Improved imputation quality of low-frequency and rare variants in European samples using the ‘Genome of The Netherlands’

Citation
Deelen, P., A. Menelaou, E. M. van Leeuwen, A. Kanterakis, F. van Dijk, C. Medina-Gomez, L. C. Francioli, et al. 2014. “Improved imputation quality of low-frequency and rare variants in European samples using the ‘Genome of The Netherlands’.” European Journal of Human Genetics 22 [11]: 1321-1326. doi:10.1038/ejhg.2014.19. http://dx.doi.org/10.1038/ejhg.2014.19.

Published Version
doi:10.1038/ejhg.2014.19

Permanent link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:13454823

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story
The Harvard community has made this article openly available. Please share how this access benefits you. Submit a story.

Accessibility
Improved imputation quality of low-frequency and rare variants in European samples using the ‘Genome of The Netherlands’

Patrick Deelen1,2, Androniki Menelaou3, Elisabeth M van Leeuwen4, Alexandros Kanterakis1,2, Freerk van Dijk1,2, Carolina Medina-Gomez5,6,7, Laurent C Francioli5, Jouke Jan Hottenga8, Lennart C Karssen4, Karol Estrada5,6,9,10, Eskil Kreiner-Møller5,6,11, Fernando Rivadeneira5,6,7, Jessica van Setten3, Javier Gutierrez-Achury1, Harm-Jan Westra1, Lude Franke1, David van Enckevort2,12, Martijn Dijkstra1,2, Heorhiy Byelas1,2, Cornelia M van Duijn5, Genome of the Netherlands Consortium16, Paul I W de Bakker3,13,14,15, Cisca Wijmenga1 and Morris A Swertz*,1,2

Although genome-wide association studies (GWAS) have identified many common variants associated with complex traits, low-frequency and rare variants have not been interrogated in a comprehensive manner. Imputation from dense reference panels, such as the 1000 Genomes Project (1000G), enables testing of ungenotyped variants for association. Here we present the results of imputation using a large, new population-specific panel: the Genome of The Netherlands (GoNL). We benchmarked the performance of the 1000G and GoNL reference sets by comparing imputation genotypes with ‘true’ genotypes typed on ImmunoChip in three European populations (Dutch, British, and Italian). GoNL showed significant improvement in the imputation quality for rare variants (MAF 0.05–0.5%) compared with 1000G. In Dutch samples, the mean observed Pearson correlation, $r^2$, increased from 0.61 to 0.71. We also saw improved imputation accuracy for other European populations (in the British samples, $r^2$ improved from 0.58 to 0.65, and in the Italians from 0.43 to 0.47). A combined reference set comprising 1000G and GoNL improved the imputation of rare variants even further. The Italian samples benefited the most from this combined reference (the mean $r^2$ increased from 0.47 to 0.50). We conclude that the creation of a large population-specific reference is advantageous for imputing rare variants and that a combined reference panel across multiple populations yields the best imputation results.

European Journal of Human Genetics (2014) 22, 1321–1326; doi:10.1038/ejhg.2014.19; published online 4 June 2014

Keywords: genotype imputation; GWAS; GoNL; rare variants; reference sets; reference panel

INTRODUCTION

Although genome-wide association studies (GWAS) have been very effective in identifying loci associated with diseases or traits,1 it has proved difficult to fine-map the association signals to causal variants.2,3 To overcome these limitations, there has been increasing interest in the interrogation of less frequent variants, especially given the enrichment of deleterious alleles at low frequencies.4–7 There are specialized chips that can assess a larger number of rare variants, like the ImmunoChip8 or MetaboChip,9 although they do not provide uniform genome-wide coverage. Hence, most investigators will use statistical imputation from SNP arrays in GWAS using dense reference panels.

Imputation using a densely typed reference set can be performed to infer untyped variants that can be used to improve the power of a GWAS,10 and there are numerous examples in which imputation has effectively enriched the results in GWAS.11,12 Although most large studies have so far been based on meta-analysis of HapMap-based imputations across cohorts, the primary limitation is that HapMap is essentially restricted to common variation (MAF > 5%). Thanks to the sequencing of larger samples, such as 1000G, more complete reference panels are now being assembled, setting off a new wave of meta-analyses.

The power of detecting an association in a GWAS is determined by its sample size and effective genome-wide coverage of the included

1University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, The Netherlands; 2University of Groningen, University Medical Center Groningen, Genomics Coordination Center, Groningen, The Netherlands; 3Department of Medical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands; 4Department of Epidemiology, Genetic Epidemiology Unit, Erasmus Medical Center, Rotterdam, The Netherlands; 5Department of Internal Medicine, Erasmus Medical Center, Rotterdam, The Netherlands; 6Department of Epidemiology, Erasmus University Medical Center, Rotterdam, The Netherlands; 7Netherlands Genomics Initiative (NGI)-sponsored Netherlands Consortium for Healthy Aging (NCHA), Rotterdam, The Netherlands; 8Department of Biological Psychology, VU University Amsterdam, Amsterdam, The Netherlands; 9Department of Medicine, Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA; 10Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA; 11COPSAC; Copenhagen Prospective Studies on Asthma in Childhood; Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark; 12NHBC BioAssist, Netherlands Bioinformatics Center, Nijmegen, The Netherlands; 13Department of Epidemiology, University Medical Center Utrecht, Utrecht, The Netherlands; 14Division of Genetics, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA, USA; 15Broad Institute of Harvard and MIT, Cambridge, MA, USA
16Genome of the Netherlands Consortium members are listed before the references.

*Correspondence: Dr MA Swertz, Genomics Coordination Center, University Medical Center Groningen, University of Groningen, Groningen 9700 RB, The Netherlands Tel: +31 50 60 65 01; Fax: +31 50 361 7230; E-mail: m.a.swertz@gmail.com

Received 13 August 2013; revised 1 November 2013; accepted 16 January 2014; published online 4 June 2014
variants, among other things.\textsuperscript{13,14} The effective coverage depends directly on the number and quality of the imputed genotypes.\textsuperscript{15}

In turn, the quality of the reference panel will depend largely on the number of samples, the quality of the haplotypes, and the number of variants included.\textsuperscript{16}

The Genome of The Netherlands (GoNL) has the potential to provide a good imputation reference panel. GoNL is a population-based sequencing project, in which 769 Dutch samples were sequenced at, on average, 14× coverage.\textsuperscript{17} In particular, the fact that GoNL sequenced trios (231) or quartets (19) has enabled improved haplotype phasing by using one of the children.\textsuperscript{18} The GoNL imputation reference set contains 998 unrelated haplotypes. In this paper, we report a quantitative analysis to assess the quality of imputed genotypes from using both GoNL and 1000G in Dutch and other European populations.

We adopted a ‘gold standard’ approach using samples genotyped on two distinct platforms, HumanHap550 and ImmunoChip. Hap550 is a commonly used genotyping chip designed to tag as many haplotypes as possible using common variants. ImmunoChip, however, is a fine-mapping chip: it contains a large number of low-frequency and rare variants for a limited number of loci (primarily selected on the basis of loci identified in immune-related traits). Starting from the Hap550-genotyped SNPs, we were able to impute a large number of variants present on ImmunoChip. We then compared these imputed genotypes with the measured (‘gold standard’) genotypes on ImmunoChip to quantify the imputation performance.

We have such a data set for three European populations: the Dutch, British, and Italians. For each population we used 745 samples genotyped on both platforms. These three populations allowed us to ascertain population-specific differences in the imputation quality of SNPs.

\section*{Materials and Methods}

\subsection*{Genome of the Netherlands}

GoNL is a project in which 769 individuals from different Dutch provinces were sequenced at, on average, 14× coverage.\textsuperscript{17} All samples are part of either one of the 231 trios or one of the 19 quartets. The phasing was performed using the trio information,\textsuperscript{18} and for the quartets one of the children was used to enhance the phasing. Because of sequence failures of two parents, from different trios, these samples were excluded from the imputation reference set. Instead, from these two trios, we used the haplotype of the child that was not present in the other parent. This resulted in an imputation reference set containing 998 unrelated haplotypes. We used GoNL release 4 for all our analyses (see http://www.nlgenome.nl). The current GoNL release 5 also contains over one million indels but did not change the SNPs.

\subsection*{Benchmarking samples}

Samples from a celiac disease patient cohort were selected, as they had been genotyped on both the Hap550 and ImmunoChip.\textsuperscript{19} The 745 Dutch and the 745 British samples were all cases, whereas the 745 Italian samples comprised 371 cases and 374 controls. The clustering for the genotype calling of the 15 imputations (five for each population).

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|}
\hline
Population & Dutch & British & Italian \\
\hline
SNPs & 2,096 & 1,992 & 2,710 \\
\hline
\end{tabular}
\caption{Observed r$^2$ for different MAF bins.}
\end{table}

The imputations were performed using Impute2 2.3.0.\textsuperscript{18} The different populations were imputed separately and in chunks of 5 Mb. For the comparison using an equal number of identical European haplotypes, we performed an imputation using all 379 European 1000G samples and a random selection of 379 GoNL samples. The random selection of GoNL samples was performed stratified on the Dutch provinces. These samples were selected using the Impute2 option: `--exclude_samples_h' .

We used MOLGENIS compute\textsuperscript{20} to implement the imputation pipeline, run the 8835 imputation chunks in parallel on a PBS compute cluster, and keep track of the 15 imputations (five for each population). All pipelines are available as open source via http://www.molgenis.org/wiki/ComputeStart.

\subsection*{Gold standard method}

As stated above, we used samples genotyped on two distinct platforms. We imputed the Hap550 genotypes from these samples and compared the imputed genotypes with the SNPs previously present only in the ImmunoChip data. We used the ImmunoChip data as our ‘gold standard’. The concordance between imputed genotypes and ImmunoChip genotypes was determined by calculating the Pearson correlation $r^2$ between the imputed dosage and ImmunoChip-observed genotypes. The mean concordances were calculated for three MAF bins: rare ($\geq 0.05\%$ and $<0.5\%$), low-frequency ($\geq 0.5\%$ and $<5\%$), and common ($>5\%$) SNPs. The MAF used to stratify the SNPs into the bins was calculated separately for each population. The results were plotted using R.
The significance of the differences between the reference sets was calculated using the Wilcoxon signed-rank test implementation in R.

Principal component analysis

The principal component analysis was performed using the EIGENSOFT 4.2 package. The components were calculated using the European 1000G, GoNL, and the 3 GWAS data sets that we used for benchmarking. Before the components were calculated, all data sets were filtered to include only variants with MAF > 5%. A joint data set, featuring variants present in all five data sets, was created. This data set was then again filtered for MAF > 5%; the merged data sets were also filtered on HWE > 1E-4 and a call rate of 95%. This data set was pruned using PLINK 1.07 with the ‘–indep-pairwise’ option, windows: 1000, step: 5, r² threshold: 0.2. The first component explained 0.33% of the variation and the second 0.10%. All subsequent components described less than 0.06%.

RESULTS

We stratified our analysis into three groups: common variants (MAF ≥ 5%), low-frequency variants (MAF 0.5–5%), and rare variants (MAF 0.05–0.5%). We focused mainly on the rare variants, as these are more difficult to impute and most can be gained in terms of imputation quality when using a better reference set. We observed a large increase in the imputation quality of rare variants when using GoNL as the reference compared with 1000G (Figure 1, Table 1). The mean observed Pearson correlation (r²) showed a significant increase from 0.61 to 0.71 for Dutch samples (Wilcoxon P-value = 7.16E-60).

The British and Italian imputations also showed a significant improvement when imputing rare variants, from 0.58 to 0.65 (P = 3.70E-35) and from 0.43 to 0.47 (P = 2.64E-13), respectively. GoNL also significantly outperformed the 1000G reference set in the imputation of variants with higher MAFs (Supplementary Figures/Supplementary Appendices S1, S2, S3).

Using a combined reference set composed of the 1000G and GoNL samples, we could improve the imputation further. The imputation of rare variants using the combined reference in Dutch and British samples showed a small increase in quality compared with GoNL-only imputation (0.02 (P = 1.16E-03) and 0.02 (P = 2.70E-05), respectively). The Italians benefitted most from the combined reference with an increase of 0.04 (P = 3.62E-30) compared with a GoNL-only reference, resulting in a mean concordance for rare variants of 0.5. The differences in imputation quality when using the combined reference set for more frequent alleles were either very small or not significant (Supplementary Figure S1, Supplementary Tables S2 and S3).

A striking trend in these results is that the imputation quality of rare variants in the Italian samples is lower than that in Dutch and British samples. The Dutch and Italian samples were genotyped at the same center and have similar call rates, and there were no indications that the genotyping quality of the Italian samples was lower. However, a principal component analysis revealed that the Italian samples were

![Figure 2](https://example.com/figure2.png)

**Figure 2** Clustering of reference and study samples. PC1 and PC2 reveal three main clusters: Tuscans from Italy (TSI), Finnish (FIN), and a Western European cluster with the CEU (Utah Residents with Northern and Western European ancestry), the GBR (British) and the GoNL samples (a). b shows that most of our GWAS samples clustered in a similar way to the corresponding 1000G/GoNL samples.
not as well represented by either 1000G or GoNL compared with the Dutch and British GWAS samples used for benchmarking (Figure 2).

We assessed whether the better performance of GoNL compared with 1000G was due to the larger number of European haplotypes in the reference set (998 vs. 758 in 1000G). We did this by performing an imputation using solely the 379 European samples in 1000G and a random subset of 379 GoNL samples. We found that the GoNL subset also significantly outperformed the European 1000G subset (Table 2).

Our experimental design also allowed us to assess the calibration of the posterior probabilities of the genotypes as they are output by Impute2. We observed that the posterior probabilities were, in general, well calibrated, although there were a few deviations from the expected accuracy (Figure 3a). For common and low-frequency variants (b and c), we observed a strong correlation ($r^2$ 0.97 and 0.91, respectively) between the Impute2 info metric and the observed $r^2$. However, for the rare variants (d), the relation between predicted and observed quality was less profound. We also observed a correlation of 0.70 and several large deviations from the diagonal.

**DISCUSSION**

We have shown that the new GoNL reference set provides higher downstream imputation accuracy than the 1000G reference set, not only for Dutch samples but also for other European populations studied in this paper. Aside from the increase in the imputation quality of rare variants in Dutch samples from 0.61 (1000G) to 0.71 (GoNL), we also observed an increase in imputation quality in British...
(0.58–0.65) and Italian (0.43–0.47) samples. We show that GoNL yielded better imputed genotypes for at least these European populations. A combined reference set, of 1000G and GoNL, increased the mean imputation quality of rare variants even further to 0.72, 0.67, and 0.50 for the Dutch, British and Italians, respectively.

By selecting an identical number of European haplotypes from 1000G and GoNL, we showed a strong added value for GoNL in all the tested populations, confirming that the trio design of GoNL and the resultant accurate haplotypes aid the downstream imputation quality. We also observed a population-specific additional value of GoNL when imputing Dutch samples. The added value (ie mean increase in imputation quality) was largest when comparing GoNL with 1000G in imputing the Dutch samples. Of course, it was already known that a better matched reference set will result in better imputed genotypes,13 however, the results from this paper were based on low-frequency variants and we show that there is also an inter-European effect of reference sets.

It is important to note that we only assessed variants present on the ImmunoChip. Although these variants were not randomly selected, we have no reason to assume that the imputation quality will be positively biased or that they do not represent low-frequency variants in general. The ImmunoChip was made to fine-map loci previously associated with autoimmune diseases using a large number of low-frequency variants and we show that there is also an inter-European effect of reference sets.

We were encouraged by the observation that the posterior probabilities were, in general, well calibrated with respect to the gold standard genotypes. We observed no adverse effects on the accuracy of the Impute2 info metrics, although for rare variants we did observe a few instances with large deviations between the predicted and observed probabilities.24 This observed inaccuracy also emphasizes the importance of validating associations from imputed genotypes.

It was shown earlier that a larger and more diverse reference set can improve the imputation of low-frequency variants.25 We observed that a combination of 1000G and GoNL showed limited added value for the imputation of rare variants in the Dutch and British samples. It was, however, interesting to observe that the imputation of the Italian samples was improved more by this combined reference panel, leading us to speculate that populations that are poorly represented in the reference panel benefit more from a large and diverse reference set. Despite the limited added value for the Dutch and British data sets, such a large reference set may still be of interest for consortia aiming to impute cohorts of both European and non-European origin. All these cohorts can be imputed using the same combined reference set and then use Impute2 to automatically select the best matching haplotypes.26 We should note that we were only able to assess variants present in both reference sets, as there are very few variants on the ImmunoChip that are unique to either GoNL or 1000G. Nonetheless, our results show that population-specific reference sets and cosmopolitan panels, such as 1000G, can augment each variant. This even holds true for the imputation of samples with ancestry other than those present in the population-specific reference sets, which provides further motivation for international efforts towards large and integrated reference sets.

CONFLICT OF INTEREST
The authors declare no conflict of interest.

ACKNOWLEDGEMENTS
This study was made possible by rainbow grant 2 from BBMRI-NL to MS, a research infrastructure financed by the Netherlands Organization for Scientific Research (NWO project 184.021.007). We thank the Target project (http://www.rug.nl/target) for providing the compute infrastructure, and the BigGrid/eBioGrid project (http://www.ebiogrid.nl) for sponsoring the pipeline implementation. We thank Jackie Senior for careful reading and editing the manuscript. This study made use of data generated by the ‘Genome of the Netherlands’ project, which is funded by the Netherlands Organization for Scientific Research (grant no. 184021007). The data were made available as a Rainbow Project of BBMRI-NL. Samples were contributed by LifeLines (http://lifelines.nl/lifelines-research/general), the Leiden Longevity Study (http://www.healthy-ageing.nl; http://www.langleven.nl), the Netherlands Twin Registry (NTR; http://www.twelingenregister.org), the Rotterdam studies (http://www.eramus-epidemiology.nl/rotterdamstudy), and the Genetic Research in Isolated Populations program (http://www.egip.nl/research/genetici/research.html#pip). The sequencing was carried out in collaboration with BGI (Shenzhen, China).

AUTHOR CONTRIBUTIONS
PD, AM, MAS, PjWdB, and CW wrote the main manuscript. All the authors contributed to the discussion of experimental design in ‘Genome of The Netherlands’ imputation working group. EMvL, AK, LCK, CM-G, JJH, and FvD revised the manuscript. PD, FvD, MD, HB, LCF, H-JW, AK, EM-K, and CM-G contributed to the implementation of the analysis.

GENOME OF THE NETHERLANDS CONSORTIUM

Analysis group: Morris A. Swertz6-7 (Co-Chair), Laurent C. Francioli1, Freerk van Dijk6-9, Androniki Menelaou1, Pieter B.T. Neerincx6-9, Sara L. Pulit1, Patrick Deelen6-9, Clara C. Elbers1, Pier Francesco Palamara4, Itsik Pe’er2-8, Abdel Abdellauoui2, Wigard P. Kloosterman1, Mannis van Oven1, Martijn Vermaat1, Mingkun Li1, Jeroen F. Laros1-3, Marja Stoneking12, Peter de Knijff13, Manfred Kayser14, Jan H. Veldink14, Leonard H. van den Berg14, Heerhuy Byelas6-7, Johan T. den Dunnen1, Martijn Dijkstra6-7, Najaf Amin1, K. Joeri van der Velde6-7, Jouke Jan Hottenga1, Jessica van Setten1, Elisabeth M. van Leeuwen15, Alexandros Kanterakis6-7, Mathijs Kattenberg1, Lennart C. Karssen1, Barbera D.C. van Schaik16, Jan Bot17, Isaac J. Nijman1, David van Enckevort18, Hailiang Mei19, Vyacheslav Koval19, Kai Ye20,21, Eric-Wubbo Lameijer21, Matthijs H. Moed21, Jayne Y. Hehir-Kwa22, Robert E. Handsaker5,23, Shamil R. Sunyaev4,5, Mashaal Sohail4,5, Fereydoun Guryev26, Paul I.W. de Bakker1,3-5 (Co-Chair);

Cohort collection and sample management group: P. Eline Slagboom23, Marian Beekman23, Anton J.M. de Craen23, H. Eka D. Soehman21, Albert Hofman23, Cornelia van Duijn15, Dorret I. Boomsma6, Gonneke Willemsen9, Bruce H. Wolfenbuttel27, Mathieu Platteeuw16, Steven J. Pitz18, Shohha Potluri18, David R. Cox28,34.

Whole-genome sequencing: Qibin Li29, Yingrui Li29, Yuanping Du29, Ruoyan Chen29, Hongzhi Cao29, Ning Li30, Sujie Cao30, Jun Wang29,31,32.

Ethical, Legal, and Social Issues: Jasper A. Bovenberg33

Scientific Research (grant no. 184021007). The data were made available as a Rainbow Project of BBMRI-NL. Samples were contributed by LifeLines (http://lifelines.nl/lifelines-research/general), the Leiden Longevity Study (http://www.healthy-ageing.nl; http://www.langleven.nl), the Netherlands Twin Registry (NTR; http://www.twelingenregister.org), the Rotterdam studies (http://www.eramus-epidemiology.nl/rotterdamstudy), and the Genetic Research in Isolated Populations program (http://www.egip.nl/research/genetici/research.html#pip). The sequencing was carried out in collaboration with BGI (Shenzhen, China).
