illumina®

Imputation Estimates Genotypes at Un-Genotyped Loci

Imputation algorithms enable genotype data estimation between marker sets with different content using the inherent correlation of SNPs in linkage disequilibrium (LD) haplotype blocks.

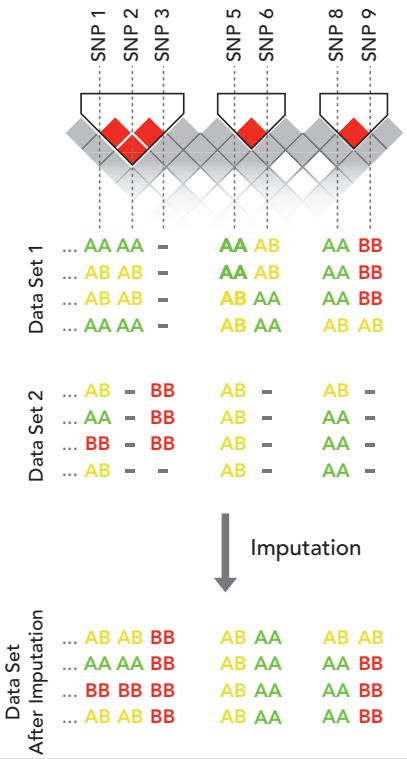
Introduction

The catalog of human genetic variation has been rapidly growing over the past few years with collaborative efforts such as the International HapMap Project. The rate of discovery about human genetic diversity will not slow any time soon, as efforts such as the 1,000 Genomes Project continue to deposit sequence data into the public domain at an unprecedented rate.

This expansion of knowledge has spurred rapid innovation in high-throughput genotyping technologies over the past few years. Today, a researcher can assay almost five million data points in a single microarray experiment. Studies of human variation and its contribution to disease have spanned many years, and microarray technology and products are evolving quickly. Thus, in the course of a genetic study, researchers collect data using different generations of microarrays in an effort to access the latest content.

When a data set is collected using two or more array types with different marker sets, some markers will not be assayed across the entire data set. This limits the total sample size for association analysis at these markers to the fraction of samples directly genotyped on the array that carried the marker of interest. Limiting the sample sizes for analysis of a marker effectively limits the study's power to detect true associations. However, recent computational advances enable researchers to use algorithms to fill in, or impute, genotypes at the markers that are not common between two genotyping arrays. In effect, imputation increases the sample size at each marker to the total number of unique individuals genotyped across the entire study (Figure 1).

Figure 1: Imputation Overview



SNPs 1–9 form three blocks of high LD, indicated by the red diamonds between the SNPs. Data Sets 1 and 2 represent a total of eight individuals genotyped using two different arrays at SNPs 1–9. The imputed data set contains genotypes for all SNP loci, with estimated genotypes filling in the missing data from Data Set 2. For example, SNP 2 is genotyped in Data Set 1 but not Data Set 2. Due to strong LD between SNPs 1–3, the individual genotypes for SNP 2 can be inferrred in Data Set 2 based on those present in Data Set 1.

Available Software

There are a number of freely available software programs that are widely used for imputation. All are command-line programs that run on Unix/Linux-based systems. Four of the most commonly used programs are listed in Table 1. User guides and tutorials are available from the respective websites for each program. Refer to the documentation of each program for instructions on download and use.

System Requirements

Imputation is a computationally intense process. A researcher interested in imputation typically needs access to the large Unix- or Linux-based clusters often available through the IT or Bioinformatics departments of research institutions. For the fastest analysis, the cluster should allow access to multiple computing nodes at a time.

Computational requirements can be reduced by performing imputation chromosome-by-chromosome or using only a few hundred samples at a time. These piecewise imputed data sets can be merged before analysis. When time or computing power is limited, researchers can focus exclusively on a single chromosome or region of interest, which will minimize the resources necessary for imputation.

Planning Considerations

Reference Population

A reference population with very dense genotyping can be used as a scaffold to align data from different experiments for imputation. The reference data provides a denser set of markers with minor allele frequency and LD information that can be used to inform imputation across data sets. The reference population should be representative of the experimental sample population[9]. For example, if the experimental data were collected from individuals of Caucasian ancestry, then a Caucasian reference sample (e.g. HapMap CEU samples) should also be used. Likewise, for samples of mixed or alternative ancestry, an appropriate reference sample should be used. Huang et al. present the optimal proportions of CEU/CHB/JPT/YRI HapMap samples for imputing diverse world populations[10].

Consistent Strand

When merging data sets, it is essential that genotypes from both data sets are presented consistently from the same strand (e.g., forward or "+" strand). Errors in this consistency will result in an inability to merge data and cryptic strand flips of A/T and C/G SNPs, which can lead to spurious results[7].

HapMap reference data are provided from a number of sources (see the Reference Data Sets section) on the forward strand. Beginning with the Infinium® HD HumanOmni1-Quad BeadChip, Illumina will provide strand annotation files for all its products, which researchers can obtain by contacting Technical Support. These strand annotation files can be used to identify markers assayed on the reverse strand. Researchers can flip reverse strand markers using a program such as PLINK before merging with reference data for imputation.

However, there are suspected strand errors in the HapMap data, so researchers should expect that a few markers (usually not more than a few thousand out of a whole-genome data set) display irreconcilable strand differences. Illumina recommends removing those SNPs from the experimental data set and proceeding with imputation.

Initial Quality Control

Before genotype imputation, Illumina recommends that researchers carry out basic data quality checks on available genotypes in the experimental data set. This generally includes removal of[7]:

- Markers with low call rate
- Large deviations from Hardy-Weinberg equilibrium
- Large numbers of discrepancies among duplicate samples
- Mendelian inconsistencies
- Markers with very low minor allele frequency (MAF)

Optimal cutoffs for these metrics vary, and researchers should use appropriate scores for their particular study.

Confidence Threshold for "Hard" Genotype Calls

When using imputation to facilitate the merging of data sets from different sources for a combined analysis, the goal is to fill in missing genotypes across two different data sets. Genotypes can be called at each missing data position for later association analysis.

Imputation provides a probability for each of the three possible genotype classes, and calls are based on the most likely genotype at each position[9]. When a hard genotype call is made, it carries with it a confidence score that corresponds to the likelihood that the called genotype was the correct choice. For example, if the genotype AA had a probability of 95% versus the genotype of AB having a probability of 3%, the confidence score for the choice of AA would reflect the overwhelming likelihood that the true genotype is AA. If the probability of AA was 40% and the probability of AB was 30%, making a hard genotype call is not as clear-cut and would be reflected in a lower confidence score. Imposing a stringent cutoff based on confidence scores will decrease the likelihood of imputation error in downstream association testing[9]. Refer to the accompanying documentation for each type of imputation software for more information on interpreting confidence scores.

Imputation Accuracy

In addition to imposing a stringent confidence score cutoff, several strategies can be employed to minimize potential errors in imputation. Watching out for these red flags will reduce the chances of inaccurate data interpretation due to imputation error.

Imputation is based on LD, so it will not predict completely independent regions of the genome. Association tests of flanking markers should show similar levels of association compared with an imputed marker. Therefore, an imputed marker with a dramatically different association statistic than the surrounding directly genotyped markers should be treated with caution and investigated carefully. An exception to this could occur when large amounts of data are missing in two or more data sets, which are merged with a reference data set. For example, if 50% of the data was missing at one SNP and 50% of the data was missing at a second neighboring SNP, then after imputation there would be nearly 100% of the genotypes for those markers across all data sets. This imputation scenario would provide such a

Table 1: Commonly Used Imputation Software Packages

Software Name	Institution	URL
Mach	University of Michigan[1,2]	http://www.sph.umich.edu/csg/abecasis/MaCH/tour/imputation.html
Beagle	University of Auckland[3]	http://faculty.washington.edu/browning/beagle/beagle.html
Impute	Oxford University[4,5]	http://mathgen.stats.ox.ac.uk/impute/impute.html
Plink	Massachusetts General Hospital / Broad Institute[6]	http://pngu.mgh.harvard.edu/~purcell/plink/

dramatic increase in power due to the increased total samples genotyped at each SNP that new associations could be discovered that were not seen in any of the individual data sets alone.

Although all imputation software uses the same fundamental phenomenon of LD across the genome, the algorithms employed by each software package differ. Likewise, each package offers differing strengths and weaknesses. Therefore, it is a good idea to use more than one software package, compare results, and investigate any major discrepancies.

Because there will always be some residual amount of error after calling genotypes with imputation, a good practice is to perform a relatively small amountof genotyping to confirm genotype calls when top association signals include imputed data. This step may include only a handful of SNPs genotyped in a few hundred individuals to confirm the imputation accuracy in these individuals for the markers of interest. Accuracy can then be extrapolated to the whole data set.

Imputation for Error-Checking Genotypes

Imputation can also be used to find potential errors in genotyping. The software package PLINK has a "drop-one" option that drops genotyped markers one-by-one across the genome, re-imputes the data, and performs an association test in real time[6]. There should be good consistency between the association statistic calculated with direct genotyping data and imputed genotyping data. When the imputed data have high confidence, large discrepancies can highlight suspected genotype errors that may have led to systematic bias.

Reference Data Sets

Most studies have successfully employed the standard HapMap panel of 60 unrelated individuals as the reference data set[3]. HapMap reference data sets in appropriate formats for use by the imputation software can be downloaded from many different web sites, including:

- http://mathgen.stats.ox.ac.uk/impute/impute.html
- http://pngu.mgh.harvard.edu/~purcell/plink/res.shtml

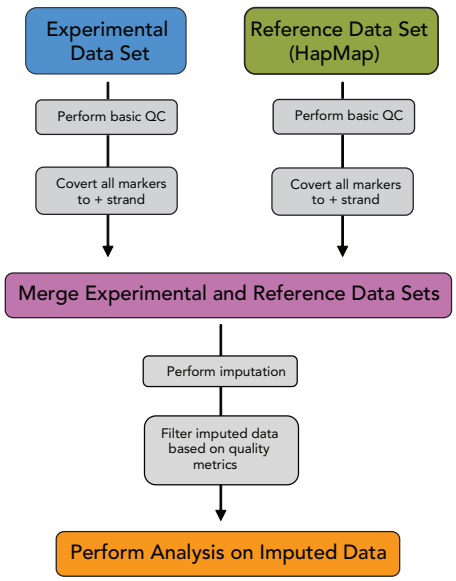
Examples Publications

Many large meta-analyses published in 2008 and 2009 were made possible due to the advancement of imputation techniques in recent years. These studies have combined data collected from many different platforms and from many different laboratories, then used imputation to fill in data that was missing due to differences in SNP content. These meta-analysis studies included tens of thousands of samples and used imputation to provide a dramatic boost in power for detecting a large number of additional associated loci. For examples of successful imputation strategies, please see the Reference Section in this document.

Imputation Using the HumanOmni1-Quad BeadChip

Illumina scientists performed the following analysis to estimate the imputation efficiency of the HumanOmni1-Quad BeadChip to other Infinium BeadChips (Table 2). HumanOmni1-Quad data were collected on 210 unrelated HapMap samples (60 YRI samples, 60 CEU samples, 90 CHB/JPT samples). The number of samples for each population was divided in half so that one half was retained for analysis—this is the Experimental Data Set (Figure 2).

Figure 2: Imputation Workflow



HapMap[2] data were downloaded from the PLINK website for the same 210 samples used to generate the HumanOmni1-Quad data. This sample set was segmented by population and the number of samples for each population was divided in half. One half was retained for analysis—this is the Reference Data Set. While the Experimental Data Set and the Reference Data Set were derived from the same 210 HapMap samples, the portion of samples retained for each data set was

Table 2: Percentage Markers on the Humanomni1-Quad Also Present on Other Infinium Beadchips

	Human1M-Duo	Human660-Quad	Human610-Quad	Human CytoSnp-12	HumanHap550	Human-hap370
Total overlap with HumanOmni1-Quad	53.25%	58.60%	59.62%	77.20%	59.54%	60.82%

The figures represent the percentage of markers available on various Infinium BeadChips that are also present on the HumanOmni1-Quad BeadChip prior to imputation.

composed of unique individuals (i.e., no individual sample appeared in both data sets).

The two data sets were then merged in a population-specific manner for imputation analysis. Those markers present in the HapMap data that were not included in the HumanOmni1-Quad data were imputed to the Experimental Data Set. For each Infinium BeadChip being evaluated, the number of imputed markers with high quality scores was taken to estimate the number of markers reclaimed by imputation. Following the imputation process, there was a significant increase in the percentage of makers that could be reclaimed from other Infinium BeadChips when starting HumanOmni1-Quad data. These results are shown in Table 3.

Conclusion

Imputation is an important and valuable method for maximizing the available information when combining genotyping data sets generated using different marker sets. By using LD and the inherent correlation between genotypes in a reference data set, the genotypes for missing markers in a data set can be confidently inferred. Several software packages are available to perform imputation, and many references describe imputation for combined analysis studies.

Table 3: Percentage Markers on the Humanomni1-Quad Also Present on Other Infinium Beadchips After Imputation Analysis*

Population	Human1M-Duo	Human660-Quad	Human610-Quad	Human-CytoSnp-12	HumanHap550	Human-hap370
CEU	73.77%	83.01%	84.31%	82.20%	86.43%	87.56%
JPT/CHB	74.29%	82.47%	83.75%	82.41%	85.76%	85.98%
YRI	72.41%	78.50%	79.85%	81.94%	81.56%	80.88%

References (Imputation algorithms)

1. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, et al. (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet 40:161–9.
2. Sanna S, Jackson AU, Nagaraja R, Willer CJ, Chen WM, et al. (2008) Common variants in the GDF5-UQCC region are associated with variation in human height. Nat Genet 40:198–203.
3. Browning BL and 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.
4. Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies via imputation of genotypes. Nat Genet 39: 906-913.
5. Howie B, 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): e100052.
6. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet, 81.
7. 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(R2): R122–8.
8. Pei YF, Li J, Zhang L, Papasian CJ, Deng HW. (2008) Analyses and comparison of accuracy of different genotype imputation methods. PLoS ONE 3(10): e3551.
9. Guan Y, Stephens M. (2008) Practical issues in imputation-based association mapping. PLoS Genet 4(12): e1000279.
10. Huang L, Li Y, Singleton AB, Hardy JA, Abecasis G, et al. (2009) Genotype-imputation accuracy across worldwide human populations. Am J Hum Genet. 84(2): 235–50.
11. Spencer CC, Su Z, Donnelly P, Marchini J. (2009) Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genet. 5(5): e1000477.
12. Padilla MA, Divers J, Vaughan LK, Allison DB, Tiwari HK. (2009) Multiple imputation to correct for measurement error in admixture estimates in genetic structured association testing. Hum Hered. 68(1): 65–72.
13. Becker T, Flaquer A, Brockschmidt FF, Herold C, Steffens M. (2009) Evaluation of potential power gain with imputed genotypes in genome-wide association studies.Hum Hered. 68(1): 23–34.
14. References (Successful Use of Imputation)
15. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, et al. (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet 40:161–169.
16. Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, et al. (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet 40:638–645.
17. Barrett JC, Hansoul S, Nicolae DL, Cho J, Duerr RH, et al. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 40:955–962.
18. Willer CJ, Speliotes EK, Loos RJF, Li S, Lindgren CM, et al. (2008) Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nat Genet 41:25–34.
19. Newton-Cheh C, Eijgelsheim M, Rice KM, de Bakker PIW, Yin X, et al. (2009) Common variants at ten loci influence QT interval duration in the QTGEN Study. Nat Genet 41:399–406.
20. Cho YS, Go MJ, Kim YJ, Heo JY, Oh JH, et al. (2009) A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. Nat Genet 41:527–534.
21. 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 May 10 [ePub ahead of print].
22. Newton-Cheh C, Johnson T, Gateva V, Tobin MD, Bochud M, et al. (2009) Genome-wide association study identifies eight loci associated with blood pressure. Nat Genet 41:666–676.
23. De Jager PL, Jia X, Wang J, deBakker PI, Ottoboni L, Aggarwal NT, et al. (2009) Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. Nat Genet 41:776–782.

Additional information

For more information about Illumina DNA Analysis tools, please visit www.illumina.com or contact us at the address below.

Illumina, Inc. • 1.800.809.4566 toll-free (U.S.) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

FOR RESEARCH USE ONLY

Pub. No. 370-2009-014 Current as of 24 June 2013

illumina®