Hum Genet (2009) 125:163–171
DOI 10.1007/s00439-008-0606-5

ORIGINAL INVESTIGATION

A comprehensive evaluation of SNP genotype imputation

Michael Nothnagel · David Ellinghaus · Stefan Schreiber · Michael Krawczak · Andre Franke

Received: 7 October 2008 / Accepted: 5 December 2008 / Published online: 17 December 2008

Abstract Genome-wide association studies have contributed significantly to the genetic dissection of complex diseases. In order to increase the power of existing marker sets even further, methods have been proposed to predict individual genotypes at un-typed loci from other marker sets by imputation, usually employing HapMap data as a reference. Although various imputation algorithms have been used in practice already, a comprehensive evaluation and comparison of these approaches, using genome-wide SNP data from one and the same population is still lacking. We therefore investigated four publicly available programs for genotype imputation (BEAGLE, IMPUTE, MACH, and PLINK) using data from 449 German individuals genotyped in our laboratory for three genome-wide SNP sets [Affymetrix 5.0 (500 k), Affymetrix 6.0 (1,000 k), and Illumina 550 k]. We observed that HapMap-based imputation in a northern European population is powerful and reliable, even in highly variable genomic regions such as the extended MHC on chromosome *6p21*. However, while genotype predictions were found to be highly accurate with all four programs, the number of SNPs for which imputation was actually carried out ('imputation efficacy') varied substantially. BEAGLE, IMPUTE, and MACH yielded nearly identical trade-offs between imputation accuracy and efficacy whereas PLINK performed consistently poorer. We nevertheless recommend either MACH or BEAGLE for practical use because these two programs are more user-friendly and generally require less memory than IMPUTE.

M. Nothnagel and D. Ellinghaus contributed equally to the manuscript.
M. Krawczak and A. Franke shared senior authorship.

Electronic supplementary material The online version of this article (doi:10.1007/s00439-008-0606-5) contains supplementary material, which is available to authorized users.

M. Nothnagel · M. Krawczak
Institute of Medical Informatics and Statistics,
Christian-Albrechts University, Kiel, Germany

D. Ellinghaus · S. Schreiber · A. Franke (✉)
Institute of Clinical Molecular Biology,
Christian-Albrechts University, Campus Kiel, House 6,
Arnold-Heller-Str. 3, 24105 Kiel, Germany
e-mail: a.franke@mucosa.de

S. Schreiber · M. Krawczak
PopGen Biobank, Christian-Albrechts University,
Kiel, Germany

Introduction

Imputation of single nucleotide polymorphism (SNP) genotypes has been proposed as a powerful means to include genetic markers into large-scale disease association studies without a need to actually genotype them (Marchini et al. 2007; Servin and Stephens 2007). In fact, multi-locus analyses using a combination of imputed and observed genotypes appear to facilitate the detection of rare causative variants (population frequency <5%) that would otherwise be overlooked (Browning and Browning 2008). The underlying computations are usually based upon those 90–120 SNP haplotypes that are provided for each of four exemplary populations by the International HapMap project (The International HapMap Consortium 2003, 2005). Although imputed genotypes have already been used in practice (Wellcome Trust Case Control Consortium 2007), an independent genome-wide validation of the approach and a comprehensive comparison of the available algorithms are still lacking. For example, Marchini et al. (2007) based the

benchmarking of their algorithm, implemented in the computer program IMPUTE, on only 10,180 coding SNPs, and a recent genome-wide investigation of imputation performance (Anderson et al. 2008) exclusively used IMPUTE. Pei et al. (2008) compared five programs using both simulated and HapMap CEU phasing data, but only small exemplary regions were analyzed and only imputation accuracy (and not efficacy) was studied. We therefore assessed systematically and at a genome-wide level how well HapMap-based imputation with one of four publicly available programs, namely BEAGLE (Browning and Browning 2007), IMPUTE (Marchini et al. 2007), MACH (Li and Abecasis 2006), and PLINK (Purcell et al. 2007), would have performed in our own collection of composite genotypes, for three genome-wide SNP sets, of 449 healthy blood donors of German descent.

Our analysis was based upon genotypes generated by means of Affymetrix 5.0 (500 k), Affymetrix 6.0 (1,000 k) and Illumina 550 k SNP arrays, respectively. Since the same DNA samples were genotyped with all three arrays, and since the arrays contained only partially overlapping marker sets, extensive genome-wide benchmarking became possible through a comparison of the imputed and observed genotypes derived with the different arrays. Prior to inclusion, individual samples and SNPs were subjected to rigorous quality control to minimize the effects of genotyping errors on imputation accuracy. Since genotype imputation requires similar patterns of linkage disequilibrium (LD) in both the study and the reference population, we also tested whether the haplotype frequencies in HapMap were representative of those observed in our own sample.

Materials and methods

Samples and reference population

DNA samples of 241 male and 208 female unrelated individuals were obtained from the PopGen biobank (Krawczak et al. 2006). The blood donors, their parents, and grandparents were all born in Germany. Written, informed consent was obtained from all study participants and all protocols were approved by the institutional ethics committee. We used the HapMap CEU samples (Frazer et al. 2007; The International HapMap Consortium 2003), comprising CEPH Utah residents of northern and western European ancestry, as a reference population for imputation.

Genotyping and quality control

Genotypes were generated with Affymetrix Genome-Wide Human SNP Arrays 5.0 (500 k) and 6.0 (1,000 k) and with the Illumina HumanHap550 Bead array (550 k), respectively. Genotyping was performed by Affymetrix (Santa Clara, CA, USA) and Illumina (San Diego, CA, USA), respectively, as a commercial service. Further genotyping details are given in the Supplementary material. Prior to inclusion, individual samples and SNPs were subjected to rigorous quality control. All sample-wise call rates were found to be >95% for the three array types. The average call rate per sample was 99.8% for Affymetrix 5.0, 99.4% for Affymetrix 6.0, and 99.9% for Illumina 550 k. Individual SNPs were required to have a call rate >95% in the German samples, a minor allele frequency >1%, and had to be in Hardy–Weinberg equilibrium ($p > 0.01$). Annotation files from the Affymetrix and Illumina arrays were used to code SNPs on the forward strand in order to match the release 22 CEU phasing data from HapMap (Frazer et al. 2007). Strand orientation was checked automatically by all imputation programs, except for SNPs with A/T and C/G alleles, and no errors were reported. All pairs of individuals had average identity-by-state (IBS) values within the threefold inter-quartile range of the array-wide IBS distribution (Tukey's outlier criterion), thus minimizing the likelihood that any individual included in our study represented a close relative of another individual, or was of different ethnic origin. Data were quality-controlled using PLINK v1.02 (Purcell et al. 2007); the total SNP numbers before and after quality control are given in Table 1. The genotype concordance rate in the overlapping marker sets exceeded 99.7% for all array pairs (Table 1) and the corresponding allele frequency distributions were found to be virtually identical in Q–Q plots (data not shown), rendering platform-specific genotyping errors negligible. Data from the three array types could be matched unambiguously as belonging to the same individual by IBS values >0.985 for the overlapping markers.

Representativeness of HapMap data

Estimates of the two linkage disequilibrium (LD) measures r^2 and $|D'|$ and of the marker allele frequencies in HapMap were obtained from the HapMap web site (http://www.hapmap.org/downloads/ld_data/2006-06/) (NCBI build 36). HaploView 4.0 (Barrett et al. 2005) was employed with default options to assess LD in the German samples.

Imputation protocols

The imputation performance of four publicly available computer programs was assessed using the release 22 CEU phasing data from HapMap as a reference. All analyses were confined to those autosomal SNPs that were present in HapMap and that passed quality control in the German samples. Using genotypes from one array type ('imputation

Table 1 SNP sets used for imputation benchmarking

Array type	No. SNPs	Array type		
		Affymetrix 5.0	Affymetrix 6.0	Illumina 550 k
Affymetrix 5.0	358,391 (80.8% of 443,816)		331,176 (99.8%)	70,716 (99.9%)
Affymetrix 6.0	656,391 (73.1% of 934,968)	24,185 (95.9% of 25,215)		135,395 (99.7%)
		260,448 (80.5% of 323,215)		
Illumina 550 k	514,883 (91.7% of 561,474)	279,869 (97. 3% of 287,675)	452,227 (86.8% of 520,996)	
		435,759 (98.1% of 444,167)	371,854 (98.0% of 379,488)	

No. SNPs number of autosomal SNPs that passed quality control and had a minor allele frequency (MAF) ≥1% in the German samples. The total number of SNPs on each array and the percentage included in the study are given in parentheses. Upper right half: number of overlapping SNPs and, in parentheses, average genotype concordance rate between array types. Lower left half: number of SNPs that were unique to an array type and had phasing information available in HapMap CEU ('imputable SNPs'); top line: imputable SNPs unique to column array type; bottom line: imputable SNPs unique to row array type. The total number of SNPs unique to each array type and the percentage included in the study are given in parentheses

basis'), the genotypes of those SNPs that were unique to the other array type ('imputation target') and that had phasing information available in HapMap CEU ('imputable SNPs') were imputed using either BEAGLE v3.0.1 (Browning and Browning 2007), IMPUTE v0.3.2 (Marchini et al. 2007), MACH v1.0.15 (Li and Abecasis 2006), or PLINK v1.02 (Purcell et al. 2007). BEAGLE was run with default settings and ten iterations of the Markov sampler (`java --Xmx2048 --jar phased=phased.input.bgl unphased=unphased.input.bgl markers=marker.ids missing=0 niterations=10 out=out_file`). IMPUTE was run using the default parameters for CEPH populations (`impute --h phased_file --l legend_file --g geno_file --m genetic_map_chr*_CEU_b36.txt --call_thresh 0.0 --Ne 11418 --i info_file --o out_file`), formatted haplotypes and legend files (release 22 CEU phasing data from HapMap), and recombination rates as provided on the IMPUTE website (http://www.stats.ox.ac.uk/~marchini/software/gwas/impute.html). MACH was run with default settings (`mach1 --s legend.txt.gz --h phase.gz ---hapmapFormat --rounds 50 ---dosage ---quality ---greedy --geno --d dat_file --p ped_file --prefix out_file`) and 50 iterations of the Markov sampler, using the haplotypes and the legend files downloaded directly from HapMap. PLINK was run with default settings (`plink ---bfile in_file ---all ---proxy-impute all ---proxy-genotypic-concordance --proxy-show-proxies ---proxy-dosage --proxy-impute-threshold 0.0 ---make-bed --out out_file`), using PLINK binary file sets of the HapMap genotype data (release 22) as offered for download on the PLINK website (http://pngu.mgh.havars.edu/purcell/plink/res.shtml). The same HapMap CEU reference data (2,543,887 SNPs) were used for all imputations. All SNPs were coded on the forward strand. The call rate and minor allele frequencies distributions of the overlapping and of the unique marker sets were highly similar for all array pairs (Figures SF1–2).

Imputation performance

Observed and imputed genotypes were deemed concordant if they matched perfectly, and the average concordance rate was calculated over all imputed SNPs in a given marker set. Each of the four programs uses some sort of 'confidence threshold' (CT) for genotype calling. However, their meaning and interpretation differ greatly between programs, rendering CT values impossible to compare between programs. Anyhow, for benchmarking purposes, we defined imputation efficacy as the proportion of imputable SNPs for which the program-specific confidence in an imputed genotype equaled or exceeded a given CT, whereas *imputation accuracy* was quantified as the concordance rate between the imputed and observed genotypes of these SNPs. We also varied the CT values in order to assess the impact of this parameter upon both imputation accuracy and efficacy.

Computation

All computations were carried out on a Linux cluster (rzcluster, Christian-Albrechts University, Kiel, Germany) comprising 74 nodes with 322 cores and 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). Genotype concordance was calculated using PLINK's `---merge-mode 7` command. Statistics and graphs were generated using R 2.6.2 (R Development Core Team 2008). Scripts and data are available at the authors' website (http://www.ikmb.uni-kiel.de/imputation/) to allow

others to optimize existing imputation algorithms or to benchmark new ones in relation to our results.

Results

Representativeness of HapMap

Since genotype imputation requires similar patterns of linkage disequilibrium (LD) in both the study and the reference population, we first compared the LD structure in the German and HapMap data for those SNPs on the Affymetrix 6.0 and Illumina 550 k arrays for which LD information was available in HapMap (Supplementary methods). Estimates of r^2 from the German and HapMap CEU genotype data were found to be in strong agreement (Pearson correlation $\rho \geq 0.95$), regardless of SNP allele frequency (Supplementary materials ST1 and SF3). The LD structures of other HapMap populations also showed a significant albeit weaker correlation with that of the German samples (ST1 and SF4–SF6). For $|D'|$, population-specific estimates were less strongly correlated with one another than for r^2.

Imputation accuracy and efficacy

While the imputation accuracy was consistently high (>93%) for all four programs employing the default CT values, their imputation efficacy varied substantially (Table 2). Whereas BEAGLE, IMPUTE, and MACH actually imputed most imputable SNPs (≥96.5%) for all array pairs, PLINK often failed and excluded up to 2/3 of the markers in some instances. Since BEAGLE does not provide a default threshold for its confidence measure (the posterior probability), we chose 0.90 for representation in Table 2. Increasing the number of iterations from 10 to 50 in BEAGLE did not alter the results considerably (data not shown). Imputation of Affymetrix genotypes using Illumina data as the imputation basis yielded higher concordance rates than vice versa (see Table 2).

The use of CT values smaller than their default value reduced the accuracy of BEAGLE, IMPUTE, and MACH only marginally, but strongly impeded imputation with PLINK (Fig. 1). On other hand, small CT values increased the imputation efficacy, particularly for PLINK. Using CT values above the default value only slightly affected the accuracy for all four programs, but often reduced the imputation efficacy dramatically. Figure 2 illustrates the trade-off between accuracy and efficacy for each program.

In terms of their accuracy and efficacy, IMPUTE and MACH were found to be superior to the other two programs, irrespective of the imputation basis. While BEAGLE was only slightly less accurate than IMPUTE and MACH, PLINK performed consistently more poorly for all array pairs. IMPUTE and MACH also yielded the best trade-off between accuracy and efficacy when varying the CT values, irrespective of the imputation basis. The imputation accuracy of all programs decreased with increasing marker heterozygosity, although this effect was weak in most instances (Supplementary Figures SF7–10). Increasing marker hterozygosity also reduced the general trade-off between accuracy and efficacy (Supplementary Figures SF11–14).

Genotype imputation can be expected to perform more poorly in regions of high inter-individual variability. The extended major histocompability complex (xMHC) on chromosome *6p21* (25.0–34.0 Mb) is known to be characterized by an extremely high haplotype diversity in some of its subregions (Raymond et al. 2005; Traherne 2008). Therefore, we assessed the accuracy and efficacy of all four programs separately for the xMHC, but using all genotypes available for chromosome 6 as the imputation basis and adhering to the default CT. The results were quite similar to those of the genome-wide assessment (Supplementary Table ST2).

Computational performance and requirements

The four programs under study differed significantly in terms of their speed and hardware requirements. On a single-processor machine (AMD-Barcelona, 2.1 GHz), the cumulative run time for the entire project was approximately 349 h for BEAGLE, 455 h for IMPUTE, 1,574 h for MACH, and 138 h for PLINK. The average computation time per imputed marker is given in Table 3 for all arrays and all programs. The working memory allocations did not exceed 2 GB RAM for BEAGLE, 16 GB for IMPUTE, 8 GB for MACH, and 4 GB for PLINK, respectively.

Discussion

Genotype imputation may potentially increase the power to detect disease associations with a given marker set, and this benefit has been shown to be largest for rare disease-associated variants (Marchini et al. 2007). Hence, imputation during the exploratory stage of a genetic disease association study may unravel associations that would otherwise be overlooked. Another instance in which imputation can prove valuable is the fine-mapping of known disease-associated regions. Here, imputation may serve to identify additional candidate SNPs worth including into a more detailed follow-up. Finally, imputation has been discussed as a method of data quality control and missing data inference in conjunction with high-throughput

Table 2 Imputation accuracy and efficacy at default CT values

Array type	Array type/Imputation algorithm											
	Affymetrix 5.0				Affymetrix 6.0				Illumina 550 k			
	BEAGLE	IMPUTE	MACH	PLINK	BEAGLE	IMPUTE	MACH	PLINK	BEAGLE	IMPUTE	MACH	PLINK
Affymetrix 5.0 (349,954)					260,448				435,759			
					81.8%	83.9%	96.8%	34.2%	77.1%	79.3%	96.5%	30.4%
					96.4%	97.0%	94.6%	93.6%	96.5%	97.0%	93.8%	94.0%
					2.2–100.0	2.2–100.0	12.7–100.0	2.2–100.0	2.9–100.0	3.1–100.0	2.7–100.0	0.9–100.0
					94.9–99.3	95.8–99.6	93.5–99.6	94.0–99.1	94.9–99.6	95.8–99.8	92.2–99.6	94.4–99.3
Affymetrix 6.0 (586,217)	24,185								371,854			
	93.9%	94.6%	99.0%	50.0%					88.4%	89.8%	98.9%	44.8%
	97.6%	98.1%	97.3%	95.2%					97.2%	97.6%	96.1%	94.6%
	0.0–100.0	0.0–100.0	0.0–100.0	3.6–100.0					4.7–100.0	2.9–100.0	4.0–100.0	2.0–100.0
	96.9–99.6	97.8–99.8	97.6–99.8	95.8–99.3					96.2–99.6	96.9–99.8	96.0–99.8	95.1–99.3
Illumina 550 k (505,844)	279,869				452,227							
	97.3%	97.9%	99.6%	61.6%	96.7%	97.5%	99.5%	58.1%				
	98.2%	98.0%	98.3%	95.7%	97.9%	98.5%	98.1%	95.3%				
	0.0–100.0	0.0–100.0	0.0–100.0	3.8–100.0	2.2–100.0	2.2–100.0	2.2–100.0	4.5–100.0				
	97.8–99.8	98.7–100.0	98.4–100.0	96.0–99.6	97.3–99.8	98.2–99.8	98.2–99.8	95.8–99.3				

Left column: array type serving as imputation basis and, in parentheses, number of quality-controlled SNPs with available HapMap phasing data. For each combination of imputation basis and target, the table contains the number of imputable SNPs (top row) as well as, for each imputation algorithm, the number of imputed SNPs relative to the number of imputable SNPs ('imputation efficacy'; centre row) Also given are the mean, range and inter-quartile range of the genotype concordance rates between imputed and observed genotypes ('imputation accuracy'; bottom row). Only SNPs with phasing data available in HapMap CEU were considered. Imputation was carried out using the software-specific default CT values (IMPUTE: 0.90 ('confidence score'), MACH: 0.30 (r^2), PLINK: 0.95 ('info score')]. Since BEAGLE does not provide a default confidence threshold, we chose 0.90 for comparison (for details, see main text)

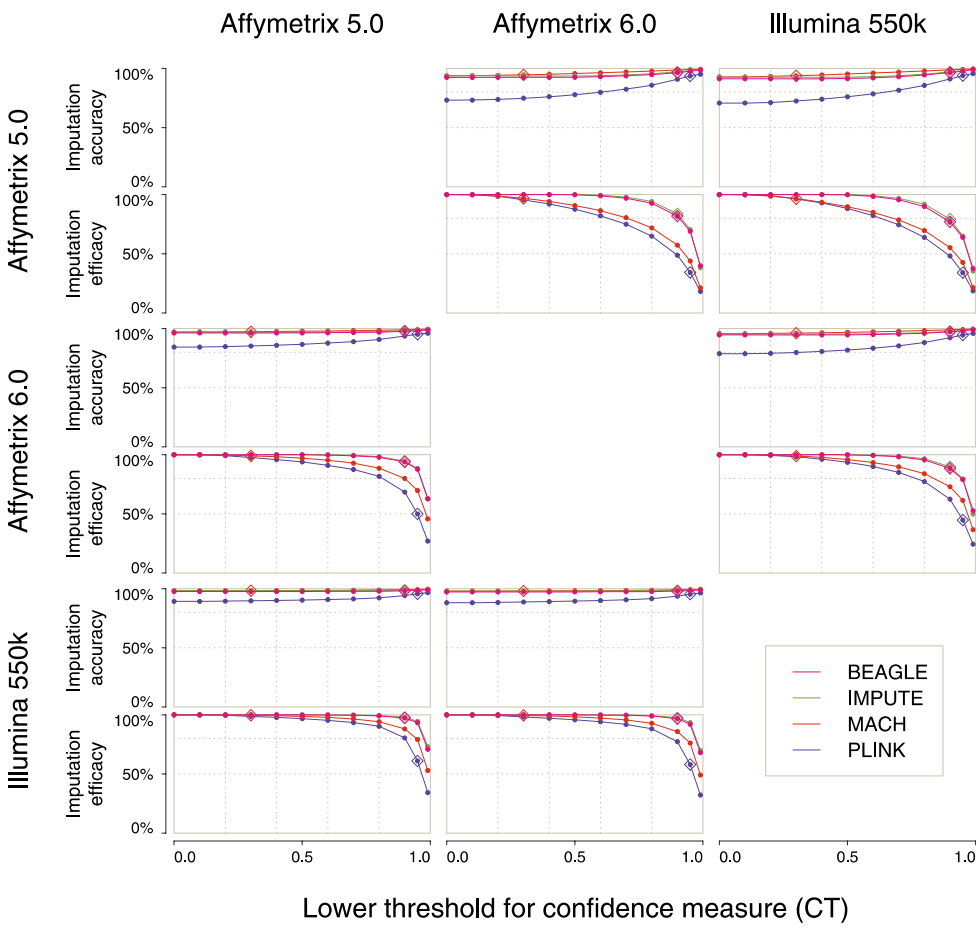

Fig. 1 Imputation performance as a function of software-specific confidence thresholds (CT). For each of the six combinations of imputation basis (*row*) and target (*column*), the imputation accuracy (percentage of correctly imputed genotypes, *top*) and the imputation efficacy (proportion of imputed SNPs relative to the number of imputable SNPs, *bottom*) are plotted against the respective CT values. Note that the CT values have program-specific interpretations and are therefore not directly comparable. Default CT values, indicated by *diamonds*, were as follows: IMPUTE, 0.90 ('confidence score'); MACH, 0.30 (r^2); PLINK, 0.95 ('info score'). Since BEAGLE does not provide a default confidence threshold, we chose 0.90 for comparison (for details, see main text)

sequencing, thereby significantly reducing the costs of such experiments (see the meeting report available at http://www.1000genomes.org/). It should be remembered, however, that imputed genotypes are not actually observed genotypes and that the ambiguity of their prediction somehow has to be included in their interpretation. One way to address this issue in the context of significance testing would be, for example, to use the posterior genotype probability, rather than discrete genotypes, in say linear or logistic regression models. The need to account for genotype ambiguity is particularly strong when whole haplotypes including un-type loci are being "estimated" (Gourraud et al. 2004). Simply using marker-wise best guesses of genotypes such instances would be meaningless.

We have shown in the present study that HapMap CEU-based imputation can reliably infer missing genotypes in a population of northern European descent, even in variable regions such as the extended MHC. This corroborates previous findings that SNPs in HapMap allow the prediction of HLA class I and II gene alleles (Leslie et al. 2008). The high imputation accuracy achieved by all four programs at their respective default CT values is not surprising, bearing in mind that these values must have been adjusted by the authors in such a way as to ensure a high posterior genotype probability (e.g.≥95%). By contrast, considerable differences were observed in terms of the number of SNPs for which genotypes were actually imputed, using default CT values. Taking both imputation accuracy and efficacy into account, MACH and IMPUTE turned out to be superior to the other two programs. At only a minor loss of accuracy, however, BEAGLE may also be a good choice for some applications, particularly because of its much smaller memory requirements. MACH and IMPUTE had nearly identical trade-offs between accuracy and efficacy. This was due to the fact that very similar statistical algorithms have been implemented in the two programs, even though minor differences nevertheless exist. For example, while MACH estimates recombination rates from the data at hand, IMPUTE relies upon user-specified recombination rates. Although the latter approach may save computation time, it renders IMPUTE sensitive to model misspecification (Browning 2008). This may be an important issue when imputation is carried out for populations that are only partially represented by HapMap CEU, namely those from southern and eastern Europe. No methodological details have been published so far about PLINK, thereby rendering a detailed discussion of the possible reasons for its poor performance difficult.

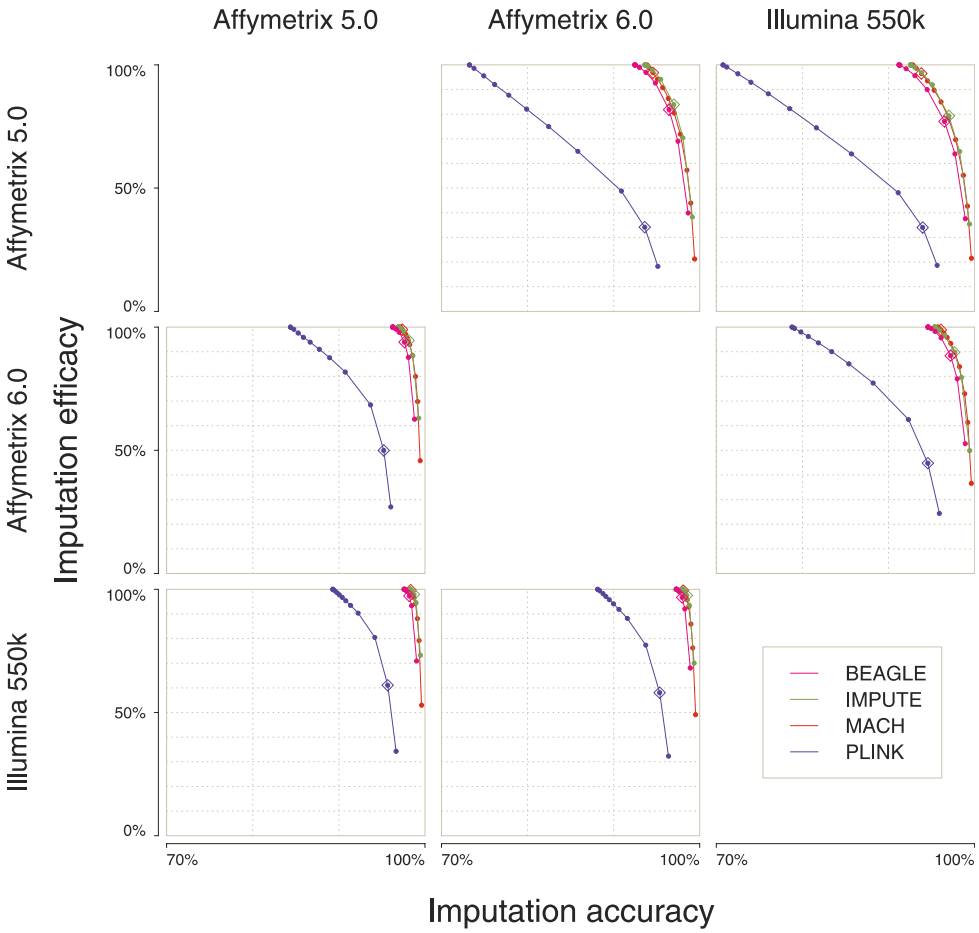

Fig. 2 Trade-off between imputation accuracy and efficacy. For each of the six combinations of imputation basis (*row*) and target (*column*), the imputation accuracy (percentage of correctly imputed genotypes) is plotted against the imputation efficacy (proportion of imputed SNPs relative to the number of imputable SNPs) at varying CT values. Performance at the default confidence threshold is indicated by *diamonds*. Default CT values were as follows: IMPUTE, 0.90 ('confidence score'); MACH, 0.30 (r^2); PLINK, 0.95 ('info score'). Since BEAGLE does not provide a default threshold for its confidence measure, i.e. the posterior probability, we chose 0.90 for comparison (for details, see main text)

In our experience, data handling is much easier with MACH and BEAGLE than with IMPUTE. Furthermore, IMPUTE occasionally caused problems during our study in terms of incomplete error feedback and excessive memory consumption.

In the present study, genotypes for Affymetrix SNPs were imputed at higher rates on the basis of Illumina SNP genotypes than vice versa. This is not surprising since SNP selection for the Illumina platform was originally based upon r^2-tagging of HapMap data. Nothing can be said, however, about the performance of imputation in regions devoid of HapMap markers. Genotypes for markers with two high-frequency alleles, i.e. those which are most useful for disease association analysis, were found to be more difficult to impute than genotypes of less polymorphic markers (Supplementary Figures SF7–14). One straightforward explanation for this difference is that the frequency of genotypes containing the common allele of a SNP increases with decreasing minor allele frequency, so that the former would be easier to predict for markers with rare alleles by chance alone. In order to clarify whether different imputation algorithms perform differently at varying levels of polymorphism, further benchmarking studies focusing on the imputation accuracy for rare genotypes are needed.

Table 3 Computation time for genotype imputation

Array type	Imputation algorithm			
	BEAGLE	IMPUTE	MACH	PLINK
Affymetrix 5.0 (349,954)	0.215	0.236	0.757	0.068
Affymetrix 6.0 (586,217)	0.204	0.264	1.065	0.073
Illumina 550 k (505,844)	0.196	0.290	1.279	0.083

Given is the average computation time (in s) per imputation and marker. Averages were taken over all imputed SNPs from the release 22 CEU HapMap phasing data that were not genotyped on the respective array type

One prerequisite for genotype imputation to be feasible is the availability of an appropriate reference data set. The HapMap CEU phasing data used here appear ideally suited for genotype imputation in European or Europe-derived populations, particularly from the North and West. This was also highlighted in our study by the strong correlation between r^2 values observed in the German and the HapMap CEU samples, and is in line with the results of a recent

genome-wide study on genetic diversity in Europe (Lao et al. 2008). One possible explanation for the weaker inter-population correlation observed for D′ than for r^2 is the small size of the HapMap CEU sample (120 chromosomes). In fact, estimates of D′ are known to be upwardly biased and to have a large variance in small samples (Ardlie et al. 2002; Teare et al. 2002; Terwilliger et al. 2002). Anyhow, further studies are required to evaluate the performance of imputation algorithms in populations other than Europeans, particular in those of African ancestry where LD is generally lower, or in mixed populations like those present in parts of northern America.

Recently, an insightful review has been published of the algorithms underlying three of the computer programs benchmarked here, namely BEAGLE, IMPUTE, and MACH (Browning 2008). Additional details on the theoretical aspects of different imputation methods can be found in that publication. Finally, we would like to point out that we were of course aware of the existence of other imputation programs such as, for example, FAMHAP (Becker and Knapp 2004) and BIMBAM (Servin and Stephens 2007). However, we chose not to include them in our benchmarking study because they are either still in the stage of development (FAMHAP) or focus upon association rather than imputation and provide no measure of imputation confidence (BIMBAM).

Acknowledgments The authors wish to thank all probands for participating in the study. We also thank Alfred Wagner and Simone Knief of the Computational Centre, Christian-Albrechts University Kiel, Germany, for their support. Thomas Wienker and Michael Steffens (IMBIE, University of Bonn, Germany) are acknowledged for performing the initial quality control of the genotype data. Marcus Will, Michael Wittig (both at the Institute of Clinical Molecular Biology, Kiel) and Olaf Junge (Institute of Medical Informatics and Statistics, Kiel) are gratefully acknowledged for expert technical help. We would like to thank Shaun Purcell (PNGU, Massachusetts General Hospital, Boston, MA, USA), Goncalo Abecasis and Yun Li (both at the Center for Statistical Genetics, University of Michigan, MI, USA), Brian Browning (Department of Statistics, University of Auckland, New Zealand), and Tim Becker (IMBIE, University of Bonn, Germany) for providing access to the latest versions of their software and for helpful discussions. This study was supported by the German Ministry of Education and Research (BMBF) through the National Genome Research Network (NGFN). The project received infrastructure support through the DFG excellence cluster "Inflammation at Interfaces".

References

Anderson CA, Pettersson FH, Barrett JC, Zhuang JJ, Ragoussis J, Cardon LR, Morris AP (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

Ardlie KG, Kruglyak L, Seielstad M (2002) Patterns of linkage disequilibrium in the human genome. Nat Rev Genet 3:299–309

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263–265

Becker T, Knapp M (2004) Maximum-likelihood estimation of haplotype frequencies in nuclear families. Genet Epidemiol 27:21–32

Browning SR (2008) Missing data imputation and haplotype phase inference for genome-wide association studies. Hum Genet 124:439–450

Browning SR, Browning BL (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 81:1084–1097

Browning BL, Browning SR (2008) Haplotypic analysis of Wellcome Trust Case Control Consortium data. Hum Genet 123:273–280

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Sun W, Wang H, Wang Y, Xiong X, Xu L, Waye MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallee C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PK, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T et al (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861

Gourraud PA, Genin E, Cambon-Thomsen A (2004) Handling missing values in population data: consequences for maximum likelihood estimation of haplotype frequencies. Eur J Hum Genet 12:805–812

Krawczak M, Nikolaus S, von Eberstein H, Croucher PJ, El Mokhtari NE, Schreiber S (2006) PopGen: population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships. Community Genet 9:55–61

Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, Balascakova M, Bertranpetit J, Bindoff LA, Comas D, Holmlund G, Kouvatsi A, Macek M, Mollet I, Parson W, Palo J, Ploski R, Sajantila A, Tagliabracci A, Gether U, Werge T, Rivadeneira F, Hofman A, Uitterlinden AG, Gieger C, Wichmann HE, Ruther A, Schreiber S, Becker C, Nurnberg P, Nelson MR, Krawczak M, Kayser M (2008) Correlation between genetic and geographic structure in Europe. Curr Biol 18:1241–1248

Leslie S, Donnelly P, McVean G (2008) A statistical method for predicting classical HLA alleles from SNP data. Am J Hum Genet 82:48–56

Li Y, 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, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39:906–913

Pei YF, Li J, Zhang L, Papasian CJ, Deng HW (2008) Analyses and comparison of accuracy of different genotype imputation methods. PLoS ONE 3:e3551

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575

R Development Core Team (2008) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Raymond CK, Kas A, Paddock M, Qiu R, Zhou Y, Subramanian S, Chang J, Palmieri A, Haugen E, Kaul R, Olson MV (2005) Ancient haplotypes of the HLA Class II region. Genome Res 15:1250–1257

Servin B, Stephens M (2007) Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet 3:e114

Teare MD, Dunning AM, Durocher F, Rennart G, Easton DF (2002) Sampling distribution of summary linkage disequilibrium measures. Ann Hum Genet 66:223–233

Terwilliger JD, Haghighi F, Hiekkalinna TS, Goring HH (2002) A bias-ed assessment of the use of SNPs in human complex traits. Curr Opin Genet Dev 12:726–734

The International HapMap Consortium (2003) The International HapMap Project. Nature 426:789–796

The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320

Traherne JA (2008) Human MHC architecture and evolution: implications for disease association studies. Int J Immunogenet 35:179–192

Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14, 000 cases of seven common diseases and 3, 000 shared controls. Nature 447:661–678