Exome sequencing of Finnish isolates enhances rare-variant association power

Locke, Adam E[1,2,3,*]; Steinberg, Karyn Meltz[2,4,*]; Chiang, Charleston WK[5,6,7,*]; Service, Susan K[5,*]; Havulinna, Aki S[8,9]; Stell, Laurel[10]; Pirinen, Matti[8,11,12]; Abel, Haley J[2,13]; Chiang, Colby C[2]; Fulton, Robert S[2]; Jackson, Anne U[3]; Kang, Chul Joo[2]; Kanchi, Krishna L[2]; Koboldt, Daniel C[2,14,15]; Larson, David E[2,13]; Nelson, Joanne[2]; Nicholas, Thomas J[2,16]; Pietilä, Arto[9]; Ramensky, Vasily[5,17]; Ray, Debashree[3,18]; Scott, Laura J[3]; Stringham, Heather M[3]; Vangipurapu, Jagadish[19]; Welch, Ryan[3]; Yajnik, Pranav[3]; Yin, Xianyong[3]; Eriksson, Johan G[20,21,22]; Ala-Korpela, Mika[23,24,25,26,27,28]; Järvelin, Marjo-Riitta[29,30,31,32,33]; Männikkö, Minna[30,34]; Laivuori, Hannele[8,35,36]; FinnGen Project; Dutcher, Susan K[2,13]; Stitziel, Nathan O[2,37]; Wilson, Richard K[2,14,15]; Hall, Ira M[1,2]; Sabatti, Chiara[10,38]; Palotie, Aarno[8,39,40]; Salomaa, Veikko[9]; Laakso, Markku[19,41]; Ripatti, Samuli[8,11,40]; Boehnke, Michael[3,†]; Freimer, Nelson B[5,†]

[1]Department of Medicine, Washington University School of Medicine, St. Louis, MO
[2]McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO
[3]Department of Biostatistics and Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI
[4]Department of Pediatrics, Washington University School of Medicine, St. Louis, MO
[5]Center for Neurobehavioral Genetics, Jane and Terry Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles, Los Angeles, CA
[6]Center for Genetic Epidemiology, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA
[7]Quantitative and Computational Biology Section, Department of Biological Sciences, University of Southern California, Los Angeles, CA
[8]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland
[9]National Institute for Health and Welfare, Helsinki, Finland
[10]Department of Biomedical Data Science, Stanford University, Stanford, CA
[11]Department of Public Health, University of Helsinki, Helsinki, Finland;
[12]Helsinki Institute for Information Technology HIIT and Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland
[13]Department of Genetics, Washington University School of Medicine, St. Louis, MO
[14]The Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH
[15]Department of Pediatrics, The Ohio State University College of Medicine, Columbus, OH
[16]USTAR Center for Genetic Discovery and Department of Human Genetics, University of Utah, Salt Lake City, UT
[17]Federal State Institution "National Medical Research Center for Preventive Medicine" of the Ministry of Healthcare of the Russian Federation, Moscow, Russia
[18]Departments of Epidemiology and Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD
[19]Institute of Clinical Medicine, Internal Medicine, University of Eastern Finland, Kuopio, Finland

[20]Department of Public Health Solutions, National Institute for Health and Welfare, Helsinki, Finland
[21]Folkhälsan Research Center, Helsinki, Finland
[22]Department of General Practice and Primary Health Care, University of Helsinki, Helsinki and Helsinki University Hospital, Helsinki, Finland
[23]Systems Epidemiology, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia
[24]Computational Medicine, Faculty of Medicine, University of Oulu and Biocenter Oulu, University of Oulu, Oulu, Finland
[25]NMR Metabolomics Laboratory, School of Pharmacy, University of Eastern Finland, Kuopio, Finland
[26]Population Health Science, Bristol Medical School, University of Bristol, Bristol, UK
[27]Medical Research Council Integrative Epidemiology Unit at the University of Bristol, Bristol, UK
[28]Department of Epidemiology and Preventive Medicine, School of Public Health and Preventive Medicine, Faculty of Medicine, Nursing and Health Sciences, The Alfred Hospital, Monash University, Melbourne, Victoria, Australia
[29]Biocenter Oulu, University of Oulu, Oulu, Finland
[30]Center for Life Course Health Research, Faculty of Medicine, University of Oulu, Oulu, Finland
[31]Unit of Primary Health Care, Oulu University Hospital, Oulu, Finland
[32]Department of Epidemiology and Biostatistics, MRC-PHE Centre for Environment and Health, School of Public Health, Imperial College London, London, UK
[33]Department of Life Sciences, College of Health and Life Sciences, Brunel University London, Uxbridge, UK
[34]Northern Finland Birth Cohorts, Faculty of Medicine, University of Oulu, Oulu, Finland
[35]Medical and Clinical Genetics, University of Helsinki and Helsinki University Hospital, Helsinki, Finland
[36]Department of Obstetrics and Gynecology, Tampere University Hospital and University of Tampere, Faculty of Medicine and Life Sciences, Tampere, Finland
[37]Cardiovascular Division, Department of Medicine, Washington University School of Medicine, St. Louis, MO
[38]Department of Statistics, Stanford University, Stanford, CA
[39]Analytical and Translational Genetics Unit (ATGU), Psychiatric & Neurodevelopmental Genetics Unit, Departments of Psychiatry and Neurology, Massachusetts General Hospital, Boston, MA
[40]Broad Institute of MIT and Harvard, Cambridge, MA
[41]Department of Medicine, Kuopio University Hospital, Kuopio, Finland

[*]These authors contributed equally to this work.
[†]These authors jointly supervised this work.

ABSTRACT

Exome sequencing studies have generally been underpowered to identify deleterious alleles with a large effect on complex traits, as such alleles are mostly rare. Because the population of northern and eastern Finland has expanded dramatically and in isolation following a series of bottlenecks, it harbors numerous deleterious alleles at relatively high frequency. Capitalizing on this circumstance, we exome sequenced nearly 20,000 individuals from these regions. Exome-wide association studies for 64 quantitative traits clinically relevant to cardiovascular and metabolic disease identified 26 newly associated deleterious alleles. Nineteen of these alleles are either unique to or >20 times more frequent in Finns than in other Europeans and show geographical clustering comparable to Mendelian disease mutations characteristic of the Finnish population. We estimate that sequencing studies in populations without this unique history would require hundreds of thousands to millions of participants to achieve comparable association power.

INTRODUCTION

Most alleles with a demonstrated deleterious effect on phenotypes directly alter protein structure or function[1,2]. Exome sequencing studies aim to discover such alleles and demonstrate their association to common diseases and disease-related quantitative traits. However, exome sequencing studies to date generally have identified few newly associated rare variants or genes[3,4]. The sample size required for such discoveries remains uncertain and theoretical analyses indicate that studies to date have been underpowered, since most deleterious variants are expected to be rare due to purifying selection[5]. These previous analyses also suggest that power to detect associations to deleterious alleles is greatest in

110 populations that have expanded in isolation after recent bottlenecks, as alleles passing
111 through the bottlenecks may rise to much higher frequencies than in other populations[6-8].
112
113 Finland exemplifies such a history. Bottlenecks occurred at the founding of early-
114 settlement regions (southern and western Finland) 2,000-4,000 years ago and again with
115 internal migration to late-settlement regions (northern and eastern Finland) in the 15th and
116 16th centuries[9]. Finland's subsequent population growth (to ~5.5 million) generated sizable
117 geographic sub-isolates in late-settlement regions.
118
119 This unique population history has resulted in "the Finnish Disease Heritage"[10], 36
120 Mendelian diseases that are much more common in Finns than in other Europeans. These
121 disorders concentrate in late-settlement regions of Finland[10], and the genes responsible for
122 them exhibit extreme enrichment of deleterious variants[11-13]. We created the FinMetSeq
123 study to capitalize on the population history of late-settlement Finland to discover rare-
124 variant associations with cardiovascular and metabolic disease-relevant quantitative traits
125 through exome sequencing of two extensively phenotyped population cohorts, FINRISK
126 and METSIM (Methods).
127
128 We successfully sequenced 19,292 FinMetSeq participants and tested the identified
129 variants for association with 64 clinically relevant quantitative traits, discovering 43 novel
130 associations with deleterious variants[14,15]: 19 associations (11 traits) in FinMetSeq alone
131 and 24 associations (20 traits) in a combined analysis of FinMetSeq with 24,776 Finns
132 from three cohorts with imputed genome-wide genotypes. Nineteen of the 26 variants

underlying these 43 associations were unique to Finland or enriched >20-fold in FinMetSeq compared to non-Finnish Europeans (NFE). These enriched alleles cluster geographically like Finnish Disease Heritage mutations, indicating that the distribution of trait-associated rare alleles may vary significantly between locations within a country.

We demonstrate that exome sequencing in a historically isolated population that expanded after recent population bottlenecks is an extraordinarily efficient strategy to discover alleles with a substantial effect on quantitative traits. As most of the novel, putatively deleterious trait-associated variants that we identified are unique to or highly enriched in Finland, we estimate that similarly powered studies of these variants in non-Finnish populations might require hundreds of thousands or millions of participants.

RESULTS

Genetic variation

In 19,292 successfully sequenced exomes, we identified 1,318,781 single nucleotide variants (SNVs) and 92,776 insertion/deletion (indel) variants (**Supplementary Tables 1-3, Supplementary Information**). Compared to NFE control exomes (gnomAD v2.1, **Extended Data Fig. 1A**), FinMetSeq exomes showed depletion of singletons and doubletons and excess variants with minor allele count (MAC)≥5, particularly for predicted-deleterious alleles (**Extended Data Fig. 1B**).

Association analyses

We tested for association between genetic variants in FinMetSeq and 64 clinically relevant quantitative traits after standard adjustments for medications and covariates and

157 transformation to normality for analyses (Methods, **Supplementary Tables 4 & 5**). Sixty-
158 two of 64 traits exhibited significant heritability with common SNVs ($P<0.05$;
159 $5\%<h^2<53\%$; **Extended Data Fig. 2A**, **Supplementary Table 6**), with substantial
160 phenotypic and genetic correlations between traits (**Extended Data Fig. 2B**).
161
162 Single-variant association tests with genetic variants with MAC≥3 among the 3,558 to
163 19,291 individuals measured for each trait (**Supplementary Tables 4 & 5**) identified 1,249
164 associations ($P<5\times10^{-7}$) at 531 variants (**Supplementary Table 7**); 53 traits associated
165 with ≥1 variant (**Fig. 1A**). All 1,249 associations remained significant after multiple testing
166 adjustment (exome-wide and across the 64 traits using a hierarchical procedure setting
167 average FDR at 5%, Methods). Using this procedure on the 531 associated variants, we
168 detected 287 more associations (**Supplementary Table 8**), most reflecting high correlation
169 between lipid traits. Of the 531 variants, those at >10x frequency in FinMetSeq compared
170 to NFE were more likely to be trait-associated (OR=4.92, $P=2.6\times10^{-5}$; **Extended Data Fig.**
171 **1C**).
172
173 After clumping associated variants within 1Mbp and with $r^2>0.5$ into single loci (Methods),
174 the 531 associated variants represented 262 distinct loci (597 trait-locus pairs,
175 **Supplementary Table 7**). The number of associated loci per trait correlated positively
176 with trait heritability ($r=0.38$, $P=8.8\times10^{-4}$), with height a notable outlier (**Fig. 1B**).
177
178 Most variants and loci (61%) associated to a single trait; 4% associated to ≥10 traits.
179 Overlapping associations (**Extended Data Fig. 3A**) reflect both phenotypic and genetic

correlations and the estimated genetic correlation of trait pairs predicts shared loci between traits (**Extended Data Fig. 3B**). Gene-based association tests revealed 54 associations with $P<3.88\times10^{-6}$ and multi-trait FDR<0.05 (Methods, **Supplementary Table 9**), including ten traits associated with *APOB* (**Extended Data Fig. 4)** and a novel association of *SECTM1* with HDL2-C (**Extended Data Fig. 5**).

To determine which of the 1,249 single-variant associations are distinct from previous GWAS findings, we repeated association analysis for each trait conditioning on published associated variants in the EBI GWAS Catalog (December 2016, Methods); 478 associations at 126 loci remained significant ($P<5\times10^{-7}$), including at least one association for 48 traits (**Supplementary Table 10**). Conditionally-associated variants were more often rare (24% vs. 11%), more likely protein-altering (31% vs. 22%), and more frequently >10x enriched in FinMetSeq relative to NFE (19% vs. 10%) than associated variants overall.

Replication and follow-up

We attempted to replicate the 478 single-variant associations (unconditional and conditional $P\leq5\times10^{-7}$) and follow up 2,120 sub-threshold associations from FinMetSeq (unconditional $5\times10^{-7}<P\leq5\times10^{-5}$ and conditional $P\leq5\times10^{-5}$) in 24,776 participants from three Finnish cohort studies: FINRISK[16,17] participants not in FinMetSeq (n=18,215), Northern Finland Birth Cohort 1966[18] (n=5,139), and Helsinki Birth Cohort[19] (n=1,412), all imputed using the Finnish SISu v2 reference panel (www.sisuproject.fi). Following association analysis within each cohort, we conducted meta-analysis of the three

imputation-based studies to test for replication of FinMetSeq variants ("replication analysis"), and four-study meta-analysis with FinMetSeq to follow up suggestive associations ("combined analysis").

Of 448 significant variant-trait associations with replication data, 392 (87.5%) replicated at $P<0.05$ (**Supplementary Table 11**). Of the 1,417 sub-threshold associations, 431 reached $P<5\times10^{-7}$ in the combined analysis (**Supplementary Table 12**); >60% of variants we could not follow up were absent in the reference panel.

Among the significant associations from FinMetSeq or combined analysis, 43 were with 26 predicted deleterious variants (six PTVs, 20 missense) that conditional analysis and literature review suggest are novel (**Table 1**). Nineteen associations (15 variants) were significant in FinMetSeq (**Table 1**; **Supplementary Table 11**); another 24 associations (16 variants) reached significance in combined analysis (**Table 1**; **Supplementary Table 12**). Of these 43 associations, 34 were with 19 variants either seen only in Finland or enriched >20-fold in FinMetSeq compared to NFE. Identifying associations for these 19 variants would have required much larger samples in NFE populations than in FinMetSeq (**Fig. 2A, B**). We provide brief summaries relating some of these associations to known biology and prior genetic evidence (**Table 1**, expanded version in **Supplementary Table 13**, **Supplementary Information**), highlighting here the most striking findings.

Anthropometric traits. A predicted damaging missense variant (p.Arg94Cys) in *THBS4* 45X more frequent in FinMetSeq than in NFE was associated in the combined analysis

with a mean 5.9 kg decrease in body weight. *THBS4* encodes thrombospondin 4, a matricellular protein found in blood vessel walls and highly expressed in heart and adipose [20]. THBS4 may regulate vascular inflammation[21] and has been implicated in heart disease risk[22].

A predicted damaging missense variant (p.Val104Met) in *DLK1* 177X more frequent in FinMetSeq than in NFE is associated in the combined analysis with a mean 1.3cm decrease in height. *DLK1* encodes Delta-Like Notch Ligand 1, an epidermal growth factor that interacts with fibronectin and inhibits adipocyte differentiation. Uniparental disomy of *DLK1* causes Temple and Kagami-Ogata Syndromes, characterized by growth restriction, hypotonia, joint laxity, motor delay, and early onset of puberty[23]. Paternally-inherited common variants near *DLK1* are associated with childhood obesity, type 1 diabetes, age at menarche, and precocious puberty[24-26]. Homozygous null mutations in the mouse ortholog *Dlk-1* lead to embryos with reduced size, skeletal length, and lean mass[27]; in Darwin's finches, SNVs at this locus have a strong effect on beak size[28].

HDL-C. A predicted deleterious missense variant p.Arg112Trp in *CD300LG* is associated in FinMetSeq with a mean 0.95 mmol/l increase in HDL-C and is associated with increased HDL2-C and ApoA1. This variant, absent in NFE, has an opposite direction of effect from a previously reported deleterious missense variant in this gene[29], which encodes a type I cell surface glycoprotein.

Amino acids. A stop gain variant (p.Arg722X) in *ALDH1L1* is associated in FinMetSeq with reduced serum glycine levels and is absent in NFE; this trait may increase risk for cardiometabolic disorders[30,31]. *ALDH1L1* encodes 10-formyltetrahydrofolate dehydrogenase, which competes with serine hydroxymethyltransferase to alter the ratio of serine to glycine in the cytosol. Gene-based tests suggest additional PTVs and missense variants in *ALDH1L1* alter glycine levels ($P=1.4\times10^{-20}$, **Extended Data Fig. 6**, **Supplementary Table 9**).

Ketone bodies. A predicted damaging missense variant (p.Phe517Ser) in *ACSS1* is associated in the combined analysis with increased serum acetate levels and is absent in NFE. *ACSS1* encodes an acyl-coenzyme A synthetase and plays a role in conversion of acetate to acetyl-CoA. In rodents, increased acetate levels lead to obesity, insulin resistance, and metabolic syndrome[32].

Trait-associations and disease endpoints

Genotype data from FinnGen[33] enabled us to test whether deleterious variants responsible for our novel trait associations contribute to related disease endpoints. We examined 22 diseases for the 25 available variants in **Table 1**; three variants were associated with diseases in FinnGen at Bonferroni threshold $P<0.05/(22\times25)=9.0\times10^{-5}$ (**Supplementary Table 14**).

A predicted damaging missense variant (p.Ser328Pro) in *KRT40*, associated in FinMetSeq with elevated HDL-C, but absent in NFE, is associated in FinnGen with increased

pancreatitis risk. While this is the first disease association reported for *KRT40*, type I keratins regulate exocrine pancreas homoeostasis[34]. A 29bp deletion causing a frameshift in *FAM151A* is associated in FinMetSeq with decreased total cholesterol in IDL and decreased IDL particle concentration, is 6.7X more frequent in FinMetSeq than NFE, and is associated in FinnGen with decreased risk of myocardial infarction. Interpretation of this association is complicated as the variant is also situated in an overlapping gene (*ACOT11*) involved in fatty acid metabolism and lies <1Mbp from a cardioprotective variant in *PCSK9*. Finally, a predicted damaging missense variant (p.Arg65Trp) in *DBH* associated with a mean 1.0 mmHg decrease in diastolic blood pressure in the combined analysis, is 23.8X more frequent in FinMetSeq than in NFE, and is associated in FinnGen with decreased risk for hypertension. Distinct loci in this gene and gene-based tests are associated with mean arterial pressure[35,36].

Replication outside Finland

To assess the generalizability of these novel associations, we attempted to replicate associations from our combined analysis in the UK Biobank (UKB). Across eight anthropometric and blood pressure traits for which UKB data are publicly available, our combined analysis identified 31 trait-variant associations, of which 23 were present in UKB. Twenty of 23 associations were to variants with MAF>1% in FinMetSeq and comparable frequency in UKB; 15 (75%) showed association in UKB at $P<0.05/23=2.2\times10^{-3}$. The three rare variants in this analysis were all >10x more frequent in FinMetSeq than UKB; none were associated in UKB (**Supplementary Table 15**). However, even after adjusting for winner's curse[37], we had <50% power to detect these

associations in UKB, consistent with the argument that extremely large samples will be needed in other populations to achieve the power for rare-variant association studies that we observed in Finland.

Enriched variants cluster geographically

Given the concentration of Finnish Disease Heritage mutations within regions of late-settlement Finland[38], we hypothesized that trait-associated variants discovered through FinMetSeq might also cluster geographically. Principal component analysis supported this hypothesis, revealing broad-scale population structure within late-settlement regions among 14,874 unrelated FinMetSeq participants with known parental birthplaces (**Extended Data Fig. 7**). Carriers of PTVs and missense alleles showed more clustering of parental birthplaces than carriers of synonymous alleles, even after adjusting for MAC (**Supplementary Tables 16A, B**).

To analyze the distribution of variants within late-settlement Finland, we delineated geographically distinct population clusters using haplotype sharing among 2,644 unrelated individuals with both parents born in the same municipality (Methods, **Extended Data Fig. 8**). We compared variant counts across functional classes and frequencies between an early-settlement reference cluster and 12 clusters containing ≥100 individuals (**Extended Data Fig. 9, Supplementary Tables 17, 18**). Clusters representing the most heavily bottlenecked late-settlement regions (Lapland and Northern Ostrobothnia) displayed a deficit of singletons and enrichment of intermediate frequency variants compared to other clusters.

317

318 Variants >10x enriched in FinMetSeq compared to NFE displayed particularly strong
319 geographical clustering (**Supplementary Table 19**). We further characterized clustering
320 for FinMetSeq-enriched trait-associated variants, by comparing mean distances between
321 birthplaces of parents of minor allele carriers to those of non-carriers (**Supplementary**
322 **Table 20**). Most such variants were highly localized. For example, for rs780671030 in
323 *ALDH1L1*, the mean distance between parental birthplaces is 135km for carriers and
324 250km for non-carriers ($P<1.0\times10^{-7}$, **Fig. 3A**).

325

326 Finally, we identified comparable geographic clustering between carriers of 35 Finnish
327 Disease Heritage mutations and carriers of FinMetSeq-enriched trait-associated variants
328 (**Fig. 3B**, Methods). Clustering was dramatically greater than that observed for non-carriers
329 of both sets of variants, suggesting that rare trait-associated variants may be much more
330 unevenly distributed geographically than previously appreciated.

331

332 **DISCUSSION**

333 We demonstrate that a well-powered exome sequencing study of deeply phenotyped
334 individuals can identify numerous rare variants associated with medically relevant
335 quantitative traits. The variants we identified provide a useful starting point for studies
336 aimed at uncovering biological mechanisms and fostering clinical translation. The power
337 of this study to discover rare-variant associations derives from the numerous deleterious
338 variants that are enriched in or unique to Finland. Prioritizing the sequencing of multiple
339 population isolates that have expanded from recent bottlenecks is a strategy for scaling up

the discovery of rare-variant associations[7,39-41]. Because genetic drift results in a different set of alleles to pass through population-specific bottlenecks, enriching some variants and depleting others, the numerous rare-variant associations that could be identified by sequencing well-phenotyped samples across multiple isolates could rapidly increase our understanding of the genetic architecture of complex traits.

Our results support recent suggestions of continuity between the genetic architectures of complex traits and disorders classically considered monogenic[42,43], by identifying numerous deleterious variants with large effects on quantitative traits that demonstrate geographical clustering comparable to that of the mutations responsible for the Finnish Disease Heritage.

Using a Finland-specific reference panel[44] to impute FinMetSeq variants into array-genotyped samples from three other Finnish cohorts enabled us to identify additional novel associations. However, the clustering in FinMetSeq of deleterious trait-associated variants within limited geographical regions and our inability to follow-up >700 sub-threshold associations from FinMetSeq for which the associated variants were absent in the Finnish imputation reference panel, emphasize the importance of representing regional subpopulations in such reference panels, to account for fine-scale population structure.

The value of rare-variant studies in population isolates will depend on the richness of phenotypes in sequenced cohorts from these populations. For example, we associated <100 of the >24,000 deleterious, highly enriched variants identified in FinMetSeq with one of

363 the 64 quantitative traits studied here. The associations we identified to disease endpoints
364 in FinnGen hint at the discoveries that will be possible when that database reaches its full
365 size of 500,000 participants. The insights gained from such efforts will accelerate the
366 implementation of precision health, informing projects in more heterogeneous populations
367 which are still at an early stage[45].

References

1 Samocha, K. E. *et al.* Regional missense constraint improves variant deleteriousness prediction. *bioRxiv*, doi:https://doi.org/10.1101/148353 (2017).

2 Marouli, E. *et al.* Rare and low-frequency coding variants alter human adult height. *Nature* **542**, 186-190, doi:10.1038/nature21039 (2017).

3 Flannick, J. *et al.* Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature* **570**, 71-76, doi:10.1038/s41586-019-1231-2 (2019).

4 Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J. & Richards, J. B. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nature reviews. Genetics* **19**, 110-124, doi:10.1038/nrg.2017.101 (2018).

5 Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci U S A* **111**, E455-464, doi:10.1073/pnas.1322563111 (2014).

6 Xue, Y. *et al.* Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations. *Nature communications* **8**, 15927, doi:10.1038/ncomms15927 (2017).

7 Southam, L. *et al.* Whole genome sequencing and imputation in isolated populations identify genetic associations with medically-relevant complex traits. *Nature communications* **8**, 15606, doi:10.1038/ncomms15606 (2017).

8 Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-753, doi:10.1038/nature08494 (2009).

9 Jakkula, E. *et al.* The genome-wide patterns of variation expose significant substructure in a founder population. *American journal of human genetics* **83**, 787-794, doi:10.1016/j.ajhg.2008.11.005 (2008).

10 Polvi, A. *et al.* The Finnish disease heritage database (FinDis) update-a database for the genes mutated in the Finnish disease heritage brought to the next-generation sequencing era. *Hum Mutat* **34**, 1458-1466, doi:10.1002/humu.22389 (2013).

11 Manning, A. *et al.* A Low-Frequency Inactivating AKT2 Variant Enriched in the Finnish Population Is Associated With Fasting Insulin Levels and Type 2 Diabetes Risk. *Diabetes* **66**, 2019-2032, doi:10.2337/db16-1329 (2017).

12 Lim, E. T. *et al.* Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS genetics* **10**, e1004494, doi:10.1371/journal.pgen.1004494 (2014).

13 Service, S. K. *et al.* Re-sequencing expands our understanding of the phenotypic impact of variants at GWAS loci. *PLoS genetics* **10**, e1004147, doi:10.1371/journal.pgen.1004147 (2014).

14 Wurtz, P. *et al.* Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Large-Scale Epidemiology: A Primer on -Omic Technologies. *American journal of epidemiology* **186**, 1084-1096, doi:10.1093/aje/kwx016 (2017).

15 Laakso, M. *et al.* The Metabolic Syndrome in Men study: a resource for studies of metabolic and cardiovascular diseases. *Journal of lipid research* **58**, 481-493, doi:10.1194/jlr.O072629 (2017).

16 Borodulin, K. *et al.* Forty-year trends in cardiovascular risk factors in Finland. *Eur J Public Health* **25**, 539-546, doi:10.1093/eurpub/cku174 (2015).

17 Abraham, G. *et al.* Genomic prediction of coronary heart disease. *Eur Heart J* **37**, 3267-3278, doi:10.1093/eurheartj/ehw450 (2016).
18 Sabatti, C. *et al.* Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature genetics* **41**, 35-46, doi:10.1038/ng.271 (2009).
19 Pulizzi, N. *et al.* Interaction between prenatal growth and high-risk genotypes in the development of type 2 diabetes. *Diabetologia* **52**, 825-829, doi:10.1007/s00125-009-1291-1 (2009).
20 Fagerberg, L. *et al.* Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics* **13**, 397-406, doi:10.1074/mcp.M113.035600 (2014).
21 Corsetti, J. P. *et al.* Thrombospondin-4 polymorphism (A387P) predicts cardiovascular risk in postinfarction patients with high HDL cholesterol and C-reactive protein levels. *Thromb Haemost* **106**, 1170-1178, doi:10.1160/TH11-03-0206 (2011).
22 Zhang, X. J. *et al.* Association between single nucleotide polymorphisms in thrombospondins genes and coronary artery disease: A meta-analysis. *Thromb Res* **136**, 45-51, doi:10.1016/j.thromres.2015.04.019 (2015).
23 Beygo, J. *et al.* New insights into the imprinted MEG8-DMR in 14q32 and clinical and molecular description of novel patients with Temple syndrome. *Eur J Hum Genet* **25**, 935-945, doi:10.1038/ejhg.2017.91 (2017).
24 Wallace, C. *et al.* The imprinted DLK1-MEG3 gene region on chromosome 14q32.2 alters susceptibility to type 1 diabetes. *Nature genetics* **42**, 68-71, doi:10.1038/ng.493 (2010).
25 Day, F. R. *et al.* Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. *Nature genetics* **49**, 834-841, doi:10.1038/ng.3841 (2017).
26 Perry, J. R. *et al.* Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* **514**, 92-97, doi:10.1038/nature13545 (2014).
27 Cleaton, M. A. *et al.* Fetus-derived DLK1 is required for maternal metabolic adaptations to pregnancy and is associated with fetal growth restriction. *Nature genetics* **48**, 1473-1480, doi:10.1038/ng.3699 (2016).
28 Chaves, J. A. *et al.* Genomic variation at the tips of the adaptive radiation of Darwin's finches. *Mol Ecol* **25**, 5282-5295, doi:10.1111/mec.13743 (2016).
29 Surakka, I. *et al.* The impact of low-frequency and rare variants on lipid levels. *Nature genetics* **47**, 589-597, doi:10.1038/ng.3300 (2015).
30 Ding, Y. *et al.* Plasma Glycine and Risk of Acute Myocardial Infarction in Patients With Suspected Stable Angina Pectoris. *J Am Heart Assoc* **5**, doi:10.1161/JAHA.115.002621 (2015).
31 Wittemans, L. B. L. *et al.* Assessing the causal association of glycine with risk of cardio-metabolic diseases. *Nature communications* **10**, 1060, doi:10.1038/s41467-019-08936-1 (2019).
32 Perry, R. J. *et al.* Acetate mediates a microbiome-brain-beta-cell axis to promote metabolic syndrome. *Nature* **534**, 213-217, doi:10.1038/nature18309 (2016).

33 Tabbassum, R. *et al.* Genetics of human plasma lipidome: Understanding lipid metabolism and its link to diseases beyond traditional lipids. *bioRxiv*, doi:10.1101/457960 (2018).
34 Casanova, M. L. *et al.* Exocrine pancreatic disorders in transsgenic mice expressing human keratin 8. *J Clin Invest* **103**, 1587-1595, doi:10.1172/JCI5343 (1999).
35 Surendran, P. *et al.* Trans-ancestry meta-analyses identify rare and common variants associated with blood pressure and hypertension. *Nature genetics* **48**, 1151-1161, doi:10.1038/ng.3654 (2016).
36 Liu, C. *et al.* Meta-analysis identifies common and rare variants influencing blood pressure and overlapping with metabolic trait loci. *Nature genetics* **48**, 1162-1170, doi:10.1038/ng.3660 (2016).
37 Palmer, C. & Pe'er, I. Statistical correction of the Winner's Curse explains replication variability in quantitative trait genome-wide association studies. *PLoS genetics* **13**, e1006916, doi:10.1371/journal.pgen.1006916 (2017).
38 Norio, R. Finnish Disease Heritage I: characteristics, causes, background. *Hum Genet* **112**, 441-456, doi:10.1007/s00439-002-0875-3 (2003).
39 Service, S. *et al.* Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. *Nature genetics* **38**, 556-560, doi:10.1038/ng1770 (2006).
40 Chiang, C. W. K. *et al.* Genomic history of the Sardinian population. *Nature genetics*, doi:10.1038/s41588-018-0215-8 (2018).
41 Rivas, M. A. *et al.* Insights into the genetic epidemiology of Crohn's and rare diseases in the Ashkenazi Jewish population. *PLoS genetics* **14**, e1007329, doi:10.1371/journal.pgen.1007329 (2018).
42 Bastarache, L. *et al.* Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. *Science* **359**, 1233-1239, doi:10.1126/science.aal4043 (2018).
43 Niemi, M. E. K. *et al.* Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. *Nature*, doi:10.1038/s41586-018-0566-4 (2018).
44 Surakka, I. S., A.-P.; Ruotsalainen, S.E.; Durbin, R.; Salomaa, V.; Daly, M.; Palotie, A.; Ripatti, S. The rate of false polymorphisms introduced when imputing genotypes from global imputation panels. *bioRxiv*, doi:10.1101/080770 (2016).
45 Collins, F. S. & Varmus, H. A new initiative on precision medicine. *N Engl J Med* **372**, 793-795, doi:10.1056/NEJMp1500523 (2015).

Supplementary Information

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements

Thanks to Terri Teshiba for coordinating ethical permissions and samples. Thanks to Sini Kerminen, Daniel Lawson, and George Busby for discussions and providing scripts to run fineSTRUCTURE. SR was supported by the Academy of Finland Center of Excellence in Complex Disease Genetics (Grant No 312062), Academy of Finland (No. 285380), the Finnish Foundation for Cardiovascular Research, the Sigrid Juselius Foundation, Biocentrum Helsinki and University of Helsinki HiLIFE Fellow grant. VR acknowledges support by RFBR, research project No. 18-04-00789 A. VS was supported by the Finnish Foundation for Cardiovascular Research. CS and LS received funding from HG006695, HL113315, and MH105578. MAK is supported by a Senior Research Fellowship from the National Health and Medical Research Council (NHMRC) of Australia (APP1158958). He also works in a unit that is supported by the University of Bristol and UK Medical Research Council (MC_UU_12013/1). The Baker Institute is supported in part by the Victorian Government's Operational Infrastructure Support Program. AUJ, DR, LJS, HMS, RW, PY, XY, and MB received funding from DK062370. SKS, CWKC, and NBF received funding from HL113315 and NS062691. The METSIM study was supported by grants from Academy of Finland (No. 321428), the Sigrid Juselius Foundation, the Finnish Foundation for Cardiovascular Research, Kuopio University Hospital, and Centre of Excellence of Cardiovascular and Metabolic Diseases supported by the Academy of Finland (ML). Sequencing was funded by 5U54HG003079, and AEL, KMS, HJB, CCC, CJK, KLK, DCK, DEL, JN, TJN, SKD, NOS, IMH, and RKW were funded by 5U54HG003079 and 5UM1HG008853-03.

Author Contributions

AEL, LJS, RKW, AaP, VS, ML, SR, MB, and NBF designed the study. AEL, KMS, HJA, RSF, DCK, DEL, JN, TJN, and JV produced and quality-controlled the sequence data. AEL, ASH, AUJ, ArP, HMS, MAK, VS, and ML collected, quality-controlled, and/or prepared the clinical data for association analysis. AEL, KMS, CWKC, SKS, ASH, LS, MP, CCC, AUJ, CJK, KK, VR, DR, JV, RW, PY, and XY analyzed data. ASH, JGE, MAK, MRJ, and MM collected, quality-controlled, and analyzed replication data. HL, SKD, NOS, IMH, CS, SR, MB, and NBF supervised experiments and analyses. AEL, KMS, CWKC, SKS, CS, MB and NBF wrote the paper. AEL, KMS, CWKC, and SKS contributed equally to this work. NBF and MB jointly supervised this work.

Author Information

Competing interests statements:

VS has participated in a conference trip sponsored by Novo Nordisk and received a honorarium from the same source for participating in an advisory board meeting. He also has ongoing research collaboration with Bayer Ltd.

HL is a member of the Nordic Expert group unconditionally supported by Gedeon Richter Nordics and has received an honorarium from Orion.

Correspondence and requests for materials should be addressed to nfreimer@mednet.ucla.edu or boehnke@umich.edu.

Data Availability: The sequence data can be accessed through dbGaP using study numbers phs000756 and phs000752. Association results can be accessed at http://pheweb.sph.umich.edu/FinMetSeq/ and are searchable via the Type 2 Diabetes Knowledge Portal (www.type2diabetesgenetics.org). Summary statistics will also be made available through the NHGRI-EBI GWAS Catalog at https://www.ebi.ac.uk/gwas/downloads/summary-statistics.

Figure Legends

Figure 1. Characterization of associations.

A) Number of genomic loci associated with each trait. Bars are subdivided into common (MAF>1%, dark blue) and rare (MAF≤1%, light blue).

B) Relationship between estimated heritability and number of loci detected per trait. Each trait is colored by trait group. Vertical bars indicate ±2 standard errors. The gray line shows the linear regression fit to indicate the general trend. The number of independent individuals used in each point is listed in **Supplementary Table 5**. Height is the notable outlier.

Figure 2. Allelic enrichment in the Finnish population and its effect on genetic discovery.

A) Relationship between MAF and estimated effect size for associations discovered in FinMetSeq. Each variant reaching significance in FinMetSeq is plotted, with associations in **Table 1** represented by dark blue points (FinMetSeq MAF) and green points (NFE MAF). Purple lines indicate 80% power curves for sample sizes of 10,000 and 20,000 at $\alpha=5x10^{-7}$.

B) Same plot as in A, highlighting the variants in **Table 1** only reaching significance in the combined analysis.

Figure 3. Geographical clustering of associated variants.

A) Example of geographical clustering for a novel trait-associated variant (**Table 1**). The map shows birth locations of all 113 parents of carriers (orange) and 113 randomly selected parents of non-carriers (blue) of the minor allele for rs780671030 in *ALDH1L1*.

B) FDH mutations (N=38) geographically cluster (by parental birthplace) similarly to trait-associated variants (**Table 1**) that are >10x more frequent in FMS than in NFE (N=12) and more than enriched variants from our combined analysis (N=7). For all variants, carriers clustered more than non-carriers (center line, median; box limits, upper and lower quartiles;

581 whiskers, 1.5 interquartile range; points, outliers).

Figure Legends (Extended Data Figures)

Extended Data Fig. 1. Allele frequency comparisons between FinMetSeq and NFE from gnomAD.

A) Distribution of allelic frequencies between FinMetSeq and gnomAD NFE. The comparison of allele frequencies shows the excess of variants at higher frequency in Finland as a result of the multiple bottlenecks experienced in Finnish population history.

B) Proportional site frequency spectra between FinMetSeq and gnomAD NFE by variant annotation class. In general, we find a depletion of the variants in the rarest frequency class, as well as enrichment of variants in the intermediate to common frequency range. The site frequency spectra were down-sampled to 18,000 chromosomes for each dataset.

C) Comparison of MAFs for trait-associated variants in FinMetSeq and NFE gnomAD. Plotted in gray background is a 2-D histogram of variants with non-zero allele frequencies in both gnomAD and FinMetSeq but no trait associations. Variants associated with at least one trait are colored and scaled inversely proportional to the logarithm of the association p-value. Variants >10x enriched in FinMetSeq compared to NFE are pink, those <10x enriched are in blue. The dashed line is the line of equal frequency. Two-sided uncorrected P-values are from a regression of trait on the count of alternative allele at each variant. The number of independent individuals used in each point is listed in **Supplementary Table 5**.

Extended Data Figure 2. Heritability of and correlations between traits. Traits are in the same order, clockwise in A, and left to right and top to bottom in B, following the trait group color key.

A) Heritability estimated in 13,342 unrelated individuals (for abbreviations see **Supplementary Table 4**), for details see **Supplementary Table 6**.

B) Heatmap of: 1) absolute Pearson correlations of standardized trait values in upper triangle; 2) absolute values of estimated pairwise genetic correlations in lower triangle.

Genetic correlations are estimated in 13,342 unrelated individuals. Values below the diagonal in gray had trait heritability less than 1.5 times the SE of heritability.

Extended Data Fig. 3. Properties of associations shared between traits.
A) Shared genomic associations by pairs of traits. For traits *x* and *y*, color in row *x* and column *y* reflects the number of loci associated with both traits divided by the number of loci associated with trait *x*. Traits are presented in the same order as in **Extended Data Figure 2A**, and the side and top color bars reflect trait groups.

B) Relationship between estimated genetic correlation and extent of sharing of genetic associations. For each trait-pair, the extent of locus sharing is defined as the number of loci associated with both traits divided by the total number of loci associated with either trait. Analysis using the absolute value of the Pearson correlation of the residual series results in a very similar pattern. The number of trait pairs in each x-axis category are as follows: 0-1%: 819; 1-10%: 204, 11-20%: 102; 21-30%: 41; 31-40%: 29; 41-50%: 16, >50%: 13. The bar within each box is the median, the box represents the upper and lower quartiles, whiskers extend to 1.5x the interquartile range, and points represent outliers.

Extended Data Fig. 4. Gene-based association of extremely rare variants in *APOB* with serum total cholesterol. The upper panel shows the distribution of the covariate adjusted and inverse-normal transformed phenotype. The lower panel displays the association statistics for each variant included in the gene-based test along with the trait value for minor allele carriers of each variant (orange triangles). SV.P is the P-value from the analysis of each variant in a single-variant analysis. The number of independent individuals in the analysis is 19,291.

Extended Data Fig. 5. Gene-based association of rare variants in *SECTM1* with HDL2 cholesterol. The upper panel shows the distribution of the covariate adjusted and inverse-normal transformed phenotype. The lower panel displays the association statistics for each variant included in the gene-based test, along with the trait value for minor allele carriers of each variant (orange triangles). SV.P is the P-value from the analysis of each variant in

a single-variant analysis. The number of independent individuals in the analysis is 10,984.

Extended Data Fig. 6. Gene-based association of extremely rare variants in *ALDH1L1* with glycine levels. The upper panel shows the distribution of the covariate adjusted and inverse-normal transformed phenotype. The lower panel displays the association statistics for each variant included in the gene-based test, along with the trait value for minor allele carriers of each variant (orange triangles). SV.P is the P-value from the analysis of each variant in a single-variant analysis. The number of independent individuals in the analysis is 8,206.

Extended Data Fig. 7. Population structure of the FinMetSeq dataset, by region. Population structure, by region, from principal components analysis of exome sequencing variant data (MAF > 1%), for 14,874 unrelated individuals known parental birthplaces. Color indicates individuals with both parents born in the same region; gray indicates individuals with different parental birth regions, or missing information for one parent. Abbreviations for the regions: Usm, Uusimaa; Swf, Southwest Finland; Stk, Satakunta; Khm, Kanta-Hame; Prk, Pirkanmaa; Phm, Paijat-Hame; Kyl, Kymenlaakso; SKa, Southern Karelia; Nka, Northern Karelia; SSv, Southern Savonia; NSv, Northern Savonia; Ctf, Central Finland; SOs, Southern Ostrobothnia; Osb, Ostrobothnia; COs, Central Ostrobothnia; NOs, Northern Ostrobothnia; Kai, Kainuu; Lap, Lapland; X, split parental birthplaces. Large solid circles represent the center of each region.

Extended Data Fig. 8. Hierarchical clustering tree produced by fineSTRUCTURE. We identified 16 subpopulations within the FinMetSeq dataset by applying a haplotype-based clustering algorithm, fineSTRUCTURE, on 2,644 unrelated individuals born by 1955 whose parents were both born in the same municipality (Methods). Each subpopulation is named based on the most common parental birth location among its members, with the following abbreviations: NKa, North Karelia; NSv, North Savonia; SOs, South Ostrobothnia; NOs, North Ostrobothnia; Kai, Kainuu; Lap, Lapland; SuK, Surrendered Karelia. A map of Finland with regions labeled is supplied for reference. If multiple subpopulations share the same location label, the subpopulation is further

distinguished with a numeral. NSv3 is used as an internal reference in enrichment analysis. See **Supplementary Table 17** for more detailed demographic descriptions of each subpopulation.

Extended Data Fig. 9. Regional variation in allele frequencies by functional annotation. Enrichment of variants by allelic class in regional sub-populations of late settlement Finland (defined in **Supplementary Table 17**). Each bin represents the ratio of variants in the subpopulation compared to the reference subpopulation (NSv3), after down-sampling the frequency spectra of all populations to 200 chromosomes. Pink cells represent an enrichment (ratio >1), blue cells represent a depletion (ratio <1). Sample sizes and confidence intervals on each enrichment ratios, and their P-values, are presented in **Supplementary Table 18**. The results are consistent with multiple bottlenecks in late settlement Finland, particularly for populations in Lapland and Northern Ostrobothnia.

METHODS

METSIM and FINRISK studies: designs, phenotypes, and sequenced participants

METSIM is a single-site study investigating cardiometabolic disorders and related traits in 10,197 men randomly selected from the population register of Kuopio, Eastern Finland, aged 45 to 73 years at initial examination from 2005 to 2010[15,46]. We attempted exome sequencing of all METSIM study participants.

FINRISK is a series of health examination surveys based on random population samples from five (six in 2002) geographical regions of Finland, carried out every five years beginning in 1972[47]. For exome sequencing, we chose 10,192 participants in the 1992-2007 FINRISK surveys from northeastern Finland (former provinces of North Karelia, Oulu, and Lapland).

All participants in both studies provided informed consent, and study protocols were approved by the Ethics Committees at participating institutions (National Public Health Institute of Finland; Hospital District of Helsinki and Uusimaa; Hospital District of Northern Savo). All relevant ethics committees approved this study.

Selection of traits, harmonization, exclusions, covariate adjustment, and transformation

Of the 257 quantitative traits measured in both METSIM and FINRISK, we selected 64 for association analysis in FinMetSeq based on clinical relevance for cardiovascular and metabolic health (**Supplementary Tables 4, 5**). We excluded individuals with type 1 diabetes and women who were pregnant at the time of phenotyping from all analyses;

individuals with T2D from analyses of glycemic traits; and individuals not fasting for at least 8 hours after their last meal for traits influenced by food consumption. A complete list of exclusions is in **Supplementary Table 5**. We adjusted measured values of systolic and diastolic blood pressures for individuals on antihypertensive medication at the time of testing[48,49], and serum lipid measures for individuals on lipid regulating medications[50,51]. Trait adjustments are listed in **Supplementary Table 5**.

We prepared quantitative traits for association analysis separately for METSIM and FINRISK by linear regression on trait-specific covariates after log transforming skewed variables. Covariates for regression analyses included: age and age^2 (METSIM); sex, age, age^2, and cohort year (FINRISK). Trait transformations and trait-specific covariates are listed in **Supplementary Table 5**. Several traits were adjusted for sex hormone treatment, which included women on contraceptives or hormone replacement therapy. We transformed residuals from these initial regression analyses to normality using inverse normal scores.

Exome sequencing

We carried out exome sequencing in two phases.

Phase 1 We quantified 10,379 DNA samples with PicoGreen (ThermoFisher Scientific) and randomly parsed samples with adequate DNA (>250ng) into cohort-specific files. We then re-arrayed samples to ensure equal numbers of METSIM and FINRISK samples on

each 96-well plate, alternating samples between studies in consecutive positions within and across plates, to minimize between-study batch effects.

Using 100-250ng input DNA, we constructed dual indexed libraries using the HTP Library Kit (KAPA Biosystems, target insert size of 250bp), pooling twelve libraries prior to hybridization to the SeqCap EZ HGSC VCRome (Roche) exome reagent. After estimating the concentration of each captured library pool by qPCR (Kapa Biosystems) to produce appropriate cluster counts for the HiSeq2000 platform (Illumina), we generated 2x100bp paired-end sequence data yielding ~6 Gb per sample to achieve a coverage depth of ≥20x for ≥70% of targeted bases for every sample.

Phase 2 We quantified, prepared, pooled, and captured 9,937 samples just as in Phase 1. Here we generated 2×125bp paired-end sequencing reads on the HiSeq2500 1T to achieve the same coverage as in Phase 1.

Contamination detection, sequence alignment, sample QC, and variant calling

We aligned sequence reads to human genome reference build 37 (bwa-mem, v0.7.7), realigned indels (GATK[52] IndelRealigner v2.4), and marked duplicates (picard MarkDuplicates, v1.113; http://broadinstitute.github.io/picard) and overlapping bases (BamUtil clipOverlap v1.0.11; http://genome.sph.umich.edu/wiki/BamUtil:_clipOverlap).

For each sample, we required SNV genotype array concordance >90% if SNV array data were available, excluding samples with estimated contamination >3% or sample swaps compared to existing genotype data (verifyBamID[53], v1.1.1; **Supplementary Table 1**).

We called SNVs and short indels with GATK[52] (v3.3, using recommended best practices) for all targeted exome bases and 500bp of sequence up and downstream of each target region using HaplotypeCaller. We merged calls in batches of 200 individuals using CombineGVCFs and recalled genotypes for all individuals at all variable sites with GenotypeGVCFs.

After merging genotypes for the 19,378 samples that passed preliminary QC checks, we filtered SNVs and indels separately using the recommended best practices for Variant Quality Score Recalibration (VQSR). We used the true positive variants in the GATK resource bundle (v2.5; build37) to train the VQSR model after restricting to sites in targeted exome regions. After assessment with VQSR, we retained variants for which we identified ≥99% of true positive sites used in the training model for both SNVs and indels.

Following initial variant filtering, we decomposed multi-allelic variants into bi-allelic variants, left-aligned indels, and dropped redundant variants using vt[54] (version 0.5). We filtered variants with >2% missing calls and/or Hardy-Weinberg p-value<10^{-6}. We additionally removed variants with an overall allele balance (alternate AC/sum of total AC) <30% in genotyped samples. We excluded 86 individuals with >2% missing variant calls yielding a final analysis set of 19,292 individuals.

761

762 **Array genotypes, genotype imputation, and integrated exome+imputation panel**

763 For all but 1,488 participants (57 METSIM, 1,431 FINRISK), previously generated array
764 genotypes were available[17,55], with which we generated three datasets: (1) a merged array-
765 based call set of all variants present in ≥90% of array-genotyped individuals across both
766 cohorts; (2) a merged array-based Haplotype Reference Consortium (HRC) v1.1 imputed
767 dataset using the Michigan Imputation Server[56,57]; (3) an integrated data set containing
768 HRC imputed genotypes and exome-sequence variants (excluding all individuals without
769 array data, and using the sequence-based genotypes where there was overlap between
770 sequenced and imputed genotypes).

771

772 **Annotation**

773 We annotated the final set of sequence variants passing QC using Ensembl's variant effect
774 predictor (VEP v76)[58] employing five *in silico* algorithms to predict the functional impact
775 of missense variants: PolyPhen2 HumDiv and HumVar[59], LRT[60], MutationTaster[61], and
776 SIFT[62].

777

778 **Association testing**

779 *Single variants*

780 We carried out single-variant association tests for transformed trait residuals with genotype
781 dosages for variants with MAC≥3 assuming an additive genetic model, using the
782 EMMAX[63] linear mixed model approach, as implemented in EPACTS (v3.3.0;
783 http://genome.sph.umich.edu/wiki/EPACTS), to account for relatedness between

784 individuals. We used genotypes for sequenced variants with MAF≥1% to construct the
785 genetic relationship matrix (GRM).
786
787 *Conditioning on associated variants from prior GWAS*
788 To differentiate association signals identified here from known associations, for each trait
789 we performed exome-wide association analysis conditioning on variants previously
790 associated ($P<10^{-7}$) with that trait in the EBI GWAS catalog
791 (https://www.ebi.ac.uk/gwas/downloads; December 4, 2016 version)[64], publications, or
792 manuscripts in preparation[55,65-67]. The keywords from the GWAS catalog we used to assign
793 known variants to each trait are in **Supplementary Table 21**. We also manually curated
794 published associations for specific metabolites[65,68].
795
796 Using the combined HRC+exome panel, we pruned each trait-specific list of associated
797 variants ("GWAS variants") based on linkage disequilibrium (LD) ($r^2>0.95$). Of 23 GWAS
798 variants absent in the HRC+exome panel, we identified a proxy ($r^2>0.80$) variant for 17;
799 we excluded the remaining six variants from conditional analysis. The variants included in
800 conditional analysis are listed in **Supplementary Table 22**. We extracted genotypes for
801 variants used in conditional analysis from the HRC+exome panel and converted dosages
802 to alternate allele counts by rounding to the nearest integer (0, 1, or 2). For conditional
803 analyses, we imputed missing genotypes for the individuals without array data using the
804 mean genotype. We then ran association analysis using the same linear mixed model
805 approach as in unconditional analysis but including the complete set of pruned GWAS

variants as covariates in the association test. We then evaluated the novelty of conditional associations by searching OMIM, ClinVar, and the literature.

Defining loci

To identify the number of distinct associations for each trait, we performed LD clumping using Swiss (https://github.com/welchr/swiss) of variants with (1) unconditional $P<5\times10^{-7}$ or (2) both unconditional and conditional $P<5\times10^{-5}$ for at least one trait. For each variant in this subset, we provided Swiss with the minimum unconditional p-value across all traits. The clumping procedure starts with the variant with the smallest p-value, merges into one locus all variants within ±1Mbp that have $r^2>0.5$ with the index variant, and iterates this process until no variants remain.

Calculating effects and variance explained of individual variants

For novel variants highlighted in **Table 1** we evaluated the effect of each variant on the trait values by calculating the mean trait value in carriers and non-carriers. As the effect estimates from our association tests are standardized, we calculated variance explained for a given variant with the equation $2f(1-f)\hat{\beta}^2$, where f is the minor allele frequency and $\hat{\beta}$ is the estimated effect size. The variance explained is in **Supplementary Table 10**.

Gene-based testing

We carried out gene-based association tests using the mixed model implementation of SKAT-O[69], considering three different, but nested, sets of variants (variant "masks"):

(1) PTVs at any allele frequency with VEP annotations: frameshift_variant, initiator_codon_variant, splice_acceptor_variant, splice_donor_variant, stop_lost, stop_gained;

(2) PTVs included in (1) plus missense variants with MAF<0.1% scored as "damaging" or "deleterious" by all five functional prediction algorithms;

(3) PTVs included in (1) plus missense variants with MAF<0.5% scored as "damaging" or "deleterious" by all five algorithms.

For each trait and mask, we only tested genes with at least two qualifying variants. Each mask contained a different number of genes with at least two qualifying variants: up to 7,996, 12,795, and 12,890 for the three masks, respectively. The exact number of genes tested varied by trait due to sample size. We first used a Bonferroni-corrected exome-wide threshold for 12,890 genes, which corresponds to a threshold of $P<3.88\times10^{-6}$. Analogous to single-variant association, we passed genes meeting this association threshold forward for additional consideration with hierarchical FDR correction, described below.

Hierarchical FDR correction for testing multiple traits and variants

To control for multiple testing across 64 traits, we adopted an FDR controlling procedure[70], using a two-stage hierarchical strategy (described in **Supplementary Information**). Stage 1 identifies the set of R variants (or genes) associated with at least one trait ($P<5\times10^{-7}$ for single-variant unconditional results and $P<3.88\times10^{-6}$ for gene-based results), controlling genome-wide FDR across all variants at 0.05. Stage 2 identifies all traits associated with the discovered variants in a manner guaranteeing an average FDR<0.05.

851

852 **Genotype validation**

853 We validated exome sequence-based genotype calls using Sanger sequencing for METSIM
854 carriers of 13 trait-associated very rare variants with MAF<0.1% in seven genes, finding
855 concordance for 107 of 108 (99.1%) non-reference genotypes evaluated.

856

857 **Replication in additional Finnish cohorts**

858 We attempted to replicate significant single-variant associations ($P<5\times10^{-7}$) and follow-up
859 suggestive single-variant associations ($P<5\times10^{-5}$) using imputed array data from up to
860 24,776 individuals from three cohort studies: Northern Finland Birth Cohort 1966
861 (NFBC1966)[18], the Helsinki Birth Cohort Study (HBCS)[19], and FINRISK study
862 participants not included in FinMetSeq[16,17].

863

864 For each cohort, prior to phasing we performed genotype quality control batch-wise using
865 standard quality thresholds. We pre-phased array genotypes with Eagle[71] (v2.3) and
866 imputed genotypes genome-wide with IMPUTE[72] (v2.3.1) using 2,690 sequenced Finnish
867 genomes and 5,092 sequenced Finnish exomes. We assessed imputation quality by
868 confirming sex, comparing sample allele frequencies with reference population estimates,
869 and examining imputation quality (INFO score) distributions. We excluded any variant
870 with INFO<0.7 within a given batch from all replication/follow-up analyses.

871

872 For each cohort, we matched, harmonized, covariate adjusted, and transformed available
873 phenotypes as described above for FinMetSeq, and ran single-variant association using the

EMMAX linear mixed model implemented in EPACTS, after generating kinship matrices from LD-pruned (command: plink --indep-pairwise 50 5 0.2) directly genotyped variants with MAF>5%.

Association to disease endpoints

From >1,100 disease endpoints available for analysis in FinnGen, we selected 22 we considered most relevant to the traits analyzed in FinMetSeq, identifying variant associations as described in Tabassum et al.[33].

Association replication in UK Biobank

For eight FinMetSeq anthropometric and blood pressure traits available in UKB (height, weight, BMI, hip circumference, waist circumference, fat percentage, systolic blood pressure, and diastolic blood pressure), we extracted, for variants reaching $P<5x10^{-7}$ in our combined analysis, trait-variant association statistics from http://www.nealelab.is/uk-biobank. Seven of the eight traits had at least one associated variant and 23 of the total of 31 variants were available in UKBB. A comparison of association results is in **Supplementary Table 15**.

Population genetic analyses

Identifying unrelated individuals

To identify nearly independent common SNVs, we removed SNVs with MAF<5% and pruned the remaining SNVs in windows of 50 SNVs, in steps of 5 SNVs, such that no pair of SNVs had $r^2>0.2$. We used KING[73] to estimate pairwise relationships among the exome-

sequenced individuals, removing one individual from each pair inferred by KING to have a relationship of 3rd degree or closer, yielding 14,874 unrelated individuals for population genetic analyses.

Enrichment of predicted-deleterious alleles in Finland

We assessed enrichment of predicted-deleterious alleles in Finland by comparing the 14,874 nearly unrelated FinMetSeq individuals to the 14,944 NFE control exomes in gnomAD (after removing NFE individuals from countries with substantial Finnish populations, Estonia and Sweden). We analyzed the two most common alleles at each site with base quality score >10, mapping quality score >20, and coverage equal to or greater than that found in ≥80% of variable sites (17.73X in FinMetSeq, 32.27X in gnomAD), resulting in ~38.6 Mbp for comparisons. We contrasted the proportional site frequency spectra for FinMetSeq and NFE for five functional variant categories (PTVs, missense, synonymous, UTR, and intronic variants) after down-sampling both datasets to 18,000 chromosomes.

We also assessed the enrichment of deleterious alleles within subpopulations of the FinMetSeq dataset. We applied Chromopainter and fineSTRUCTURE on 2,644 unrelated FinMetSeq individuals whose parents were both born in the same municipality to identify 16 sub-population clusters[74] (**Supplementary Information**). Of the 16 clusters, we used as the reference population a cluster for which the highest proportion of the parents of its members were from early-settlement Finland (NSv3, **Supplementary Table 17**). We used the twelve clusters with >100 members in subsequent analyses (**Supplementary Table**

920 **17**). We then compared the ratio of the site frequency spectra to the reference for PTVs,
921 missense, and synonymous variants, down-sampling both datasets to 200 haploid
922 chromosomes. For each comparison, we computed statistical evidence for enrichment or
923 depletion at a given allele count bin by exact binomial test against a null of equal number
924 of variants found in both the test and reference cluster.
925
926 *Geographical clustering of predicted functionally deleterious alleles*
927 We first generated a distance matrix tabulating the pairwise geographical distance between
928 the birthplaces of all available parents of unrelated sequenced individuals. For each variant
929 of interest, we computed for the minor allele carriers in FinMetSeq the mean distance
930 among all parent pairs. We evaluated statistical significance of geographical clustering by
931 comparing the observed mean distance to mean distances for up to 10,000,000 sets of
932 randomly drawn non-carrier individuals matched by cohort status and number of parents
933 with birthplace information available. Birthplaces of carrier and non-carrier individuals
934 were plotted on a map of Finland, including regions that were ceded prior to WW2 (©
935 Karttakeskus Oy, 2001).
936
937 To assess whether PTVs or missense variants may be more geographically clustered than
938 synonymous variants, we first identified a set of near-independent variants (r^2>0.02) with
939 MAC≥3 and MAF≤5% among the 14,874 unrelated individuals. For each variant, we
940 computed the mean pairwise geographical distance between the birthplaces across all pairs
941 of the available parents of carriers of the minor allele and regressed this mean distance on
942 variant class (PTVs, missense, or synonymous) and MAC, MAC^2, and MAC^3

(**Supplementary Table 16**). For those variants in gnomAD, we also assessed whether variants enriched in FinMetSeq compared to NFE are more likely to be geographically clustered. As above, we computed the mean pairwise distances among parents of carriers of the minor allele and regressed mean distance on the logarithm of enrichment and MAC, MAC^2, and MAC^3 (**Supplementary Table 19**). In both analyses we assessed a model with the interaction terms but report only the model without interactions if the interactions were not significant.

Heritability estimates and genetic correlations

We used genome-wide array genotype data on the 13,326 unrelated individuals for whom both exome sequence and array data were available to estimate heritability and genetic correlations for the 64 traits. We constructed a GRM with PLINK[75] (v.1.90b, https://www.cog-genomics.org/plink2) by applying additional filters for MAF>1% and genotype missingness rate <2% to the set of previously-used genotyped SNVs, leaving 205,149 SNVs for GRM calculation. We used the exact mixed model approach of biMM[76] (v.1.0.0, http://www.helsinki.fi/~mjxpirin/download.html) to estimate the heritability of our 64 traits and the genetic correlation of the 2,016 trait pairs.

Methods References

46 Stancáková, A. *et al.* Changes in insulin sensitivity and insulin release in relation to glycemia and glucose tolerance in 6,414 Finnish men. *Diabetes* **58**, 1212-1221, doi:10.2337/db08-1607 (2009).

47 Borodulin, K. *et al.* Cohort Profile: The National FINRISK Study. *Int J Epidemiol*, doi:10.1093/ije/dyx239 (2017).

48 Wu, J. *et al.* A summary of the effects of antihypertensive medications on measured blood pressure. *Am J Hypertens* **18**, 935-942, doi:10.1016/j.amjhyper.2005.01.011 (2005).

49 Tobin, M. D., Sheehan, N. A., Scurrah, K. J. & Burton, P. R. Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and

systolic blood pressure. *Statistics in medicine* **24**, 2911-2935, doi:10.1002/sim.2165 (2005).
50 Liu, D. J. *et al.* Exome-wide association study of plasma lipids in >300,000 individuals. *Nature genetics*, doi:10.1038/ng.3977 (2017).
51 Friedewald, W. T., Levy, R. I. & Fredrickson, D. S. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clin Chem* **18**, 499-502 (1972).
52 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491-498, doi:10.1038/ng.806 (2011).
53 Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.* **91**, 839-848, doi:10.1016/j.ajhg.2012.09.004 (2012).
54 Tan, A., Abecasis, G. R. & Kang, H. M. Unified representation of genetic variants. *Bioinformatics* **31**, 2202-2204, doi:10.1093/bioinformatics/btv112 (2015).
55 Davis, J. P. *et al.* Common, low-frequency, and rare genetic variants associated with lipoprotein subclasses and triglyceride measures in Finnish men from the METSIM study. *PLoS genetics* **13**, e1007079, doi:10.1371/journal.pgen.1007079 (2017).
56 Das, S. *et al.* Next-generation genotype imputation service and methods. *Nature genetics* **48**, 1284-1287, doi:10.1038/ng.3656 (2016).
57 McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics* **48**, 1279-1283, doi:10.1038/ng.3643 (2016).
58 McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122, doi:10.1186/s13059-016-0974-4 (2016).
59 Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature methods* **7**, 248-249, doi:10.1038/nmeth0410-248 (2010).
60 Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. *Genome research* **19**, 1553-1561, doi:10.1101/gr.092619.109 (2009).
61 Schwarz, J. M., Cooper, D. N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nature methods* **11**, 361-362, doi:10.1038/nmeth.2890 (2014).
62 Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols* **4**, 1073-1081, doi:10.1038/nprot.2009.86 (2009).
63 Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nature genetics* **42**, 348-354, doi:10.1038/ng.548 (2010).
64 Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005-D1012, doi:10.1093/nar/gky1120 (2019).
65 Kettunen, J. *et al.* Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nature communications* **7**, 11122, doi:10.1038/ncomms11122 (2016).

66 Kettunen, J. *et al.* Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nature genetics* **44**, 269-276, doi:10.1038/ng.1073 (2012).
67 Teslovich, T. 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* **27**, 1664-1674, doi:10.1093/hmg/ddy067 (2018).
68 Inouye, M. *et al.* Novel Loci for metabolic networks and multi-tissue expression studies reveal genes for atherosclerosis. *PLoS Genet.* **8**, e1002907, doi:10.1371/journal.pgen.1002907 (2012).
69 Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *American journal of human genetics* **91**, 224-237, doi:10.1016/j.ajhg.2012.06.007 (2012).
70 Peterson, C. B., Bogomolov, M., Benjamini, Y. & Sabatti, C. Many Phenotypes Without Many False Discoveries: Error Controlling Strategies for Multitrait Association Studies. *Genet. Epidemiol.* **40**, 45-56, doi:10.1002/gepi.21942 (2016).
71 Loh, P. R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nature genetics* **48**, 1443-1448, doi:10.1038/ng.3679 (2016).
72 Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics* **5**, e1000529, doi:10.1371/journal.pgen.1000529 (2009).
73 Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867-2873, doi:10.1093/bioinformatics/btq559 (2010).
74 Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS genetics* **8**, e1002453, doi:10.1371/journal.pgen.1002453 (2012).
75 Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7, doi:10.1186/s13742-015-0047-8 (2015).
76 Pirinen, M. *et al.* biMM: efficient estimation of genetic variances and covariances for cohorts with high-dimensional phenotype measurements. *Bioinformatics* **33**, 2405-2407, doi:10.1093/bioinformatics/btx166 (2017).