ARTICLE

https://doi.org/10.1038/s41586-019-1457-z

Exome sequencing of Finnish isolates enhances rare-variant association power

Adam E. Locke[1,2,3,43], Karyn Meltz Steinberg[2,4,43], Charleston W. K. Chiang[5,6,7,43], Susan K. Service[5,43], Aki S. Havulinna[8,9], Laurel Stell[10], Matti Pirinen[8,11,12], Haley J. Abel[2,13], Colby C. Chiang[2], Robert S. Fulton[2], Anne U. Jackson[3], Chul Joo Kang[2], Krishna L. Kanchi[2], Daniel C. Koboldt[2,14,15], David E. Larson[2,13], Joanne Nelson[2], Thomas J. Nicholas[2,16], Arto Pietilä[9], Vasily Ramensky[5,17], Debashree Ray[3,18], Laura J. Scott[3], Heather M. Stringham[3], Jagadish Vangipurapu[19], Ryan Welch[3], Pranav Yajnik[3], Xianyong Yin[3], Johan G. Eriksson[20,21,22], Mika Ala-Korpela[23,24,25,26,27,28], Marjo-Riitta Järvelin[29,30,31,32,33], Minna Männikkö[30,34], Hannele Laivuori[7,35,36], FinnGen Project[37]Susan K. Dutcher[2,13], Nathan O. Stitziel[2,38], Richard K. Wilson[2,14,15], Ira M. Hall[1,2], Chiara Sabatti[9,39], Aarno Palotie[7,40,41], Veikko Salomaa[9], Markku Laakso[19,42], Samuli Ripatti[7,11,41], Michael Boehnke[3,44]* & Nelson B. Freimer[5,44]*

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 considerably and in isolation following a series of bottlenecks, individuals of these populations have numerous deleterious alleles at a relatively high frequency. Here, using exome sequencing of nearly 20,000 individuals from these regions, we investigate the role of rare coding variants in clinically relevant quantitative cardiometabolic traits. Exome-wide association studies for 64 quantitative traits identified 26 newly associated deleterious alleles. Of these 26 alleles, 19 are either unique to or more than 20 times more frequent in Finnish individuals than in other Europeans and show geographical clustering comparable to Mendelian disease mutations that are characteristic of the Finnish population. We estimate that sequencing studies of populations without this unique history would require hundreds of thousands to millions of participants to achieve comparable association power.

Most alleles with demonstrated deleterious effects on phenotypes directly alter the structure or function of a protein[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 that is required for such discoveries remains uncertain and theoretical analyses indicate that studies to date have been underpowered, as most deleterious variants are expected to be rare owing to purifying selection[5]. These previous analyses also suggest that the power to detect associations to deleterious alleles is highest in populations that have expanded in isolation after recent bottlenecks, as alleles passing through the bottlenecks may increase to much higher frequencies than in other populations[6–8].

Finland exemplifies such a history. Bottlenecks occurred at the founding of early-settlement regions (southern and western Finland) 2,000–4,000 years ago and again with internal migration to late-settlement regions (northern and eastern Finland) in the fifteenth and sixteenth centuries[9]. Finland's subsequent population growth (to approximately 5.5 million) generated sizable geographical sub-isolates in late-settlement regions.

This unique population history has resulted in 'the Finnish Disease Heritage'[10], 36 Mendelian diseases that are much more common in

[1]Department of Medicine, Washington University School of Medicine, St Louis, MO, USA. [2]McDonnell Genome Institute, Washington University School of Medicine, St Louis, MO, USA. [3]Department of Biostatistics and Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI, USA. [4]Department of Pediatrics, Washington University School of Medicine, St Louis, MO, USA. [5]Center for Neurobehavioral Genetics, Jane and Terry Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles, Los Angeles, CA, USA. [6]Center for Genetic Epidemiology, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. [7]Quantitative and Computational Biology Section, Department of Biological Sciences, University of Southern California, Los Angeles, CA, USA. [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, USA. [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, USA. [14]The Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH, USA. [15]Department of Pediatrics, The Ohio State University College of Medicine, Columbus, OH, USA. [16]USTAR Center for Genetic Discovery and Department of Human Genetics, University of Utah, Salt Lake City, UT, USA. [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, USA. [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, London, 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]A list of participants and their affiliations appears in the Supplementary Information. [38]Cardiovascular Division, Department of Medicine, Washington University School of Medicine, St Louis, MO, USA. [39]Department of Statistics, Stanford University, Stanford, CA, USA. [40]Analytical and Translational Genetics Unit (ATGU), Psychiatric & Neurodevelopmental Genetics Unit, Departments of Psychiatry and Neurology, Massachusetts General Hospital, Boston, MA, USA. [41]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [42]Department of Medicine, Kuopio University Hospital, Kuopio, Finland. [43]These authors contributed equally: Adam E. Locke, Karyn Meltz Steinberg, Charleston W. K. Chiang, Susan K. Service. [44]These authors jointly supervised this work: Michael Boehnke, Nelson B. Freimer. *e-mail: boehnke@umich.edu; nfreimer@mednet.ucla.edu

Finnish individuals than in other Europeans. These disorders concentrate in late-settlement regions of Finland[10], and the genes responsible for them exhibit extreme enrichment of deleterious variants[11–13]. We created the Finnish Metabolic Sequencing (FinMetSeq) study to capitalize on the population history of late-settlement Finland to discover rare-variant associations with cardiovascular and metabolic disease-relevant quantitative traits through exome sequencing of two extensively phenotyped population cohorts, FINRISK and METSIM (Methods).

We successfully sequenced 19,292 FinMetSeq participants and tested the identified variants for association with 64 clinically relevant quantitative traits, discovering 43 novel associations with deleterious variants[14,15]: 19 associations (11 traits) in FinMetSeq alone and 24 associations (20 traits) in a combined analysis of FinMetSeq with 24,776 Finns from three cohorts with imputed genome-wide genotypes. Of the 26 variants that underlie these 43 associations, 19 were unique to Finland or enriched more than 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 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 would require hundreds of thousands or millions of participants.

Genetic variation

In 19,292 successfully sequenced exomes, we identified 1,318,781 single-nucleotide variants and 92,776 insertion or deletion variants (Supplementary Tables 1–3 and Supplementary Information). Compared to NFE control exomes (gnomAD v.2.1, 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 transformation to normality for analyses (Methods, Supplementary Tables 4, 5). Out of 64 traits, 62 exhibited significant heritability with common single-nucleotide variants ($P < 0.05$; $5\% < h^2 < 53\%$; Extended Data Fig. 2a, Supplementary Table 6), with substantial phenotypic and genetic correlations between traits (Extended Data Fig. 2b).

Single-variant association tests with genetic variants with MAC ≥ 3 among the 3,558 to 19,291 individuals measured for each trait (Supplementary Tables 4, 5) identified 1,249 associations ($P < 5 \times 10^{-7}$) at 531 variants (Supplementary Table 7); 53 traits were associated with at least one variant (Fig. 1a). All 1,249 associations remained significant after adjustment for multiple testing (exome-wide and across the 64 traits using a hierarchical procedure setting average the false discovery rate (FDR) to 5%; see Methods). Using this procedure on the 531 associated variants, we detected 287 more associations (Supplementary Table 8), most of which reflected a high correlation between lipid traits. Of the 531 variants, those with a greater than 10× frequency in FinMetSeq compared to NFE were more likely to be trait-associated (odds ratio = 4.92, $P = 2.6 \times 10^{-5}$; Extended Data Fig. 1c).

After clumping associated variants within 1 megabase (Mb) and with $r^2 > 0.5$ into single loci (Methods), the 531 associated variants represented 262 distinct loci (597 trait–locus pairs; Supplementary Table 7). The number of associated loci per trait correlated positively with trait heritability ($r = 0.38$, $P = 8.8 \times 10^{-4}$), although height was a notable outlier (Fig. 1b).

Most variants and loci (61%) were associated with a single trait; 4% were associated with ≥10 traits. 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 $P < 3.88 \times 10^{-6}$ and multi-trait FDR-corrected $P < 0.05$ (Methods and Supplementary Table 9), including 10 traits associated with *APOB* (Extended Data Fig. 4) and a novel association of *SECTM1* with high density lipoprotein cholesterol subfraction 2 (HDL2-C) (Extended Data Fig. 5).

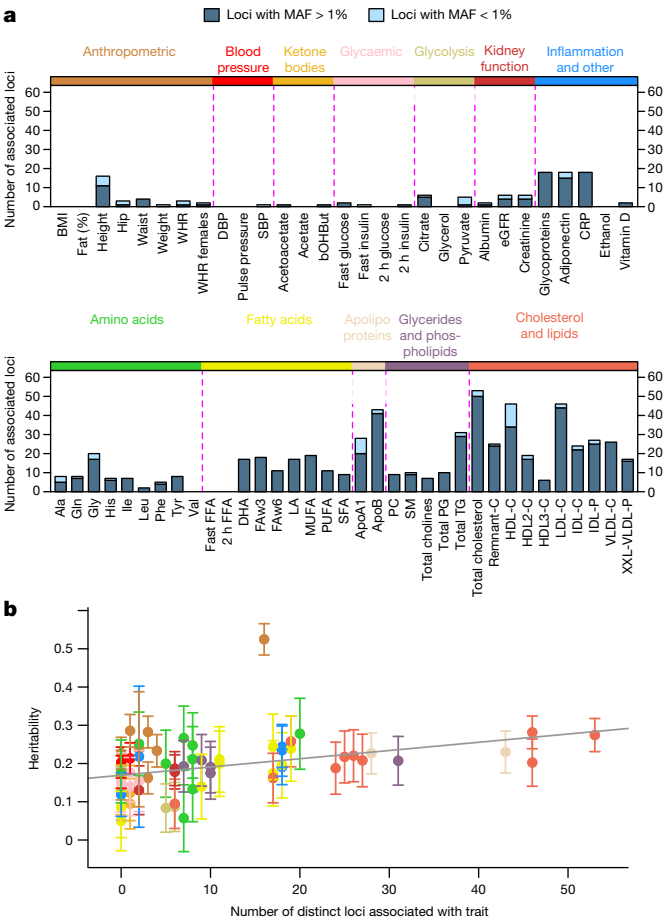


Fig. 1 | Characterization of associations. a, Numbers of genomic loci associated with each trait. Bars are subdivided into common (MAF > 1%, dark blue) and rare (MAF ≤ 1%, light blue) variants. **b**, Relationship between estimated heritability and number of loci detected per trait. Each trait is coloured by trait group. Data are mean ± s.e.m. The grey 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. See Supplementary Table 4 for abbreviations.

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

Replication and follow-up

We attempted to replicate the 478 single-variant associations (unconditional and conditional $P \leq 5 \times 10^{-7}$) and follow up on 2,120 sub-threshold associations from FinMetSeq (unconditional $5 \times 10^{-7} < P \leq 5 \times 10^{-5}$ and conditional $P \leq 5 \times 10^{-5}$) in 24,776

Table 1 | Associations with predicted deleterious variants from FinMetSeq or combined analysis

Chromosome: position	Gene	FinMetSeq MAF	NFE MAF	MAF ratio (95% CI)	Trait	FinMetSeq P	FinMetSeq β	Replication or combined *P*	Replication or combined β
1:55,076,137	*FAM151A*	0.099	0.0147	6.7 (6.1–7.5)	IDL-C	5.4×10^{-16}	−0.187	$\mathbf{2.1 \times 10^{-17}}$	−0.191
					IDL-P	8.9×10^{-14}	−0.172	$\mathbf{1.9 \times 10^{-16}}$	−0.185
2:120,848,049	*EPB41L5*	0.085	0.044	1.9 (1.8–2.1)	eGFR[a]	1.7×10^{-6}	−0.093	4.8×10^{-12}	−0.107
					Creatinine[a]	2.5×10^{-6}	0.091	2.5×10^{-12}	0.098
3:125,831,672	*ALDH1L1*	0.0026	0	∞	Gly	1.8×10^{-8}	−0.873	$\mathbf{4.5 \times 10^{-4}}$	−0.827
4:13,612,630	*BOD1L1*	0.0001	0	∞	WHR	4.7×10^{-7}	−2.501	NA	NA
5:79,336,091	*THBS4*	0.0045	0.0001	45 (14.4–140.9)	Weight[a]	6.7×10^{-7}	−0.377	3.2×10^{-7}	−0.252
5:140,181,423	*PCDHA3*	0.0001	NA	NA	WHR	2.7×10^{-7}	2.559	NA	NA
9:107,548,661	*ABCA1*	0.00023	0	∞	HDL-C	4.8×10^{-10}	−2.046	NA	NA
9:136,501,728	*DBH*	0.05	0.0021	23.8 (18.4–30.4)	DBP[a]	1.5×10^{-6}	−0.115	2.8×10^{-12}	−0.11
11:47,282,929	*NR1H3*	0.0042	0.00003	140 (19.5–1004.4)	HDL-C	1.4×10^{-7}	0.425	$\mathbf{6.7 \times 10^{-7}}$	0.435
					HDL2-C[a]	3.2×10^{-6}	0.473	1.3×10^{-8}	0.458
					VLDL-C[a]	4.0×10^{-6}	−0.469	3.1×10^{-7}	−0.412
11:116,692,293	*APOA4*	0.0096	0.012	0.8 (0.7–0.9)	HDL-C[a]	2.2×10^{-5}	0.225	1.5×10^{-7}	0.196
11:117,352,857	*DSCAML1*	0.016	0.0002	80 (35.7–179.3)	VLDL-C	4.1×10^{-8}	0.299	$\mathbf{2.0 \times 10^{-3}}$	0.162
14:101,198,426	*DLK1*	0.023	0.00013	177 (66.3–472.4)	Height[a]	2.7×10^{-5}	−0.149	1.2×10^{-10}	−0.163
16:55,862,682	*CES1*	0.0018	0.00003	60 (8.3–432.0)	HDL-C	1.1×10^{-10}	0.771	$\mathbf{3.8 \times 10^{-6}}$	0.793
					ApoA1[a]	1.9×10^{-6}	0.668	4.0×10^{-9}	0.718
16:56,996,009	*CETP*	0.0017	0.00003	56.7 (7.9–408.3)	ApoA1	2.6×10^{-8}	0.834	$\mathbf{1.8 \times 10^{-4}}$	1.034
					HDL-C	1.1×10^{-14}	0.946	$\mathbf{8.8 \times 10^{-21}}$	1.217
16:68,013,570	*DPEP3*	0.0099	0.00044	22.5 (12.9–39.1)	HDL-C	1.6×10^{-7}	−0.295	$\mathbf{7.2 \times 10^{-15}}$	−0.373
					ApoA1[a]	5.2×10^{-6}	−0.294	4.0×10^{-7}	−0.253
16:68,732,169	*CDH3*	0.0044	0.00064	6.9 (4.2–11.2)	Pyruvate[a]	3.7×10^{-5}	0.417	6.6×10^{-10}	0.471
17:6,599,157	*SLC13A5*	0.00091	0	∞	Citrate	1.3×10^{-9}	1.294	$\mathbf{9.5 \times 10^{-12}}$	1.309
17:7,129,898	*DVL2*	0.02	0.02	1.0 (0.9–1.1)	Val[a]	4.2×10^{-5}	−0.239	5.7×10^{-9}	−0.232
17:39,135,270	*KRT40*	0.00013	0	∞	HDL-C	3.2×10^{-8}	2.416	NA	NA
17:41,062,979	*G6PC*	0.025	0	∞	MUFA	4.4×10^{-7}	0.275	3.5×10^{-1}	0.067
					Glycerol[a]	5.8×10^{-6}	0.218	4.1×10^{-7}	0.183
					CRP[a]	1.6×10^{-5}	0.175	4.0×10^{-9}	0.185
					Total TG[a]	1.0×10^{-6}	0.23	1.3×10^{-7}	0.197
17:41,926,216	*CD300LG*	0.00034	0	∞	HDL-C	4.8×10^{-14}	2.061	$\mathbf{4.9 \times 10^{-2}}$	0.801
					HDL2-C	1.3×10^{-7}	2.154	NA	NA
					ApoA1	8.1×10^{-8}	1.694	NA	NA
18:47,091,686	*LIPG*	0.0025	0	∞	HDL2-C[a]	1.2×10^{-5}	0.579	5.6×10^{-10}	0.624
					PC[a]	3.1×10^{-6}	0.624	1.1×10^{-8}	0.578
					Total PG[a]	9.0×10^{-6}	0.594	1.1×10^{-7}	0.538
19:10,683,762	*AP1M2*	0.015	0.00009	167 (41.6–668.5)	ApoB	5.8×10^{-8}	−0.282	$\mathbf{1.5 \times 10^{-3}}$	−0.199
					IDL-C[a]	1.1×10^{-6}	−0.289	6.9×10^{-14}	−0.319
					IDL-P[a]	2.1×10^{-6}	−0.281	8.5×10^{-14}	−0.318
					Remnant-C[a]	8.0×10^{-6}	−0.268	2.7×10^{-12}	−0.301
19:11,350,904	*ANGPTL8*	0.0025	0	∞	HDL2-C[a]	3.4×10^{-6}	0.564	1.1×10^{-8}	0.574
19:49,318,380	*HSD17B14*	0.046	0.05	0.9 (0.8–1.0)	Val[a]	3.4×10^{-5}	−0.152	2.1×10^{-7}	−0.144
20:24,994,201	*ACSS1*	0.0026	0	∞	Acetate[a]	1.3×10^{-5}	0.626	2.1×10^{-12}	0.631

Chromosome positions were based on GRCh37. NFE MAFs were taken from gnomAD v.2.1 control exomes excluding Estonian or Swedish individuals. MAF: 0, variant present in gnomAD, but not in NFE controls; NA, variant not present in gnomAD. Replication values with $P < 0.05$ are highlighted in bold. 95% CI, 95% confidence interval. See Supplementary Table 4 for trait abbreviations.
[a]Associated traits that only reach significance in combined analysis.

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 v.2 reference panel (www.sisuproject.fi). Following association analysis within each cohort, we conducted a meta-analysis of the three imputation-based studies to test for replication of FinMetSeq variants (replication analysis) and a four-study meta-analysis with FinMetSeq to follow up on 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 $P < 5 \times 10^{-7}$ in the combined analysis (Supplementary Table 12); more than 60% of the variants were absent from the reference panel and thus could not be tested further.

Among the significant associations from FinMetSeq or the combined analysis, 43 associations were with 26 predicted deleterious variants (6 protein truncating variants (PTVs) and 20 missense variants) that

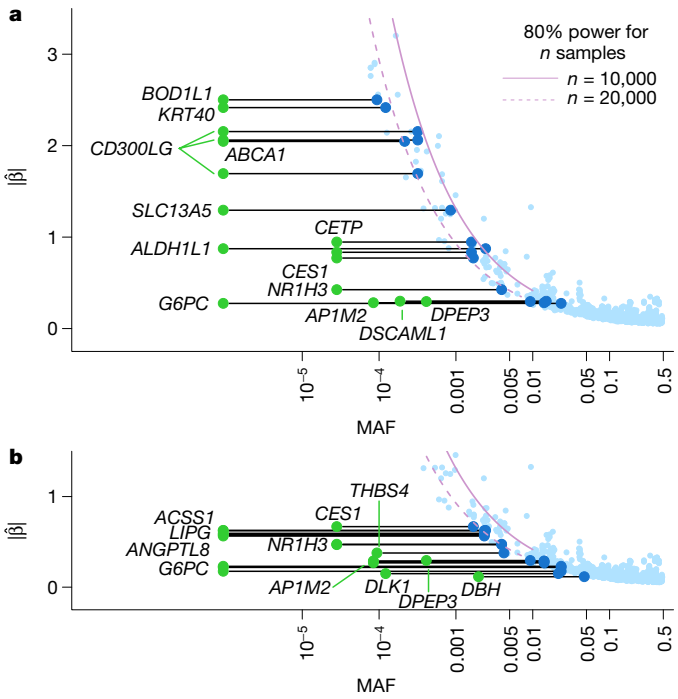


Fig. 2 | Allelic enrichment in the Finnish population and its effect on genetic discovery. a, Relationship between MAF and estimated effect size for associations discovered in FinMetSeq. Each variant that reached significance in FinMetSeq was plotted, with associations in Table 1 represented by dark-blue points (FinMetSeq MAFs) and green points (NFE MAFs). Purple lines indicate 80% power curves for sample sizes of $n = 10{,}000$ and $n = 20{,}000$ at $\alpha = 5 \times 10^{-7}$. **b**, Same plot as in **a**, highlighting the variants in Table 1 that only reached significance in the combined analysis.

conditional analysis and literature review suggest are novel (Table 1). Of those, 19 associations (15 variants) were significant in FinMetSeq (Table 1 and Supplementary Table 11); another 24 associations (16 variants) reached significance in the combined analysis (Table 1 and Supplementary Table 12). Furthermore, 34 out of 43 associations were with 19 variants either found only in Finland or enriched more than 20-fold in FinMetSeq compared to NFE. The identification of 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 previously described genetic evidence (Table 1, expanded version in Supplementary Table 13; see Supplementary Information), highlighting here the most notable findings.

Anthropometric traits

A predicted damaging missense variant (Arg94Cys) in *THBS4*, which was 45× 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 that is found in blood vessel walls and highly expressed in heart and adipose tissues[20]. THBS4 may regulate vascular inflammation[21] and has been implicated in the risk of heart disease[22].

A predicted damaging missense variant (Val104Met) in *DLK1*, which was 177× more frequent in FinMetSeq than in NFE, was associated in the combined analysis with a mean 1.3 cm 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, which are 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 orthologue *Dlk1* lead to embryos with reduced size, skeletal length and lean mass[27]; in Darwin's finches, single-nucleotide variants at this locus have a strong effect on beak size[28].

High-density lipoprotein cholesterol

A predicted deleterious missense variant (Arg112Trp) in *CD300LG* is associated in FinMetSeq with a mean 0.95 mmol l^{-1} increase in high-density lipoprotein cholesterol (HDL-C) and is associated with increased HDL2-C and ApoA1. This variant, which is absent from 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 (Arg722X) in *ALDH1L1* is associated in FinMetSeq with reduced serum glycine levels and is absent from 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 that additional PTVs and missense variants in *ALDH1L1* alter glycine levels ($P = 1.4 \times 10^{-20}$; Extended Data Fig. 6 and Supplementary Table 9).

Ketone bodies

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

Trait–associations and disease end points

Genotype data from FinnGen[33] enabled us to test whether deleterious variants responsible for our novel trait associations contributed to related disease end points. We examined 22 diseases for the 25 available variants shown in Table 1; 3 variants were associated with diseases in FinnGen at a Bonferroni threshold value of $P < 0.05/(22 \times 25) = 9.0 \times 10^{-5}$ (Supplementary Table 14).

A predicted damaging missense variant (Ser32Pro) in *KRT40*, which is associated in FinMetSeq with elevated HDL-C but is absent in NFE, is associated in FinnGen with increased risk of pancreatitis. Although this is the first disease association reported for *KRT40*, type-I keratins regulate exocrine pancreas homeostasis[34]. A 29-bp deletion that causes a frameshift in *FAM151A* is associated in FinMetSeq with decreased total cholesterol in intermediate-density lipoproteins (IDL-C) and decreased concentration of IDL particles, is 6.7× 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*), which is involved in fatty acid metabolism and lies <1Mb from a cardioprotective variant in *PCSK9*. Finally, a predicted damaging missense variant (Arg65Trp) in *DBH*, which is associated with a mean 1.0 mm Hg decrease in diastolic blood pressure in the combined analysis, is 23.8× more frequent in FinMetSeq than in NFE, and is associated in FinnGen with decreased risk of 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 with data from the UK Biobank. Across 8 anthropometric and blood pressure traits for which UK Biobank data are publicly available, our combined analysis identified 31 trait–variant associations, of which 23 were present in the UK Biobank. Of the 23 associations, 20 were to variants with a minor allele frequency (MAF) > 1% in FinMetSeq and a comparable frequency in UK Biobank; 15 (75%) showed association in UK Biobank at $P < 0.05/23 = 2.2 \times 10^{-3}$. The three rare variants in this

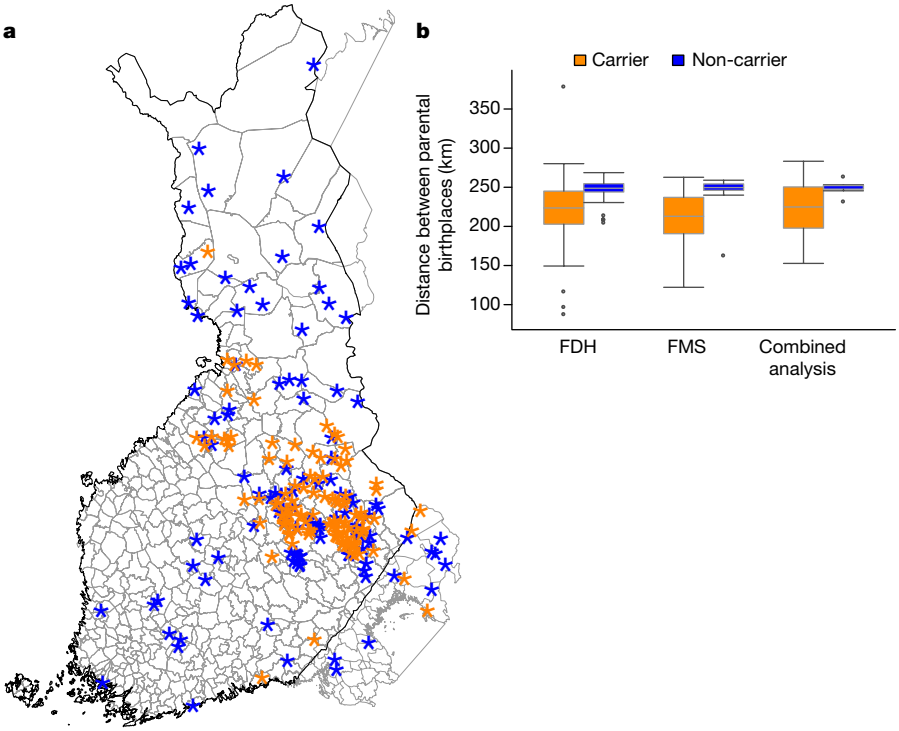


Fig. 3 | Geographical clustering of associated variants. a, Example of geographical clustering for a novel trait-associated variant (Table 1). The map shows birth locations of all 113 parents of carriers (orange) and 113 randomly selected parents of non-carriers (blue) of the minor allele for rs780671030 in *ALDH1L1*. **b**, Mutations in the Finnish Disease Heritage (FDH) genes ($n = 38$) geographically cluster (by parental birthplace) similarly to trait-associated variants (Table 1) that are >10× more frequent in FinMetSeq 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 (centre line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, outliers). Birthplaces of carrier and non-carrier individuals were plotted on a map of Finland, including regions that were ceded before the Second World War (© Karttakeskus Oy, 2001).

analysis were all more than 10× more frequent in FinMetSeq than in UK Biobank; none were associated in UK Biobank (Supplementary Table 15). However, even after adjusting for winner's curse[37], we had <50% power to detect these associations in UK Biobank, 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 would also cluster geographically. Principal component analysis supported this hypothesis, revealing a 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 Table 16a, b).

To analyse 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 and 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 and Supplementary Tables 17, 18). Clusters that represent 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.

Variants that were more than 10× enriched in FinMetSeq compared to NFE displayed particularly strong geographical clustering (Supplementary Table 19). We further characterized clustering for FinMetSeq-enriched trait-associated variants, by comparing mean distances between birthplaces of parents of minor allele carriers to those of non-carriers (Supplementary Table 20). Most of these variants were highly localized. For example, for rs780671030 in *ALDH1L1*, the mean distance between parental birthplaces is 135 km for carriers and 250 km for non-carriers ($P < 1.0 \times 10^{-7}$, Fig. 3a).

Finally, we identified comparable geographical clustering between carriers of 35 Finnish Disease Heritage mutations and carriers of FinMetSeq-enriched trait-associated variants (Fig. 3b and Methods). Clustering was considerably greater in carriers than clustering observed for non-carriers of both sets of variants, suggesting that rare trait-associated variants may be much more unevenly distributed geographically than has previously been appreciated.

Discussion

We demonstrate that a well-powered exome-sequencing study of deeply phenotyped individuals can identify numerous rare variants that are associated with medically relevant quantitative traits. The variants that we identified provide a useful starting point for studies aimed at uncovering biological mechanisms and fostering clinical translation. The power of this study to discover rare-variant associations derives from the numerous deleterious variants that are enriched in or unique to Finland. Prioritizing the sequencing of multiple population isolates that have expanded from recent bottlenecks is a strategy for increasing the scale of 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, thus enriching some variants and depleting others, the numerous rare-variant associations that could be identified by sequencing of 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 that are classically considered monogenic[42,43], by identifying numerous deleterious variants with large effects on quantitative traits that demonstrate geographical clustering comparable to the clustering 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 on more than 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 structures.

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 fewer than 100 of the more than 24,000 deleterious, highly enriched variants identified in FinMetSeq with any of the 64 quantitative traits studied here. The associations that we identified to disease end points in FinnGen hint at the discoveries that will be possible when that database reaches its full size of 500,000 participants. The insights gained from such efforts will accelerate the implementation of precision health, informing projects in more heterogeneous populations that are still at an early stage[45].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1457-z.

Received: 5 November 2018; Accepted: 2 July 2019;
Published online: 31 July 2019

1. Samocha, K. E. et al. Regional missense constraint improves variant deleteriousness prediction. Preprint at https://www.bioRxiv.org/content/10.1101/148353v1 (2017).
2. Marouli, E. et al. Rare and low-frequency coding variants alter human adult height. *Nature* **542**, 186–190 (2017).
3. Flannick, J. et al. Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature* **570**, 71–76 (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. *Nat. Rev. Genet.* **19**, 110–124 (2018).
5. Zuk, O. et al. Searching for missing heritability: designing rare variant association studies. *Proc. Natl Acad. Sci. USA* **111**, E455–E464 (2014).
6. Xue, Y. et al. Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations. *Nat. Commun.* **8**, 15927 (2017).
7. Southam, L. et al. Whole genome sequencing and imputation in isolated populations identify genetic associations with medically-relevant complex traits. *Nat. Commun.* **8**, 15606 (2017).
8. Manolio, T. A. et al. Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
9. Jakkula, E. et al. The genome-wide patterns of variation expose significant substructure in a founder population. *Am. J. Hum. Genet.* **83**, 787–794 (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 (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 (2017).
12. Lim, E. T. et al. Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genet.* **10**, e1004494 (2014).
13. Service, S. K. et al. Re-sequencing expands our understanding of the phenotypic impact of variants at GWAS loci. *PLoS Genet.* **10**, e1004147 (2014).
14. Würtz, P. et al. Quantitative serum nuclear magnetic resonance metabolomics in large-scale epidemiology: a primer on -omic technologies. *Am. J. Epidemiol.* **186**, 1084–1096 (2017).
15. Laakso, M. et al. The Metabolic Syndrome in Men study: a resource for studies of metabolic and cardiovascular diseases. *J. Lipid Res.* **58**, 481–493 (2017).
16. Borodulin, K. et al. Forty-year trends in cardiovascular risk factors in Finland. *Eur. J. Public Health* **25**, 539–546 (2015).
17. Abraham, G. et al. Genomic prediction of coronary heart disease. *Eur. Heart J.* **37**, 3267–3278 (2016).
18. Sabatti, C. et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet.* **41**, 35–46 (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 (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 (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 (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 (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 (2017).
24. Wallace, C. et al. The imprinted DLK1-MEG3 gene region on chromosome 14q32.2 alters susceptibility to type 1 diabetes. *Nat. Genet.* **42**, 68–71 (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. *Nat. Genet.* **49**, 834–841 (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 (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. *Nat. Genet.* **48**, 1473–1480 (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 (2016).
29. Surakka, I. et al. The impact of low-frequency and rare variants on lipid levels. *Nat. Genet.* **47**, 589–597 (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**, e002621 (2015).
31. Wittemans, L. B. L. et al. Assessing the causal association of glycine with risk of cardio-metabolic diseases. *Nat. Commun.* **10**, 1060 (2019).
32. Perry, R. J. et al. Acetate mediates a microbiome–brain–β-cell axis to promote metabolic syndrome. *Nature* **534**, 213–217 (2016).
33. Tabbassum, R. et al. Genetics of human plasma lipidome: understanding lipid metabolism and its link to diseases beyond traditional lipids. Preprint at https://www.biorxiv.org/content/10.1101/457960v1 (2018).
34. Casanova, M. L. et al. Exocrine pancreatic disorders in transsgenic mice expressing human keratin 8. *J. Clin. Invest.* **103**, 1587–1595 (1999).
35. Surendran, P. et al. Trans-ancestry meta-analyses identify rare and common variants associated with blood pressure and hypertension. *Nat. Genet.* **48**, 1151–1161 (2016).
36. Liu, C. et al. Meta-analysis identifies common and rare variants influencing blood pressure and overlapping with metabolic trait loci. *Nat. Genet.* **48**, 1162–1170 (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 Genet.* **13**, e1006916 (2017).
38. Norio, R. Finnish Disease Heritage I: characteristics, causes, background. *Hum. Genet.* **112**, 441–456 (2003).
39. Service, S. et al. Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. *Nat. Genet.* **38**, 556–560 (2006).
40. Chiang, C. W. K. et al. Genomic history of the Sardinian population. *Nat. Genet.* **50**, 1426–1434 (2018).
41. Rivas, M. A. et al. Insights into the genetic epidemiology of Crohn's and rare diseases in the Ashkenazi Jewish population. *PLoS Genet.* **14**, e1007329 (2018).
42. Bastarache, L. et al. Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. *Science* **359**, 1233–1239 (2018).
43. Niemi, M. E. K. et al. Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. *Nature* **562**, 268–271 (2018).
44. Surakka, I. The rate of false polymorphisms introduced when imputing genotypes from global imputation panels. Preprint at https://www.biorxiv.org/content/10.1101/080770v1 (2016).
45. Collins, F. S. & Varmus, H. A new initiative on precision medicine. *N. Engl. J. Med.* **372**, 793–795 (2015).

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Study designs, phenotypes, and sequenced participants of the METSIM and FINRISK studies. 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. We attempted exome sequencing of all METSIM study participants[15,46].

FINRISK is a series of health examination surveys[47] based on random population samples from five (six in 2002) geographical regions of Finland, carried out every five years beginning in 1972. 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 type 2 diabetes from analyses of glycaemic traits; and individuals who had not fasted for at least 8 h after their last meal for traits influenced by food consumption. A complete list of exclusions can be found 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 (>250 ng) 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 100–250 ng input DNA, we constructed dual-indexed libraries using the HTP Library Kit (KAPA Biosystems, target insert size of 250 bp), pooling 12 libraries before 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 2× 100-bp paired-end sequencing data, yielding approximately 6 Gb per sample to achieve a coverage depth of ≥20× for ≥70% of targeted bases for every sample.

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

Contamination detection, sequence alignment, sample quality control and variant calling. We aligned sequence reads to the human genome reference build 37 (bwa-mem, v.0.7.7), realigned insertions or deletions (indels) (GATK[52] IndelRealigner v.2.4) and marked duplicates (Picard MarkDuplicates, v.1.113; http://broadinstitute.github.io/picard) and overlapping bases (BamUtil clipOverlap v.1.0.11; http://genome.sph.umich.edu/wiki/BamUtil:_clipOverlap).

For each sample, we required single-nucleotide variant (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] v.1.1.1; Supplementary Table 1).

We called SNVs and short indels with GATK[52] (v.3.3, using recommended best practices) for all targeted exome bases and 500 bp 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 quality-control 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 (v.2.5; 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] (v.0.5). We filtered variants with >2% missing calls and/or Hardy–Weinberg $P< 10^{-6}$. We additionally removed variants with an overall allele balance (alternate allele count/sum of total allele count) < 30% in genotyped samples. We excluded 86 individuals with >2% missing variant calls yielding a final analysis set of 19,292 individuals.

Array genotypes, genotype imputation and integrated exome + imputation panel. For all except 1,488 participants (57 METSIM, 1,431 FINRISK), previously generated array genotypes were available[17,55], with which we generated three datasets: (1) a merged array-based call set of all variants present in ≥90% of array-genotyped individuals across both cohorts; (2) a merged array-based Haplotype Reference Consortium (HRC) v.1.1 imputed dataset using the Michigan Imputation Server[56,57]; (3) an integrated dataset containing HRC imputed genotypes and exome-sequence variants (excluding all individuals without array data, and using the sequence-based genotypes in cases in which there was overlap between sequenced and imputed genotypes).

Annotation. We annotated the final set of sequence variants that passed quality control using variant effect predictor (VEP v.76)[58] of Ensembl using five in silico algorithms to predict the functional impact of missense variants: PolyPhen2 HumDiv and HumVar[59], LRT[60], MutationTaster[61] and SIFT[62].

Association testing. *Single variants.* We carried out single-variant association tests for transformed trait residuals with genotype dosages for variants with MAC ≥ 3 assuming an additive genetic model, using the EMMAX[63] linear mixed model approach, as implemented in EPACTS (v.3.3.0; http://genome.sph.umich.edu/wiki/EPACTS), to account for relatedness between individuals. We used genotypes for sequenced variants with MAF ≥ 1% to construct the genetic relationship matrix.

Conditioning on associated variants from previous GWAS. To differentiate association signals identified here from known associations, we performed exome-wide association analysis for each trait conditioning on variants previously associated ($P < 10^{-7}$) with that trait in the EBI GWAS catalogue (https://www.ebi.ac.uk/gwas/downloads; 4 December 2016 version)[64], publications[55,65–67] or manuscripts in preparation. The keywords from the GWAS catalogue that we used to assign known variants to each trait can be found in Supplementary Table 21. We also manually curated published associations for specific metabolites[65,68].

Using the combined HRC and exome panel, we pruned each trait-specific list of associated variants (GWAS variants) based on linkage disequilibrium ($r^2 > 0.95$). Of the 23 GWAS variants that were absent from the HRC and exome panel, we identified a proxy ($r^2 > 0.80$) variant for 17; we excluded the remaining 6 variants from the conditional analysis. The variants included in the conditional analysis are listed in Supplementary Table 22. We extracted genotypes for variants used in conditional analysis from the HRC and exome panel and converted dosages to alternate allele counts by rounding to the nearest integer (0, 1 or 2). For conditional analyses, we imputed missing genotypes for the individuals without array data using the mean genotype. We then ran association analysis using the same linear mixed model 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 linkage disequilibrium clumping using Swiss (https://github.com/welchr/swiss) of variants with unconditional $P < 5 \times 10^{-7}$ or both unconditional and conditional $P < 5 \times 10^{-5}$ 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 ±1Mb 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 var. exp. $= 2f(1-f)\hat{\beta}^2$, where f is the MAF and $\hat{\beta}$ is the estimated effect size. The variance explained is included 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 owing to sample size. We first used a Bonferroni-corrected exome-wide threshold for 12,890 genes, which corresponds to a threshold of $P < 3.88 \times 10^{-6}$. Analogous to single-variant association, we passed genes that met this association threshold for additional consideration with hierarchical false-discovery rate (FDR) correction, as 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 the Supplementary Information). Stage 1 identifies the set of R variants (or genes) associated with at least one trait ($P < 5 \times 10^{-7}$ for single-variant unconditional results and $P < 3.88 \times 10^{-6}$ for gene-based results), controlling genome-wide FDR across all variants at $P = 0.05$. Stage 2 identifies all traits associated with the discovered variants in a manner that guarantees an average FDR $P < 0.05$.

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

Replication in additional Finnish cohorts. We attempted to replicate significant single-variant associations ($P < 5 \times 10^{-7}$) and follow up suggestive single-variant associations ($P < 5 \times 10^{-5}$) using imputed array data from up to 24,776 individuals from three cohort studies: Northern Finland Birth Cohort 1966[18], the Helsinki Birth Cohort Study[19] and FINRISK study participants not included in FinMetSeq[16,17].

For each cohort, before phasing we performed genotype quality control batch-wise using standard quality thresholds. We pre-phased array genotypes with Eagle[71] (v.2.3) and imputed genotypes genome-wide with IMPUTE[72] (v.2.3.1) using 2,690 sequenced Finnish genomes and 5,092 sequenced Finnish exomes. We assessed imputation quality by confirming sex, comparing sample allele frequencies with reference population estimates and examining imputation quality (INFO score) distributions. We excluded any variant with INFO < 0.7 within a given batch from all replication/follow-up analyses.

For each cohort, we matched, harmonized, covariate adjusted and transformed available 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 linkage disequilibrium-pruned (command: plink –indep-pairwise 50 5 0.2) directly genotyped variants with MAF > 5%.

Association to disease end points. From >1,100 disease end points available for analysis in FinnGen, we selected 22 that we considered most relevant to the traits analysed in FinMetSeq, identifying variant associations as described previously[33].

Association replication in UK Biobank. For eight FinMetSeq anthropometric and blood pressure traits available in UK Biobank (height, weight, body mass index, hip circumference, waist circumference, fat percentage, systolic blood pressure and diastolic blood pressure), we extracted, for variants reaching $P < 5 \times 10^{-7}$ in our combined analysis, trait-variant association statistics from http://www.nealelab.is/uk-biobank. Of the 8 traits, 7 had at least one associated variant and 23 of the total of 31 variants were available in UK Biobank. 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 third 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 analysed 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 (17.73× in FinMetSeq, 32.27× in gnomAD), resulting in around 38.6 Mb for comparisons. We contrasted the proportional site frequency spectra for FinMetSeq and NFE for five functional variant categories (PTVs, missense, synonymous, untranslated regions 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 to 2,644 unrelated FinMetSeq individuals whose parents were both born in the same municipality to identify 16 subpopulation 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 (Northern Savonia population 3 (NSv3), Supplementary Table 17). We used the twelve clusters with >100 members in subsequent analyses (Supplementary Table 17). We then compared the ratio of the site frequency spectra to the reference for PTVs, missense and synonymous variants, down-sampling both datasets to 200 haploid chromosomes. For each comparison, we computed statistical evidence for enrichment or depletion at a given allele count bin by exact binomial test against a null of equal number of variants found in both the test and reference cluster.

Geographical clustering of predicted functionally deleterious alleles. We first generated a distance matrix tabulating the pairwise geographical distance between the birthplaces of all available parents of unrelated sequenced individuals. For each variant of interest, we computed for the minor allele carriers in FinMetSeq the mean distance among all parent pairs. We evaluated statistical significance of geographical clustering by comparing the observed mean distance to mean distances for up to 10,000,000 sets of randomly drawn non-carrier individuals matched by cohort status and number of parents with birthplace information available.

To assess whether PTVs or missense variants may be more geographically clustered than synonymous variants, we first identified a set of near-independent variants ($r^2 > 0.02$) with MAC ≥ 3 and MAF ≤ 5% among the 14,874 unrelated individuals. For each variant, we computed the mean pairwise geographical distance between the birthplaces across all pairs of the available parents of carriers of the minor allele and regressed this mean distance on 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 sequencing and array data were available to estimate heritability and genetic correlations for the 64 traits. We constructed a genetic relationship matrix 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 genetic relationship matrix 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.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The sequencing data can be accessed through dbGaP (https://www.ncbi.nlm.nih.gov/gap/) 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 (http://www.type2diabetesgenetics.org/). Summary statistics are also available through the NHGRI-EBI GWAS Catalog at https://www.ebi.ac.uk/gwas/downloads/summary-statistics.

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 (2009).
47. Borodulin, K. et al. Cohort profile: the National FINRISK Study. *Int. J. Epidemiol.* **47**, 696–696i (2017).
48. Wu, J. et al. A summary of the effects of antihypertensive medications on measured blood pressure. *Am. J. Hypertens.* **18**, 935–942 (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. *Stat. Med.* **24**, 2911–2935 (2005).
50. Liu, D. J. et al. Exome-wide association study of plasma lipids in >300,000 individuals. *Nat. Genet.* **49**, 1758–1766 (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 (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 (2012).
54. Tan, A., Abecasis, G. R. & Kang, H. M. Unified representation of genetic variants. *Bioinformatics* **31**, 2202–2204 (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 Genet.* **13**, e1007079 (2017).
56. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
57. The Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
58. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
59. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
60. Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. *Genome Res.* **19**, 1553–1561 (2009).
61. Schwarz, J. M., Cooper, D. N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods* **11**, 361–362 (2014).
62. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009).
63. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (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 (2019).
65. Kettunen, J. et al. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat. Commun.* **7**, 11122 (2016).
66. Kettunen, J. et al. Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat. Genet.* **44**, 269–276 (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 (2018).
68. Inouye, M. et al. Novel loci for metabolic networks and multi-tissue expression studies reveal genes for atherosclerosis. *PLoS Genet.* **8**, e1002907 (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. *Am. J. Hum. Genet.* **91**, 224–237 (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 (2016).
71. Loh, P. R. et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (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 Genet.* **5**, e1000529 (2009).
73. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
74. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).
75. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (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 (2017).

Acknowledgements We thank T. Teshiba for coordinating ethical permissions and samples; S. Kerminen, D. Lawson and G. Busby for discussions and providing scripts to run fineSTRUCTURE. S.R. was supported by the Academy of Finland Center of Excellence in Complex Disease Genetics (312062), Academy of Finland (285380), the Finnish Foundation for Cardiovascular Research, the Sigrid Juselius Foundation, Biocentrum Helsinki and University of Helsinki HiLIFE Fellow grant. V.R. acknowledges support by RFBR, research project 18-04-00789 A. V.S. was supported by the Finnish Foundation for Cardiovascular Research. C.S. and L.S. received funding from HG006695, HL113315 and MH105578. M.A.-K. is supported by a Senior Research Fellowship from the National Health and Medical Research Council (NHMRC) of Australia (APP1158958) and 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. A.U.J., D.R., L.J.S., H.M.S., R.W., P.Y., X.Y. and M.B. received funding from DK062370. S.K.S., C.W.K.C. and N.B.F. received funding from HL113315 and NS062691. The METSIM study was supported by grants from Academy of Finland (321428), the Sigrid Juselius Foundation, the Finnish Foundation for Cardiovascular Research, Kuopio University Hospital and the Centre of Excellence of Cardiovascular and Metabolic Diseases is supported by the Academy of Finland (M.L.). Sequencing was funded by 5U54HG003079. A.E.L., K.M.S., H.J.A., C.C.C., C.J.K., K.L.K., D.C.K., D.E.L., J.N., T.J.N., S.K.D., N.O.S., I.M.H. and R.K.W. were funded by 5U54HG003079 and 5UM1HG008853-03.

Author contributions A.E.L., L.J.S., R.K.W., A. Palotie, V.S., M.L., S.R., M.B. and N.B.F. designed the study. A.E.L., K.M.S., H.J.A., R.S.F., D.C.K., D.E.L., J.N., T.J.N. and J.V. produced and quality-controlled the sequence data. A.E.L., A.S.H., A.U.J., A. Pietilä, H.M.S., M.A.-K., V.S. and M.L. collected, quality-controlled and/or prepared the clinical data for association analysis. A.E.L., K.M.S., C.W.K.C., S.K.S., A.S.H., L.S., M.P., C.C.C., A.U.J., C.J.K., K.L.K., V.R., D.R., J.V., R.W., P.Y. and X.Y. analysed data. A.S.H., J.G.E., M.A.-K., M.-R.J. and M.M. collected, quality-controlled and analysed replication data. H.L., S.K.D., N.O.S., I.M.H., C.S., S.R., M.B. and N.B.F. supervised experiments and analyses. A.E.L., K.M.S., C.W.K.C., S.K.S., C.S., M.B. and N.B.F. wrote the paper.

Competing interests : V.S. 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. H.L. is a member of the Nordic Expert group unconditionally supported by Gedeon Richter Nordics and has received an honorarium from Orion. All other authors have no competing interests.

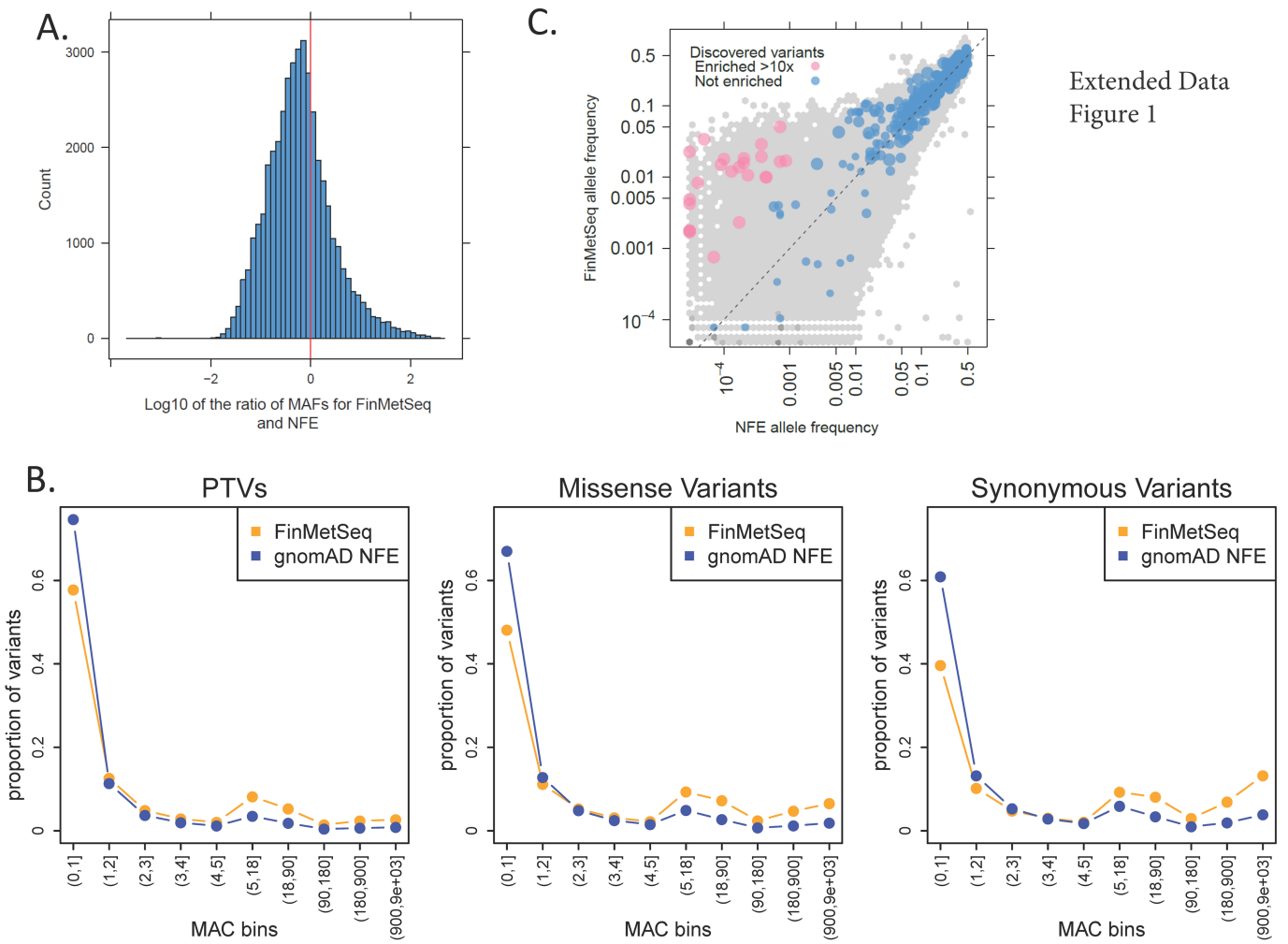
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1457-z.

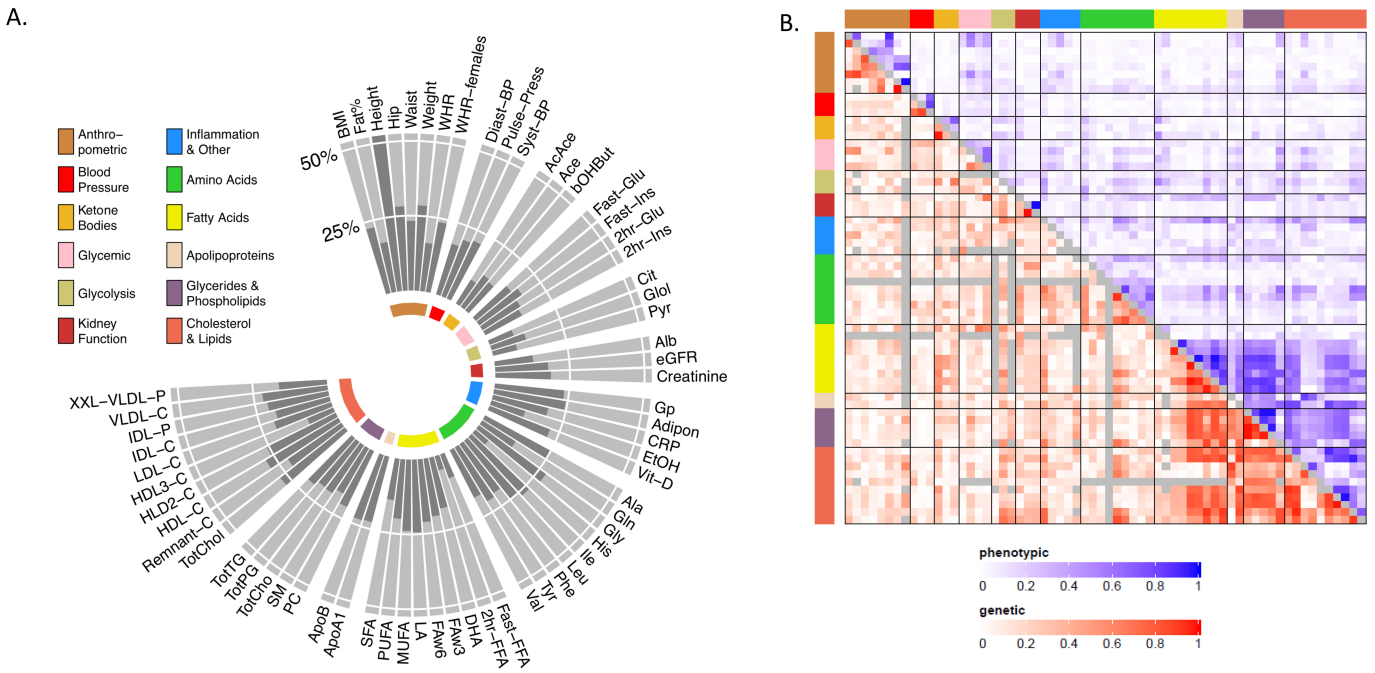
Correspondence and requests for materials should be addressed to M.B. or N.B.F.

Peer review information *Nature* thanks Timothy Frayling, Alan Shuldiner, André G. Uitterlinden, Daniel E. Weeks for their contribution to the peer review of this work.

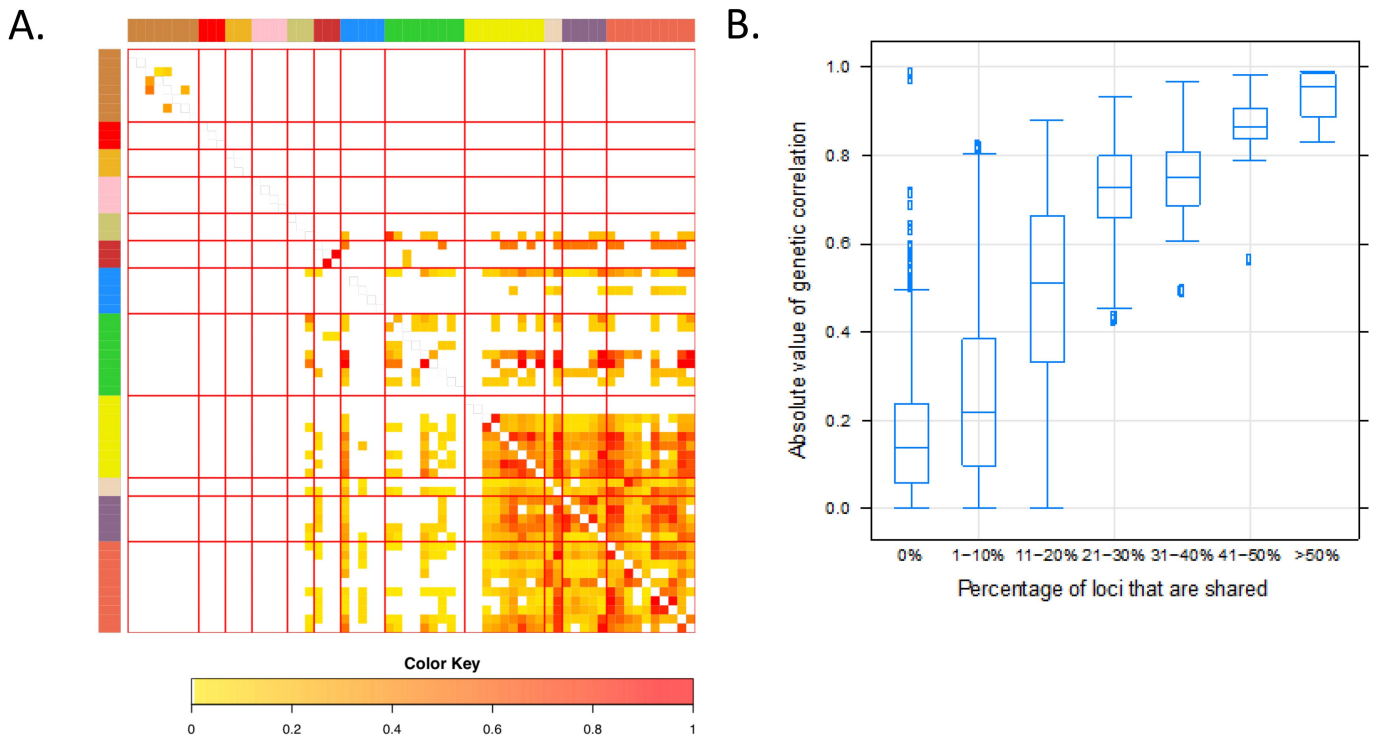
Extended Data Figure 1



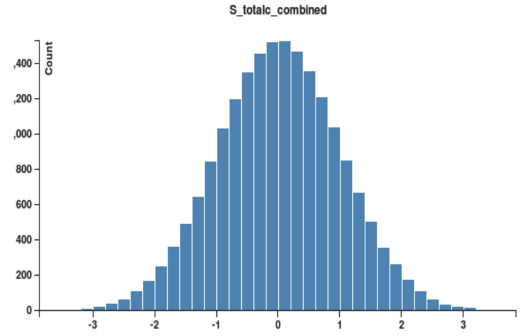
Extended Data Fig. 1 | Allele frequency comparisons between FinMetSeq and NFE from gnomAD. a, Distribution of allelic frequencies between FinMetSeq and gnomAD NFE. The comparison of allele frequencies shows the excess of variants at higher frequency in Finland as a result of the multiple bottlenecks 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 data set. **c**, Comparison of MAFs for trait-associated variants in FinMetSeq and NFE gnomAD. Plotted in the grey background is a two-dimensional 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 coloured and scaled inversely proportional to the logarithm of the association *P* value. Variants >10× enriched in FinMetSeq compared to NFE are pink, those <10× 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 Fig. 2 | Heritability of and correlations between traits. **a**, **b**, Traits are in the same order, clockwise in **a**, and left to right and top to bottom in **b**, following the trait group colour key. **a**, Heritability estimated in 13,342 unrelated individuals (for abbreviations see Supplementary Table 4; for details see Supplementary Table 6). **b**, Heat map of the absolute Pearson correlations of standardized trait values (top right triangle) and the absolute values of estimated pairwise genetic correlations (bottom left triangle). Genetic correlations are estimated in 13,342 unrelated individuals. Values in grey below the diagonal had trait heritability less than 1.5× the s.e. of heritability.



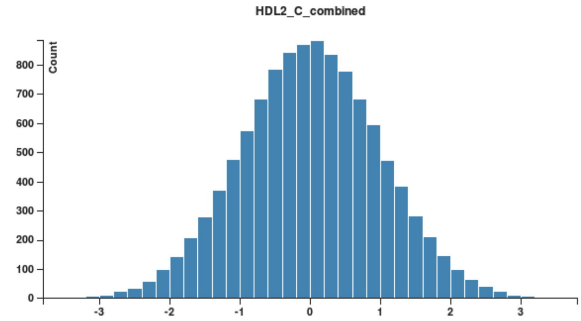
Extended Data Fig. 3 | Properties of associations shared between traits. **a**, Shared genomic associations by pairs of traits. For traits *x* and *y*, colour in row *x* and column *y* reflects the number of loci associated with both traits divided by the number of loci associated with trait *x*. Traits are presented in the same order as in Extended Data Fig. 2a, and the side and top colour 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 is 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 1.5× the interquartile range and points represent outliers.



Symbol	Meaning
▲	Rare homozygote
▼	Heterozygote
⋮	Mean trait value (all samples)
|	Mean trait value (variant carriers)

GENE	mmskat.P	VARIANT	SV.P	BETA	MAF	MAC
APOB	2.0e-13	2:21228642_G/A	0.93	0.089	0.000030	1.0
		2:21230094_AT/A	0.0000013	-2.8	0.000080	3.0
		2:21230336_AT/A	0.33	-0.96	0.000030	1.0
		2:21230583_C/A	0.00011	-2.7	0.000050	2.0
		2:21231852_GA/G	0.0065	-2.7	0.000030	1.0
		2:21233909_TGA/T	0.16	-1.4	0.000030	1.0
		2:21234140_C/CG	0.052	-1.9	0.000030	1.0
		2:21234858_C/CG	0.0098	-2.5	0.000030	1.0
		2:21251412_T/C	0.23	-1.2	0.000030	1.0

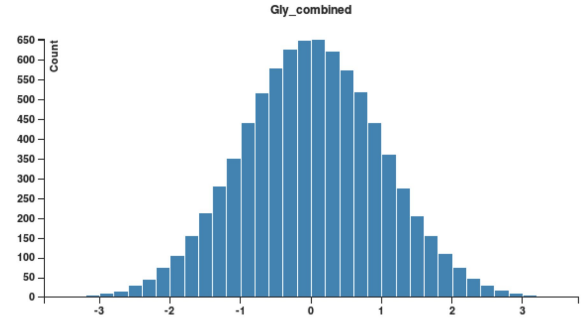
Extended Data Fig. 4 | Gene-based association of extremely rare variants in *APOB* with serum total cholesterol. Top, the distribution of the covariate-adjusted and inverse-normal transformed phenotype. Bottom, the association statistics for each variant included in the gene-based test along with the 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.



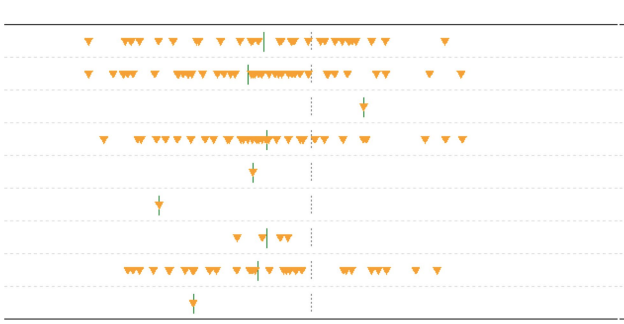
Symbol	Meaning
▲	Rare homozygote
▼	Heterozygote
⋮	Mean trait value (all samples)
|	Mean trait value (variant carriers)

GENE	mmskat.P	VARIANT	SV.P	BETA	MAF	MAC
SECTM1	5.6e-7	17:80280768_G/A	0.0088	-2.6	0.000050	1.0
		17:80280876_G/A	0.000020	-2.9	0.000090	2.0

Extended Data Fig. 5 | Gene-based association of rare variants in *SECTM1* with HDL2 cholesterol. Top, the distribution of the covariate-adjusted and inverse-normal transformed phenotype. Bottom, 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.

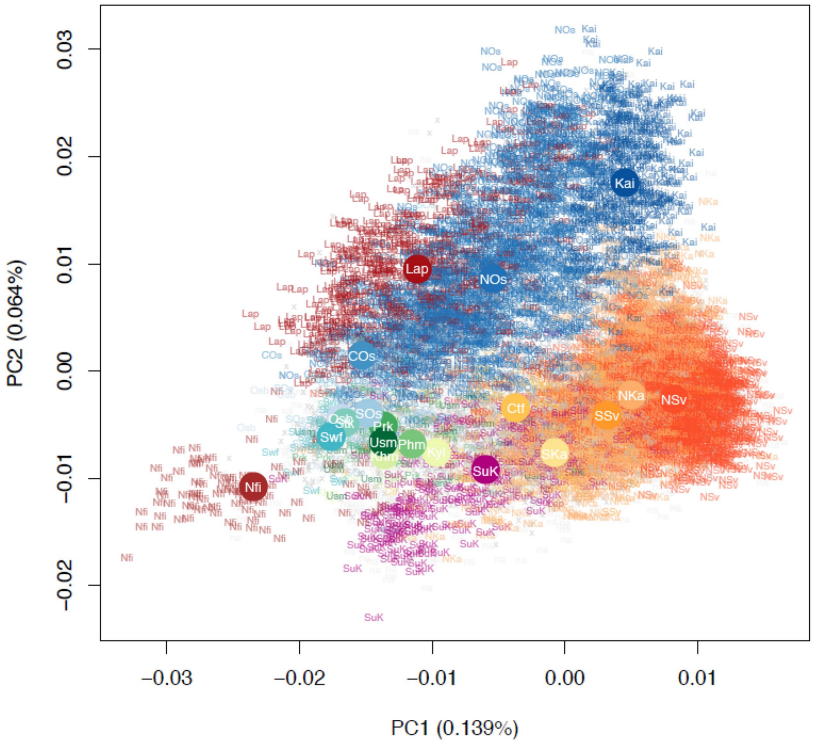


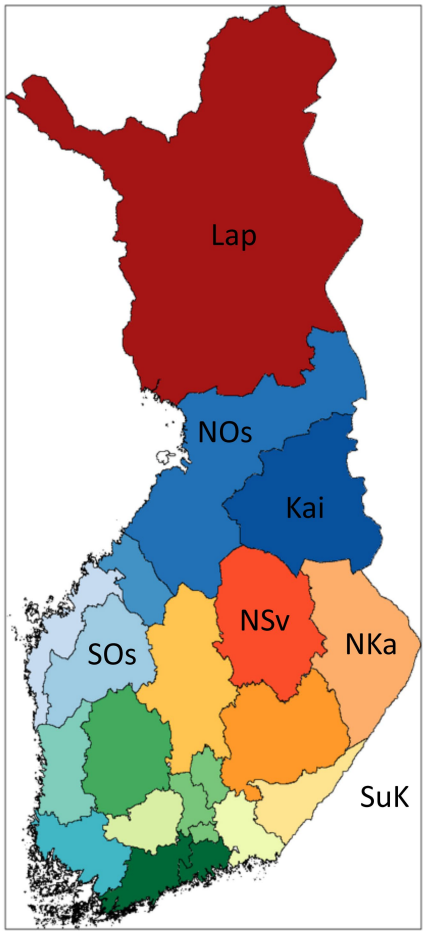
Symbol	Meaning
▲	Rare homozygote
▼	Heterozygote
⋮	Mean trait value (all samples)
|	Mean trait value (variant carriers)



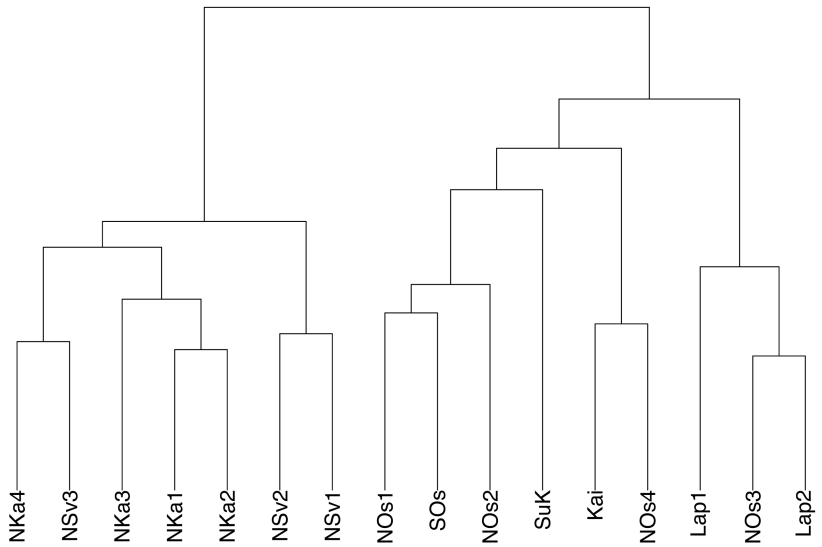
GENE	mmskat.P	VARIANT	SV.P	BETA	MAF	MAC
ALDH1L1	3.9e-21	3:125828918_G/A	0.000029	-0.73	0.0020	33
		3:125831672_G/A	1.8e-8	-0.87	0.0026	42
		3:125850306_T/C	0.53	0.62	0.000060	1.0
		3:125854377_C/T	0.000072	-0.64	0.0024	39
		3:125873443_C/T	0.44	-0.76	0.000060	1.0
		3:125877290_G/A	0.039	-2.1	0.000060	1.0
		3:125879703_G/T	0.19	-0.66	0.00024	4.0
		3:125879755_C/T	0.000050	-0.70	0.0021	34
		3:125879807_T/A	0.13	-1.5	0.000060	1.0

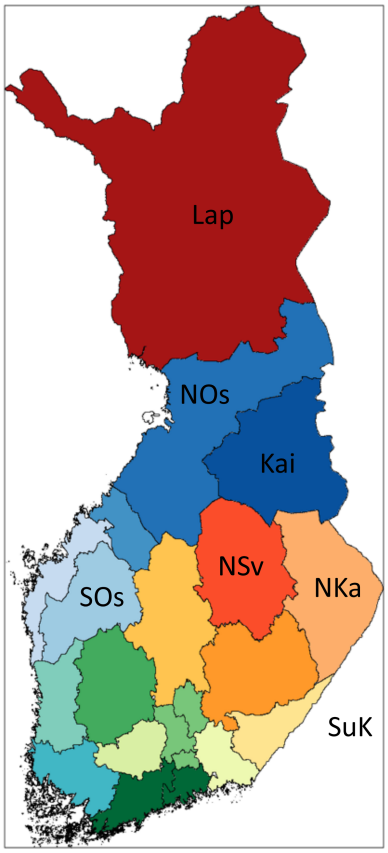
Extended Data Fig. 6 | Gene-based association of extremely rare variants in *ALDH1L1* with glycine levels. Top, the distribution of the covariate-adjusted and inverse-normal transformed phenotype. Bottom, the association statistics for each variant included in the gene-based test, along with the trait 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.



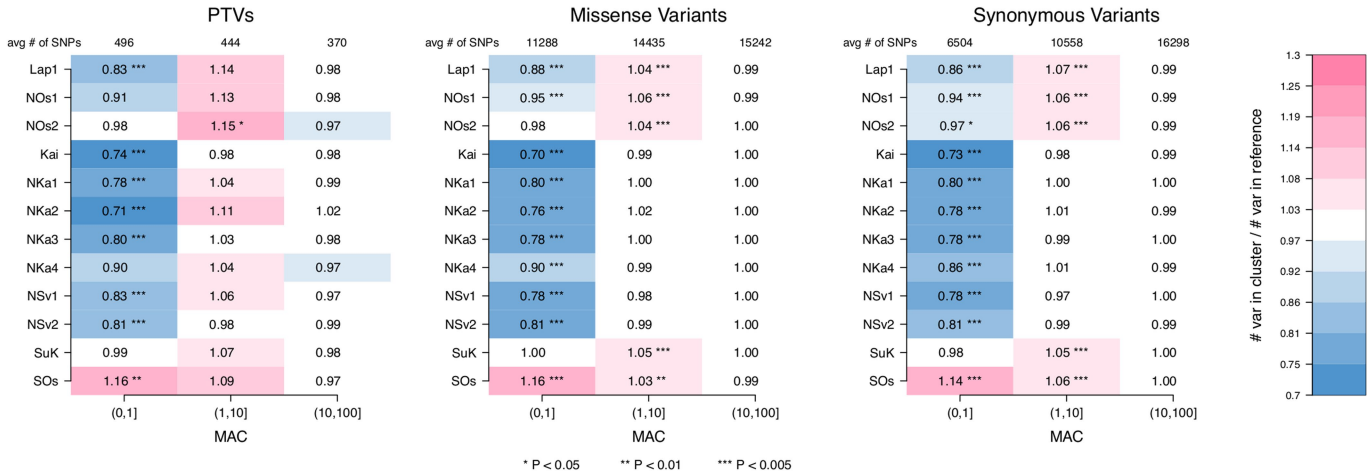


Extended Data Fig. 7 | Population structure of the FinMetSeq dataset, by region. Population structure, by region, from a principal component analysis of exome-sequencing variant data (MAF > 1%) for 14,874 unrelated individuals with known parental birthplaces. Colour indicates individuals with both parents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





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. Kai, Kainuu; Lap, Lapland; NKa, North Karelia; NOs, North Ostrobothnia; NSv, North Savonia; SOs, South Ostrobothnia; SuK, Surrendered Karelia. A map of Finland with regions labelled 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 for the 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 subpopulations 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 enrichment (ratio >1), blue cells represent depletion (ratio <1). Sample sizes and confidence intervals for each enrichment ratio and the associated *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. $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.005$.

nature research

Corresponding author(s): Freimer, NB

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	bwa-mem v0.7.7 picard v1.113 GATK - IndelRealigner v2.4 BamUtil - clipOverlap v1.0.11 verifyBamID v1.1.1 GATK v3.3 VQSR vt v0.5 VEP v76 CADD v1.2 (PolyPhen2 (v2.2.2), LRT (11/09 release), MutationTaster (2013 release), SIFT (09/11 release)) as in dbNSFP v2.4
Data analysis	EMMAX EPACTS v3.3.0 PLINK v1.9 R 3.4.0 Swiss v1.0.0 SKAT-O

Eagle v2.3
IMPUTE2 v2.3.1
KING v2.0
SHAPEIT v2, r837
ChromoPainter v2
fineStructure v2.0.8
biMM v1.0.0

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The sequence data can be accessed through dbGaP using the following study numbers: FINRISK: phs000756, METSIM: phs000752. Association results can be accessed at http://pheweb.sph.umich.edu/FinMetSeq/. NOTE: METSIM phs000752 is the correct accession number, however dbGaP has not yet released the data. We are working to resolve this

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	All available subjects in two extensive population cohorts of Finnish subjects
Data exclusions	We excluded 126 individuals, 92 with type 1 diabetes and 34 women who were pregnant at the time of phenotyping, from all analyses. Pregnancy is known to dramatically alter metabolic profiles and type 1 diabetics also represent an altered profile compared to the general population, and thus both might obscure variant-trait relationships present in the rest of the population. Both represent a very small fraction of the overall sample. Though these samples were sequenced, they were excluded prior to any gene/trait association testing. We also excluded 3,088 individuals with T2D from analyses of glycemic traits. For traits influenced by food consumption (amino acids, fatty acids, LDL cholesterol, total triglycerides, and glycemic traits), we excluded individuals not fasting for at least 8 hours after their last meal. A complete list of exclusions can be found in Supplementary Table 4. All exclusion criteria were determined before any analyses were conducted.
Replication	We performed replication analysis of significant single-variant associations (P<5×10-7) and follow-up analysis of suggestive single-variant associations (P<5×10-5) in up to 24,776 individuals from three GWAS cohort studies: Northern Finland Birth Cohort 1966 (NFBC1966), the Helsinki Birth Cohort Study (HBCS), and FINRISK study participants not included in the exome sequencing portion of FinMetSeq. We also did look ups of our discoveries in UK Bio Bank (for some of the same quantitative traits) and FinnGen (a Finnish Biobank, for disease endpoints).
Randomization	no experimental treatments in our study
Blinding	no experimental treatments in our study

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	METSIM is a single-site study comprised of 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. FINRISK is a series of health examination surveys carried out by the National Institute for Health and Welfare (formerly National Public Health Institute) of Finland every five years beginning in 1972. The surveys are based on random population samples from five (six in 2002) geographical regions of Finland. Participants were selected by 10-year age group, sex, and study area. Survey sample sizes have varied from 7,000 to 13,000 individuals and participation rates from 60% to 90%. The age-range was 25 to 64 years until 1992 and 25 to 74 years since 1997.
Recruitment	FINRISK - Multi-site national health examination of adults executed every 5 years since 1972 representing a geographically diverse cross-section of the country. No major exclusions. METSIM - Single site population cohort representing older (>= 45 at recruitment) adult males in the city of Kuopio in eastern Finland. Though a population cohort, recruited only older men due to their increased risk for cardiovascular and metabolic disease.