Genotype SNP Imputation Methods Manual, Version 0.1.0 (May 14, 2010)

Practical and methodological considerations with three SNP genotype imputation software programs

Available online at: www.ImmPort.org

David L Morris[1], Patricia P Ramsay[2], Kim E Taylor[3], Lindsey A Criswell[3], Tim J Vyse[1], Glenys Thomson[4], Lisa F. Barcellos[2]

[1] Rheumatology Section, Faculty of Medicine, Imperial College, London, W12 0NN, UK, e-mails: david.morris@imperial.ac.uk, t.vyse@imperial.ac.uk

[2] Division of Epidemiology, School of Public Health, University of California, Berkeley, CA 94720-7356, USA, e-mails: pramsay@genepi.berkeley.edu, barcello@genepi.berkeley.edu

[3] Division of Rheumatology, University of California, San Francisco, CA 94143, USA, e-mails: ktaylor@medicine.ucsf.edu, lindsey.criswell@ucsf.edu

[4] Department of Integrative Biology, 3060 Valley Life Sciences Building MC #3140, University of California, Berkeley, CA 94720-3140, USA, e-mail: glenys@berkeley.edu

I. Overview

A. IMPUTE

B. MACH

C. BEAGLE

II. Methodological considerations

A. General
B. Population Structure
C. Accuracy
D. Which data to use? Imputed or typed?
E. Power

III. Quick start guides

A. IMPUTE2
    a. IMPUTE2 – Notes
B. BEAGLE
    a. BEAGLE – Notes
C. MACH
    a. MACH – Notes

IV. Imputation Example

I OVERVIEW

Several software programs are now available for the imputation of untyped SNPs or missing data in a genetic dataset. Imputation can be used to replace missing data when genotyping has failed in a percentage of the typed SNPs, to combine study populations that have been genotyped on different platforms, or to expand the coverage of SNPs in a certain population beyond what has been genotyped. Three of the more widely used programs are: BEAGLE v3.0.4 (Browning BL, 2009; Browning SR, 2007; Browning BL, 2007), IMPUTE v1 (Marchini J, 2007) v2 (Howie BN, 2009), and MACH v1.03 (Li 2006). All can be applied in a variety of situations, though some have attributes that make them suitable for one purpose over another.

Several articles have described comparisons of imputation methods with respect to computational efficiency and the accuracy of results (Pei YF, 2008; Yu Z, 2007; Nothnagel M, 2009). Overall, IMPUTE, MACH, and BEAGLE have been shown to have similar accuracy (given similar parameters), and all of these programs have been shown to outperform other methods for imputation such as fastPHASE (Scheet P, 2006) and PLINK (Purcell S, 2007).

It is possible to impute both un-typed SNPs (SNPs not typed in any samples) and missing data (typing failed for some individuals). Section IV has an example of a typical imputation setup. It is important to remember however that when imputing missing data, the genotypes for a SNP will be a mixture of calls and estimates (imputed). If the imputation is very accurate and the actual calls are not, or vice-versa, then this could lead to spurious associations. For each SNP it is therefore advisable to check for consistency when imputing missing data. Testing the allele frequency between imputed and actual is one approach but better would be to impute all (overwrite genotype calls and impute missing data) and compare this to the data from imputing just missing (actual genotype calls and imputed missing data). With respect to this issue, IMPUTE v2 gives the option of imputing just missing data or imputing all samples (overwrite genotype calls).

A IMPUTE

IMPUTE estimates unobserved genotypes in genome-wide case-control studies. The first version (v1) of IMPUTE employed a haplotype reference dataset only (such as HapMap (International HapMap Consortium, 2003) population haplotypes) to impute data for a supplied study dataset. Pre-formatted files for the Hap Map populations' haplotypes are provided on the IMPUTE website. Although reference panels for other populations may be used, creating one's own haplotype input files requires additional work (the phased data must be presented with one SNP per row and one Haplotype per column. Alleles in the Haplotype files are coded as 1 or 0 and a legend file must be constructed to show the allele codes and genome positions for each SNP). The latest version of IMPUTE (v2.1) allows the use of genotype reference files along with haplotype reference data. There is a

software package (GTOOL v0.5.0) available for converting files from ped format to IMPUTE format, however this can only be used with genotype files.

IMPUTE employs a Hidden Markov Model (HMM) method to compare the set of genotypes for each individual in the study dataset to the reference haplotypes/genotypes in order to resolve the untyped SNPs. A HMM comprises some unknown (hidden) states, such as haplotype backgrounds, and points in time (points in the genome (loci) for genetic data). The movement between states is a markov chain (the probability of belonging to state *j (j=1..k)* at time *i* depends on time *i-1)*. The k states are the k possible haplotype backgrounds (taken from the reference panel data), and the probability of moving between states is based on recombination rates. A HMM also has 'emission probabilities' which determine what an outcome could be (allele in our case) when in state *j* at point *i*. The Markov chain leads to points along the chain being more independent the further apart they are.

Cases and controls may be combined in the input files, and it has been shown (Howie BN, 2009) that using controls as reference genotypes and subsequently in a case/control analysis of the imputed data does not introduce spurious associations. This option will be useful when the cases have been typed on a lower density array while the controls come from a higher density and reliable source (say the WTCCC2 data). IMPUTE's computational intensity does require considerable time; large datasets may need to be broken down into subsets and then reconstructed, and in v1 chromosomes must be analyzed individually.

The program output contains a file of posterior probabilities of each potential genotype for each individual in the study dataset, as well as a file providing an estimate of the quality of imputation for each SNP. In the file of genotype probabilities there are three data points per SNP/individual (representing the probability of genotypes AA, Aa and aa, where a is the minor allele). As the output file format is not readily recognized by most standard genetic analysis software packages, the program GTOOL v0.5.0 can be used to transform the dataset into a typical linkage-style ped file. The user has flexibility when creating this file to include only the genotype probabilities that fall above a selected threshold (0.90 is the default). Genotypes below the selected threshold are set to missing. Alternatively, a case-control association analysis can be performed directly on the IMPUTE output files using the software package SNPTEST v2.1. SNPTEST provides several analysis options, some specifically designed for interpreting probabilistic data. Additionally, SNPTEST is capable of Bayesian analysis which is well suited to the uncertainly in imputation.

The latest version of IMPUTE is v2.1 (Howie BN, 2009), though v1 is still being supported. While similar in most ways to v1, v2 has been designed to allow increased options for the reference panel data. Populations other than those from HapMap can be more easily included as reference sets and datasets with multiple chromosomes need not necessarily be broken up. In addition, v2 allows the option of including a subset of the individuals in the study dataset as part of the reference set. In v2.0, the subset of the study data should be genotyped at least to the same density as the larger study dataset. IMPUTE

v2.0 has a strict hierarchical design, where SNPs in the inference data must be a subset of the SNPs within the reference genotypes and these reference genotype SNPs must be a subset of the SNPs within the reference haplotypes. Any SNPs that fail this hierarchy are removed from the analysis. Version 2.1, however, has overcome this limitation by using SNPs in the inference data to make inference on SNPs that are not in one of the reference panels. (This feature can be turned off if the user does not want to 'fill in' the reference data.)

B MACH

The imputation module of MACH v1.0 imputes unobserved genotypes primarily for case-control studies. It also employs a reference dataset to impute data for the study dataset of interest, using a Hidden Markov Model approach. The user can provide any phased population dataset as a reference set, including the HapMap. In addition, users can employ MACH's haplotyping module to prepare a phased reference panel from an unphased panel. Input files required for MACH are similar to those required for Merlin v1.12 (Abecasis GR, 2002).

The program outputs multiple files. It can output a file of the "best guess" genotypes (those with probability of 0.50 or greater). MACH also outputs a file that provides an estimated dosage of each imputed genotype along with the dosage of observed genotypes. The dosage is a range between 0-2, 0 representing no copies of the SNP reference allele, and 2 representing 2 copies of the reference allele. Therefore, while observed genotypes have discrete values included in [0 1 2], imputed genotypes are an estimate of the number of copies of the reference allele, represented by a decimal falling anywhere between 0-2. MACH also outputs a file of posterior probabilities and a file with quality scoring for each SNP. While the "best-guess" genotypes can be analyzed just as observed genotypes, with any standard statistical genetics software package, caution should be used in interpretation of results due to the uncertainty of the imputed data. The dosage file is particularly well-suited for analysis methods that take this into account, such as logistic and linear regression.

C BEAGLE

Earlier versions of the program BEAGLE relied on clustering methods for inference, without a reference set of known, phased haplotypes, and therefore when used for imputation were best employed for the inference of sporadic missing SNPs in a dataset of observed genotypes. The latest version, v3.0.4, employs a Hidden Markov Model method similar to IMPUTE and MACH and, with the use of a reference panel, can be

used to impute unobserved SNPs in a genome-wide study dataset. One unique feature of BEAGLE v3.0.4 is that it can handle datasets of both unrelated individuals and trio families. Parent-child relationships are taken into account and employed in the imputation methodology applied to trios.

Formatting of files for BEAGLE input can be made easier using the formatting module of WASPv0.82 (http://chgr.mc.vanderbilt.edu/wasp/), or the utility programs that are available with BEAGLE. BEAGLE output files include posterior probabilities as well as "best guess" inferred allelic genotypes. BEAGLE also contains a genome-wide association analysis module.

II METHODOLOGICAL CONSIDERATIONS

A General

There are several issues specific to the imputation of unobserved SNPs as well as the follow-up analysis of imputed data. One consideration to keep in mind is that the quality of the imputation is dependent on the quality of both the reference panel supplied (if used) and the set of observed genotypes in the study dataset. Standard quality control measures such as minor allele frequency, Hardy-Weinberg equilibrium, and genotype rate should be considered and SNPs falling beneath acceptable thresholds should be eliminated from the input files. The density of typed SNPs is also a consideration; as imputation methods necessarily rely on LD (correlation) among SNPs for inference, more densely typed SNPs will produce higher imputation accuracy.

Whenever analyzing more than one genetic dataset it is very important to ensure that the genotypes are read from the same strand of DNA. For example the same SNP may have alleles A/G in one dataset and T/C in the other because the first typed the forward strand and the last the reverse (or vice-versa). If this was the case then all observed T's in the second dataset should be changed to A's and C's to G's (Or the data in the first dataset should be changed). This is referred to as 'Strand Flipping'. This is not so straight forward, however, for A/T or C/G SNPs and in this case the minor allele frequency should be checked. For a C/G SNP (for example), if the minor allele is G in one dataset and C in another then the strand should be flipped in one of the datasets. Special care must be taken with A/T or C/G SNPs (ambiguous SNPs), however, as the observed MAF is a random variable (when looking at a sample of data) and when the observed MAF is close to 0.5 you can never be sure if flipping is required or not. It is advisable to check manufacturer's descriptions for strand and either correct for this or supply a file (IMPUTE accepts 'strand files') to the imputation algorithm.

Strand-flipping issues should be checked and resolved whenever imputation is used to combine datasets, especially if these have been typed on separate platforms. Strand-

flipping is also an issue when choosing a reference panel for the imputation; your data must be on the same strand as the reference panel data, or information on stranding needs to be supplied. Both IMPUTE and MACH have internal checks for handling this, though users may be responsible for specifying it as an option. BEAGLE will crash if any SNPs have strand errors, however there is an additional python script available as a utility for strand checking.

B POPULATION STRUCTURE

The imputation process uses either an external dataset to infer missing genotypes, or jointly models your data (learns from all observations within the data). Thus, it is important that the population structure within the data is known and correctly matched. The HapMap now has genotype data for various populations and it is important to match these to your data. If your dataset is a mixture of populations then it would not be advisable to impute together in one stage, unless you have good reason to believe that the haplotype structure is conserved across the populations for the region you are imputing.

C ACCURACY

It is important to remember that imputation is an estimation process and the imputed data should not be treated in the same manner as typed SNPs. Models for association studies are generally conditioned on known genotypes and it would be wrong to present such results on imputed data. There are methods for accounting for imputation uncertainty in further analysis; SNPTEST for example uses genotype probabilities to average over the uncertainty.

A very useful, and coherent, feature of IMPUTE is that it returns its imputed data as probabilities, rather than actual genotype calls. This naturally accounts for uncertainty which can be used in association analyses (default in SNPTEST). BEAGLE can also return genotype probabilities in the '.gprobs' file option, and MACH provides probabilities in the '.mlprob' file.

IMPUTE returns a general measure of imputation accuracy based on a cross validation procedure. This gives the user a measure of accuracy based on its predictive strength for the known genotype data which can be used to infer accuracy for the un-typed SNPs. MACH also allows for accuracy assessment using cross validation with its 'mask' function.

Planning the imputation to get the most accurate results is perhaps more important than using measures of accuracy after imputation. As mentioned above, the choice of reference data will affect accuracy, with respect to differing haplotype structures across

populations. However, regardless of population issues you may also have further options; the Hapmap has extensive coverage but the sample sizes are not very large. It may be better to use a reference dataset (perhaps from another study that you have performed) that has a larger sample size. This will improve imputation accuracy but with the trade off that you will have fewer imputed SNPs.

D Which data to use? Imputed or typed?

Q) For a particular SNP with a percentage of missing data, should you impute the missing genotype data and analyze this together with the true typed data?

A) This will depend on the balance between imputation accuracy and confidence in your data:

A SNP may have a lot of missing data and this could indicate that the called genotypes may not actually be very reliable. If you were to fill in the gaps with imputation you may get high quality imputed genotypes (based on the HapMap reference data, for example) mixed with low quality called genotypes and this may lead to false associations.
It is advisable to carefully compare (using allele frequencies) the imputed genotypes at a SNP with the genotypes you have at that SNP:

- If there is good agreement (allele frequencies in typed data agree with the imputed data) then it is probably acceptable to fill in the missing genotypes at that SNP with imputed genotypes.
- If there is not good agreement (allele frequencies in typed data disagree with imputed), and the imputed genotypes are very uncertain then it would be best not to fill in missing genotypes.
- If there is not good agreement and the imputed genotypes are certain (very high probabilities) then it is not clear what to do. Whether to use the genotypes from the assay or the ones from imputation or neither is a matter of quality control and it might be best to omit this SNP from further analysis.

Generally it is best only to impute sporadic missing data when you are confident about the genotypes you have for that SNP, otherwise it would be best to impute all samples for the SNP. Any bias in the genotype data (if it is not of high quality) may follow though to the imputation. In this case the reference data may do a better job of informing on all samples for the SNP.

E POWER

Reduction in power due to imputation uncertainty is an important consideration in any further analysis of the data. Imputed SNPs must be treated differently than typed SNPs in any upstream analysis, as the reduction in power affects how any resulting p-values should be treated. Imputed SNPs with poor accuracy should have lower p-value

thresholds than those with high accuracy, as a reduction in power lowers the posterior probability of the alternative hypothesis.

III QUICK START QUIDES:

A IMPUTE2.1

i) What do you want to impute?	**Commands**
- Only impute un-typed snps	-os 0 1 2 3
- Only impute missing data (Sporadic missing data)	-pgs_miss –os 3
- Impute both untyped SNPs and missing data	-pgs_miss –os 0 1 2 3

ii) Data required

Files required (* = optional)

1 Haplotype reference file
2 Genotype reference file*
3 Genotype-ref-Strand file*
4 Genotype data file
5 Genotype-data-strand file*

iii) Program parameters (default/recommended)

1	Effective population size (-Ne)	(10k – 15k)
2	Imputation region boundaries, in bp (-int): (lower and upper)	
3	Maximum imputation segment size, in bp:	
4	Interval end buffer, in kbp (-buffer):	(250)
5	Exclude SNPs from imputation (-exclude_snps):	
6	SNP types to be included in the output file (-os):	
	0 Only SNPs in haploid ref panel	
	1 SNPs in both the diploid and haploid ref panel	
	2 SNPs in both reference and inference data genotype files	
	3 SNPs only in the inference genotypes file	
7	Total number of MCMC iterations (-iter):	(30)
8	Number of (-iter) to discard as burn-in (-burnin):	(10)
9	Number of conditioning states (-k):	(40)
10	–fix_strand (check and flip strands (if no strand file is supplied)	

a) IMPUTE2 - NOTES:

1) Reference hapmap data can be downloaded from the IMPUTE website. This is conveniently on the + strand so you can either strand your genotype-ref and genotype-inference data file on the + strand or supply a strand file.

B BEAGLE

Formatting your data

Your data Files must be in BEAGLE format. The simplest way to achieve this is to have the data in .ped format and convert them to .bgl using the utility program 'linkage2beagle'
Reference data (phased Hapmap data, for example) must also be in beagle format. It is straight forward to convert phased haplotype data to beagle format using the utility program 'phased2beagle'.
Importantly, all your data files must be on the same strand (see note 1).

What do you want to impute?

- Sporadic missing genotype data?
    - Given a data set 'fileA.bgl' of unphased unrelated individuals, the code to impute missing data is:
        - 'java -Xmx1000m -jar beagle.jar unphased=fileA.bgl missing=?'
        - For trio data replace 'unphased=fileA.bgl 'with 'trios=fileA.bgl'
        - For parent offspring pair data replace 'unphased=fileA.bgl' with 'pairs=fileA.bgl'

- Un-typed SNPS?
    - Given a data set 'fileA.bgl' of unphased unrelated individuals and a reference phased dataset 'fileB.ble', the code is
        - java -Xmx1000m -jar beagle.jar unphased=fileA.bgl phased=fileB.bgl markers=markers.txt missing=?
        - For trio data replace 'unphased=fileA.bgl 'with 'trios=fileA.bgl'

a) BEAGLE - NOTES:

1) Your genotype data and reference data must be on the same strand. BEAGLE will not switch the alleles for any trivial SNPs and will not run if there are any. The BEAGLE website has a link to a python script which will check the strand of your files and switch alleles when appropriate.

C MACH

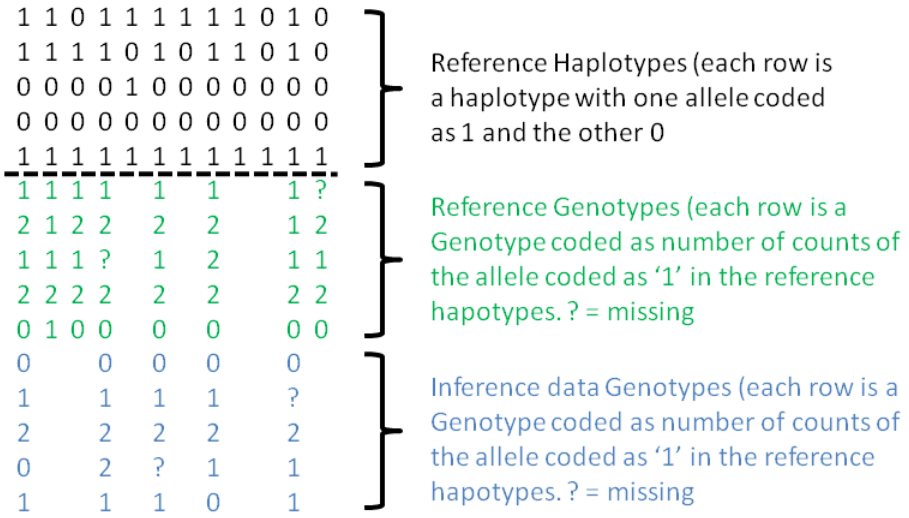
What do you want to impute?

- Sporadic missing genotype data?
    - Given a data set 'sample.ped' and a merlin format data file (map file) 'sample.dat', the code is:
        - 'mach –d sample.dat –p sample.ped –rounds 50 –greedy --geno'

- Un-typed SNPS?
    - Given a data set 'sample.ped', a merlin format data file 'sample.dat', a reference haplotype file 'hapmap.haplos' and a list of SNPs in the reference dataset 'hapmap.snps' , the code is:
        - 'mach –d sample.dat –p sample.ped –h hapmap.haplos –s hapmap.snps –rounds 50 –greedy --geno'

    Your also have the option of using reference genotype data in which case you merge this data with you own and treat the problem as if you had sporadic missing data.

IV IMPUTATION EXAMPLE:



Typical set up for an imputation study. The Blue data are study genotypes and the aim is to estimate genotypes up to the density of the reference haplotypes (in black). In some studies however the aim may be to impute only up to the density of some reference genotypes (in green). It is possible to have only haplotype reference data, genotype reference data or both.

Missing data (coded here as '?') can also be imputed, both in the inference data and reference data genotypes.

Acknowledgements

We would like to thank Bryan Howie and Jonathan Marchini for helpful comments.

References

Abecasis GR, Cherny SS, Cookson WO and Cardon LR (2002) Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet **30**:97-101.

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.

Browning SR and Browning BL (2007) Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. Am J Hum Genet 81:1084-1097.

Browning BL and Browning SR (2007) Efficient multilocus association mapping for whole genome association studies using localized haplotype clustering. Genet Epidemiol 31:365-375.

deBakker PI, Ferreira M, Jia X, Neale B, Raychaudurio S and Voight B (2008) Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum Mol Genet 17(R2):R122-8.

Howie BN, Donnelly P and Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genetics 5(6).

Li Y and Abecasis GR (2006) Mach 1.0: Rapid Haplotype Reconstruction and Missing Genotype Inference. Am J Hum Genet **S79** 2290.

Marchini J, Howie B, Myers S, McVean G and Donnelly P (2007) A new multipoint method for genome-wide association studies via imputation of genotypes. Nature Genetics 39: 906-913.

Nothnagel M, Ellinghaus D, Schreiber S, Krawczak M and Franke A (2009) A comprehensive evaluation of SNP genotype imputation . Hum Genet 125(2):163-71.

http://chgr.mc.vanderbilt.edu/wasp/

International HapMap Consortium (2003). The International HapMap Project. Nature 426(6968):789-96.

Pei YF, Li J, Zhang L, Papasian CJ and Deng HW (2008) Analyses and comparison of accuracy of different genotype imputation methods. PLoS ONE 3(10):e3551.
Yu Z and Schaid DJ (2007). Methods to impute missing genotypes for population data. Hum Genet 122(5): 495-504.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet 81(3): 559-75.

Scheet P and Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78(4): 629-44.