We chose it because we deal with huge amounts of data.
Besides, it sounds really cool.
— Larry Page, co-founder of Google Inc.

2

Meta-imputation of reference data to increase accuracy and power in association analysis

Contents

2.1 Introduction

Genome-wide association (GWA) studies have identified thousands of genetic risk factors that influence disease susceptibility and complex disease phenotypes. A contributing factor to this success is the ability to statistically estimate, or *impute* genotypes that have not been observed in a study sample. Genotype imputation has become a standard technique in GWA studies where it is used to increase the number of variants to achieve higher power in association analysis as well as to facilitate meta-analysis of association results across different studies (Marchini *et al.*, 2007; Marchini and Howie, 2010). Methods for genotype imputation match patterns of genetic variation observed in a study sample with a more densely typed set of haplotypes in a reference panel. The extent of shared variation is informative for estimating the most likely genotypic states at other, unobserved variant sites in the same individuals. Commonly employed imputation methods are, for example, `Beagle` (Browning and Browning, 2016), `MACH` (Li *et al.*, 2010), and `IMPUTE2` (Howie *et al.*, 2009, 2011a).

Genotypes can be imputed with remarkably high accuracy, allowing researchers to assay only a modest number of markers in sampled individuals, which makes large-scale data collection feasible and cost-effective (Li *et al.*, 2009). The accuracy of imputation is dependent on several factors. These include the number of genotyped markers in the study sample, ~~the number of individuals sampled,~~ the size of the reference panel, and the genetic similarity between sampled and reference individuals (Howie *et al.*, 2009; Roshyara and Scholz, 2015). The coverage of the reference panel further influences the power to find significant associations. The availability and choice of reference data therefore becomes crucial in considerations of statistical power in study design.

One of the first larger sets of publicly available reference genomes was established by the International HapMap Project (HapMap), which identified 3.1 million variants through genotyping of 270 individuals from four continental populations (International HapMap Consortium *et al.*, 2007; International HapMap 3 Consortium *et al.*, 2010). More recently, the 1000 Genomes Project (1000G) released reference data in three phases at progressively increasing sample size, currently reaching over 88 million variants from low-coverage whole-genome sequencing (WGS) of 2,504 individuals from 26 populations (1000 Genomes Project Consortium *et al.*, 2012, 2015). Due to ongoing advances in next-generation sequencing (NGS) technologies and reductions in costs, large-scale WGS studies have become routine. However, genetic variation generally shows extensive stratification dependent on geography and ethnicity. Also, disease risk factors can be segregated on a much finer scale. Therefore, any study may only capture the variation present in the population or study cohort sampled, particularly among lower frequency and rare variants.

To increase the chance of detecting significant risk variants through GWA methods, it would be desirable to combine sequencing data from different studies to generate a single, large reference panel for imputation. However, the integration of independently produced datasets is not straightforward due to differences arising from different sequencing platforms, coverage, and strategies to filter and call variant genotypes. It is not directly feasible, for example, to compile an unbiased union of variant calls across studies, because monomorphic sites cannot be distinguished from sites that were filtered or missed. Conversely, retaining the intersection of variants that are present in all panels would dispose of much information.

One solution would be to re-process raw sequence or genotype data from multiple studies together, where variants are jointly called and phased over a combined set of samples. For example, in a large-scale collaborative effort, the Haplotype Reference Consortium (HRC) has recently created a reference panel from study data of 20 participating cohorts, which included a total of 64,976 human haplotypes in its first release (McCarthy *et al.*, 2016). This dataset currently represents the largest single resource of human genetic variation, but currently only includes samples of European ancestry. Although data are not accessible publicly, an online service has been provided for imputation and phasing from the internally stored reference dataset*.

Here, I propose an alternate solution in which multiple reference panels are separately imputed into a given study sample after which the genotype datasets produced are merged. Because imputed data may only differ in variant coverage, while the sample set is identical, it is feasible to merge data and integrate genotype information at overlapping sites. The underlying intuition is that the accuracy of an imputed genotype is indicated by its posterior probability or other metrics that result from the imputation process; for example *allelic* R^2 in `Beagle`, $\hat{r}^2$ in `Mach`, and *info-score* in `IMPUTE2`. The presented method applies such information to select from or assign higher weights to candidate genotypes, thereby indirectly leveraging information across different reference panels.

The following section (2.2) describes the approach by which sets of imputed genotype data are combined to form an integrated, larger genotype dataset; the method is referred

* Haplotype Reference Consortium: http://www.haplotype-reference-consortium.org [Date accessed: 2017-02-05]

to as *meta-imputation*. I considered several strategies to combine data based on different summary metrics. To be able to efficiently evaluate this method, as well as for application to genomic datasets on a larger scale, I implemented the method as a computational tool written in `C++` called `meta-impute`.[*] For assessment of meta-imputation, I constructed multiple, smaller reference panels from a larger dataset, which enabled comparisons between meta-imputation and direct imputations from both single and whole reference data. An additional analysis was conducted using data from several independent studies. The composition of reference data is described in Section 2.3 (page 56). The performance of meta-imputation was evaluated in regards to genotype accuracy and power to detect significant association signals. An accuracy analysis was conducted in Section 2.4 (page 57). Statistical power was analysed in a series of association experiments using simulated case-control data, which is described in Section 2.5 (page 72). Results are jointly discussed in Section 2.6 (page 81).

2.2 Approach

There are several ways by which genotypes imputed from independent sources can be combined at overlapping sites. To provide the means to explore a range of possibilities, the presented solution is implemented as a two-step process. First, a *score metric* is obtained for each genotype which, second, informs a *merge operation*. The general approach of meta-imputation is based on the assumption that a given metric is informative for distinguishing candidate genotypes that are more or less likely to reflect the underlying, true genotypic state. Here, several score metrics (Section 2.2.2, page 53) and two merge operations (Section 2.2.3, page 55) were considered, which are described after introducing principal notation and the general algorithm below.

2.2.1 Description of the method

It is convenient to think of genotype data as being arranged in a matrix, G, of size $M \times N$ where M is the number of observed variant sites and N is the diploid sample size. Let g_{ij} denote the genotype observed at marker i in individual j, such that g_i refers to the vector

[*] Meta-imputation software (`meta-impute`): `https://github.com/pkalbers/meta-impute`

of genotypes of size N at the ith site, and g_j the vector of genotypes of size M belonging to individual j. Meta-imputation combines the information contained across several such genotype matrices. Let L denote the number of available genotype datasets imputed from different reference panels, such that $G_1, \ldots, G_L$ are available, and $k \in \{1, \ldots, L\}$ is used to identify a particular matrix; note that $L \geq 2$ is assumed. Because reference data were imputed into the same study sample, the number of individuals, N, is constant in each matrix but M_k may vary due to differences in coverage per reference panel.

Meta-imputation combines available genotype matrices in an aggregated matrix, A, of size $M_A \times N \times L$ where M_A is the number of variants in the combined set of sites across imputed panels. The algorithm merges genotype information at overlapping variants by gathering those that correspond to the same genomic position per chromosome. Here, the word *analogue* is used to refer to the set of available data vectors that correspond to the same variant. Let a_i denote an analogue variant, *i.e.* the set of overlapping genotype vectors at the ith site in the aggregated matrix, and a_{ij} an analogue genotype, *i.e.* the set of overlapping genotypes at this site in individual j. Note that the number of genotypes referred to by a_{ij} may vary dependent on presence in the reference panel. Let l denote the number of overlapping variants in an analogue, where $1 \leq l \leq L$, such that l_i refers to the size of a_i.

Genotype formats may differ according to the type of data available. Note that the following considers single-nucleotide polymorphisms (SNP) specifically. In generic terms, a genotype can be observed in one of three possible states; homozygous for the reference allele, heterozygous, or homozygous for the alternate allele, which can be encoded by the alternate allele count (*allele dosage*); that is 0, 1, or 2, respectively. Imputed genotypes are typically expressed by the uncertainty associated with the imputation process. Here, an imputed genotype is considered as a tuple (p_0, p_1, p_2) of sum 1, representing the inferred posterior probability per genotypic state. Hence, a_{ijk} refers to a genotype tuple at the ith site in individual j taken from G_k.

The meta-imputation algorithm assigns a score value, s_{ijk}, to each a_{ijk}; *i.e.* each candidate genotype per analogue variant. A *meta-imputed* genotype is formed, denoted by $\hat{g}_{ij}$, by merging candidate genotype data conditional on the score assigned. At sites where

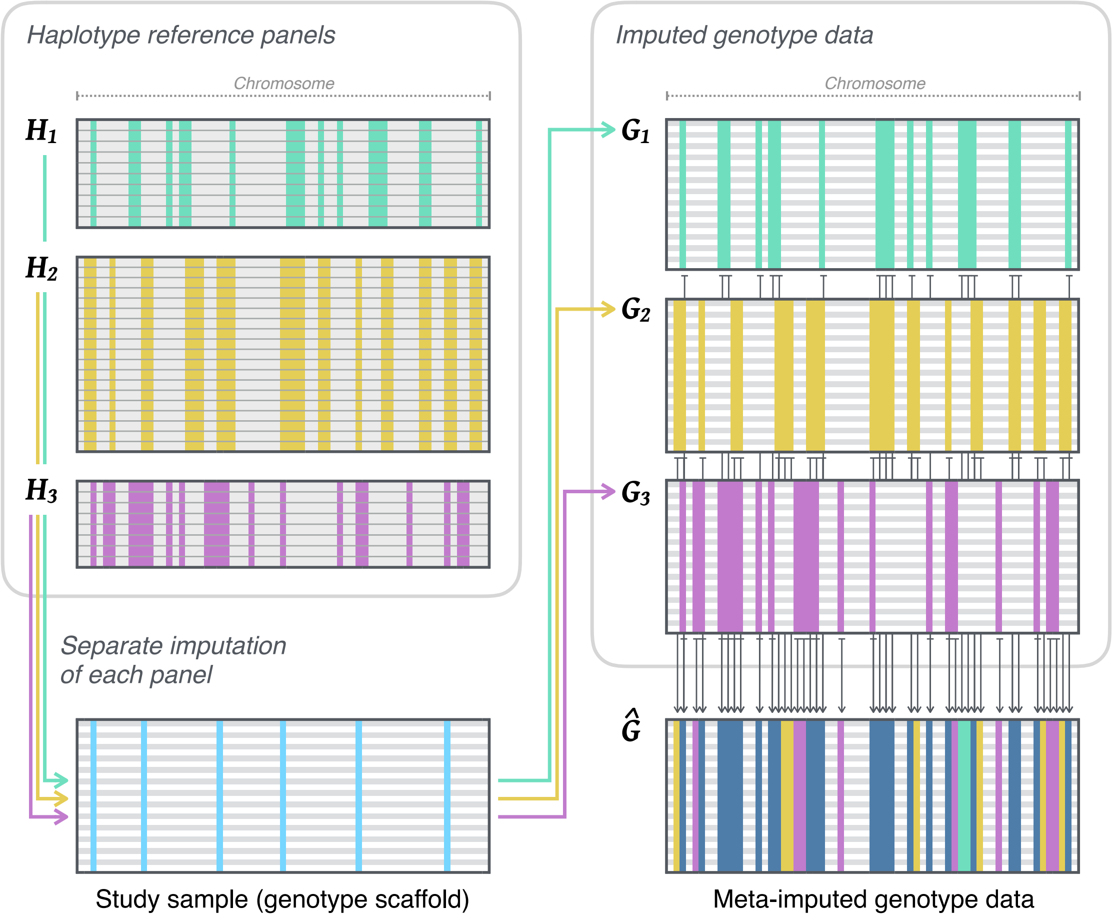
Haplotype reference panels
Chromosome
H_1
H_2
H_3
Separate imputation of each panel
Study sample (genotype scaffold)
Imputed genotype data
Chromosome
G_1
G_2
G_3
$\hat{G}$
Meta-imputed genotype data

Figure 2.1: Illustration of the meta-imputation concept. An example of three haplotype reference panels is shown; denoted by H_1, H_2, and H_3, where haplotypes are indicated by row (*grey*) and observed variant sites are indicated by column. Each panel may vary in sample size and coverage. Reference data are separately imputed into the same study sample, which is a "scaffold" of typed genotype markers, where individual genotypes are indicated by row (alternating *grey-white*) and observed markers by column (*light-blue*). Each imputation returns an imputed genotype dataset, denoted by G_1, G_2, and G_3, containing marker genotypes as present in the corresponding reference panel, but where the number of individuals, N, is the same as in the study sample in each imputed dataset. Imputed data are combined through meta-imputation, such that the resulting genotype dataset, $\hat{G}$, contains the union of variant sites across panels. Variants merged across multiple datasets are indicated (*dark-blue*); the markers specific to a given panel are indicated by their corresponding colour.

$l_i = 1$, that is a given variant was imputed from only one reference panel, genotype data are retained as is, to capture as much variation as available from each separate imputation. The resulting genotype matrix, $\hat{G}$, contains the union of variants across input datasets. A simplified illustration of the meta-imputation concept is given in Figure 2.1 (this page).

2.2.2 Score metrics

The score metrics considered in this work are described below; asserted 2-letter codes are used for the remainder of this chapter.

Maximum probability (`MP`). The mode of the probability distribution of a candidate genotype is taken as the value of the genotype's score; that is the maximum value in the tuple of posterior probabilities, which takes values in $[0, 1]$. The score is separately obtained for each candidate genotype, a_{ijk}, such that

$$s_{ijk} \;=\; \max\Big[(p_0, p_1, p_2)_{ijk}\Big]\;. \tag{2.1}$$

`IMPUTE2` information score (`IS`). The information score (or *info-score*) is used, which is a quality metric of the difference between observed and expected information, dependent on the imputed genotype distribution and estimated allele frequency; see definition below (Marchini and Howie, 2010, S3, eq. 16; modified here to correspond to present notation).

$$\mathrm{I}_{ik} \;=\; \begin{cases} 1 - \frac{\sum_{j=1}^{N} f_{ijk}\; e_{ijk}^2}{2N\hat{\theta}_{ik}(1-\hat{\theta}_{ik})} & \text{if } \hat{\theta}_{ik} \in (0,1) \\ 1 & \text{if } \hat{\theta}_{ik} = 0, \hat{\theta}_{ik} = 1 \end{cases} \tag{2.2}$$

where $e_{ijk} = p_{1\,ijk} + 2p_{2\,ijk}$ is the expected allele dosage, similarly $f_{ijk} = p_{1\,ijk} + 4p_{2\,ijk}$, and $\hat{\theta}_{ik}$ is an estimate of the unknown population allele frequency, calculated as

$$\hat{\theta}_{ik} \;=\; \frac{\sum_{j=1}^{N} e_{ijk}}{2N}\;. \tag{2.3}$$

The `IMPUTE2` info-score takes values in $[0, 1]$ where values close to 0 or 1 indicate low or high certainty, respectively. This and other information measures (*e.g.* `Beagle` R^2 or `Mach` $\hat{r}^2$) are commonly used as a filter criterion in quality control (QC) of imputed GWA data. Because meta-imputation was evaluated using `IMPUTE2` for imputations (see Section 2.4.1, page 59), it is justifiable to use this information measure as a score metric. Since the info-score is calculated per imputed variant, the same score value is assigned to each candidate genotype imputed from a given reference panel at each site. Its value is assigned to each candidate genotype at a given imputed variant; that is

$$s_{ijk} = \mathrm{I}_{ik} \; \forall j \, . \tag{2.4}$$

Sample certainty (SC). A simple measure of imputation certainty is calculated per individual, such that a score value is assigned to genotypes across variants. This metric is calculated as the proportion of an individual's genotypes which have a maximum probability that satisfies a threshold rule, defined as

$$s_{ijk} = \frac{\sum_{i=1}^{M} \mathrm{I}_{ijk}}{M} \tag{2.5}$$

where

$$\mathrm{I}_{ijk} = \begin{cases} 1 & \text{if } \max\left[(p_0, p_1, p_2)_{ijk}\right] \geq r \\ 0 & \text{otherwise} \end{cases} \tag{2.6}$$

where r is an arbitrarily defined value. In the present implementation, this threshold was set to $r = 0.9$. The intention of the SC metric is to prioritise imputations from reference haplotypes which show a closer fit to the genetic variation observed per individual in the study sample, which is assumed to be captured by the posterior probability at imputed genotypes. It must be noted that more sophisticated approaches for the estimation of genetic similarity exist, which provide summary statistics that could be used in place of the present score metric. Possible examples range from multi-locus statistics to fine-scale measures of population structure and demographic history (*e.g.* McVean *et al.*, 2004; Lawson *et al.*, 2012).

Random score (RS). In addition, the option to assign random score values to candidate genotypes was included, to be considered as a control against which the above metrics were compared. The score was explicitly calculated as

$$s_{ijk} = \frac{\mathrm{rand}(R)}{100} \, , \qquad R \in \{1, 2, \ldots, 99\} \tag{2.7}$$

where rand $(\cdot)$ is a function which uniformly selects one value from R at random, such that $0 < s_{ijk} < 1$.

2.2.3 Merge operations

Any operation to merge the information available per analogue genotype can be divided into one of two conceptually distinct approaches; either one candidate genotype is selected and others are discarded, or a new genotype tuple is mathematically derived from available data. Accordingly, I considered the following two operations; note that the specified 3-letter codes are used henceforth.

Maximum score selection (MSS). A candidate genotype is selected by using score metrics as a ranking criterion, where the genotype tuple with the highest assigned score is selected from an analogue genotype in a_{ij} and retained as is in $\hat{g}_{ij}$; see below.

$$\hat{k} \;=\; \underset{k\in\{1,\dots,l_i\}}{\arg\max}\left[s_{ij}\right] \qquad \textbf{s.t.} \qquad \hat{g}_{ij} \;=\; a_{ij\hat{k}} \tag{2.8}$$

If the highest score value is equal in more than one candidate genotypes, one is selected at random from those with the highest score.

Weighted linear combination (WLS). Tuple values of the meta-imputed genotype are derived from candidate genotypes as a linear combination of their posterior probability per genotypic state. This is calculated as the weighted average over analogue genotype probabilities, using corresponding score values as weights. Each candidate genotype thereby contributes to the resulting probability distribution in $\hat{g}_{ij}$~~, except for genotypes with $s_{ijk} = 0$~~. Probability values in each tuple a_{ijk} are multiplied by their assigned s_{ijk} after normalising scores such that values in s_{ij} sum to 1. The tuple of the meta-imputed genotype is then constructed by calculating the sum over the weighted probabilities at each genotypic state; see below (the mathematical definition follows Stone (1961)).

$$\hat{g}_{ij} \;=\; (\hat{p}_0, \hat{p}_1, \hat{p}_2)_{ij} \;=\; \sum_{k=1}^{l_i} (p_0, p_1, p_2)_{ijk}\; s_{ijk} \tag{2.9}$$

Implicitly, the resulting probability distribution in $\hat{g}_{ij}$ sums to 1. In contrast to MSS above, the weighted linear combination of genotype data does not discard available information. But note that tuple values may not be regarded as posterior probabilities when candidate genotypes were combined using WLS, but rather as “pseudo-probabilities”.

2.3 Generation of reference datasets

Multiple reference panels were derived from the 1000 Genomes Project (1000G) Phase I dataset, which comprises both low-coverage whole-genome sequencing and whole-exome sequencing data of 1,092 individuals from 14 populations of European, East-Asian, African, and admixed American ancestries.* This original dataset was split into non-overlapping subsets in two scenarios, A and B, reflecting situations when reference data of similar or distinct ethnic backgrounds would be available for imputation into a given study sample; see details below.

Scenario A included four panels composed of individuals belonging to European sub-populations (CEU, FIN, GBR, and TSI) as an example use case when different reference data of similar ethnic background are available.

Scenario B included four panels from different continental populations (AFR, AMR, ASN, and EUR) as an example use case when panels of distinct-ancestry samples are available.

Because sample sizes of the population groups considered in Scenario B differed in 1000G (more than in Scenario A), extracted individuals were randomly drawn from each group to create panels of equal size. Note that this was done to be able to better compare imputation accuracy among the generated panels, but which is not a requirement when using the meta-imputation method. Monomorphic sites and singletons were removed in each generated panel to more closely resemble data from independently conducted studies, where singleton or monomorphic variant calls are likely to be removed in the final dataset. In the following, the term *split panel* is used to denote subset reference data from 1000G. Throughout, analyses were limited to data from one chromosome, namely chromosome 20. This was done to allow for a larger number of replicate analyses, as will be described in Section 2.5 (page 72).

Generated split panels were used for separate imputations and subsequent integration of estimated genotype data through meta-imputation. To compare meta-imputed

* Note that I completed work on the *meta-imputation* project prior to the release of 1000G Phase III (1000 Genomes Project Consortium *et al.*, 2015).

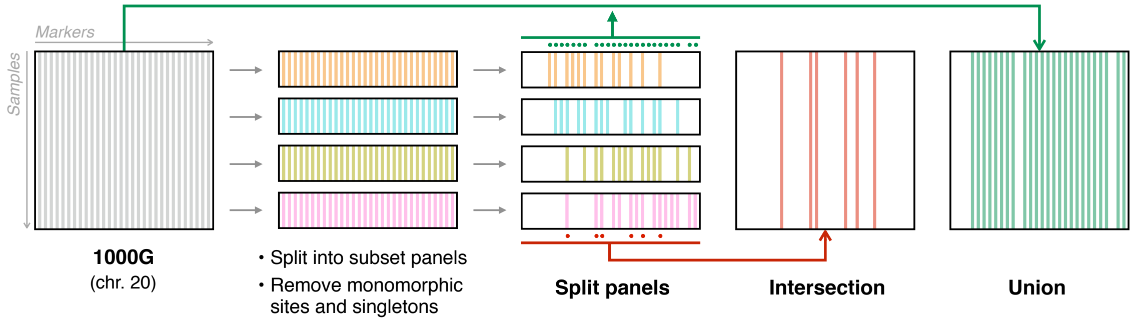
Markers
Samples
1000G
(chr. 20)
• Split into subset panels
• Remove monomorphic sites and singletons
Split panels
Intersection
Union

Figure 2.2: Generation of reference panels in each scenario. The original 1000 Genomes dataset (Phase I, chromosome 20) was used to generate multiple, smaller panels for imputation. This was done in two scenarios to create data of similar or distinct ethnic backgrounds. In each scenario, data were split into four *split* panels of approximately equal size. Monomorphic sites and singletons were removed in each split panel. Two additional panels were generated from the obtained split panels per scenario; one *intersection* panel and one *union* panel, both of which contained the union of individuals across split panels, but where the intersection panel only included sites if captured in all split panels, and the union panel included all sites as observed in the original dataset (except monomorphic or singleton sites as per the individuals included).

genotypes to those that were directly imputed from a unified panel, two additional reference datasets were generated from 1000G per scenario, which combined samples across respective split panels; referred to as the *intersection* panel and the *union* panel. The union reference contained variation as present in the original dataset, but for the individuals contained across split panels in a given scenario, and with monomorphic and singleton variants removed. The intersection reference contained the same set of individuals as the union panel, but where variant sites not shared across all split panels were removed. Unlike the split panels, from which imputed data were combined in meta-imputation, the genotype datasets obtained in imputations from the intersection and union panels were used in direct comparisons to meta-imputed data. The process of reference data generation is illustrated in Figure 2.2 (this page). A summary of the final reference datasets in each scenario is given in Table 2.1 (next page).

2.4 Accuracy of estimated genotypes

Evaluation of genotype accuracy was done in two parts. First, each combination of score metric and merge operation was tested and compared to select the best performing setting for downstream analyses. Second, meta-imputed genotypes generated under the selected

Table 2.1: Dimensions of generated reference data used for imputations. Panels included in Scenarios A and B were generated from the 1000G Phase I dataset. These "split" panels are named after their respective population codes in 1000G. Only data from chromosome 20 were considered. Note that split panels in Scenario B were reduced to match the size of the smallest panel in that scenario. Both the *intersection* and the *union* panels were created from the combined set of individuals across panels in each scenario.

Scenario A			Scenario B		
Panel	Samples	Variants	Panel	Samples	Variants
CEU	85	197,252	*AFR*	181	429,088
FIN	93	205,093	*AMR*	181	307,454
GBR	89	202,707	*ASN*	181	209,209
TSI	98	207,583	*EUR*	181	233,527
Intersection	365	168,744	Intersection	724	144,259
Union	365	253,852	Union	724	559,172

setting were examined in comparison to genotype data imputed from each split reference panel, as well as the intersection and union imputations. Details about the methods used are given in the section below. Results are presented in Section 2.4.2 (page 61).

2.4.1 Methods

Calculation of genotype accuracy requires that the true genotypic states at untyped variants in a study sample are known. This was done by using a larger dataset from which a subset of variants was drawn to form an imputation scaffold. Missing variants were then re-imputed from available reference panels. The generation of the genotype scaffold is described below, followed by details about imputation, quality control, and the calculation of genotype accuracy.

Generation of genotype scaffold data (study sample)

The study sample used for imputations was extracted as a scaffold from data of the Genetics of Type 2 Diabetes Project (GoT2D), consisting of 2,657 individuals of Central and Northern European descent (Fuchsberger *et al.*, 2016).* The dataset is composed of data obtained on several platforms; low-coverage whole-genome sequencing (~ 5×), high-coverage whole-exome sequencing (~ 82×), and genotyping data using the *Illumina Omni2.5 Array*. To maintain a congruent set of markers in the genotype scaffold, variants

* GoT2D Consortium: http://www.type2diabetesgenetics.org/projects/got2d [Date accessed: 2016-12-02]

typed on the latter were extracted from the larger GoT2D dataset, yielding 40,255 variants of in total 387,499 SNPs on chromosome 20 in GoT2D, after removing monomorphic sites and singletons. Remaining sites were masked for comparison after imputation, where imputed variants were matched to their corresponding sites in the masked dataset to calculate genotype accuracy.

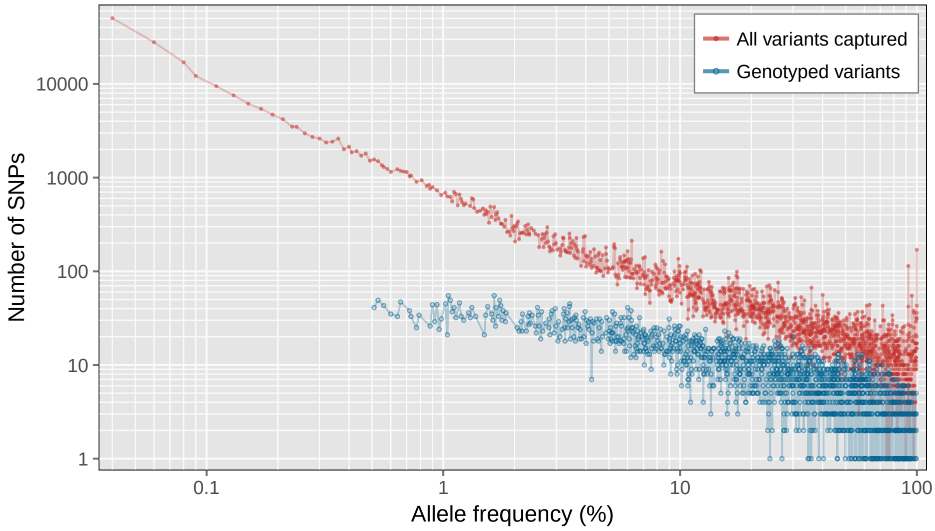

Figure 2.3: Site frequency spectrum of variants captured in GoT2D (chromosome 20). The site frequency spectrum (SFS) is shown for SNPs as captured in the full GoT2D dataset (*red*) and SNPs genotyped using *Illumina Omni2.5 Array* (*blue*), given the allele frequencies observed in the full GoT2D sample for chromosome 20, after removing singletons and monomorphic sites. `ADDED`

Figure 2.3 (this page) shows the site frequency spectrum (SFS) of variants captured in the GoT2D dataset, highlighting the discrepancy between the frequency distribution observed at all captured variants and those obtained through genotyping only. Rare variants (*e.g.*at allele frequency ≤ 1%) are underrepresented in the genotyped set of SNPs and, thus, in the extracted genotype scaffold. Imputation from a reference panel into the scaffold attempts to fill these gaps, including sites with alleles occurring at lower frequencies.

Imputation and quality control

Imputations were performed using `IMPUTE2` version 2.3.0 (Howie *et al.*, 2009), and executed in consecutive chunks of 5 Megabases (Mb). The GoT2D dataset comprises already phased haplotypes, so imputations were carried out on pre-phased genotypes (command line argument `-use_prephased_g` in `IMPUTE2`). Because meta-imputation is indirectly based on information from more reference haplotypes than available in each separate imputation, the number of haplotypes that inform the imputation process was set to the maximum number present in a given reference panel (command line argument `-k_hap` in `IMPUTE2`). This was done to minimise potential biases in comparisons between meta-imputed and imputed genotypes, but is not a requirement for general applications of this approach.

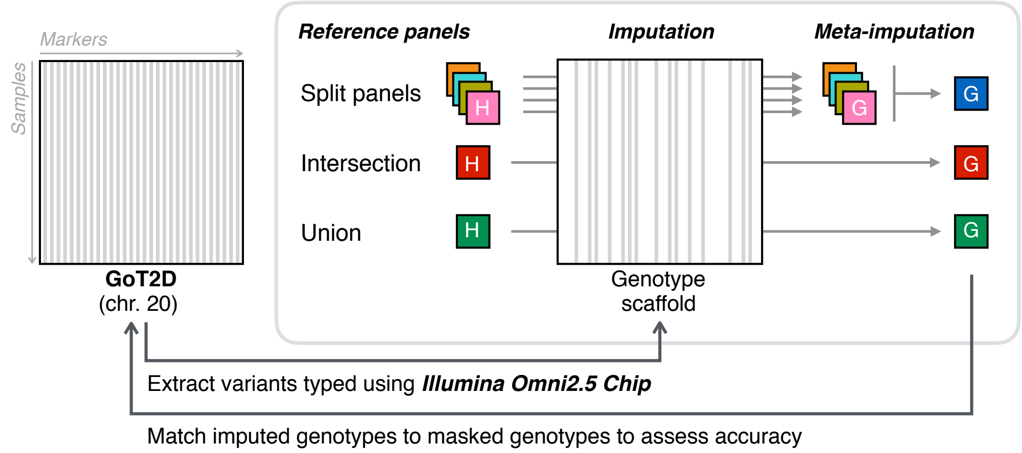

Figure 2.4: Illustration of the accuracy assessment process. Imputations were performed on the same genotype scaffold, which consisted of genetic markers obtained through genotyping using *Illumina Omni2.5 Chip*, which was part of the GoT2D dataset. This scaffold was extracted from GoT2D data, where remaining markers were masked for subsequent calculation of accuracy (squared Pearson correlation coefficient, r^2) at corresponding sites after imputation. Several reference panels were available, which were imputed into the same scaffold. Meta-imputation was applied to the imputed datasets obtained from split panels, which were generated as distinct subsets from the 1000G dataset. The intersection and union panels were separately imputed into the scaffold and subsequently compared to meta-imputed data on corresponding variant sets.

Imputed and meta-imputed genotype data were filtered in QC, removing variants at `IMPUTE2` info-score < 0.4 and at deviations from Hardy-Weinberg equilibrium (HWE) at p-value $< 1 \times 10^{-4}$. These metrics were computed using `QCTOOL`,* which was performed on both directly imputed and meta-imputed datasets. Imputed data were filtered before the assessment of imputation accuracy, but not before integration through meta-imputation. The proportion of variants retained after QC was used as an indicator for data quality in comparisons between imputed and meta-imputed data, as well as to characterise the

* `QCTOOL`: `http://www.well.ox.ac.uk/~gav/qctool` [`Date accessed: 2016-12-02`]

quality achieved though using different meta-imputation settings. Hence, QC results were separately reported for each part of the analysis. A summary of the described analysis is illustrated in Figure 2.4 (page 60).

Calculation of genotype accuracy

Genotype accuracy was calculated as the squared Pearson correlation coefficient, r^2, as a measure for the strength of the linear relationship between imputed and masked genotype vectors, such that r^2 was computed per site. This was done after conversion of genotypes to allelic dosage, calculated as $d = 0p_0 + 1p_1 + 2p_2$ where $d \in \{0, 1, 2\}$ for masked genotypes or $0 \leq d \leq 2$ when calculated from imputed genotype probabilities. Note that the Pearson correlation coefficient is defined as the covariance divided by the product of the standard deviation (SD) of two random variables. This is problematic if SD = 0, which is the case when variant genotypes are imputed as being monomorphic. To compensate for this loss in precision towards lower frequencies, the coefficient was set to $r^2 = 0$ for monomorphic variants. Imputed and masked genotype data were sorted into minor allele frequency (MAF) bins, based on their population frequency (MAF in the GoT2D dataset). In the following, accuracy is reported as mean r^2 calculated at corresponding variants per MAF bin.

2.4.2 Results

Accuracy of meta-imputed genotypes was explored for each combination of score metric and merge operation. The best performing setting was then chosen for comparison to direct imputations, as well as further analysis in Section 2.5 (page 72).

Comparison of meta-imputation settings

Each combination of score metric and merge operation produced an identical set of variants; that is, the combined set of variants across imputed panels. In total, 253,852 variants were returned from each meta-imputation in Scenario A (European subpopulations) and 559,172 in Scenario B (continental populations); *i.e.* the same number as captured by the union panel. Meta-imputed datasets were further reduced to the set of variants that matched to masked variants in the original GoT2D dataset. Variants contained in the genotype scaffold were removed, as these were not imputed. This retained 181,561 and 196,300 variants in Scenarios A and B, respectively.

First, I report the impact of quality control (QC) to characterise the different meta-imputation settings by the number of retained sites. Briefly, recall that QC was carried out to remove variant sites at IMPUTE2 info-score < 0.4 and at deviations from HWE at p-value $< 1 \times 10^{-4}$. I then report genotype accuracy measured on sites retained after QC for each setting.

Table 2.2: Variants retained after quality control per meta-imputation setting. The number of variants retained after quality control (QC), n, per meta-imputation setting (combination of score metric and merge operation) in Scenario A and B. The percentage is given relative to the set of sites matched to masked variants in the GoT2D dataset and after removing sites contained in the imputation scaffold; 181,561 and 196,300 in A and B, respectively. Variants were removed at IMPUTE2 info-score < 0.4 and at deviations from HWE at p-value $< 1 \times 10^{-4}$.

Merge	Score	Scenario A		Scenario B	
		n retained	(%)	n retained	(%)
MSS	MP	168,595	(92.9)	178,034	(90.7)
	IS	169,455	(93.3)	179,677	(91.5)
	SC	168,686	(92.9)	179,449	(91.4)
	RS	165,877	(91.4)	171,517	(87.4)
WLC	MP	161,079	(88.7)	166,458	(84.8)
	IS	162,511	(89.5)	169,860	(86.5)
	SC	160,464	(88.4)	165,907	(84.5)
	RS	160,369	(88.3)	165,787	(84.5)

The number of variants retained after QC differed among meta-imputation settings; see Table 2.2 (this page). Merge operations had a higher impact on the quality of meta-imputed genotypes than score metrics. In Scenario A, on average 92.6 % (±0.431 % SE) of variants were retained when MSS (maximum score selection) was used as the merge operation, with fewer retained using WLC (weighted linear combination), where 88.7 % (±0.272 % SE) were retained on average. This was similar in Scenario B, where 90.3 % (±0.977 % SE) and 85.1 % (±0.491 % SE) were retained on average under MSS and WLC, respectively. Most of the variants removed in either setting were low in frequency. For instance at MAF ≤ 1%, 74.6 % (±0.705 % SE) and 68.3 % (±0.495 % SE) passed QC in Scenario A when using MSS and WLC, respectively, as well as 72.2 % (±1.73 % SE) and 61.7 % (±0.962 % SE) in

Scenario B, respectively. Among score metrics, the number of variants that passed QC was lowest for RS (random scores) in each comparison; for example, 67.5 % and 60.5 % at MAF ≤ 1% in A and B, respectively.

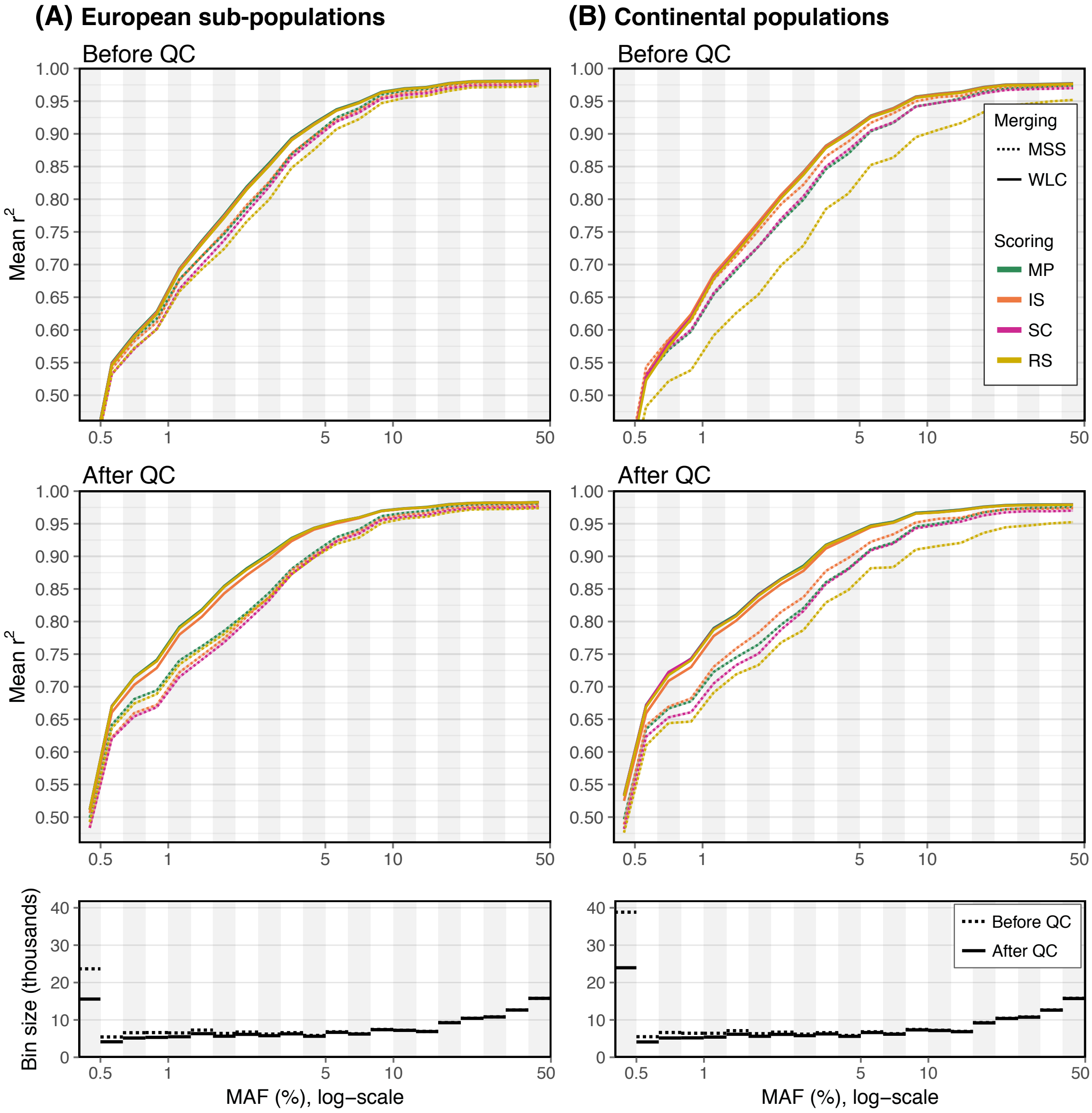

Figure 2.5: Accuracy comparison of score metrics and merge operations in meta-imputation. Each combination of merge operation (MSS and WLC) and score metric (MP, IS, SC, and RS) was examined in Scenarios A and B. Accuracy was measured as mean r^2 calculated between meta-imputed variants and variants masked in the GoT2D dataset. Results are shown both before and after QC. Bin sizes were defined on log-scale where *grey-white* bars indicate boundaries. The panels at the bottom indicate the number of variants per bin before QC (*dotted*) and the average number of variants per bin after QC (*solid*).

Although MSS overall preserved a relatively large proportion of markers after QC, the accuracy of retained genotypes was overall lower compared to data produced under WLC.

Imputation accuracy improved after QC as illustrated in Figure 2.5 (page 63), which shows mean r^2 calculated in MAF bins of equal size on log-scale. The differences among settings were small, in particular among score metrics when WLC was used, but where differences in accuracy become more pronounced after QC, which highlighted a clear distinction between merge operations. Throughout, mean r^2 was higher for genotype data produced under WLC. In Scenario A, for example, mean r^2 at MAF $\leq$ 1% before QC was 0.472 (±0.886×10^{-3} SE) in WLC and 0.464 (±0.894×10^{-3} SE) in MSS, but showed a larger difference after QC, namely 0.605 (±1.01×10^{-3} SE) and 0.472 (±0.886×10^{-3} SE) in WLC and MSS, respectively. This was also seen in Scenario B, where mean r^2 at MAF $\leq$ 1% was 0.428 (±0.796×10^{-3} SE) and 0.418 (±0.811×10^{-3} SE) before QC in WLC and MSS, respectively, as well as 0.600 (±0.951×10^{-3} SE) in WLC and 0.548 (±0.914×10^{-3} SE) in MSS after QC. Accuracy differences between merge operations were more pronounced at higher MAF; as seen in Figure 2.5. For example, at MAF $\leq$ 5% after QC, mean r^2 was 0.873 (±0.442×10^{-3} SE) and 0.811 (±0.580×10^{-3} SE) in Scenario A for WLC and MSS, respectively, as well as 0.862 (±0.472×10^{-3} SE) and 0.793 (±0.649×10^{-3} SE) in Scenario B, respectively.

Accuracy as measured for each setting after QC is given in Table 2.3 (next page), which shows mean r^2 computed in three broader MAF bins to summarise accuracy levels at rare variants (here defined at MAF $\in [0.00, 0.01]$), low-frequency (MAF $\in (0.01, 0.05]$), and common variants (MAF $\in (0.05, 0.50]$). The RS score metric overall resulted in less accurate genotype data compared to other metrics, in particular in Scenario B where RS was least accurate in all comparisons. This was not the case in Scenario A, where it showed a higher accuracy than IS and SC at rare and low-frequency variants. However, note that accuracy differences among score metrics were low overall in Scenario A (see Table 2.3), due to the presumed higher genetic similarity between sample individuals and reference haplotypes (recall that the GoT2D sample is composed of individuals of Central and Northern European descent).

Regardless, MP (maximum probability) outperformed other score metrics in most comparisons; except in Scenario B, for low-frequency variants under MSS, where it was outperformed by IS. The MP score metric was found to further improve accuracy under

Table 2.3: Accuracy measured for each meta-imputation setting. Accuracy was measured as mean r^2 (±SE) per MAF bin; defined to reflect average levels of accuracy measured at rare, low-frequency, and common variants. Reported values were measured after QC for each meta-imputation setting (combination of merge operation and score metric), in Scenarios A and B. The setting with the highest accuracy per MAF bin and per scenario is highlighted (**bold**).

MAF bin	Merge	Score	Scenario A		Scenario B	
			Mean r^2(± SE*)	n	Mean r^2(± SE*)	n
[0.00, 0.01]	`MSS`	`MP`	0.585 (1.947)	31,694	0.557 (1.823)	42,101
		`IS`	0.567 (1.968)	32,099	0.554 (1.819)	42,769
		`SC`	0.565 (1.952)	31,646	0.544 (1.769)	42,335
		`RS`	0.578 (1.989)	30,693	0.536 (1.901)	38,468
	`WLC`	`MP`	**0.608** (2.019)	28,818	**0.603** (1.912)	35,020
		`IS`	0.600 (2.004)	29,461	0.592 (1.870)	37,049
		`SC`	0.606 (2.027)	28,607	0.603 (1.911)	34,793
		`RS`	0.606 (2.031)	28,545	0.601 (1.918)	34,748
(0.01, 0.05]	`MSS`	`MP`	0.818 (1.169)	43,330	0.799 (1.325)	42,865
		`IS`	0.809 (1.168)	43,787	0.814 (1.220)	43,341
		`SC`	0.804 (1.146)	43,499	0.790 (1.231)	43,570
		`RS`	0.813 (1.157)	41,848	0.769 (1.416)	40,173
	`WLC`	`MP`	**0.876** (0.871)	39,309	**0.865** (0.936)	39,240
		`IS`	0.867 (0.903)	40,107	0.856 (0.962)	40,376
		`SC`	0.874 (0.878)	39,000	0.863 (0.936)	38,992
		`RS`	0.874 (0.879)	38,972	0.863 (0.941)	38,943
(0.05, 0.50]	`MSS`	`MP`	0.970 (0.280)	93,571	0.960 (0.372)	93,068
		`IS`	0.967 (0.287)	93,569	0.962 (0.344)	93,567
		`SC`	0.964 (0.288)	93,541	0.956 (0.342)	93,544
		`RS`	0.962 (0.301)	93,336	0.931 (0.488)	92,876
	`WLC`	`MP`	**0.977** (0.181)	92,952	**0.973** (0.193)	92,198
		`IS`	0.976 (0.183)	92,943	0.972 (0.196)	92,435
		`SC`	0.976 (0.179)	92,857	0.972 (0.191)	92,122
		`RS`	0.976 (0.179)	92,852	0.972 (0.191)	92,096

[*] Standard error (SE) ×10^{-3}

`WLC`, such that the combination of `MP` and `WLC` was seen to yield the highest accuracy in each MAF bin and in both scenarios (as highlighted in Table 2.3). Therefore, in the following, `WLC` was chosen as merge operation and `MP` as score metric; hence, the combination of `MP` and `WLC` is implied when referring to meta-imputation below.

Improvements of accuracy in comparison to direct imputations

Available split panels were imputed into the generated study sample and imputed genotype data were then combined through meta-imputation. The union and intersection panels were separately imputed for subsequent comparison to meta-imputed genotypes. Before accuracy was measured, all data were subjected to QC and variants were removed when not matched to masked variants or when contained in the imputation scaffold. For simplicity, imputed datasets are referred to by the panel from which they were estimated.

Comparisons were based on mean r^2 calculated at corresponding (meta-)imputed and masked variants pooled by MAF bin. In addition, significant differences in the MAF distribution of imputed and corresponding meta-imputed variants were determined using the two-sample Kolmogorov—Smirnov (KS) test. However, significance was determined from the median of the KS test statistic, here denoted by $\widetilde{D}$, calculated at $n = 500$ randomly selected sites over 1,000 repeated draws. This was done to account for varying subset sizes retained in each comparison, and to avoid potential biases due to correlations of linkage disequilibrium (LD) at nearby markers. MAF distributions were significantly different if

$$\widetilde{D}_n \; > \; c(\alpha)\sqrt{\frac{2n}{n^2}} \tag{2.10}$$

for significance levels $c(0.05) = 1.36$ and $c(0.01) = 1.63$. A similar approach was applied by Pasaniuc *et al.* (2014) to compare signatures of functional enrichment in imputed data.

Table 2.4: Effect of quality control on imputed genotype data. The number (percent) of variants retained after QC for direct imputations (*i.e.* four split panels, intersection panel, and union panel) and meta-imputation. Numbers refer to variants retained after removing unmatched sites and those contained in the imputation scaffold.

Panel	Scenario A			Scenario B		
	Split	*n* retained	(%)	Split	*n* retained	(%)
Split panel (1)	CEU	135,218	(95.4)	AFR	123,662	(91.8)
Split panel (2)	FIN	141,017	(96.6)	AMR	155,266	(93.6)
Split panel (3)	GBR	137,277	(95.0)	ASN	99,531	(94.3)
Split panel (4)	TSI	138,613	(94.0)	EUR	161,364	(95.3)
Meta-imputed (1–4)	–	161,079	(88.7)	–	166,458	(84.8)
Intersection panel	–	116,980	(99.8)	–	92,312	(99.8)
Union panel	–	174,229	(96.0)	–	184,158	(93.8)

The numbers of retained variants for each panel are given in Table 2.4 (this page). Meta-imputed data showed the highest proportion of variants removed through QC. In Scenario A, 11.3 % of meta-imputed variants were removed, whereas only 0.197 % of variants in the intersection and 4.04 % in the union panel were removed, compared to an average of 4.74 % (±0.536 % SE) among split panels. Note that only 3.40 % of markers did not pass QC after imputation from the FIN sub-population. The proportion of meta-imputed genotypes removed after QC was also highest in Scenario B (15.2 %) which is compared to only 0.163 % in the intersection and 6.19 % in the union panel, as well as 6.23 % (±0.735 % SE) on average in split panels, where the lowest proportion of removed variants was seen for the EUR panel (4.66 %). However, the number of retained variants in meta-imputed data (161,076 and 166,458 in A and B, respectively) exceeded those retained in any split panel or the intersection panel; see Table 2.4.

Each of the imputed datasets was compared separately to meta-imputation, on the same set of variants retained after QC. The distribution of accuracy (mean r^2) measured by MAF is shown in Figure 2.6 (next page); average accuracy measured for each imputation strategy in comparison to meta-imputation is given in Table 2.5 (page 69), where accuracy was measured by MAF to distinguish rare variants ($\text{MAF} \in [0.00, 0.01]$), low-frequency ($\text{MAF} \in (0.01, 0.05]$), and common variants ($\text{MAF} \in (0.05, 0.50]$).

In Scenario A, meta-imputation showed an improvement in accuracy over imputations from split panels. For example, for rare variants, the highest improvement among split panel comparisons was seen with the GBR sample, where mean r^2 was 0.637 ($\pm 3.37\times10^{-3}$ SE) for GBR and 0.659 ($\pm 3.30\times10^{-3}$ SE) for meta-imputed data. Differences were larger at low-frequency, where the highest improvement was seen in comparison with the TSI sample; 0.865 ($\pm 1.16\times10^{-3}$ SE) and 0.907 ($\pm 0.808\times10^{-3}$ SE) for TSI and meta-imputation, respectively. Only the union panel was higher in accuracy than meta-imputation; *e.g.* 0.697 ($\pm 1.87\times10^{-3}$ SE) and 0.627 ($\pm 2.00\times10^{-3}$ SE) for rare variants, respectively, and 0.893 ($\pm 0.801\times10^{-3}$ SE) and 0.877 ($\pm 0.859\times10^{-3}$ SE) at low-frequency variants, respectively. Meta-imputation showed approximately equal levels of accuracy as the union panel at common variants, where the difference in mean r^2 was 0.00318 ($\pm 0.811\times10^{-4}$ SE).

Genotype accuracy showed higher differences in Scenario B, where meta-imputation improved accuracy in most split panel comparisons. For rare variants, the highest difference was seen to genotype data imputed from the AFR split panel, where mean r^2 was 0.703 ($\pm 3.24\times10^{-3}$ SE), compared to 0.745 ($\pm 3.07\times10^{-3}$ SE) for meta-imputed genotypes. However, meta-imputation showed similar accuracy as the imputation from the EUR

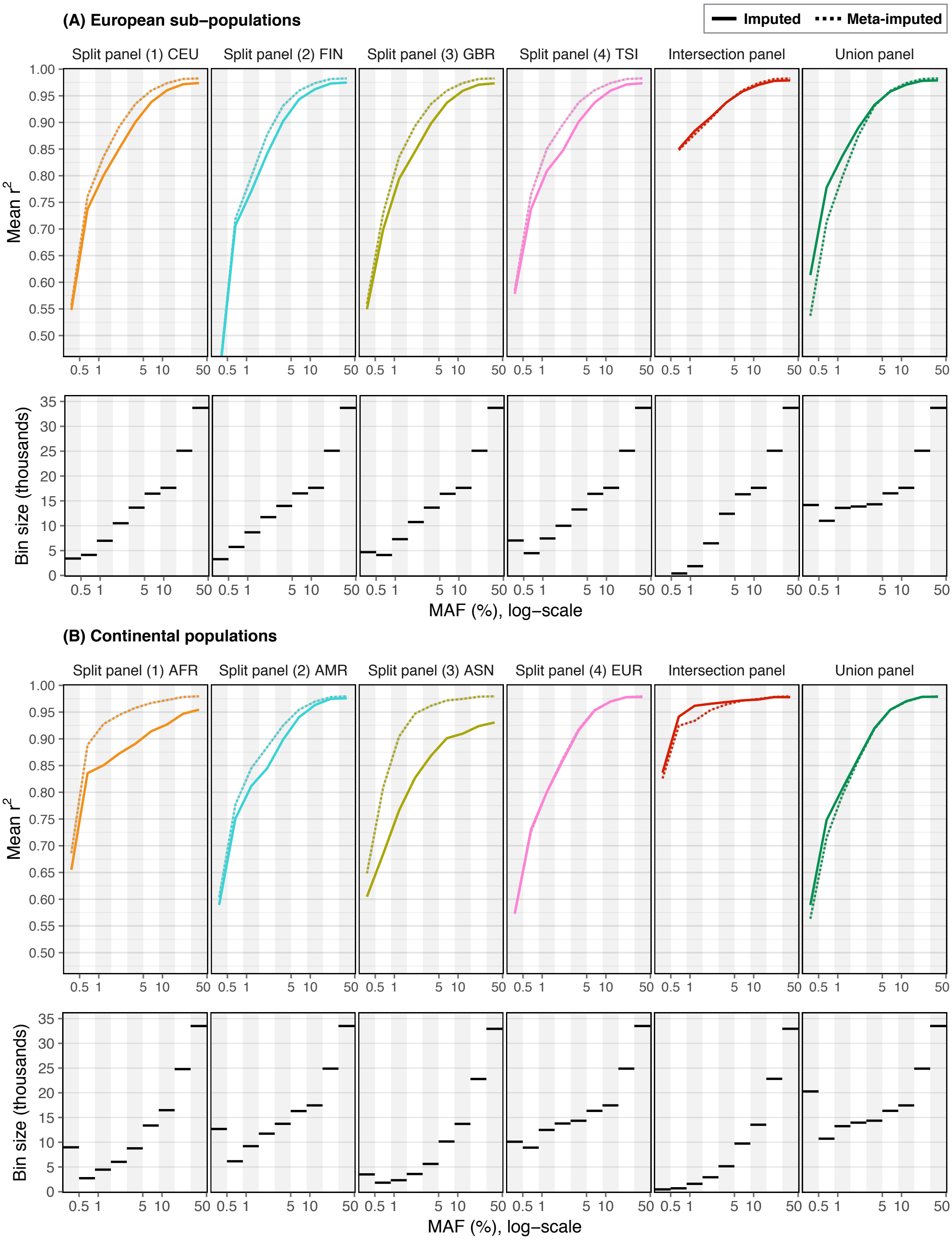

Figure 2.6: Accuracy comparison between meta-imputation and direct imputations. Accuracy was measured as mean r^2 per MAF bin, defined on log-scale where *grey-white* bars indicate boundaries. Each imputed panel (imputations from the four split panels, the intersection panel, and the union panel) was separately compared to meta-imputation on the same set of variants per bin (*i.e.*on the intersected set of SNPs); shown for variants retained after QC, in Scenarios A and B. MAF bins were defined on the actual allele frequencies as determined by the GoT2D dataset. Note that mean r^2 is not shown if the number of markers dropped below 50 per MAF bin. Panels at the bottom show the number of variants compared per bin.

Table 2.5: Accuracy of imputation strategies at rare, low-frequency, and common variants. Accuracy was calculated as mean r^2 per MAF bin on the same set of variants retained after QC in each comparison between meta-imputation and direct imputation, where n denotes the number of variants compared. The imputation strategy with the highest accuracy is highlighted (**bold**). The median of KS test statistic, $\widetilde{D}_{500}$, determined whether imputed and meta-imputed MAF distributions were significantly different; see Equation (2.10) on page 66.

MAF bin	Panel	n	Imputation		Meta-imputation		KS test†
			Mean r^2	(± SE‡)	Mean r^2	(± SE‡)	$\widetilde{D}_{500}$
Scenario A	(European sub-populations)						
[0.00, 0.01]	Split panel, CEU	8,636	0.665	(3.491)	**0.683**	(3.402)	0.046
	Split panel, FIN	10,416	0.619	(3.292)	**0.630**	(3.255)	0.024
	Split panel, GBR	10,023	0.637	(3.371)	**0.659**	(3.296)	0.034
	Split panel, TSI	12,763	0.654	(2.974)	**0.670**	(2.909)	0.098*
	Intersection panel	546	0.823	(9.901)	**0.824**	(9.832)	0.028
	Union panel	27,712	**0.697**	(1.873)	0.627	(2.001)	0.264**
(0.01, 0.05]	Split panel, CEU	30,012	0.866	(1.092)	**0.902**	(0.813)	0.040
	Split panel, FIN	32,969	0.853	(1.076)	**0.885**	(0.887)	0.032
	Split panel, GBR	30,442	0.860	(1.119)	**0.901**	(0.813)	0.036
	Split panel, TSI	29,445	0.865	(1.164)	**0.907**	(0.808)	0.036
	Intersection panel	20,604	**0.925**	(0.850)	0.923	(0.803)	0.034
	Union panel	39,213	**0.893**	(0.801)	0.877	(0.859)	0.088*
(0.05, 0.50]	Split panel, CEU	92,885	0.964	(0.269)	**0.977**	(0.180)	0.014
	Split panel, FIN	92,936	0.966	(0.250)	**0.977**	(0.181)	0.012
	Split panel, GBR	92,845	0.964	(0.273)	**0.977**	(0.181)	0.012
	Split panel, TSI	92,840	0.964	(0.283)	**0.977**	(0.180)	0.012
	Intersection panel	92,751	0.973	(0.212)	**0.977**	(0.180)	0.012
	Union panel	92,938	0.973	(0.212)	**0.977**	(0.181)	0.012
Scenario B	(Continental populations)						
[0.00, 0.01]	Split panel, AFR	12,495	0.703	(3.238)	**0.745**	(3.070)	0.030
	Split panel, AMR	20,416	0.653	(2.480)	**0.672**	(2.388)	0.040
	Split panel, ASN	5,661	0.640	(4.972)	**0.714**	(4.586)	0.082
	Split panel, EUR	21,223	**0.658**	(2.227)	0.656	(2.207)	0.048
	Intersection panel	1,364	**0.907**	(4.420)	0.892	(4.430)	0.172**
	Union panel	33,430	**0.653**	(1.871)	0.626	(1.894)	0.148**
(0.01, 0.05]	Split panel, AFR	18,468	0.879	(1.631)	**0.948**	(0.767)	0.048
	Split panel, AMR	33,093	0.861	(1.158)	**0.894**	(0.855)	0.036
	Split panel, ASN	11,181	0.837	(2.414)	**0.948**	(0.954)	0.114**
	Split panel, EUR	38,417	**0.867**	(0.969)	0.870	(0.910)	0.032
	Intersection panel	9,443	**0.967**	(0.811)	0.957	(0.816)	0.062
	Union panel	39,140	**0.871**	(0.949)	0.866	(0.924)	0.058
(0.05, 0.50]	Split panel, AFR	88,152	0.941	(0.426)	**0.976**	(0.172)	0.022
	Split panel, AMR	92,144	0.967	(0.263)	**0.973**	(0.191)	0.016
	Split panel, ASN	79,581	0.921	(0.519)	**0.978**	(0.165)	0.024
	Split panel, EUR	92,188	0.972	(0.218)	**0.973**	(0.193)	0.016
	Intersection panel	79,051	0.976	(0.201)	**0.977**	(0.168)	0.016
	Union panel	92,187	**0.973**	(0.214)	0.973	(0.193)	0.016

† Median of the Kolmogorov–Smirnov (KS) test statistic, $\widetilde{D}$; empirical CDF of imputed and meta-imputed MAF tested at $\alpha = 0.05$ ($*$) and $\alpha = 0.01$ ($**$).

‡ Standard error (SE) $\times 10^{-3}$.

sample, where the difference in mean r^2 was 0.00208 (±0.716×10^{-4} SE). Likewise, at low-frequency, mean r^2 was 0.879 (±1.63×10^{-3} SE) for AFR and 0.948 (±0.767×10^{-3} SE) for meta-imputation, and the difference in accuracy was 0.00249 (±0.367×10^{-3} SE) with regard to the EUR split panel. As in Scenario A, differences were smaller for common variants, such that the difference in mean r^2 was below 0.001 in comparisons to imputations from the AFR, AMR, and EUR panels, but where the ASN sample showed the highest difference; mean r^2 was 0.921 (±0.519×10^{-3} SE) for ASN and 0.978 (±0.165×10^{-3} SE) for meta-imputation. The union panel was similar in accuracy as meta-imputation, where the overall difference in mean r^2 was 0.0106 (±8.47×10^{-3} SE).

Imputations from the intersection panel in Scenario A and B showed approximately equal levels of accuracy to meta-imputation. The difference in mean r^2 averaged to 0.000811 (±1.43×10^{-4} SE) across MAF in Scenario A, and 0.00824 (±4.82×10^{-3} SE) in Scenario B. However, note that the number of variants in the intersection panel was the lowest among available panel data in both scenarios (Table 2.4), and was further reduced as accuracy was measured on the same sets of variants retained in both the intersection and the meta-imputed datasets. For example, the comparison between the intersection panel and meta-imputation included only 546 variants at MAF ≤ 1% in Scenario A and 1,364 variants in Scenario B, whereas each split panel and the union panel were compared on several thousands of variants at this frequency range. The high accuracy of genotypes imputed from the intersection panel may result from retaining only those variants that are "cosmopolitan" within the scope of the present evaluation.

Further, the empirical cumulative distribution function (CDF) of MAF at imputed and meta-imputed variants was compared per MAF bin. Differences are illustrated in Figure 2.7 (next page), which shows the CDF of compared variants in relation to the known population frequencies at masked variants in the GoT2D dataset; calculated by subtracting (meta-)imputed frequencies from masked frequencies (ΔMAF) at the same set of markers. Notably, meta-imputed frequencies showed high consistency with imputed frequencies at rare variants (MAF $\in [0.00, 0.01]$) across split panel imputations, but were skewed in comparison to imputations from the union panel in both scenarios. Significant differences were found for rare variant imputations from the TSI sample ($\widetilde{D} = 0.098$) in Scenario A,

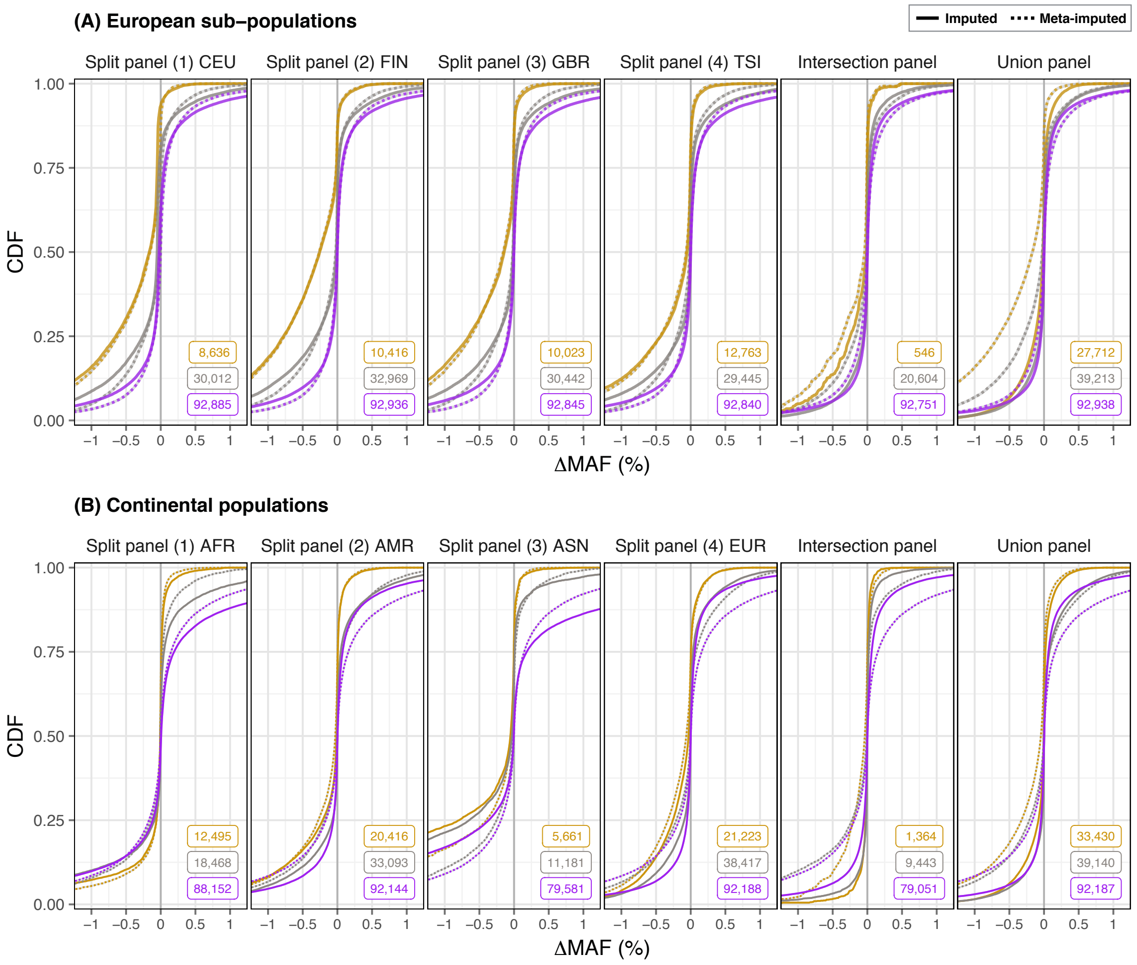

Figure 2.7: Difference between imputed and masked minor allele frequency. Comparison of imputed and meta-imputed MAF in relation to known population frequencies, compared on the same set as retained after QC in each comparison. Frequency difference, ΔMAF, was calculated as the MAF observed at a masked variant minus MAF at the corresponding (meta-)imputed variant, pooled in three MAF bins; rare variants (MAF $\in [0.00, 0.01]$; *yellow*), low-frequency (MAF $\in (0.01, 0.05]$; *grey*), and common variants (MAF $\in (0.05, 0.50]$; *purple*). Numbers per MAF bin per comparison are given in each panel (*colour-coded*).

as well as for the union panel at rare and low-frequency variants (0.264 and 0.088, respectively). In Scenario B, imputed and meta-imputed differences were significantly different for rare variant imputations from the intersection panel and the union panel (0.172 and 0.148, respectively) and for the union panel at low-frequency variants (0.114). These results suggested that meta-imputation was able to correctly reproduce realistic allele frequency distributions from the combination of imputed genotypes from different sources, while achieving higher or similar accuracy compared to direct imputations from split panels. Results of KS tests in each comparison are given in Table 2.5 (page 69).

In summary, split panel imputations were either outperformed or similar levels of accuracy were achieved in direct comparisons to meta-imputed data; see Table 2.5 for a complete summary of genotype accuracy measured in each comparison. Although imputations from the union panel outperformed meta-imputation, such differences may be expected given that the union panel contained all the information which meta-imputation had to leverage indirectly from several data sources. Nonetheless, the present evaluation of genotype accuracy was limited with regard to coverage; for instance, genotype data imputed from the intersection panel was found to be relatively high in accuracy and similar with regard to meta-imputed data, but the low number of variants present in the intersection panel may not yield similar improvements under realistic conditions in association analyses. Therefore, to provide a comprehensive assessment of the meta-imputation method and to account for a potential tradeoff between accuracy and coverage, I conducted a more extensive power analysis in the following section.

2.5 Power to detect significant risk signals

The power of meta-imputation to detect disease risk factors in association tests was evaluated using simulated sample data. This was done in consideration of expected power when causal risk factors vary in their allele frequency as well as risk effect size. In particular, a series of simulated case-control association experiments was conducted, from which the power to detect significant association signals was determined, at specified allele frequencies and effect size of simulated risk factors. The description of the methods used is provided below (Section 2.5.1, this page), followed by the presentation of results (Section 2.5.2, page 77).

2.5.1 Methods

The same regime to carry out imputation and QC was followed as described in Section 2.4.1 (page 58). An additional set of haplotype reference data was available from four independent sequencing studies, which were included here as Scenario C; see below.

Finns. A Finnish cohort composed of data from the Sequencing Initiative Suomi Project (*SISu*) and the *Finrisk* Project (Vartiainen *et al.*, 2010; Pajunen *et al.*, 2010; Lim *et al.*, 2014; Borodulin *et al.*, 2015); 4x depth; sample size and number of SNPs considered here were N = 1,941 and M = 283,654, respectively.

GoNL. The Genome of the Netherlands Project (Boomsma *et al.*, 2013; Deelen *et al.*, 2014; Genome of the Netherlands Consortium, 2014); 12x depth, consisting of a representative sample of 250 trio-families; N = 748, M = 362,694.

ORCADES. The Orkney Complex Disease Study of genetic epidemiology of an isolated population in northern Scotland (McQuillan *et al.*, 2008); 4x depth, family-based data; N = 399, M = 236,755.

UK10K. The *UK10K* Genome Sequencing Project (UK10K Consortium *et al.*, 2015); 6.5x depth; N = 3,642, M = 527,199.

Also, an intersection panel was prepared from these four datasets, but no union panel. As before, only data from chromosome 20 were considered. Note that the above datasets were part of the early stage HRC testing phase (McCarthy *et al.*, 2016).[*]

Simulation of study sample data

Simulations were performed using HAPGEN version 2.2.0 (Su *et al.*, 2011), which requires a *template* dataset of haplotypes to reproduce realistic variant data in HWE, such that LD patterns in the simulated dataset are consistent with the haplotype sample. Individual sites can be simulated to independently act as causal disease variants with specified relative risk. The simulation generates two GWA samples of individuals that are affected (*cases*) or not affected (*controls*) by a disease phenotype. Data are identical in coverage as the template dataset.

Here, simulations were performed using GoT2D data (chromosome 20) to serve as the template dataset. The size of simulated case and control samples was fixed to 2,500 individuals each. Although a larger sample would have been beneficial in terms of

[*] Acknowledgement: Data provided by Professor Jonathan Marchini, Department of Statistics, University of Oxford; prior to the release of the HRC dataset.

signal detection through association testing, exceeding the size of the template dataset ($N = 2{,}657$) was expected to result in factitious allele frequency changes. For example, an iterative re-sampling strategy could be applied to introduce new low-frequency variants (*e.g.* following Moutsianas *et al.*, 2015). However, this was not done here because the effect size of risk variants (as defined during simulation) would likely be affected by such a sampling process.

A series of simulation experiments was conducted in which one variant was selected per simulation to act as a causal risk factor. Its relative risk (RR) was defined for heterozygous genotypes (RR_{het}) in a log-additive disease model (*i.e.* multiplicative on linear scale); the following three risk categories were defined.

$$\begin{aligned}
\textit{Low risk}: &\quad RR_{het} = 1.2 \quad (RR_{hom} = 1.44) \\
\textit{Modest risk}: &\quad RR_{het} = 1.6 \quad (RR_{hom} = 2.56) \\
\textit{High risk}: &\quad RR_{het} = 2.0 \quad (RR_{hom} = 4.00)
\end{aligned}$$

The analysis was performed by conducting 300 replicate simulations per risk category, where variants occurring at different frequencies were selected in three defined MAF intervals; very low frequency (MAF $\in [0.5, 1]\,\%$), low frequency (MAF $\in (1, 5]\,\%$), and high frequency (MAF $\in (5, 50]\,\%$), such that 100 variants were drawn from each interval and simulated as risk variants.

Note that variant selection was done at random, regardless of presence or absence of the selected variant in any of the available reference panels, so as to mirror conditions encountered under realistic GWA settings; *i.e.* when a causal variant itself is absent in an imputation reference, its risk effects may be detectable through LD at neighbouring sites.

To generate a study sample for imputations, a variant scaffold was extracted from each simulation replicate. Because the set of simulated variants mirrored those in the GoT2D dataset, sites that matched with variants typed on *Illumina Omni2.5 Array* were identified and extracted. A scaffold thus contained 40,255 variants into which available reference panels were imputed. Note that simulations produced two datasets; one case and one corresponding control dataset. These were concatenated before imputation to ensure

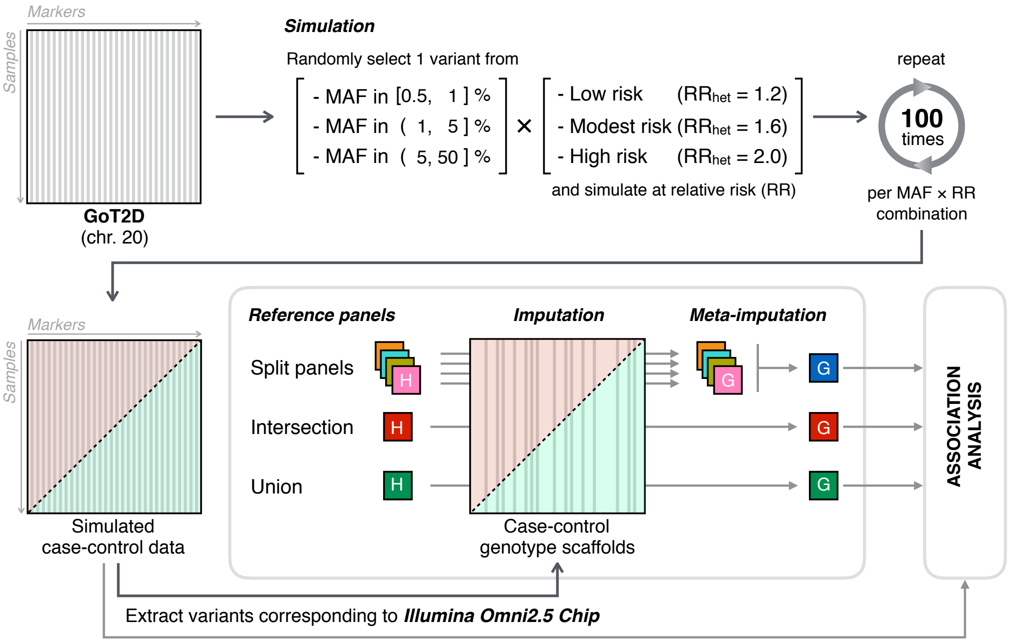

Figure 2.8: Illustration of the simulation process. Meta-imputation was assessed in terms of statistical power to detect significant risk association signals in a series of simulated case-control experiments. The GoT2D dataset was used as a template for simulations using HAPGEN (Su *et al.*, 2011), where one variant was randomly selected within one of three defined MAF intervals. The selected variant was then simulated to act as a causal disease variant in the simulated case-control dataset, where relative risk (RR_{het}) was defined according to one of three defined risk categories. In total, 100 replicate simulations were conducted per combination of MAF interval and risk category (per scenario). Simulated data were used to extract a genotype scaffold into which available reference panels were imputed, followed by meta-imputation of imputed datasets. Imputed and meta-imputed datasets were then subjected to association analysis, including the simulated (not imputed) datasets for comparison.

consistency in the imputation analysis. Imputed data were again separated into case and control samples prior to association analysis (described below). Because HAPGEN2 produces haplotype data, imputations were executed on pre-phased genotypes. A summary of the simulation process is illustrated in Figure 2.8 (this page).

Association analysis in imputed genotype data

Imputed case and control datasets were analysed using a frequentist score test under an additive model of association, implemented in SNPTEST version 2.5 (Marchini *et al.*, 2007). In contrast to the previous analysis (Section 2.4 on page 57), in which the variants not included in the extracted scaffold were masked to measure accuracy after imputation, here, the simulated case-control dataset was retained and separately examined in association analysis. This was done to enable comparisons of meta-imputed and imputed data to a non-imputed benchmark result for each simulation replicate.

The genomic control inflation factor, λ_{GC}, was calculated to investigate if systematic biases are present in association results, which is defined as the median of χ^2 test statistics resulting from case-control association tests divided by the expected median of the χ^2 distribution (Devlin *et al.*, 2001). Because the frequentist score test was used, λ_{GC} was calculated on basis of the resulting p-values from which the χ^2 statistic was calculated with one degree of freedom.

Calculation of power in replicate simulation experiments

Significant association signals were identified in each simulation and pooled by MAF interval and risk category, according to which variants were selected and simulated. The proportion of datasets in which significance was reached at the known risk variant was taken as a simple estimate for the statistical power to detect genetic risk effects. Note that the position of the simulated risk variant was known through simulation, but the variant itself may not be retained after imputation or QC. Therefore, signal detection was performed within a 1 Mb region around the position of the simulated risk variant, for any site reaching significance with this region.

Significance was defined at a nominal threshold of p-value $\leq 1 \times 10^{-6}$. Note that this threshold is higher (thus, less conservative) than commonly applied genome-wide thresholds, *e.g.* at 5×10^{-8} (*e.g.*, see Risch and Merikangas, 1996), because analyses were conducted on data from chromosome 20 only. However, to provide additional detail, power was estimated under a moving significance threshold; between p-value $\leq 1 \times 10^{-8}$ and p-value $\leq 1 \times 10^{-4}$. As a comparative measure between association results produced from the different imputation strategies, the difference in power between the non-imputed simulation dataset and a given (meta-)imputed dataset is reported, denoted by Δ_P, which is calculated as the average difference along the moving significance threshold.

2.5.2 Results

A number of 100 variants were selected per MAF interval such that there were 300 variants in total. Each was then simulated at the three defined risk categories such that 900 simulations were conducted from which a genotype scaffold was extracted for imputation. Given the four split panels, the intersection panel, and the union panel available per Scenario A and B, as well as the four independent reference datasets and the generated intersection panel in Scenario C, a total of 15,300 imputation analyses were performed. Imputed data were then combined in meta-imputation (except the intersection and union panels), resulting in 900 additional genotype datasets. Each dataset was then subjected to association analysis, including the non-imputed simulated case-control sample, which was used as a benchmark for comparisons. Hence, a total of 17,100 association analyses were conducted, where each was treated as an independent GWA study. All analyses were performed on whole-chromosome data (chromosome 20).

Association results were inspected with regard to inflation before and after QC; the difference is shown in Figure 2.9 (next page) where λ_{GC} is shown as the average per MAF interval. Inflation was slightly increased at higher risk variant frequencies. The difference of λ_{GC} before and after QC was small in Scenarios A and C, but noticeable in Scenario B, where association results of meta-imputed data were deflated ($\lambda_{GC} < 1$) before QC. After QC, λ_{GC} values of all imputed and meta-imputed datasets were approximately equal to inflation measured for the non-imputed simulation dataset in each scenario.

Association results were at $\lambda_{GC} \approx 1$ on average in each scenario when the simulated risk variant was very low in frequency (MAF $\in [0.5, 1]$ %), but increased to $\lambda_{GC} \approx 1.05$ for risk variants at higher frequencies (MAF $\in [5, 50]$ %). Although these results suggested no major inflation, higher values of λ_{GC} are generally expected in presence of population sub-structure, including cryptic relationships among individuals in the sample. One explanation why λ_{GC} increased at higher risk variant frequencies is that individuals in the case sample appeared to be more related to each other than the individuals in the control sample.

Association results for each imputation strategy (referring to results obtained on genotype data imputed from available reference panels and meta-imputation) were

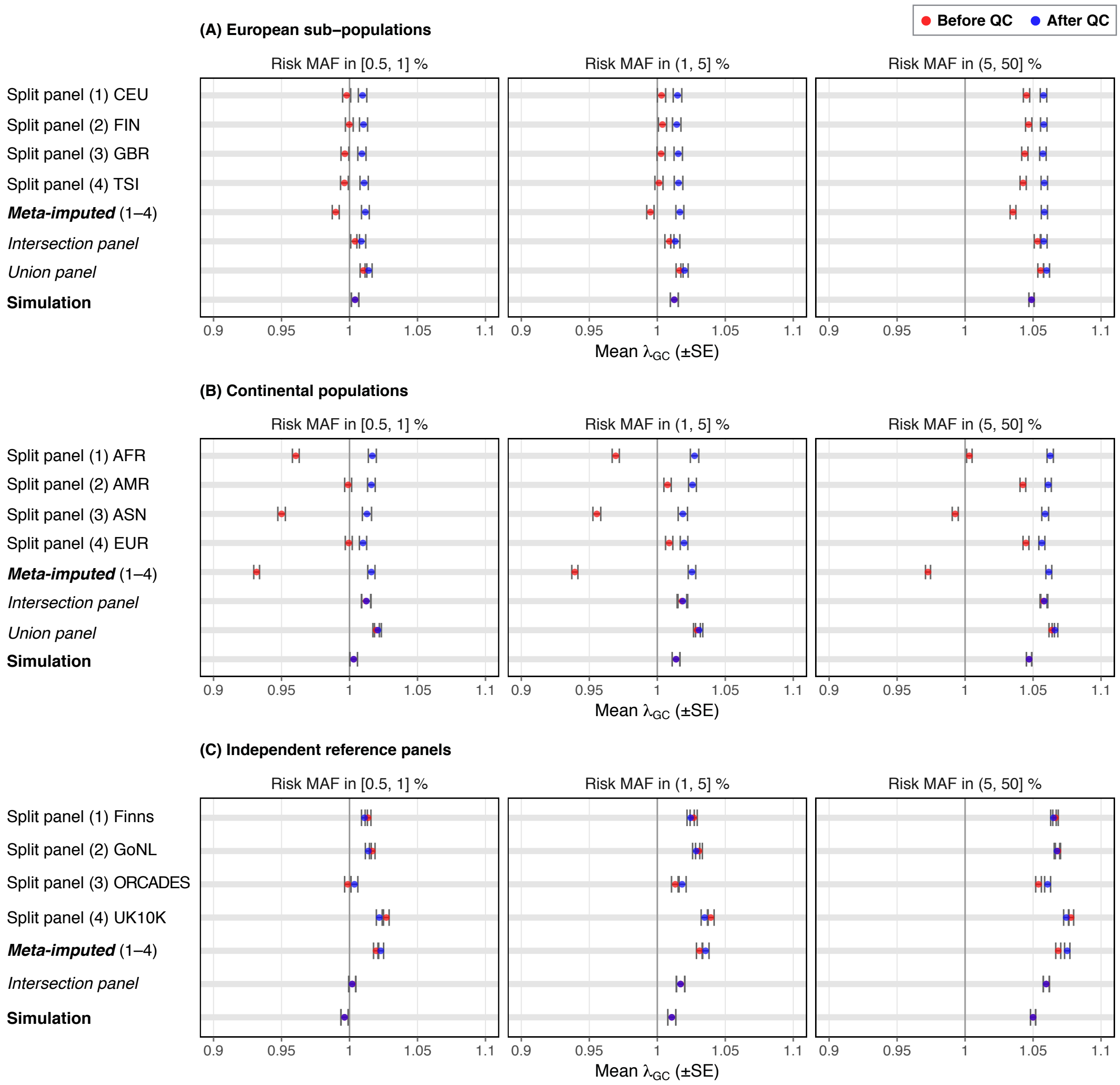

Figure 2.9: Inflation observed in simulated case-control experiments. Genomic control inflation factor calculated before (*red*) and after (*blue*) variants were filtered in QC, reported as mean λ_{GC} over replicate association results by MAF of the simulated risk variants.

separately evaluated with regard to each combination of risk category and the MAF interval from which simulated risk variants were selected. The distribution of power measured under a moving significance threshold (between p-value $\leq 1 \times 10^{-8}$ and p-value $\leq 1 \times 10^{-4}$) is shown in Figure 2.10; for Scenario A (next page), B (page 80), and C (page 81). The results are summarised in Table 2.6, for power measured at the nominal significance threshold (p-value $\leq 1 \times 10^{-6}$) and the average difference to the non-imputed simulation benchmark (Δ_P) along the moving threshold, averaged per MAF interval of simulated risk variants; for Scenario A (page 82), B (page 83), and C (page 84).

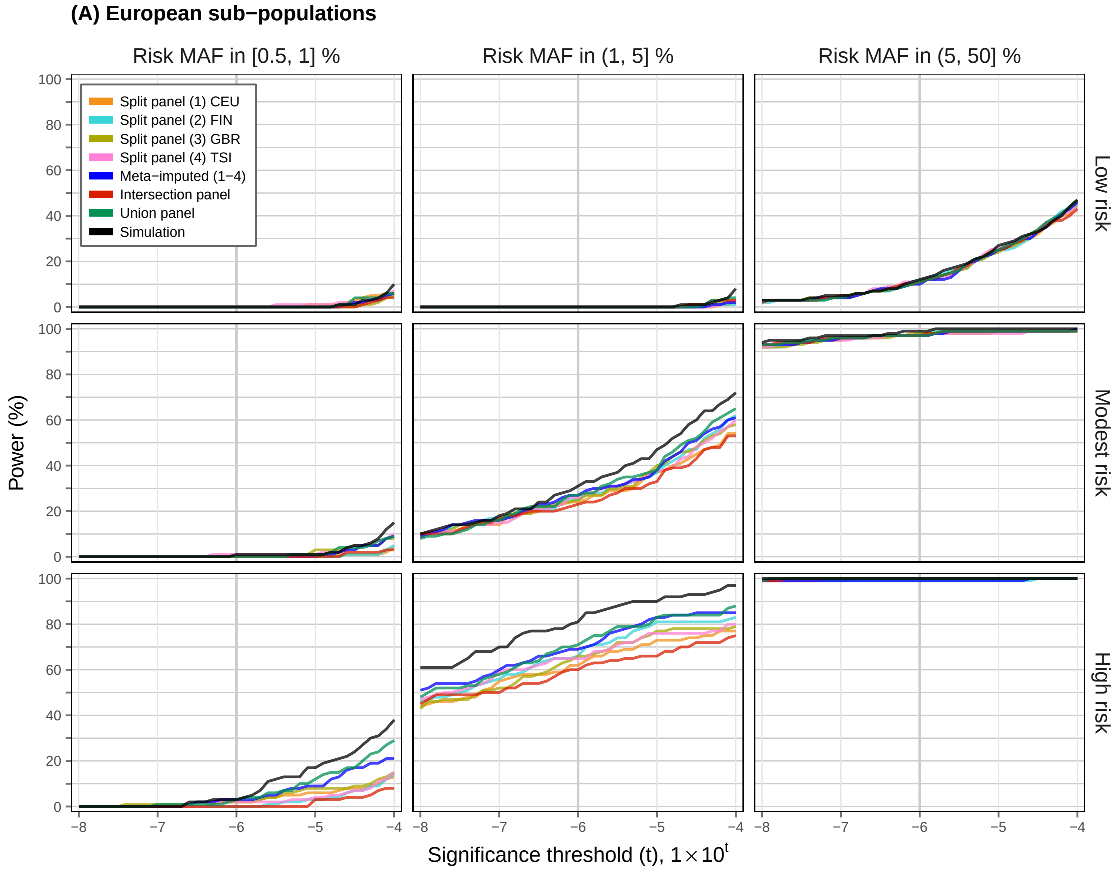

Figure 2.10: Power measured under a moving significance threshold. Power was calculated as the proportion of replicate association analyses ($n = 100$, per combination of risk category and MAF interval) in which any signal reached significance within 1 Mb around the position of a simulated risk variant. A moving significance threshold between p-value $\leq 1 \times 10^{-8}$ and p-value $\leq 1 \times 10^{-4}$ was applied to each association dataset. CORRECTED

The union panel was seen with the lowest average difference in power at very low frequencies of the simulated risk variant (MAF $\in [0.5, 1]\,\%$) in Scenario A, where Δ_P was 1.05 (± 0.206 SE). Meta-imputation showed the lowest average difference at very low MAF in Scenario B, $\Delta_P = 1.23\,\%$ ($\pm 0.247\,\%$ SE), as well as Scenario C, $\Delta_P = 1.32\,\%$ ($\pm 0.248\,\%$ SE); but recall that Scenario C (independent reference panels) did not contain a union panel. However, even in the high risk category in each scenario, estimated power did not exceed 3% for any imputation strategy when the simulated risk variant was very low in frequency, such that observed differences were negligible as these could be attributed to stochastic noise. Similarly, observed differences were small in each risk category when causal variants were selected from the high frequency interval (MAF $\in [5, 50]\,\%$), where the lowest average difference in power was recorded

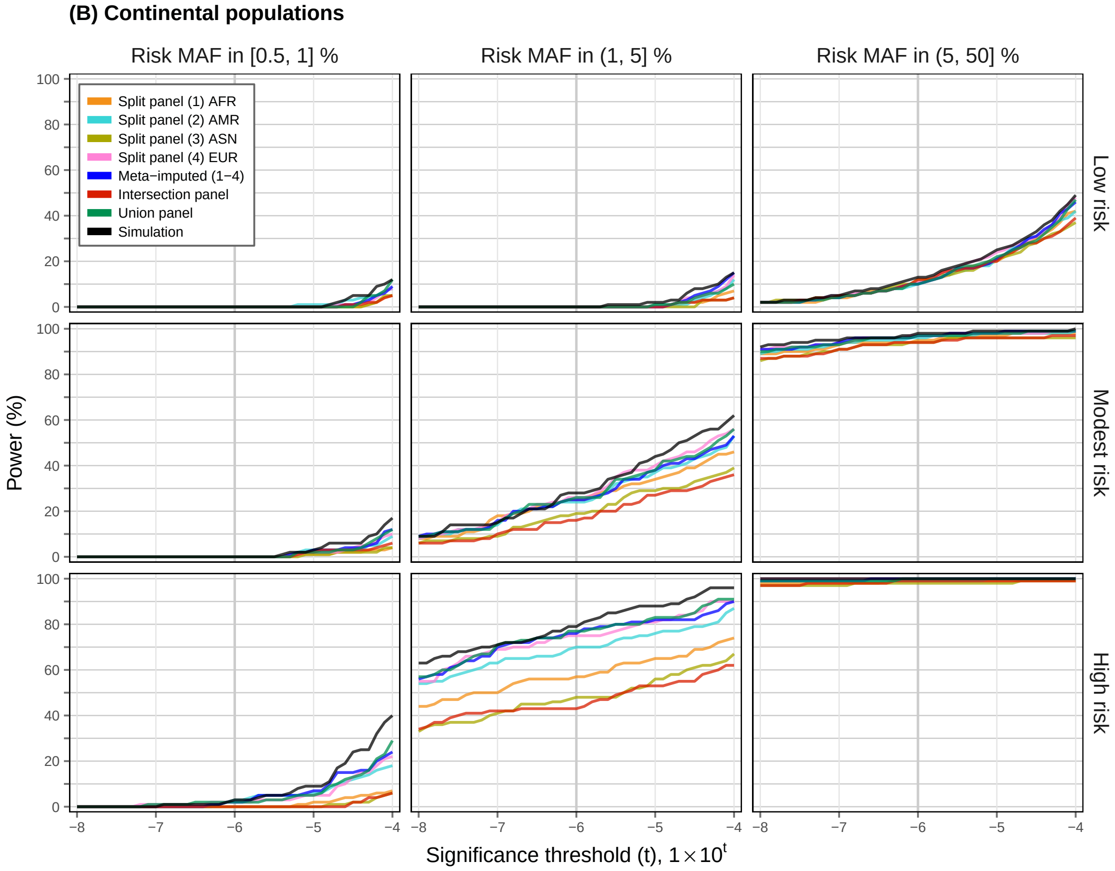

Figure 2.10: Continued. CORRECTED

for the union panel in Scenario A, $\Delta_P = 0.537\,\%$ ($\pm 0.067{,}6\,\%$ SE), the EUR split panel in B, 0.650 % (±0.072,1 % SE), and the intersection panel in C, 0.545 % (±0.091,8 % SE). However, note that $\Delta_P < 1\%$ in each strategy at high risk MAF in Scenarios A and C, but where some of the strategies showed larger differences in Scenario B, *e.g.* the ASN split panel and the intersection panel; 2.83 % (±0.177 % SE) and 2.56 % (±0.173 % SE), respectively.

Noticeable differences were seen among imputation strategies for simulated risk variants selected at low frequency ($\text{MAF} \in [1, 5]\,\%$). The union panel was recorded with the lowest difference in power relative to the non-imputed simulation benchmark in Scenarios A and B, 4.85 % (±0.416 % SE) and 2.65 % (±0.248 % SE), respectively, whereas the intersection panel had the highest difference, 9.56 % (±0.846 % SE) and 15.5 % (±1.25 % SE), respectively. Notably, meta-imputation was similarly close as the union panel and outperformed the other imputation strategies in Scenario A, 4.96 % (±0.431 % SE). For example, at a nominal threshold (p-value $\leq 1 \times 10^{-6}$), the

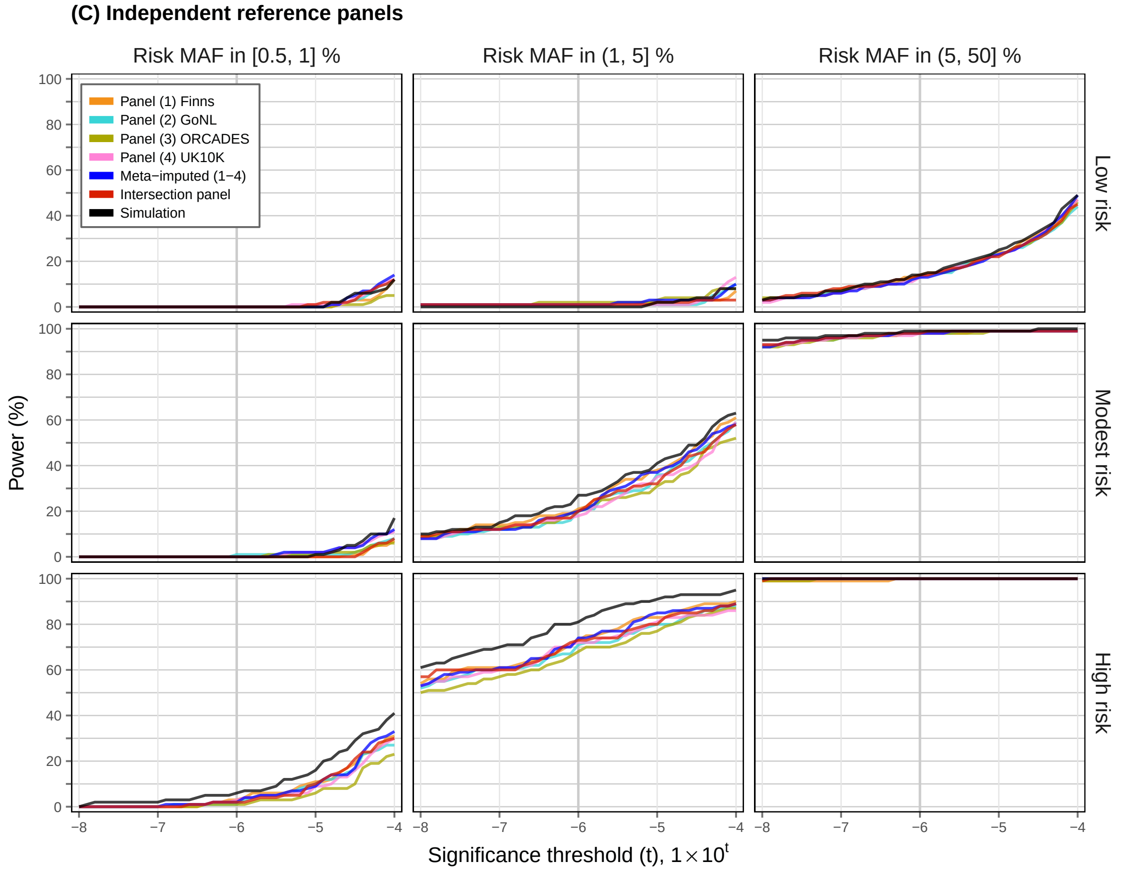

Figure 2.10: Continued. CORRECTED

union panel reached 71% power and meta-imputation 69% in the high risk category. In Scenario B, the power observed for meta-imputed data was high by comparison, *e.g.* 76% power at high risk, compared to 77% for the union panel and 43% for the intersection panel; however, Δ_P measured for meta-imputation was 3.07 % (±0.295 % SE), which was lower in the EUR split panel, 2.72 % (±0.259 % SE), reaching 76% in the high risk category. In Scenario C, the *Finns* panel showed the lowest difference in power, 3.11 % (±0.343 % SE), and the *ORCADES* panel the highest, 5.76 % (±0.572 % SE); yet, meta-imputation ranked 2nd best among the strategies compared, 3.51 % (±0.357 % SE), but 1st in the high risk category with 74% power (compared to 73% and 68% for *Finns* and *ORCADES*, respectively).

2.6 Discussion

Meta-imputation was presented as a novel approach to integrate reference data after imputation into a common study sample, but the idea of combining genotype data imputed

Table 2.6: Estimated power per imputation strategy. Power was estimated as the proportion of significant association signals found among replicate simulation experiments, for which one variant per simulation was selected at random from three MAF intervals (as specified in the table). Each of the selected variants was simulated to act as a causal risk factor, where relative risk was simulated in three categories; low ($RR_{het} = 1.2$), modest ($RR_{het} = 1.6$), and high risk ($RR_{het} = 2.0$). Power at a nominal significance threshold (p-value $\leq 1 \times 10^{-6}$) is reported at each combination of MAF interval and risk category. The average difference (Δ_P) in relation to the non-imputed simulation benchmark is given per MAF interval for each imputation strategy; the lowest average difference is highlighted (**bold**). This table shows the results obtained for imputed and meta-imputed data in Scenario A; results for Scenario B (next page) and Scenario C (page 84) are shown separately.

(A) European sub-populations

Risk MAF (%)	Panel	Power (%), p-value $\leq 1 \times 10^{-6}$			Δ_P (%)*	
		Low	Modest	High	Mean	(± SE)
$[0.5, 1]$	Split panel (1) CEU	0	0	2	2.537	(0.487)
	Split panel (2) FIN	0	0	0	3.041	(0.520)
	Split panel (3) GBR	0	0	2	2.098	(0.444)
	Split panel (4) TSI	0	1	2	2.309	(0.509)
	Meta-imputed (1-4)	0	0	3	1.553	(0.289)
	Intersection panel	0	0	0	3.366	(0.593)
	Union panel	0	0	3	**1.049**	(0.206)
$(1, 5]$	Split panel (1) CEU	0	24	62	8.561	(0.729)
	Split panel (2) FIN	0	25	66	6.374	(0.543)
	Split panel (3) GBR	0	25	66	7.431	(0.670)
	Split panel (4) TSI	0	26	65	7.122	(0.605)
	Meta-imputed (1-4)	0	27	69	4.959	(0.431)
	Intersection panel	0	23	60	9.561	(0.846)
	Union panel	0	27	71	**4.846**	(0.416)
$(5, 50]$	Split panel (1) CEU	11	98	100	0.780	(0.068)
	Split panel (2) FIN	12	98	99	0.894	(0.070)
	Split panel (3) GBR	11	98	100	0.821	(0.083)
	Split panel (4) TSI	11	97	100	0.748	(0.090)
	Meta-imputed (1-4)	10	97	99	0.959	(0.069)
	Intersection panel	11	97	100	0.634	(0.073)
	Union panel	11	97	100	**0.537**	(0.068)

[*] Average difference in power between simulated and (meta-)imputed association results (Δ_P); averaged over risk category (low, modest, and high risk) and association signals detected at a moving significance threshold; between p-value $\leq 1 \times 10^{-8}$ and p-value $\leq 1 \times 10^{-4}$.

from different reference panels has been investigated before. Chen *et al.* (2013) used low to high-depth sequencing data as references for imputations into a given study sample, where imputed data have been matched and combined at overlapping sites, but such that variant genotypes imputed from the high-quality panel were included preferentially. They have shown that this approach improved overall accuracy compared to each separately imputed dataset. Here, I considered several variations of this approach which I evaluated

Table 2.6: Continued.

(B) Continental populations

Risk MAF (%)	Panel	Power (%), p-value $\leq 1 \times 10^{-6}$			Δ_P (%)*	
		Low	Modest	High	Mean	(± SE)
$[0.5, 1]$	Split panel (1) AFR	0	0	0	3.000	(0.541)
	Split panel (2) AMR	0	0	3	1.488	(0.337)
	Split panel (3) ASN	0	0	0	3.203	(0.583)
	Split panel (4) EUR	0	0	2	1.553	(0.296)
	Meta-imputed (1-4)	0	0	2	**1.228**	(0.247)
	Intersection panel	0	0	0	3.049	(0.579)
	Union panel	0	0	2	1.301	(0.247)
$(1, 5]$	Split panel (1) AFR	0	25	57	9.366	(0.866)
	Split panel (2) AMR	0	24	70	5.220	(0.442)
	Split panel (3) ASN	0	19	48	14.618	(1.197)
	Split panel (4) EUR	0	26	75	2.715	(0.259)
	Meta-imputed (1-4)	0	25	76	3.065	(0.295)
	Intersection panel	0	16	43	15.545	(1.248)
	Union panel	0	26	77	**2.650**	(0.248)
$(5, 50]$	Split panel (1) AFR	11	95	99	1.984	(0.104)
	Split panel (2) AMR	10	96	99	1.407	(0.109)
	Split panel (3) ASN	12	94	98	2.829	(0.177)
	Split panel (4) EUR	11	97	100	**0.650**	(0.072)
	Meta-imputed (1-4)	10	97	100	0.951	(0.089)
	Intersection panel	12	94	99	2.561	(0.173)
	Union panel	10	97	100	1.138	(0.093)

* See Table 2.6A (page 82).

using several reference datasets as available in different use case scenarios. Notably, the meta-imputation method does not require prior knowledge to guide the merging process (such as high or low quality of each dataset considered), which instead is determined by summary information derived directly from imputed genotype data.

The results I presented in this chapter showed that the combination of genotype data may indeed result in an increase of accuracy across the allele frequency spectrum, but where the largest improvements were seen for low-frequency variants (*e.g.* 1–5% MAF). I showed that meta-imputation improved genotype accuracy such that single-reference imputations were outperformed (*e.g.* in Scenario A), but also that meta-imputed genotype data may not further increase accuracy if a reference is highly accurate by itself (*e.g.* the EUR sample in Scenario B). Nonetheless, the inclusion of other, more distantly related reference haplotypes may not affect the accuracy of the resulting meta-imputed dataset (*e.g.* the AFR or ASN samples for imputation into the European sample in Scenario B).

Table 2.6: Continued.

(C) Independent reference panels

Risk MAF (%)	Panel	Power (%), p-value $\leq 1 \times 10^{-6}$			Δ_P (%)*	
		Low	Modest	High	Mean	(± SE)
$[0.5, 1]$	Panel (1) Finns	0	0	3	1.919	(0.249)
	Panel (2) GoNL	0	1	1	1.862	(0.290)
	Panel (3) ORCADES	0	0	1	2.805	(0.416)
	Panel (4) UK10K	0	0	3	1.683	(0.305)
	Meta-imputed (1-4)	0	0	2	**1.317**	(0.248)
	Intersection panel	0	0	2	1.854	(0.258)
$(1, 5]$	Panel (1) Finns	1	21	73	**3.114**	(0.343)
	Panel (2) GoNL	1	20	71	4.943	(0.436)
	Panel (3) ORCADES	2	20	68	5.764	(0.572)
	Panel (4) UK10K	1	18	72	4.789	(0.428)
	Meta-imputed (1-4)	1	20	74	3.512	(0.357)
	Intersection panel	1	20	73	4.195	(0.387)
$(5, 50]$	Panel (1) Finns	14	98	100	0.683	(0.067)
	Panel (2) GoNL	13	98	100	0.821	(0.103)
	Panel (3) ORCADES	14	98	100	0.911	(0.096)
	Panel (4) UK10K	13	98	100	0.780	(0.079)
	Meta-imputed (1-4)	13	98	100	0.748	(0.079)
	Intersection panel	14	98	100	**0.545**	(0.092)

* See Table 2.6A (page 82).

Meta-imputed genotype data were contrasted with data obtained in imputations from corresponding, larger datasets, which contained the unified sample across the datasets considered in meta-imputation; *i.e.* the intersection and the union of variants present across the other reference datasets, respectively. Although meta-imputation did not perform markedly better in terms of accuracy (measured at the same variant sites), I showed that meta-imputation generally outperformed the intersection panel, in terms of power to detect significant association signals, due to the low coverage retained at the intersection of variants across available reference data. However, note that meta-imputation combined data such that the resulting coverage was identical to the coverage of the union panel; meta-imputed data was overall similar to using the union reference for imputation, with regard to both accuracy and power.

In conclusion, these results suggest that meta-imputation is a viable approach to combine genotype data such that a larger, unified dataset of imputed genotypes is available for association analysis. However, it is unlikely to increase accuracy and power further than possible with imputation from a large, canonical reference; *e.g.* the reference dataset provided by the Haplotype Reference Consortium (HRC). Yet, future GWA studies may benefit from meta-imputation, for example, in situations when researchers have to choose from a collection of available reference datasets, or to increase the coverage of imputed data in general. The meta-imputation algorithm, as presented in this chapter, is available as a computational tool which I implemented in `C++`.*

* Meta-imputation software (`meta-impute`): `https://github.com/pkalbers/meta-impute`

The key test for an acronym is to ask whether it helps or hurts communication.

— Elon Musk

Abbreviations

1000G	1000 Genomes Project
CDF	Cumulative distribution function
GoT2D	Genetics of Type 2 Diabetes Project
GWA	Genome-wide association
HapMap	International HapMap Project
HRC	Haplotype Reference Consortium
HWE	Hardy-Weinberg equilibrium
LD	Linkage disequilibrium
MAF	Minor allele frequency
Mb	Megabase
NGS	Next-generation sequencing
QC	Quality control
SFS	Site frequency spectrum
SNP	Single-nucleotide polymorphism
WGS	Whole-genome sequencing

My definition of a scientist is that you can complete the following sentence: 'he or she has shown that . . . '

— E. O. Wilson

Bibliography

1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., and McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**(7422), 56–65.

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., and Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, **526**(7571), 68–74.

Abecasis, G. R., Cherny, S. S., and Cardon, L. R. (2001). The impact of genotyping error on family-based analysis of quantitative traits. *European Journal of Human Genetics*, **9**(2), 130–134.

Albrechtsen, A., Korneliussen, T. S., Moltke, I., Hansen, T. v. O., Nielsen, F. C., and Nielsen, R. (2009). Relatedness Mapping and Tracts of Relatedness for Genome-Wide Data in the Presence of Linkage Disequilibrium. *Genetic Epidemiology*, **33**(3), 266–274.

Altshuler, D., Daly, M. J., and Lander, E. S. (2008). Genetic Mapping in Human Disease. *Science*, **322**(5903), 881–888.

Altshuler, D. L., Bentley, D. R., Chakravarti, A., Collins, F. S., Donnelly, P., Gabriel, S. B., Knoppers, B. M., Lander, E. S., Lehrach, H., Mardis, E. R., Nickerson, D. A., Peltonen, L., Wang, J., Wilson, R. K., Gibbs, R. A., Deiros, D., Metzker, M., Wheeler, D., Li, J., Jian, M., Li, G., Li, R., Liang, H., Tian, G., Wang, B., Wang, W., Yang, H., Zhang, X., Zheng, H., Ambrogio, L., Bloom, T., Sougnez, C. L., Gormley, N., Humphray, S., Kingsbury, Z., Koko-Gonzales, P., Stone, J., McKernan, K. J., Costa, G. L., Ichikawa, J. K., Lee, C. C., Borodina, T. A., Dahl, A., Davydov, A. N., Marquardt, P., Mertes, F., Nietfeld, W., ROSENSTIEL, P., Schreiber, S., Soldatov, A. V., Timmermann, B., Tolzmann, M., Egholm, M., Affourtit, J., Ashworth, D., Attiya, S., Bachorski, M., Buglione, E., Burke, A., Caprio, A., Celone, C., Clark, S., Conners, D., Gu, L., Guccione, L., Kao, K., Kebbel, A., Knowlton, J., Labrecque, M., McDade, L., Mealmaker, C., Minderman, M., Nawrocki, A., Niazi, F., Pareja, K., Ramenani, R., Riches, D., Song, W., Turcotte, C., Wang, S., Fulton, L., Weinstock, G., Burton, J., Carter, D. M., Churcher, C., Coffey, A., Cox, A., Quail, M., Skelly, T., Stalker, J., Swerdlow, H. P., Turner, D., De Witte, A., Giles, S., Sabo, A., Yu, J., Fang, X., Guo, X., Li, Y., Luo, R., Tai, S., Wu, H., Zheng, H., Zheng, X., Zhou, Y., Wang, J., Huang, W., Indap, A., Kural, D., Lee, W.-P., Stromberg, M. P., Ward, A. N., Lee, C., Mills, R. E., Shi, X., Daly, M. J., DePristo, M. A., Ball, A. D., Browning, B. L., Cibulskis, K., Garimella, K. V., Grossman, S. R., Hanna, M., Jaffe, D. B., Kernytsky, A. M., Li, H., Maguire, J. R., McCarroll, S. A., McKenna, A., Nemesh, J. C., Philippakis, A. A., Poplin, R. E., Price, A., Rivas, M. A., Sabeti, P. C., Schaffner, S. F., Shefler, E., Shlyakhter, I. A., Cooper, D. N., Ball, E. V., Mort, M., Phillips, A. D., Stenson, P. D., Sebat, J., Makarov, V., Ye, K., Yoon, S. C., Clark, A. G., Boyko, A., Degenhardt, J., Gutenkunst, R. N., Kaganovich, M., Keinan, A., Lacroute, P., Ma, X., Reynolds, A., Clarke, L., Cunningham, F., Herrero, J., Keenen, S., Kulesha, E., Leinonen, R., McLaren, W. M., Radhakrishnan, R., Smith, R. E., Zalunin, V., Zheng-Bradley, X., Korbel, J. O., Stütz, A. M., Bauer, M., Keira Cheetham, R., Cox, T., Eberle, M., James, T., Kahn, S., Murray, L., Ye, K., De La Vega, F. M., Fu, Y., Hyland, F. C. L., Manning, J. M., McLaughlin, S. F., Peckham, H. E., Sakarya, O., Sun, Y. A., Tsung, E. F., Batzer, M. A., Konkel, M. K., Walker, J. A., Sudbrak, R., Albrecht, M. W., Amstislavskiy, V. S., Herwig, R., Parkhomchuk, D. V., Agarwala, R., Khouri, H. M., Morgulis, A. O., Paschall, J. E., Phan, L. D., Rotmistrovsky, K. E., Sanders, R. D., Shumway, M. F., Xiao, C., Lunter, G., Marchini, J. L., Moutsianas, L., Myers, S., Tumian, A., Desany, B., Knight, J., Winer, R., Craig, D. W., Beckstrom-Sternberg, S. M., Christoforides, A., Kurdoglu, A. A., Pearson, J. V., Sinari, S. A., Tembe, W. D., Haussler, D., Hinrichs, A. S., Katzman, S. J., Kern, A., Kuhn, R. M., Przeworski, M., Hernandez, R. D., Howie, B., Kelley, J. L., Cord Melton, S., Abecasis, G. R., Li, Y., Anderson, P., Blackwell, T., Chen, W., Cookson, W. O., Ding, J., Min Kang, H., Lathrop, M., Liang, L., and Moffatt, M. F. a. (2010). A map of human genome variation from population-scale sequencing. *Nature*, **467**(7319), 1061–1073.

Bodmer, W. and Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genetics*, **40**(6), 695–701.

Boehnke, M. and Cox, N. J. (1997). Accurate Inference of Relationships in Sib-Pair Linkage Studies. *The American Journal of Human Genetics*, **61**(2), 423–429.

Boomsma, D. I., Wijmenga, C., Slagboom, E. P., Swertz, M. A., Karssen, L. C., Abdellaoui, A., Ye, K., Guryev, V., Vermaat, M., van Dijk, F., Francioli, L. C., Hottenga, J. J., Laros, J. F. J., Li, Q., Li, Y., Cao, H., Chen, R., Du, Y., Li, N., Cao, S., van Setten, J., Menelaou, A., Pulit, S. L., Hehir-Kwa, J. Y., Beekman, M., Elbers, C. C., Byelas, H., de Craen, A. J. M., Deelen, P., Dijkstra, M., den Dunnen, J. T., de Knijff, P., Houwing-Duistermaat, J., Koval, V., Estrada, K., Hofman, A., Kanterakis, A., van Enckevort, D., Mai, H., Kattenberg, M., van Leeuwen, E. M., Neerincx, P. B. T., Oostra, B., Rivadeneira, F., Suchiman, E. H. D., Uitterlinden, A. G., Willemsen, G., Wolffenbuttel, B. H., Wang, J., de Bakker, P. I. W., van Ommen, G.-J., and van Duijn, C. M. (2013). The Genome of the Netherlands: design, and project goals. *European Journal of Human Genetics*, **22**(2), 221–227.

Borodulin, K., Vartiainen, E., Peltonen, M., Jousilahti, P., Juolevi, A., Laatikainen, T., Mannisto, S., Salomaa, V., Sundvall, J., and Puska, P. (2015). Forty-year trends in cardiovascular risk factors in Finland. *European Journal of Public Health*, **25**(3), 539–546.

Botstein, D. and Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics*, **33**(3s), 228–237.

Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American journal of human genetics*, **32**(3), 314–331.

Browning, B. L. and Browning, S. R. (2011). A fast, powerful method for detecting identity by descent. *American journal of human genetics*, **88**(2), 173–182.

Browning, B. L. and Browning, S. R. (2013). Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, **194**(2), 459–471.

Browning, B. L. and Browning, S. R. (2016). Genotype Imputation with Millions of Reference Samples. *American journal of human genetics*, **98**(1), 116–126.

Browning, S. R. (2008). Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes. *Genetics*, **178**(4), 2123–2132.

Browning, S. R. and Browning, B. L. (2010). High-Resolution Detection of Identity by Descent in Unrelated Individuals. *The American Journal of Human Genetics*, **86**(4), 526–539.

Browning, S. R. and Browning, B. L. (2012). Identity by Descent Between Distant Relatives: Detection and Applications. *Annual Review of Genetics*, **46**(1), 617–633.

Buetow, K. H. (1991). Influence of aberrant observations on high-resolution linkage analysis outcomes. *The American Journal of Human Genetics*, **49**(5), 985–994.

Burdett, T., Hall, P., Hasting, E., Hindorff, L., Junkins, H., Klemm, A., MacArthur, J., Manolio, T., Morales, J., Parkinson, H., *et al.* (2016). The nhgri-ebi catalog of published genome-wide association studies. *Available at: www.ebi.ac.uk/gwas. Accessed 2017-01-20, version 1.0.*

Bustamante, C. D., Burchard, E. G., and De La Vega, F. M. (2011). Genomics for the world. *Nature*, **475**(7355), 163–165.

Cai, L., Fisher, A. L., Huang, H., and Xie, Z. (2016). CRISPR-mediated genome editing and human diseases. *Genes & Diseases*, **3**(4), 244–251.

Chakravarti, A. (1999). Population genetics - making sense out of sequence. *Nature Genetics*, **21**, 56–60.

Chen, J., Zhang, J.-G., Li, J., Pei, Y.-F., and Deng, H.-W. (2013). On Combining Reference Data to Improve Imputation Accuracy. *PloS one*, **8**(1).

Choi, Y., Wijsman, E. M., and Weir, B. S. (2009). Case-control association testing in the presence of unknown relationships. *Genetic Epidemiology*, **33**(8), 668–678.

Cirulli, E. T. and Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Publishing Group*, **11**(6), 415–425.

Colombo, R. (2007). Dating mutations. *eLS*.

Correns, K. F. J. (1899). Untersuchungen über die Xenien bei *Zea mays*. *Berichte der Deutschen Botanischen Gesellschaft*, **17**, 410–418.

Cotterman, C. W. (1940). *A calculus for statistico-genetics*. Ph.D. thesis, The Ohio State University.

Coventry, A., Bull-Otterson, L. M., Liu, X., Clark, A. G., Maxwell, T. J., Crosby, J., Hixson, J. E., Rea, T. J., Muzny, D. M., Lewis, L. R., Wheeler, D. A., Sabo, A., Lusk, C., Weiss, K. G., Akbar, H., Cree, A., Hawes, A. C., Newsham, I., Varghese, R. T., Villasana, D., Gross, S., Joshi, V., Santibanez, J., Morgan, M., Chang, K., Hale IV, W., Templeton, A. R., Boerwinkle, E., Gibbs, R., and Sing, C. F. (2010). Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature communications*, **1**(8), 131–6.

Cox, D. G. and Kraft, P. (2006). Quantification of the Power of Hardy-Weinberg Equilibrium Testing to Detect Genotyping Error. *Human heredity*, **61**(1), 10–14.

Crow, J. F. (1954). Breeding structure of populations. ii. effective population number. *Statistics and mathematics in biology*, **543**, 556.

Crow, J. F. and Kimura, M. (1970). An introduction to population genetics theory. *An introduction to population genetics theory.*

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., and Lander, E. S. (2001). High-resolution haplotype structure in the human genome. *Nature Genetics*, **29**(2), 229–232.

de Bakker, P. I. W., Yelensky, R., Pe'er, I., Gabriel, S. B., Daly, M. J., and Altshuler, D. (2005). Efficiency and power in genetic association studies. *Nature Genetics*, **37**(11), 1217–1223.

De Vries, H. M. (1900). Sur la loi de disjonction des hybrides. *Comptes rendus de l'Académie des Sciences*, **130**, 845–847.

Deelen, P., Kanterakis, A., van Dijk, F., Medina-Gomez, C., Kreiner-Møller, E., Rivadeneira, F., Gutierrez-Achury, J., van Enckevort, D., Dijkstra, M., Byelas, H., Genome of Netherlands Consortium, de Bakker, P. I. W., and Swertz, M. A. (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.

Delaneau, O., Coulonges, C., and Zagury, J.-F. (2008). Shape-IT: new rapid and accurate algorithm for haplotype inference. **9**(1), 540.

Delaneau, O., Marchini, J., and Zagury, J.-F. (2011). A linear complexity phasing method for thousands of genomes. *Nature methods*, **9**(2), 179–181.

Delaneau, O., Zagury, J.-F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nature methods*, **10**(1), 5–6.

Devlin, B., Roeder, K., and Wasserman, L. (2001). Genomic control, a new approach to genetic-based association studies. *Theoretical population biology*, **60**(3), 155–166.

Donnelly, K. P. (1983). The probability that related individuals share some section of genome identical by descent. *Theoretical population biology*, **23**(1), 34–63.

Douglas, J. A., Boehnke, M., and Lange, K. (2000). A Multipoint Method for Detecting Genotyping Errors and Mutations in Sibling-Pair Linkage Data. *The American Journal of Human Genetics*, **66**(4), 1287–1297.

Douglas, J. A., Skol, A. D., and Boehnke, M. (2002). Probability of Detection of Genotyping Errors and Mutations as Inheritance Inconsistencies in Nuclear-Family Data. *The American Journal of Human Genetics*, **70**(2), 487–495.

Eberle, M. A., Fritzilas, E., Krusche, P., Kallberg, M., Moore, B. L., Bekritsky, M. A., Iqbal, Z., Chuang, H.-Y., Humphray, S. J., Halpern, A. L., Kruglyak, S., Margulies, E. H., McVean, G., and Bentley, D. R. (2016). A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Research*, **27**(1), 1–9.

Ewens, W. J. (2012a). *Mathematical Population Genetics 1: Theoretical Introduction*, volume 27. Springer Science & Business Media.

Ewens, W. J. (2012b). *Mathematical population genetics 1: theoretical introduction*, volume 27. Springer Science & Business Media.

Fisher, R. (1930a). The genetical theory of natural selection.

Fisher, R. A. (1930b). *The genetical theory of natural selection*. Oxford University Press, Oxford.

Fisher, R. A. (1949). The theory of inbreeding. *The theory of inbreeding*.

Fisher, R. A. (1954). A fuller theory of "junctions" in inbreeding. *Heredity*, **8**(2), 187–197.

Forney, G. D. (1973). The Viterbi Algorithm. In *Proceedings of the IEEE*, pages 268–278.

Frazer, K. A., Murray, S. S., Schork, N. J., and Topol, E. J. (2009). Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, **10**(4), 241–251.

Freedman, M. L., Reich, D., Penney, K. L., McDonald, G. J., Mignault, A. A., Patterson, N., Gabriel, S. B., Topol, E. J., Smoller, J. W., Pato, C. N., Pato, M. T., Petryshen, T. L., Kolonel, L. N., Lander, E. S., Sklar, P., Henderson, B., Hirschhorn, J. N., and Altshuler, D. (2004). Assessing the impact of population stratification on genetic association studies. *Nature Genetics*, **36**(4), 388–393.

Fu, W., O'Connor, T. D., Jun, G., Kang, H. M., Abecasis, G., Leal, S. M., Gabriel, S., Altshuler, D., Shendure, J., Nickerson, D. A., Bamshad, M. J., Project, N. E. S., and Akey, J. M. (2012). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, **493**(7431), 216–220.

Fu, Y. X. (1995). Statistical Properties of Segregating Sites. *Theoretical population biology*, **48**(2), 172–197.

Fu, Y.-X. and Li, W.-H. (1999). Coalescing into the 21st Century: An Overview and Prospects of Coalescent Theory. *Theoretical population biology*, **56**(1), 1–10.

Fuchsberger, C., Flannick, J., Teslovich, T. M., Mahajan, A., Agarwala, V., Gaulton, K. J., Ma, C., Fontanillas, P., Moutsianas, L., McCarthy, D. J., Rivas, M. A., Perry, J. R. B., Sim, X., Blackwell, T. W., Robertson, N. R., Rayner, N. W., Cingolani, P., Locke, A. E., Tajes, J. F., Highland, H. M., Dupuis, J., Chines, P. S., Lindgren, C. M., Hartl, C., Jackson, A. U., Chen, H., Huyghe, J. R., van de Bunt, M., Pearson, R. D., Kumar, A., Müller-Nurasyid, M., Grarup, N., Stringham, H. M., Gamazon, E. R., Lee, J., Chen, Y., Scott, R. A., Below, J. E., Chen, P., Huang, J., Go, M. J., Stitzel, M. L., Pasko, D., Parker, S. C. J., Varga, T. V., Green, T., Beer, N. L., Day-Williams, A. G., Ferreira, T., Fingerlin, T., Horikoshi, M., Hu, C., Huh, I., Ikram, M. K., Kim, B.-J., Kim, Y., Kim, Y. J., Kwon, M.-S., Lee, J., Lee, S., Lin, K.-H., Maxwell, T. J., Nagai, Y., Wang, X., Welch, R. P., Yoon, J., Zhang, W., Barzilai, N., Voight, B. F., Han, B.-G., Jenkinson, C. P., Kuulasmaa, T., Kuusisto, J., Manning, A., Ng, M. C. Y., Palmer, N. D., Balkau, B., áková, A. S., Abboud, H. E., Boeing, H., Giedraitis, V., Prabhakaran, D., Gottesman, O., Scott, J., Carey, J., Kwan, P., Grant, G., Smith, J. D., Neale, B. M., Purcell, S., Butterworth, A. S., Howson, J. M. M., Lee, H. M., Lu, Y., Kwak, S.-H., Zhao, W., Danesh, J., Lam, V. K. L., Park, K. S., Saleheen, D., So, W. Y., Tam, C. H. T., Afzal, U., Aguilar, D., Arya, R., Aung, T., Chan, E., Navarro, C., Cheng, C.-Y., Palli, D., Correa, A., Curran, J. E., Rybin, D., Farook, V. S., Fowler, S. P., Freedman, B. I., Griswold, M., Hale, D. E., Hicks, P. J., Khor, C.-C., Kumar, S., Lehne, B., Thuillier, D., Lim, W. Y., Liu, J., van der Schouw, Y. T., Loh, M., Musani, S. K., Puppala, S., Scott, W. R., Yengo, L., Tan, S.-T., Taylor, H. A., Thameem, F., Wilson, G., Wong, T. Y., Njølstad, P. R., Levy, J. C., Mangino, M., Bonnycastle, L. L., Schwarzmayr, T., Fadista, J., Surdulescu, G. L., Herder, C., Groves, C. J., Wieland, T., Bork-Jensen, J., Brandslund, I., Christensen, C., Koistinen, H. A., Doney, A. S. F., Kinnunen, L., Esko, T., Farmer, A. J., Hakaste, L., Hodgkiss, D., Kravic, J., Lyssenko, V., Hollensted, M., Jørgensen, M. E., Jørgensen, T., Ladenvall, C., Justesen, J. M., Käräjämäki, A., Kriebel, J., Rathmann, W., Lannfelt, L., Lauritzen, T., Narisu, N., Linneberg, A., Melander, O., Milani, L., Neville, M., Orho-Melander, M., Qi, L., Qi, Q., Roden, M., Rolandsson, O., Swift, A., Rosengren, A. H., Stirrups, K., Wood, A. R., Mihailov, E., Blancher, C., Carneiro, M. O., Maguire, J., Poplin, R., Shakir, K., Fennell, T., DePristo, M., de Angelis, M. H., Deloukas, P., Gjesing, A. P., Jun, G., Nilsson, P., Murphy, J., Onofrio, R., Thorand, B., Hansen, T., Meisinger, C., Hu, F. B., Isomaa, B., Karpe, F., Liang, L., Peters, A., Huth, C., O'Rahilly, S. P., Palmer, C. N. A., Pedersen, O., Rauramaa, R., Tuomilehto, J., Salomaa, V., Watanabe, R. M., Syvänen, A.-C., Bergman, R. N., Bharadwaj, D., Bottinger, E. P., Cho, Y. S., Chandak, G. R., Chan, J. C. N., Chia, K. S., Daly, M. J., Ebrahim, S. B., Langenberg, C., Elliott, P., Jablonski, K. A., Lehman, D. M., Jia, W., and Ma, R. (2016). The genetic architecture of type 2 diabetes. *Nature*, **536**(7614), 41–47.

Genome of the Netherlands Consortium (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics*, **46**(8), 818–825.

Gibson, G. (2012). Rare and common variants: twenty arguments. *Nature Reviews Genetics*, **13**(2), 135–145.

Gordon, D., Heath, S. C., Liu, X., and Ott, J. (2001). A Transmission/Disequilibrium Test That Allows for Genotyping Errors in the Analysis of Single-Nucleotide Polymorphism Data. *The American Journal of Human Genetics*, **69**(2), 371–380.

Gordon, D., Finch, S. J., Nothnagel, M., and Ott, J. u. r. (2002). Power and Sample Size Calculations for Case-Control Genetic Association Tests when Errors Are Present: Application to Single Nucleotide Polymorphisms. *Human heredity*, **54**(1), 22–33.

Gore, A., Li, Z., Fung, H.-L., Young, J. E., Agarwal, S., Antosiewicz-Bourget, J., Canto, I., Giorgetti, A., Israel, M. A., Kiskinis, E., Lee, J.-H., Loh, Y.-H., Manos, P. D., Montserrat, N., Panopoulos, A. D., Ruiz, S., Wilbert, M. L., Yu, J., Kirkness, E. F., Izpisua Belmonte, J. C., Rossi, D. J., Thomson, J. A., Eggan, K., Daley, G. Q., Goldstein, L. S. B., and Zhang, K. (2011). Somatic coding mutations in human induced pluripotent stem cells. *Nature*, **471**(7336), 63–67.

Griffiths, R. C. (1991). The Two-Locus Ancestral Graph. In *Selected Proceedings of the Sheffield Symposium on Applied Probability*, pages 100–117. Institute of Mathematical Statistics, Hayward, CA.

Griffiths, R. C. and Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology*, **3**(4), 479–502.

Griffiths, R. C. and Marjoram, P. (1997a). An ancestral recombination graph. *Institute for Mathematics and its Applications*, **87**, 257.

Griffiths, R. C. and Marjoram, P. (1997b). Progress in population genetics and human evolution.

Griffiths, R. C. and Tavaré, S. (1998). The age of a mutation in a general coalescent tree. *Communications in Statistics. Stochastic Models*, **14**(1-2), 273–295.

Gusev, A., Lowe, J. K., Stoffel, M., Daly, M. J., Altshuler, D., Breslow, J. L., Friedman, J. M., and Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. *Genome Research*, **19**(2), 318–326.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009). Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genetics*, **5**(10), e1000695–11.

Haldane, J. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *J Genet*, **8**(29), 299–309.

Hardy, G. H. (1908). Mendelian Proportions in a Mixed Population. *Science*, **28**(706), 49–50.

Harris, K. and Nielsen, R. (2013). Inferring Demographic History from a Spectrum of Shared Haplotype Lengths. *PLoS Genetics*, **9**(6).

Hayes, B. J., Visscher, P. M., McPartlan, H. C., and Goddard, M. E. (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Research*, **13**(4), 635–643.

Hein, J., Schierup, M., and Wiuf, C. (2004). *Gene genealogies, variation and evolution: a primer in coalescent theory*. Oxford University Press, Oxford.

Henn, B. M., Gutenkunst, R. N., Indap, A. R., Marth, G. T., Clark, A. G., Yu, F., Gibbs, R. A., 1000 Genomes Project, Durbin, R. M., Flicek, P., Gabriel, S. B., Lander, E. S., Wheeler, D., Cibulskis, K., Fennell, T. J., Jaffe, D. B., Shefler, E., Egholm, M., Fulton, R., Bainbridge, M., Challis, D., Sabo, A., Wang, J., Lee, C., Korn, J. M., Sudbrak, R., Auton, A., Iqbal, Z., Desany, B., Dooling, D., Hurles, M. E., MacArthur, D. G., Abyzov, A., Zhang, Z., Garrison, E. P., Banks, E., Handsaker, R. E., Hartl, C., De La Vega, F. M., Alkan, C., Snyder, M., Muzny, D., Reid, J., Quinlan, A. R., Stewart, C., Wu, J., Gravel, S., Sherry, S. T., McVean, G. A., Abecasis, G. R., Koboldt, D. C., Palotie, A., Bustamante, C. D., Schafer, A. J., and Brooks, L. D. (2011). Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences*, **108**(29), 11983–11988.

Howie, B., Marchini, J., and Stephens, M. (2011a). Genotype imputation with thousands of genomes. *G3 (Bethesda, Md.)*, **1**(6), 457–470.

Howie, B., Marchini, J., and Stephens, M. (2011b). Genotype Imputation with Thousands of Genomes. *G3 (Bethesda, Md.)*, **1**(6), 457–470.

Howie, B. N., Donnelly, P., and Marchini, J. (2009). A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genetics*, **5**(6), e1000529.

Hudson, R. (1990). Gene genealogies and the coalescent process. *Oxford surveys in evolutionary biology*, **7**(1), 44.

Hudson, R. and Kaplan, N. L. (1985). Statistical Properties of the Number of Recombination Events in the History of a Sample of Dna-Sequences. *Genetics*, **111**(1), 147–164.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical population biology*, **23**(2), 183–201.

Hudson, R. R. (2001). Two-locus sampling distributions and their application. *Genetics*, **159**(4), 1805–1817.

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**(2), 337–338.

International HapMap 3 Consortium, Schaffner, S. F., Peltonen, L., Dermitzakis, E., Bonnen, P. E., Altshuler, D. M., de Bakker, P. I. W., Deloukas, P., Gabriel, S. B., Gwilliam, R., Hunt, S., Inouye, M., Jia, X., Palotie, A., Parkin, M., Whittaker, P., Yu, F., Chang, K., Hawes, A., Lewis, L. R., Ren, Y., Wheeler, D., Gibbs, R. A., Muzny, D. M., Barnes, C., Darvishi, K., Hurles, M., Korn, J. M., Kristiansson, K., Lee, C., McCarrol, S. A., Nemesh, J., Keinan, A., Montgomery, S. B., Pollack, S., Soranzo, N., Gonzaga-Jauregui, C., Anttila, V., Brodeur, W., Daly, M. J., Leslie, S., McVean, G., Moutsianas, L., Nguyen, H., Zhang, Q., Ghori, M. J. R., McGinnis, R., McLaren, W., Price, A. L., Takeuchi, F., Grossman, S. R., Shlyakhter, I., Hostetter, E. B., Sabeti, P. C., Adebamowo, C. A., Foster, M. W., Gordon, D. R., Licinio, J., Manca, M. C., Marshall, P. A., Matsuda, I., Ngare, D., Wang, V. O., Reddy, D., Rotimi, C. N., Royal, C. D., Sharp, R. R., Zeng, C., Brooks, L. D., and McEwen, J. E. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**(7311), 52–58.

International HapMap Consortium (2003). The International HapMap Project. *Nature*, **426**(6968), 789–796.

International HapMap Consortium, Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, H., Zhao, H., Zhou, J., Gabriel, S. B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Shen, Y., Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Shen, Y., Sun, W., Wang, H., Wang, Y., Wang, Y., Xiong, X., Xu, L., Waye, M. M. Y., Tsui, S. K. W., Xue, H., Wong, J. T.-F., Galver, L. M., Fan, J.-B., Gunderson, K., Murray, S. S., Oliphant, A. R., Chee, M. S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J.-F., Phillips, M. S., Roumy, S., Sallée, C., Verner, A., Hudson, T. J., Kwok, P.-Y., Cai, D., Koboldt, D. C., Miller, R. D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L.-C., Mak, W., Song, Y. Q., Tam, P. K. H., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., Ohnishi, Y., Sekine, A., Tanaka, T., Tsunoda, T., Deloukas, P., Bird, C. P., Delgado, M., Dermitzakis, E. T., Gwilliam, R., Hunt, S., Morrison, J., Powell, D., Stranger, B. E., Whittaker, P., Bentley, D. R., de Bakker, P. I. W., Barrett, J., Chretien, Y. R., Maller, J., McCarroll, S., Patterson, N., Pe'er, I., Price, A., Purcell, S., Sabeti, P., Saxena, R., Schaffner, S. F., Sham, P. C., Varilly, P., Altshuler, D., Stein, L. D., Krishnan, L., Smith, A. V., Tello-Ruiz, M. K., Thorisson, G. A., Chakravarti, A., Chen, P. E., Cutler, D. J., Kashuk, C. S., Lin, S., Abecasis, G. R., Guan, W., Li, Y., Munro, H. M., Qin, Z. S., Thomas, D. J., McVean, G., Auton, A., Bottolo, L., Cardin, N., Eyheramendy, S., Freeman, C., Marchini, J., Myers, S., Spencer, C., Stephens, M., Donnelly, P., Cardon, L. R., Clarke, G., Evans, D. M., Morris, A. P., Weir, B. S., Mullikin, J. C., Sherry, S. T., Feolo, M., Skol, A., Zhang, H., Matsuda, I., Fukushima, Y., Macer, D. R., Suda, E., Rotimi, C. N., Adebamowo, C. A., Ajayi, I., Aniagwu, T., Marshall, P. A., Nkwodimmah, C., Royal, C. D. M., Leppert, M. F., Dixon, M., Peiffer, A., Qiu, R., Kent, A., Kato, K., Niikawa, N., Adewole, I. F., Knoppers, B. M., Foster, M. W., Clayton, E. W., Watkin, J., Muzny, D., Nazareth, L., Sodergren, E., Weinstock, G. M., Yakub, I., Onofrio, R. C., Richter, D. J., Birren, B. W., Daly, M. J., Wilson, R. K., Fulton, L. L., Rogers, J., Burton, J., Carter, N. P., Clee, C. M., Griffiths, M., Jones, M. C., McLay, K., Plumb, R. W., Ross, M. T., Sims, S. K., Willey, D. L., Chen, Z., Han, H., Kang, L., Godbout, M., Wallenburg, J. C., L'Archevêque, P., Bellemare, G., Saeki, K., Wang, H., An, D., Fu, H., Li, Q., Wang, Z., Wang, R., Holden, A. L., Brooks, L. D., McEwen, J. E., Guyer, M. S., Wang, V. O., Peterson, J. L., Shi, M., Spiegel, J., Sung, L. M., Zacharia, L. F., Collins, F. S., Kennedy, K., Jamieson, R., and Stewart, J. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**(7164), 851–861.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), 860–921.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature*, **431**(7011), 931–945.

Kaiser, J. (2008). A plan to capture human diversity in 1000 genomes (Science (395)). *Science*, **319**(5868), 1336.

Kang, S. J., Gordon, D., and Finch, S. J. (2004). What SNP genotyping errors are most costly for genetic association studies? *Genetic Epidemiology*, **26**(2), 132–141.

Keinan, A. and Clark, A. G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, **336**(6082), 740–743.

Kelleher, J., Etheridge, A. M., and McVean, G. (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS computational biology*, **12**(5), e1004842–22.

Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, **61**(4), 893–903.

Kimura, M. and Ota, T. (1973). The age of a neutral mutant persisting in a finite population. *Genetics*, **75**(1), 199–212.

Kingman, J. F. C. (1982a). Exchangeability and the evolution of large populations. *Exchangeability in Probability and Statistics*, pages 97—-112.

Kingman, J. F. C. (1982b). On the genealogy of large populations. *Journal of Applied Probability*, **19**(A), 27–43.

Kingman, J. F. C. (1982c). The coalescent. *Stochastic processes and their applications*, **13**(3), 235–248.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., and Hoh, J. (2005). Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science*, **308**(5720), 385–389.

Kong, A., Masson, G., Frigge, M. L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P. I., Ingason, A., Steinberg, S., Rafnar, T., Sulem, P., Mouy, M., Jonsson, F., Thorsteinsdottir, U., Gudbjartsson, D. F., Stefansson, H., and Stefansson, K. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genetics*, **40**(9), 1068–1075.

Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics*, **22**(2), 139–144.

Kryukov, G. V., Pennacchio, L. A., and Sunyaev, S. R. (2007). Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *The American Journal of Human Genetics*, **80**(4), 727–739.

Labuda, M., Labuda, D., Korab-Laskowska, M., Cole, D. E., Zietkiewicz, E., Weissenbach, J., Popowska, E., Pronicka, E., Root, A. W., and Glorieux, F. H. (1996). Linkage disequilibrium analysis in young populations: pseudo-vitamin d-deficiency rickets and the founder effect in french canadians. *American journal of human genetics*, **59**(3), 633.

Lander, E. S. (1996). The new genomics: Global views of biology. *Science*, **274**(5287), 536–539.

Lawson, D. J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genetics*, **8**(1), e1002453.

Leutenegger, A.-L., Prum, B., Génin, E., Verny, C., Lemainque, A., Clerget-Darpoux, F., and Thompson, E. A. (2003). Estimation of the Inbreeding Coefficient through Use of Genomic Data. *American journal of human genetics*, **73**(3), 516–523.

Li, N. and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, **165**(4), 2213–2233.

Li, W. H. (1975). The first arrival time and mean age of a deleterious mutant gene in a finite population. *The American Journal of Human Genetics*, **27**(3), 274–286.

Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype imputation. *Annual review of genomics and human genetics*, **10**, 387–406.

Li, Y., Willer, C. J., Ding, J., Scheet, P., and Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, **34**(8), 816–834.

Lim, E. T., Würtz, P., Havulinna, A. S., Palta, P., Tukiainen, T., Rehnström, K., Esko, T., Mägi, R., Inouye, M., Lappalainen, T., Chan, Y., Salem, R. M., Lek, M., Flannick, J., Sim, X., Manning, A., Ladenvall, C., Bumpstead, S., Hämäläinen, E., Aalto, K., Maksimow, M., Salmi, M., Blankenberg, S., Ardissino, D., Shah, S., Horne, B., McPherson, R., Hovingh, G. K., Reilly, M. P., Watkins, H., Goel, A., Farrall, M., Girelli, D., Reiner, A. P., Stitziel, N. O., Kathiresan, S., Gabriel, S., Barrett, J. C., Lehtimäki, T., Laakso, M., Groop, L., Kaprio, J., Perola, M., McCarthy, M. I., Boehnke, M., Altshuler, D. M., Lindgren, C. M., Hirschhorn, J. N., Metspalu, A., Freimer, N. B., Zeller, T., Jalkanen, S., Koskinen, S., Raitakari, O., Durbin, R., MacArthur, D. G., Salomaa, V., Ripatti, S., Daly, M. J., Palotie, A., and for the Sequencing Initiative Suomi (SISu) Project (2014). Distribution and Medical Impact of Loss-of-Function Variants in the Finnish Founder Population. *PLoS Genetics*, **10**(7), e1004494–12.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of Next-Generation Sequencing Systems. *Journal of biomedicine & biotechnology*, **2012**(7), 1–11.

Loh, P.-R., Palamara, P. F., and Price, A. L. (2016a). Fast and accurate long-range phasing in a uk biobank cohort. *Nature genetics*.

Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., Reshef, Y. A., Finucane, H. K., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G. R., Durbin, R., and Price, A. L. (2016b). Reference-based phasing using the Haplotype Reference Consortium panel. Technical report.

Malécot, G. (1948). Mathematics of heredity. *Les mathematiques de l'heredite.*

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F. C., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, **461**(7265), 747–753.

Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. **11**(7), 499–511.

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, **39**(7), 906–913.

Mardis, E. R. (2006). Anticipating the 1,000 dollar genome. *Genome biology*, **7**(7), 112.

Mardis, E. R. (2017). DNA sequencing technologies: 2006–2016. *Nature Protocols*, **12**(2), 213–218.

Marth, G. T., Yu, F., Indap, A. R., Garimella, K., Gravel, S., Leong, W. F., Tyler-Smith, C., Bainbridge, M., Blackwell, T., Zheng-Bradley, X., Chen, Y., Challis, D., Clarke, L., Ball, E. V., Cibulskis, K., Cooper, D. N., Fulton, B., Hartl, C., Koboldt, D., Muzny, D., Smith, R., Sougnez, C., Stewart, C., Ward, A., Yu, J., Xue, Y., Altshuler, D., Bustamante, C. D., Clark, A. G., Daly, M., DePristo, M., Flicek, P., Gabriel, S., Mardis, E., Palotie, A., Gibbs, R., and 1000 Genomes Project (2011). The functional spectrum of low-frequency coding variation. *Genome biology*, **12**(9), R84.

Maruyama, T. (1974). The age of an allele in a finite population. *Genetical research*, **23**(2), 137–143.

Mathieson, I. and McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. *Nature Genetics*, **44**(3), 243–U29.

Mathieson, I. and McVean, G. (2014). Demography and the Age of Rare Variants. *PLoS Genetics*, **10**(8), e1004528.

Maynard Smith, J. (1989). *Evolutionary genetics.* Oxford University Press.

McCarroll, S. A., Kuruvilla, F. G., Korn, J. M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M. H., de Bakker, P. I. W., Maller, J. B., Kirby, A., Elliott, A. L., Parkin, M., Hubbell, E., Webster, T., Mei, R., Veitch, J., Collins, P. J., Handsaker, R., Lincoln, S., Nizzari, M., Blume, J., Jones, K. W., Rava, R., Daly, M. J., Gabriel, S. B., and Altshuler, D. (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics*, **40**(10), 1166–1174.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Publishing Group*, **9**(5), 356–369.

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., Kang, H. M., Fuchsberger, C., Danecek, P., Sharp, K., Luo, Y., Sidore, C., Kwong, A., Timpson, N., Koskinen, S., Vrieze, S., Scott, L. J., Zhang, H., Mahajan, A., Veldink, J., Peters, U., Pato, C., van Duijn, C. M., Gillies, C. E., Gandin, I., Mezzavilla, M., Gilly, A., Cocca, M., Traglia, M., Angius, A., Barrett, J. C., Boomsma, D., Branham, K., Breen, G., Brummett, C. M., Busonero, F., Campbell, H., Chan, A., Chen, S., Chew, E., Collins, F. S., Corbin, L. J., Smith, G. D., Dedoussis, G., Dorr, M., Farmaki, A.-E., Ferrucci, L., Forer, L., Fraser, R. M., Gabriel, S., Levy, S., Groop, L., Harrison, T., Hattersley, A., Holmen, O. L., Hveem, K., Kretzler, M., Lee, J. C., McGue, M., Meitinger, T., Melzer, D., Min, J. L., Mohlke, K. L., Vincent, J. B., Nauck, M., Nickerson, D., Palotie, A., Pato, M., Pirastu, N., McInnis, M., Richards, J. B., Sala, C., Salomaa, V., Schlessinger, D., Schoenherr, S., Slagboom, P. E., Small, K., Spector, T., Stambolian, D., Tuke, M., Tuomilehto, J., Van den Berg, L. H., Van Rheenen, W., Volker, U., Wijmenga, C., Toniolo, D., Zeggini, E., Gasparini, P., Sampson, M. G., Wilson, J. F., Frayling, T., de Bakker, P. I. W., Swertz, M. A., McCarroll, S., Kooperberg, C., Dekker, A., Altshuler, D., Willer, C., Iacono, W., Ripatti, S., Soranzo, N., Walter, K., Swaroop, A., Cucca, F., Anderson, C. A., Myers, R. M., Boehnke, M., McCarthy, M. I., Durbin, R., Abecasis, G., and Marchini, J. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, **48**(10), 1279–1283.

McClellan, J. and King, M.-C. (2010). Genetic Heterogeneity in Human Disease. *Cell*, **141**(2), 210–217.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome biology*, pages 1–14.

McQuillan, R., Leutenegger, A.-L., Abdel-Rahman, R., Franklin, C. S., Pericic, M., Barac-Lauc, L., Smolej-Narancic, N., Janicijevic, B., Polasek, O., Tenesa, A., MacLeod, A. K., Farrington, S. M., Rudan, P., Hayward, C., Vitart, V., Rudan, I., Wild, S. H., Dunlop, M. G., Wright, A. F., Campbell, H., and Wilson, J. F. (2008). Runs of Homozygosity in European Populations. *The American Journal of Human Genetics*, **83**(3), 359–372.

McVean, G., Awadalla, P., and Fearnhead, P. (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*.

McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science*, **304**(5670), 581–584.

Mendel, G. (1866). Versuche über Pflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brünn*, **4**, 3–47.

Metzker, M. L. (2009). Sequencing technologies — the next generation. *Nature Publishing Group*, **11**(1), 31–46.

Milligan, B. G. (2003). Maximum-Likelihood Estimation of Relatedness. *Genetics*, **163**(3), 1153–1167.

Moore, G. E. (1965). Cramming more components onto integrated circuits. *Electronics*, **38**(8).

Morgan, T. H. (1911). Random segregation versus coupling in Mendelian inheritance. *Science*, **34**(873), 384–384.

Morral, N., Bertranpetit, J., Estivill, X., and Nunes, V. (1994). The origin of the major cystic fibrosis mutation (DF508) in European populations. *Nature*.

Morris, A. and Cardon, L. (2007). Whole genome association. *Handbook of Statistical Genetics, Third Edition*, pages 1238–1263.

Moskvina, V. and Schmidt, K. M. (2006). Susceptibility of Biallelic Haplotype and Genotype Frequencies to Genotyping Error. *Biometrics*, **62**(4), 1116–1123.

Moskvina, V., Craddock, N., Holmans, P., Owen, M., and O'Donovan, M. (2005). Minor genotyping error can result in substantial elevation in type i error rate in haplotype based case control analysis. In *American Journal of Medical Genetics Part B-Neuropsychiatric Genetics*, volume 138, pages 19–19.

Moutsianas, L., Agarwala, V., Fuchsberger, C., Flannick, J., Rivas, M. A., Gaulton, K. J., Albers, P. K., GoT2D Consortium, McVean, G., Boehnke, M., Altshuler, D., and McCarthy, M. I. (2015). The Power of Gene-Based Rare Variant Methods to Detect Disease-Associated Variation and Test Hypotheses About Complex Disease. *PLoS Genetics*, **11**(4), e1005165–24.

Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science*.

Naidoo, N., Pawitan, Y., Soong, R., Cooper, D. N., and Ku, C.-S. (2011). Human genetics and genomics a decade after the release of the draft sequence of the human genome. *Human Genomics*, **5**(6), 577–622.

Neuhauser, C. (2001). Mathematical models in population genetics. *Handbook of statistical genetics*.

Nordborg, M. (2001). Coalescent theory. *Handbook of statistical genetics*.

O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., Traglia, M., Huang, J., Huffman, J. E., Rudan, I., McQuillan, R., Fraser, R. M., Campbell, H., Polasek, O., Asiki, G., Ekoru, K., Hayward, C., Wright, A. F., Vitart, V., Navarro, P., Zagury, J.-F., Wilson, J. F., Toniolo, D., Gasparini, P., Soranzo, N., Sandhu, M. S., and Marchini, J. (2014). A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PLoS Genetics*, **10**(4), e1004234–21.

Ott, J. (1999). *Analysis of human genetic linkage*. JHU Press.

Pajunen, P., Rissanen, H., Härkänen, T., Jula, A., Reunanen, A., and Salomaa, V. (2010). The metabolic syndrome as a predictor of incident diabetes and cardiovascular events in the Health 2000 Study. *Diabetes and Metabolism*, **36**(5), 395–401.

Palamara, P. F. and Pe'er, I. (2013). Inference of historical migration rates via haplotype sharing. *Bioinformatics*, **29**(13), i180–i188.

Palamara, P. F., Lencz, T., Darvasi, A., and Pe'er, I. (2012). Length Distributions of Identity by Descent Reveal Fine-Scale Demographic History. *The American Journal of Human Genetics*, **91**(5), 809–822.

Palin, K., Campbell, H., Wright, A. F., Wilson, J. F., and Durbin, R. (2011). Identity-by-descent-based phasing and imputation in founder populations using graphical models. *Genetic Epidemiology*, **35**(8), 853–860.

Pasaniuc, B., Zaitlen, N., Shi, H., Bhatia, G., Gusev, A., Pickrell, J., Hirschhorn, J., Strachan, D. P., Patterson, N., and Price, A. L. (2014). Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics (Oxford, England)*, **30**(20), 2906–2914.

Pe'er, I., de Bakker, P. I. W., Maller, J., Yelensky, R., Altshuler, D., and Daly, M. J. (2006). Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nature Genetics*, **38**(6), 663–667.

Pennisi, E. (2007). Human Genetic Variation. *Science*, **318**(5858), 1842–1843.

Powell, J. E., Visscher, P. M., and Goddard, M. E. (2010). Reconciling the analysis of IBD and IBS in complex trait studies. *Nature Publishing Group*, **11**(11), 800–805.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, **38**(8), 904–909.

Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *The American Journal of Human Genetics*, **69**(1), 124–137.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., and Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, **81**(3), 559–575.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286. Alcatel-Lucent Bell Labs, Murray, United States.

Rasmussen, M. D., Hubisz, M. J., Gronau, I., and Siepel, A. (2014). Genome-Wide Inference of Ancestral Recombination Graphs. *PLoS Genetics*, **10**(5), e1004342–27.

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., and Lander, E. S. (2001). Linkage disequilibrium in the human genome. *Nature*, **411**(6834), 199–204.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, **273**(5281), 1516–1517.

Risch, N., de Leon, D., Ozelius, L., Kramer, P., Almasy, L., Singer, B., Fahn, S., Breakefield, X., and Bressman, S. (1995). Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. *Nature Genetics*, **9**(2), 152–159.

Risch, N. J. (2000). Searching for genetic determinants in the new millennium. *Nature*, **405**(6788), 847–856.

Roach, J. C., Glusman, G., Smit, A. F. A., Huff, C. D., Hubley, R., Shannon, P. T., Rowen, L., Pant, K. P., Goodman, N., Bamshad, M., Shendure, J., Drmanac, R., Jorde, L. B., Hood, L., and Galas, D. J. (2010). Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. *Science*, **328**(5978), 636–639.

Roshyara, N. R. and Scholz, M. (2015). Impact of genetic similarity on imputation accuracy. *BMC Genetics*, **16**(1), 1–16.

Rousset, F. (2002). Inbreeding and relatedness coefficients: What do they measure? *Heredity*, **88**(5), 371–380.

Sanger, F. and Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, **94**(3), 441–448.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74**(12), 5463–5467.

Scally, A. and Durbin, R. (2012). Revising the human mutation rate: implications for understanding human evolution. *Nature Reviews Genetics*, **13**(10), 745–753.

Schork, N. J., Murray, S. S., Frazer, K. A., and Topol, E. J. (2009). Common vs. rare allele hypotheses for complex diseases. *Current opinion in genetics & development*, **19**(3), 212–219.

Schroff, M. H. (2016). *Genealogical properties of rare variation and their implications for demographic inference*. Ph.D. thesis, University of Oxford.

Serre, J. L., Simon-Bouy, B., Mornet, E., Jaume-Roig, B., Balassopoulou, A., Schwartz, M., Taillandier, A., Boue, J., and Boue, A. (1990). Studies of RFLP closely linked to the cystic fibrosis locus throughout Europe lead to new considerations in populations genetics. *Human genetics*, **84**(5), 449–454.

Shields, D. C., Collins, A., Buetow, K. H., and Morton, N. E. (1991). Error filtration, interference, and the human linkage map. *Proceedings of the National Academy of Sciences of the United States of America*, **88**(15), 6501–6505.

Simons, Y. B., Turchin, M. C., Pritchard, J. K., and Sella, G. (2014). The deleterious mutation load is insensitive to recent population history. *Nature Publishing Group*, **46**(3), 220–224.

Slack, J. (2014). *Genes: A Very Short Introduction*. Oxford University Press, Oxford.

Slatkin, M. (2000). Allele age and a test for selection on rare alleles. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **355**(1403), 1663–1668.

Slatkin, M. (2008a). Inbreeding coefficients and coalescence times. *doi.org*, pages 1–9.

Slatkin, M. (2008b). Linkage disequilibrium – understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, **9**(6), 477–485.

Slatkin, M. and Bertorelle, G. (2001). The use of intraallelic variability for testing neutrality and estimating population growth rate. *Genetics*, **158**(2), 865–874.

Slatkin, M. and Rannala, B. (2000). Estimating allele age. *Annual review of genomics and human genetics*, **1**(1), 225–249.

Sobel, E., Papp, J. C., and Lange, K. (2002). Detection and integration of genotyping errors in statistical genetics. *The American Journal of Human Genetics*, **70**(2), 496–508.

Stam, P. (1980). The distribution of the fraction of the genome identical by descent in finite random mating populations. *Genetical research*, **35**(02), 131.

Stone, M. (1961). The Opinion Pool. *The Annals of Mathematical Statistics*, **32**(4), 1339–1342.

Sturtevant, A. H. (1913). The linear arrangement of six sex-linked factors in drosophila, as shown by their mode of association. *Journal of Experimental Zoology Part A: Ecological Genetics and Physiology*, **14**(1), 43–59.

Su, Z., Marchini, J., and Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics*, **27**(16), 2304–2305.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**(2), 437–460.

Tajima, F. (1993). Measurement of dna polymorphism. *Mechanisms of molecular evolution*, pages 37–59.

Takahata, N. (1993). Allelic genealogy and human evolution. *Molecular Biology and Evolution*, **10**(1), 2–22.

Tavaré, S. (2004). Part I: Ancestral inference in population genetics. In *Lectures on probability theory and statistics*, pages 1–188. Springer.

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, **145**(2), 505–518.

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., Kang, H. M., Jordan, D., Leal, S. M., Gabriel, S., Rieder, M. J., Abecasis, G., Altshuler, D., Nickerson, D. A., Boerwinkle, E., Sunyaev, S., Bustamante, C. D., Bamshad, M. J., Akey, J. M., GO, B., GO, S., and Project, N. E. S. (2012). Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science*, **337**(6090), 64–69.

Thompson, E. A. (1974). Gene Identities and Multiple Relationships. *Biometrics*, **30**(4), 667.

Thompson, E. A. (1975). The estimation of pairwise relationships. *Annals of human genetics*, **39**(2), 173–188.

Thompson, E. A. (1976). Estimation of age and rate of increase of rare variants. *The American Journal of Human Genetics*, **28**(5), 442–452.

Thompson, E. A. (2008). The IBD process along four chromosomes. *Theoretical population biology*, **73**(3), 369–373.

Thompson, E. A. (2013). Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics*, **194**(2), 301–326.

Tschermak, E. (1900). Über künstliche Kreuzung bei *Pisum sativum*. *Berichte der Deutschen Botanischen Gesellschaft*, **18**, 232–239.

UK10K Consortium, Walter, K., Crooks, L., Memari, Y., Perry, J. R. B., Xu, C., Futema, M., Lawson, D., Iotchkova, V., Schiffels, S., Hendricks, A. E., Li, R., Floyd, J., Wain, L. V., Humphries, S. E., Barrett, J. C., Plagnol, V., Richards, J. B., Greenwood, C. M. T., Timpson, N. J., Soranzo, N., Danecek, P., Barroso, I., McCarthy, S., Tachmazidou, I., Durbin, R., Hurles, M. E., Kennedy, K., Palotie, A., Zeggini, E., Cocca, M., Huang, J., and Min, J. L. (2015). The UK10K project identifies rare variants in health and disease. *Nature*, **526**(7571), 82–90.

Vartiainen, E., Laatikainen, T., Peltonen, M., Juolevi, A., Mannisto, S., Sundvall, J., Jousilahti, P., Salomaa, V., Valsta, L., and Puska, P. (2010). Thirty-five-year trends in cardiovascular risk factors in Finland. *International Journal of Epidemiology*, **39**(2), 504–518.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Miklos, G. L. G., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R.-R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang,

Z. Y., Wang, A., Wang, X., Wang, J., Wei, M.-H., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S. C., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Angela Center, Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.-H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigo, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.-H., Coyne, M., Dahlke, C., Mays, A. D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., and Majoros... (2001). The Sequence of the Human Genome. *Science*, **291**(5507), 1304–1351.

Viterbi, A. J. (1967). Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *IEEE Transactions on Information Theory*, **13**(2), 260–269.

Voight, B. F. and Pritchard, J. K. (2005). Confounding from Cryptic Relatedness in Case-Control Association Studies. *PLoS Genetics*, **1**(3), e32–10.

Wakeley, J. (2008). *Coalescent Theory: An Introduction*. W. H. Freeman.

Wakeley, J. and Wilton, P. (2016). Coalescent and models of identity by descent. In R. M. Kliman, editor, *Encyclopedia of Evolutionary Biology*, pages 287 – 292. Academic Press, Oxford.

Wang, J. (2005). Estimation of effective population sizes from data on genetic markers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **360**(1459), 1395–1409.

Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, **171**(4356), 737–738.

Watterson, G. (1996). Motoo Kimura's Use of Diffusion Theory in Population Genetics. *Theoretical population biology*, **49**(2), 154–188.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical population biology*, **7**(2), 256–276.

Watterson, G. A. (1976). Reversibility and the age of an allele. i. moran's infinitely many neutral alleles model. *Theoretical Population Biology*, **10**(3), 239–253.

Weinberg, W. (1908). Über den Nachweis der Vererbung beim Menschen. *Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg.*, **64**, 368–382.

Weissenbach, J. (1993). A second generation linkage map of the human genome based on highly informative microsatellite loci. *Gene*, **135**(1-2), 275–278.

Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G. T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C. L., Irzyk, G. P., Lupski, J. R., Chinault, C., Song, X.-z., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D. M., Margulies, M., Weinstock, G. M., Gibbs, R. A., and Rothberg, J. M. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**(7189), 872–876.

Winkler, H. (1920). *Verbreitung und Ursache der Parthenogenesis im Pflanzen- und Tierreich*. Verlag G. Fischer, Jena.

Wiuf, C. and Hein, J. (1997). On the number of ancestors to a DNA sequence. *Genetics*, **147**(3), 1459–1468.

Wright, S. (1921). Systems of Mating. I. the Biometric Relations between Parent and Offspring. *Genetics*, **6**(2), 111–123.

Wright, S. (1922). Coefficients of Inbreeding and Relationship. *The American Naturalist*, **56**(645), 330–338.

Wright, S. (1931a). Evolution in Mendelian Populations. *Genetics*, **16**(2), 97–159.

Wright, S. (1931b). Evolution in Mendelian Populations. *Genetics*, **16**(2), 97–159.

Yu, N., Zhao, Z., Fu, Y. X., Sambuughin, N., Ramsay, M., Jenkins, T., Leskinen, E., Patthy, L., Jorde, L. B., Kuromori, T., and Li, W. H. (2001). Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1. *Molecular Biology and Evolution*, **18**(2), 214–222.

Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M. J., Neale, B. M., Sunyaev, S. R., and Lander, E. S. (2014). Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences*, **111**(4), E455–64.

1. *I have told you more than I know* [...].
2. *What I have told you is subject to change without notice.*
3. *I hope I raised more questions than I have given answers.*
4. *In any case, as usual, a lot more work is necessary.*

– Fuller Albright