The effect of genome-wide association scan quality control on imputation outcome for common variants

Eleftheria Zeggini, Lorraine Southam, Kalliope Panoutsopoulou, Nigel W Rayner, Kay Chapman, Caroline Durrant, Teresa Ferreira, Nigel Arden, Andrew Carr, Panos Deloukas, et al.

▶ **To cite this version:**

Eleftheria Zeggini, Lorraine Southam, Kalliope Panoutsopoulou, Nigel W Rayner, Kay Chapman, et al.. The effect of genome-wide association scan quality control on imputation outcome for common variants. European Journal of Human Genetics, 2011, 10.1038/ejhg.2010.242 . hal-00611254

HAL Id: hal-00611254
https://hal.science/hal-00611254

Submitted on 26 Jul 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The effect of genome-wide association scan quality control on imputation outcome for common variants

Lorraine Southam[1], Kalliope Panoutsopoulou[2], N William Rayner[3,4], Kay Chapman[1], Caroline Durrant[3], Teresa Ferreira[3], Nigel Arden[5,6], Andrew Carr[1], Panos Deloukas[2], Michael Doherty[7], John Loughlin[8], Andrew McCaskie[8,9], William ER Ollier[10], Stuart Ralston[11], Timothy D Spector[12], Ana M Valdes[12], Gillian A Wallis[13], J Mark Wilkinson[14,15], the arcOGEN consortium, Jonathan Marchini[16], Eleftheria Zeggini[*,2]

[1]Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK, [2]Wellcome Trust Sanger Institute, Hinxton, UK, [3]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK, [4]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford, UK, [5]NIHR Biomedical Research Unit, University of Oxford, Oxford, UK, [6]MRC Epidemiology Resource Centre, University of Southampton, Southampton, UK, [7]Academic Rheumatology, University of Nottingham, Nottingham, UK, [8]Institute of Cellular Medicine, Musculoskeletal Research Group, Newcastle University, Newcastle upon Tyne, UK, [9]The Newcastle upon Tyne Hospitals NHS Trust Foundation Trust, The Freeman Hospital, Newcastle upon Tyne, UK, [10]Centre for Integrated Genomic Medical Research, University of Manchester, Manchester, UK, [11]Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK, [12]Department of Twin Research and Genetic Epidemiology, King's College London, London, UK, [13]Wellcome Trust Centre for Cell Matrix Research, University of Manchester, Manchester, UK, [14]Academic Unit of Bone Metabolism, Department of Human Metabolism, University of Sheffield, Sheffield, UK, [15]Sheffield NIHR Bone Biomedical Research Unit, Centre for Biomedical Research, Northern General Hospital, Sheffield, UK, [16]Department of Statistics, University of Oxford, Oxford, UK.
*Correspondence: Dr. Eleftheria Zeggini, Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1HH, UK. Tel: +44 1223 496868; Fax: +44 1223 496826; Email: eleftheria@sanger.ac.uk.

Running title: The effect of GWAS QC on imputation outcome

Abstract

Imputation is an extremely valuable tool in conducting and synthesising genome-wide association studies (GWAS). Directly-typed SNP quality control is thought to affect imputation quality. It is therefore common practice to use quality-controlled (QCed) data as input for imputing genotypes. This study aims to determine the effect of commonly-applied QC steps on imputation outcomes. We performed several iterations of imputing SNPs across chromosome 22 in a dataset consisting of 3,177 samples with Illumina 610k GWAS data, applying different QC steps each time. The imputed genotypes were compared to the directly-typed genotypes. In addition, we investigated the correlation between alternatively QCed data. We also applied a series of post-imputation QC steps

balancing elimination of poorly-imputed SNPs and information loss. We found that the difference between the unQCed data and the fully-QCed data on imputation outcome was minimal. Our study shows that imputation of common variants is generally very accurate and robust to GWAS QC, which is not a major factor affecting imputation outcome. A minority of common-frequency SNPs with particular properties cannot be accurately imputed regardless of QC stringency. These findings may not generalize to the imputation of low frequency and rare variants.

Keywords
genome-wide association study, imputation, quality control, single nucleotide polymorphism

Introduction

Genome-wide association scans (GWAS) have proven to be a successful strategy for detecting common variants exerting modest effects on complex disease risk. Currently available commercial platforms focus on common variants and capture the majority of HapMap (1) SNPs with minor allele frequency (MAF) >0.05 in European populations (2). Several large-scale consortia have been formed in order to carry out GWAS meta-analyses for various phenotypes, with successful outcome (for example 3, 4, 5, 6, 7). To enable the combination of data across studies carried out on different platforms, and to enable *in silico* fine mapping of association signals, imputation approaches were proposed a few years ago (8) as a means of statistically inferring genotypes at untyped loci using a reference set, for example the HapMap (~2,500,000 SNPs).

An important aspect of any GWAS analysis is the implementation of a series of rigorous quality control (QC) steps prior to testing for association. These QC procedures help guard against genotyping error, population stratification, sample duplication and other confounders that can affect the analysis results. QC steps are typically applied at the sample- and SNP-specific level. Sample-level QC includes filtering out samples with low call rates, evidence for different ethnic origin, high heterozygosity, relatedness/ duplication, gender discrepancies and genotyping batch effects. SNP-level QC includes filtering out SNPs with low call rates and deviation from Hardy-Weinberg equilibrium

(HWE) at pre-determined thresholds. It is generally believed that datasets should be stringently QCed at the marker level before applying imputation approaches. For this reason, lower MAF SNPs tend to also be excluded, as their accuracy can be hampered by poor clustering properties and incorrect automated genotype calling (at least with currently widely-used algorithms). Even though such weight is placed on pre-imputation SNP QC, the effects of applying different criteria and thresholds to the starting dataset have not been investigated thus far. In this report, we evaluate the effect of GWAS QC on imputation outcome, and find that imputation works very well for common variants irrespective of QC and that a minority of some common-frequency SNPs with particular properties cannot be accurately imputed regardless of QC stringency.

Materials and Methods

We used an empirical GWAS dataset to assess the effect of QC on imputation outcome. We focused on chromosome 22, (n=9,038 directly-typed SNPs) from 3,177 osteoarthritis (OA) cases from the UK, typed on the Illumina 610k quad chip as part of the arcOGEN consortium GWAS (manuscript submitted). Chromosome 22 is representative of the genome in terms of the proportion of directly-typed to imputed SNPs. All samples included in our analysis had passed standard sample-level QC (based on call rate, heterozygosity, relatedness, ethnicity and gender discrepancies). We imputed genotypes at variants on the basis of HapMap phase II release 22 CEU data (n=33,815 SNPs on chr22) using IMPUTE v1 (8). We performed each imputation in duplicate, with and without the IMPUTE v1 -pgs (predict genotyped SNPs) flag, which resulted in one set of imputed data

containing the original genotypes and in the other imputed genotypes. To assess the effect of varying levels of QC, we carried out several rounds of imputation, using differently QCed OA SNP data as the starting point.

Initially, we imputed on the basis of no SNP-level QC, including all directly-typed SNPs, regardless of MAF, call rate and HWE. We also imputed on the basis of only those SNPs that passed stringent QC thresholds (call rate>95% for SNPs with a MAF≥5% and call rate>99% for SNPs with a MAF<5%, HWE exact p>0.0001, MAF >0.01 and removing all SNPs with GC or TA alleles) (Table 1). Although imputation biases can occur due to poorly-clustering SNPs with miscalled genotypes in the starting dataset, cluster plot checking is not feasible at the genome-wide scale and therefore it is not implemented in standard GWAS QC.

1 We evaluated the accuracy of imputed genotypes by comparing allele frequencies at the same SNP between imputed and true, directly typed data. For each QC-imputation iteration, we performed an allele frequency comparison between the actual directly-typed and imputed SNPs. Under perfect imputation, we would expect to see alignment with the null hypothesis of no association. We used SNPTEST (9) to investigate differences between directly-typed and imputed genotypes at the same variants within the same samples, taking into account the distribution of genotype probabilities for each individual. For the purposes of our comparison, we used those SNPs that were directly genotyped in OA cases and also present in the HapMap reference samples. Table 1 summarizes the number of these SNPs for each QC threshold.

When comparing directly-typed with imputed allele frequencies at the same variant in the same individuals, we arbitrarily considered $p<10^{-6}$ as significantly different. We calculated the correlation between imputed and directly-typed MAF, using the expected counts to allow for genotype-associated probabilities. We also applied a series of post-imputation QC steps in order to eliminate unreliably imputed SNPs, aiming to filter out as many of these SNPs as possible whilst retaining a good proportion of non-significant SNPs. We compared two alternative methods for post imputation QC filtering, firstly the IMPUTE-info score, which is associated with the imputed allele frequency estimate which ranges from 1, indicating high confidence, to 0 suggesting decreased confidence, and secondly the freq-add-proper-info score provided by SNPTEST, a relative statistical score ranging from 0 to 1, representing no information to complete information respectively. The SNPTEST freq-add-proper-info score has been shown to be highly correlated with the IMPUTE-info score under the additive model (10). In both scenarios we also filtered out SNPs with MAF <5%. Figure 1 illustrates the effects of altering post-imputation QC filters on the QCed data. Based on these results we chose to use the IMPUTE-info score with a filtering threshold <0.8 and MAF <5% which effectively eliminated ~79% of the significant SNPs whilst retaining ~85% of the non-significant ones (SNPTEST freq-add-proper-info <0.9 and MAF 5% would be roughly equivalent to this eliminating ~73% of the significant SNPs whilst retaining ~89% of the non-significant ones). We applied this post-imputation filter to each of our datasets and compared the results. We looked at the unQCed and QCed datasets first, as synopsised in Table 1. For each scenario, we examined frequency

differences between the directly-typed and the imputed genotypes as described above. In addition, we compared the imputed genotypes at imputed SNPs only for the unQCed and the fully QCed (QCed data with all poorly-clustering markers removed) strategies.

Results

Table 1 summarises the number of SNPs with significantly ($p<10^{-6}$) different allele frequencies between the directly-typed and imputed data in the same set of individuals for each of the different QC sets. Correlation plots and R^2 values for the comparisons of the QCed and unQCed datasets are presented in Figure 2. The difference between the unQCed (R^2=0.993) and QCed data (R^2=0.994) was minimal. After post-imputation filtering there were 77 SNPs with significantly different (imputed v. directly-typed) allele frequencies in the unQCed data compared with 67 significant SNPs in the QCed data. In an attempt to improve imputation for the small subset of poorly-imputed SNPs in the QCed data we excluded all SNPs with MAF<5% and, subsequently, also SNPs with MAF<10%. We found that eliminating these lower MAF SNPs prior to imputation had little effect overall. The R^2 for the post-imputation QC filtered comparison with the QCed data was virtually identical both when excluding all SNPs with MAF<5% (R^2 = 0.994) and when excluding all SNPs with MAF<10% (R^2 = 0.991).

Given this apparent minimal influence of input data QC on imputation outcome, we investigated further the small set of SNPs displaying significant allele frequency differences for the presence of a common characteristic that could conceivably be used as a post-imputation filter. In order to rule out poor genotyping as the cause of these

significant differences, we examined all cluster plots for the unfiltered significant SNPs ($p<1x10^{-6}$, n=325). Fourteen poorly-clustering SNPs were removed and the data were re-imputed. After post-imputation QC, 3 additional SNPs were not significant and 6 were less significant. We then inspected the cluster plots for 10 SNPs on either side of the 61 SNPs remaining significantly different to rule out poor imputation due to flanking SNP poor clustering properties. We examined the cluster plots for 1,008 SNPs and found that 36 of these were poor; these resided in the proximity of 35 of the significant SNPs. We subsequently removed these SNPs and re-imputed. We found that following post-imputation QC filtering, only 3 of the 61 SNPs were no longer significant and the R^2 remained the same as for the QCed data ($R^2 = 0.994$) for the post-imputation QC filtered data. When we repeated comparisons using IMPUTE v2 with the HapMap3 (CEU, release #2 Feb 2009) and data from the 1,000 genomes project (Pilot 1 genotypes released Mar 2010; phased haplotypes released Jun 2010) as the reference panels, we observed qualitatively similar results.

Differences in region-specific recombination rates may account for the few remaining significant SNPs, as variants in areas of especially high recombination rate may be more challenging to impute accurately regardless of QC. To investigate this, we firstly examined the QCed unfiltered data and found that when the data were dichotomized into those markers with lower (<1cm/Mb) and higher (≥1cM/Mb) recombination rates there were more significant SNPs present in the higher recombination rate group compared to the lower recombination group ($p=1.85x10^{-27}$, average recombination rates of 12.8 and 3.04

respectively). When we examined the QCed data post-imputation QC, this difference disappeared (p=0.526). This clearly indicates that application of the post-imputation QC filter successfully identifies the majority of significant SNPs with high recombination rates. Therefore, to include recombination rate as an extra filter would not be prudent, for example using the QCed post-imputation QC filtered data and applying a further filter using a recombination rate threshold of >1cM/Mb would eliminate 2,075 SNPs, only 24 of which are significantly different.

Discussion

The imputation accuracy of common variants does not appear to be substantially affected by GWAS QC steps. Our data demonstrate that there is little difference in imputation accuracy observed in unQCed GWAS data when compared with QCed GWAS data. Furthermore, the implementation of additional QC steps (e.g. filtering out variants with MAF<0.05 and <0.10) does not considerably improve overall imputation accuracy. Missing variants and directly typed variants that fail pre-imputation QC checks are imputed and these data are used for downstream analyses. Post-imputation QC successfully eliminates a good proportion of inaccurately-imputed SNPs. Specifically, by applying a very stringent post-imputation QC threshold a smaller set of variants with more accurately predicted genotypes remain. The IMPUTE-info threshold of <0.8 and MAF≤5% criterion successfully filtered out the majority of poorly-imputed SNPs. However, the application of these strict filters in GWAS data could result in many SNPs being excluded from the data and thus potential true association signals could be missed. Some of the inaccurately-imputed

variants were due to poor clustering properties. It is plausible that the handful of variants that still remained inaccurately imputed could be due to differences in ethnicity between our data and the HapMap CEU reference panel from which the genotypes were predicted. We have used IMPUTE, but do not expect our results and conclusions to qualitatively differ with different imputation methods, for example BEAGLE and MACH exhibit similar imputation accuracy to IMPUTE (11). Differences in population structure between the reference panel and target dataset can be a source of imputation inaccuracy. Imputation accuracy for common SNPs may be further increased by using larger reference panels with data on denser sets of variants. Our results show that GWAS QC is not of paramount importance for the imputation of common variants. This may be different for the imputation of low frequency and rare variants based on emerging reference panels such as the 1000 genomes (www.1000genomes.org) and UK10k (wwww.uk10k.org) projects. In summary, our study demonstrates that imputation of common variants is generally very accurate and robust to GWAS QC, which is not a major factor affecting imputation outcome.

Acknowledgements
EZ is supported by the Wellcome Trust (WT088885/Z/09/Z), LS is supported by the European Community Framework 7 large collaborative project grant TREAT-OA, KC is supported by a Botnar Fellowship and by the Wellcome Trust (WT079557MA), NWR is supported by the Wellcome Trust (WT079557MA), JMW is supported by the Higher Education Funding Council for England. JL receives support from the UK NIHR Biomedical

Research Centre for Ageing and Age-related disease award to the Newcastle upon Tyne Hospitals NHS Foundation Trust. The arcOGEN consortium is funded by a special purpose grant from Arthritis Research UK (grant 18030).

Conflict of Interest Statement

The authors declare no conflict of interest.

References

1. The International HapMap Consortium: The International HapMap Project. *Nature 2003;* **426**: 789-796

2. Barrett JC, Cardon LR: Evaluating coverage of genome-wide association studies. *Nat Genet 2006*; **38**: 659-662

3. Zeggini E, Scott LJ, Saxena R, *et al*: Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet 2008*; **40**: 638-45

4. Prokopenko I, Langenberg C, Florez JC, *et al*: Variants in MTNR1B influence fasting glucose levels. *Nat Genet 2009*; **41**: 77-81

5. Franke A, Balschun T, Karlsen TH, *et al*: Replication of signals from recent studies of Crohn's disease identifies previously unknown disease loci for ulcerative colitis. *Nat Genet 2008*; **40**: 713-715

6. Barret JC, Clayton DG, Concannon P, *et al:* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes, *Nat Genet 2009*; **41**: 703-707

7. Soranzo N, Spector TD, Mangino M, *et al*: A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat Genet 2009*; **41**: 1182 - 1190

8. Marchini J, Howie B, Myers S, *et al*: A new multipoint method for genome-wide association studies via imputation of genotypes. Nat Genet 2007; **39**: 906-913

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

10 . Marchini J, Howie B: Genotype imputation for genome-wide association studies. *Nat Rev Genet* 2010; **11**(7):499-511

3 11. Nothnagel M, Ellinghaus D, Schreiber S, et al. A comprehensive evaluation of SNP genotype imputation. Hum Genet 2009; 125:163–171

Titles and legends to figures

Table 1.

arcOGEN data for chromosome 22 detailing the different pre-imputation QC steps. A breakdown of the SNP number for each QC threshold is indicated both with and without the post-imputation QC.

Figure 1.

A. Imputation results for the QCed data indicating the total number of SNPs filtered for different QC thresholds using the IMPUTE-info and freq-add-proper-info scores. The SNPs remaining after the filter (red bar) have been subdivided into SNPs that are significant (green bar) and not significant (yellow bar). B. The same data as percentage of significant and non significant SNPs removed for each threshold. Both methods of filtering appear to be equivalent, but the freq_add_proper_info is shifted to the right for the same numerical threshold; we chose the impute-info <0.8 for further analysis (similar to a freq_add_proper_info <0.9).

Figure 2.

Correlation plots and the associated R^2 for (A) The unQCed and the QCed with and without post-imputation QC filtering (IMPUTE-info <0.8 and MAF <5%). (B) The imputed-only markers in the unQCed and fully QCed data (QCed data with all poorly-clustering markers removed) without post-imputation QC filtering.

Figure 1.

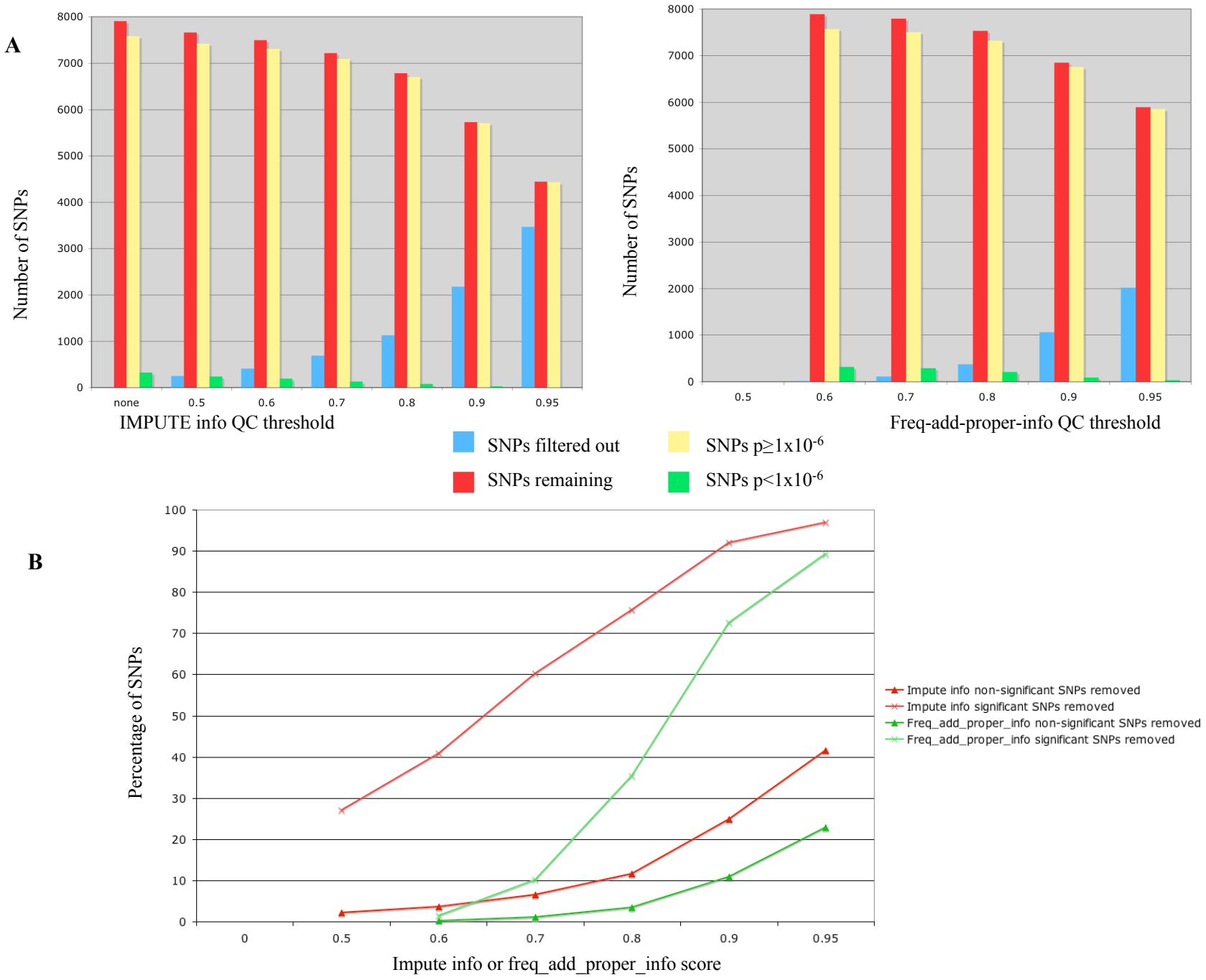
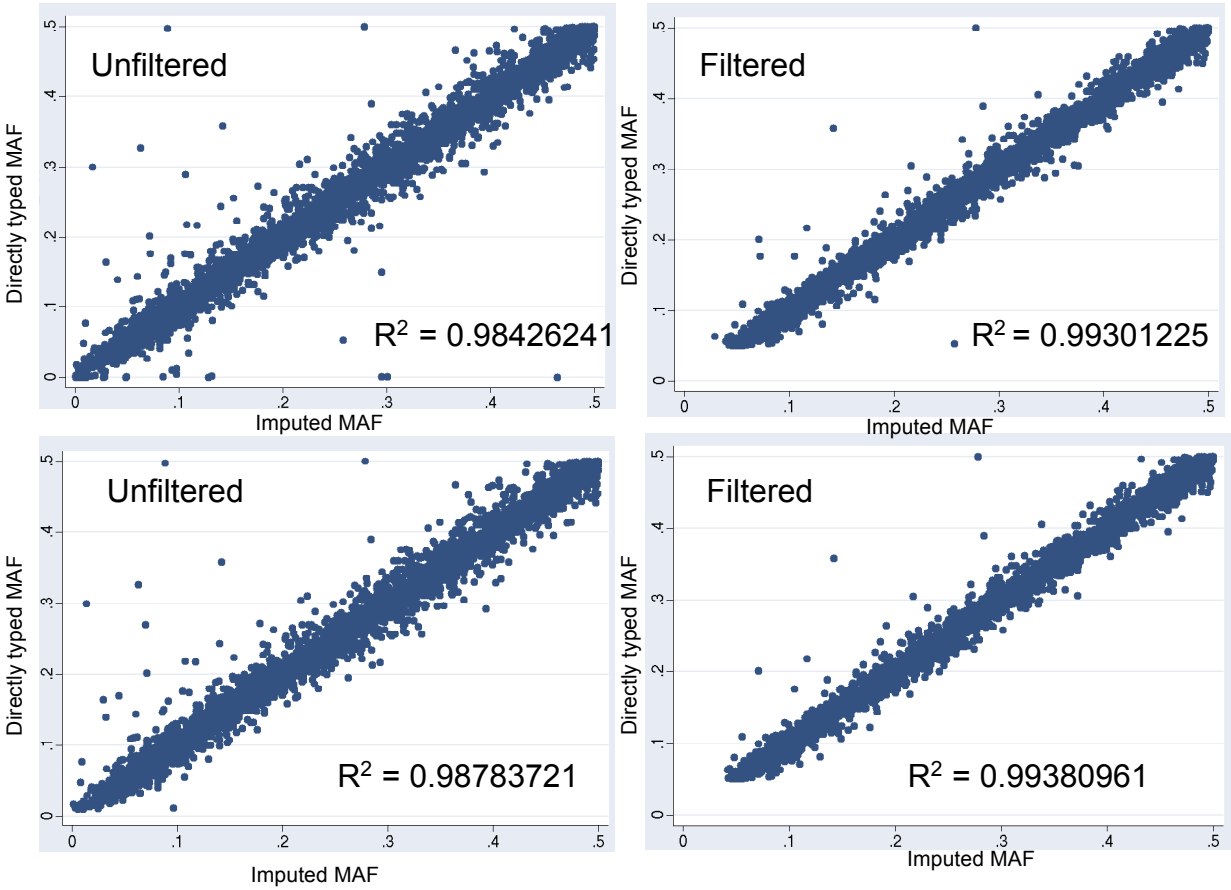

A
Number of SNPs
8000
7000
6000
5000
4000
3000
2000
1000
0
none
0.5
0.6
0.7
0.8
0.9
0.95
IMPUTE info QC threshold
Number of SNPs
Freq-add-proper-info QC threshold
SNPs filtered out
SNPs remaining
SNPs p≥1x10-6
SNPs p<1x10-6
B
Percentage of SNPs
100
90
80
70
60
50
40
30
20
10
0
Impute info or freq_add_proper_info score
Impute info non-significant SNPs removed
Impute info significant SNPs removed
Freq_add_proper_info non-significant SNPs removed
Freq_add_proper_info significant SNPs removed

Figure 2.

(A)



UnQCed data

QCed data

(B)

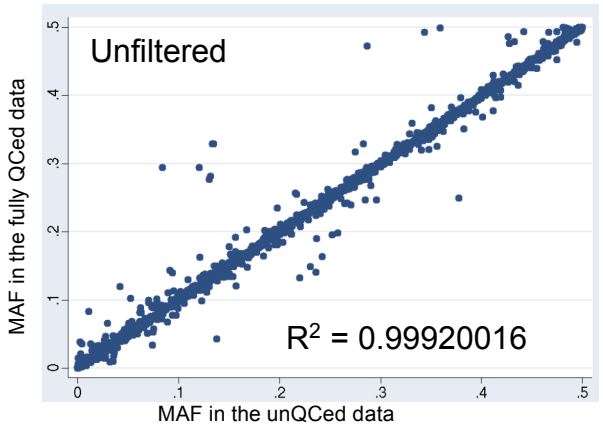


Table 1. Summary of QC steps and related SNP number breakdown.

Pre-impute QC threshold applied	**Directly typed SNPs also present in HapMap**	**Post-imputation unfiltered SNPs**		**Post-imputation QC filtered[1] SNPs**	
		NS	**S**	**NS**	**S**
None ("unQCed" dataset)	8064[2]	7689	375	6498	77
Typical GWAS QC[3] ("QCed" dataset)	7910	7585	325	6446	67
As above plus 14[4] significant SNPs removed with poor cluster plots	7896	7592	304	6449	61
As above plus 36[5] additional SNPs removed with poor cluster plots	7860	7557	303	6419	58
Typical GWAS QC[3] plus MAF <5%	7554	7269	285	6434	65
Typical GWAS QC[3] plus MAF <10%	6544	6287	257	5569	53

[1]Filtering is based on removal of SNPs with an IMPUTE-info score of <0.8 and MAF <5%.

[2]There were 8082 SNPs in the unQCed data, of which 18 were monomorphic in the arcOGEN cases but polymorphic in HapMap; these SNPs were removed by IMPUTE.

[3]Typical GWAS QC was MAF ≤5% with call rate <95% and MAF <5% with call rate <99%, HWE $p<1x10^{-4}$, and exclusion of GCAT and MAF <1% SNPs, applied as an additional post-association analysis and pre-imputation QC step.

[4]Significant SNPs with poor cluster plots removed.

[5]Those SNPs flanking the significant SNPs with poor cluster plots removed.

NS, not significant ($p\geq1x10^{-6}$); **S**, significant SNPs ($p<1x10^{-6}$)