Population-specific genotype imputations using minimac or IMPUTE2

Elisabeth M van Leeuwen[1], Alexandros Kanterakis[2,3], Patrick Deelen[2,3], Mathijs V Kattenberg[4], The Genome of the Netherlands Consortium[5], P Eline Slagboom[6,7], Paul I W de Bakker[8,9], Cisca Wijmenga[2,3], Morris A Swertz[2,3], Dorret I Boomsma[4], Cornelia M van Duijn[1], Lennart C Karssen[1,10,11] & Jouke Jan Hottenga[4,11]

[1]Department of Epidemiology, Erasmus Medical Center, Rotterdam, the Netherlands. [2]University of Groningen, University Medical Center Groningen, Genomics Coordination Center, Groningen, the Netherlands. [3]University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, the Netherlands. [4]Department of Biological Psychology, VU University, Amsterdam, the Netherlands. [5]A full list of members and their affiliations is available in the **Supplementary Note**. [6]Department of Molecular Epidemiology, Leiden University Medical Center, Leiden, the Netherlands. [7]Netherlands Consortium for Healthy Ageing, Leiden University Medical Center, Leiden, the Netherlands. [8]Department of Medical Genetics, University Medical Center Utrecht, Utrecht, the Netherlands. [9]Department of Epidemiology, University Medical Center Utrecht, Utrecht, the Netherlands. [10]PolyOmica, Groningen, the Netherlands. [11]These authors contributed equally to this work. Correspondence should be addressed to C.M.v.D. (c.vanduijn@erasmusmc.nl).

Published online 30 July 2015; doi:10.1038/nprot.2015.077

In order to meaningfully analyze common and rare genetic variants, results from genome-wide association studies (GWASs) of multiple cohorts need to be combined in a meta-analysis in order to obtain enough power. This requires all cohorts to have the same single-nucleotide polymorphisms (SNPs) in their GWASs. To this end, genotypes that have not been measured in a given cohort can be imputed on the basis of a set of reference haplotypes. This protocol provides guidelines for performing imputations with two widely used tools: minimac and IMPUTE2. These guidelines were developed and used by the Genome of the Netherlands (GoNL) consortium, which has created a population-specific reference panel for genetic imputations and used this reference to impute various Dutch biobanks. We also describe several factors that might influence the final imputation quality. This protocol, which has been used by the largest Dutch biobanks, should take approximately several days, depending on the sample size of the biobank and the computer resources available.

INTRODUCTION

Data from GWASs of different cohorts can be combined into a meta-analysis even when the samples of the cohorts have been typed on different genotyping platforms. By imputing missing genotypes, a homogeneous data set for meta-analysis can be created. Genotype imputation allows the estimation of genotypes in a target data set, based on one or more available reference sets of SNPs, and it is based on searching common haplotypes between an individual's genome and a reference panel with a high density of genotyped SNPs, such as those provided by the HapMap[1], 1000 Genomes[2] and the GoNL[3–5] projects. Missing genotypes are then inferred from common haplotypes that are found in the reference set. Implementation of these methods usually results in estimates of the posterior probability distributions $P_g = (P_{AA}, P_{AB}, P_{BB})$ of the genotypes based on the available data[6].

Weaknesses in both genotype calling and imputation of missing genotypes can lead to biases in GWASs and subsequently in meta-analysis. Therefore, Anderson *et al.*[7] have previously published a protocol dealing with quality control of genotype data, and our work can be seen as an extension of that protocol. A guideline for imputations with the Beagle[8] and IMPUTE2 (ref. 9) tools, as well as postimputation quality control, has also been published by Verma *et al.*[10], and a protocol for doing meta-analysis of GWAS results for large numbers of cohorts is described in Winkler *et al.*[11].

In this protocol, we show how to perform genotype imputations with a population-specific reference panel, including how to deal with factors that may adversely affect the imputation result (e.g., how to properly split up large data sets for imputation). This protocol differs from the previous guidelines in the study by Verma *et al.*[10], providing instructions for imputations with IMPUTE2 (ref. 9) and minimac[12]. We describe the different pipelines for imputations using the genome-wide SNP data provided by Anderson *et al.*[7] as a target data set. We will start with the quality control of this target set using the pipeline from Anderson *et al.*[7]. We will show how to lift the target set over to the correct National Center for Biotechnology Information (NCBI) build and then provide pipelines for imputation using minimac[12] and IMPUTE2 (ref. 9; **Fig. 1**). All pipelines are developed for GNU/Linux-based computer resources, and all commands should be typed at the Bash shell prompt, where Bash variables are indicated by '${variablename}'. This protocol does not include commands to submit compute intensive tasks to a job scheduling system such as OpenPBS (see 'Computer Resources' section), as different computer clusters may use different scheduling systems.

This protocol has been used to impute the genotypes of individuals of various Dutch biobanks using the GoNL reference panel. This has resulted in the discovery of five novel associations at four loci for cholesterol levels including a rare missense variant in the *ABCA6* gene, which is predicted to be deleterious[13].

GoNL reference set

The construction of a novel imputation reference data set is a complex procedure that requires dense genotyping and accurate estimation of haplotypes from genotype data (known as phasing) of samples from a specific population. The most thoroughly documented and widely available imputation reference sets come from the HapMap[1] and 1000 Genomes projects[2]. Both projects contain samples from various populations, and consequently a given genotype of a low-frequency variant may not be represented adequately in the reference data set. Moreover, when the percentage of samples belonging to a different geographical population is beyond a certain proportion, the imputation quality does not

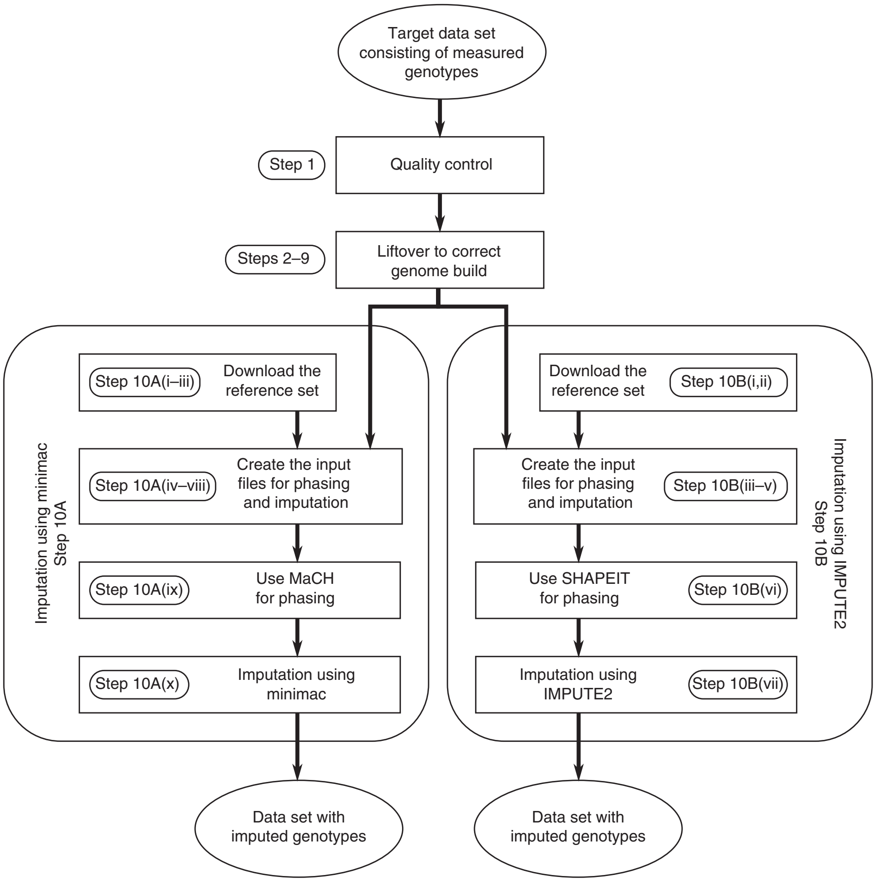

Target data set consisting of measured genotypes
Step 1
Quality control
Steps 2–9
Liftover to correct genome build
Imputation using minimac Step 10A
Step 10A(i–iii)
Download the reference set
Step 10A(iv–viii)
Create the input files for phasing and imputation
Step 10A(ix)
Use MaCH for phasing
Step 10A(x)
Imputation using minimac
Data set with imputed genotypes
Imputation using IMPUTE2 Step 10B
Download the reference set
Step 10B(i,ii)
Create the input files for phasing and imputation
Step 10B(iii–v)
Use SHAPEIT for phasing
Step 10B(vi)
Imputation using IMPUTE2
Step 10B(vii)
Data set with imputed genotypes

Figure 1 | Workflow of the imputation protocol for imputations of unobserved genotypes with the GoNL reference panel. The first stage of the protocol is to perform quality control of the target data set consisting of measured genotypes; this followed by performing the liftover to the correct human genome build. The human genome build of the GoNL reference panel is UCSC hg19. These steps are independent of the tools that are used for the actual phasing and imputation. The next step is to download the reference set, which is necessary to create the correct input file for phasing and imputations. The reference set file format is different for each tool. Next, MaCH or SHAPEIT is used for phasing, followed by minimac and IMPUTE2 for the imputations.

improve. Jostins *et al.*[14] found that when imputing samples from the 1958 British Birth Cohort, the accuracy starts to fall off when the proportion of non-CEU (Northern Europeans from Utah; refs. 1,2) samples exceed 20%, as the effect of increased diversity is outweighed by the effect of mismatching. This relationship is specific to low-frequency variants. Moreover, Pistis *et al.*[15] found that the effectiveness of population-specific reference panels can be appreciable for other populations, but that effectiveness will vary depending on the size of the panels and the demographic history of the isolate.

As interest in the field of genetic epidemiology is shifting toward low-frequency variants, the GoNL consortium has created a population-specific reference set for imputation with the goal of identifying associations between various phenotypes and low-frequency genetic variants. To this end, 231 parent-offspring trios and 19 parent-offspring quartets of Dutch descent had their complete genome sequenced with at least 12× coverage[3–5]. The strength of this reference set comes from several factors. The first is the trio design, which improves the haplotype quality. The second is the coverage, which is higher than that of the 1000 Genomes Project, and the third factor is the sequencing of samples from a homogeneous population. The quality of the haplotypes boosts imputation accuracy in independent samples, especially for lower-frequency alleles[4].

The GoNL reference set is available by applying through http://www.nlgenome.nl/, menu option 'Request data', which leads to the application form. After filling in the form, the request will be evaluated by the GoNL steering committee. After a positive evaluation, a data access agreement needs to be signed and, subsequently, the reference panel can be downloaded in Variant Call Format (VCF). For this protocol, the fourth release of the GoNL reference panel was used, which contains 499 individuals of Dutch ancestry and 19,562,004 autosomal SNPs.

Tools for imputation

The three most commonly used tools for genotype imputation are minimac[12], IMPUTE2 (ref. 9) and Beagle[8]. Multiple aspects of the three tools, e.g., their imputation accuracy, error rates and computational performance, have been compared previously[6,10,16,17]. The choice for a given tool depends on the target set that is to be imputed and on the type of computational resources available, as discussed in this paper. Within the GoNL[3–5] consortium, only minimac and IMPUTE2 were used for imputations, and therefore Beagle will not be discussed in this manuscript. It is, however, possible to impute samples with the GoNL reference panel using Beagle. Minimac can be downloaded freely from the web, and its source code is available under an open-source license. IMPUTE2 is available for download for academic use only; no source code is provided.

IMPUTE2 performs both the phasing and the imputation, whereas minimac only imputes data sets that have been phased by MaCH[18] or SHAPEIT2 (ref. 19). However, although IMPUTE2 can perform phasing, its authors recommend using SHAPEIT2 (ref. 19), followed by using IMPUTE2 for the imputations. Of the three tools, only IMPUTE2 can combine two reference panels. This allows imputation with both the 1000 Genomes reference panel and the GoNL reference panel, which has been shown to improve imputation quality[3]. MaCH and minimac make their own recombination maps on the basis of input data; IMPUTE2 requires a recombination map.

The requested file format of the reference set is also different among the tools. The GoNL project[3–5], the 1000 Genomes project[2] and the HapMap project[1] provide their data in VCF format[20]. The VCFtools[20] software package can convert these VCF files into phased haplotypes in IMPUTE2 reference-panel format. The authors of IMPUTE2 also provided a Perl script to perform this conversion. Minimac can handle the original VCF files without conversion.

Both tools produce several output files. The first one is the so-called 'info file,' which contains the SNP name, the base-pair positions, the frequencies of the alleles and the R^2. Here R^2 is the estimated squared correlation (between zero and one) between the allele dosage with highest posterior probability in the genotype probabilities file and the true allele dosage for the marker; larger values of allelic R^2 indicate a more accurate genotype imputation. In a second file, IMPUTE2 gives the probabilities of the three genotypes *AA*, *AB* and *BB*, whereas minimac gives the probability of a homozygote for allele 1 and the probability of the heterozygote. Only minimac has the option to output best-guess alleles. Dosage files are produced only by minimac; however, it takes only one additional step to convert the genotype probabilities from IMPUTE2 into dosages. If a sample has genotype probabilities (P_{AA}, P_{AB}, P_{BB}) for a marker, then the estimated *B*-allele dosage (d_{B}) is $d_{\mathrm{B}} = P_{AB} + 2\ P_{BB}$. All formats can be converted using fcGENE[21].

Quality control of the target data set

To achieve a high-quality imputation standard, GWAS quality control filters need to be applied to the target data set and, if necessary, also to the reference set before imputation. The purpose of these filters is to exclude both markers and samples with low-quality data. Anderson *et al.*[7] and Verma *et al.*[10] provide a detailed protocol that deals with both per-maker and per-individual filtering.

Other factors that influence the imputation quality are the type of arrays used for genotyping, and strand and build issues. Present-day high-density arrays are of high quality; however, the low-density arrays used in the beginning of the GWAS era were less so. It is therefore useful to check the type of array that was used for genotyping of the target set. The genotype calls from the arrays are aligned to a specific strand[22]. In order to obtain high-quality imputations, it is important to correct possible strand alignment issues. Although IMPUTE2 and MaCH have options to fix misaligned alleles between the study and the reference panel by inverting the alleles when possible, the alignment of the target set should be fixed before imputing the target set with, e.g., SHAPEIT2 (ref. 19). This only holds for ambiguous strands (e.g., AT and TA); detecting and correcting the strand of the non-ambiguous SNPs (e.g., AT and GC) is more of a challenge. Deelen *et al.*[23] have published a method for solving the strand issues of nonambiguous SNPs. For imputation purposes, the alleles should be aligned to the forward strand, as the imputation tools assume that the target set is on the same strand as the reference panel, which is the forward strand.

It is important for imputation that both the target set and the reference set are on the same NCBI build, as SNP names may change or SNPs may be relocated or merged between builds. Release 4 of the GoNL reference set uses (NCBI) build 37 (human genome 19, hg19). If the reference and the target set are aligned using a different genome assembly, it is recommend to re-align the target panel to the assembly of the reference rather than the other way around. This is because the phased haplotype structure of the reference panel will be distorted if the position of the markers is altered. Moreover, re-aligning of the target set takes less time compared with re-aligning of the reference panel. The liftOver tool from the University of California, Santa Cruz (UCSC)[24] converts genome positions between different genome builds (see 'Performing quality control' section and http://genome.sph.umich.edu/wiki/LiftOver).

A major pitfall of genotype imputation is a difference between groups of individuals, which, after imputation, can be (falsely) associated with a phenotype. Array differences or quality differences (e.g., call rates) between cases and controls should be avoided. Therefore, the most ideal situation would be to genotype all individuals on the same array. If this is not possible, it is highly advised to apply strict quality control. The type of array also influences the imputations; chunking the observed genotypes of low-density arrays as discussed in 'Handling large target data sets' below may lead to empty chunks. High-density genotype arrays are therefore advised. Other important imputation pitfalls are monomorphic and extremely rare SNPs[25]; therefore, these should be removed from both the target set and the reference panel.

After performing all quality control steps, the target data set needs to be converted into the correct input format (**Box 1**) for the imputation tool of choice.

Quality metrics

The quality of an imputation experiment can be assessed by various metrics[10]. These metrics can be divided into two categories on the basis of whether true genotypes are available or not. The most common imputation metric is the R^2 that represents the correlation between the imputed and the real genotypes.

When the true genotypes are unknown, various statistics can be used to estimate the R^2. Marchini and Howie[6] present a thorough review of the R^2 metrics used by MaCH, Beagle, SNPTEST and IMPUTE2. Comparison of these measures showed that they are highly correlated. Another R^2 metric[26] is the ratio of the variance of the imputed allele dosage and the variance of the true allele dosage. Although the variance of the true allele dosage is unknown, it can be estimated as $2p(1–p)$ under Hardy-Weinberg equilibrium, where p is the estimated allele frequency. To illustrate how well rare and common SNPs were imputed, a plot can be made with the percentage of SNPs at various cutoffs for the R^2 for various minor allele frequency bins[8,27].

When the true genotypes are available, the quality of the imputation can also be evaluated by calculating the false-positive and false-negative genotypes[4]. False-positive genotypes are those that have a high imputation R^2, but that were in fact imputed incorrectly. False-negative genotypes are those that have a low R^2, but that were actually imputed correctly. Another qualitative metric is the concordance between real and imputed genotypes. A graph of the percentage of discordance versus the percentage of missing genotypes for various thresholds of the genotype probability can be used to compare different imputation methods[9].

Handling large target data sets

To successfully identify rare variants associated with particular phenotypes, large sample sizes are needed. Moreover, the number of variants in the reference panels are increasing, both leading to increasing computation times. Splitting up the target sets and distributing the computational burden of phasing and imputation over several computers allows imputation of such large sets to finish within a reasonable time frame. Splitting up the target set reduces the time to finish the imputations (**Supplementary Fig. 1**); however, it requires a computer cluster. A target set can be split up in two ways: it can be split into subsets of samples

Box 1 | Input files for imputations

The input files for the various imputation tools

For MaCH and minimac, the target set that will be imputed needs to be stored per chromosome in Merlin[28] format. The Merlin pedigree file contains the relationships, the phenotypes and the genotypes per individual per row. The first columns of the pedigree file contain the family identifier, the individual identifier, the father and mother identifiers, and the sex of the individual (with females decoded as 2 and the males decoded as 1). The subsequent columns can encode phenotypes for discrete and quantitative traits followed by the genotypes. The alleles should be coded as 'A', 'C', 'G' or 'T' and missing alleles should be encoded with 'N', 'X' or '0'. As MaCH and minimac assume samples to be unrelated, both the father and mother identifiers should be zero. The description of the columns is stored in the data file, with one row per column, indicating the data type (encoded as M, marker; A, affection status; T, quantitative trait; and C, covariate) and providing a one-word label for each column.

For IMPUTE2, the genotype information should be stored in a one-line-per-SNP format. The first five entries of each line should be the SNP ID, rs ID of the SNP, base-pair position of the SNP, the allele coded *A* and the allele coded *B*. The subsequent columns contain the prior probabilities for the three genotypes *AA*, *AB* and *BB* for each individual in the target set. This format allows for genotype uncertainty, and therefore the probabilities for a given individual need not sum to 1. The order of samples in the genotype file should match the order of the samples in the sample file. The sample file has three parts: (i) a header line detailing the names of the columns in the file, (ii) a line detailing the types of variables stored in each column and (iii) a line for each individual detailing the information for that individual (more details on the IMPUTE2 file formats can be found at http://www.stats.ox.ac.uk/~marchini/software/gwas/file_format.html).

PLINK format to store genotyped data

The most commonly used file format for storing genotype data of the samples in the target set is the PLINK format (http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml#ped). The pedigree file (extension .ped) in PLINK format is a headerless white space (space or tab)-delimited file that contains the pedigree information, the phenotype information and the genotype information for all samples in the data set. Every row corresponds to one individual and contains at least six columns, which contain the family identifier, the individual identifier, the paternal and maternal identifier, the sex of the samples (with males encoded as 1 and females encoded as 2) and the phenotype of the sample, just like the Merlin format. Genotypes (column 7 onward) can be any character (e.g., 1, 2, 3 and 4 or A, C, G and T or anything else) except 0, which is, by default, the missing genotype character. All markers should be biallelic. All SNPs (whether haploid or not) must have two alleles specified, and either both or neither alleles should be missing. The SNPs are described in the map file (extension .map); each line of this file describes a single marker, and it must contain exactly four columns: the chromosome, the SNP identifier, the genetic distance in Morgans and the base-pair position in base-pair units. The ped and map file can be converted into a more memory- and time-efficient binary file with the extensions .bed, .bim and .fam.

or split into chunks of chromosomes. The division into groups of samples can be done randomly, although the distribution of cases and controls should be similar in the subgroups. However, as imputations are mostly done once per cohort followed by the subsequent analysis of many phenotypes using the same imputed genotype data, splitting a target set into equal proportions of cases and controls provides a challenge, and we therefore do not recommend this. This only holds for the imputations and not for phasing, as the samples do not affect each other in phasing. Splitting up into samples may, however, be helpful to optimize the capacity utilization of a compute cluster.

The second, more useful, strategy for dividing up the target set is to split the chromosomes into chunks of a few Mb. Depending on the imputation tool, the strategy to split up into chunks is different. When using minimac, the ChunkChromosome tool (http://genome.sph.umich.edu/wiki/ChunkChromosome) can be used to split each chromosome before imputation (see Step 10A(viii)). When imputing with IMPUTE2, it is not necessary to first split up the chromosome, as one of the command-line arguments of IMPUTE2 is the position interval to impute.

To evaluate the quality of the imputations after the chromosome is split into chunks, we imputed chromosome 21 of all 5,974 samples of the Rotterdam Study cohort I with the European part of the 1000 Genomes reference set (release August 2010) using minimac after phasing with MaCH using two approaches. In both approaches, the data set was split up before phasing with MaCH. The first approach was to split the SNPs on chromosome 21 into chunks of 500 kb, 1, 2, 3, 4, 5, 7.5 and 10 Mb, respectively, each with an overlap of 5% on each side of the chunk. The second approach was to split the same chromosome into chunks of 5 Mb with an overlap of 2.5% (250 kb), 5% (500 kb), 7.5% (750 kb), 10% (1 Mb) and 12.5% (1.25 Mb) on each side, respectively. **Figure 2** shows that the target set can be split into subsets of at least 5 Mb with an overlap of at least 250 kb without decreasing the imputation quality.

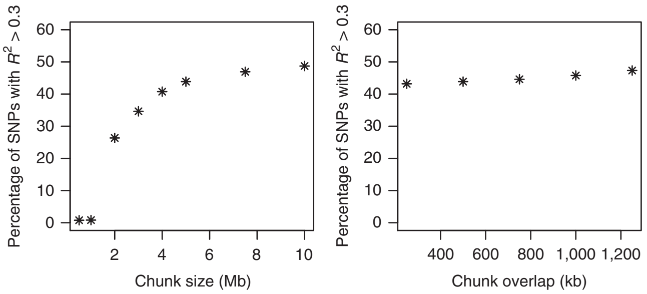


Figure 2 | The percentage of SNPs with R^2 > 0.3 after imputing chromosome 21 of 5,974 samples of Rotterdam Study cohort I when the target set is split into several chunks of chromosomes and the percentage overlap between chunks is 10%, and when the chromosome of the target set is split into 5 Mb chunks and the size of the overlap is varied. This figure illustrates that the target set can be split into subsets of at least 5 Mb with an overlap of at least 250 kb without decreasing the imputation quality. Asterisks indicate data points on the graphs.

MATERIALS

EQUIPMENT

▲ CRITICAL This protocol assumes that the computer uses GNU/Linux as its operating system (which is the case for most, if not all, computer clusters), and that the analyst uses Bash as his/her shell (which is the default on most GNU/Linux systems).

Data

- Genome-wide SNP data (raw-GWA-data.tgz). See the supplementary material in Anderson *et al.*[7] for an example data set
- GoNL reference panel for imputations. The reference set is available by applying through http://www.nlgenome.nl/

Software

- Several tools such as gawk, sort, uniq, wget, tar, sed and head, which are usually installed by default on a GNU/Linux system
- PLINK v1.07 (ref. 22): the binaries compiled for various platforms and installation instructions can be downloaded from http://pngu.mgh.harvard.edu/~purcell/plink/download.shtml
- liftOver: this tool can be used to lift over from one human genome build to the other, and it can be downloaded from http://pngu.mgh.harvard.edu/~purcell/plink/download.shtml#download
- VCFtools v0.1.12b: this tool can be downloaded from http://sourceforge.net/projects/vcftools/files/latest/download/vcftools_0.1.12b.tar.gz
- ChunkChromosome (release 2014-05-27): this tool can be downloaded from http://www.sph.umich.edu/csg/cfuchsb/generic-ChunkChromosome-2014-05-27.tar.gz
- MaCH (release 1.0): this tool can be downloaded from http://www.sph.umich.edu/csg/abecasis/MaCH/download/mach.1.0.18.Linux.tgz
- Minimac (release 2013.7.17): this tool can be downloaded from http://www.sph.umich.edu/csg/cfuchsb/minimac-beta-2013.7.17.tgz
- SHAPEIT v2.790: this tool can be downloaded from https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.v2.r790.RHELS_5.4.static.tar.gz
- IMPUTE2 v2.3.1: this tool can be downloaded from https://mathgen.stats.ox.ac.uk/impute/impute_v2.3.1_x86_64_static.tgz

EQUIPMENT SETUP

Computer resources Imputing SNPs in data sets of several thousands of samples using reference sets consisting of several millions of SNPs (e.g., HapMap[1]) up to several tens of millions of SNPs (GoNL project[3–5] or the 1000 Genomes project[2]) cannot be done on a commodity desktop computer, as that would take months of time and it requires more memory (RAM) than is usually available. As discussed earlier, the answer lies in splitting the imputation task into smaller pieces and running these subtasks on a computer cluster.

The work described in this paper was done on two such clusters. The Lisa cluster at SARA (https://userinfo.surfsara.nl/systems/lisa/) is a heterogeneous cluster that consists of more than 500 machines with a total of at least 6,000 cores and 16–24 GB of RAM each, running Debian Linux (http://www.debian.org). The Millipede cluster at Groningen University is a heterogeneous cluster with 252 nodes with a total of 3,216 cores and 24–128 GB of RAM each. It runs RedHat Enterprise Linux 5 (http://www.redhat.com/en/technologies/linux-platforms/enterprise-linux). Both clusters use the OpenPBS (http://www.mcs.anl.gov/research/projects/openpbs/) system to schedule tasks across their nodes.

The memory requirements for MaCH are ~100 MB. The minimac protocol requires 3 GB, whereas SHAPEIT requires ~1.5 MB and IMPUTE2 requires ~3 GB.

PROCEDURE

Performing quality control ● TIMING ~8 h

1| The first step is to perform standard quality control on the target set. To do this, complete the protocol for quality control, as described by Anderson *et al.*[7]. We assume that the genotypes have been called by a genotyping center and returned in PLINK format named 'raw-GWA-data.ped, raw-GWA-data.map'. All genotypes are annotated to the forward strand. After performing quality control of this genome-wide SNP data, 1,919 samples and 313,878 markers remain. The resulting files are named 'clean-GWA-data.bed, clean-GWA-data.bim' and 'clean-GWA-data.fam'.

Converting the target set to the correct genome build ● TIMING ~20 min

2| If the target set is on another genome build than the reference set, it is important to lift the target set over to the same build as the reference set. The following steps show how to convert the target set from UCSC hg17 (NCBI build 35) to UCSC hg19 (Genome Reference Consortium GRCh37).

First download the chain file

```
wget

http://hgdownload.cse.ucsc.edu/goldenPath/hg17/liftOver/hg17ToHg19.over.chain.gz
```

Next, type the following to unzip the chain file.

```
gunzip hg17ToHg19.over.chain.gz
```

3| Start the liftover by converting the target set with PLINK to a map and ped file. This will create the 'clean-GWA-data.map' and 'clean-GWA-data.ped' files.

```
plink --noweb --bfile clean-GWA-data --recode --out clean-GWA-data
```

4| The next step is to create a BED file based on the map file using the following command:

```
gawk '{print "chr"$1, $4, $4+1, $2}' OFS="\t" clean-GWA-data.map > clean-
GWA-data_HG17.BED
```

5| Perform the liftover:

```
./liftOver -bedPlus=4 clean-GWA-data_HG17.BED hg17ToHg19.over.chain clean-
GWA-data.HG19.BED clean-GWA-data_unmapped.txt
```

6| Use the resulting file 'clean-GWA-data_unmapped.txt' to create a list of unmapped SNPs:

```
gawk '/^[^#]/ {print $4}' clean-GWA-data_unmapped.txt > clean-GWA-
data_unmappedSNPs.txt
```

7| Create a mapping file using the new BED file:

```
gawk '{print $4, $2}' OFS="\t" clean-GWA-data.HG19.BED > clean-GWA-
data.HG19.mapping.txt
```

8| Use PLINK to remove the unmapped SNPs from the target data set:

```
plink --noweb --file clean-GWA-data --exclude clean-GWA-
data_unmappedSNPs.txt --update-map clean-GWA-data.HG19.mapping.txt --make-
bed --out clean-GWA-data.HG19.temp
plink --noweb --bfile clean-GWA-data.HG19.temp --recode --out clean-GWA-data.HG19
```

9| Create a new SNP list for the data set:

```
gawk '{print $2}' clean-GWA-data.HG19.map > clean-GWA-data.HG19.snplist
```

The resulting files that are produced after quality control and after lifting over the data set to the correct build are named 'clean-GWA-data.HG19.map' and 'clean-GWA-data.HG19.ped'. In this case, the data set was lifted over from build 35 to build 37; however, other liftovers are also possible. The UCSC Genome Browser website provides multiple chain files.

Imputations with minimac or IMPUTE2

10| SNP imputations can be performed using either a combination of MaCH/minimac (option A) or IMPUTE2 (option B).

(A) MaCH/minimac ● TIMING ~60 h

(i) *Downloading the reference set for minimac*. This pipeline for imputations with MaCH and minimac imputes the target set after quality control and if the target set is on another genome build than the reference set, the target set is lifted over to the same build of the GoNL reference panel release 4. First, create a new directory for the reference set: 'mkdir reference-GoNL-v4'. The zipped VCF files of the GoNL reference panel should be placed in this directory. In this protocol, we assume that the names of the files are as follows: 'gonl.chr{1-22}.release4.gtc.vcf.gz'.

(ii) Use VCFtools to create info files for all chromosomes by running the following command:

```
for chr in {1..22}; do
    vcftools --gzvcf reference-GoNL-
    v4/gonl.chr${chr}.release4.gtc.vcf.gz --get-INFO NS --out
    reference-GoNL-v4/gonl.chr${chr}.release4.gtc;
done
```

(iii) Create a file with all the positions that are in the reference set:

```
rm -f snps-reference.txt
for i in reference-GoNL-v4/gonl.chr*.release4.gtc.INFO; do
    gawk '$1!="CHROM" {print $1"_"$2}' $i >> snps-reference.txt;
done
```

(iv) *Creating the input files for phasing and imputation*. To get a list of positions of SNPs that are in the target set and/or in the reference set, use the following commands:

```
gawk '{print $1"_"$4}' clean-GWA-data.HG19.map > snps-reference-and-rawdata
```

and

```
sort snps-reference.txt | uniq >> snps-reference-and-rawdata
```

To get only those SNPs that are in both the target set and reference set, use the following command:

```
sort snps-reference-and-rawdata | uniq -d | gawk -F "_" '{$3=$2+1;
print $1, $2, $3, "R"NR}' > snps-reference-and-rawdata-duplicates
```

? TROUBLESHOOTING

(v) The names of the SNPs that are in both the target set and in the reference set need to be extracted from the target set. Use PLINK to do this as follows:

```
plink --noweb --file clean-GWA-data.HG19 --extract snps-reference-and-rawdata-
duplicates --range --make-bed --out clean-GWA-data.HG19.for-impute.plink
```

(vi) MaCH and minimac need one file per chromosome. Extract SNPs for each chromosome:

```
for chr in {1..22}; do
    plink --noweb --bfile clean-GWA-data.HG19.for-impute.plink-
    chr ${chr} --recode --out clean-GWA-data.HG19.for-
    impute.plink.chr${chr};
done
```

(vii) Convert the resulting PLINK sets into merlin file format, as minimac requests this:

```
for chr in {1..22}; do
    gawk '{$6=0; print $0}' clean-GWA-data.HG19.for-
    impute.plink.chr${chr}.ped > clean-GWA-data.HG19.for-
    impute.merlin.chr${chr}.ped;
    echo "T faket1" > clean-GWA-data.HG19.for-
    impute.merlin.chr${chr}.dat;
    gawk '$2="M"$2 {print $2}' clean-GWA-data.HG19.for-
    impute.plink.chr${chr}.map >> clean-GWA-data.HG19.for
    impute.merlin.chr${chr}.dat;
    echo "chromosome markername position" > clean-GWA-data.HG19.for-
    impute.merlin.chr${chr}.map;
    gawk '{print $1, $2, $4}' clean-GWA-data.HG19.for-
    impute.plink.chr${chr}.map >> clean-GWA-data.HG19.for-
    impute.merlin.chr${chr}.map;
done
```

(viii) Split the merlin files so that they contain 2,500 markers with a 500-marker overlap using the ChunkChromosome tool:

```
for chr in {1..22}; do
     ./generic-ChunkChromosome/executables/ChunkChromosome -d
     clean-GWA-data.HG19.for-impute.merlin.chr${chr}.dat -n 2500 -o 500;
done
```

(ix) *Using MaCH for phasing*. Use MaCH to phase the haplotypes in each chunk:

```
for chunk in chunk*.dat; do
     machfile="${chunk%.*}";
     merlinfile="${machfile#*-}.ped";
     executables/mach1 -d ${chunk} -p ${merlinfile} --rounds 20 -
     states 200 --phase --interim 5 --sample 5 --compact --prefix
     ${machfile};
done
```

? TROUBLESHOOTING

(x) *Imputation with minimac*. Execute the following commands to impute all chunks using minimac:

```
for chunk in chunk*.dat; do
filename1="${chunk%.*}";
filename2="${filename1#*-}.ped";
chr=`echo "${filename1##*.}" | sed 's/chr//'`;
minimac --vcfReference --rs --refHaps reference-GoNL-
v4/gonl.chr${chr}.release4.gtc.vcf.gz --haps ${filename1}.gz
--snps ${filename1}.dat.snps --rounds 5 --states 200 --autoClip
autoChunk-clean-GWA-data.HG19.for-impute.merlin.chr${chr}.dat
--gzip --phased --probs --prefix ${filename1};
done
```

? TROUBLESHOOTING

(B) IMPUTE2 ● TIMING ~7 h

(i) *Downloading the reference set for IMPUTE2*. This pipeline for imputations with IMPUTE2 imputes the target set after quality control and if the target set is on another genome build than the reference set, the target set is lifted over to the same build of the GoNL reference panel release 4. First, create a new directory for the reference set: 'mkdir reference-GoNL-v4'. All files of the GoNL reference panel should be placed in this directory. In this protocol, we assume that the names of the files are as follows: 'gonl.chr{1–22}.release4.gtc. {hap.gz, legend.gz, geneticmap.txt}'.

(ii) Now create a file with all the SNP names that are in the reference set:

```
rm -r snps-reference.txt;
for chr in {1..22}; do
     gunzip -c reference-GoNL-
     v4/gonl.chr${chr}.release4.gtc.legend.gz | gawk -v chr=${chr}
     '$5=="SNP" && $1!="id" {print chr"_"$2}' >> snps-
     reference.txt;
done
```

(iii) *Creating the input files for phasing and imputation*. Use the following commands to get a list of positions of SNPs that are in the target set and/or in the reference set:

```
gawk '{print $1"_"$4}' clean-GWA-data.HG19.map > snps-reference-and-rawdata
```

and

```
sort snps-reference.txt | uniq >> snps-reference-and-rawdata
```

To get only those SNPs that are in both the target set and reference set, use the following command:

```
sort snps-reference-and-rawdata | uniq -d | gawk -F "_" '{$3=$2+1; print $1, $2,
$3, "R"NR}' > snps-reference-and-rawdata-duplicates
```

? TROUBLESHOOTING

(iv) The names of the SNPs that are in both the target set and in the reference set need to be extracted from the target set. Use PLINK to run the following command. This creates the following files per chromosome: 'clean-GWA-data.HG19.for-impute.plink.chr${chr}.ped' and 'clean-GWA-data.HG19.for-impute.plink.chr${chr}.map'.

```
plink --noweb --file clean-GWA-data.HG19 --extract snps-reference-and-rawdata-
duplicates --range --make-bed --out clean-GWA-data.HG19.for-impute.plink
```

(v) As we will phase per chromosome, split the PLINK file into 22 files:

```
for chr in {1..22}; do
   plink --bfile clean-GWA-data.HG19.for-impute.plink --chr $chr --recode --out
   clean-GWA-data.HG19.for-impute.plink.chr${chr};
done
```

? TROUBLESHOOTING

(vi) *Using SHAPEIT for phasing*. For every chromosome, phase the haplotypes using SHAPEIT:

```
for chr in {1..22}; do
       namefile="clean-GWA-data.HG19.for-impute.plink.chr${chr}";
            ./shapeit.v2.r790.RHELS_5.4.static --input-ped
            ${namefile}.ped ${namefile}.map --input-map reference-
            GoNL-v4/gonl.chr${chr}.release4.gtc.geneticmap.txt --
            output-max ${namefile}.phased --thread 8 --output-log
            ${namefile}.phased;
done
```

(vii) *Imputation with IMPUTE2*. For every chromosome, perform imputations in chunks of 5 Mb:

```
refdir="reference-GoNL-v4";
for chr in {1.22}; do
    namefile="clean-GWA-data.HG19.for-
    impute.plink.chr${chr}.phased";
    maxPos=$(gawk '$1!="position" {print $1}'
    ${refdir}/gonl.chr${chr}.release4.gtc.geneticmap.txt | sort -n
    | tail -n 1);
    nrChunk=$(expr ${maxPos} "/" 5000000);
    nrChunk2=$(expr ${nrChunk} "+" 1);
    start="0";
    for chunk in $(seq 1 $nrChunk2); do
          endchr=$(expr $start "+" 5000000);
          startchr=$(expr $start "+" 1);
          ./impute_v2.3.1_x86_64_static/impute2
          -known_haps_g ${namefile}.haps
          -m ${refdir}/gonl.chr${chr}.release4.gtc.geneticmap.txt
          -h ${refdir}/gonl.chr${chr}.release4.gtc.hap.gz
          -l ${refdir}/gonl.chr${chr}.release4.gtc.legend.gz
          -int ${startchr} ${endchr} -Ne 20000 -o
          ${namefile}.chunk${chunk}.impute2;
    start=${endchr};
    done
done
```

? TROUBLESHOOTING

(viii) Convert the files with the probabilities for the three genotypes into dosage files:

```
   for chr in {1.22}; do
     namefile="clean-GWA-data.HG19.for-
     impute.plink.chr${chr}.phased";
     maxPos=$(gawk '$1!="position" {print $1}'
     ${refdir}/gonl.chr${chr}.release4.gtc.geneticmap.txt | sort -n
     | tail -n 1);
     nrChunk=$(expr ${maxPos} "/" 5000000);
     nrChunk2=$(expr ${nrChunk} "+" 1);
     for chunk in $(seq 1 $nrChunk2); do
           gawk '{tp = $1 " " $2 " " $3 " " $4 " " $5; for (i=6;
     i<=NF; i+=3) tp = tp " " $(i+1) + 2.0*$(i+2); print tp }'
     ${namefile}.chunk${chunk}.impute2 >
     ${namefile}.chunk${chunk}.impute2.dosage;
     done
done
```

? TROUBLESHOOTING

It is likely that many of the tools used in this protocol will be updated as time passes; we therefore recommend checking whether there are new versions of the tools each time the protocol is run and what the changes between versions are.

Imputation with MaCH and minimac, Step 10A(iv), and imputation with IMPUTE2, Step 10B(iii). This step checks the concordance between SNPs within the target set and the reference panel based on the position on the chromosome, by assuming that the SNP names are equal in both. This requires both panels to be aligned to the correct human genome build. Another option is to leave the SNPs that are in the target set and not in the reference panel. In that case, Step 10A(iv,v) (for MaCH and minimac) or Step 10B(iii, iv (for IMPUTE2) can be replaced by 'plink --noweb --file clean-GWA-data.HG19 --make-bed --out clean-GWA-data.HG19.for-impute.plink.' It is also important to have both the target set and the reference panel on the same human genome build, as IMPUTE2 links the two panels according to chromosome and position, not SNP name.

Imputation with MaCH and minimac, Step 10A(ix). The command-line parameters '--interim 5' (to save intermediate results), '--sample 5' (random (but plausible) sets of haplotypes for each individual should be drawn every five iterations) and '--compact' (reduces memory use at the cost of runtime) can be removed from the command line to save time and disk space.

Imputation with MaCH and minimac, Step 10A(x). The command-line parameter '--rs' allows the use of rs GWAS SNP identifiers in the target set. This command-line parameter can be removed if the target set does not include rs identifiers.

Imputation with IMPUTE2, Step 10B(v). To increase the speed of the IMPUTE2 protocol, the target set could be reformatted into binary PLINK format (**Box 1**); therefore, the '--recode' command should be replaced by `'--make-bed'`. The follow-up Step 10B(vi, vii) should be adjusted for binary files in that case.

Imputation with IMPUTE2, Step 10B(vii). When the analyst wants to use two phased reference panels, the IMPUTE2 command should be replaced with

```
./impute_v2.3.1_x86_64_static/impute2 -known_haps_g
${namefile}.haps -m ${refdir}/gonl.chr${chr}.release4.gtc.geneticmap.txt -h
${refdir}/gonl.chr${chr}.release4.gtc.hap.gz
${refdir}/1000g.chr${chr}.release4.gtc.hap.gz -l
```

```
    ${refdir}/gonl.chr${chr}.release4.gtc.legend.gz
    ${refdir}/1000g.chr${chr}.release4.gtc.legend.gz -int ${startchr} ${endchr} -Ne
    20000 -o ${namefile}.chunk${chunk}.impute2;
```

When combining several of the commands into Bash shell script files, be sure to add 'set –e' and 'set –u' as the first two actual commands in the script. This makes sure that the script halts on errors and when undefined variables are being used, respectively. If additional debugging of Bash scripts is required, running a script such as 'bash -x scriptfile.sh' will run the script in debug mode, showing the value of variables and so on. Alternatively, if only a certain part of a Bash script is to be debugged, adding 'set –x' before and 'set +x' after the problematic part will enable debugging only for that part.

● TIMING

Step 1, performing quality control: ~8 h
Steps 2–9, converting the target set to the correct build: ~20 min
Step 10, imputations with minimac or IMPUTE2:
Step 10A, MaCH/minimac: ~60 h
Step 10A(i–iii), downloading the reference set for minimac: ~15 min
Step 10A(iv–viii), creating the input files for imputation: ~5 min
Step 10A(ix), using MaCH for phasing per chunk: ~15 h
Step 10A(x), imputation with minimac: ~45 h
Step 10B, IMPUTE2: ~7 h
Step 10B(i,ii), downloading the reference set for IMPUTE2: ~10 min
Step 10B(iii–v), creating the input files for imputation: ~10 min
Step 10B(vi), using SHAPEIT for phasing per chromosome: varies per chromosome from 1.5 h to 5.5 h
Step 10B(vii,viii), imputation with IMPUTE2 per chunk: ~1 h

Inexperienced analysts will typically require more time. The estimated times and memory requirements are based on the target and reference sets used in this protocol; the estimates may also vary with different cohort designs. Moreover, given the computational nature of this protocol, timing will also heavily depend on the computational resources that are available to the analyst, and to a lesser extent on the versions of the tools. The phasing and imputation steps are the most time-consuming steps.

ANTICIPATED RESULTS

Converting the target set to the correct build

The genome-wide SNP data used in this protocol consists of 1,919 samples and 313,878 markers after performing quality control. After lifting this data set over from hg17 to hg19, the data set consists of 1,919 samples and 304,930 markers.

Imputation with MaCH and minimac

Imputation with minimac results in eight files per chunk. Each file is a compressed (zipped) file. If needed, such a file can be decompressed by running 'gunzip -c filename.gz > filename'. Given the command for minimac specified earlier, the names of the output files start with 'hunk1-clean-GWA-data.HG19.for-impute.merlin.chr1' for chunk 1 of chromosome 1:

- a file with the extension '.dose.gz', which contains the imputed dosage for each genotype. Each row in the output will include one column per marker.
- a file with the extension '.erate.gz', which contains the error rate per marker.
- a file with the extension '.hapDose.gz', which contains the dosage for each haplotype separately.
- a file with the extension '.haps.gz', which contains the most likely alleles for each haplotype separately.
- a file with the extension '.info.draft', which contains the reference allele, nonreference allele and frequency per marker. It also lists the markers that were genotyped.
- a file with the extension '.info.gz', which contains the information about reference allele, frequencies and quality of imputations per marker. It also lists the markers that were genotyped.
- a file with the extension '.prob.gz', which contains the imputed probabilities for each genotype. Each row in the output will include two columns per marker. The first of these columns denotes the probability of a homozygote for allele 1. The second column denotes the probability of a heterozygote.
- a file with the extension '.rec.gz', which contains the switch error rate per interval.

Imputation with IMPUTE2

Imputation with IMPUTE2 results in five files per chunk. Given the command for IMPUTE2 specified earlier, the names of the output files start with 'clean-GWA-data.HG19.for-impute.plink.chr1.phased.chunk1.impute2' for chunk 1 of chromosome 1:

- a file without any extra extension: this file contains the main results of the imputations. The first five entries of each line should be the SNP ID, rs ID of the SNP, base-pair position of the SNP, the allele coded *A* and the allele coded *B*. The subsequent columns contain the probabilities for the three genotypes *AA*, *AB* and *BB* for the each individual in the target set. This format allows for genotype uncertainty, and therefore the probabilities for a given individual need not sum to 1.
- a file with the extension '_info': this file contains the following columns— SNP identifier, rsID, base-pair position, expected frequency of allele coded 1, measure of the observed statistical information associated with the allele frequency estimate, average certainty of best-guess genotypes and the internal 'type' assigned to SNP.
- a file with the extension '_info_by_sample', which contains the concordance and the R^2 per sample.
- a file with the extension '_summary', which contains a summary of the screen output.
- a file with the extension '_warnings', which contains all warnings generated by IMPUTE2.

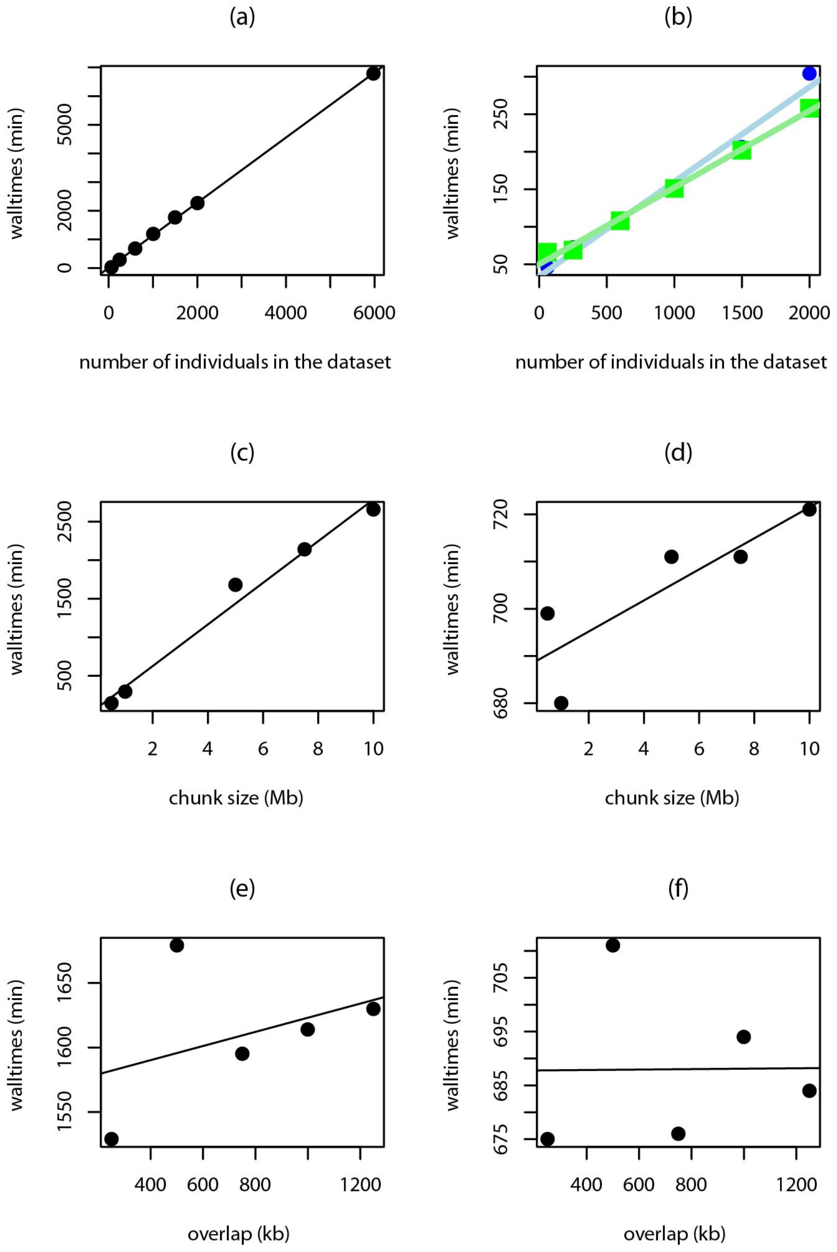
Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS We acknowledge the Genetic Cluster Computer (http://www.geneticcluster.org), which is financially supported by the Netherlands Scientific Organization (NWO 480-05-003) along with a supplement from the Dutch Brain Foundation and the VU University Amsterdam. We thank SURFsara Computing and Networking Services (http://www.surfsara.nl) for their support in using the Lisa Compute Cluster. This work was supported by the BioAssist Biobanking Task Force of the Netherlands Bioinformatics Centre, which is supported by the Netherlands Genomics Initiative. This work is part of the program of BiG Grid, the Dutch e-Science Grid, which is financially supported by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (Netherlands Organisation for Scientific Research, NWO). This work was financed as a Rainbow Project of the Biobanking and Biomolecular Research Infrastructure Netherlands (BBMRI-NL, RP-2), a Research Infrastructure financed by the Dutch government (NWO 184.021.007). The work of L.C.K. was partially funded by the European Union FP7 (2007–2013) program under grant agreement numbers 305280 (MIMOmics) and 602736 (PainOmics).

AUTHOR CONTRIBUTIONS E.M.v.L., A.K., L.C.K. and J.J.H. wrote the first draft of the article. A.K., E.M.v.L., M.V.K. and J.J.H. performed analyses. E.M.v.L., P.D. and M.V.K. designed the protocol. D.I.B. performed study design and genotyping of the Netherlands Twin Registry. A.K., P.D., M.V.K., P.I.W.d.B., C.W., M.A.S., D.I.B., C.M.v.D., L.C.K., P.E.S. and J.J.H. revised the article.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

1. International HapMap 3 Consortium. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
2. 1000 Genomes Project Consortium. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
3. Boomsma, D.I. *et al.* The Genome of the Netherlands: design, and project goals. *Eur. J. Hum. Genet.* **22**, 221–227 (2014).
4. Deelen, P. *et al.* Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'. *Eur. J. Hum. Genet.* **22**, 1321–1326 (2014).
5. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46**, 818–825 (2014).
6. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).
7. Anderson, C.A. *et al.* Data quality control in genetic case-control association studies. *Nat. Protoc.* **5**, 1564–1573 (2010).
8. Browning, B.L. & Browning, S.R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**, 210–223 (2009).
9. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
10. Verma, S.S. *et al.* Imputation and quality control steps for combining multiple genome-wide datasets. *Front. Genet.* **5**, 370 (2014).
11. Winkler, T.W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat. Protoc.* **9**, 1192–1212 (2014).
12. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & calo R Abecasis, G. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).
13. van Leeuwen, E.M. *et al.* Genome of the Netherlands population-specific imputations identify an *ABCA6* variant associated with cholesterol levels. *Nat. Commun.* **6**, 6065 (2015).
14. Jostins, L., Morley, K.I. & Barrett, J.C. Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. *Eur. J. Hum. Genet.* **19**, 662–666 (2011).
15. Pistis, G. *et al.* Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. *Eur. J. Hum. Genet.* **23**, 975–983 (2014).
16. Browning, S.R. & Browning, B.L. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* **12**, 703–714 (2011).
17. Nho, K. *et al.* The effect of reference panels and software tools on genotype imputation. *AMIA Annu. Symp. Proc.* **2011**, 1013–1018 (2011).
18. Li, Y., Willer, C.J., Ding, J., Scheet, P. & calo R Abecasis, G. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816–834 (2010).
19. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).
20. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
21. Roshyara, N.R. & Scholz, M. fcGENE: a versatile tool for processing and transforming SNP datasets. *PLoS ONE* **9**, e97589 (2014).
22. Nelson, S.C., Doheny, K.F., Laurie, C.C. & Mirel, D.B. Is 'forward' the same as 'plus'?...and other adventures in SNP allele nomenclature. *Trends Genet.* **28**, 361–363 (2012).
23. Deelen, P. *et al.* Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. *BMC Res. Notes* **7**, 901 (2014).
24. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
25. Sulovari, A. & Li, D. Gact: a genome build and allele definition conversion tool for SNP imputation and meta-analysis in genetic association studies. *BMC Genomics* **15**, 610 (2014).
26. de Bakker, P.I.W. *et al.* Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* **17**, R122–R128 (2008).
27. Wang, Z. *et al.* Improved imputation of common and uncommon SNPs with a new reference set. *Nat. Genet.* **44**, 6–7 (2012).
28. Abecasis, G.R., Cherny, S.S., Cookson, W.O. & Cardon, L.R. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* **30**, 97–101 (2002).



Supplementary Figure 1

The walltimes when splitting up the data set.

The walltimes per job for MaCH (a, c, e) and minimac (b, d, f) for various ways of splitting up the data set. The walltime is the time as measured by a clock on the wall (CPU time, disk writing etcetera) required to impute the target set. The walltime per job for running MaCH fits the linear regression models *t=8.6 + 1.13n* (Figure a), *t=86.49 + 270.02n* (Figure c) and *t=1568.3 + 2.7n* (Figure e). The walltime per job for running minimac fits the linear regression model *t=33.8 + 0.13n* (split before MaCH (blue circles)), *t=50.2 + 0.10n* (split after MaCH (green squares)) (Figure b), *t=688.6 + 3.29n* (Figure d) and *t=687.7 + 0.02n* (Figure f). *t* is the walltime in minutes and *n* the number of samples (a, b), the size of the chunks in Mb (c, d) and the percentage of overlap (e, f). The percentage overlap is 10% in Figure c and d and the chunk size is 5Mb in Figure e and f.

Supplementary Note: the Genome of the Netherlands consortium members:

Analysis group: Morris A. Swertz[6,7] (Co-Chair), Laurent C. Francioli[1], Freerk van Dijk[6,7], Androniki Menelaou[1], Pieter B.T. Neerincx[6,7], Sara L. Pulit[1], Patrick Deelen[6,7], Clara C. Elbers[1], Pier Francesco Palamara[2], Itsik Pe'er[2,8], Abdel Abdellaoui[9], Wigard P. Kloosterman[1], Mannis van Oven[10], Martijn Vermaat[11], Mingkun Li[12], Jeroen F.J. Laros[11], Mark Stoneking[12], Peter de Knijff[13], Manfred Kayser[10], Jan H. Veldink[14], Leonard H. van den Berg[14], Heorhiy Byelas[6,7], Johan T. den Dunnen[11], Martijn Dijkstra[6,7], Najaf Amin[15], K. Joeri van der Velde[6,7], Jouke Jan Hottenga[9], Jessica van Setten[1], Elisabeth M. van Leeuwen[15], Alexandros Kanterakis[6,7], Mathijs Kattenberg[9], Lennart C. Karssen[15], Barbera D.C. van Schaik[16], Jan Bot17, Isaäc J. Nijman[1], David van Enckevort[18], Hailiang Mei[18], Vyacheslav Koval[19], Kai Ye[20,21], Eric-Wubbo Lameijer[21], Matthijs H. Moed[21], Jayne Y. Hehir-Kwa[22], Robert E. Handsaker[5,23], Shamil R. Sunyaev[4,5], Mashaal Sohail[4,5], Fereydoun Hormozdiari[24], Tobias Marschall[25], Alexander Schönhuth[25], Victor Guryev[26], Paul I.W. de Bakker[1,3-5] (Co-Chair);

Cohort collection and sample management group: P. Eline Slagboom[21], Marian B. Beekman[21], Anton J.M. de Craen[21], H. Eka D. Suchiman[21], Albert Hofman[15], Cornelia van Duijn[15], Dorret I. Boomsma[9], Gonneke Willemsen[9], Bruce H. Wolffenbuttel[27], Mathieu Platteel[6], Steven J. Pitts[28], Shobha Potluri[28], David R. Cox[28,34];

Whole-genome sequencing: Qibin Li[29], Yingrui Li[29], Yuanping Du[29], Ruoyan Chen[29], Hongzhi Cao[29], Ning Li[30], Sujie Cao[30], Jun Wang[29,31,32];

Ethical, Legal, and Social Issues: Jasper A. Bovenberg[33];

Steering committee: Cisca Wijmenga[6,7] (Principal Investigator), Morris A. Swertz[6,7], Cornelia M. van Duijn[15], Dorret I. Boomsma[9], P. Eline Slagboom[21], Gertjan B. van Ommen[11], Paul I.W. de Bakker[1,3-5]

[1] Department of Medical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands. [2] Department of Computer Science, Columbia University, New York, NY, USA. [3] Department of Epidemiology, University Medical Center Utrecht, Utrecht, The Netherlands. [4] Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [5] Broad Institute of Harvard and MIT, Cambridge, MA, USA. [6] Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands. [7] Genomics Coordination Center, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands. [8] Department of Systems Biology, Columbia University, New York, NY, USA. [9] Department of Biological Psychology, VU University Amsterdam, Amsterdam, The Netherlands. [10] Department of Forensic Molecular Biology, Erasmus Medical Center, Rotterdam, The Netherlands. [11] Leiden Genome Technology Center, Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands. [12] Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany. [13] Forensic Laboratory for DNA Research, Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands. [14] Department of Neurology, University Medical Center Utrecht, Utrecht, The Netherlands. [15] Department of Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands. [16] Bioinformatics Laboratory, Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Amsterdam Medical Center, Amsterdam, The Netherlands. [17] SURFsara, Science Park, Amsterdam, The Netherlands. [18] Netherlands Bioinformatics Centre, Nijmegen, The Netherlands. [19] Department of Internal Medicine, Erasmus Medical Center, Rotterdam, The Netherlands. [20] The Genome Institute, Washington University, St. Louis, MO, USA. [21] Section of Molecular Epidemiology, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands. [22] Department of Human Genetics, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands. [23] Department of Genetics, Harvard Medical School, Boston, MA, USA. [24] Department of Genome Sciences, University of Washington, Seattle, WA, USA. [25] Centrum voor Wiskunde en Informatica, Life Sciences Group, Amsterdam, The Netherlands. [26] European Research Institute for the Biology of Ageing, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands. [27] Department of Endocrinology, University Medical Center Groningen, Groningen, The Netherlands. [28] Rinat-Pfizer Inc, South San Francisco, CA, USA. [29] BGI-Shenzhen, Shenzhen, China. [30] BGI-Europe, Copenhagen, Denmark. [31] Department of Biology, University of Copenhagen, Copenhagen, Denmark. [32] The Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen, Denmark. [33] Legal Pathways Institute for Health and Bio Law, Aerdenhout, The Netherlands. [34] Deceased.