# Advantages of genotype imputation with ethnically matched reference panel for rare variant association analyses

Mart Kals[1,2*], Tiit Nikopensius[1], Kristi Läll[1,3], Kalle Pärn[2], Timo Tõnis Sikka[1,4], Jaana Suvisaari[5], Veikko Salomaa[5], Samuli Ripatti[2,6], Aarno Palotie[2,6], Andres Metspalu[1], Tõnu Esko[1,6], Priit Palta[1,2¶], Reedik Mägi[1¶]

[1] Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia

[2] Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland

[3] Institute of Mathematics and Statistics, University of Tartu, Tartu, Estonia

[4] Department of Biotechnology, Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia

[5] National Institute for Health and Welfare, Helsinki, Finland

[6] Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

* Corresponding author

E-mail: mart.kals@ut.ee (MK)

[¶] These authors jointly supervised this work

# Abstract

Genotype imputation has become a standard procedure prior genome-wide association studies (GWASs). For common and low-frequency variants, genotype imputation can be performed sufficiently accurately with publicly available and ethnically heterogeneous reference datasets like 1000 Genomes Project (1000G) and Haplotype Reference Consortium panels. However, the imputation of rare variants has been shown to be significantly more accurate when ethnically matched reference panel is used. Even more, greater genetic similarity between reference panel and target samples facilitates the detection of rare (or even population-specific) causal variants. Notwithstanding, the genome-wide downstream consequences and differences of using ethnically mixed and matched reference panels have not been yet comprehensively explored.

We determined and quantified these differences by performing several comparative evaluations of the discovery-driven analysis scenarios. A variant-wise GWAS was performed on seven complex diseases and body mass index by using genome-wide genotype data of ~37,000 Estonians imputed with ethnically mixed 1000G and ethnically matched imputation reference panels. Although several previously reported common (minor allele frequency; MAF > 5%) variant associations were replicated in both resulting imputed datasets, no major differences were observed among the genome-wide significant findings or in the fine-mapping effort. In the analysis of rare (MAF < 1%) coding variants, 46 significantly associated genes were identified in the ethnically matched imputed data as compared to four genes in the 1000G panel based imputed data. All resulting genes were consequently studied in the UK Biobank data.

These associations provide a solid example of how rare variants can be efficiently analysed to discover novel, potentially functional genetic variants in relevant phenotypes. Furthermore, our work serves as proof of a cost-efficient study design, demonstrating that the usage of ethnically matched imputation reference panels can enable substantially improved imputation of rare variants, facilitating novel high-confidence findings in rare variant GWAS scans.

## Author summary

Over the last decade, genome-wide association studies (GWASs) have been widely used for detecting genetic biomarkers in a wide range of traits. Typically, GWASs are carried out using chip-based genotyping data, which are then combined with a more densely genotyped reference panel to infer untyped genetic variants in chip-typed individuals. The latter method is called genotype imputation and its accuracy depends on multiple factors. Publicly available and ethnically heterogeneous imputation reference panels (IRPs) such as 1000 Genomes Project (1000G) are sufficiently accurate for imputation of common and low-frequency variants, but custom ethnically matched IRPs outperform these in case of rare variants. In this work, we systematically compare downstream association analysis effects on eight complex traits in ~37,000 Estonians imputed with ethnically mixed and ethnically matched IRPs. We do not observe major differences in the single variant analysis, where both imputed datasets replicate previously reported significant loci. But in the gene-based analysis of rare protein-coding variants we show that ethnically matched panel clearly outperforms 1000G panel based imputation, providing 10-fold increase in significant gene-trait associations. Our study demonstrates empirically that imputed data based on ethnically matched panel is very promising for rare variant analysis – it captures more population-specific variants and makes it possible to efficiently identify novel findings.


## Introduction

Genome-wide association studies (GWASs) have been successfully implemented to capture genetic variants with small to modest effect sizes and have identified thousands of common variants robustly associated with different complex traits and diseases [1]. However, even in aggregate, these explain only a small fraction of the heritability of studied diseases.

The sample size of a GWAS can be increased through relatively cheap chip-based genotyping and subsequent genotype imputation. Imputation is a commonly used computational method for lending information from a densely genotyped reference panel of phased haplotypes, allowing to study variants that have not been directly genotyped in target samples and thereby this approach not only

increases the power but also the resolution of GWAS [2–4]. Genotype imputation can also facilitate better fine-mapping association signals through the increase of genetic variant density in candidate genomic regions [5].

Publicly accessible imputation reference panels like 1000 Genomes Project (1000G) [6] and Haplotype Reference Consortium (HRC) [7] have been frequently used for imputation in advance to GWAS. Nevertheless, both of these ethnically heterogeneous reference panels have only limited capacity to provide complete and accurate imputation of rare (minor allele frequency; MAF < 1%) variants [8], suggested to contribute to the missing heritability [9]. During the last few years it has been shown that using an ethnically matched reference panel can greatly improve the 'completeness' and accuracy of genotype imputation [10–15], resulting in higher imputation accuracy compared to the 1000G panel even in case of smaller panel size [16,17]. In addition, several recent studies have demonstrated the utility of ethnically matched datasets for the discovery of disease or trait-associated rare variants [18–25].

Imputed datasets based on ethnically matched reference panels are considered to be powerful tools to discover previously unidentified rare variants. However, the typical approaches for testing associations of genetic variants with phenotypes based on simple regression models, and are underpowered for rare variants in most studies due to their low frequencies and large numbers [26]. To overcome these issues, different methods have been proposed to increase statistical power in rare variant association studies, typically by combining information across multiple rare variants within a specific genomic region or functional unit (e.g. gene) [27–29]. Often these methods focus on certain categories of variation (e.g. missense or loss-of-function (LoF) variants) [30], and have been applied successfully in several studies [31–33]. Therefore, gene-based tests allow to capture the joint contribution of multiple rare variants, improve power and enable to identify novel disease associated genes encompassing putatively functional variants [34–37].

In the current study, we impute 51,886 chip-typed Estonians with both ethnically matched Estonian-Finnish (EstFin) and ethnically mixed 1000G imputation reference panels (IRPs) to determine and quantify the differences in analysis results of eight complex traits. In particular, we evaluate two

97 analysis scenarios: 1) a variant-wise GWAS; 2) a gene-wise analysis to determine the joint
98 contribution of rare (MAF < 1%) nonsynonymous (NS) and LoF variants which we validate in the UK
99 Biobank data.

## 100 Results

101 First, we developed a high-coverage (~30×) whole genome sequencing (WGS) based imputation
102 reference panel comprising of ethnically closely related 2,279 Estonians and 1,856 Finns, resulting in
103 8,270 haplotypes in the EstFin IRP. Secondly, we imputed 51,886 chip-genotyped Estonians with the
104 EstFin and 1000G IRPs (S1 Appendix, S1 Table, S1 and S2 Figs). The EstFin IRP provided 13.86
105 million (M) and the 1000G IRP 9.06 M confidently imputed variants (imputation INFO-value > 0.8)
106 with MAF > 0.05% in 36,716 unrelated individuals, which were further used to carry out a
107 comparative GWAS and gene-wise association testing of rare variants with eight complex traits.
108 Finally, the identified significant gene-trait associations were studied in the UK Biobank data (Fig 1).

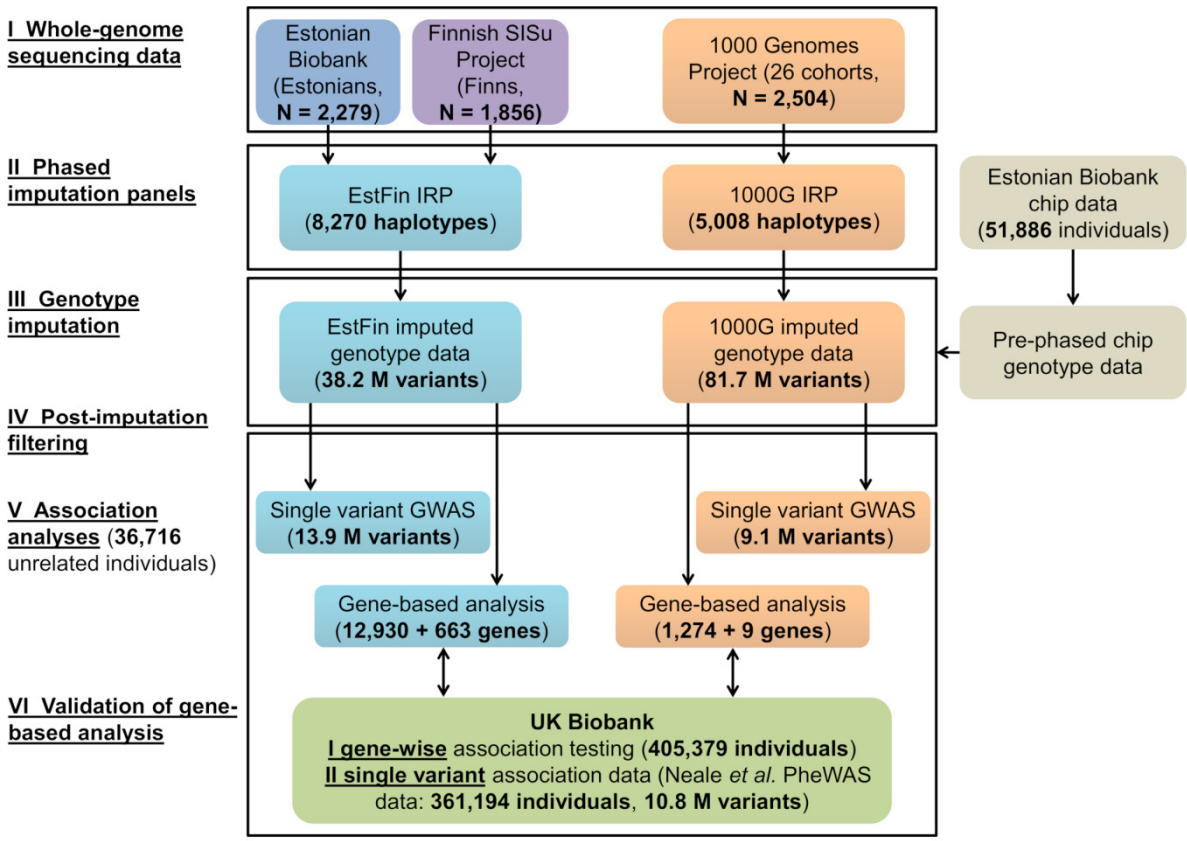


**Fig 1. Schematic overview of used datasets and main processing and analysis steps.** This scheme gives an overview of used imputation reference panels, chip-based and imputed genotype datasets, and comparative association analyses, for which gene-based results were validated in the UK Biobank data.

## Single variant analysis

Genome-wide association testing was performed separately in both imputed datasets with eight complex traits: body mass index and seven complex diseases of major public health importance [38] – bipolar disorder (BD), coronary artery disease (CAD), Crohn's disease (CD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D), and type 2 diabetes (T2D) (S2 Table). We detected 12 and 13 genome-wide significant ($P < 6.25 \times 10^{-9}$) loci based on EstFin and 1000G IRPs, respectively. Results of variant-wise GWA studies are summarized in Table 1 and S3 Fig. In both datasets we discovered eleven identical loci and three IRP-specific associations, all of which have been previously reported [1] (S3 Table). Autoimmune diseases RA and T1D demonstrated common variant associations in the HLA-region, BMI and T2D revealed long established association with *FTO* gene (Fig 2). Although lead variants did not overlap (except rs11102694 and rs9273363 with T1D), top hits from both datasets were in close proximity and in high linkage disequilibrium (S4 Table and S4 Fig).

IRP-specific associations included *BCL2L15* association with T1D in case of the EstFin panel based imputation, whereas an intergenic locus at chromosome 6p12.3 was associated with BMI and *TLE1* locus with T2D in the 1000G-based imputed data only. Although, for both associations, the lowest *P* values in the other IRP-based imputed data were very close to the genome-wide significance level (S4 Table and S3 Fig).

**Table 1. Study-wide significant loci detected in single variant association analysis.** Confidently imputed variants (INFO > 0.8) were tested for associations with complex traits (BMI – body mass index, CAD – coronary artery disease, RA – rheumatoid arthritis, T1D – type 1 diabetes, T2D – type 2 diabetes). Analyses were conducted separately in the EstFin-based and the 1000G-based imputed datasets. The study-wide significance threshold after correction for multiple testing was set to $P < 6.25 \times 10^{-9}$ and for each locus, associated gene containing variant with the lowest $P$ value was reported. In the last two columns, results of fine-mapping analysis are shown with the numbers of putative causal variants per genomic region.
(1) Locus including significant ($P < 3.27 \times 10^{-8}$, 10-fold enrichment in Estonians) variants detected in MAF-enriched analysis in comparison of 503 European individuals from the 1000G phase 3 data.

| Trait | Chromosomal locus | Associated gene | | Number of putative causal variants | |
|---|---|---|---|---|---|
| | | EstFin IRP | 1000G IRP | EstFin IRP | 1000G IRP |
| BMI | 1p13.3 | *AMPD2* | *GPR61* | 1 | 1 |
| BMI | 1q25.2(1) | *SEC16B* | Intergenic | 1 | 1 |
| BMI | 2p25.3(1) | Intergenic | Intergenic | 2 | 2 |
| BMI | 6p12.3 | - | Intergenic | - | 1 |
| BMI | 12q13.12 | *FAIM2* | Intergenic | 1 | 1 |
| BMI | 16q12.2 | *FTO* | *FTO* | 1 | 1 |
| BMI | 18q21.32(1) | Intergenic | Intergenic | 1 | 1 |
| CAD | 9p21.3 | Intergenic | Intergenic | 1 | 1 |
| RA | 6p21.32(1) | Intergenic | Intergenic | 1 | 2 |
| T1D | 1p13.2 | *BCL2L15* | - | 1 | - |
| T1D | 6p21.32(1) | Intergenic | Intergenic | 5 | 3 |
| T2D | 9q21.32 | - | *TLE1* | - | 1 |
| T2D | 10q25.2-25.3 | *TCF7L2* | *TCF7L2* | 1 | 1 |
| T2D | 16q12.2 | *FTO* | *FTO* | 1 | 1 |
| **Total** | | **12** | **13** | **17** | **17** |

To identify the likely causal variant at each locus, we performed fine-mapping analysis in all significant genomic regions discovered in genome-wide association scan. All but three (BMI at 2p25.3, RA and T1D at the HLA-region) significant regions demonstrated only one likely causal variant (Table 1). We also tested variants having allele frequency enrichment in Estonians as compared to the 503 European individuals from the 1000G data. Genome-wide significant ($P < 3.27 \times 10^{-8}$ for 10-fold enrichment) variants were detected for BMI at 1q25.2, 2p25.3, and 18q21.32 loci and for RA and T1D in the HLA-region in both imputed datasets (Table 1).

In conclusion, single variant association analyses did not indicate major differences in results based on data imputed with ethnically matched and mixed IRPs.

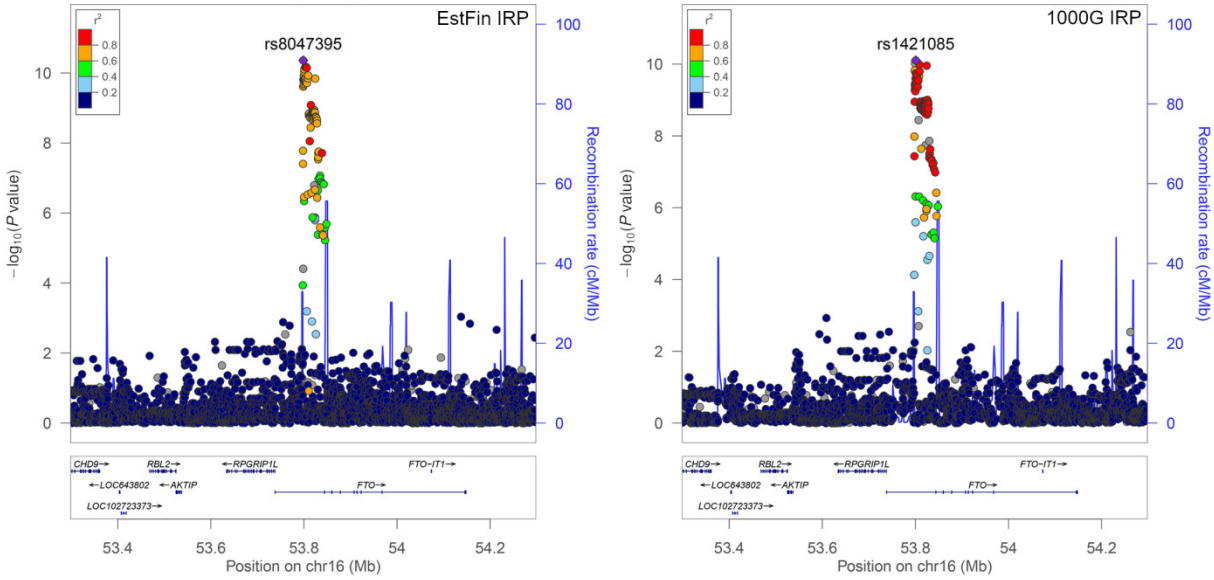


**Fig 2. Regional association plots of type 2 diabetes at the *FTO* locus.** The left panel of regional plot shows the genome-wide association analysis results for the EstFin-based imputed data, while the right panel shows results for the 1000G IRP imputed data. The purple symbol represents the lead variant, and the rest of the colour-coded variants denote LD with the lead variant estimated by $r^2$ from the 1000G phase 3 (EUR population) data. Comparison of analysis results of both imputed datasets indicates that variants with the lowest *P* value do not overlap, but are highly correlated. Imputed data based on the EstFin panel provides evidence for an association between T2D and alleles of the *FTO* locus with the lowest *P* value at rs8047395 ($P = 4.4 \times 10^{-11}$). The same SNV shows a significant association in the data imputed with 1000G IRP ($P = 1.5 \times 10^{-10}$), but the lowest *P* value is at rs1421085 ($P = 7.9 \times 10^{-11}$, Pearson's $r^2$=0.82 between rs8047395 and rs1421085).

## Gene-based analysis

In addition to single variant analyses, we conducted gene-based association testing of rare (MAF < 1%) nonsynonymous (NS) and loss-of-function (LoF) variants for eight complex traits in both imputed datasets. When considering genes with at least two confidently imputed (INFO > 0.8) rare NS variants, we observed noteworthy differences in the number of genes analysed – EstFin IRP outperformed 1000G, providing 12,930 and 1,274 unique genes, respectively. In the analysis of rare LoF variants, we identified even more drastic differences – 663 genes were tested in the EstFin panel imputation and only six genes in the 1000G panel imputation.

Consequently, whilst testing the genes including rare NS variants, we detected 38 significant gene-trait associations ($P_{NS} < 4.83 \times 10^{-7}$) in the EstFin-based imputed data and four in the 1000G panel based imputation (Table 2). At significance level $P_{LoF} < 9.40 \times 10^{-6}$ we detected 10 genes including rare LoF variants based on the EstFin imputed data and none in the 1000G-based imputation (Table 2). Comparative results of gene-based analysis are presented in Figures 3 and S5. While none of these

147 associated genes were implicated in our single variant GWAS, 122 NS and 22 LoF variants were
148 involved in gene-wise analysis (S5 Table). We determined that the large majority (45 out of 52) of
149 significant gene-trait associations relied on two or three NS/LoF variants. Seven out of 52 associations
150 relied on four or more NS/LoF variants and the signal of joint contribution of rare variants was driven
151 by multiple variants ($P < 0.05$) for 22/52 tests (S5 Table).

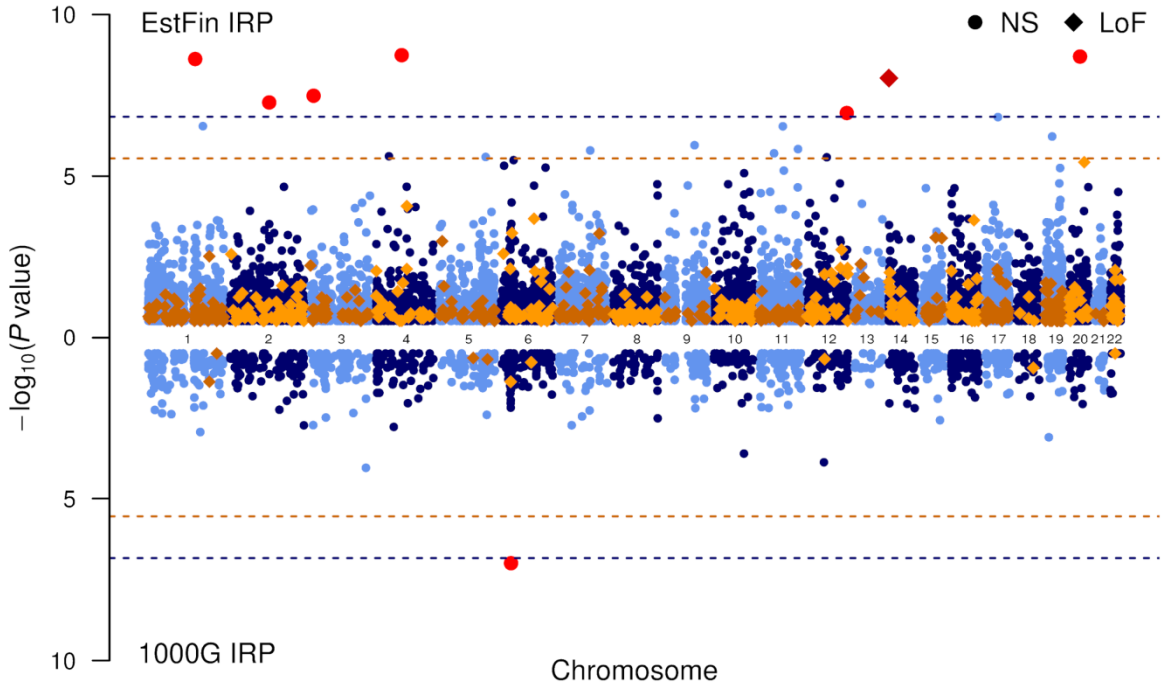


**Fig 3. Miami plot of the gene-based association testing of bipolar disorder.** The top panel shows the gene-based association analysis results using the EstFin-based imputed data, while the bottom part shows results for the 1000G IRP imputed data. Blue dots represent tested genes including NS variants and orange squares LoF variants. Dashed lines indicate significance levels after correction for multiple testing: $P < 4.83 \times 10^{-7}$ for NS variants (blue) and $P < 9.40 \times 10^{-6}$ for LoF variants (orange). Red symbols denote significant genes. In the analysis of NS variants we identify six genes based on the EstFin IRP data and one significant gene in the 1000G-based data. In the EstFin-based imputed data we detect a single gene-trait association of LoF substitutions, whereas none of the significant associations is observed in the data based on ethnically mixed 1000G IRP.

**Table 2. Identified significant gene-trait associations.** Genes with at least two confidently imputed (INFO > 0.8) rare (MAF < 1%) nonsynonymous (NS) and loss-of-function (LoF) variants are tested for association with complex traits (BMI – body mass index, BD – bipolar disorder, RA – rheumatoid arthritis, T1D – type 1 diabetes, T2D – type 2 diabetes). Analyses are performed separately in the EstFin-based and the 1000G-based imputed datasets. Multiple testing corrected significance levels are applied based on the number of genes tested in both datasets: $P_{NS} < 4.83 \times 10^{-7}$ for genes containing NS variants and $P_{LoF} < 9.40 \times 10^{-6}$ for genes containing LoF variants. First, the results of gene-based analysis in 36,716 Estonian Biobank (EBB) individuals are presented. Next, the gene-trait associations are validated in 405,379 UK Biobank (UKBB) individuals. Finally, for each significant gene-trait result detected in the EBB data, variant-trait association with the smallest *P* value from the UKBB single variant GWAS is provided.

| | | | | | **Gene-based analysis of EBB data** | **Gene-based analysis of UKBB data** | **Single variant analysis of UKBB data** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **IRP** | **Functional annotation** | **Trait** | **Gene** | **Chr** | ***P* value** | ***P* value** | **Lead variant** | **Minor allele** | **MAF** | **Beta** | **Se** | ***P* value** |
| 1000G | NS | BD | *GGNBP1* | 6 | $3.35 \times 10^{-7}$ | 1 | rs141041358 | A | $1.32 \times 10^{-3}$ | 0.0032 | 0.0011 | $5.08 \times 10^{-3}$ |
| | | CD | *HTR3D* | 3 | $9.48 \times 10^{-10}$ | 0.676 | rs570697703 | G | $4.02 \times 10^{-7}$ | 4.5451 | 1.4028 | $1.20 \times 10^{-3}$ |
| | | CD | *GPRC6A* | 6 | $4.35 \times 10^{-7}$ | 0.749 | rs150641887 | A | $3.84 \times 10^{-6}$ | 0.4667 | 0.0380 | $1.30 \times 10^{-34}$ |
| | | T1D | *FAM186B* | 12 | $6.19 \times 10^{-8}$ | 0.220 | rs140980069 | T | $3.56 \times 10^{-5}$ | 0.0385 | 0.0080 | $1.60 \times 10^{-6}$ |
| EstFin | NS | BMI | *CYLD* | 16 | $1.68 \times 10^{-7}$ | - | rs190787930 | A | $4.17 \times 10^{-3}$ | 0.0592 | 0.0190 | $1.89 \times 10^{-3}$ |
| | | BD | *SHE* | 1 | $7.88 \times 10^{-9}$ | 0.179 | rs79480105 | A | $4.49 \times 10^{-2}$ | 0.0008 | 0.0003 | $1.01 \times 10^{-2}$ |
| | | BD | *AMMECR1L* | 2 | $1.74 \times 10^{-7}$ | 0.375 | rs137977337 | T | $1.72 \times 10^{-3}$ | 0.0045 | 0.0015 | $2.85 \times 10^{-3}$ |
| | | BD | *VGLL4* | 3 | $1.08 \times 10^{-7}$ | 1 | rs528386411 | C | $1.16 \times 10^{-5}$ | 0.1132 | 0.0323 | $4.49 \times 10^{-4}$ |
| | | BD | *LIN54* | 4 | $6.00 \times 10^{-9}$ | 0.324 | rs142253468 | C | $3.78 \times 10^{-5}$ | 0.0825 | 0.0110 | $6.05 \times 10^{-14}$ |
| | | BD | *FZD10* | 12 | $3.69 \times 10^{-7}$ | 0.711 | rs1046893 | C | $3.76 \times 10^{-1}$ | 0.0001 | 0.0001 | $2.75 \times 10^{-1}$ |
| | | BD | *SAMHD1* | 20 | $6.62 \times 10^{-9}$ | 0.755 | rs566610995 | G | $1.58 \times 10^{-3}$ | 0.0029 | 0.0010 | $4.59 \times 10^{-3}$ |
| | | CD | *HTR3D* | 3 | $2.38 \times 10^{-8}$ | 0.676 | rs570697703 | G | $4.02 \times 10^{-7}$ | 4.5451 | 1.4028 | $1.20 \times 10^{-3}$ |
| | | CD | *PLA2G12A* | 4 | $1.44 \times 10^{-7}$ | 0.468 | rs763365177 | G | $1.61 \times 10^{-3}$ | 0.0033 | 0.0017 | $5.25 \times 10^{-2}$ |
| | | CD | *HLA-G* | 6 | $4.33 \times 10^{-7}$ | - | rs80153902 | A | $3.60 \times 10^{-5}$ | 0.0360 | 0.0106 | $6.83 \times 10^{-4}$ |
| | | CD | *RAPGEF5* | 7 | $1.55 \times 10^{-7}$ | 0.227 | rs578001462 | T | $2.16 \times 10^{-4}$ | 0.0191 | 0.0048 | $6.65 \times 10^{-5}$ |
| | | CD | *ZNF92* | 7 | $5.86 \times 10^{-10}$ | 0.046 | rs144227733 | A | $2.22 \times 10^{-4}$ | 0.0109 | 0.0045 | $1.58 \times 10^{-2}$ |
| | | CD | *ORC5* | 7 | $1.78 \times 10^{-10}$ | 1 | rs76304209 | T | $4.26 \times 10^{-3}$ | 0.0037 | 0.0011 | $2.23 \times 10^{-4}$ |
| | | CD | *R3HCC1* | 8 | $6.58 \times 10^{-9}$ | 0.368 | rs375458319 | A | $2.45 \times 10^{-3}$ | 0.0027 | 0.0014 | $4.82 \times 10^{-2}$ |
| | | CD | *TMEM64* | 8 | $5.18 \times 10^{-10}$ | 0.083 | rs185086305 | T | $2.93 \times 10^{-5}$ | 0.0500 | 0.0138 | $2.98 \times 10^{-4}$ |
| | | CD | *NEU3* | 11 | $3.61 \times 10^{-7}$ | 0.876 | rs35872360 | G | $2.86 \times 10^{-1}$ | -0.0004 | 0.0001 | $1.12 \times 10^{-2}$ |
| | | CD | *EED* | 11 | $1.87 \times 10^{-8}$ | - | rs534451904 | A | $2.44 \times 10^{-3}$ | 0.0040 | 0.0014 | $2.97 \times 10^{-3}$ |
| | | CD | *KCNA1* | 12 | $1.16 \times 10^{-7}$ | - | rs149959487 | A | $3.80 \times 10^{-4}$ | -0.0032 | 0.0036 | $3.68 \times 10^{-1}$ |
| | | CD | *TAOK2* | 16 | $7.01 \times 10^{-9}$ | 0.792 | rs10445105 | G | $5.37 \times 10^{-2}$ | -0.0007 | 0.0003 | $1.76 \times 10^{-2}$ |
| | | CD | *ANKRD30B* | 18 | $3.03 \times 10^{-8}$ | 0.766 | rs9675858 | T | $4.99 \times 10^{-1}$ | -0.0004 | 0.0001 | $4.03 \times 10^{-4}$ |
| | | CD | *TRIP10* | 19 | $1.36 \times 10^{-7}$ | 0.288 | rs61757561 | C | $1.68 \times 10^{-2}$ | 0.0012 | 0.0005 | $2.20 \times 10^{-2}$ |
| | | CD | *GRWD1* | 19 | $1.59 \times 10^{-7}$ | 0.022 | rs199819631 | G | $1.11 \times 10^{-3}$ | 0.0065 | 0.0021 | $2.13 \times 10^{-3}$ |
| | | CD | *WRB* | 21 | $1.37 \times 10^{-7}$ | 0.105 | rs553712185 | G | $2.23 \times 10^{-1}$ | 0.0004 | 0.0002 | $1.01 \times 10^{-2}$ |
| | | CD | *NIPSNAP1* | 22 | $1.05 \times 10^{-7}$ | - | rs549356766 | A | $2.36 \times 10^{-6}$ | -0.0049 | 0.0530 | $9.27 \times 10^{-1}$ |
| | | RA | *CREBRF* | 5 | $3.24 \times 10^{-9}$ | 0.260 | rs78586862 | G | $5.66 \times 10^{-2}$ | -0.0015 | 0.0005 | $5.51 \times 10^{-3}$ |
| | | RA | *HSPB1* | 7 | $3.97 \times 10^{-7}$ | - | rs145206720 | C | $2.62 \times 10^{-4}$ | -0.0094 | 0.0081 | $2.47 \times 10^{-1}$ |
| | | RA | *PRPF31* | 19 | $2.83 \times 10^{-7}$ | 0.783 | rs187106635 | A | $2.40 \times 10^{-3}$ | -0.0066 | 0.0027 | $1.58 \times 10^{-2}$ |

| | | | | | Gene-based analysis of EBB data | Gene-based analysis of UKBB data | Single variant analysis of UKBB data | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **IRP** | **Functional annotation** | **Trait** | **Gene** | **Chr** | ***P* value** | ***P* value** | **Lead variant** | **Minor allele** | **MAF** | **Beta** | **Se** | ***P* value** |
| EstFin | NS | RA | *THOC5* | 22 | $1.47 \times 10^{-7}$ | 0.284 | rs571323390 | C | $1.08 \times 10^{-3}$ | 0.0040 | 0.0012 | $5.44 \times 10^{-4}$ |
| | | T1D | *RNF13* | 3 | $6.88 \times 10^{-9}$ | 0.586 | rs140200425 | C | $2.90 \times 10^{-2}$ | 0.0009 | 0.0003 | $1.25 \times 10^{-3}$ |
| | | T1D | *CENPU* | 4 | $7.30 \times 10^{-9}$ | 0.617 | rs776959139 | A | $2.24 \times 10^{-3}$ | 0.0030 | 0.0008 | $7.98 \times 10^{-5}$ |
| | | T1D | *HIST1H1C* | 6 | $3.63 \times 10^{-7}$ | 0.371 | rs201637343 | T | $5.73 \times 10^{-4}$ | 0.0065 | 0.0021 | $2.08 \times 10^{-3}$ |
| | | T1D | *BAK1* | 6 | $5.96 \times 10^{-11}$ | 0.583 | rs11757379 | T | $2.26 \times 10^{-1}$ | 0.0003 | 0.0001 | $4.96 \times 10^{-3}$ |
| | | T1D | *FAM170B* | 10 | $1.14 \times 10^{-8}$ | 0.464 | rs75297145 | T | $2.09 \times 10^{-1}$ | 0.0002 | 0.0001 | $3.59 \times 10^{-2}$ |
| | | T1D | *SLC22A8* | 11 | $1.61 \times 10^{-7}$ | 0.210 | rs11568481 | A | $2.33 \times 10^{-4}$ | 0.0110 | 0.0023 | $1.41 \times 10^{-6}$ |
| | | T1D | *C11orf30* | 11 | $1.32 \times 10^{-7}$ | 0.576 | rs74904466 | A | $2.70 \times 10^{-2}$ | 0.0008 | 0.0002 | $2.09 \times 10^{-4}$ |
| | | T1D | *ISLR* | 15 | $3.63 \times 10^{-7}$ | 0.249 | rs1052622 | G | $3.23 \times 10^{-1}$ | 0.0002 | 0.0001 | $4.14 \times 10^{-2}$ |
| | | T1D | *HOXB6* | 17 | $3.23 \times 10^{-7}$ | 1 | rs33990581 | T | $1.22 \times 10^{-2}$ | -0.0006 | 0.0004 | $1.70 \times 10^{-1}$ |
| | | T1D | *ZNF701* | 19 | $1.40 \times 10^{-7}$ | 0.324 | rs370776009 | G | $3.43 \times 10^{-1}$ | 0.0002 | 0.0001 | $2.91 \times 10^{-2}$ |
| EstFin | LoF | BD | *OR11H6* | 14 | $3.07 \times 10^{-8}$ | - | rs143225754 | T | $1.28 \times 10^{-3}$ | 0.0021 | 0.0011 | $6.67 \times 10^{-2}$ |
| | | CD | *HLA-G* | 6 | $2.67 \times 10^{-7}$ | - | rs80153902 | A | $3.60 \times 10^{-5}$ | 0.0360 | 0.0106 | $6.83 \times 10^{-4}$ |
| | | CD | *IQCE* | 7 | $2.51 \times 10^{-6}$ | - | rs80187333 | A | $6.83 \times 10^{-3}$ | -0.0023 | 0.0008 | $6.29 \times 10^{-3}$ |
| | | CD | *YME1L1* | 10 | $1.47 \times 10^{-7}$ | - | rs558886293 | T | $2.33 \times 10^{-3}$ | 0.0043 | 0.0015 | $3.71 \times 10^{-3}$ |
| | | CD | *ANKRD30B* | 18 | $9.83 \times 10^{-8}$ | - | rs9675858 | T | $4.99 \times 10^{-1}$ | -0.0004 | 0.0001 | $4.03 \times 10^{-4}$ |
| | | CD | *FERMT1* | 20 | $1.43 \times 10^{-7}$ | - | rs78566304 | A | $1.06 \times 10^{-3}$ | 0.0069 | 0.0021 | $1.09 \times 10^{-3}$ |
| | | T1D | *ALLC* | 2 | $1.42 \times 10^{-6}$ | - | rs573758969 | C | $1.45 \times 10^{-3}$ | 0.0036 | 0.0010 | $2.38 \times 10^{-4}$ |
| | | T1D | *LPP* | 3 | $9.76 \times 10^{-8}$ | - | rs186012592 | G | $6.79 \times 10^{-5}$ | 0.0224 | 0.0045 | $8.11 \times 10^{-7}$ |
| | | T1D | *ZIC2* | 13 | $1.44 \times 10^{-11}$ | - | rs13542 | A | $2.23 \times 10^{-1}$ | 0.0002 | 0.0001 | $4.98 \times 10^{-2}$ |
| | | T1D | *ZNF83* | 19 | $2.73 \times 10^{-8}$ | 0.609 | rs329940 | A | $1.09 \times 10^{-4}$ | 0.0168 | 0.0056 | $2.50 \times 10^{-3}$ |

152 **Validation of gene-based analysis**

153 First, all 52 unique genes that demonstrated significant gene-trait associations in gene-wise analysis
154 were selected and gene-based association testing was repeated using 405,379 individuals from the UK
155 Biobank [39,40] (Table 2). The strongest gene-trait associations were detected with CD in *GRWD1* (*P*
156 = 0.022) and *ZNF92* (*P* = 0.046) genes, but neither of these were significant after correcting for
157 multiple testing.

158 Secondly, we used variant-wise GWAS results of the UK Biobank in 361,194 individuals [41].
159 Although approximately only one fourth of the tested rare NS and LoF variants overlapped between
160 the Estonian Biobank and the UK Biobank data, the UKBB GWAS analysis confirmed association
161 signals ($P < 10^{-5}$) in five detected genes. Considering the lowest *P* values in our candidate gene
162 regions, we detected two significant ($P = 1.30 \times 10^{-34}$, CD in *GPRC6A* with the 1000G imputation; *P*
163 $= 6.05 \times 10^{-14}$, BD in *LIN54* with the EstFin imputation) and three suggestive ($P = 8.11 \times 10^{-7}$, T1D in
164 *LPP* with the EstFin imputation, $P = 1.41 \times 10^{-6}$, T1D in *SLC22A8* with the EstFin imputation, $P = 1.60$
165 $\times 10^{-6}$, T1D in *FAM186B* with the 1000G imputation) associations in the UKBB GWAS data (Table
166 2). Significant gene-based findings with plausible evidence of biological meaningfulness including
167 *VGLL4*, *LPP* and *HLA-G* as well as few other loci are discussed in S1 Appendix. Relevant GWAS
168 Catalog entries related to significant findings from gene-based analysis are presented in S6 Table.

169 Gene-based analysis of rare variants demonstrated that our ethnically matched panel clearly
170 outperformed the 1000G-based imputation, providing approximately 10-fold increase in the number of
171 tested genes and significant findings. Further literature and GWAS Catalog mining of these significant
172 associations determined that most of these associated genes have been previously indicated, but there
173 were also several novel findings.

## 174 Discussion

175 Over the past few years, several ethnically matched imputation reference panels have been developed
176 to complement the widely used cosmopolitan 1000G and HRC panels. The former panels have showed
177 great improvement in imputation accuracy, but their effect to the downstream analyses have not been

178 well examined. In the current study, we performed a comparison of ethnically matched EstFin and
179 ethnically mixed 1000G-based imputed genotypes in the Estonian Biobank study cohort of ~52,000
180 individuals. In addition to the single variant analysis, we examined downstream differences as a
181 measure of identified associations in MAF-enriched and rare-variant analysis. We demonstrate that
182 ethnically matched reference panel empowers the detection of rare variant signals, enabling to identify
183 clinically significant novel loci for complex diseases that will be further discussed below.

## 184 Ethnically matched reference panel lead to considerable gain in downstream analyses 185 for rare variants

186 Ethnically matched panel provides a significantly higher proportion of confidently imputed variants
187 compared to the 1000G panel (S1 Table). The difference increases with the decrease of MAF, because
188 of the insufficient representation of rare (or even population-specific) variation in the mixed-ethnicity
189 panels like 1000G IRP. Correspondingly, we observed that single variant GWAS identified a similar
190 number of genome-wide significant findings in these two imputed datasets and we did not detect any
191 major differences in fine-mapping of these loci. At the same time, it should be taken into consideration
192 that in the current analyses we rely on a relatively small number of disease cases, resulting in limited
193 statistical power. It is likely one of the main factors why we did not observe any major differences in
194 single variant GWA analyses. Possibly these results would be different with significantly larger
195 cohorts as, at some point, one should start detecting low-frequency and rare variants that have been
196 imputed confidently and therefore can be tested with the ethnically matched IRPs. Secondly, 1000G
197 reference panel contains European haplotypes, and therefore it can be a relatively good imputation
198 reference for common variation present in the Estonian population. But the results can differ for those
199 populations, which are more distant from the populations used in transethnic imputation reference sets.

200 Rare variant analyses demonstrated great differences, where the EstFin-based imputation clearly out-
201 performed the 1000G imputation, allowing for the identification of 10 times more genome-wide
202 significant genes (Table 2). The EstFin IRP includes a larger number of haplotypes close to the target
203 samples, deriving unique variants from genomes not included in the 1000G panel. This improves the
204 chances of a rare variant being effectively tagged by a haplotype. Moreover, including haplotypes

205 from ethnically distant populations may not accurately capture LD patterns of population-specific
206 variants or imputation can introduce polymorphic variants in the target samples that are actually
207 monomorphic as observed previously [10]. Our results empirically demonstrate the contribution of
208 rare variants in complex traits analysis using ethnically matched panel, as compared to ethnically
209 mixed population reference.

210 **Validation of gene-based analysis**

211 Validation of gene-based analysis results in the UK Biobank individual-level data detected two
212 significant gene-trait signals ($P < 0.05$), but neither remained significant after multiple testing
213 correction. For gene-trait associations detected in the EBB data, we were not able to validate 6 (out of
214 42) and 9 (out of 10) genes containing NS and LoF variants, respectively. This was accounted for the
215 UKBB data containing less than two NS and LoF variants within these genes (Table 2). A likely
216 explanation is that the ethnically matched panel captures a significantly larger number of such rare
217 variants which are not well-captured through the imputation with more heterogeneous reference
218 panels. Therefore we argue that failing to validate most of the gene-based analysis results in the UK
219 Biobank data can be due to the population-specific nature of the rare variant findings.

220 Nevertheless, some of the associations were validated by matching the observed significant genes with
221 the variants located in the same gene regions in the UKBB single variant association analysis, as well
222 as many of the genes detected by us were associated with relevant traits in literature (S1 Appendix).
223 We hypothesize about a causative role for *LPP* variants conferring susceptibility to T1D – an
224 assumption being initially rejected in a study involving both celiac disease and T1D patients [42].
225 Pleiotropic effects have been reported for *LPP* in association studies involving diverse autoimmune di-
226 seases where shared susceptibility factors outside the HLA-region are widely recognized. In addition,
227 *LPP* mRNA and protein are expressed in multiple tissues, including islets of Langerhans and pancreas,
228 and *LPP* gene is relatively intolerant of LoF variation (ExAC pLI = 0.58) [43].

229 In conclusion, ethnically matched IRPs enable considerably more confident imputation of low-
230 frequency and rare variants as compared to ethnically mixed IRPs. Due to the size of our study dataset
231 (36,716 individuals), small number of disease cases and consequently limited statistical power, we did

232 not observe major differences in single variant GWAS results. However, in the gene-based analysis we
233 observed a substantial advantage of ethnically matched IRP-based imputation over unmatched mixed
234 reference panel-based imputation, enabling significantly improved gene-based analysis of also low-
235 frequency and rare variants, allowing more efficient discovery of rare disease-associated genes and
236 variants.

# 237 Materials and methods

## 238 Study cohorts

239 **Estonian Biobank.** The Estonian Biobank (EBB) is a population-based biobank of the Estonian
240 Genome Center at the University of Tartu. EBB contains almost 52,000 individuals of the Estonian
241 population (aged ≥18 years), which closely reflects the age, sex and geographical distribution of the
242 Estonian adult population [44]. At baseline, the general practitioners performed a standardized health
243 examination of the participants, who also donated blood samples for DNA, white blood cells and
244 plasma tests and filled out a questionnaire on health-related topics. All biobank participants have
245 signed a broad informed consent form, which allows periodical linking to national registries,
246 electronic health record databases and hospital information systems. The majority of biobank
247 participants have been analysed using genotyping arrays. High-coverage whole genome sequencing
248 data is available for the 2,535 individuals, selected randomly by county of birth. The project was
249 approved by the Research Ethics Committee of the University of Tartu (application number 234/T-12).

250 **FINRISK.** FINRISK is a series of health examination surveys carried out by the National Institute
251 for Health and Welfare (formerly National Public Health Institute) of Finland every five years since
252 1972. The surveys are based on random population samples from five (or six in 2002) specified
253 geographical areas of Finland. The samples have been stratified by 10-year age group, sex and study
254 area. The sample sizes have varied from approximately 7,000 to 13,000 individuals and the
255 participation rates from 60% to 90% in different study years. The age-range was 25-64 years until
256 1992 and 25-74 since 1997. The survey included a self-administered questionnaire, a standardized
257 clinical examination carried out by specifically trained study nurses and drawing of a blood sample.

258 Details of the examination have been previously described [45,46]. DNA has been collected since the
259 1992 survey from approximately 34,000 participants. The surveys have appropriate ethical approvals
260 following the usual practices of each survey-year and the participants have signed an informed
261 consent. The validity of clinical diagnoses in these registers has been documented in several
262 publications [47–50].

263 **Finnish Migraine Families collection.** The families were collected over a period of 25 years from
264 six headache clinics in Finland (Helsinki, Turku, Jyväskylä, Tampere, Kemi, and Kuopio) and through
265 advertisements on the national migraine patient organization web page (http://migreeni.org/).
266 Geographically, family members are represented from across the entire country. The current collection
267 consists of 1,589 families, which included a complete range of pedigree sizes from small to large (e.g.,
268 1,023 families had 1–4 related individuals and 566 families had 5+ related individuals). Currently, the
269 collection consists of 8,319 family members, of whom 5,317 have a migraine diagnosis based on the
270 third edition of the established International Classification for Headache Disorders (ICHD-3) criteria
271 [51].

272 **MESTA.** The Living Conditions and Physical Health of Outpatients with Schizophrenia study
273 recruited 276 outpatients with schizophrenia spectrum disorder (ICD-10 F20–F29) from the psychosis
274 outpatient clinics of three municipalities in Finland (Järvenpää, Mäntsälä, and Tuusula). The study
275 protocol consisted of a questionnaire and interview assessing current symptoms, functioning, lifestyle,
276 and a comprehensive health examination. DNA samples were collected as a part of the study based on
277 a separate informed consent [52,53].

278 **Health 2000.** The Finnish Health 2000 Survey was based on a nationally representative sample of
279 8,028 persons aged 30 years or over living in mainland Finland. A two-stage stratified cluster
280 sampling design was used. The sampling frame was regionally stratified according to the five
281 university hospital regions, and from each university hospital region 16 health care districts were
282 sampled as clusters (altogether 80 health care districts). Persons within the health care districts were
283 selected by systematic sampling, and persons aged 80 years and over were oversampled by doubling
284 the sampling fraction. The field work took place between September 2000 and June 2001, and

285 consisted of a home interview and a health examination at the local health centre, or a condensed inter-
286 view and health examination of non-respondents at home. In addition, several questionnaires were
287 used to assess symptoms, lifestyle, and exposures related to different health problems. Of the study
288 sample, 88% were interviewed, 80% attended a comprehensive health examination and 5% attended a
289 condensed examination at home [54].

290 **Ethnically matched imputation reference panel**

291 WGS data for Estonian and Finnish samples were generated and jointly processed at the Broad
292 Institute of MIT and Harvard. WGS samples had PCR-free DNA preparation (Estonian Biobank,
293 FINRISK, Finnish Migraine Families collection, Health 2000) and PCR-amplified preparation
294 (MESTA), followed by sequencing on the Illumina HiSeq X platform with the use of 151-bp paired-
295 end reads with mean coverage of ~30×. Sequenced reads were aligned to the GRCh37 human
296 reference genome assembly using BWA-MEM v0.7.7 [55]; PCR duplicates were marked using Picard
297 v1.136 (http://broadinstitute.github.io/picard/), and the Genome Analysis Toolkit (GATK) v3.4-46
298 [56,57] best-practice guideline was applied for further BAM processing and variant calling.

299 Samples were excluded based on high contamination (>5%), high proportion of chimeric alignment
300 (>5%), low genotype quality (GQ < 50), low coverage (<20×), high coverage (> mean + 3 sd),
301 relatedness (identity-by-descent (IBD) > 0.1), sex mismatches, high genotype discordance (>5%)
302 between sequenced and chip-based data. Additionally, samples were filtered (mean ± 3 sd) based on
303 total number of variants, non-reference variants, singletons, heterozygous/homozygous variants ratio
304 (single nucleotide variants (SNVs) and indels were tested separately in the above-mentioned cases),
305 insertion/deletion ratio for novel indels, insertion/deletion ratio for indels observed in dbSNP, and
306 transition/transversion ratio. After filtering and exclusion of duplicates, the WGS datasets were
307 merged, containing 4,135 individuals (2,279 Estonians and 1,856 Finns).

308 The following variants were set to missing: GQ < 20, read depth > 200×, phred-scaled genotype
309 likelihood of reference allele < 20 for heterozygous and homozygous variant calls, and allele balance
310 <0.2 or >0.8 for heterozygous calls. The GATK Variant Quality Score Recalibration (VQSR) was
311 used to filter variants with a truth sensitivity of 99.8% for SNVs and of 99.9% for indels. Variants

312 with inbreeding coefficient < –0.3, quality by depth < 2 for SNVs and < 3 for indels, call rate < 90%,
313 and Hardy-Weinberg equilibrium (HWE) *P* value < $1\times10^{-9}$ were removed. Monomorphic, multi-allelic
314 variants, and low-complexity regions [58] were further excluded. The final IRP contains 38,226,084
315 variants.

316 Autosomal chromosomes and GRCh37 (hg19) human reference genome assembly was used for all
317 analysis.

## 318 Chip-based genotype data

319 The EBB participants have been analysed using Illumina genotyping arrays: 1) Global Screening
320 Array (GSA, N=33,277), 2) HumanCoreExome (CE, N=7,832), 3) HumanOmniExpress (OMNI,
321 N=8,137), and 4) 370K (N=2,640). Individuals with missing phenotype data were excluded. Final set
322 of genotyped data contained 48,163 unique individuals. The genotype calling for the microarrays was
323 performed using Illumina's GenomeStudio v2010.3 software. The genotype calls for rare variants on
324 the GSA array were corrected using the zCall software (version May 8th, 2012). After variant calling,
325 the data was filtered using PLINK v.1.90 [59] by sample (call rate > 95%, no sex mismatches between
326 phenotype and genotype data, heterozygosity < mean ± 3 sd) and marker-wise (HWE *P* value > 1 ×
327 $10^{-6}$, call rate > 95%, and for the GSA array additionally by Illumina GenomeStudio GenTrain score >
328 0.6, Cluster Separation Score > 0.4). Before the imputation, variants with MAF < 1% and C/G or T/A
329 polymorphisms as well as indels were removed, as these genotype calls do not allow precise phasing
330 and imputation.

## 331 Phasing and imputation

332 The WGS-based imputation reference panel was phased using Eagle v2.3 [60,61] with default
333 parameters except the *Kpbwt* parameter that was set to 20000 to increase accuracy. Pre-phasing of
334 genotyped data was performed in similar manner for all four arrays separately with Eagle and imputed
335 with Beagle v4.1 [62]. All pre-phased genotype datasets were imputed twice using the following
336 reference panels: 1) EstFin IRP containing 8,270 reference haplotypes and 38.2 M autosomal variants;
337 2) 1000G IRP holding 5,008 reference haplotypes and 81.7 M autosomal markers. All four imputed
338 arrays were merged by IRP with BCFtools v1.6 (https://samtools.github.io/bcftools/bcftools.html).

339 Imputation information measure (INFO-value) [4] were added using BCFtools plugin 'impute-info'.
340 Monomorphic, multi-allelic and directly genotyped variants were excluded for all downstream
341 analyses. Only confidently imputed variants (INFO-value > 0.8) with MAF > 0.05% were considered:
342 13,859,717 (12,872,515 SNVs, 987,202 indels) variants imputed with the EstFin IRP and 9,058,236
343 (8,232,261 SNVs, 825,975 indels) variants imputed with the 1000G IRP (Fig 1).

344 **Phenotypes**

345 In the association analysis, only unrelated individuals were included (IBD sharing < 0.2). Samples
346 were excluded by choosing the minimal list of related individuals to break all kinship ties and, if
347 possible, cases were preferred over controls using RELOUT5 tool from Allele
348 (http://www.toomashaller.com/allele.html). Questionnaire-based data was linked to the electronic
349 health records (the Estonian Health Insurance database, data available for years 2003–2015) and other
350 health-related databases like the Estonian Causes of Death Registry (2003–2015), and the Estonian
351 Cancer Registry (2003–2013).

352 After linking, dead people without time of death, participants without records from registries, and
353 individuals older than 80 years at recruitment were excluded. The latter because diagnoses in the
354 elderly people are often related to significant risk-altering comorbidities (cancer or cardiovascular
355 diseases). Associations of body mass index, three cardiometabolic (coronary artery disease,
356 hypertension, type 2 diabetes) and four autoimmune (bipolar disorder, Crohn's disease, rheumatoid
357 arthritis, type 1 diabetes) diseases were analyzed in 36,716 Estonians (S2 Table).

358 **Single variants analysis**

359 Single variant analysis was conducted with Hail 0.1 (http://broadinstitute.github.io/picard/). Linear
360 regression was used to test each variant's allelic dosage additive effect with body mass index, and
361 Firth [63] logistic regression with seven diseases. Models were adjusted for age, sex, first ten principal
362 components (PC1-10), and genotype array. Only confidently imputed variants (INFO > 0.8) with MAF
363 > 0.05% were considered. A multiple testing corrected significance level ($5 \times 10^{-8}$ / 8 phenotypes) =
364 $6.25 \times 10^{-9}$ were used.

365 All genome-wide significant loci were visualized by regional association plots using LocusZoom
366 v0.4.8 [64] with the 1000G phase 3 European population LD reference panel. Pairwise examination of
367 quantile-quantile plots of GWAS *P* values indicated that the distribution of the test statistics were
368 nearly identical for both datasets, and did not demonstrate significant genomic inflation (S6 Fig). All
369 significantly associated loci were compared to the National Human Genome Research Institute
370 (NHGRI-EBI) GWAS Catalog [1] (April 10, 2018) data.

371 **Fine-mapping**

372 To identify causal variants that denote molecular mechanisms behind the associations, we performed
373 fine-mapping analysis using FINEMAP v1.3 [65] around (± 500 kilobase (kb)) genome-wide
374 significant loci detected by variant-wise analysis (Table 1). FINEMAP was applied with default
375 parameters, allowing for at most five causal variants and the highest posterior probability for the
376 number of causal signals was used.

377 **Enrichment analysis**

378 Enriched variants in the EstFin imputed data were detected in comparison of 503 European (EUR)
379 individuals from the 1000G phase 3 data. Enrichment rates were calculated as MAF in Estonians (Est)
380 divided by MAF in 1000G EUR individuals:

381 $$\text{Enrichment} = \frac{\text{MAF}_{\text{Est}}}{\text{MAF}_{\text{1000G_EUR}}}.$$

382 Corresponding Bonferroni corrected significance level for variants enriched 10-fold in Estonians
383 (enrichment > 10) was [0.05 / (191,099 variants × 8 phenotypes)] = $3.27 \times 10^{-8}$.

384 **Gene-based analysis**

385 To determine the joint contribution of rare variants on eight complex traits, we implemented gene-
386 based SKAT-O [27] tests with EPACTS (https://genome.sph.umich.edu/wiki/EPACTS). All variants
387 were annotated with EPACTS module 'anno' (GENCODE v14 [66]). Only nonsynonymous
388 (nonsynonymous, normal splice site or stop gain) and loss-of-function (stop gain, essential splice site
389 or frameshift) variants with INFO > 0.8 and 0.000001% < MAF < 1% were included. Models were
390 adjusted for age, sex, PC1-10 and genotype arrays. Results were post-filtered that each gene contained

391 at least two NS or LoF variants. We identified 12,930 (663) and 1,274 (9) NS (LoF) genes in the
392 EstFin and the 1000G IRP-based imputed data, respectively. Bonferroni corrected significance levels
393 based on the number of identified genes in both imputed datasets were used: [0.05 / (12,951 × 8
394 phenotypes)] = $4.83 \times 10^{-7}$ and [0.05 / (665 × 8 phenotypes)] = $9.40 \times 10^{-6}$ for NS and LoF genes,
395 respectively.

396 **UK Biobank data**

397 The UK Biobank enrolled about 500,000 people aged between 40-69 years in 2006-2010 from across
398 the United Kingdom [67]. Two approaches were used to validate significant gene-trait associations
399 from the UKBB data:

400 1) We used genotyped and imputed individual-level data as released by UK Biobank in March 2018
401 [39,40]. In the analysis we used 405,379 unrelated (IBD sharing < 0.2) individuals with European
402 origin and confidently imputed variants (INFO > 0.8). Diagnosis of prevalent disease was based on
403 International Classification of Diseases (ICD-10) diagnosis codes and self-reported data. The same
404 SKAT-O models were applied as used to discover 52 significant gene-trait associations in the
405 Estonian Biobank data (Table 2).

406 2) We used GWAS analysis results of the UK Biobank in 361,194 individuals provided by the Neale
407 lab [41] and selected variant-trait results with the lowest *P* value for each significant gene-trait
408 association (gene ± 5 kb) detected in the Estonian Biobank data (Table 2).

409 ## Acknowledgments

410 This research is financially supported by EU H2020 grant 692145, the European Regional
411 Development Fund, Center of Excellence in Genomics GENTRANSMED (Project No. 2014-
412 2020.4.01.15-0012), the Estonian Research Council Grants PUTJD817, IUT20-60, OUT1665P. We
413 would like to acknowledge the International SISu Project for sharing the Finnish anonymised
414 imputation reference panel data. We would like to acknowledge the High Performance Computing
415 Center of the University of Tartu. PP and KP were supported by the NIASC - Nordic Information for
416 Action eScience Center (a Nordic Center of Excellence; financed by NordForsk; Project No. 62721)
417 grant to AP and SR. We would like to thank William Rayner for providing necessary files for

genotype data preparation in his website. We would like to thank the UK Biobank for sharing the data (application No. 17085).

## References


1. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. Oxford University Press; 2019;47: D1005–D1012. doi:10.1093/nar/gky1120
2. Li Y, Willer C, Sanna S, Abecasis G. Genotype Imputation. Annu Rev Genomics Hum Genet. 2009;10: 387–406. doi:10.1146/annurev.genom.9.081307.164242
3. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet. Nature Publishing Group; 2007;39: 906–913. doi:10.1038/ng2088
4. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. Nature Publishing Group; 2010;11: 499–511. doi:10.1038/nrg2796
5. Liu JZ, Tozzi F, Waterworth DM, Pillai SG, Muglia P, Middleton L, et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nat Genet. Nature Publishing Group; 2010;42: 436–440. doi:10.1038/ng.572
6. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. Nature Publishing Group; 2015;526: 68–74. doi:10.1038/nature15393
7. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. Nature Publishing Group; 2016;48: 1279–1283. doi:10.1038/ng.3643
8. Saunders EJ, Dadaev T, Leongamornlert DA, Jugurnauth-Little S, Tymrakiewicz M, Wiklund F, et al. Fine-Mapping the HOXB Region Detects Common Variants Tagging a Rare Coding Allele: Evidence for Synthetic Association in Prostate Cancer. Gibson G, editor. PLoS Genet. Public Library of Science; 2014;10: e1004129. doi:10.1371/journal.pgen.1004129
9. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. Nature Publishing Group; 2010;11: 446–450. doi:10.1038/nrg2809

10. Mitt M, Kals M, Pärn K, Gabriel SB, Lander ES, Palotie A, et al. Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. Eur J Hum Genet. Nature Publishing Group; 2017;25: 869–876. doi:10.1038/ejhg.2017.51

11. Pistis G, Porcu E, Vrieze SI, Sidore C, Steri M, Danjou F, et al. Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. Eur J Hum Genet. Macmillan Publishers Limited; 2015;23: 975–983. doi:10.1038/ejhg.2014.216

12. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, et al. Large-scale whole-genome sequencing of the Icelandic population. Nat Genet. Nature Publishing Group; 2015;47: 435–444. doi:10.1038/ng.3247

13. Deelen P, Menelaou A, Van Leeuwen EM, Kanterakis A, Van Dijk F, Medina-Gomez C, et al. Improved imputation quality of low-frequency and rare variants in European samples using the "Genome of the Netherlands." Eur J Hum Genet. 2014;22: 1321–1326. doi:10.1038/ejhg.2014.19

14. Huang J, Howie B, McCarthy S, Memari Y, Walter K, Min JL, et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. Nat Commun. 2015;6. doi:10.1038/ncomms9111

15. Sidore C, Busonero F, Maschio A, Porcu E, Naitza S, Zoledziewska M, et al. Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. Nat Genet. 2015;47: 1272–1281. doi:10.1038/ng.3368

16. Zhou W, Fritsche LG, Das S, Zhang H, Nielsen JB, Holmen OL, et al. Improving power of association tests using multiple sets of imputed genotypes from distributed reference panels. Genet Epidemiol. 2017;41: 744–755. doi:10.1002/gepi.22067

17. Lin Y, Liu L, Yang S, Li Y, Lin D, Zhang X, et al. Genotype imputation for Han Chinese population using Haplotype Reference Consortium as reference. Hum Genet. 2018;137: 431–436. doi:10.1007/s00439-018-1894-z

18. Holm H, Gudbjartsson DF, Sulem P, Masson G, Helgadottir HT, Zanon C, et al. A rare variant

in MYH6 is associated with high risk of sick sinus syndrome. Nat Genet. 2011;43: 316–323. doi:10.1038/ng.781

19. Nioi P, Sigurdsson A, Thorleifsson G, Helgason H, Agustsdottir AB, Norddahl GL, et al. Variant *ASGR1* Associated with a Reduced Risk of Coronary Artery Disease. N Engl J Med. 2016;374: 2131–2141. doi:10.1056/NEJMoa1508419

20. Jonsson T, Stefansson H, Steinberg S, Jonsdottir I, Jonsson P V., Snaedal J, et al. Variant of *TREM2* Associated with the Risk of Alzheimer's Disease. N Engl J Med. Massachusetts Medical Society ; 2013;368: 107–116. doi:10.1056/NEJMoa1211103

21. Helgason H, Sulem P, Duvvari MR, Luo H, Thorleifsson G, Stefansson H, et al. A rare nonsynonymous sequence variant in C3 is associated with high risk of age-related macular degeneration. Nature Genetics. 2013. pp. 1371–1376. doi:10.1038/ng.2740

22. Steinthorsdottir V, Thorleifsson G, Sulem P, Helgason H, Grarup N, Sigurdsson A, et al. Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. Nat Genet. 2014;46: 294–298. doi:10.1038/ng.2882

23. Styrkarsdottir U, Thorleifsson G, Sulem P, Gudbjartsson DF, Sigurdsson A, Jonasdottir A, et al. Nonsense mutation in the LGR4 gene is associated with several human diseases and other traits. Nature. 2013;497: 517–520. doi:10.1038/nature12124

24. Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, et al. The UK10K project identifies rare variants in health and disease. Nature. 2015;526: 82–89. doi:10.1038/nature14962

25. Timpson NJ, Walter K, Min JL, Tachmazidou I, Malerba G, Shin SY, et al. A rare variant in APOC3 is associated with plasma triglyceride and VLDL levels in Europeans. Nat Commun. 2014;5. doi:10.1038/ncomms5871

26. Moutsianas L, Agarwala V, Fuchsberger C, Flannick J, Rivas MA, Gaulton KJ, et al. The Power of Gene-Based Rare Variant Methods to Detect Disease-Associated Variation and Test Hypotheses About Complex Disease. PLoS Genet. 2015;11. doi:10.1371/journal.pgen.1005165

27. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control

whole-exome sequencing studies. Am J Hum Genet. 2012;91: 224–237. doi:10.1016/j.ajhg.2012.06.007

28. Li B, Leal SM. Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. Am J Hum Genet. 2008;83: 311–321. doi:10.1016/j.ajhg.2008.06.024

29. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol. 2010;34: 188–193. doi:10.1002/gepi.20450

30. Bomba L, Walter K, Soranzo N. The impact of rare and low-frequency genetic variants in common disease. Genome Biol. BioMed Central; 2017;18: 77. doi:10.1186/s13059-017-1212-4

31. Singh T, Kurki MI, Curtis D, Purcell SM, Crooks L, McRae J, et al. Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. Nat Neurosci. 2016;19: 571–577. doi:10.1038/nn.4267

32. Do R, Stitziel NO, Won HH, Jørgensen AB, Duga S, Merlini PA, et al. Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. Nature. 2015;518: 102–106. doi:10.1038/nature13917

33. Ganna A, Satterstrom FK, Zekavat SM, Das I, Kurki MI, Churchhouse C, et al. Quantifying the Impact of Rare and Ultra-rare Coding Variation across the Phenotypic Spectrum. Am J Hum Genet. 2018;102: 1204–1211. doi:10.1016/j.ajhg.2018.05.002

34. Purcell SM, Moran JL, Fromer M, Ruderfer D, Solovieff N, Roussos P, et al. A polygenic burden of rare disruptive mutations in schizophrenia. Nature. 2014;506: 185–190. doi:10.1038/nature12975

35. Fritsche LG, Igl W, Bailey JNC, Grassmann F, Sengupta S, Bragg-Gresham JL, et al. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. Nat Genet. 2016;48: 134–143. doi:10.1038/ng.3448

36. Teslovich TM, Kim DS, Yin X, Stančáková A, Jackson AU, Wielscher M, et al. Identification of seven novel loci associated with amino acid levels using single-variant and gene-based tests in 8545 Finnish men from the METSIM study. Hum Mol Genet. 2018;27: 1664–1674. doi:10.1093/hmg/ddy067

37. Wessel J, Chu AY, Willems SM, Wang S, Yaghootkar H, Brody JA, et al. Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. Nat Commun. Nature Publishing Group; 2015;6: 5897. doi:10.1038/ncomms6897

38. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447: 661–678. doi:10.1038/nature05911

39. UK Biobank Phasing and Imputation Documentation. 2015. Available from: http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/imputation_documentation_May2015.pdf. Cited 10 September 2018.

40. UK Biobank Genotyping and Imputation Data Release. 2018. Available from: http://www.ukbiobank.ac.uk/wp-content/uploads/2018/03/UKB-Genotyping-and-Imputation-Data-Release-FAQ-v3-2.pdf. Cited 10 September 2018.

41. Neale lab. Updated (second) round of GWAS results of the UK Biobank. 2018; Available from: http://www.nealelab.is/uk-biobank/. Cited 10 September 2018.

42. Smyth DJ, Plagnol V, Walker NM, Cooper JD, Downes K, Yang JHM, et al. Shared and distinct genetic variants in type 1 diabetes and celiac disease. N Engl J Med. Massachusetts Medical Society; 2008;359: 2767–2777. doi:10.1056/NEJMoa0807917

43. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. Nature Publishing Group; 2016;536: 285–291. doi:10.1038/nature19057

44. Leitsalu L, Haller T, Esko T, Tammesoo ML, Alavere H, Snieder H, et al. Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. Int J Epidemiol. 2015;44: 1137–1147. doi:10.1093/ije/dyt268

45. Borodulin K, Tolonen H, Jousilahti P, Jula A, Juolevi A, Koskinen S, et al. Cohort Profile: The National FINRISK Study. Int J Epidemiol. 2017;47: 696–696i. doi:10.1093/ije/dyx239

46. Borodulin K, Vartiainen E, Peltonen M, Jousilahti P, Juolevi A, Laatikainen T, et al. Forty-year trends in cardiovascular risk factors in Finland. Eur J Public Health. 2015;25: 539–546. doi:10.1093/eurpub/cku174

47. Mähönen M, Jula A, Harald K, Antikainen R, Tuomilehto J, Zeller T, et al. The validity of heart failure diagnoses obtained from administrative registers. Eur J Prev Cardiol. SAGE PublicationsSage UK: London, England; 2013;20: 254–259. doi:10.1177/2047487312438979

48. Sund R. Quality of the Finnish Hospital Discharge Register: a systematic review. Scand J Public Health. SAGE PublicationsSage UK: London, England; 2012;40: 505–515. doi:10.1177/1403494812456637

49. Pajunen P, Koukkunen H, Ketonen M, Jerkkola T, Immonen-Raiha P, Karja-Koskenkari P, et al. The validity of the Finnish Hospital Discharge Register and Causes of Death Register data on coronary heart disease. Eur J Cardiovasc Prev Rehabil. SAGE PublicationsSage UK: London, England; 2005;12: 132–137. doi:10.1097/01.hjr.0000140718.09768.ab

50. Tolonen H, Salomaa V, Torppa J, Sivenius J, Immonen-Räihä P, Lehtonen A, et al. The validation of the Finnish Hospital Discharge Register and Causes of Death Register data on stroke diagnoses. Eur J Cardiovasc Prev Rehabil. SAGE PublicationsSage UK: London, England; 2007;14: 380–385. doi:10.1097/01.hjr.0000239466.26132.f2

51. Headache Classification Committee of the International Headache Society (IHS). The International Classification of Headache Disorders, 3rd edition (beta version). Cephalalgia. 2013;33: 629–808. doi:10.1177/0333102413485658

52. Eskelinen S, Sailas E, Joutsenniemi K, Holi M, Suvisaari J. Clozapine use and sedentary lifestyle as determinants of metabolic syndrome in outpatients with schizophrenia. Nord J Psychiatry. 2015;69: 339–345. doi:10.3109/08039488.2014.983544

53. Eskelinen S, Sailas E, Joutsenniemi K, Holi M, Koskela TH, Suvisaari J. Multiple physical healthcare needs among outpatients with schizophrenia: findings from a health examination study. Nord J Psychiatry. 2017;71: 448–454. doi:10.1080/08039488.2017.1319497

54. Aromaa A, Koskinen S, editors. Health and functional capacity in Finland. Baseline results of the Health 2000 health examination survey. Publications of the National Public Health Institute: Helsinki, Finland; 2004. Available from: http://urn.fi/URN:NBN:fi-fe201204193452. Cited 10 September 2018.

55. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform.

Bioinformatics. 2010;26: 589–595. doi:10.1093/bioinformatics/btp698

56. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20: 1297–1303. doi:10.1101/gr.107524.110

57. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Current Protocols in Bioinformatics. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2013. p. 11.10.1-11.10.33. doi:10.1002/0471250953.bi1110s43

58. Li H, Wren J. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 2014. pp. 2843–2851. doi:10.1093/bioinformatics/btu356

59. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007;81: 559–575. doi:10.1086/519795

60. Loh P-R, Palamara PF, Price AL. Fast and accurate long-range phasing in a UK Biobank cohort. Nat Genet. Nature Publishing Group; 2016;48: 811–816. doi:10.1038/ng.3571

61. Loh P-R, Danecek P, Palamara PF, Fuchsberger C, A Reshef Y, K Finucane H, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. Nature Publishing Group; 2016;48: 1443–1448. doi:10.1038/ng.3679

62. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet. 2008;84: 210–223. doi:10.1016/j.ajhg.2009.01.005

63. Firth D. Bias reduction of maximum likelihood estimates. Biometrika. 1993;80: 27–38. doi:10.1093/biomet/80.1.27

64. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 2011. pp. 2336–2337. doi:10.1093/bioinformatics/btq419

65. Benner C, Spencer CCA, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics.

2016;32: 1493–1501. doi:10.1093/bioinformatics/btw018

66. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for the ENCODE project. Genome Res. 2012;22: 1760–1774. doi:10.1101/gr.135350.111

67. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med. 2015;12. doi:10.1371/journal.pmed.1001779

## Supporting information

**S1 Fig. Number of confidently imputed variants.** Venn diagrams of confidently imputed variants (INFO > 0.8) using EstFin (blue) and 1000G (orange) IRPs in four minor allele frequency categories. **(A)** 0.05% < MAF ≤ 0.5%, **(B)** 0.5% < MAF ≤ 1%, **(C)** 1% < MAF ≤ 5%, **(D)** MAF > 5%.

**S2 Fig. Distributions of imputation INFO-values.** Distribution of imputation INFO-values for the EstFin and the 1000G IRPs imputed data measured on chromosome 20 in four minor allele frequency categories. **(A)** EstFin IRP, **(B)** 1000G IRP.

**S3 Fig. Single variant association analysis results.** The top panel of Miami plot shows the single variant association analysis results using the EstFin-based imputed data, while the bottom part shows GWAS results for the 1000G IRP imputed data. Red dots denote significant regions and the genome-wide significance threshold after correction for multiple testing ($P < 6.25 \times 10^{-9}$) is indicated by a red dashed line.

**S4 Fig. Significant loci from single variant association analysis.** Regional plots show all significant genomic regions from the single variant analysis. The left panel indicates results for the EstFin-based imputed data, while the right panel shows results for the 1000G IRP imputed data. The purple symbol represents the lead variant, and the rest of the colour-coded variants denote LD with the lead variant estimated by $r^2$ from the 1000G phase 3 (EUR population) data.

**S5 Fig. Gene-based association analysis results.** The top panel of Miami plot shows the gene-based association analysis results using the EstFin-based imputed data, while the bottom part shows results for the 1000G IRP imputed data. Blue dots represent tested genes including NS variants and orange asterisks LoF variants. Dashed lines indicate significance levels after correction for multiple testing: $P < 4.83 \times 10^{-7}$ for NS variants (blue) and $P < 9.40 \times 10^{-6}$ for LoF variants (orange). Red symbols denote significant genes.

**S6 Fig. Quantile-quantile plots for single variant association analysis.** Quantile-quantile plots of the GWAS *P* values based on the EstFin (left panel, blue dots) and the 1000G (right panel, purple dots) IRPs imputed data. Region in gray dashed lines is the 95% confidence band.

**S1 Table. Number of imputed variants.** Number of overall, well-imputed (INFO > 0.4) and confidently imputed (INFO > 0.8) variants with the EstFin and the 1000G IRPs. The last column indicates confidently imputed variants common for both IRP-based imputations.

**S2 Table. An overview of seven complex diseases.** ICD-10 diagnosis codes, number of cases and controls for seven complex diseases in 36,716 Estonian Biobank individuals used in the association analysis.

**S3 Table. Previously known associations for significant single variant analysis results.** Relevant genome-wide associations from the GWAS Catalog for significant genomic regions (around genes (± 50 kb) with the lowest *P* value) detected by variant-wise analysis. Gray background refers to direct relationship between studied trait and GWAS Catalog entry.

**S4 Table. Summary statistics of lead variants at significant loci identified in single variant GWAS results.** Confidently imputed variants (INFO > 0.8) are tested for associations with complex traits (BMI – body mass index, CAD – coronary artery disease, RA – rheumatoid arthritis, T1D – type 1 diabetes, T2D – type 2 diabetes). Analyses are conducted separately for the EstFin and the 1000G IRP-based imputed datasets. Multiple testing corrected significance level ($P < 6.25 \times 10^{-9}$) is used. For each significant loci, lead variant with single variant GWAS summary statistics are provided.

**S5 Table. Overview of genetic variants involved in significant gene-trait associations.** A list of all single variants involved in significant gene-wise associations with single variant GWAS summary

statistics. In the last column, relevant references are provided, where particular gene-trait association is previously identified.

**S6 Table. Previously known associations for significant gene-based analysis results.** Relevant genome-wide associations from the GWAS Catalog for significant genes (± 50 kb) detected by gene-wise analysis. Gray background refers to direct relationship between studied trait and GWAS Catalog entry.

**S1 Appendix. Overview of genotype imputation and examples of disease associated genes.** A detailed overview of genotype imputation with the EstFin and the 1000G IRPs are provided. Five examples of significant gene-trait associations from the EstFin-based imputation results with potential underlying biological mechanisms are considered in more details.