Current software for genotype imputation

David Ellinghaus,[1] Stefan Schreiber,[1] Andre Franke[1] and Michael Nothnagel[2]*

[1]Institute of Clinical Molecular Biology, Christian-Albrechts University, Kiel, Germany
[2]Institute of Medical Informatics and Statistics, Christian-Albrechts University, Kiel, Germany
**Correspondence to*: Tel: +49 (0) 431 597 1963; Fax: +49 (0) 431 597 1842; E-mail: d.ellinghaus@ikmb.uni-kiel.de

Date received: 11th April 2009

Abstract

Genotype imputation for single nucleotide polymorphisms (SNPs) has been shown to be a powerful means to include genetic markers in exploratory genetic association studies without having to genotype them, and is becoming a standard procedure. A number of different software programs are available. In our experience, user-friendliness is often the deciding factor in the choice of software to solve a particular task. We therefore evaluated the usability of three publicly available imputation programs: BEAGLE, IMPUTE and MACH. We found all three programs to perform well with HapMap reference data, with little effort needed for data preparation and subsequent association analysis. Each of them has different strengths and weaknesses, however, and none is optimal for all situations.

Keywords: *genotype imputation software, genome-wide association study, HapMap, single nucleotide polymorphism*

Introduction

Genotype imputation for single nucleotide polymorphisms (SNPs) has been shown to be a powerful means to include genetic markers in large-scale disease association studies without the need to actually genotype them.[1,2] Imputation is therefore becoming a standard procedure in exploratory genetic association studies. There are a number of areas in which imputation could possibly be beneficial. Imputation may single out an untyped SNP as having the strongest signal of association in a given region, with implications for the follow-up strategy. Case-control cohorts that were genotyped on different platforms (eg by Illumina and Affymetrix) can also be combined in a joint analysis (if the study design allows for such a combination) and for meta-analyses. Imputation might also serve as a means of quality control by highlighting likely genotyping errors. Finally, imputation could help in the reconstruction of missing genotypes in untyped family members in pedigree data.

The computations that underlie genotype imputation are based on a haplotype reference; the publicly available haplotype reference data that are provided by the International HapMap Project are usually employed.[3,4] More and different reference datasets can be expected in the future. In particular, the 1000 Genomes Project (http://www.1000genomes.org) will further augment our knowledge of the haplotype structure of SNPs with a minor allele frequency greater than one per cent.

A number of different software programs are available for genotype imputation, so the researcher must decide which program to use. Besides the accuracy and efficacy of imputation, which we addressed in a recent paper,[5] the best program for an interested user should be fast, easy to install and to handle, should have meaningful default options, checks for errors in the input and in the parameter

settings, and should feed back relevant information to the user. In our experience, user-friendliness is often the deciding factor in the choice of software to solve a particular task. We therefore evaluated the usability of three publicly available programs: BEAGLE,[6,7] IMPUTE[1] and MACH.[8] We did not consider PLINK[9] because of its inferior performance,[5] nor BIMBAM[2] or FAMHAP[10] because they were not included in our afore-mentioned benchmarking study. The URL and citations for the programs are listed in Table 1.

Implemented imputation methods

All three programs considered here make use of a hidden Markov model (HMM) to predict the missing genotypes of SNP markers. BEAGLE uses a localised haplotype-cluster model.[6] In this model, the reference haplotypes are grouped into clusters at each SNP. This allows for a reduction in complexity at different locations. IMPUTE and MACH implement variants of the 'product of approximate conditionals' (PAC) model.[11] The performance of both programs with regard to precision of prediction (accuracy) and efficacy is indistinguishable for populations that are well represented by HapMap. There are, however, subtle differences between the algorithms. For example, IMPUTE relies on user-specified recombination rates, whereas their estimation is part of the algorithm with MACH. Although the approach of IMPUTE may save computation time, it renders it sensitive to model misspecification.[12] This may be an important issue when imputation is carried out for populations that are less well represented by HapMap. All considered programs also differ in the methods used to infer haplotype phase and/or model recombination and mutation events. An insightful review of the underlying algorithms has been published recently.[12]

Table 1. List of programs that can be used for genotype imputation, together with citations and the websites from which they can be obtained. Only the first three programs were considered here

Software	Version	URL	Ref.
BEAGLE	3.0.2	http://www.stat.auckland.ac.nz/~browning/beagle/beagle.html	6,7
IMPUTE	0.5.0	http://www.stats.ox.ac.uk/~marchini/software/gwas/impute.html	1
MACH	1.10.16	http://www.sph.umich.edu/csg/abecasis/MaCH/	8
PLINK	1.05	http://pngu.mgh.harvard.edu/~purcell/plink/	9
BIMBAM	0.99	http://stephenslab.uchicago.edu/software.html	2
FAMHAP	18	http://famhap.meb.uni-bonn.de/	10

Criteria for software evaluation

We considered the most recent versions of each of the three programs for this evaluation: BEAGLE 3.0.2, IMPUTE 0.5.0 and MACH 1.10.16. Note that these versions differ from those considered in our recent benchmarking paper.[5] We grouped the considerable number of software features into five main groups, based on the following questions:

- **Accessibility**
  For which computer platforms are program binaries available? Are the sources available? What is the quality of the documentation? How responsive are the software authors to questions and requests? Is a graphical user interface (GUI) available?
- **Input**
  What is the workload for the preparation of datasets before imputation? How and in which format are the reference data available? Do the programs provide assistance in the preparation?
- **Processing**
  What are the memory demands and the runtime for the imputation? Can these demands be optimised? Are errors in the data well handled by the programs and do the programs give sufficient error feedback?
- **Output**
  What is the workload for analysing the

programs' outputs with other software? What type of information is reported, including posterior genotype probabilities and data summaries?

- **Miscellaneous**
  Are there any other special features of the programs?

A summary of the results is listed in Table 2.

Memory consumption and runtimes of the programs were assessed using an Affymetrix SNP array 6.0 (1000 k) dataset comprising 449 healthy blood donors of German decent. Further details on the dataset are described elsewhere.[5] The phased HapMap CEU haplotype data[13] (Centre d'Etudes du Polymorphisme Humain [CEPH]; Utah residents with ancestry from northern and western Europe) were used as the imputation reference. Annotation files from Affymetrix were used to map SNP markers to the forward strand. For illustrative purposes, analyses were limited to chromosome 6. For this chromosome, 43,265 SNPs with genotypes were available in our sample. The HapMap CEU reference contained 182,381 SNP markers.

Accessibility

Software platforms and licence

Many potential users will depend on pre-compiled, ready-to-use binaries for their respective platforms to carry out imputation. Platform incompatibilities can therefore represent a serious obstacle for the application of the programs. All three studied tools are offered as pre-compiled versions on their corresponding websites (see Table 1). BEAGLE is written in Java and thus runs on all major computing platforms with the appropriate java interpreter. IMPUTE binaries are available for all major platforms, including Linux, MacOS X, Solaris and Windows. MACH only supports Linux and MacOS X, but offers support for other platforms on demand. Unfortunately, source code is not available for any of the programs. MACH developers have announced that they will share their sources at some point in the future.

BEAGLE software is free, without any restrictions. IMPUTE remains the property of the University of Oxford and is distributed solely for non-commercial use. Software licensing for the MACH program is unclear, but the developers have announced that this will be resolved soon. Optionally, MACH and IMPUTE ask users to register for receiving electronic notifications about future software updates.

Documentation

The extent of software documentation varies considerably between the programs. IMPUTE and MACH provide the user with README files and short web tutorials on how to carry out the main task with the programs. There is a lack of documentation for the MACH program, so the average user is incapable of using MACH with all its features appropriately. By contrast, BEAGLE comes with comprehensive and informative documentation that contains many real-world examples of how to prepare input files, how to deal with memory management, how to handle data and how to analyse the output. Each software package includes example input files.

Graphical user interface and web service

All tools are executed through the command line — that is, graphical user interfaces (GUI) are not available. This is likely to represent a hurdle for some users interested in these programs, and we recommend that this should be addressed by future software development. Imputation with MACH can also be performed without a local installation, using an online version at the HapMap website (http://www.hapmap.org/); however, the online version imposes a limit of 5000 HapMap reference SNPs.

Authors' responsiveness

In our experience, getting help from the authors was not a problem. The response from the authors after e-mail requests was always quick, if sometimes lacking in detail. Brian Browning (co-author of

Table 2. Summary of a number of software features for three imputation programs. For more details, see the main text

Group	Feature	BEAGLE	IMPUTE	MACH
Accessibility	Operating system	Java (platform independent)	Linux, Windows, MacOS X, Solaris	Linux, MacOS X
	Licence	Free	Free for academic use	Not clear
	Source code	Not available	Not available	Availability announced
	Documentation	Commendable	Clearly structured	Incomplete
	Authors' response	Quick and detailed	Quick	Quick
	GUI	No	No	No
Input	Genotype format	Discrete; custom format	Probabilities; custom format	QTDT (Linkage)
	Reference format	Custom format	Custom format; prepared HapMap reference available	HapMap format (custom format)
	Conversion utilities	Yes	Yes	No
Processing	Target of imputation	Chromosomes	Chromosomes or segments	Chromosomes
	Memory-saving mode	Yes	No	Yes
	Known checking errors	None	Missing probability and input check	Problematic handling of missing reference
	Runtime [chr. 6]	350 minutes	433 minutes	2781 minutes
	Maximum memory allocation	2 GB	14 GB [<1 GB with ~10 MB segments]	7 GB
	Memory-saving mode	Yes	No	Yes
	Strand orientation	Check	Check + autoflip	Check + autoflip

Continued

Table 2. Continued

Group	Feature	BEAGLE	IMPUTE	MACH
Output	Genotypes	Posterior probability	Posterior probability	Posterior probability; allele dosage
	Quality measure for imputation	Allelic R^2	Information measure, average of the maximum posterior probabilities	R^2; average of the maximum posterior probabilities
	Output file size [chr. 6]	1476 MB	533 MB	938 MB
Miscellaneous	X chromosome imputation	No	Yes	No
	Trio data	Yes	No	No
	Multi-allelic markers	Yes	No	No
	Accuracy estimation	No	Yes	No

chr., chromosome.

BEAGLE) gave by far the most comprehensive and in-depth responses.

Input

Input data format

Many users with an interest in imputation will have only limited experience with scripting and programming of tasks. This includes format conversions before and after imputation, the latter being only the first step in a sequence of analyses. The use of (quasi-) standard data formats that can be generated and processed with a variety of software programs is therefore a prerequisite for widespread software use in genetic epidemiological studies. This includes the Linkage (pre-makeped; http://linkage.rockefeller.edu/soft/linkage/) format, HapMap format (http://www.hapmap.org/) and the increasingly popular PLINK format (http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml), which is a variant of the Linkage format.

Pedigree data

MACH requires to have the genotype data in QTDT format, which is similar to the Linkage format (http://www.sph.umich.edu/csg/abecasis/qtdt/docs/input.html). BEAGLE uses its own genotype data format, but provides java utilities for convenient transformation of phased and unphased data files from the QTDT format. IMPUTE's unique genotype format allows for genotype uncertainty and was designed to work seamlessly with other software tools from the University of Oxford. Thus, genotypes in QTDT or PLINK format have to be converted into genotype probabilities. This can be done with the GTOOL utility (http://www.stats.ox.ac.uk/~marchini/software/gwas/gtool.html).

Reference haplotype data

With MACH, the phased reference haplotypes files from HapMap Phase II (http://ftp.hapmap.org/phasing/) can immediately be used after download. IMPUTE uses its own reference format, but prepared HapMap Phase II reference data are available

for download from the IMPUTE website. BEAGLE also uses its own reference format, but provides the utility PHASED2BEAGLE within the BEAGLE package for converting the HapMap reference haplotypes files to this format.

Processing

Computational demands

Typically, genotype imputation runs are started on a per chromosome basis. Earlier versions of IMPUTE required a huge amount of working memory (RAM) for this task. Since version 0.4.0, it has been possible (and recommended) to carry out genome-wide imputation in chromosomal sub-regions, instead of imputing whole chromosomes. In order to do this, input files do not have to be split manually. Instead, the region of imputation can be specified by command line arguments, which is very convenient. There is also an additional option to avoid edge effects at the borders of the imputed sub-regions. Afterwards, the imputed sub-regions can be easily concatenated to generate imputed files for complete chromosomes, for example by using the 'cat' command under Linux or MacOS X, while redirecting the output into a text file. BEAGLE and MACH do not offer the imputation of particular chromosomal regions with a special treatment of region borders. Their memory requirements are much lower compared to IMPUTE, however, and they have implemented alternative algorithms which pass memory costs to runtime in order to reduce memory usage. The BEAGLE software also supplies the user with a tool and detailed instructions on how to divide the sample cohort (not the reference panel) into sub-samples and perform imputation on each sub-sample separately.

Memory consumption

In general, it is difficult for non-technical users to predict the working memory and runtime requirements for particular datasets. MACH and IMPUTE provide an estimation of memory allocation (main memory consumption) at runtime, so one should check the memory message while the programs start. BEAGLE does not show memory information while running, but devotes a short but informative chapter of its documentation to this problem. The main memory allocations for chromosome 6 did not exceed 2 gigabytes (GB) for BEAGLE (<1 GB in memory-saving mode), 14 GB for IMPUTE (<1 GB for each of the 18 chunks of ~10 megabases [Mb] size) and 7 GB for MACH (<1 GB in memory-saving mode), respectively.

Runtime

We used the data from chromosome 6 for illustrating the runtime differences between the programs. All programs ran on a single AMD-Shanghai 2.4 GHz processor machine, providing a maximum of 32 GB shared RAM, using the AMD64-variant of CentOS-5 (Linux distribution based on Red Hat Enterprise Linux) and the batch processing system PBSPro (Altair Engineering). BEAGLE's cumulative runtime was the shortest of all three programs (350 minutes; 366 minutes in memory-saving mode [5 per cent increase]). IMPUTE required a considerably longer time (433 minutes [24 per cent higher than that of BEAGLE]; 464 minutes when split into 18 chromosomal segments of ~10 Mb [7 per cent increase]), while MACH was by far the slowest program (2781 minutes [695 per cent higher than that of BEAGLE] — that is, about two days; 4421 minutes in memory-saving mode [59 per cent increase]).

Strand orientation

Strand orientation of the alleles has to be consistent between the observed genotypes and the haplotype reference data, which is the responsibility of the user. All three programs check for strand concordance, however, SNP markers with C/G and A/T alleles cannot be tested for orientation. BEAGLE automatically stops when strand errors occur. A python script from the author can be used to switch the respective alleles if necessary. IMPUTE and MACH can automatically flip SNP markers to the other allele, when called with an additional option. By default, IMPUTE drops erroneous markers, while MACH quits when strand errors occur. IMPUTE additionally provides the user with strand files for the

Affymetrix GeneChip 500K Mapping Array Set and SNP Array 6.0. When run with such a strand file, IMPUTE automatically flips SNP markers where necessary. With Illumina genotype data, which contains hardly any C/G and A/T SNP markers, the use of the auto-flip option with IMPUTE and MACH is sufficient for automatic correction.

Error handling

Proper error handling improves the usability of software enormously, since the user is not forced to investigate the sometimes lengthy process of error detection. Adequate handling of errors should include helpful error warnings and the reason(s) for program termination when fatal errors occur. In general, we found only a few errors that are, in our view, mishandled by the programs. The error handling of BEAGLE is exemplary; in our experience, the program always stopped with an appropriate error message when running with incorrect input. IMPUTE does not (sufficiently) check genotype probabilities, accepting negative values or those exceeding 1.0. Also, it does not terminate when the genotype input file cannot be found (eg due to an incorrect path or filename). Instead, IMPUTE enters an infinite loop, requiring a manual termination. If MACH is unable to find the reference input files, it gives a warning but does not terminate. Instead, MACH starts to infer haplotypes from the genotypes without any reference, resulting in the allocation of more than tenfold the amount of main memory usually used. In many instances, this will cause the computer to crash when the warning by MACH is overlooked by the user, which can easily happen when MACH runs in a batch processing system, as is generally the case for large computing clusters. We are, of course, aware that more errors might have escaped our attention, and this list of issues is not likely to be comprehensive.

Output

Posterior genotype probabilities

Imputed genotypes are predictions, not actual observations, as obtained from genotyping. Subsequent analyses, such as testing for phenotypic association, should incorporate the uncertainty of these predictions, to avoid spurious results. A general approach for this is the use of posterior probabilities for the imputed genotypes. A less general approach, albeit probably well suited for screening purposes, is to use allele dosage, which is defined as the estimated number of minor SNP alleles for a genotype. Imputation programs should always report values for at least one of these approaches. Fortunately, all three considered programs report the posterior probabilities of genotype calls. In the corresponding files, the posterior probability of observed genotypes (ie those without uncertainty due to imputation) will be 1.0. IMPUTE's genotype output file has exactly the same format as the input file. For each SNP and individual, the prior genotype probabilities have been replaced by the posterior probabilities after imputation. MACH and BEAGLE both have their own file format for the posterior probabilities, which are nevertheless very similar to that of IMPUTE. MACH also reports allele dosages.

Of note is the size of the posterior genotype probability files, which can be very large, depending on the number of imputed samples and SNPs. BEAGLE and MACH report these probabilities with a fixed accuracy of three decimal places. BEAGLE reports the posterior probabilities of all three possible SNP genotypes, resulting in a size of 1476 Mb (100 per cent) for chromosome 6. MACH reports only two of the three probabilities, reducing the size to 938 Mb (63.6 per cent). IMPUTE reports only two decimal places by default, although more can be specified through the respective command-line option, and no decimal places at all if the probability equals zero or one. The resulting output file had a size of 533 Mb (36.1 per cent). Due to the enormous redundancy in the output files, they can be compressed to a tenth of their size using gzip or similar utilities.

Prediction quality

Statistical predictions should always be accompanied by measures of their accuracy. IMPUTE generates an accuracy information file containing two measures. First, the confidence score, which is the average of the maximum posterior probabilities of the imputed genotypes for a SNP, and, second, the information

measure, a measure of the observed statistical information associated with the estimate of the allele frequency, are given for each imputed and non-imputed SNP. MACH also produces two slightly different estimates of imputation quality. Its quality score is identical to the confidence score of IMPUTE, while the ratio of the variances of the observed and the estimated allele counts is denoted by R^2 (also termed r^2 or OEvar). MACH uses this measure to assess the imputation performance of a SNP. BEAGLE assesses the quality of imputed genotypes by estimating allelic R^2 for each SNP, which is the squared correlation of the allele dosage with the highest posterior probability and the true allele dosage. As with MACH, this metric can be used to exclude SNPs with poor imputation accuracy.

Association testing

Genotype imputation is typically the first step for subsequent testing of phenotypic association in the exploratory, hypothesis-generating stage of a genetic epidemiological study. Each of the three programs considered here generates output that can readily be used by other programs for such an analysis. Output from BEAGLE can be used with PRESTO, which implements permutation testing of order statistics.[14] For IMPUTE, the SNPTEST tool (available from the IMPUTE website) allows for numerous association tests of binary and quantitative features, as well as for covariate adjustment and the calculation of Bayes factors. MACH has recently been complemented by the programs MACH2DAT and MACH2QTL (available from the MACH website), which can be used to test for association with binary and quantitative traits via regression models. The analysis of imputed genotypes with the software environment R,[15] (eg for model selection purposes) requires the split of the posterior probability files into smaller files of a few thousand SNPs per file for convenient use inside the R environment.

Miscellaneous

X chromosome imputation

While all considered programs can impute autosomal genotypes, IMPUTE is the only program so far that supports X chromosome imputation. It also properly handles the pseudoautosomal, as well as the non-pseudoautosomal regions of chromosome X. The gender of individuals (male/female) must be specified in an additional file. IMPUTE reports posterior genotype probabilities for females and allele probabilities for males. X chromosome imputation has not yet been implemented in BEAGLE or MACH. There will be an announcement that MACH is to incorporate X chromosome imputation in the next release (Y. Li, personal communication). BEAGLE can be expected to include this feature in the near future (B. Browning, personal communication).

Trio data

A unique feature of BEAGLE is its ability to process offspring-parent trio and offspring-parent pair data alone or in combination with unphased or phased data of unrelated individuals. At this point, MACH and IMPUTE can only handle data from unrelated individuals. MACH is expected to handle family data in the near future (Y. Li; personal communication).

Imputation of multi-allelic markers

So far, the principal targets for marker genotype information have been SNP markers. For most SNPs, only two common alleles exist, however, tri-allelic SNPs do exist and multi-allelic markers—such as short tandem repeats (microsatellites)—are still in use. An example is marker rs2032582,[16] which is listed as tri-allelic in the SNP database (dbSNP), but only as biallelic in HapMap. While IMPUTE and MACH can only handle biallelic markers, BEAGLE is able to process markers with up to 128 different alleles. This makes it the only imputation program suitable for multi-allelic markers.

Estimation of imputation accuracy

Only IMPUTE reports on the estimated distribution of imputation accuracy measures when quitting. IMPUTE implements a leave-one-out approach to estimate the prediction error for SNPs with observed genotypes.

'Correcting' genotypes

In some cases, MACH 'corrects' observed (ie genotyped) SNP genotypes based on the HapMap reference when used with the -mle option. The fraction of corrected genotypes is relatively low (in our experience, more than 99.7 per cent of observed genotypes remain untouched), but still significant. Correcting a highly likely genotyping error could be a rationale for this action, however, neither the feature nor the reasoning is documented. In our view, this lack of information should be addressed by the developers.

Large reference panels

Currently, most genotype imputation is performed using small reference panels from HapMap Phase II. This implies a limitation for imputation accuracy, particularly for markers with a low minor allele frequency. A considerable strength of BEAGLE is its flexibility and capability to use reference panels much larger than the 90 or 120 phased HapMap haplotypes used so far. BEAGLE's computational runtime depends on the combined size of the sample and the reference panel,[7] whereas computational times for MACH and IMPUTE are approximately linear with the size of the sample and quadratic with the size of the reference.[1,8] A comparison of BEAGLE and IMPUTE for increasingly large reference panels has been published recently.[7] Larger reference panels for many more populations are already available or will become so soon — for example, from the HapMap Phase III data and from the 1000 Genomes Project (10–15 Mio. SNPs estimated). For such large reference panels, BEAGLE might become the program of choice in the near future.

Conclusions

All three programs considered here perform well in imputing genotypes in populations well represented by HapMap. All three can be used with little effort required for data preparation and subsequent association analysis. Each of them has different strengths and weaknesses, however, and none is optimal for all situations. BEAGLE is slightly less accurate than IMPUTE and MACH, but is by far the fastest of the three and also very well documented. Only BEAGLE can handle multi-allelic markers. It may also be the most suitable program for use with the larger reference panels that will soon be available. IMPUTE and MACH are most accurate, but IMPUTE is embedded in a whole pipeline for data analysis and is much faster than MACH. So far, it is the only software among the three for X chromosome imputation. On the other hand, the long runtime of MACH is the price that comes with estimating the recombination rates from the dataset itself. This feature makes MACH less prone than IMPUTE to model misspecification in situations where the population sample to be imputed may not be well represented by the reference. MACH can be expected to perform better than IMPUTE under these conditions. BEAGLE and MACH are generally less memory-consuming than IMPUTE; however, the latter can be run for chromosomal subregions, which considerably reduces memory consumption, with an extra feature of avoiding border effects. All three programs would benefit from a graphical user interface, which would make them accessible to a wider range of users.

Acknowledgments

This work was supported by the German Ministry of Education and Research (BMBF) through a grant from the National Genome Research Network (NGFN). The project received infrastructure support through the DFG excellence cluster 'Inflammation at Interfaces'.

References

1. Marchini, J., Howie, B., Myers, S., McVean, G. *et al.* (2007), 'A new multipoint method for genome-wide association studies by imputation of genotypes', *Nat. Genet.* Vol. 39, pp. 906–913.
2. Servin, B. and Stephens, M. (2007), 'Imputation-based analysis of association studies: Candidate regions and quantitative traits', *PLoS Genet.* Vol. 3, p. e114.
3. The International HapMap Consortium (2003), 'The International HapMap Project', *Nature* Vol. 426, pp. 789–796.
4. The International HapMap Consortium (2005), 'A haplotype map of the human genome', *Nature* Vol. 437, pp. 1299–1320.
5. Nothnagel, M., Ellinghaus, D., Schreiber, S., Krawczak, M. *et al.* (2009), 'A comprehensive evaluation of SNP genotype imputation', *Hum. Genet.* Vol. 125, pp. 163–171.
6. Browning, S.R. and Browning, B.L. (2007), 'Rapid and accurate haplotype phasing and missing-data inference for whole-genome association

studies by use of localized haplotype clustering', *Am. J. Hum. Genet.* Vol. 81, pp. 1084–1097.
7. Browning, B.L. and Browning, S.R. (2009), 'A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals', *Am. J. Hum. Genet.* Vol. 84, pp. 210–223.
8. Li, Y. and Abecasis, G.R. (2006), 'Mach 1.0: Rapid haplotype reconstruction and missing genotype inference', *Am. J. Hum. Genet.* Vol. S79, p. 2290.
9. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L. *et al.* (2007), 'PLINK: A tool set for whole-genome association and population-based linkage analyses', *Am. J. Hum. Genet.* Vol. 81, pp. 559–575.
10. Becker, T. and Knapp, M. (2004), 'Maximum-likelihood estimation of haplotype frequencies in nuclear families', *Genet. Epidemiol.* Vol. 27, pp. 21–32.
11. Li, N. and Stephens, M. (2003), 'Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data', *Genetics* Vol. 165, pp. 2213–2233.
12. Browning, S.R. (2008), 'Missing data imputation and haplotype phase inference for genome-wide association studies', *Hum. Genet.* Vol. 124, pp. 439–450.
13. Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A. *et al.* (2007), 'A second generation human haplotype map of over 3.1 million SNPs', *Nature* Vol. 449, pp. 851–861.
14. Browning, B.L. (2008), 'PRESTO: Rapid calculation of order statistic distributions and multiple-testing adjusted P-values via permutation for one and two-stage genetic association studies', *BMC Bioinformatics*, Vol. 9, p. 309.
15. R Development Core Team (2008), 'R: A language and environment for statistical computings', Vienna, Austria. Available at http://www.R-project.org (accessed 17th June 2009).
16. Huebner, C., Petermann, I., Browning, B.L., Shelling, A.N. *et al.* (2007), 'Triallelic single nucleotide polymorphisms and genotyping error in genetic epidemiology studies: MDR1 (ABCB1) G2677/T/A as an example', *Cancer Epidemiol. Biomarkers Prev.* Vol. 16, pp. 1185–1192.