wide Association Studies

Contributed by Dr. Jeffrey C. Barrett, Group Leader, Wellcome Trust Sanger Institute

Introduction

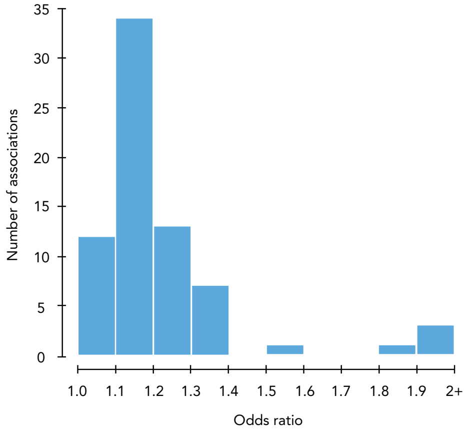
Genome-wide association studies (GWAS) have delivered hundreds of bona-fide associations to complex human disease[1]. This surge has led to new insights about the etiology of many diseases and is redefining scientific aims and approaches across human genetics. Despite these successes, it is increasingly clear that the majority of common alleles associated with disease contribute weakly to overall disease susceptibility, with only a handful conferring >25% increase in risk (Figure 1). Indeed, GWAS have only uncovered a fraction of the genetic architecture of human disease (e.g. 40 loci for height explain less than 10% of genetic variance)[2], and a great deal of interest is currently focused on explaining this “missing heritability.”

Emerging technologies such as next-generation sequencing and very high-density microarrays present opportunities to further understand the genetic variation that underpins phenotypic variation. This technological evolution, along with the availability of data from public resources like the HapMap and 1000 Genomes projects, offers a compelling motivation to extract maximum value from existing GWAS data through additional analysis and expansion of sample numbers. Genotype imputation is a statistical approach that can be used in concert with large-scale reference projects to increase the power of existing GWAS and further the discovery of novel associations. During the imputation process, GWAS genotypes at a few hundred thousand sites are analyzed in conjunction with a reference sample genotyped at millions of sites.

Imputation uses the correlation between markers present in the reference sample for making predictions of the genotypes present in an experimental sample. An imputation algorithm uses both the dense information from the reference data set and the less-dense genotype information from the experimental sample to infer genotypes at SNPs not directly genotyped in the experiment. The correlation between markers is described by linkage disequilibrium (LD) patterns across the genome, and is similar across the reference and experimental samples. This phenomenon provides the bridge that allows imputation to be successful. Therefore, assuming the information in a reference sample is fixed, the choice of markers initially genotyped in an experimental data set can strongly influence the success of imputation downstream. Illumina’s BeadChips are designed to efficiently capture, or tag, nearly all common variants across the genome. This tag SNP approach is designed to leverage known correlations of markers across the genome, making the data from these arrays well suited to cross-chip imputation. The use of LD yields higher statistical power to detect associations, and increased imputation accuracy in comparison to arrays with randomly selected SNP content.

Because commercial genotype platforms differ in their SNP content, and updated content is continually being added to newer products, imputation serves as a crucial bridge when merging distinct studies genotyped on different platforms, combining different versions of the same platform, or adopting a new platform during the course of a study. For GWAS, such meta-analyses are necessitated by the need for large sample sizes to discover modest genetic effects (Figure 2). This article presents a detailed description of genome-wide imputation applied to such meta-analyses by way of two examples: Crohn’s disease (CD)3 and Type 1 diabetes (T1D)[4], and a brief discussion of the future utility of imputation in testing association to rare variants.

Figure 1: Distribution of Odds Ratios for Risk Alleles Identified in Two Different Complex Diseases

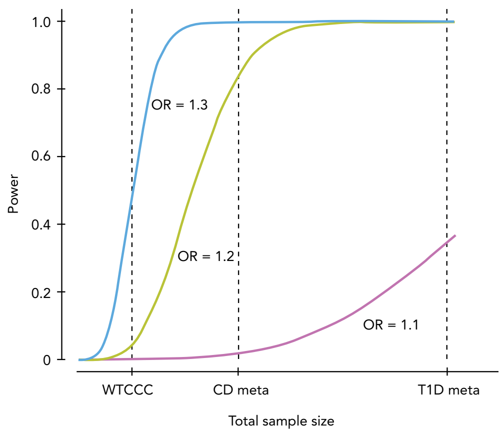


Odds ratios for each copy of the risk allele for 72 confirmed associations to Crohn’s disease3 and type 1 diabetes4

Methods

GWAS data from three separate scans for CD were assembled as part of a meta-analysis aimed at identifying common alleles of modest effect: 1450 Belgian and French samples6 genotyped on the Illumina HumanHap300, 1923 US and Canadian samples[7] also genotyped on the Illumina HumanHap300, and 4686 UK samples genotyped on the Affymetrix GeneChip 500K as part of the Wellcome Trust Case Control Consortium (WTCCC)[8]. Standard QC metrics including missing data rate, heterozygosity, allele frequency, and Hardy-Weinberg equilibrium were applied to each data set separately to obtain clean data sets for imputation. All genotypes were aligned to the + strand of build 35 of the human genome in order to match the reference set (details below). While the choice of build and orientation is arbitrary, this alignment step is critically important to avoid imputation errors, especially for SNPs with complement alleles (i.e. A/T, C/G). Different imputation programs use different file formats, but all accept either the pedigree-style genotype files used by standard analysis programs such as PLINK9 (e.g. MACH, BEAGLE), or provide specially designed file format conversion tools (e.g. IMPUTE). An excellent discussion of

Figure 2: Power of Meta-Analysis to Detect Disease Association



Power for a combined analysis to demonstrate genome-wide significant association ($P<5\times10^{-8}$) for various effect sizes (odds ratios 1.1–1.3) for a 20% allele. Total sample sizes (GWAS plus replication) are shown for the WTCCC original design (N=5,000), the CD meta-analysis[3] (N=14,871), and the T1D meta-analysis[4] (N=34,180). Very large sample sizes attained via meta-analysis are necessary to detect weak effects.

these key practicalities in imputation analysis is provided by de Bakker et al[10]. They provide detailed examples about annotating build, strand and allele information for SNPs on commercial platforms, advice on aligning these data to the HapMap, and information about correctly performing association tests on imputed data.

The initial CD meta-analysis used the HapMap2 data set consisting of 2.6 million SNPs in 60 individuals of European ancestry as the reference. Despite the success of this meta-analysis, it is likely that ever larger sample sizes will allow the detection of additional modest risk factors. A new extension to this study will add three additional sets of CD samples genotyped on the Illumina HumanHap550-Duo (comprising 3094 cases and 10,225 controls). This second-generation meta-analysis will use the new HapMap3 data release as a reference, consisting of roughly 1.5 million SNPs from two commercially available platforms (the Illumina Human1M-Duo and Affymetrix Human SNP array 6.0) genotyped in 200 European individuals. Despite having fewer total SNPs than HapMap2, this new resource is recommended for imputation for several reasons. First, data quality affects the accuracy of imputation and HapMap3 is more accurate than HapMap2. While most HapMap2 data were of high quality, the data set contained a small number of poorly performing SNPs, which can have adverse effects when cases and controls have been genotyped on different chips. The HapMap3 has extended the set of populations sampled (which is crucial for GWAS in non-European samples), and increased the number of individuals from each population. The larger reference sample size of HapMap3 offers substantial gains in imputation accuracy, especially for SNPs with <10% frequency. Finally, nearly all of the remaining common SNPs in HapMap2 are highly correlated with one or more SNPs in HapMap3; therefore, the additional SNPs in HapMap2 provide little additional information. The high-quality genotypes and larger sample panel in HapMap3 make it the current state-of-the-art reference set.

A number of different statistical frameworks have been used to tackle the problem of genotype imputation, each of which has advantages and drawbacks. The initial CD meta-analysis used the popular programs MACH11 and IMPUTE12. These programs yield high accuracy of imputed genotypes via a hidden Markov model that captures certain aspects of population history such as the local recombination rate. The trade off for the complexity of these programs is that they run slowly and require a large amount of memory, making them less suitable for the large HapMap3 reference set. The current extension of the CD project is using another HMM-based program, BEAGLE, which achieves nearly the same imputation accuracy but runs faster and can scale more readily to reference sets with hundreds of samples13. BEAGLE's speed and the ease with which it incorporates the HapMap3 data make it a good choice for current imputation analysis, but different tools may be better suited to specific problems, as discussed by Ellinghaus et al[14]. Finally, it is worth noting that the developers of these algorithms are constantly improving their programs (e.g. IMPUTE v215) to enable quicker run-times or to provide new features.

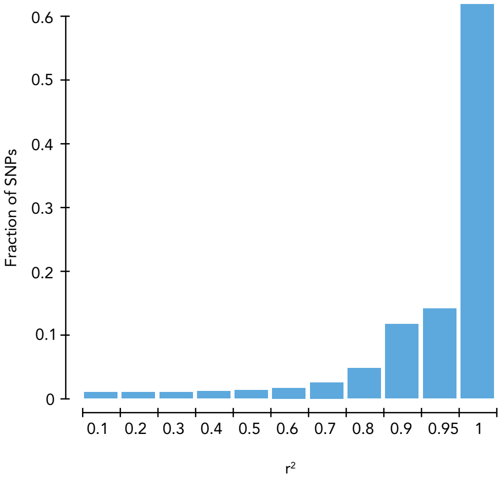
The time required to impute the CD extension data set scaled approximately linearly with the number of GWAS samples. For example, BEAGLE required approximately 2000 CPU-hours to impute the HapMap3 SNPs into the 4686 WTCCC CD samples and 622 hours to impute into the 1452 Belgian/French samples. Imputation can easily be parallelized across sections of chromosomes and subsets of the sample, but each sample subset must contain a consistent mixture of cases and controls[13] to avoid introducing differential bias in the imputed allele frequency estimates. High memory requirements (>8GB) can pose problems, but both IMPUTE v2 and BEAGLE have configurable trade offs between memory usage and processing time. With today's technology, genome-wide imputation cannot be carried out on a standard desktop computer. However, given that it must only be done once for a given experimental data set, and it adds considerable value to expensive GWAS data sets, imputation is a tractable task for groups with even modest computational resources.

Fast Imputation in a Meta-Analysis of Type 1 Diabetes

In a similar experiment, a GWAS of T1D involving 8000 UK samples was undertaken with the Illumina HumanHap550-Duo BeadChip. We chose the Illumina HumanHap550-Duo because it provided extremely high-quality genotypes and excellent coverage of common variation in European samples. A meta-analysis was then completed using these data and two previous GWAS run with the Affymetrix 500K chips. By using imputation, these data could be confidently integrated across platforms, allowing the selection of the higher coverage Illumina chip[16] for the second GWAS.

Approximately 1500 control samples from the 1958 British Birth Cohort overlapped between the individual GWAS samples and were genotyped on both chips. This set of samples allowed a much simpler imputation method to be employed, where linear regressions of nearby SNPs were used as predictors (implemented in snpMatrix[17]) and with similar accuracy to the HMM methods described above. Imputation accuracy was assessed in the 1958 BBC samples that were typed on both the Affymetrix 500K and Illumina HumanHap550-Duo platforms. The SNPs on the Affymetrix chip were used to impute SNPs on the Illumina chip, and vice versa. The predictions were then compared to the true genotypes on the unused chip in each case.

Figure 3: Imputation Delivers Highly Accurate Results



Imputation results were found to be extremely accurate for the majority of SNPs. Over 80% of imputed SNPs were found to have an r^2>0.9. r^2 is a measure of accuracy with 1 being the most accurate and 0 being inaccurate.

Association Analysis of Imputed Data

Regardless of what software or reference sets are used to generate imputed data, some care is required in the subsequent association analysis. While genotype platforms generally produce exact genotype calls (i.e. each individual is assigned genotype AA, AB, or BB), imputation programs generate probabilities for each of the three possible genotypes. The simplest means of analyzing this output is to take the 'best guess' genotypes (i.e. the genotype with the highest probability) and analyze in the classical fashion, but this approach will ultimately lose power and incorrectly weight different constituent scans in a meta-analysis because it does not account for the uncertainty in genotype assignment[10]. Luckily, nearly all genome-wide association packages (including snpMatrix[17], SNPTEST and PLINK v1.079) can analyze the genotype dosages (the expected number of copies of a specified allele, from 0 to 2) that are produced by imputation programs.

Results

For common variation, imputation results are extremely accurate and allow for seamless integration of data sets in meta-analyses. Figure 3 shows that nearly all SNPs with frequency >0.01 were predicted with high accuracy. Imputation based on Illumina HumanHap550-Duo genotypes was more accurate (87% $r^2 > 0.9$) than imputation from Affymetrix 500K genotypes (60% $r^2 > 0.9$). Better coverage of common variation and higher data quality contribute to this difference in accuracy; newer versions of these chips are likely to perform even better. The large fraction of accurately imputed variants enables powerful joint analysis of nearly all SNPs on any of the SNP chips used in the individual CD or T1D scans.

Joint association analysis of the imputed data sets did not substantially inflate the overall test statistics in either CD (λGC=1.16) or T1D (λGC=1.12), indicating that imputation did not induce any systematic biases. Across both studies, 25 associations which had been previously uncovered in the analyses of the constituent scans were confirmed, and a further 36 novel regions with meta-analysis $P<10^{-6}$ were taken forward to replication. Of these, 30 reached genome-wide significance after replication, and several others showed nominal evidence of replication, representing a doubling of confirmed loci via imputation based meta-analysis. The extremely high rate of replication of the significant initial findings implies that very few strong artifactual associations were introduced by imputation and underscores the reliability of the method.

Discussion

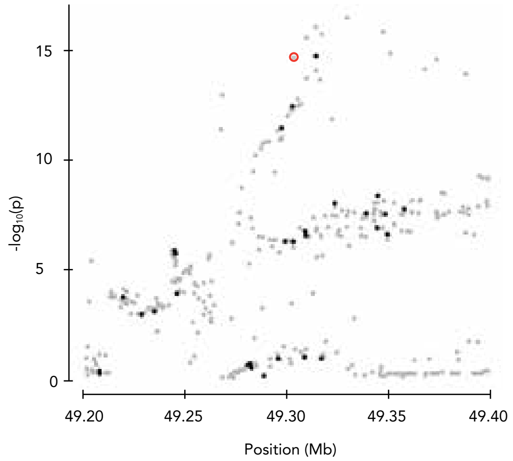
Genotype imputation is becoming a de rigeur part of GWAS, and it has been used in meta-analyses of many different diseases and traits. Individual GWAS are rarely large enough to discover associations with odds ratios <1.2 (Figure 2). By combining data sets from multiple complementary studies, researchers can increase the sample size of their analysis to uncover such associations. However, without imputation, the SNP sets from disparate chips act as an impediment to combining data sets or to adopting new platforms. In the studies described within this application note, the increase in power from using larger meta-analysis sample sizes doubled the number of identified loci in both CD and T1D. The high accuracy and reliability of current imputation methods have led to widespread acceptance and have established the technique as a standard practice for the analysis of disease genetics data sets. As a testament to the extent of imputation adoption, the manuscripts describing popular methods such as IMPUTE and BEAGLE have been cited hundreds of times throughout the literature.

The rapid pace of technological development in genome-wide SNP chips has spured human disease genetics research, with the density of chips increasing by an order of magnitude in the last few years. At first glance, this presents a challenge to the analysis of experiments that include data from current cutting-edge products together with data sets that, while only a few years old, were generated with a different version of a product or even a completely different platform. However, dozens of high-profile GWAS meta-analyses incorporating data from many different chips have all been enabled by imputation, and clearly demonstrate that previous GWAS need never be abandoned, but can be repeatedly drawn on in dissecting the genetic architecture of disease.

While imputation is already ubiquitous in analyses that combine data sets across different genotype platforms, it will become increasingly important for future analyses when reference sets with much larger numbers of SNPs (including a large proportion of rare variants) will be available. Because common SNPs are so well covered by the current generation of SNP chips, there have not been many examples of associations discovered via imputation within a single study[18] (as opposed to the meta-analyses discussed within this document, which combined information across studies). However, once projects like the 1000 Genomes generate reference sets with nearly all variants at >1% frequency (compared to the HapMap which is only complete to 5–10%), the usage of imputation within single studies will become much more prominent. For example, Figure 4 shows WTCCC CD genotypes and 1000 Genomes imputation at the well characterized *NOD2* locus, where two rare missense and a frameshift mutation have been shown to be causal[19], accounting for the signal at nearby

AAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAA
AATCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACC
AACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTA
TTAACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGA
AAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAA
AAGATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGA
AACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACA

Figure 4: Identification of a Rare Mutation by Imputation Analysis



Association results in WTCCC CD cases near the NOD2 gene. Black points represent genotyped markers, and grey points represent imputation from the 1000 Genomes Project. The point highlighted in red is rs2066844, a known functional polymorphism[19].

common SNPs. While GWAS capture an extremely strong signal at this locus, they give little insight into the underlying causal mutations. However, the imputed data immediately identify one of the missense mutations being among the strongest local associations. Given its known function, the identified association would make this mutation an attractive candidate for further study. Interestingly, the other two causal mutations are too rare to be observed in this early release of 1000 Genomes data, and highlight the possibility for greater discoveries once the complete data are available.

The future utility of genome-wide imputation will rest on parallel applications: keeping archival GWAS current by allowing them to be integrated into new studies, and maximizing the power of GWAS to begin to test for association to rare variations. These methods have the potential to trigger a renaissance of GWAS discoveries as the 1000 Genomes Project releases higher quality data on more samples and SNPs (as well as indels and larger copy-number polymorphisms). Through meta-analysis across disparate genotype platforms and in the application of exciting new reference sets, imputation allows researchers to probe more deeply into the allelic architecture of disease.

Imputation Web Resources

Online Resource	Web Address
HapMap Project	http://www.hapmap.org
1000 Genomes Project	http://1000genomes.org
BEAGLE	http://www.stat.auckland.ac.nz/~bbrowning/beagle/beagle.html
IMPUTE	http://mathgen.stats.ox.ac.uk/impute/impute.html
MACH	http://www.sph.umich.edu/csg/abecasis/mach
snpMatrix	http://www.bioconductor.org/packages/bioc/html/snpMatrix.html
PLINK	http://pngu.mgh.harvard.edu/purcell/plink

Acknowledgements

I would like to thank all the members of the International IBD Genetics Consortium, and the Type 1 Diabetes Genetics Consortium, with whom the meta analyses were conducted, and Carl Anderson and Kate Morley for comments on this manuscript.

References

1. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP , et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. U.S.A., 106: 9362–9367.
2. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature, 461: 747–753.
3. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, et al. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat. Genet., 40: 955–962.
4. Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, et al. (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat. Genet.
5. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science, 324: 387–389.
6. Libioulle C, Louis E, Hansoul S, Sandor C, Farnir F, et al. (2007) Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. PLoS Genet, 3(4): e58.
7. Rioux JD, Xavier RJ, Taylor KD, Silverberg MS, Goyette P, et al. (2007) Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat Genet, 39(5): 596–604.
8. Wellcome Trust Case Control Consortium. (2007) A Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature, 447(7145): 661–78.
9. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet, 81(3): 559–75.
10. de Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, et al. (2008) Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum. Mol. Genet., 17: R122–128.
11. Li Y, Abecasis GR. (2006) Mach 1.0: Rapid haplotype reconstruction and missing genotype inference. Am J Hum Genet, S79: 2290.
12. Marchini J, Howie B, Myers S, McVean G, and Donnelly P. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet, 39(7): 906–13.
13. 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.
14. Ellinghaus D, Schreiber S, Franke A, Nothnagel M. (2009) Current software for genotype imputation. Hum. Genomics, 3: 371–380.
15. 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.
16. Barrett JC, Cardon LR. (2006) Evaluating coverage of genome-wide association studies. Nat Genet, 38(6): 659–62.
17. Clayton D, Leung HT. (2007) An R package for analysis of whole-genome association studies. Hum. Hered., 64: 45–51.
18. Anderson CA, Pettersson GH, Barrett JC, Zhuang JJ, Ragoussis J, et al. (2008) Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. Am. J. Hum. Genet., 83: 112–119.
19. Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, et al. (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature, 411(6837): 599–603.

Illumina, Inc. • 9885 Towne Centre Drive, San Diego, CA 92121 USA • 1.800.809.4566 toll-free • 1.858.202.4566 tel • techsupport@illumina.com • illumina.com

FOR RESEARCH USE ONLY