

1 **Exome sequencing identifies high-impact trait-associated alleles enriched in Finns**

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

13

14 <sup>1</sup>Department of Medicine, Washington University School of Medicine, St. Louis, MO

15 <sup>2</sup>McDonnell Genome Institute, Washington University School of Medicine, St. Louis,  
16 MO

17 <sup>3</sup>Department of Biostatistics and Center for Statistical Genetics, University of Michigan  
18 School of Public Health, Ann Arbor, MI

19 <sup>4</sup>Department of Pediatrics, Washington University School of Medicine, St. Louis, MO

20 <sup>5</sup>Center for Neurobehavioral Genetics, Jane and Terry Semel Institute for Neuroscience  
21 and Human Behavior, University of California Los Angeles, Los Angeles, CA

22 <sup>6</sup>Center for Genetic Epidemiology, Department of Preventive Medicine, Keck School of  
23 Medicine, University of Southern California, Los Angeles, CA

24 <sup>7</sup>Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki,  
25 Finland

26 <sup>8</sup>National Institute for Health and Welfare, Helsinki, Finland

27 <sup>9</sup>Department of Biomedical Data Science, Stanford University, Stanford, CA

28 <sup>10</sup>Department of Public Health, University of Helsinki, Helsinki, Finland;

29 <sup>11</sup>Helsinki Institute for Information Technology HIIT and Department of Mathematics  
30 and Statistics, University of Helsinki, Helsinki, Finland

31 <sup>12</sup>Department of Genetics, Washington University School of Medicine, St. Louis, MO

32 <sup>13</sup>The Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH

33 <sup>14</sup>Department of Pediatrics, The Ohio State University College of Medicine, Columbus,  
34 OH

35 <sup>15</sup>USTAR Center for Genetic Discovery and Department of Human Genetics, University  
36 of Utah, Salt Lake City, UT

37 <sup>16</sup>Federal State Institution "National Medical Research Center for Preventive Medicine"  
38 of the Ministry of Healthcare of the Russian Federation, Moscow, Russia

39 <sup>17</sup>Departments of Epidemiology and Biostatistics, Bloomberg School of Public Health,  
40 Johns Hopkins University, Baltimore, MD

41 <sup>18</sup>Institute of Clinical Medicine, Internal Medicine, University of Eastern Finland,  
42 Kuopio, Finland

43 <sup>19</sup>Department of Public Health Solutions, National Institute for Health and Welfare,  
44 Helsinki, Finland

45 <sup>20</sup>Folkhälsan Research Center, Helsinki, Finland

46 <sup>21</sup>Department of General Practice and Primary Health Care, University of Helsinki,  
47 Helsinki and Helsinki University Hospital, Helsinki, Finland  
48 <sup>22</sup>Systems Epidemiology, Baker Heart and Diabetes Institute, Melbourne, Victoria,  
49 Australia  
50 <sup>23</sup>Computational Medicine, Faculty of Medicine, University of Oulu and Biocenter Oulu,  
51 University of Oulu, Oulu, Finland  
52 <sup>24</sup>NMR Metabolomics Laboratory, School of Pharmacy, University of Eastern Finland,  
53 Kuopio, Finland  
54 <sup>25</sup>Population Health Science, Bristol Medical School, University of Bristol, Bristol, UK  
55 <sup>26</sup>Medical Research Council Integrative Epidemiology Unit at the University of Bristol,  
56 Bristol, UK  
57 <sup>27</sup>Department of Epidemiology and Preventive Medicine, School of Public Health and  
58 Preventive Medicine, Faculty of Medicine, Nursing and Health Sciences, The Alfred  
59 Hospital, Monash University, Melbourne, Victoria, Australia  
60 <sup>28</sup>Biocenter Oulu, University of Oulu, Oulu, Finland  
61 <sup>29</sup>Center for Life Course Health Research, Faculty of Medicine, University of Oulu, Oulu,  
62 Finland  
63 <sup>30</sup>Unit of Primary Health Care, Oulu University Hospital, Oulu, Finland  
64 <sup>31</sup>Department of Epidemiology and Biostatistics, MRC-PHE Centre for Environment and  
65 Health, School of Public Health, Imperial College London, London, UK  
66 <sup>32</sup>Department of Life Sciences, College of Health and Life Sciences, Brunel University  
67 London, Uxbridge, UK  
68 <sup>33</sup>Northern Finland Birth Cohorts, Faculty of Medicine, University of Oulu, Oulu,  
69 Finland  
70 <sup>34</sup>Medical and Clinical Genetics, University of Helsinki and Helsinki University Hospital,  
71 Helsinki, Finland  
72 <sup>35</sup>Department of Obstetrics and Gynecology, Tampere University Hospital and University  
73 of Tampere, Faculty of Medicine and Life Sciences, Tampere, Finland  
74 <sup>36</sup>Cardiovascular Division, Department of Medicine, Washington University School of  
75 Medicine, St. Louis, MO  
76 <sup>37</sup>Department of Statistics, Stanford University, Stanford, CA  
77 <sup>38</sup>Analytical and Translational Genetics Unit (ATGU), Psychiatric &  
78 Neurodevelopmental Genetics Unit, Departments of Psychiatry and Neurology,  
79 Massachusetts General Hospital, Boston, MA  
80 <sup>39</sup>Broad Institute of MIT and Harvard, Cambridge, MA  
81 <sup>40</sup>Department of Medicine, Kuopio University Hospital, Kuopio, Finland  
82  
83 \*These authors contributed equally to this work.  
84 †These authors jointly supervised this work.

85

86 **ABSTRACT**

87 As yet undiscovered rare variants are hypothesized to substantially influence an  
88 individual's risk for common diseases and traits, but sequencing studies aiming to  
89 identify such variants have generally been underpowered. In isolated populations that  
90 have expanded rapidly after a population bottleneck, deleterious alleles that passed  
91 through the bottleneck may be maintained at much higher frequencies than in other  
92 populations. In an exome sequencing study of nearly 20,000 cohort participants from  
93 northern and eastern Finnish populations that exemplify this phenomenon, most novel  
94 trait-associated deleterious variants are seen only in Finland or display frequencies more  
95 than 20 times higher than in other European populations. These enriched alleles underlie  
96 34 novel associations with 21 disease-related quantitative traits and demonstrate a  
97 geographical clustering equivalent to that of Mendelian disease mutations characteristic  
98 of the Finnish population. Sequencing studies in populations without this unique history  
99 would require hundreds of thousands to millions of participants for comparable power for  
100 these variants.

101

102 **INTRODUCTION**

103 Genotyping studies of common genetic variants (defined here as minor allele frequency  
104 [MAF]>1%) have identified tens of thousands of genome-wide significant associations  
105 with common diseases and disease-related quantitative traits<sup>1</sup>. For most traits, however,  
106 these associations account for only a modest fraction of trait heritability, and the  
107 mechanisms through which associated variants contribute to biological processes remain  
108 mostly unknown. These observations have led to the expectation that rare variants

109 (defined here as  $MAF \leq 1\%$ ) which are not well-tagged by the single-nucleotide  
110 polymorphisms (SNPs) on genome-wide genotyping arrays are probably responsible for  
111 much of the heritability that remains unexplained<sup>2</sup>. Additionally, because purifying  
112 selection acts to remove deleterious alleles from the population, most variants that exert a  
113 sizable effect on complex traits, and that likely offer the best prospect for revealing  
114 biological mechanisms, should be particularly rare.

115

116 Rare variants are unevenly distributed between populations and difficult to represent  
117 effectively on commercial genotyping arrays, as evidenced by relatively sparse  
118 association findings even from large array-based studies of coding variants<sup>3-6</sup>.  
119 Discovering rare variant associations will therefore almost certainly require exome or  
120 genome sequencing of very large numbers of individuals. However, the sample size  
121 required to reliably identify rare-variant associations remains uncertain; most sequencing  
122 studies to date have identified few novel associations, and theoretical analyses confirm  
123 that they have been underpowered to do so<sup>7</sup>. These analyses also suggest that power to  
124 detect rare variant associations varies enormously between populations that have  
125 expanded in isolation from recent bottlenecks compared to those that have not.

126

127 In isolated populations that expand rapidly following a bottleneck, alleles that pass  
128 through the bottleneck often rise to a much higher frequency than in other populations<sup>8-10</sup>.  
129 If the bottleneck was recent, even deleterious alleles under negative selection may remain  
130 relatively frequent in these populations, resulting in increased power to detect association  
131 with disease-related traits. The Finnish population exemplifies this type of history. It

132 grew from bottlenecks occurring 2,000-4,000 years ago in the founding of the early-  
133 settlement regions of southern and western Finland; internal migration in the 15<sup>th</sup> and 16<sup>th</sup>  
134 centuries to the late-settlement regions of northern and eastern Finland created additional  
135 bottlenecks<sup>11</sup>. The subsequent rapid growth of the Finnish population (to ~5.5 million,  
136 larger than any other human isolate) generated sizable geographic sub-isolates in late-  
137 settlement regions.

138

139 Geneticists have long noted that the bottlenecks that were so prominent in Finland's  
140 recent history caused 36 Mendelian disorders to be much more common in Finland than  
141 in other European countries, while several other disorders are much less common, a  
142 phenomenon termed "the Finnish Disease Heritage"<sup>12</sup>. The identification of mutations for  
143 35 of these disorders has confirmed that they mostly concentrate in late settlement  
144 regions<sup>12</sup>. Additional studies demonstrated, in these regions, an overall enrichment of  
145 deleterious variants more extreme compared to other isolates or to Finland generally<sup>13-15</sup>.  
146 We reasoned that this enrichment would enable exome sequencing studies of late-  
147 settlement Finland to be better powered than studies in other populations to  
148 systematically investigate the impact of low-frequency variants on disease-related  
149 quantitative traits. Based on this expectation, we formed such a sample ("FinMetSeq")  
150 from two Finnish population-based cohort studies: FINRISK and METSIM (see  
151 Methods).

152

153 Using >1.4 M variants identified and genotyped by successful exome sequencing of  
154 19,292 FinMetSeq participants, we conducted single-variant association analysis with 64

155 clinically relevant quantitative traits<sup>16,17</sup>. We identified 43 novel associations with  
156 deleterious variants in 25 traits: 19 associations (11 traits) in FinMetSeq and 24  
157 associations (20 traits) in a combined analysis of FinMetSeq with an additional 24,776  
158 Finns from three cohorts for which imputed array-based genome-wide genotype data  
159 were available. Nineteen of the 26 variants underlying these 43 novel associations were  
160 unique to Finland or enriched >20-fold in FinMetSeq compared to non-Finnish  
161 Europeans (NFE).

162

163 We demonstrate that (1) a well-powered exome sequencing study can identify numerous  
164 rare alleles, each of which has a substantial effect on one or more traits in the individuals  
165 who carry them, and (2) exome sequencing in a population that has expanded after recent  
166 population bottlenecks is an extraordinarily efficient strategy to discover such effects. As  
167 most of the novel putatively deleterious trait-associated variants that we identified are  
168 unique to or highly enriched in Finland, similarly powered studies of these variants in  
169 non-Finnish populations might require hundreds of thousands or even millions of  
170 participants. Additionally, the geographical clustering of these enriched alleles, like the  
171 Finnish Disease Heritage mutations, demonstrates that the distribution of trait-associated  
172 rare alleles may vary significantly between locales within a country.

173

## 174 **RESULTS**

### 175 **Genetic variation**

176 We attempted to sequence the protein-coding regions of 23,585 genes covering 39 MB of  
177 genomic sequence in 20,316 FinMetSeq participants. After extensive quality control, we  
178 included in downstream analysis 19,292 individuals sequenced to 47x mean depth

179 (Methods, **Supplementary Table 1**). We identified 1,318,781 single nucleotide variants  
180 (SNVs) and 92,776 insertion/deletion (indel) variants, with a mean of 20,989 SNVs and  
181 604 indel variants per individual. The majority (87.5%) of SNVs identified were rare  
182 (MAF<1%); 40.5% were singletons (**Table 1**). Each participant carried 15 singleton  
183 variants on average, 17 rare (MAF≤1%) protein truncating variants (PTVs; annotated as  
184 stop gain, essential splice site, start loss, or frameshift) alleles, and 171 common  
185 (MAF>1%) PTVs (**Supplementary Table 2**). Frameshift indels accounted for the largest  
186 proportion of PTVs (31% of rare, 42% of common), while stop gain variants were the  
187 most frequent type of protein truncating SNVs (29% of rare, 20% of common).

188

189 We compared variant allele frequencies in FinMetSeq to those of NFE control exomes  
190 from the Genome Aggregation Database (gnomAD v2.1, **Extended Data Fig. 1**). As in  
191 previous smaller-scale comparisons of Finnish and NFE exomes, in FinMetSeq we  
192 observe a depletion of the rarest alleles (singletons and doubletons) and a relative excess  
193 of more common variants (minor allele count, MAC  $\geq 5$ ) compared to NFE for all classes  
194 of variants. This effect is particularly marked for alleles predicted to be deleterious  
195 (**Extended Data Fig. 2**).

196

## 197 **Single-variant association analyses**

198 We tested for association between genetic variants in FinMetSeq and 64 clinically  
199 relevant quantitative traits measured in members of both FINRISK and METSIM  
200 (**Supplementary Table 3**). We adjusted lipid and blood pressure traits for lipid lowering  
201 and antihypertensive medication use, respectively, adjusted all traits for covariates using

202 linear regression (**Supplementary Table 4**), and inverse normalized trait residuals to  
203 generate normally distributed traits for genetic association analysis that assumed an  
204 additive model (Methods). Based on common variants, 62 of 64 traits exhibited  
205 significant heritability ( $P<0.05$ ;  $h^2$  range 5.0-52.5%; **Fig. 1A, Supplementary Table 5**),  
206 and there was substantial correlation between traits, phenotypically and genetically (**Fig.**  
207 **1B**).

208

209 We tested the 64 traits for single-variant associations with the 362,996 to 602,080 genetic  
210 variants with MAC  $\geq 3$  among the 3,558 to 19,291 individuals measured for each trait  
211 (**Supplementary Tables 3 & 4**). Association results are available for download and can  
212 be explored interactively with PheWeb (<http://pheweb.sph.umich.edu/FinMetSeq/>) and  
213 via the Type 2 Diabetes Knowledge Portal ([www.type2diabetesgenetics.org](http://www.type2diabetesgenetics.org)). We  
214 identified 1,249 trait-variant associations ( $P<5\times 10^{-7}$ ) at 531 variants (**Supplementary**  
215 **Table 6**), with 53 of 64 traits associated with at least one variant (**Fig. 2A**). All 1,249  
216 associations remained significant after multiple testing adjustment across the exome and  
217 across the 64 traits with a hierarchical procedure setting average FDR at 5% (Methods).  
218 Using the hierarchical FDR procedure, we detected an additional 287 trait-variant  
219 associations at these 531 variants (**Supplementary Table 7**). These additional  
220 associations reflect the high correlation between a subset of lipid traits, e.g. high-density  
221 lipoprotein cholesterol (HDL-C) and apolipoprotein A1 (ApoA1). Given the diversity of  
222 traits assessed in these cohorts, these associations may shed additional light on the  
223 biology of measures that have been less frequently assayed in large GWAS, such as  
224 intermediate density lipoproteins (IDL) and very-low-density lipoprotein (VLDL)

225 particles. Of the 531 associated variants, 59 (11%) were rare ( $MAF \leq 1\%$ ); by annotation,  
226 200 (38%) were coding, 108 (20%) missense, and 11 (2%) protein truncating.  
227 Furthermore, minor alleles at  $>10$ -fold increased frequency in FinMetSeq compared to  
228 NFE are substantially more likely to be associated with a trait compared to variants with  
229 similar or lower MAF in FinMetSeq compared to NFE ( $OR=4.92$ ,  $P=2.6 \times 10^{-5}$ ; **Extended**  
230 **Data Fig. 3**).

231

232 We clumped associated variants within 1 Mb and with  $r^2 > 0.5$  into a single locus,  
233 irrespective of the associated traits (Methods). After clumping, the 531 associated  
234 variants represented 262 distinct loci (597 trait-locus pairs, **Supplementary Table 6**);  
235 158 of the 262 loci (60%) consisted of a single trait-associated variant. As expected, the  
236 number of associated loci per trait was positively correlated with trait heritability ( $r=0.38$ ,  
237  $P=8.8 \times 10^{-4}$ ). Height was a noticeable outlier, with relatively few associations considering  
238 its high estimated heritability (**Fig. 2B**).

239

240 The majority of variants and loci (61%) were associated with a single trait; 4% were  
241 associated with  $\geq 10$  traits. Overlapping associations (**Fig. 2C**) strongly reflect the  
242 relationships exhibited by both trait and genetic correlations (**Fig. 1B**). For example,  
243 rs113298164, a missense variant in *LIPC* (p.Thr405Met), is associated with 11 traits,  
244 including cholesterol, fatty acids, apolipoproteins, and cholines. Similarly, the estimated  
245 genetic correlation of trait pairs is a strong predictor of the probability for a trait pair to  
246 share associated loci (**Fig. 2D**).

247

248 To determine which of the 1,249 single-variant associations were distinct from known  
249 GWAS associations for the same traits, we repeated association analysis for each trait  
250 conditional on published associated variants ( $P < 10^{-7}$ ) for the corresponding trait in the  
251 EBI GWAS Catalog (December 2016 release). Of the 1,249 trait-variant associations,  
252 478 (at 213 of 531 variants) remained significant ( $P < 5 \times 10^{-7}$ ) after conditional analysis,  
253 representing 126 of the original 262 loci, including at least one conditionally significant  
254 locus for each of 48 traits (**Supplementary Table 8**). The conditionally-associated  
255 variants were more often rare (24% vs. 11%), more likely to alter or truncate the resulting  
256 protein (31% vs. 22%), and more frequently  $>10$ x enriched in FinMetSeq relative to NFE  
257 (19% vs. 10%) compared to the full set of associated variants.

258

### 259 **Gene-based association analyses**

260 To identify genes associated with the 64 traits, we performed aggregate tests of protein  
261 coding variants, grouping variants using three different masks. Mask 1 comprised PTVs  
262 of any frequency; Masks 2 and 3 also included missense variants with  $MAF < 0.1\%$  or  
263 0.5% predicted to be deleterious by five algorithms (Methods). We identified 54 gene-  
264 based associations with  $P < 3.88 \times 10^{-6}$  (adjusting for testing a maximum of 12,890 genes  
265 containing at least two qualifying variants) and with multi-trait  $FDR < 0.05$ , analogous to  
266 the threshold used for single-variant association testing (Methods). Fifteen of these  
267 associations required  $\ge 2$  variants to achieve significance (i.e. the association was not  
268 driven by a single strongly associated variant; **Supplementary Table 9**). Extremely rare  
269 ( $MAC \le 3$ ) PTVs drove the association of eight traits with *APOB* (**Extended Data Fig. 4**).  
270 We found a novel association between two very rare stop gain variants in *SECTM1* and

271 HDL2 cholesterol ( $P=7.2\times10^{-7}$ , **Extended Data Fig. 5**). *SECTM1* encodes an interferon-  
272 induced transmembrane protein that is negatively regulated by bacterial  
273 lipopolysaccharide (LPS)<sup>18</sup>. The association could reflect the role of HDL particles in  
274 binding and neutralizing LPS in infections and sepsis<sup>19</sup>.

275

276 **Replication and follow-up of single-variant associations in three additional Finnish**  
277 **cohorts: Identification of novel coding, deleterious variant associations**

278 We attempted to replicate the 478 single-variant associations from FinMetSeq  
279 (unconditional and conditional  $P\leq5\times10^{-7}$ ) and to follow-up the 2,120 suggestive but sub-  
280 threshold associations from FinMetSeq (unconditional  $5\times10^{-7} < P \leq 5\times10^{-5}$ , conditional  
281  $P\leq5\times10^{-5}$ ) in 24,776 participants from three Finnish cohort studies for which varying  
282 subsets of the 64 FinMetSeq traits were available: FINRISK<sup>20,21</sup> participants not  
283 sequenced in FinMetSeq (n=18,215), the Northern Finland Birth Cohort 1966<sup>22</sup>  
284 (n=5,139), and the Helsinki Birth Cohort<sup>23</sup> (n=1,412). For each of the three cohorts, we  
285 carried out genotype imputation using the Finnish-specific SISu v2 reference panel  
286 (<http://www.sisuproject.fi>), which is comprised of 5,380 haplotypes from whole-genome  
287 based sequencing and 10,184 haplotypes from whole-exome based sequencing in coding  
288 regions, and then used the same single-variant association analysis strategy employed in  
289 FinMetSeq. We then carried out meta-analysis of the three imputation-based studies to  
290 test for replication of associated FinMetSeq variants (“replication analysis”) and four-  
291 study meta-analysis with FinMetSeq to follow-up suggestive associations (“combined  
292 analysis”; Methods).

293

294 We obtained data for 448 of the 478 significant variant-trait associations (191 of the 213  
295 requested variants). Of the 448 associations for which we had replication data, 439  
296 (98.0%) had the same direction of effect in replication analysis as in FinMetSeq; 392 of  
297 the 448 replicated at  $P<0.05$  (87.5%; **Supplementary Table 10**). We also obtained data  
298 to follow up 1,417 of the 2,120 sub-threshold associations (1,014 of the 1,554 requested  
299 variants); >60% of the variants that we could not follow up were very rare in FinMetSeq  
300 and were not present in the SISu reference panel. Of the 1,417 sub-threshold trait-variant  
301 associations, 431 reached  $P<5\times10^{-7}$  in the combined analysis (**Supplementary Table**  
302 **11**).

303

304 Among the significant results from FinMetSeq or combined analysis, 43 associations  
305 were with 26 predicted deleterious variants that conditional analysis and literature review  
306 suggest are novel (**Table 2**). Nineteen such associations, at 15 deleterious coding  
307 variants, were significant in FinMetSeq (**Table 2**; **Supplementary Table 10**). Twelve of  
308 these associations replicated ( $P<0.05$ ) in the replication analysis and remained significant  
309 in the combined analysis; for the other seven associations we either did not have  
310 replication data (six associations) or did not replicate but had very low power (<5%) in  
311 the replication analysis (one association). Four of the 15 variants were PTVs; 11 were  
312 missense variants predicted to be deleterious by at least one of five prediction algorithms.  
313 Another 24 associations, with 16 variants (two PTVs and 14 missense variants predicted  
314 to be deleterious), only reached significance in the combined analysis (**Table 2**;  
315 **Supplementary Table 11**). Five variants with significant associations in FinMetSeq  
316 alone were associated with additional traits in combined analysis (**Table 2**).

317

318 Of the 43 associations shown in **Table 2**, 34 were with 19 variants either seen only in  
319 Finland or enriched by >20-fold in FinMetSeq compared to NFE (13 of 15 variants in  
320 FinMetSeq and 11 of 16 variants in combined analysis with five variants overlapping).  
321 Identifying associations for these 19 variants would have required much larger samples in  
322 NFE populations than in FinMetSeq (**Fig. 3A & B**). We provide brief summaries relating  
323 each of these highly enriched associations to known biology and prior genetic evidence  
324 relating to the respective genes in **Supplementary Information**. We highlight a few of  
325 the most striking findings, below.

326

327 *Anthropometric traits.* As a group these are among the most extensively investigated  
328 quantitative traits, with thousands of common variant associations reported, most of very  
329 small effect<sup>24-28</sup>. We identified several rare, large effect variants for these traits, including  
330 a predicted damaging missense variant (rs200373343, p.Arg94Cys) in *THBS4* 45X more  
331 frequent in FinMetSeq than in NFE and associated in the combined analysis with a mean  
332 decrease in body weight of 5.9 kg (**Table 2**). *THBS4* encodes thrombospondin 4, a  
333 matricellular protein found in blood vessel walls and highly expressed in heart and  
334 adipose tissue<sup>29</sup>. *THBS4* is involved in local signaling in the developing and adult nervous  
335 system, and may function in regulating vascular inflammation<sup>30</sup>. Coding variants in  
336 *THBS4* and other thrombospondin genes have been implicated in increased risk for heart  
337 disease<sup>31-33</sup>.

338

339 We identified a predicted damaging missense variant (rs2273607, p.Val104Met) in *DLK1*  
340 that is 177X more frequent in FinMetSeq than in NFE and is associated in the combined  
341 analysis with a mean decrease in height of 1.3 cm (**Table 2**). *DLK1* encodes Delta-Like  
342 Notch Ligand 1, an epidermal growth factor that interacts with fibronectin and inhibits  
343 adipocyte differentiation. Uniparental disomy of *DLK1* causes Temple Syndrome and  
344 Kagami-Ogata Syndrome, characterized by pre- and postnatal growth restriction,  
345 hypotonia, joint laxity, motor delay, and early onset of puberty<sup>34-36</sup>. Paternally-inherited  
346 common variants near *DLK1* have been associated with child and adolescent obesity,  
347 type 1 diabetes, age at menarche, and central precocious puberty in girls<sup>37-39</sup>.  
348 Homozygous null mutations in the mouse ortholog Dlk-1 lead to embryos with reduced  
349 size, skeletal length, and lean mass<sup>40</sup>, while in Darwin's finches, SNVs at this locus have  
350 a strong effect on beak size<sup>41</sup>.

351  
352 *HDL-C*. Two novel variants with large effects on HDL-C in FinMetSeq are absent in  
353 NFE. The predicted deleterious missense variant rs750623950 (p.Arg112Trp) in  
354 *CD300LG* is associated in FinMetSeq with a mean increase in HDL-C of 0.95 mmol/l,  
355 and also associated with HDL2-C and ApoA1 (**Table 2**). *CD300LG* encodes a type I cell  
356 surface glycoprotein. A missense variant in *ABCA1* (rs765246726, p.Cys2107Arg) is  
357 associated in FinMetSeq with a mean reduction in HDL-C of 0.64 mmol/l (**Table 2**).  
358 Fifteen more variants (including ten which are absent in NFE) contributed to a strong  
359 *ABCA1* gene-based association signal (P=2.2×10<sup>-13</sup>; **Supplementary Table 9, Extended**  
360 **Data Fig. 6**). *ABCA1* encodes the cholesterol efflux regulatory protein, which regulates  
361 cholesterol and phospholipid metabolism. Individuals who are homozygotes or

362 compound heterozygotes for any of several *ABCA1* mutations produce very little HDL-C  
363 and experience the manifestations of severe hypercholesterolemia.

364

365 *Amino Acids.* A stop gain variant (rs780671030, p.Arg722X) in *ALDH1L1* is associated  
366 in FinMetSeq with a mean reduction in serum glycine levels of 0.03 mmol/l but is not  
367 observed in NFE (**Table 2**); this effect may increase risk for several cardiometabolic  
368 disorders<sup>42,43</sup>. *ALDH1L1* encodes 10-formyltetrahydrofolate dehydrogenase, which  
369 competes with the enzyme serine hydroxymethyltransferase to alter the ratio of serine to  
370 glycine in the cytosol. Although rs780671030 was the strongest associated variant, gene-  
371 based association tests suggest that additional PTVs and missense variants in *ALDH1L1*  
372 also alter glycine levels ( $P=1.4\times10^{-20}$ , **Extended Data Fig. 7, Supplementary Table 9**).

373

374 *Ketone bodies.* A predicted damaging missense variant (rs201013770, p.Phe517Ser) in  
375 *ACSSI1* is associated in the combined analysis with mean increased serum acetate level of  
376 0.005 mmol/l but is not observed in NFE (**Table 2**). *ACSSI1* encodes an acyl-coenzyme A  
377 synthetase and plays a role in the conversion of acetate to acetyl-CoA. In rodents,  
378 increased acetate levels lead to obesity, insulin resistance, and metabolic syndrome,  
379 mediated by activation of the parasympathetic nervous system<sup>44</sup>.

380

### 381 **Associated variants and disease endpoints**

382 Newly available GWAS data from the FinnGen project<sup>45</sup> enabled us to test the hypothesis  
383 that deleterious variants responsible for our novel quantitative trait associations (**Table 2**)  
384 could also contribute to disease endpoints related to these traits. FinnGen has particularly

385 rich data on such endpoints as the samples are largely drawn from Finnish hospital  
386 biobanks. In total, we examined 22 disease endpoint phenotypes for all 25 available  
387 variants in **Table 2**. Three variants were associated with disease endpoints in FinnGen at  
388 a Bonferroni-corrected threshold of  $P<0.05/(22\times25)=9.0\times10^{-5}$  (**Supplementary Table**  
389 **12**).

390

391 A predicted damaging missense variant (17:39135270:A/G; p.Ser32Pro) in *KRT40* which  
392 is not observed in NFE and associated in FinMetSeq with a mean elevation in HDL-C of  
393 1.07 mmol/l (**Table 2**), is associated in FinnGen with increased risk for pancreatitis.  
394 While this is the first disease association reported for this gene, the type I keratin family,  
395 of which *KRT40* is a member, is believed to play an important role in regulating exocrine  
396 pancreas homoeostasis<sup>46</sup>. A 29 bp deletion on chromosome 1 causes a frameshift in  
397 *FAM151A* which is 6.7X more frequent in FinMetSeq than NFE and associated in  
398 FinMetSeq with both decreased total cholesterol in IDL and decreased IDL particle  
399 concentration (**Table 2**), is associated in FinnGen with decreased risk of myocardial  
400 infarction. The interpretation of this association is complicated by the fact that the variant  
401 is also present in an overlapping transcript (*ACOT11*), a gene that plays a role in fatty  
402 acid metabolism and lies <1 MB from a well-known cardioprotective variant in *PCSK9*.  
403 Finally, a predicted damaging missense variant (rs77273740; p.Arg65Trp) in *DBH* that is  
404 23.8X more frequent in FinMetSeq than in NFE and is associated with a mean decrease  
405 of 1 mmHg in diastolic blood pressure in our combined analysis (**Table 2**), is associated  
406 in FinnGen with decreased risk for hypertension. Distinct loci in this gene have

407 previously been shown with mean arterial pressure and this variant was included in a  
408 gene-based association with mean arterial pressure<sup>5,6</sup>.

409

#### 410 **Replication outside of Finland: UK Biobank**

411 To begin to assess the generalizability outside of Finland of the novel associations that  
412 we detected, we attempted to replicate associations from our combined Finnish analyses  
413 in the UK Biobank (UKBB), a European sample that is approximately ten-fold larger.  
414 Across eight anthropometric and blood pressure traits for which UKBB data are publicly  
415 available, our Finnish combined analysis had identified 31 trait-variant associations  
416 reaching  $P<5\times10^{-7}$ . More than a quarter of these variants (8 of 31) were not present in the  
417 UKBB database. Of the remaining 23 associations, 20 were to variants that were common  
418 in FinMetSeq (MAF > 1%) and had a comparable frequency in UKBB; 15 (75%) of these  
419 variants showed association in UKBB at  $P<0.05/23=2.2\times10^{-3}$  (Bonferroni correction for  
420 23 tests). Of the three rare variants in this analysis, all of which were enriched at  $>10\times$   
421 frequency in FinMetSeq compared to UKBB, none showed association in UKBB  
422 (**Supplementary Table 13**). Even after adjusting for winner's curse<sup>47</sup> and with a sample  
423 size of 340,000-360,000, we had <50% power to detect all three of these associations in  
424 UKBB (**Supplementary Table 13**). This comparison supports the argument that  
425 extremely large samples will be needed in most other populations to achieve the power  
426 for rare variant association studies that we have observed in Finland.

427

#### 428 **Geographical clustering of associated variants**

429 Given the concentration within sub-regions of northern and eastern Finland of most  
430 Finnish Disease Heritage mutations<sup>48</sup>, we hypothesized that the distribution of rare trait-  
431 associated variants discovered through FinMetSeq might also display geographical  
432 clustering. In support of this hypothesis, principal component analysis revealed broad-  
433 scale population structure within the late-settlement region among 14,874 unrelated  
434 FinMetSeq participants whose parental birthplaces are known (**Extended Data Fig. 8**).  
435 Consistent with our hypothesis, parental birthplaces were significantly more  
436 geographically clustered for carriers of PTVs and missense alleles than for carriers of  
437 synonymous alleles, even after adjusting for MAC (**Supplementary Tables 14A, 14B**).  
438

439 To enable finer scale analysis of the distribution of variants within late-settlement  
440 Finland, we delineated geographically distinct population clusters using patterns of  
441 haplotype sharing among 2,644 unrelated individuals with both parents known to be born  
442 in the same municipality (Methods, **Extended Data Fig. 9**). Taking the cluster that is  
443 most genetically similar to early-settlement Finland as a reference, we compared variant  
444 counts for different functional classes and frequencies between this reference cluster and  
445 each of the other 12 clusters that contained  $\geq 100$  individuals (**Fig. 4, Supplementary**  
446 **Tables 15, 16**). In the two clusters that represent the most heavily bottlenecked late-  
447 settlement regions (Lapland and Northern Ostrobothnia), we observed a marked deficit of  
448 singletons and significant enrichment of variants at intermediate frequency compared to  
449 other clusters. This pattern is most significant for missense variants, which are present in  
450 the exome in large numbers; PTVs show consistently greater enrichment, but with less  
451 statistical significance likely due to very small counts (**Fig. 4**). Two clusters in which we

452 observed marked enrichment of singletons, Surrendered Karelia and South Ostrobothnia,  
453 showed the highest levels of genetic diversity across the frequency spectrum, likely  
454 reflecting relatively recent gene flow into these regions from neighboring countries  
455 (Russia and Sweden, respectively, **Fig. 4**).

456

457 We observed particularly strong geographical clustering among variants >10X enriched  
458 in FinMetSeq compared to NFE (**Fig. 5A, Extended Data Fig. 10, Supplementary**  
459 **Table 17**). We further characterized geographical clustering for FinMetSeq-enriched  
460 trait-associated variants, by comparing the mean distances between birthplaces for  
461 parents of minor allele carriers to those of non-carriers (**Supplementary Table 18**). Most  
462 such variants were highly localized. For example, for variant rs780671030 in *ALDH1L1*,  
463 which may be unique to Finns, the mean distance between parental birthplaces is 135 km  
464 for carriers and 250 km for non-carriers ( $P<1.0\times10^{-7}$ , **Fig. 5B**). In contrast, few of the  
465 variants that displayed sub-threshold association in FinMetSeq but that showed  
466 significant associations in the combined analysis were significantly geographically  
467 clustered within Finland (**Supplementary Table 18**).

468

469 Finally, we compared the geographic clustering of FinMetSeq-enriched trait-associated  
470 variants to that of 35 Finnish Disease Heritage mutations carried by  $\geq 3$  FinMetSeq  
471 individuals with known parental birthplaces. FinMetSeq carriers of monogenic Finnish  
472 Disease Heritage mutations and FinMetSeq carriers of trait-associated variants identified  
473 in FinMetSeq displayed a comparable degree of geographic clustering. This clustering  
474 was dramatically greater than that observed for the non-carriers of both sets of variants

475 (Fig. 5C), suggesting that rare variants associated with complex traits may be much more  
476 unevenly distributed geographically than has been appreciated to date.

477

## 478 **DISCUSSION**

479 We have demonstrated that a well-powered exome sequencing study of deeply  
480 phenotyped individuals can identify numerous rare variants associated with medically  
481 relevant quantitative traits. The variants that we identified may provide a useful starting  
482 point for studies aimed at uncovering biological mechanisms or fostering clinical  
483 translation. For example, further investigation of the p.Arg722X variant in *ALDH1L1*  
484 associated with reduced serum glycine could help elucidate the role of this gene in  
485 astrocyte function, a topic of growing interest in neurobiology. Glycine is a key  
486 inhibitory neurotransmitter localized to astrocytes<sup>49</sup>, while the mouse ortholog, *Aldh1l1*,  
487 is the primary marker for astrocytes in experimental research, since it is strongly  
488 expressed in astrocytes, but not in neurons<sup>50</sup>.

489

490 The substantial power of this study for discovering rare variant associations derives from  
491 the occurrence, in the recently expanded and heavily bottlenecked populations of  
492 northern and eastern Finland, of a large pool of deleterious variants that appear unique to  
493 Finland or at frequencies orders of magnitude greater than in NFE. This observation  
494 motivates a strategy for scaling up the discovery of rare variant associations by  
495 prioritizing the sequencing of populations beyond Finland that have expanded in isolation  
496 from recent bottlenecks. Examples of other such populations include those of  
497 Ashkenazim<sup>51</sup>, Iceland<sup>52</sup>, Quebec<sup>53</sup>, highland regions of Latin America<sup>54</sup>, and

498 geographically isolated regions of larger European countries such as Sardinia<sup>55</sup> and  
499 Crete<sup>10</sup>. In each of these populations, genetic drift has almost certainly caused a different  
500 set of alleles to pass through the corresponding population-specific bottlenecks, enriching  
501 some variants while depleting others. The numerous rare-variant associations that could  
502 be identified by sequencing available, phenotyped samples across multiple population  
503 isolates could rapidly increase our understanding of the genetic architecture of complex  
504 traits. One caveat is that the extended LD blocks that are typical in such populations may  
505 make it difficult to identify the causative variant from among multiple deleterious  
506 variants within an association region<sup>56</sup>.

507

508 Recent studies have suggested a continuity between the genetic architectures of complex  
509 traits and disorders classically considered monogenic<sup>57,58</sup>. Our results offer strong support  
510 for this continuity, not only in identifying numerous deleterious variants with large  
511 effects on quantitative traits, but in demonstrating that such variants show geographical  
512 clustering comparable to that of the mutations responsible for the Finnish Disease  
513 Heritage.

514

515 The use of a Finland-specific genotype reference panel<sup>59</sup> to impute FinMetSeq variants  
516 into array-genotyped samples from three other Finnish cohorts enabled us to identify  
517 many additional novel associations. This result suggests that using sequence data from a  
518 subset of individuals in each population to impute variants in array-genotyped samples  
519 from the same population is a cost-effective strategy for detecting rare-variant  
520 associations. However, the clustering in FinMetSeq of deleterious trait-associated

521 variants within limited geographical regions and our inability to follow-up >700 sub-  
522 threshold associations from FinMetSeq for which the associated variants were not present  
523 in the Finnish imputation reference panel, emphasize the importance of extensively  
524 representing regional subpopulations when designing such reference panels, to account  
525 for fine-scale population structure.

526

527 To fully realize the value of large-scale sequencing studies in population isolates, it will  
528 be necessary to increase the richness of phenotypes available in sequenced cohorts from  
529 these populations. For example, we associated <100 of the >24,000 deleterious, highly  
530 enriched variants identified in FinMetSeq with one of the 64 cardiometabolic-related  
531 quantitative traits studied here. In Finland, the national health care system and the  
532 population's willingness to participate in biomedical research mean that extensive  
533 medical records and population registries are available for mining additional phenotype  
534 data, and create an opportunity for callback by genotype for further phenotyping and  
535 collection of biological samples<sup>60</sup>. Notably, the associations we identified to disease  
536 endpoints in FinnGen give a hint of the discoveries that will be possible when that  
537 database reaches its full size of 500,000 participants. The insights gained from such  
538 efforts will accelerate the implementation of precision health, informing projects in  
539 larger, more heterogeneous populations which are still at an early stage<sup>61</sup>.

540 **References**

541 1 MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide  
542 association studies (GWAS Catalog). *Nucleic Acids Res* **45**, D896-D901,  
543 doi:10.1093/nar/gkw1133 (2017).

544 2 Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature*  
545 **461**, 747-753, doi:10.1038/nature08494 (2009).

546 3 Mahajan, A. *et al.* Refining the accuracy of validated target identification through  
547 coding variant fine-mapping in type 2 diabetes. *Nature genetics* **50**, 559-571,  
548 doi:10.1038/s41588-018-0084-1 (2018).

549 4 Turcot, V. *et al.* Protein-altering variants associated with body mass index  
550 implicate pathways that control energy intake and expenditure in obesity. *Nature*  
551 *genetics* **50**, 26-41, doi:10.1038/s41588-017-0011-x (2018).

552 5 Liu, C. *et al.* Meta-analysis identifies common and rare variants influencing blood  
553 pressure and overlapping with metabolic trait loci. *Nature genetics* **48**, 1162-  
554 1170, doi:10.1038/ng.3660 (2016).

555 6 Surendran, P. *et al.* Trans-ancestry meta-analyses identify rare and common  
556 variants associated with blood pressure and hypertension. *Nature genetics* **48**,  
557 1151-1161, doi:10.1038/ng.3654 (2016).

558 7 Zuk, O. *et al.* Searching for missing heritability: designing rare variant association  
559 studies. *Proc Natl Acad Sci U S A* **111**, E455-464, doi:10.1073/pnas.1322563111  
560 (2014).

561 8 Xue, Y. *et al.* Enrichment of low-frequency functional variants revealed by  
562 whole-genome sequencing of multiple isolated European populations. *Nature*  
563 *communications* **8**, 15927, doi:10.1038/ncomms15927 (2017).

564 9 Ganna, A. *et al.* Quantifying the Impact of Rare and Ultra-rare Coding Variation  
565 across the Phenotypic Spectrum. *American journal of human genetics* **102**, 1204-  
566 1211, doi:10.1016/j.ajhg.2018.05.002 (2018).

567 10 Southam, L. *et al.* Whole genome sequencing and imputation in isolated  
568 populations identify genetic associations with medically-relevant complex traits.  
569 *Nature communications* **8**, 15606, doi:10.1038/ncomms15606 (2017).

570 11 Jakkula, E. *et al.* The genome-wide patterns of variation expose significant  
571 substructure in a founder population. *Am J Hum Genet* **83**, 787-794,  
572 doi:10.1016/j.ajhg.2008.11.005 (2008).

573 12 Polvi, A. *et al.* The Finnish disease heritage database (FinDis) update-a database  
574 for the genes mutated in the Finnish disease heritage brought to the next-  
575 generation sequencing era. *Hum Mutat* **34**, 1458-1466, doi:10.1002/humu.22389  
576 (2013).

577 13 Manning, A. *et al.* A Low-Frequency Inactivating AKT2 Variant Enriched in the  
578 Finnish Population Is Associated With Fasting Insulin Levels and Type 2  
579 Diabetes Risk. *Diabetes* **66**, 2019-2032, doi:10.2337/db16-1329 (2017).

580 14 Lim, E. T. *et al.* Distribution and medical impact of loss-of-function variants in  
581 the Finnish founder population. *PLoS genetics* **10**, e1004494,  
582 doi:10.1371/journal.pgen.1004494 (2014).

583 15 Service, S. K. *et al.* Re-sequencing expands our understanding of the phenotypic  
584 impact of variants at GWAS loci. *PLoS genetics* **10**, e1004147,  
585 doi:10.1371/journal.pgen.1004147 (2014).

586 16 Wurtz, P. *et al.* Metabolite profiling and cardiovascular event risk: a prospective  
587 study of 3 population-based cohorts. *Circulation* **131**, 774-785,  
588 doi:10.1161/CIRCULATIONAHA.114.013116 (2015).

589 17 Laakso, M. *et al.* The Metabolic Syndrome in Men study: a resource for studies of  
590 metabolic and cardiovascular diseases. *Journal of lipid research* **58**, 481-493,  
591 doi:10.1194/jlr.O072629 (2017).

592 18 Huyton, T., Gottmann, W., Bade-Doding, C., Paine, A. & Blasczyk, R. The T/NK  
593 cell co-stimulatory molecule SECTM1 is an IFN "early response gene" that is  
594 negatively regulated by LPS in human monocytic cells. *Biochim Biophys Acta*  
595 **1810**, 1294-1301, doi:10.1016/j.bbagen.2011.06.020 (2011).

596 19 Pirillo, A., Catapano, A. L. & Norata, G. D. HDL in infectious diseases and  
597 sepsis. *Handb Exp Pharmacol* **224**, 483-508, doi:10.1007/978-3-319-09665-0\_15  
598 (2015).

599 20 Borodulin, K. *et al.* Forty-year trends in cardiovascular risk factors in Finland.  
600 *Eur J Public Health* **25**, 539-546, doi:10.1093/eurpub/cku174 (2015).

601 21 Abraham, G. *et al.* Genomic prediction of coronary heart disease. *Eur Heart J* **37**,  
602 3267-3278, doi:10.1093/eurheartj/ehw450 (2016).

603 22 Sabatti, C. *et al.* Genome-wide association analysis of metabolic traits in a birth  
604 cohort from a founder population. *Nature genetics* **41**, 35-46, doi:10.1038/ng.271  
605 (2009).

606 23 Pulizzi, N. *et al.* Interaction between prenatal growth and high-risk genotypes in  
607 the development of type 2 diabetes. *Diabetologia* **52**, 825-829,  
608 doi:10.1007/s00125-009-1291-1 (2009).

609 24 Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for  
610 obesity biology. *Nature* **518**, 197-206, doi:10.1038/nature14177 (2015).

611 25 Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat  
612 distribution. *Nature* **518**, 187-196, doi:10.1038/nature14132 (2015).

613 26 Wood, A. R. *et al.* Defining the role of common variation in the genomic and  
614 biological architecture of adult human height. *Nature genetics* **46**, 1173-1186,  
615 doi:10.1038/ng.3097 (2014).

616 27 Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and  
617 body mass index in approximately 700000 individuals of European ancestry. *Hum  
618 Mol Genet*, doi:10.1093/hmg/ddy271 (2018).

619 28 Pulit, S. L. *et al.* Meta-analysis of genome-wide association studies for body fat  
620 distribution in 694,649 individuals of European ancestry. *Hum Mol Genet*,  
621 doi:10.1093/hmg/ddy327 (2018).

622 29 Fagerberg, L. *et al.* Analysis of the human tissue-specific expression by genome-  
623 wide integration of transcriptomics and antibody-based proteomics. *Mol Cell  
624 Proteomics* **13**, 397-406, doi:10.1074/mcp.M113.035600 (2014).

625 30 Corsetti, J. P. *et al.* Thrombospondin-4 polymorphism (A387P) predicts  
626 cardiovascular risk in postinfarction patients with high HDL cholesterol and C-  
627 reactive protein levels. *Thromb Haemost* **106**, 1170-1178, doi:10.1160/TH11-03-  
628 0206 (2011).

629 31 Cui, J. *et al.* Thrombospondin-4 1186G>C (A387P) is a sex-dependent risk factor  
630 for myocardial infarction: a large replication study with increased sample size

631 from the same population. *Am Heart J* **152**, 543 e541-545,  
632 doi:10.1016/j.ahj.2006.06.002 (2006).

633 32 Wessel, J., Topol, E. J., Ji, M., Meyer, J. & McCarthy, J. J. Replication of the  
634 association between the thrombospondin-4 A387P polymorphism and myocardial  
635 infarction. *Am Heart J* **147**, 905-909, doi:10.1016/j.ahj.2003.12.013 (2004).

636 33 Zhang, X. J. *et al.* Association between single nucleotide polymorphisms in  
637 thrombospondins genes and coronary artery disease: A meta-analysis. *Thromb  
638 Res* **136**, 45-51, doi:10.1016/j.thromres.2015.04.019 (2015).

639 34 Beygo, J. *et al.* New insights into the imprinted MEG8-DMR in 14q32 and  
640 clinical and molecular description of novel patients with Temple syndrome. *Eur J  
641 Hum Genet* **25**, 935-945, doi:10.1038/ejhg.2017.91 (2017).

642 35 Prats-Puig, A. *et al.* The placental imprinted DLK1-DIO3 domain: a new link to  
643 prenatal and postnatal growth in humans. *Am J Obstet Gynecol* **217**, 350 e351-350  
644 e313, doi:10.1016/j.ajog.2017.05.002 (2017).

645 36 Rosenfeld, J. A. *et al.* Clinical features associated with copy number variations of  
646 the 14q32 imprinted gene cluster. *Am J Med Genet A* **167A**, 345-353 (2015).

647 37 Wallace, C. *et al.* The imprinted DLK1-MEG3 gene region on chromosome  
648 14q32.2 alters susceptibility to type 1 diabetes. *Nature genetics* **42**, 68-71,  
649 doi:10.1038/ng.493 (2010).

650 38 Day, F. R. *et al.* Genomic analyses identify hundreds of variants associated with  
651 age at menarche and support a role for puberty timing in cancer risk. *Nature  
652 genetics* **49**, 834-841, doi:10.1038/ng.3841 (2017).

653 39 Perry, J. R. *et al.* Parent-of-origin-specific allelic associations among 106  
654 genomic loci for age at menarche. *Nature* **514**, 92-97, doi:10.1038/nature13545  
655 (2014).

656 40 Cleaton, M. A. *et al.* Fetus-derived DLK1 is required for maternal metabolic  
657 adaptations to pregnancy and is associated with fetal growth restriction. *Nature  
658 genetics* **48**, 1473-1480, doi:10.1038/ng.3699 (2016).

659 41 Chaves, J. A. *et al.* Genomic variation at the tips of the adaptive radiation of  
660 Darwin's finches. *Mol Ecol* **25**, 5282-5295, doi:10.1111/mec.13743 (2016).

661 42 Ding, Y. *et al.* Plasma Glycine and Risk of Acute Myocardial Infarction in  
662 Patients With Suspected Stable Angina Pectoris. *J Am Heart Assoc* **5**,  
663 doi:10.1161/JAHA.115.002621 (2015).

664 43 Wittemans, L. B. L. *et al.* Assessing the causal association of glycine with risk of  
665 cardio-metabolic diseases. *Nature communications* **10**, 1060, doi:10.1038/s41467-  
666 019-08936-1 (2019).

667 44 Perry, R. J. *et al.* Acetate mediates a microbiome-brain-beta-cell axis to promote  
668 metabolic syndrome. *Nature* **534**, 213-217, doi:10.1038/nature18309 (2016).

669 45 Tabbassum, R. *et al.* Genetics of human plasma lipidome: Understanding lipid  
670 metabolism and its link to diseases beyond traditional lipids. *bioRxiv*,  
671 doi:10.1101/457960 (2018).

672 46 Casanova, M. L. *et al.* Exocrine pancreatic disorders in transgenic mice  
673 expressing human keratin 8. *J Clin Invest* **103**, 1587-1595, doi:10.1172/JCI5343  
674 (1999).

675 47 Palmer, C. & Pe'er, I. Statistical correction of the Winner's Curse explains  
676 replication variability in quantitative trait genome-wide association studies. *PLoS*  
677 *genetics* **13**, e1006916, doi:10.1371/journal.pgen.1006916 (2017).

678 48 Norio, R. Finnish Disease Heritage I: characteristics, causes, background. *Hum*  
679 *Genet* **112**, 441-456, doi:10.1007/s00439-002-0875-3 (2003).

680 49 Zeisel, A. *et al.* Molecular Architecture of the Mouse Nervous System. *Cell* **174**,  
681 999-1014 e1022, doi:10.1016/j.cell.2018.06.021 (2018).

682 50 Chai, H. *et al.* Neural Circuit-Specialized Astrocytes: Transcriptomic, Proteomic,  
683 Morphological, and Functional Evidence. *Neuron* **95**, 531-549 e539,  
684 doi:10.1016/j.neuron.2017.06.029 (2017).

685 51 Rivas, M. A. *et al.* Insights into the genetic epidemiology of Crohn's and rare  
686 diseases in the Ashkenazi Jewish population. *PLoS genetics* **14**, e1007329,  
687 doi:10.1371/journal.pgen.1007329 (2018).

688 52 Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic  
689 population. *Nature genetics* **47**, 435-444, doi:10.1038/ng.3247 (2015).

690 53 Peischl, S. *et al.* Relaxed Selection During a Recent Human Expansion. *Genetics*  
691 **208**, 763-777, doi:10.1534/genetics.117.300551 (2018).

692 54 Service, S. *et al.* Magnitude and distribution of linkage disequilibrium in  
693 population isolates and implications for genome-wide association studies. *Nature*  
694 *genetics* **38**, 556-560, doi:10.1038/ng1770 (2006).

695 55 Chiang, C. W. K. *et al.* Genomic history of the Sardinian population. *Nature*  
696 *genetics*, doi:10.1038/s41588-018-0215-8 (2018).

697 56 Peltonen, L., Palotie, A. & Lange, K. Use of population isolates for mapping  
698 complex traits. *Nature reviews. Genetics* **1**, 182-190, doi:10.1038/35042049  
699 (2000).

700 57 Bastarache, L. *et al.* Phenotype risk scores identify patients with unrecognized  
701 Mendelian disease patterns. *Science* **359**, 1233-1239, doi:10.1126/science.aal4043  
702 (2018).

703 58 Niemi, M. E. K. *et al.* Common genetic variants contribute to risk of rare severe  
704 neurodevelopmental disorders. *Nature*, doi:10.1038/s41586-018-0566-4 (2018).

705 59 Surakka, I. S., A.-P.; Ruotsalainen, S.E.; Durbin, R.; Salomaa, V.; Daly, M.;  
706 Palotie, A.; Ripatti, S. The rate of false polymorphisms introduced when imputing  
707 genotypes from global imputation panels. *bioRxiv*, doi:10.1101/080770 (2016).

708 60 Latva-Rasku, A. *et al.* A Partial Loss-of-Function Variant in AKT2 Is Associated  
709 With Reduced Insulin-Mediated Glucose Uptake in Multiple Insulin-Sensitive  
710 Tissues: A Genotype-Based Callback Positron Emission Tomography Study.  
711 *Diabetes* **67**, 334-342, doi:10.2337/db17-1142 (2018).

712 61 Collins, F. S. & Varmus, H. A new initiative on precision medicine. *N Engl J Med*  
713 **372**, 793-795, doi:10.1056/NEJMp1500523 (2015).

714  
715 **Acknowledgements**

716 Thanks to Terri Teshiba for coordinating ethical permissions and samples. Thanks to Sini  
717 Kerminen, Daniel Lawson, and George Busby for discussions and providing scripts to

718 run fineSTRUCTURE. SR was supported by the Academy of Finland Center of  
719 Excellence in Complex Disease Genetics (Grant No 312062), Academy of Finland (No.  
720 285380), the Finnish Foundation for Cardiovascular Research, the Sigrid Juselius  
721 Foundation, Biocentrum Helsinki and University of Helsinki HiLIFE Fellow grant. VR  
722 acknowledges support by RFBR, research project No. 18-04-00789 A. VS was supported  
723 by the Finnish Foundation for Cardiovascular Research. CS and LS received funding  
724 from HG006695, HL113315, and MH105578. MAK is supported by a Senior Research  
725 Fellowship from the National Health and Medical Research Council (NHMRC) of  
726 Australia (APP1158958). He also works in a unit that is supported by the University of  
727 Bristol and UK Medical Research Council (MC\_UU\_12013/1). The Baker Institute is  
728 supported in part by the Victorian Government's Operational Infrastructure Support  
729 Program. AUJ, DR, LJS, HMS, RW, PY, XY, and MB received funding from  
730 DK062370. SKS, CWKC, and NBF received funding from HL113315 and NS062691.  
731 The METSIM study was supported by grants from Academy of Finland (No. 321428),  
732 the Sigrid Juselius Foundation, the Finnish Foundation for Cardiovascular Research,  
733 Kuopio University Hospital, and Centre of Excellence of Cardiovascular and Metabolic  
734 Diseases supported by the Academy of Finland (ML). Sequencing was funded by  
735 5U54HG003079, and AEL, KMS, HJB, CCC, CJK, KLK, DCK, DEL, JN, TJN, SKD,  
736 NOS, IMH, and RKW were funded by 5U54HG003079 and 5UM1HG008853-03.  
737

### 738 **Author Contributions**

739 AEL, LJS, RKW, AaP, VS, ML, SR, MB, and NBF designed the study. AEL, KMS,  
740 HJA, RSF, DCK, DEL, JN, TJN, and JV produced and quality-controlled the sequence

741 data. AEL, AUJ, ArP, HMS, MAK, VS, and ML produced and quality-controlled the  
742 clinical data. AEL, KMS, CWKC, SKS, ASH, LS, MP, CCC, AUJ, CJK, KK, VR, DR,  
743 JV, RW, PY, and XY analyzed data. JGE, MAK, MRJ, and MM provided replication  
744 data. HL, SKD, NOS, IMH, CS, SR, MB, and NBF supervised experiments and analyses.  
745 AEL, KMS, CWKC, SKS, CS, MB and NBF wrote the paper. AEL, KMS, CWKC, and  
746 SKS contributed equally to this work. NBF and MB jointly supervised this work.

747

748 Competing interests statements:

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

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

754

755 Correspondence and requests for materials should be addressed to  
756 [nfreimer@mednet.ucla.edu](mailto:nfreimer@mednet.ucla.edu) or [boehnke@umich.edu](mailto:boehnke@umich.edu).

757

758 Data Availability: The sequence data can be accessed through dbGaP using the following  
759 study numbers: FINRISK: phs000756, METSIM: phs000752. Association results can be  
760 accessed at <http://pheweb.sph.umich.edu/FinMetSeq/>.

**Table 1.** Sequence variants identified using whole-exome sequencing of 19,292 FinMetSeq participants. Percentages are the percent of all variants in the given category to either have MAF <1% or to be singleton variants.

Variant type	All variants	MAF<1%	Singleton variants
SNV	1,318,781	87.5%	40.5%
Insertion/Deletion	92,776	87.0%	43.1%
Predicted LoF	33,156	96.4%	55.0%
Non-synonymous	353,228	92.2%	46.4%

Variant Annotation	All variants	MAF<1%	Singleton variants
Splice Acceptor	3,180	95.4%	50.8%
Splice Donor	3,795	96.2%	53.3%
Stop Gain	11,382	97.3%	54.3%
Frameshift	12,845	96.6%	58.2%
Stop Loss	621	88.1%	48.1%
Initiator Codon/Start Loss	1,333	93.6%	49.1%
Inframe Insertion	1,673	90.3%	44.5%
Inframe Deletion	4,936	92.9%	46.8%
Missense	353,228	92.3%	46.4%
Splice Region	40,248	87.1%	41.2%
Incomplete Terminal Codon	16	81.3%	50.0%
Stop Retained	217	86.2%	42.4%
Synonymous	180,104	85.7%	40.0%
Coding Sequence	78	88.5%	41.0%
Mature miRNA	239	92.9%	48.5%
5' UTR	35,572	87.8%	38.2%
3' UTR	66,539	86.2%	38.6%
Non-coding Exonic	82,126	85.8%	37.8%
Intronic	601,362	85.1%	37.4%
Upstream	8,820	86.5%	38.3%
Downstream	3,050	84.6%	38.3%
Intergenic	193	85.5%	31.1%

Variant annotation refers to the "most deleterious" annotation for a given variant across all Ensembl (v88) transcripts, following the order defined by VEP ([https://useast.ensembl.org/info/genome/variation/prediction/predicted\\_data.html](https://useast.ensembl.org/info/genome/variation/prediction/predicted_data.html)).

**Table 2.** Associations with predicted deleterious variants that conditional analysis and literature review suggest are novel. These associations reach exome-wide significance in FinMetSeq alone or in a combined analysis of FinMetSeq with three replication cohorts.

Chr:Pos (GRCh37)	Gene	Anno: Prediction^	FMS MAF	NFE MAF#	MAF Ratio (95% CI)	Trait	FMS P	FMS Beta	Repl. or combined P**	Repl. or combined Beta	Mean in carriers   non-carriers
1:55076137	<i>FAM151A</i>	FS:PTV	.099	.0147	6.7 (5.6-7.8)	Total Chol. in IDL IDL Particle Conc.	$5.4 \times 10^{-16}$ $8.9 \times 10^{-14}$	-0.187 -0.172	<b>2.1 <math>\times 10^{-17}</math></b> <b>1.9 <math>\times 10^{-16}</math></b>	-0.191 -0.185	.84   .87 mmol/l .130   .134 umol/l
2:120848049	<i>EPB41L5</i>	MS:T;B;B;N;D	.085	.044	1.9 (0.9-3)	eGFR* Creatinine*	$1.7 \times 10^{-6}$ $2.5 \times 10^{-6}$	-0.093 0.091	$4.8 \times 10^{-12}$ $2.5 \times 10^{-12}$	-0.107 0.098	88.6   89.9 SI 81.62   80.64 umol/l
3:125831672	<i>ALDH1L1</i>	SG:PTV	.0026	0	$\infty$	Glycine	$1.8 \times 10^{-8}$	-0.873	<b>4.5 <math>\times 10^{-4}</math></b>	-0.827	.24   .27 mmol/l
4:13612630	<i>BODIL1</i>	MS:D;D;D;N;D	.0001	0	$\infty$	WHR adj. BMI	$4.7 \times 10^{-7}$	-2.501	NA	NA	.88   .93
5:79336091	<i>THBS4</i>	MS:D;D;D;D;D	.0045	.0001	45 (41.9-48.1)	Weight*	$6.7 \times 10^{-7}$	-0.377	$3.2 \times 10^{-7}$	-0.252	74.6   80.5 kg
5:140181423	<i>PCDHA3</i>	FS:PTV	.0001	NA	NA	WHR adj. BMI	$2.7 \times 10^{-7}$	2.559	NA	NA	.14   .93
9:107548661	<i>ABCA1</i>	MS:D;D;D;D;D	.00023	0	$\infty$	Serum HDL Chol.	$4.8 \times 10^{-10}$	-2.046	NA	NA	.80   1.44 mmol/l
9:136501728	<i>DBH</i>	MS:D;D;P;N;N	.05	.0021	23.8 (22.4-25)	Diastolic BP*	$1.5 \times 10^{-6}$	-0.115	$2.8 \times 10^{-12}$	-0.11	83.1   84.1 mmHg
						Serum HDL Chol.	$1.4 \times 10^{-7}$	0.425	<b>6.7 <math>\times 10^{-7}</math></b>	0.435	1.59   1.44 mmol/l
11:47282929	<i>NRIH3</i>	MS:D;P;P;D;D	.0042	.00003	140 (132.8-147.2)	HDL2 Chol.* VLDL Chol.*	$3.2 \times 10^{-6}$ $4.0 \times 10^{-6}$	0.473 -0.469	$1.3 \times 10^{-8}$ $3.1 \times 10^{-7}$	0.458 -0.412	1.07   .92 mmol/l .75   .91 mmol/l
11:116692293	<i>APOA4</i>	MS:T;D;P;N;N	.0096	.012	0.8 (-0.4-2)	Serum HDL Chol.*	$2.2 \times 10^{-5}$	0.225	$1.5 \times 10^{-7}$	0.196	1.51   1.44 mmol/l
11:117352857	<i>DSCAML1</i>	MS:T;B;B;.;D	.016	.0002	80 (77.8-82.2)	VLDL Chol.	$4.1 \times 10^{-8}$	0.299	<b>2.0 <math>\times 10^{-3}</math></b>	0.162	1.01   .90 mmol/l
14:101198426	<i>DLK1</i>	MS:T;B;B;N;D	.023	.00013	177 (174.3-179.6)	Height*	$2.7 \times 10^{-5}$	-0.149	$1.2 \times 10^{-10}$	-0.163	170.7   172.0 cm
16:55862682	<i>CES1</i>	MS:D;D;D;D;D	.0018	.00003	60 (52.8-67.2)	Serum HDL Chol. Serum ApoA1*	$1.1 \times 10^{-10}$ $1.9 \times 10^{-6}$	0.771 0.668	<b>3.8 <math>\times 10^{-6}</math></b> $4.0 \times 10^{-9}$	0.793 0.718	1.77   1.44 mmol/l 1.65   1.47 g/l
16:56996009	<i>CETP</i>	SD:PTV	.0017	.00003	56.7 (49.4-63.9)	Serum ApoA1 Serum HDL Chol.	$2.6 \times 10^{-8}$ $1.1 \times 10^{-14}$	0.834 0.946	<b>1.8 <math>\times 10^{-4}</math></b> <b>8.8 <math>\times 10^{-21}</math></b>	1.034 1.217	1.70   1.47 g/l 1.84   1.44 mmol/l
16:68013570	<i>DPEP3</i>	MS:T;B;B;N;D	.0099	.00044	22.5 (20.8-24.2)	Serum HDL Chol. Serum ApoA1*	$1.6 \times 10^{-7}$ $5.2 \times 10^{-6}$	-0.295 -0.294	<b>7.2 <math>\times 10^{-15}</math></b> $4.0 \times 10^{-7}$	-0.373 -0.253	1.33   1.44 mmol/l 1.40   1.47 g/l
16:68732169	<i>CDH3</i>	MS:D;D;D;D;D	.0044	.00064	6.9 (5.2-8.5)	Pyruvate*	$3.7 \times 10^{-5}$	0.417	$6.6 \times 10^{-10}$	0.471	.08   .07 mmol/l
17:6599157	<i>SLC13A5</i>	MS:D;D;D;D;D	.00091	0	$\infty$	Citrate	$1.3 \times 10^{-9}$	1.294	<b>9.5 <math>\times 10^{-12}</math></b>	1.309	.14   .11 mmol/l
17:7129898	<i>DVL2</i>	MS:D;D;D;D;D	.02	.02	1 (-0.2-2.1)	Valine*	$4.2 \times 10^{-5}$	-0.239	$5.7 \times 10^{-9}$	-0.232	.210   .217 mmol/l
17:39135270	<i>KRT40</i>	MS:D;P;P;N;D	.00013	0	$\infty$	Serum HDL Chol.	$3.2 \times 10^{-8}$	2.416	NA	NA	2.51   1.44 mmol/l
						Total MUFA Glycerol* Plasma CRP* Triglycerides*	$4.4 \times 10^{-7}$ $5.8 \times 10^{-6}$ $1.6 \times 10^{-5}$ $1.0 \times 10^{-6}$	0.275 0.218 0.175 0.23	$3.5 \times 10^{-1}$ $4.1 \times 10^{-7}$ $4.0 \times 10^{-9}$ $1.3 \times 10^{-7}$	0.067 0.183 0.185 0.197	3.88   3.62 mmol/l .092   .088 mmol/l 2.47   2.17 mg/l 1.60   1.46 mmol/l
17:41062979	<i>G6PC</i>	MS:T;P;P;D;D	.025	0	$\infty$	Serum HDL Chol.	$4.8 \times 10^{-14}$	2.061	<b>4.9 <math>\times 10^{-2}</math></b>	0.801	2.39   1.44 mmol/l
17:41926216	<i>CD300LG</i>	MS:T;D;P;N;N	.00034	0	$\infty$	HDL2 Chol. Serum ApoA1	$1.3 \times 10^{-7}$ $8.1 \times 10^{-8}$	2.154 1.694	NA NA	NA NA	1.88   .92 mmol/l 2.04   1.47 g/l
18:47091686	<i>LIPG</i>	SA:PTV	.0025	0	$\infty$	HDL2 Chol.* Phosphocholines* Phosphoglycerides*	$1.2 \times 10^{-5}$ $3.1 \times 10^{-6}$ $9.0 \times 10^{-6}$	0.579 0.624 0.594	$5.6 \times 10^{-10}$ $1.1 \times 10^{-8}$ $1.1 \times 10^{-7}$	0.624 0.578 0.538	1.13   .92 mmol/l 2.44   2.20 mmol/l 2.50   2.25 mmol/l

19:10683762	<i>APIM2</i>	MS:D;D;D;D;D	.015	.00009	167 (162.7-170.7)	Serum ApoB	$5.8 \times 10^{-8}$	-0.282	<b>1.5x10<sup>-3</sup></b>	-0.199	.96   1.02 g/l
						Total Chol. in IDL*	$1.1 \times 10^{-6}$	-0.289	$6.9 \times 10^{-14}$	-0.319	.81   .87 mmol/l
						IDL Particle Conc.*	$2.1 \times 10^{-6}$	-0.281	$8.5 \times 10^{-14}$	-0.318	.125   .133 umol/l
						Remnant Chol.*	$8.0 \times 10^{-6}$	-0.268	$2.7 \times 10^{-12}$	-0.301	1.65   1.77 mmol/l
19:11350904	<i>ANGPTL8</i>	SG:PTV	.0025	0	$\infty$	HDL2 Chol.*	$3.4 \times 10^{-6}$	0.564	$1.1 \times 10^{-8}$	0.574	1.06   .92 mmol/l
19:49318380	<i>HSD17B14</i>	MS:D;D;D;D;D	.046	.05	0.9 (-0.2-2)	Valine*	$3.4 \times 10^{-5}$	-0.152	$2.1 \times 10^{-7}$	-0.144	.212   .217 mmol/l
20:24994201	<i>ACSSI</i>	MS:D;D;D;D;D	.0026	0	$\infty$	Acetate*	$1.3 \times 10^{-5}$	0.626	$2.1 \times 10^{-12}$	0.631	.046   .041 mmol/l

<sup>^</sup>Annotations are from VEP: FS=Frameshift; SG=Stop Gain; SD=Splice Donor; SA=Splice Acceptor; MS=Missense. All but MS are PTV. Predictions for missense variants are presented for five different prediction algorithms, each separated by semi-colon: SIFT (D=Damaging; T=Tolerated); PolyPhen2 - human diversity (D=Probably Damaging; P=Possibly Damaging; B=Benign); PolyPhen2 - hum variation (D=Probably Damaging; P=Possibly Damaging; B=Benign); Mutation Taster (A=Automatic Disease Causing; D=Disease Causing; N=Polymorphism; P=Automatic Polymorphism); and LRT (D=Deleterious; N=Neutral; U=Unknown).

# Non-Finnish European (NFE) MAF are taken from exomes of gnomAD v2.1 control individuals that were not from Estonia or Sweden. A variant with frequency 0 indicates that the variant was present in some subset of gnomAD, but was not found in NFE controls. NA indicates the variant was not present in gnomAD.

Minor Allele Frequency Ratio (MAF Ratio) is MAF in FinMetSeq/MAF in gnomAD NFE.

\*Indicates an association only reaching significance in meta-analysis combining FinMetSeq and the three replication cohorts. If unlabeled, the association was significant in FinMetSeq alone.

\*\* Replication P-values <0.05 are highlighted in bold.

## METHODS

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

762 **METSIM** is a single-site study comprised of 10,197 men randomly selected from the  
763 population register of Kuopio, Eastern Finland, aged 45 to 73 years at initial examination  
764 from 2005 to 2010<sup>17,62</sup>. The goal of METSIM is to investigate genetic and non-genetic  
765 factors associated with Type 2 Diabetes (T2D), cardiovascular disease (CVD), insulin  
766 resistance, and related traits. The METSIM study protocol includes collection of data on  
767 CVD history and risk factors, measurements of height, weight, waist, hip, blood pressure,  
768 and collection of a blood sample for laboratory measurements and DNA extraction.  
769 Diagnoses of myocardial infarction<sup>63</sup>, stroke<sup>64</sup>, and peripheral vascular disease were  
770 verified based on medical records at baseline. We attempted exome sequencing of all  
771 METSIM study participants.

772  
773 **FINRISK** is a series of health examination surveys carried out by the National Institute  
774 for Health and Welfare (formerly National Public Health Institute) of Finland every five  
775 years beginning in 1972<sup>65</sup>. The surveys are based on random population samples from  
776 five (six in 2002) geographical regions of Finland. Participants were selected by 10-year  
777 age group, sex, and study area. Survey sample sizes have varied from 7,000 to 13,000  
778 individuals and participation rates from 60% to 90%. The age-range was 25 to 64 years  
779 until 1992 and 25 to 74 years since 1997. The survey includes a self-administered  
780 questionnaire, a standardized clinical examination carried out by specifically trained  
781 study nurses, and collection of a blood sample for laboratory measurements and DNA  
782 extraction<sup>66</sup>. For exome sequencing, we chose 10,192 participants from the 1992, 1997,  
783 2002, and 2007 FINRISK surveys from northeastern Finland (former provinces of North

784 Karelia, Oulu, and Lapland). This selection was based on the hypothesis that the rapid  
785 growth in isolation of the populations of this region from severe bottlenecks in the 16<sup>th</sup>-  
786 17<sup>th</sup> centuries might have resulted in deleterious variants rising to a much higher  
787 frequency than in other populations.

788

789 METSIM participants fasted for more than 10 hours prior to blood draw. FINRISK  
790 participants were instructed to fast for four hours before the scheduled examination and  
791 to avoid heavy meals earlier in the day; duration of fasting was recorded. Laboratory  
792 measurements in METSIM included an oral glucose tolerance test with samples at 0, 30,  
793 and 120 minutes (only fasting measurements in known diabetics) for glucose, insulin,  
794 proinsulin, and free fatty acids, as well as fasting laboratory measurements including  
795 lipids, lipoproteins, apolipoproteins, adiponectin, and hs-CRP. Most of these  
796 measurements were carried out in FINRISK non-fasting samples; two-hour oral glucose  
797 tolerance tests with glucose and insulin measured at 0 and 120 minutes were carried out  
798 in subsets of FINRISK 1992, 2002 and 2007 participants. Both studies include proton  
799 NMR metabolomics measurements of lipoprotein subclasses, their lipid concentrations  
800 and composition, apolipoprotein A-I and B, multiple cholesterol and triglyceride  
801 measures, albumin, various fatty acids, and numerous low-molecular-weight metabolites,  
802 including amino acids, glycolysis related measures and ketone bodies<sup>67,68</sup>.

803

804 METSIM and FINRISK participants are followed up regularly via record linkage using  
805 the Finnish health registries, allowing for near complete follow-up of multiple disease  
806 diagnoses; participants may also be called back in person for additional studies.

807 Participants in both studies provided informed consent, and all study protocols were  
808 approved by the Ethics Committees at the participating institutions (FINRISK cohorts  
809 1992 & 1997: National Public Health Institute of Finland; FINRISK 2002, 2007, & 2012:  
810 Ethical Review Board of the Hospital District of Helsinki and Uusimaa; METSIM:  
811 Research Ethics Committee, Hospital District of Northern Savo IRB #1).

812

813 **Selection of traits, harmonization, exclusions, covariate adjustment, and**  
814 **transformation**

815 Of the 257 quantitative metabolic and cardiovascular traits measured in both METSIM  
816 and FINRISK, we selected 64 traits for association analysis in the current study based on  
817 clinical relevance for cardiovascular and metabolic health (**Supplementary Tables 3, 4**).

818

819 *Exclusions*

820 We excluded 126 individuals, 92 with type 1 diabetes and 34 women who were pregnant  
821 at the time of phenotyping, from all analyses, and 3,088 individuals with T2D from  
822 analyses of glycemic traits. For traits influenced by food consumption (amino acids, fatty  
823 acids, LDL cholesterol, total triglycerides, and glycemic traits), we excluded individuals  
824 not fasting for at least 8 hours after their last meal. A complete list of exclusions can be  
825 found in **Supplementary Table 4**.

826

827 *Trait adjustments*

828 For individuals on antihypertensive medication at the time of testing, we added 15 and 10  
829 mmHg to the measured values of systolic and diastolic blood pressures, respectively<sup>69,70</sup>.

830 For individuals on lipid regulating medications, we multiplied the measured total  
831 cholesterol by 1.25 <sup>71</sup>. For FINRISK participants, we calculated LDL cholesterol (LDL-  
832 C) levels using the Friedewald equation (LDL-C = adjusted total cholesterol – HDL-C –  
833 (triglycerides/2.2)) and excluded individuals with elevated triglycerides (>2.5mmol/l);  
834 LDL-C was measured directly in METSIM participants and for participants on lipid  
835 regulating medication, values were divided by 0.7 <sup>72</sup>. All trait adjustments are listed in  
836 **Supplementary Table 4**.

837

838 *Trait transformations and adjustment for covariates*

839 We prepared quantitative traits for association analysis separately for METSIM and  
840 FINRISK participants by linear regression on trait-specific covariates; skewed variables  
841 were log transformed prior to linear regression analysis. Both before and after log  
842 transformation, we examined all variables for normality and for outliers. Log  
843 transformation was adequate in all cases to approximate normality for initial covariate  
844 adjustment. Outliers, of which there were no more than 5 individuals with values >5SD  
845 for any trait prior to adjustment and inverse normalization, were not removed. Covariates  
846 for these regression analyses always included covariates age and age<sup>2</sup> for METSIM and  
847 sex, age, age<sup>2</sup>, and cohort year for FINRISK. Trait transformations and trait-specific  
848 covariates are listed in **Supplementary Table 4**. Several traits were adjusted for sex  
849 hormone treatment, which included women on contraceptives or hormone replacement  
850 therapy. We transformed residuals from these initial regression analyses to normality  
851 using inverse normal scores.

852

853 **Exome sequencing**

854 We carried out exome sequencing in two phases.

855

856 Phase 1 We quantified the 10,379 Phase 1 DNA samples with the Quant-iT PicoGreen  
857 dsDNA reagent on a Varioskan Microplate Reader (ThermoFisher Scientific) and  
858 randomly parsed samples with adequate DNA (>250ng) into cohort specific files. We  
859 then re-arrayed samples using the BioMicroLab XL20 (USA Scientific) to ensure equal  
860 numbers of METSIM and FINRISK samples on each 96-well plate, alternating samples  
861 between studies in consecutive positions within and across plates, to reduce opportunities  
862 for between-study batch effects.

863

864 We constructed dual indexed libraries using 100-250ng of genomic DNA and the KAPA  
865 HTP Library Kit (KAPA Biosystems) on the SciClone NGS instrument (Perkin Elmer).  
866 The target insert size was 250 bp. We then pooled twelve libraries prior to hybridization  
867 with the SeqCap EZ HGSC VCRome (Roche) reagent that targets 45.1 Mb (23,585 genes  
868 and 189,028 non-overlapping exons) of the human genome. Each library pool contained  
869 samples from both cohorts and was comprised of 300-400 ng of each individual library  
870 for a total library input of 3.6-4.8  $\mu$ g into the initial hybridization. We estimated the  
871 concentration of each captured library pool by qPCR (Kapa Biosystems) to produce  
872 appropriate cluster counts for the HiSeq2000 platform (Illumina). We then generated  
873 2x100bp paired-end sequence data yielding ~6 Gb per sample to achieve a coverage  
874 depth of  $\geq$ 20x for  $\geq$ 70% of targeted bases for every sample.

875

876 Phase 2 We quantified, prepared, pooled, and captured the 9,937 Phase 2 samples just as  
877 in Phase 1. Here we generated 2×125 bp paired-end sequencing reads on the HiSeq2500  
878 1T to again achieve a coverage depth of  $\geq 20x$  for  $\geq 70\%$  of targeted bases for every  
879 sample.

880

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

882 We aligned sequence reads to human genome reference build 37 using bwa-mem  
883 (v0.7.7), marked duplicates with picard MarkDuplicates (v1.113;  
884 <http://broadinstitute.github.io/picard>), and realigned indels with the GATK<sup>73</sup>  
885 IndelRealigner (v2.4). We used BamUtil clipOverlap (v1.0.11;  
886 [http://genome.sph.umich.edu/wiki/BamUtil:\\_clipOverlap](http://genome.sph.umich.edu/wiki/BamUtil:_clipOverlap)) to mark overlapping bases  
887 from paired reads resulting from short insert fragments.

888

889 For each sample, we required  $\geq 70\%$  of target bases sequenced at  $\geq 20x$  depth, and SNV  
890 genotype array concordance  $>90\%$  if SNV array data were available. We used  
891 verifyBamID<sup>74</sup> (v1.1.1) to detect and exclude samples with estimated contamination  
892  $>3\%$ . Where available, we also used existing genotype data with verifyBamID to detect  
893 and exclude sample swaps. Of 20,316 samples attempted, 13 failed sequencing, 35 were  
894 sample swaps, 760 either had low genotype concordance, sex mismatch, or estimated  
895 contamination  $>3\%$ , and four had discrepancies between reported and genotype-estimated  
896 relationships (**Supplementary Table 1**).

897

898 We called SNVs and short indels using the recommended best practices for cohort-level  
899 calling with GATK<sup>73</sup> (v3.3). For each individual, we called bases and identified variant  
900 sites for all targeted exome bases and 500 bp of sequence up and downstream of each  
901 target region using HaplotypeCaller, resulting in calling substantial numbers of non-  
902 exonic variants. We merged these calls in batches of 200 individuals using  
903 CombineGVCFs and recalled genotypes for all individuals at all variable sites with  
904 GenotypeGVCFs.

905

906 After merging genotypes for the 19,378 samples that passed preliminary QC checks, we  
907 filtered SNVs and indels separately using the recommended best practices for Variant  
908 Quality Score Recalibration (VQSR). We used the set of true positive variants provided  
909 in the GATK resource bundle (v2.5; build37) for training the VQSR model after  
910 restricting to sites in targeted exome regions. After assessment with VQSR, we retained  
911 variants for which we identified  $\geq 99\%$  of true positive sites used in the training model  
912 (i.e. truth sensitivity) for both SNVs and indels.

913

914 Following initial variant filtering, we decomposed multi-allelic variants into bi-allelic  
915 variants, left-aligned indels, and dropped redundant variants using vt<sup>75</sup> (version 0.5). We  
916 filtered variants with  $>2\%$  missing calls and/or Hardy-Weinberg p-value $<10^{-6}$ . We  
917 applied an additional filter removing variants with an overall allele balance (AB; alternate  
918 AC/sum of total AC)  $<30\%$  in genotyped samples. We then excluded 86 individuals with  
919  $>2\%$  missing variant calls yielding a final analysis set of 19,292 individuals.

920

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

922 METSIM participants were previously genotyped with the Illumina OmniExpress array;  
923 genotyping and quality control are described elsewhere<sup>76</sup>. FINRISK participants were  
924 previously genotyped in several batches on different arrays<sup>21</sup>. We lacked genotype array  
925 data for 1,488 participants (57 METSIM, 1,431 FINRISK). From the available genotype  
926 array data, we generated three datasets: 1) a merged array-based genotype call set of all  
927 variants present in ≥90% of array-genotyped individuals across both cohorts; 2) a merged  
928 Haplotype Reference Consortium (HRC) v1.1 imputed data set of the array-based  
929 genotypes; 3) an integrated data set containing genotyped and imputed array-based  
930 variants and exome sequence variants (HRC+exome). Where there was overlap between  
931 the sequence and imputed genotypes, we kept the sequence-based genotypes. We  
932 excluded the 1,488 individuals with no array data from the HRC+exome panel.

933

934 We prepared array genotypes for imputation using the Imputation Preparation and  
935 Checking tool ([http://www.well.ox.ac.uk/~wrayner/tools/HRC-1000G-check-  
936 bim.v4.2.5.zip](http://www.well.ox.ac.uk/~wrayner/tools/HRC-1000G-check-bim.v4.2.5.zip)) and used the Michigan Imputation Server<sup>77</sup>  
937 ([www.imputationserver.sph.umich.edu](http://www.imputationserver.sph.umich.edu)) to impute genotypes using the HRC (v1.1)  
938 reference panel<sup>78</sup>. METSIM samples were imputed in a single batch. FINRISK samples  
939 were imputed in batches based on the genotyping array and/or center where genotypes  
940 were generated.

941

942 **Annotation**

943 We annotated the final set of variants passing QC using Ensembl's variant effect  
944 predictor (VEP v76)<sup>79</sup> and Combined Annotation-Dependent Depletion<sup>80</sup> (CADD v1.2).  
945 We employed five *in silico* algorithms implemented in dbNSFP v2.4  
946 (<https://sites.google.com/site/jpopgen/dbNSFP>) to predict the functional impact of  
947 missense variants: PolyPhen2 HumDiv and HumVar<sup>81</sup>, LRT<sup>82</sup>, MutationTaster<sup>83</sup>, and  
948 SIFT<sup>84</sup>.

949

## 950 **Association testing**

### 951 *Single variants*

952 We carried out single-variant association tests for transformed trait residuals with  
953 genotype dosages for variants with  $MAC \geq 3$  assuming an additive genetic model. We  
954 used the EMMAX<sup>85</sup> linear mixed model approach, as implemented in EPACTS (v3.3.0;  
955 <http://genome.sph.umich.edu/wiki/EPACTS>), to account for relatedness between  
956 individuals. We used genotypes for sequenced variants with  $MAF \geq 1\%$  to construct the  
957 genetic relationship matrix (GRM).

958

### 959 *Conditioning on associated variants from prior GWAS*

960 To differentiate association signals identified in this study from known association  
961 signals, for each trait we performed exome-wide association analysis conditioning on  
962 variants previously associated ( $P < 10^{-7}$ ) with that trait. We compiled a list of known  
963 variants for each trait from the EBI GWAS catalog  
964 (<https://www.ebi.ac.uk/gwas/downloads>; December 4, 2016 version), from recent papers,  
965 and from manuscripts in preparation of which we were aware<sup>76,86-88</sup>. The keywords from

966 the GWAS catalog we used to assign known variants to each of our traits are listed in  
967 **Supplementary Table 19**. We also manually curated the associations from Inouye et  
968 al.<sup>89</sup> and Kettunen et al.<sup>86</sup> to attribute “blood metabolite” associations to the specific  
969 associated metabolites.

970

971 Using the combined HRC+exome panel (see above), we pruned each trait-specific list of  
972 associated variants (“GWAS variants”) based on linkage disequilibrium (LD) ( $r^2 > 0.95$ ).  
973 Twenty-three GWAS variants were not present in the HRC+exome panel. For 17 of these  
974 23 variants, we identified a proxy ( $r^2 > 0.80$ ) variant instead; we excluded the remaining  
975 six variants from conditional analysis. The list of variants included in conditional analysis  
976 for each trait is included in **Supplementary Table 20**. We extracted genotypes for  
977 variants used in conditional analysis from the integrated HRC+exome panel and  
978 converted dosages to alternate allele counts by rounding to the nearest integer (0, 1, or 2).  
979 We imputed missing genotypes for the 1,488 individuals without array data using the  
980 mean genotype dosage for purposes of conditional analysis.

981

982 For conditional analysis for each trait, we ran association analysis using the same linear  
983 mixed model approach as in unconditional analysis but including the complete set of  
984 pruned GWAS variants as covariates in the association test. Following conditional  
985 association, we further determined novelty based on absence of the variant from OMIM  
986 descriptions, ClinVar, and extensive literature searches.

987

988 *Defining loci*

989 The set of >1.4M variants tested for association includes variants in LD. To identify the  
990 number of distinct associations for each trait, we performed LD clumping using Swiss  
991 (<https://github.com/welchr/swiss>). We selected the subset of variants with (1)  
992 unconditional  $P < 5 \times 10^{-7}$  or (2) both unconditional and conditional  $P < 5 \times 10^{-5}$  for at least  
993 one trait. For each variant in this subset, we provided Swiss with the minimum  
994 unconditional p-value across all traits. The clumping procedure starts with the variant  
995 with the smallest p-value (index variant), and merges into one locus all variants within  $\pm 1$   
996 Mb that have  $r^2 > 0.5$  with the index variant. The procedure is repeated iteratively until no  
997 variants remain. Trait associations with variants in the same locus are considered to  
998 represent the same signal and trait associations with variants in different loci to be  
999 distinct signals.

1000

1001 *Calculating effects and variance explained of individual variants*

1002 For novel variants highlighted in **Table 2** we evaluated the effect of each variant on the  
1003 trait values. We did this by calculating the mean trait value in carriers and non-carriers,  
1004 assuming no homozygous carriers. Differences noted are the difference in the two means.

1005

1006 Given that the effect estimates from our association tests are standardized, we calculated  
1007 variance explained for a given variant with the equation  $2f(1-f)\hat{\beta}^2$ , where  $f$  is the minor  
1008 allele frequency and  $\hat{\beta}$  is the estimated effect size. The variance explained is included in  
1009 **Supplementary Table 8**.

1010

1011 *Gene-based testing*

1012 We carried out gene-based association tests using the mixed model implementation of  
1013 SKAT-O<sup>90</sup>, which tests for the optimal mixture of burden and dispersion-style multi-  
1014 marker tests while adjusting for relatedness between individuals using the same GRM  
1015 calculated for the single-variant tests. EMMA and the mixed model version of SKAT-O  
1016 (mmskat) are implemented in EPACTS.

1017

1018 We implemented gene-based tests using three different, but nested, sets of variants  
1019 (variant “masks”):

1020 (1) PTVs at any allele frequency with VEP annotations: frameshift\_variant,  
1021 initiator\_codon\_variant, splice\_acceptor\_variant, splice\_donor\_variant, stop\_lost,  
1022 stop\_gained;

1023 (2) PTVs included in (1) plus missense variants with MAF<0.1% scored as “damaging”  
1024 or “deleterious” by all five functional prediction algorithms;

1025 (3) PTVs included in (1) plus missense variants with MAF<0.5% scored as “damaging”  
1026 or “deleterious” by all five functional prediction algorithms.

1027

1028 For each trait and mask, we only tested genes with at least two qualifying variants. Each  
1029 mask contained a different number of genes with at least two qualifying variants: up to  
1030 7,996, 12,795, and 12,890 for the three masks, respectively. The exact number of genes  
1031 tested varied by trait due to sample size. We first used a Bonferroni-corrected exome-  
1032 wide threshold for 12,890 genes, which corresponds to a threshold of  $P<3.88\times10^{-6}$ .  
1033 Analogous to single-variant association, we passed genes meeting this association

1034 threshold forward for additional consideration with hierarchical FDR correction as  
1035 described below.

1036

### 1037 **Hierarchical FDR correction for testing multiple traits and variants**

1038 In controlling for multiple testing our goal was to make sure that, by looking across 64  
1039 traits, we did not increase the proportion of falsely discovered variants. To accomplish  
1040 this, we adopted a FDR controlling procedure described in Peterson et al.<sup>91</sup>, which uses a  
1041 hierarchical strategy to increase power while controlling type I error (**Supplementary**  
1042 **Methods**). This procedure has two stages. Stage 1 identifies the set of R variants that are  
1043 associated with at least one trait, controlling genome-wide FDR across all variants at  
1044 0.05. Stage 2 identifies all the traits that are associated with the discovered variants in a  
1045 manner that guarantees an average FDR<0.05.

1046

1047 In Stage 1 we restricted ourselves to the R=531 variants that have an unconditional  
1048 association  $P<5\times10^{-7}$  with at least one trait. For these, we calculated a p-value for the  
1049 hypothesis of no association between the variant and any of the 64 traits using Simes'  
1050 rule<sup>92</sup>, a combination rule that is robust to dependence between phenotypes. To account  
1051 for the fact that we did an initial selection of these R variants from the total number of  
1052 variants tested (T), we passed the Simes p-values to a Benjamini-Hochberg (BH)  
1053 procedure that controls FDR at target level  $0.05\times R/T$ , a modification<sup>93</sup> which guarantees  
1054 that the FDR in the set of S variants discovered to be associated with at least one trait is  
1055 less than 0.05.

1056

1057 In stage 2, to determine which traits are associated with the set of the S selected variants  
1058 we apply the Benjamini and Bogomolov<sup>94</sup> procedure. This procedure applies a  
1059 multiplicity correction variant by variant, passing the 64 trait association p-values from  
1060 each of the S selected variants and all 64 traits to a BH procedure that controls FDR at  
1061 target level  $0.05 \times S/T$ .

1062

1063 We applied this hierarchical correction to two different sets of results: the set of single-  
1064 variant unconditional results and the set of gene-based test results. The gene-based tests  
1065 used a threshold of  $P < 3.88 \times 10^{-6}$  to identify the R nominally significant genes in the first  
1066 stage of the hierarchical procedure.

1067

### 1068 **Genotype validation**

1069 We validated exome sequence-based genotype calls using Sanger sequencing for  
1070 METSIM carriers of 13 trait-associated very rare variants with  $MAF < 0.1\%$  in seven  
1071 genes. All but one of 108 (99.1%) non-reference genotypes validated were concordant.

1072

### 1073 **Association replication in additional Finnish cohorts**

1074 We performed replication analysis of significant single-variant associations ( $P < 5 \times 10^{-7}$ )  
1075 and follow-up analysis of suggestive single-variant associations ( $P < 5 \times 10^{-5}$ ) in up to  
1076 24,776 individuals from three GWAS cohort studies: Northern Finland Birth Cohort 1966  
1077 (NFBC1966), the Helsinki Birth Cohort Study (HBCS), and FINRISK study participants  
1078 not included in the exome sequencing portion of FinMetSeq.

1079

1080 A detailed description of the NFBC1966 study has been published previously and  
1081 additional information is available at: <http://www.oulu.fi/nfbc/node/18091><sup>22</sup>. The data  
1082 used here, including clinical measurements and blood samples for genetic and NMR  
1083 metabolite analyses, were collected at the 31-year follow-up in 1997. NFBC1966 samples  
1084 (n=5,139) were genotyped on the Illumina 370k array.

1085

1086 The HBCS includes participants born in Helsinki from 1934-1944 and has been described  
1087 elsewhere<sup>23</sup>; a basic description is available at <https://thl.fi/fi/web/thlfi-en/research-and-expertwork/projects-and-programmes/helsinki-birth-cohort-study-hbcs-idefix>. HBCS  
1088 samples (n=1,412) were genotyped on the Illumina 610k array.

1090

1091 The FINRISK cohort was described in detail above, and participants (replication  
1092 n=18,125) were genotyped in several batches on the Illumina 610k, CoreExome, or  
1093 OmniExpress arrays<sup>20,21</sup>.

1094

1095 For each replication cohort, prior to phasing we performed genotype quality control  
1096 batch-wise using standard quality thresholds for both sample-wise and variant-wise  
1097 filtering: call rate>95%, HWE>10<sup>-6</sup>, MAF>5%. We pre-phased array genotypes with  
1098 Eagle<sup>95</sup> (v2.3) and imputed genotypes genome-wide with IMPUTE<sup>96</sup> (v2.3.1) using the  
1099 SISu v2 reference panel consisting of 2,690 sequenced Finnish genomes and 5,092  
1100 sequenced Finnish exomes. Following imputation, we assessed imputation quality by  
1101 confirming sex, comparing sample allele frequencies with reference population estimates,

1102 and examining imputation quality (INFO score) distributions. We excluded any variant  
1103 with INFO<0.7 within a given batch from all replication/follow-up analyses.

1104

1105 For each of the three cohorts, we matched, harmonized, covariate adjusted, and  
1106 transformed available phenotypes as described above for FinMetSeq. We used the same  
1107 covariates as for FINRISK. For each cohort, we ran single-variant association using the  
1108 EMMAX linear mixed model implemented in EPACTS after generating kinship matrices  
1109 from LD-pruned (command: plink --indep-pairwise 50 5 0.2) directly genotyped variants  
1110 with MAF>5%.

1111

## 1112 **Association to disease endpoints in FinnGen**

1113 From a list of >1,100 disease endpoints available for analysis in the FinnGen project, we  
1114 selected 22 we considered most likely to be related to the quantitative traits analyzed in  
1115 FinMetSeq. As described in detail in Tabassum et al.<sup>45</sup>, variant associations with disease  
1116 endpoints in FinnGen biobank participants were tested using SPAtest R package and  
1117 adjusting for age, sex, and first 10 PCs in up to ~97,000 individuals.

1118

## 1119 **Association replication in UK Biobank**

1120 For the eight traits analyzed in FinMetSeq that were also available in the current UKBB  
1121 release (height, weight, BMI, hip circumference, waist circumference, fat percentage,  
1122 systolic blood pressure, and diastolic blood pressure), we extracted trait-variant  
1123 association statistics for variants reaching  $P<5\times10^{-7}$  in the FinMetSeq combined analysis  
1124 from the analysis of unrelated white British individuals generated by the Neale lab

1125 (<http://www.nealelab.is/uk-biobank>). Seven of the eight traits had at least one associated  
1126 variant and 23 of the total of 31 variants were available in UKBB. A comparison of  
1127 association results is in **Supplementary Table 13**.

1128

1129 **Population genetic analyses**

1130 *Identifying unrelated individuals*

1131 To identify a set of nearly independent common autosomal SNVs, we removed SNVs  
1132 with  $MAF < 5\%$  and pruned the remaining SNVs in windows of 50 SNVs, in steps of 5  
1133 SNVs, such that no pair of SNVs had  $r^2 > 0.2$ . We used the resulting 26,036 SNVs to  
1134 estimate pairwise relationships among the 19,292 exome-sequenced individuals using  
1135 KING<sup>97</sup>. We then removed one individual from each of the 4,418 pairs inferred by KING  
1136 to have a relationship of 3rd degree or closer, resulting in a set of 14,874 (nearly)  
1137 unrelated individuals for population genetic analyses.

1138

1139 *Identifying sub-population clusters in FinMetSeq*

1140 We first combined exome sequence variants and a genome-wide set of 220,798 SNVs  
1141 from GWAS arrays to provide a genome-wide backbone to aid in phasing and computing  
1142 haplotype sharing. After removing variants with  $MAC < 3$ , variants in known regions of  
1143 long range LD<sup>98</sup> and variants with  $HWE < 10^{-4}$ , we phased the remaining 764,696 variants  
1144 using SHAPEIT<sup>99</sup> (version 2, r837). To assess the substructure in our dataset while  
1145 minimizing the effect of mixing due to recent population mobility, we focused on the  
1146 2,644 unrelated individuals born by 1955 whose parents were both born in the same  
1147 municipality (irrespective of the birth location of the individual).

1148

1149 We identified sub-populations of the 2,644 individuals using ChromoPainter (version 2)  
1150 and fineSTRUCTURE<sup>100</sup> (version 2.0.8). We first used ChromoPainter to generate a  
1151 pairwise co-ancestry matrix, which represents each individual's DNA as a count of  
1152 haplotype blocks copied from every other individual in the dataset. Following previous  
1153 practices<sup>101</sup>, for computational efficiency, we estimated and fixed the switch and global  
1154 emission rates as input for ChromoPainter on a subset of the data; cluster inference is  
1155 known to be robust to up to 10-fold deviations of the estimated switch and emission  
1156 rates<sup>102</sup>. For further computational speedup, we generated an initial clustering by  
1157 applying a normal mixture model clustering<sup>103</sup> (mclust package in R, version 5.1) to the  
1158 top ten principal components of the coancestry matrix and used this initial cluster  
1159 solution as seed to the fineSTRUCTURE analysis. We conducted 1 million Markov chain  
1160 Monte Carlo (MCMC) iterations retaining one sample for every 1,000 iterations after  
1161 discarding 3 million iterations as burn-in. After MCMC, we used fineSTRUCTURE to  
1162 perform *post-hoc* refinement of cluster membership; we started with the MCMC sample  
1163 with the highest posterior probability and reassigned membership, taking into account the  
1164 cluster membership at each of the recorded MCMC samples<sup>102</sup>.

1165

1166 In total, we ran five MCMC chains using fineSTRUCTURE, retaining the configuration  
1167 with highest posterior probability for further analysis. We confirmed convergence of the  
1168 fineSTRUCTURE MCMC runs by calculating Geweke's convergence diagnostic using  
1169 the coda package (version 0.18) in R to compare the number of inferred clusters in the  
1170 first 10% and last 50% of the MCMC chain, and visual inspections of the general

1171 consistency of cluster memberships between independent MCMC chains. In total, we  
1172 inferred 245 sub-population clusters among the 2,644 individuals.

1173

1174 We inspected the initial clustering solution from fineSTRUCTURE by examining for  
1175 each individual the estimated proportion of their haplotype length derived from each of  
1176 the inferred clusters using non-negative least squares<sup>102,104</sup>. This approach showed many  
1177 individuals derived a substantial proportion of their haplotype length not from the cluster  
1178 initially assigned by fineSTRUCTURE, but instead from a different but related sub-  
1179 cluster on the fineSTRUCTURE hierarchical clustering tree, suggesting redundancy in  
1180 fineSTRUCTURE-inferred clusters. We therefore combined related clusters by  
1181 successively merging pairs of clusters that resulted in the smallest decrease in the  
1182 posterior probability of the fineSTRUCTURE hierarchical clustering tree. At each merge,  
1183 we reorganized individuals into merged cluster memberships and re-estimated the  
1184 haplotype-sharing profile for each individual. We iteratively merged the hierarchical tree  
1185 until  $\geq 90\%$  of individuals were assigned to the cluster where they also derive the highest  
1186 proportion of haplotype sharing, resulting in 16 clusters for the 2,644 reference  
1187 individuals, each named based on the most common parental birthplaces of its members  
1188 (**Supplementary Table 15**).

1189

1190 *Enrichment of predicted functionally deleterious alleles in Finland*

1191 We assessed enrichment of predicted functionally deleterious alleles in Finland by  
1192 comparing the 14,874 nearly unrelated (pairwise kinship coefficient  $<0.0448$ ) FinMetSeq  
1193 individuals to the 14,944 NFE control exomes in gnomAD, excluding from the NFE

1194 individuals from the neighboring countries of Estonia and Sweden in which substantial  
1195 numbers of Finns reside. We analyzed sites with base quality score >10, mapping quality  
1196 score >20, and coverage equal to or greater than that found in  $\geq 80\%$  of variable sites  
1197 (17.73X in FinMetSeq, 32.27X in gnomAD), resulting in  $\sim 38.6$  Mbp for comparisons.  
1198 We considered only the two most common alleles at each site. We contrasted the  
1199 proportional site frequency spectra for FinMetSeq and NFE for five functional variant  
1200 categories (PTVs, missense, synonymous, UTR, and intronic variants) after accounting  
1201 for sample size differences between datasets by down-sampling both datasets to 18,000  
1202 chromosomes.

1203

1204 We also assessed the enrichment of functional alleles within subpopulations of the  
1205 FinMetSeq dataset. Of the 16 sub-population clusters identified by fineSTRUCTURE, we  
1206 used as the reference population a cluster for which the highest proportion of the parents  
1207 of its members were from the southwestern, “early-settlement” part of Finland (NSv3,  
1208 **Supplementary Table 15**). Twelve of the remaining 15 clusters also have >100 members  
1209 and were used in subsequent analyses (**Supplementary Table 15**). We then compared  
1210 the ratio of the site frequency spectra to the reference for PTVs, missense, and  
1211 synonymous variants, again down-sampling both datasets to 200 haploid chromosomes to  
1212 account for sample size differences. For a given comparison, we computed statistical  
1213 evidence for enrichment or depletion at a given allele count bin by exact binomial test  
1214 against a null of equal number of variants found in both the test and reference cluster.

1215

1216 *Geographical clustering of predicted functionally deleterious alleles*

1217 We first generated a distance matrix tabulating the pairwise geographical distance in  
1218 kilometers between the birthplaces of all available parents of unrelated sequenced  
1219 individuals. For each variant of interest, we computed for the minor allele carriers in  
1220 FinMetSeq the mean distance among all parent pairs. For example, for a variant with  
1221 three carriers with information for five (of the possible six) parents, we computed the  
1222 mean for all (5-choose-2 = 10) distances. We evaluated statistical significance of  
1223 geographical clustering by comparing the mean distance to the means for up to  
1224 10,000,000 sets of randomly drawn non-carrier individuals matched by cohort status and  
1225 number of parents with birthplace information available.

1226

1227 To assess whether PTVs or missense variants may be more geographically clustered than  
1228 synonymous variants, we first identified a set of near-independent variants ( $r^2 > 0.02$ ) with  
1229  $MAC \geq 3$  and  $MAF \leq 5\%$  among the 14,874 unrelated individuals. This set included 4,312  
1230 PTVs, 91,851 missense variants, and 49,842 synonymous variants. For each variant, we  
1231 computed the mean pairwise geographical distance in kilometers between the birthplaces  
1232 across all pairs of the available parents of carriers of the minor allele and regressed this  
1233 mean distance on variant class (PTVs, missense, or synonymous) and MAC,  $MAC^2$ , and  
1234  $MAC^3$  (**Supplementary Table 14**).

1235

1236 We also assessed whether variants showing stronger enrichment (compared to NFE) are  
1237 more likely to be geographically clustered. Starting with the three functional classes of  
1238 variants identified above, we further restricted analysis to those variants found in  
1239 gnomAD so we could calculate the enrichment in frequency over gnomAD NFE. We

1240 included 1,540 PTVs, 46,953 missense, and 28,912 synonymous variants in this analysis  
1241 after pruning variants for LD with PLINK. As above, we computed the mean pairwise  
1242 distances among parents of carriers of the minor allele and regressed mean distance on  
1243 the logarithm of enrichment and MAC, MAC<sup>2</sup>, and MAC<sup>3</sup> (**Supplementary Table 17**). In  
1244 both analyses, we first assessed a model with the interaction terms but reported only the  
1245 model without interactions if the interactions were not significant.

1246

1247 *Heritability estimates and genetic correlations*

1248 We used genome-wide array genotype data on the 13,326 unrelated individuals for whom  
1249 both exome sequence and array data were available to estimate heritability and genetic  
1250 correlations for the 64 traits. We constructed a GRM with PLINK<sup>105</sup> (v.1.90b,  
1251 <https://www.cog-genomics.org/plink2>) by applying additional filters for MAF>1% and  
1252 genotype missingness rate <2% to the set of previously-used genotyped SNVs, leaving  
1253 205,149 SNVs for GRM calculation. We used the exact mixed model approach of  
1254 biMM<sup>106</sup> (v.1.0.0, <http://www.helsinki.fi/~mjxpirin/download.html>) to estimate the  
1255 heritability of our 64 traits and the genetic correlation of the 2,016 trait pairs.

1256 **Methods References**

1257 62 Stancáková, A. *et al.* Changes in insulin sensitivity and insulin release in relation  
1258 to glycemia and glucose tolerance in 6,414 Finnish men. *Diabetes* **58**, 1212-1221,  
1259 doi:10.2337/db08-1607 (2009).

1260 63 Thygesen, K. *et al.* Third universal definition of myocardial infarction. *J. Am.*  
1261 *Coll. Cardiol.* **60**, 1581-1598, doi:10.1016/j.jacc.2012.08.001 (2012).

1262 64 European Stroke Initiative Executive, C. *et al.* European Stroke Initiative  
1263 Recommendations for Stroke Management-update 2003. *Cerebrovasc. Dis.* **16**,  
1264 311-337 (2003).

1265 65 Borodulin, K. *et al.* Cohort Profile: The National FINRISK Study. *Int J*  
1266 *Epidemiol.* doi:10.1093/ije/dyx239 (2017).

1267 66 Borodulin, K. *et al.* Forty-year trends in cardiovascular risk factors in Finland.  
1268 *Eur. J. Public Health* **25**, 539-546, doi:10.1093/eurpub/cku174 (2015).

1269 67 Soininen, P., Kangas, A. J., Würtz, P., Suna, T. & Ala-Korpela, M. Quantitative  
1270 serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology  
1271 and genetics. *Circ. Cardiovasc. Genet.* **8**, 192-206,  
1272 doi:10.1161/CIRCGENETICS.114.000216 (2015).

1273 68 Wurtz, P. *et al.* Quantitative Serum Nuclear Magnetic Resonance Metabolomics  
1274 in Large-Scale Epidemiology: A Primer on -Omic Technologies. *American*  
1275 *journal of epidemiology* **186**, 1084-1096, doi:10.1093/aje/kwx016 (2017).

1276 69 Wu, J. *et al.* A summary of the effects of antihypertensive medications on  
1277 measured blood pressure. *Am J Hypertens* **18**, 935-942,  
1278 doi:10.1016/j.amjhyper.2005.01.011 (2005).

1279 70 Tobin, M. D., Sheehan, N. A., Scurrah, K. J. & Burton, P. R. Adjusting for  
1280 treatment effects in studies of quantitative traits: antihypertensive therapy and  
1281 systolic blood pressure. *Statistics in medicine* **24**, 2911-2935,  
1282 doi:10.1002/sim.2165 (2005).

1283 71 Liu, D. J. *et al.* Exome-wide association study of plasma lipids in >300,000  
1284 individuals. *Nature genetics* **49**, 1758-1766, doi:10.1038/ng.3977 (2017).

1285 72 Friedewald, W. T., Levy, R. I. & Fredrickson, D. S. Estimation of the  
1286 concentration of low-density lipoprotein cholesterol in plasma, without use of the  
1287 preparative ultracentrifuge. *Clin Chem* **18**, 499-502 (1972).

1288 73 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using  
1289 next-generation DNA sequencing data. *Nat. Genet.* **43**, 491-498,  
1290 doi:10.1038/ng.806 (2011).

1291 74 Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in  
1292 sequencing and array-based genotype data. *Am. J. Hum. Genet.* **91**, 839-848,  
1293 doi:10.1016/j.ajhg.2012.09.004 (2012).

1294 75 Tan, A., Abecasis, G. R. & Kang, H. M. Unified representation of genetic  
1295 variants. *Bioinformatics* **31**, 2202-2204, doi:10.1093/bioinformatics/btv112  
1296 (2015).

1297 76 Davis, J. P. *et al.* Common, low-frequency, and rare genetic variants associated  
1298 with lipoprotein subclasses and triglyceride measures in Finnish men from the  
1299 METSIM study. *PLoS genetics* **13**, e1007079, doi:10.1371/journal.pgen.1007079  
1300 (2017).

1301 77 Das, S. *et al.* Next-generation genotype imputation service and methods. *Nature*  
1302 *genetics* **48**, 1284-1287, doi:10.1038/ng.3656 (2016).

1303 78 McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype  
1304 imputation. *Nature genetics* **48**, 1279-1283, doi:10.1038/ng.3643 (2016).

1305 79 McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122,  
1306 doi:10.1186/s13059-016-0974-4 (2016).

1307 80 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity  
1308 of human genetic variants. *Nat. Genet.* **46**, 310-315, doi:10.1038/ng.2892 (2014).

1309 81 Adzhubei, I. A. *et al.* A method and server for predicting damaging missense  
1310 mutations. *Nature methods* **7**, 248-249, doi:10.1038/nmeth0410-248 (2010).

1311 82 Chun, S. & Fay, J. C. Identification of deleterious mutations within three human  
1312 genomes. *Genome research* **19**, 1553-1561, doi:10.1101/gr.092619.109 (2009).

1313 83 Schwarz, J. M., Cooper, D. N., Schuelke, M. & Seelow, D. MutationTaster2:  
1314 mutation prediction for the deep-sequencing age. *Nature methods* **11**, 361-362,  
1315 doi:10.1038/nmeth.2890 (2014).

1316 84 Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-  
1317 synonymous variants on protein function using the SIFT algorithm. *Nature  
1318 protocols* **4**, 1073-1081, doi:10.1038/nprot.2009.86 (2009).

1319 85 Kang, H. M. *et al.* Variance component model to account for sample structure in  
1320 genome-wide association studies. *Nature genetics* **42**, 348-354,  
1321 doi:10.1038/ng.548 (2010).

1322 86 Kettunen, J. *et al.* Genome-wide study for circulating metabolites identifies 62  
1323 loci and reveals novel systemic effects of LPA. *Nature communications* **7**, 11122,  
1324 doi:10.1038/ncomms11122 (2016).

1325 87 Kettunen, J. *et al.* Genome-wide association study identifies multiple loci  
1326 influencing human serum metabolite levels. *Nature genetics* **44**, 269-276,  
1327 doi:10.1038/ng.1073 (2012).

1328 88 Teslovich, T. M. *et al.* Identification of seven novel loci associated with amino  
1329 acid levels using single-variant and gene-based tests in 8545 Finnish men from  
1330 the METSIM study. *Hum Mol Genet* **27**, 1664-1674, doi:10.1093/hmg/ddy067  
1331 (2018).

1332 89 Inouye, M. *et al.* Novel Loci for metabolic networks and multi-tissue expression  
1333 studies reveal genes for atherosclerosis. *PLoS Genet.* **8**, e1002907,  
1334 doi:10.1371/journal.pgen.1002907 (2012).

1335 90 Lee, S. *et al.* Optimal unified approach for rare-variant association testing with  
1336 application to small-sample case-control whole-exome sequencing studies.  
1337 *American journal of human genetics* **91**, 224-237, doi:10.1016/j.ajhg.2012.06.007  
1338 (2012).

1339 91 Peterson, C. B., Bogomolov, M., Benjamini, Y. & Sabatti, C. Many Phenotypes  
1340 Without Many False Discoveries: Error Controlling Strategies for Multitrait  
1341 Association Studies. *Genet. Epidemiol.* **40**, 45-56, doi:10.1002/gepi.21942 (2016).

1342 92 Simes, R. J. An Improved Bonferroni Procedure for Multiple Tests of  
1343 Significance. *Biometrika* **73**, 751-754, doi:Doi 10.2307/2336545 (1986).

1344 93 Brzyski, D. *et al.* Controlling the Rate of GWAS False Discoveries. *Genetics* **205**,  
1345 61-75, doi:10.1534/genetics.116.193987 (2017).

1346 94 Benjamini, Y. & Bogomolov, M. Selective inference on multiple families of  
1347 hypotheses. *J. R. Stat. Soc. Series B Stat. Methodol.* **76**, 297-318,  
1348 doi:10.1111/rssb.12028 (2014).

1349 95 Loh, P. R. *et al.* Reference-based phasing using the Haplotype Reference  
1350 Consortium panel. *Nature genetics* **48**, 1443-1448, doi:10.1038/ng.3679 (2016).

1351 96 Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype  
1352 imputation method for the next generation of genome-wide association studies.  
1353 *PLoS genetics* **5**, e1000529, doi:10.1371/journal.pgen.1000529 (2009).

1354 97 Manichaikul, A. *et al.* Robust relationship inference in genome-wide association  
1355 studies. *Bioinformatics* **26**, 2867-2873, doi:10.1093/bioinformatics/btq559 (2010).

1356 98 Price, A. L. *et al.* Long-range LD can confound genome scans in admixed  
1357 populations. *American journal of human genetics* **83**, 132-135; author reply 135-  
1358 139, doi:10.1016/j.ajhg.2008.06.005 (2008).

1359 99 Delaneau, O., Zagury, J. F. & Marchini, J. Improved whole-chromosome phasing  
1360 for disease and population genetic studies. *Nat Methods* **10**, 5-6,  
1361 doi:10.1038/nmeth.2307 (2013).

1362 100 Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population  
1363 structure using dense haplotype data. *PLoS genetics* **8**, e1002453,  
1364 doi:10.1371/journal.pgen.1002453 (2012).

1365 101 Kerminen, S. *et al.* Fine-Scale Genetic Structure in Finland. *G3* **7**, 3459-3468,  
1366 doi:10.1534/g3.117.300217 (2017).

1367 102 Leslie, S. *et al.* The fine-scale genetic structure of the British population. *Nature*  
1368 **519**, 309-314, doi:10.1038/nature14230 (2015).

1369 103 Fraley, C. & Raftery, A. E. Model-based clustering, discriminant analysis, and  
1370 density estimation. *J Am Stat Assoc* **97**, 611-631, doi:Doi  
1371 10.1198/016214502760047131 (2002).

1372 104 Busby, G. B. *et al.* Admixture into and within sub-Saharan Africa. *eLife* **5**,  
1373 doi:10.7554/eLife.15266 (2016).

1374 105 Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger  
1375 and richer datasets. *Gigascience* **4**, 7, doi:10.1186/s13742-015-0047-8 (2015).

1376 106 Pirinen, M. *et al.* biMM: efficient estimation of genetic variances and covariances  
1377 for cohorts with high-dimensional phenotype measurements. *Bioinformatics* **33**,  
1378 2405-2407, doi:10.1093/bioinformatics/btx166 (2017).

1379

## Figure Legends

**Figure 1. Characterization of traits by heritability, Pearson correlation, and genetic correlation.** Traits in both figures are in the same order, clockwise in A, and left to right and top to bottom in B, and following the trait group color key.

A) Estimated heritability ( $h_x^2$ ) for each of the 64 traits included in association analysis. Heritability is based on ~205,000 common variants from GWAS arrays available in 13,342 unrelated individuals. Height has the highest heritability estimate at 52.5%. Estimates of trait heritability for metabolic measures are somewhat lower than previous reports (Kettunen, 2012) because estimates are from population-level data as opposed to twin studies and heritability was estimated from covariate adjusted and inverse normal transformed residuals, rather than raw trait values. Trait abbreviations are listed in Supplementary Table 3. All traits are significantly heritable except for 2hr-FFA (Fatty Acid) and His (Amino Acid), see Supplementary Table 5 for estimates, SEs, and P-values.

B) Heatmap of: 1) absolute Pearson correlations of standardized trait values in upper triangle, and 2) absolute values of the genetic correlation,  $\rho_G(x,y)$ , in lower triangle, where  $\rho_G(x,y)$  is the estimated genetic correlation of traits  $x$  and  $y$ . Values below the diagonal in gray had non-estimable genetic correlations.

## Figure 2. Characterization of discovered associations.

A) Number of genomic loci associated with each trait. Each bar is subdivided into common (MAF>1%, dark blue) and rare (MAF<1%, light blue). Traits are sorted by group as in Figure 1.

B) Relationship between estimated heritability and number of genomic loci detected for each trait. Each trait is colored by trait group following the trait group color key. Vertical bars indicated  $\pm 2$  standard errors of the heritability estimate. The gray line shows the linear regression fit, shown to indicate the general trend.

C) Heatmap of shared genomic associations by pairs of traits. For traits  $x$  and  $y$ , the 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 2A, and the side and top color bars reflect the trait groups.

D) Relationship between estimated genetic correlation and extent of sharing of genetic associations. For each pair of traits, 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. The bar within each box is the median, the box represents

the inter-quartile range, whiskers extend up to 1.5x the interquartile range, and outliers are presented as individual points. Analysis using the absolute value of the Pearson correlation of the residual series results in a very similar pattern.

**Figure 3. 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 exomes alone. Each variant reaching significance in FinMetSeq is plotted. Those associations highlighted in Table 2 are represented with a dark blue point (FinMetSeq MAF) and a corresponding brown point reflecting the NFE MAF (gnomAD). The purple lines indicate the 80% power curves for significance at  $5 \times 10^{-7}$  for sample sizes of 10,000 and 20,000. The right end of the power curve for N=20,000 terminates at MAF = 0.007. Plots show the dramatic increase in power due to higher relative frequency in Finland.

B) Relationship between MAF and estimated effect size for associations discovered in the combined analysis. Same plot as in A, highlighting the variants in Table 2 only reaching significance in the combined analysis.

**Figure 4. Regional variation in allele frequencies by functional annotation.** Enrichment of functional allelic class in sub-populations (regions) of Northern and Eastern Finland. For each minor allele count bin, we computed the ratio of number of variants found in each subpopulation to an internal 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  $<1$ ). The 12 sub-populations with sample size  $>100$  are shown. The results are consistent with multiple independent bottlenecks followed by subsequent drift in Northern and Eastern Finland, particularly for populations in Lapland and Northern Ostrobothnia. Abbreviations for regions: Kainuu (Kai), Lapland (Lap1, Lap2), Northern Karelia (NKa1, NKa2, NKa3, NKa4), Northern Ostrobothnia (NOs1, NOs2, NOs3, NOs4), Northern Savonia (NSv1, NSv2, NSv3), Southern Ostrobothnia (SOs), and Surrendered Karelia (SuK). For more detailed information on region definitions see Supplementary Table 15. Confidence intervals on the enrichment ratios, and their P-values, are presented in Supplementary Table 16.

**Figure 5. Geographical clustering of associated variants.**

A) Geographical clustering of PTVs as a function of MAC and frequency enrichment over NFE from gnomAD. For each PTV ( $r^2 \leq 0.02$ ,  $MAC \geq 3$ ,  $MAF \leq 0.05$ ) we computed the mean distance between birth places of available

parents of all carriers of the minor allele. We compared the frequency of the minor allele in FinMetSeq to gnomAD NFE. Blue and pink colors denote the frequency is lower or higher in FinMetSeq than in gnomAD NFE, respectively. The size of the point is proportional to the logarithm of the frequency ratio difference. In general, we observe that variants enriched in FinMetSeq are more geographically clustered.

B) Example of geographical clustering for a trait associated variant. The birth locations of all parents of carriers (orange) and a matching number of parents of non-carriers (blue) of the minor allele for variant chr3:125831672 (rs780671030, p.Arg722X) in *ALDH1L1* are displayed on a map of Finland. This variant is associated with serum glycine levels in FinMetSeq and has a frequency of 0 in NFE samples from gnomAD. The parents of carriers are born on average 135 km apart, the parents of non-carriers on average 250 km apart ( $P < 10^{-7}$  by permutation).

C) Comparison of geographical clustering between Finnish Disease Heritage (FDH) mutations and trait-associated variants that are  $>10x$  more frequent in FinMetSeq than in NFE. The degree of geographical clustering (based on parental birthplace) is comparable between carriers of those variants that showed significant associations in FinMetSeq alone (FMS) and carriers of FDH mutations, and greater than that seen in carriers of variants that showed significant association only in the combined analysis (FMS+Replication). For all variants, carriers of the minor allele displayed greater clustering than non-carriers. The bar within each box is the median, the box represents the inter-quartile range, whiskers extend up to 1.5x the interquartile range, and outliers are presented as individual points.

## Figure Legends (Extended Data Figures)

**Extended Data Fig. 1. Comparison of allele frequencies of variants in FinMetSeq and NFE from gnomAD.** 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.

**Extended Data Fig. 2. 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.

**Extended Data Fig. 3. 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 significantly associated with at least one trait are colored and scaled proportionately to the association p-value, with more significant associations having a larger symbol. 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. Variants unique to Finns and absent in gnomAD are not plotted.

**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.

**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.

**Extended Data Fig. 6. Gene-based association of extremely rare variants in *ABCA1* with serum HDL**

**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.

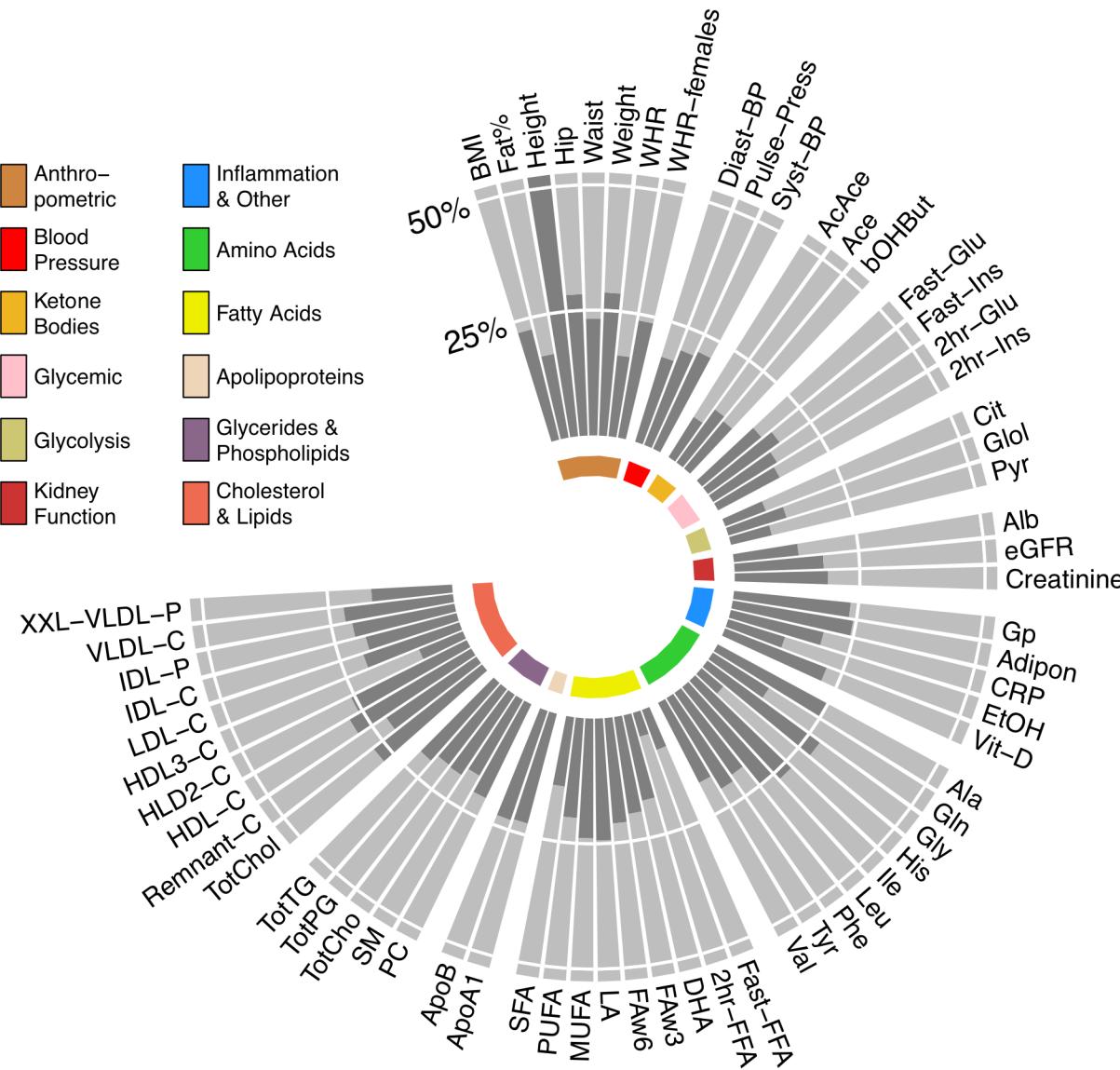
**Extended Data Fig. 7. Gene-based association of extremely rare variants in *ALDH1L1* with glycine levels.** The upper panel shows the distribution of the covariate adjusted and inverse-normal transformed phenotype. The lower panel displays the association statistics for each variant included in the 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.

**Extended Data Fig. 8. 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 whose parental birthplaces were known. 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. A map of Finland with regions labeled is supplied for reference.

**Extended Data Fig. 9. 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 15** for more detailed demographic descriptions of each subpopulation.

**Extended Data Fig. 10. Geographical clustering of missense and synonymous variants as a function of minor allele count and frequency enrichment over gnomAD NFE.** This represents the same analysis as Figure 5A, but for missense and synonymous variants rather than PTVs. Similar to PTVs, missense and synonymous variants that show greater enrichment in FinMetSeq are more likely to be geographically clustered. Blue and pink colors denote the frequency is lower or higher in FinMetSeq than in gnomAD NFE, respectively. The size of the point is proportional to the logarithm of the frequency ratio difference.

A.



B.

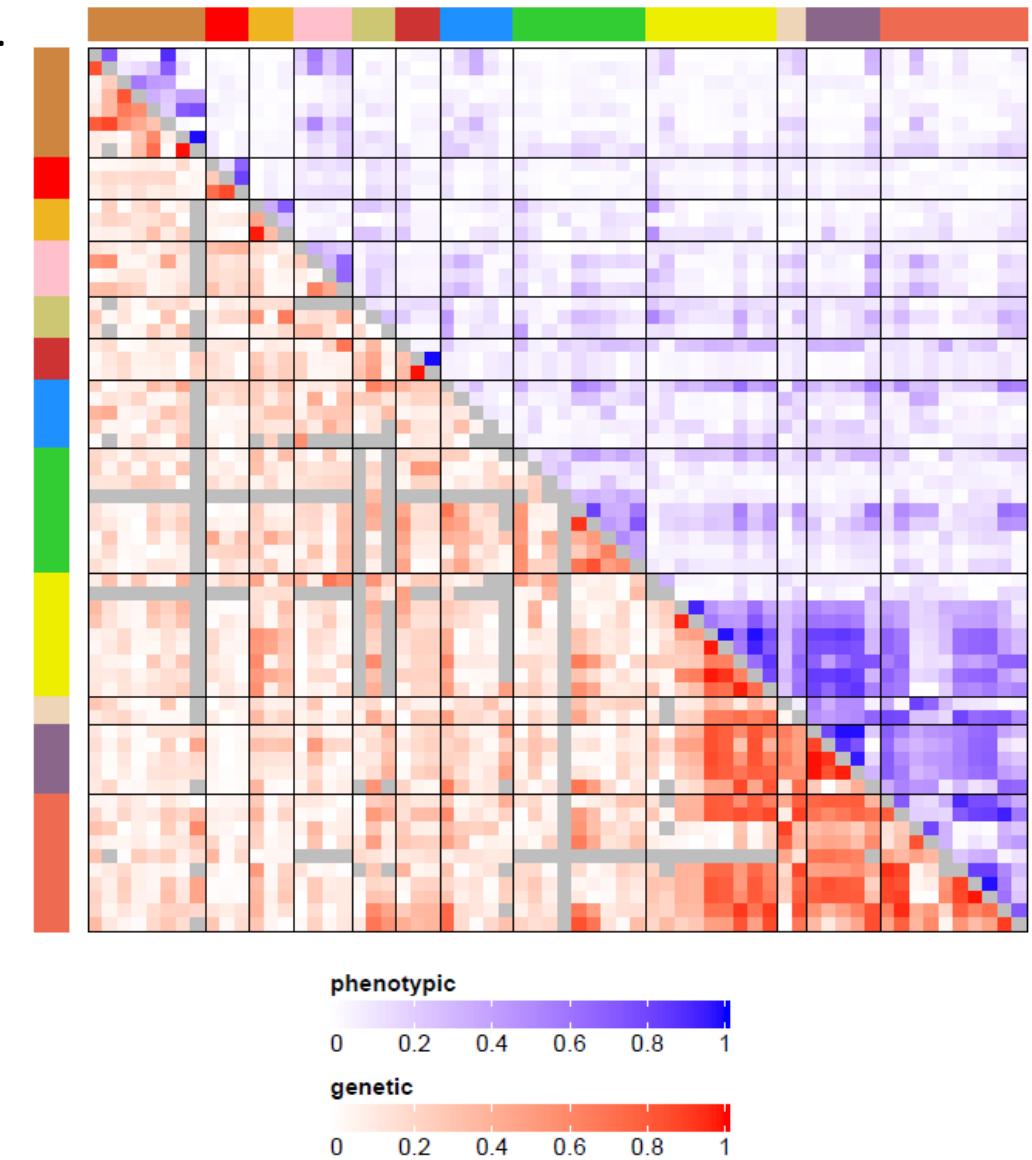
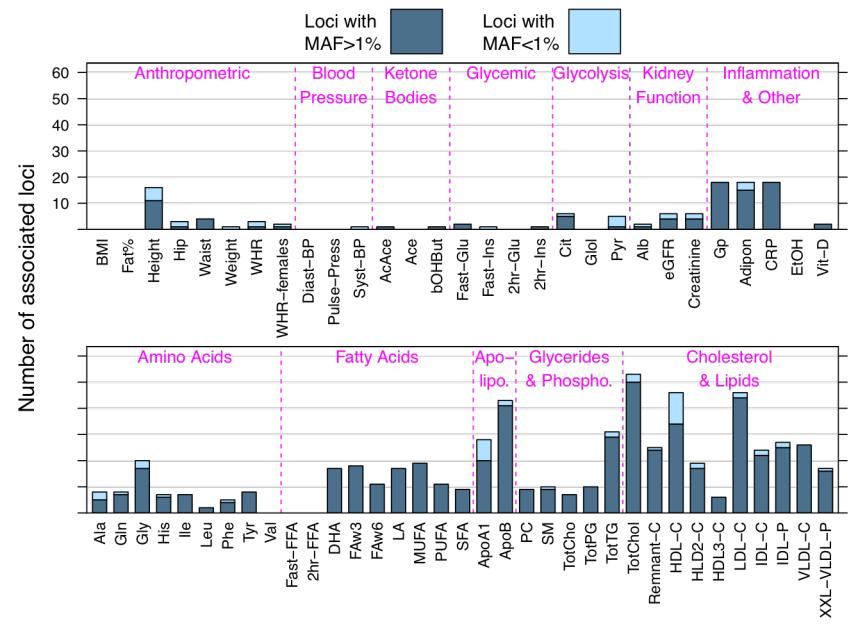


Figure 1. Characterization of traits by heritability, Pearson correlation, and genetic correlation. Traits in both figures are in the same order, clockwise in A, and left to right and top to bottom in B, and following the trait group color key.

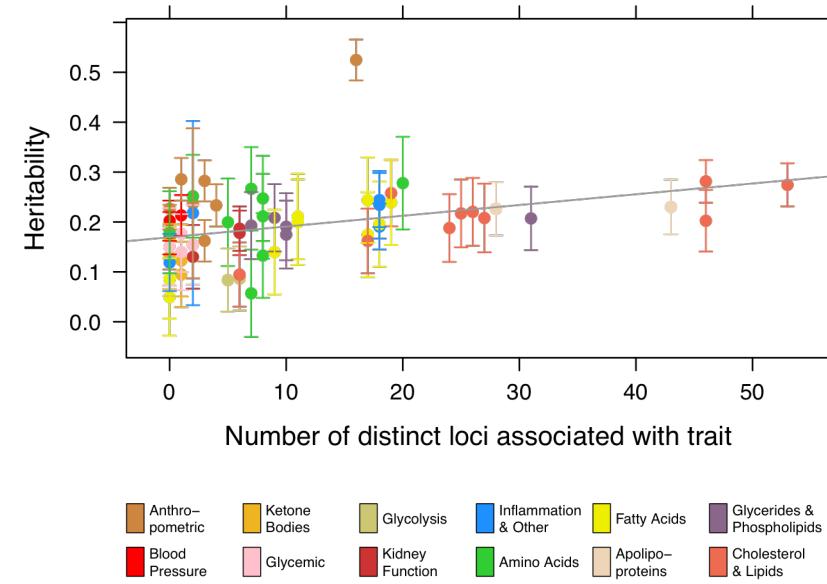
A) Estimated heritability ( $h_x^2$ ) for each of the 64 traits included in association analysis. Heritability is based on ~205,000 common variants from GWAS arrays available in 13,342 unrelated individuals. Height has the highest heritability estimate at 52.5%. Estimates of trait heritability for metabolic measures are somewhat lower than previous reports (Kettunen, 2012) because estimates are from population-level data as opposed to twin studies and heritability was estimated from covariate adjusted and inverse normal transformed residuals, rather than raw trait values. Trait abbreviations are listed in Supplementary Table 3. All traits are significantly heritable except for 2hr-FFA (Fatty Acid) and His (Amino Acid), see Supplementary Table 5 for estimates, SEs, and P-values.

B) Heatmap of: 1) absolute Pearson correlations of standardized trait values in upper triangle, and 2) absolute values of the genetic correlation,  $pg(x,y)$ , in lower triangle, where  $pg(x,y)$  is the estimated genetic correlation of traits x and y. Values below the diagonal in gray had non-estimable genetic correlations.

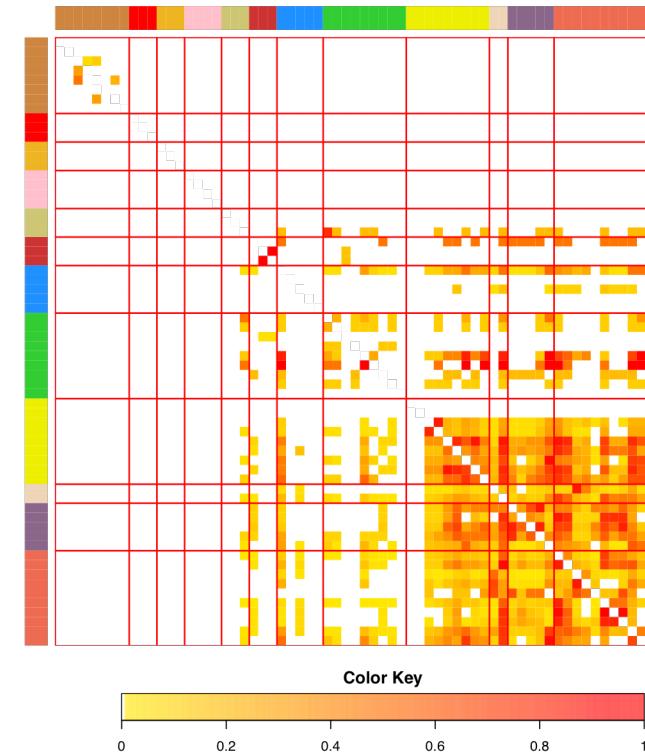
A.



B.



C.



D.

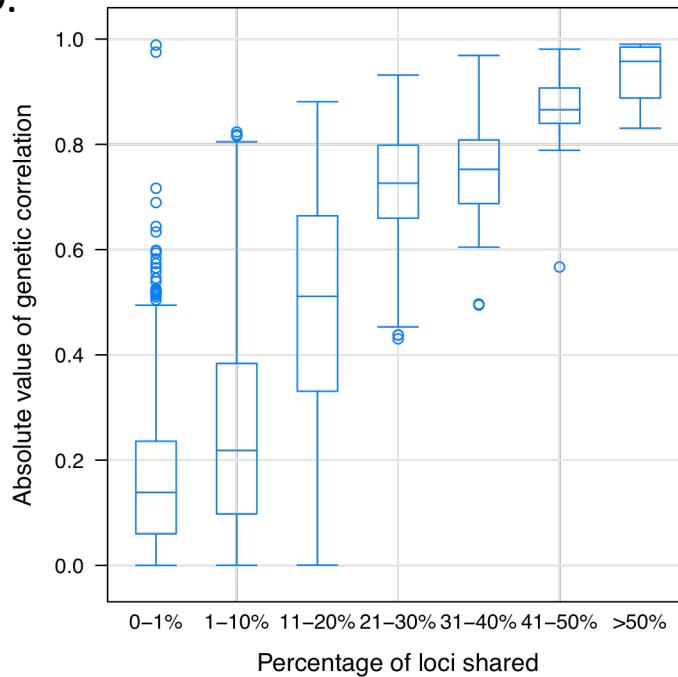


Figure 2. Characterization of discovered associations.

A) Number of genomic loci associated with each trait. Each bar is subdivided into common (MAF>1%, dark blue) and rare (MAF<1%, light blue). Traits are sorted by group as in Figure 1.

B) Relationship between estimated heritability and number of genomic loci detected for each trait. Each trait is colored by trait group following the trait group color key. Vertical bars indicated  $\pm 2$  standard errors of the heritability estimate. The gray line shows the linear regression fit, shown to indicate the general trend.

C) Heatmap of shared genomic associations by pairs of traits. For traits x and y, the 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 2A, and the side and top color bars reflect the trait groups.

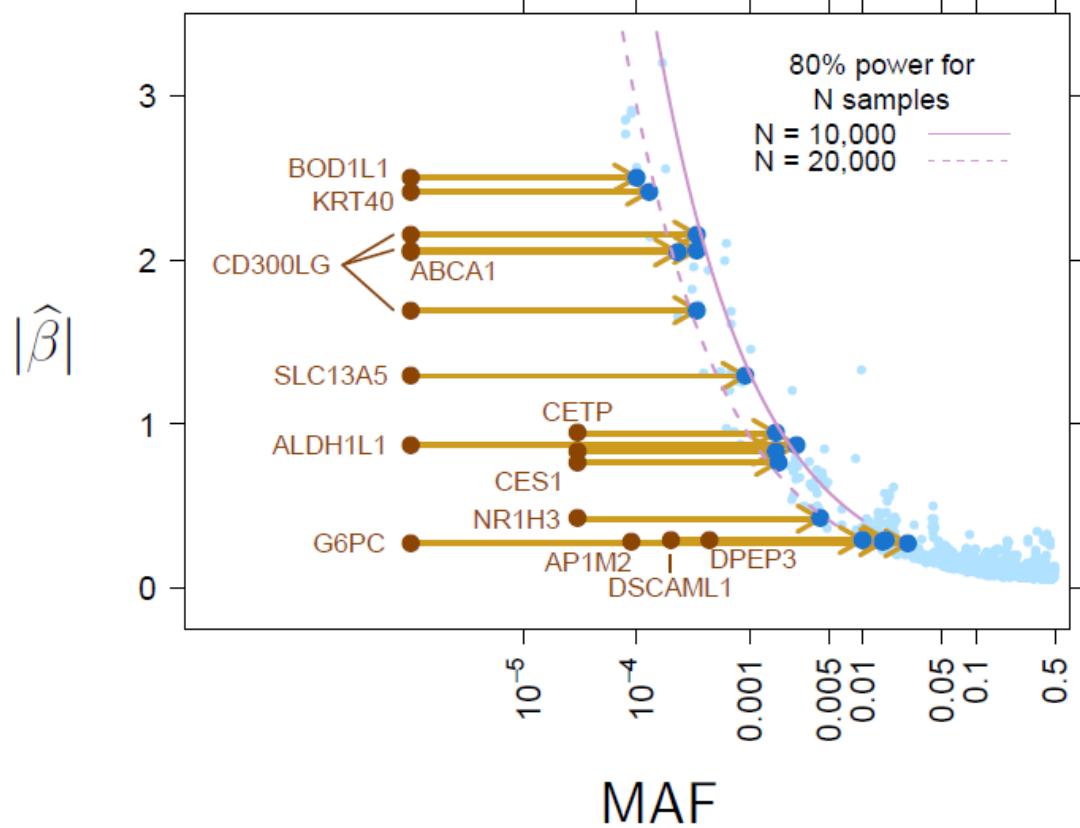
D) Relationship between estimated genetic correlation and extent of sharing of genetic associations. For each pair of traits, the extent of locus sharing is defined as the number of loci associated with both traits divided by the total number of loci associated with either trait. The bar within each box is the median, the box represents the inter-quartile range, whiskers extend up to 1.5x the interquartile range, and outliers are presented as individual points. Analysis using the absolute value of the Pearson correlation of the residual series results in a very similar pattern.

Figure 3. Allelic enrichment in the Finnish population and its effect on genetic discovery.

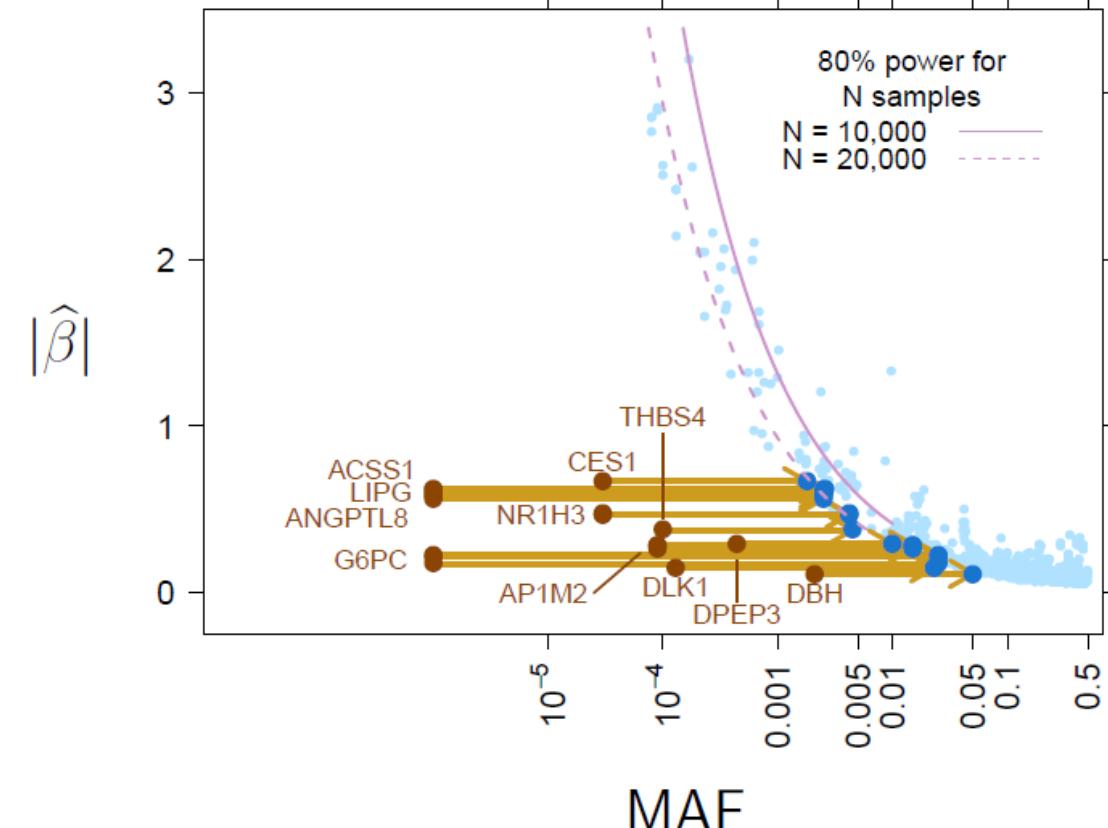
A) Relationship between MAF and estimated effect size for associations discovered in FinMetSeq exomes alone. Each variant reaching significance in FinMetSeq is plotted. Those associations highlighted in Table 2 are represented with a dark blue point (FinMetSeq MAF) and a corresponding brown point reflecting the NFE MAF (gnomAD). The purple lines indicate the 80% power curves for significance at  $5 \times 10^{-7}$  for sample sizes of 10,000 and 20,000. The right end of the power curve for N=20,000 terminates at MAF = 0.007. Plots show the dramatic increase in power due to higher relative frequency in Finland.

B) Relationship between MAF and estimated effect size for associations discovered in the combined analysis. Same plot as in A, highlighting the variants in Table 2 only reaching significance in the combined analysis.

A



B



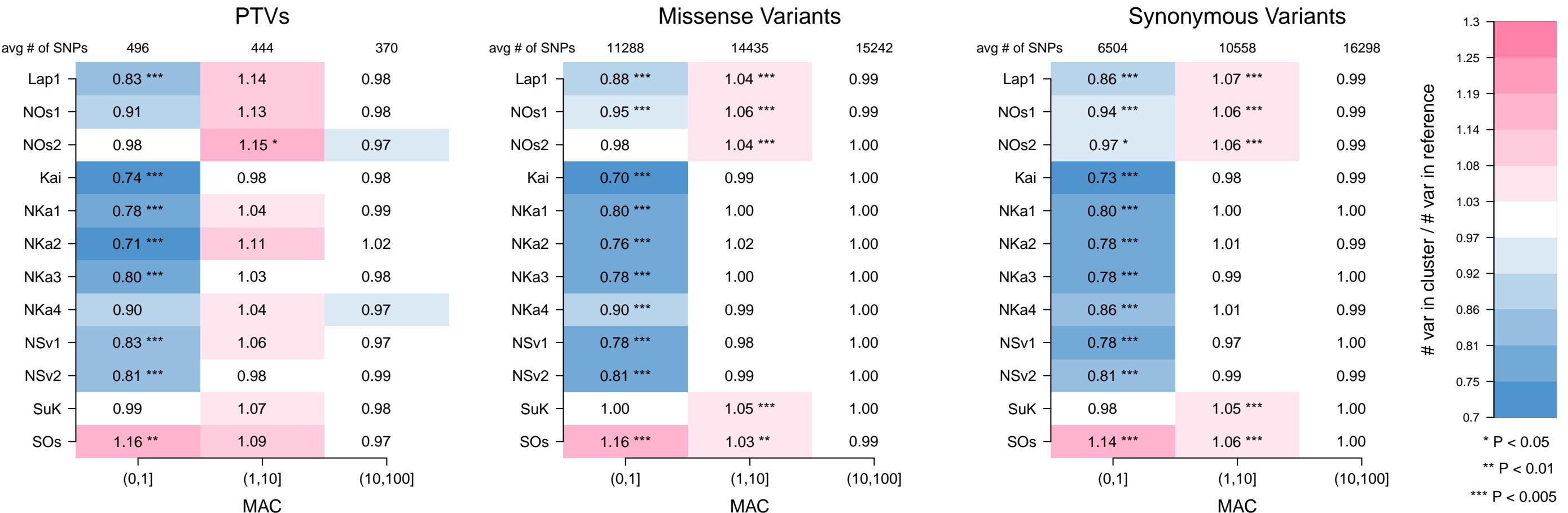
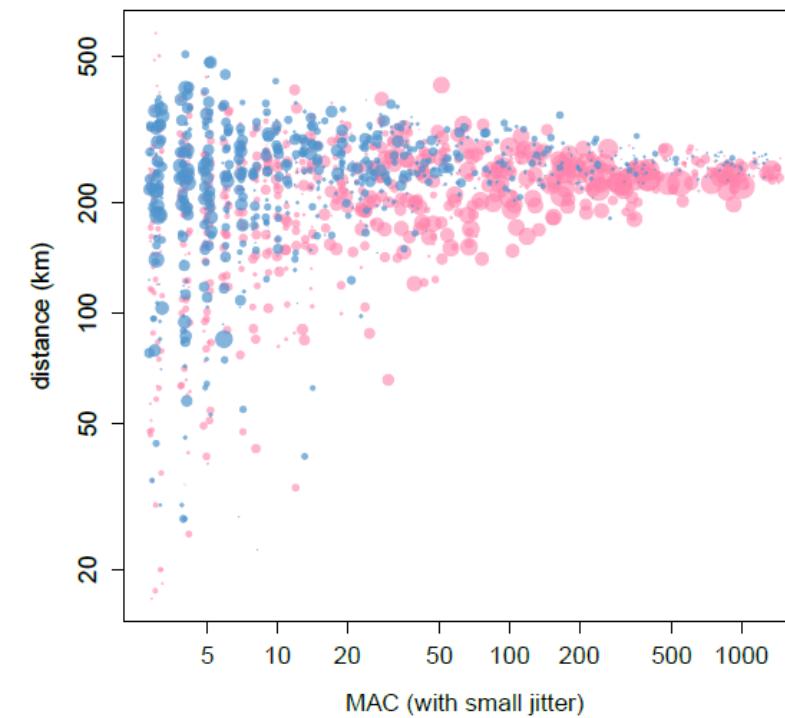
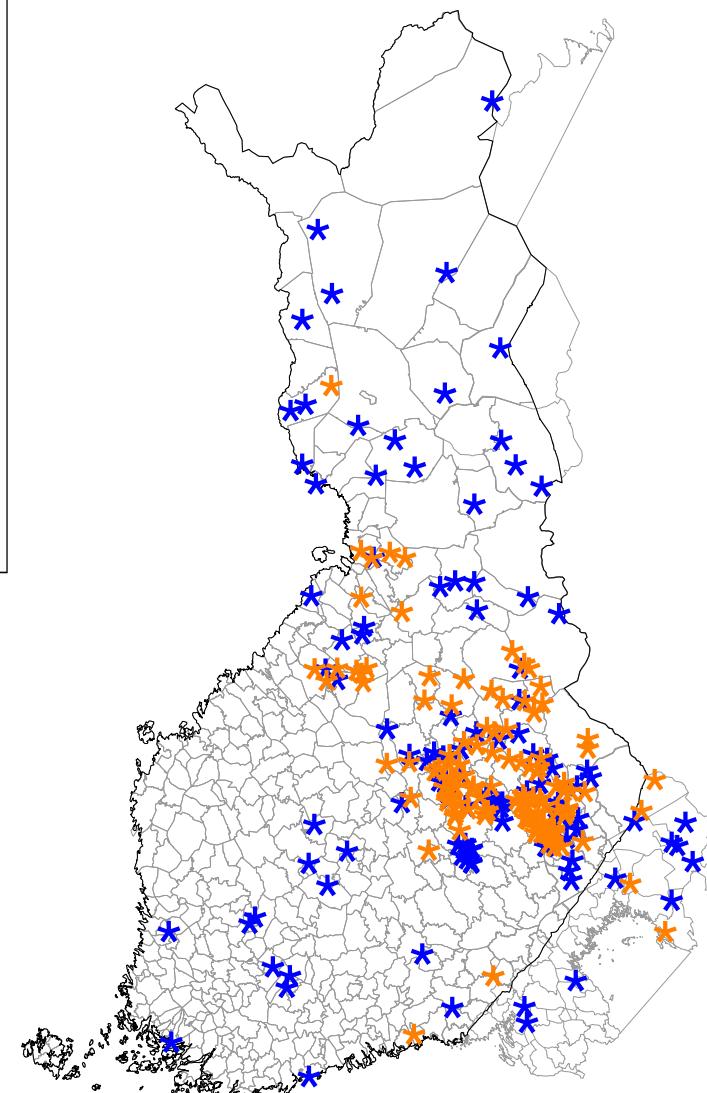


Figure 4. Regional variation in allele frequencies by functional annotation. Enrichment of functional allelic class in sub-populations (regions) of Northern and Eastern Finland. For each minor allele count bin, we computed the ratio of number of variants found in each subpopulation to an internal reference 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). The 12 sub-populations with sample size >100 are shown. The results are consistent with multiple independent bottlenecks followed by subsequent drift in Northern and Eastern Finland, particularly for populations in Lapland and Northern Ostrobothnia. Abbreviations for regions: Kainuu (Kai), Lapland (Lap1, Lap2), Northern Karelia (NKa1, NKa2, NKa3, NKa4), Northern Ostrobothnia (NOs1, NOs2, NOs3, NOs4), Northern Savonia (NSv1, NSv2, NSv3), Southern Ostrobothnia (SOs), and Surrendered Karelia (SuK). For more detailed information on region definitions see Supplementary Table 15. Confidence intervals on the enrichment ratios, and their P-values, are presented in Supplementary Table 16.

A.



B.



C.

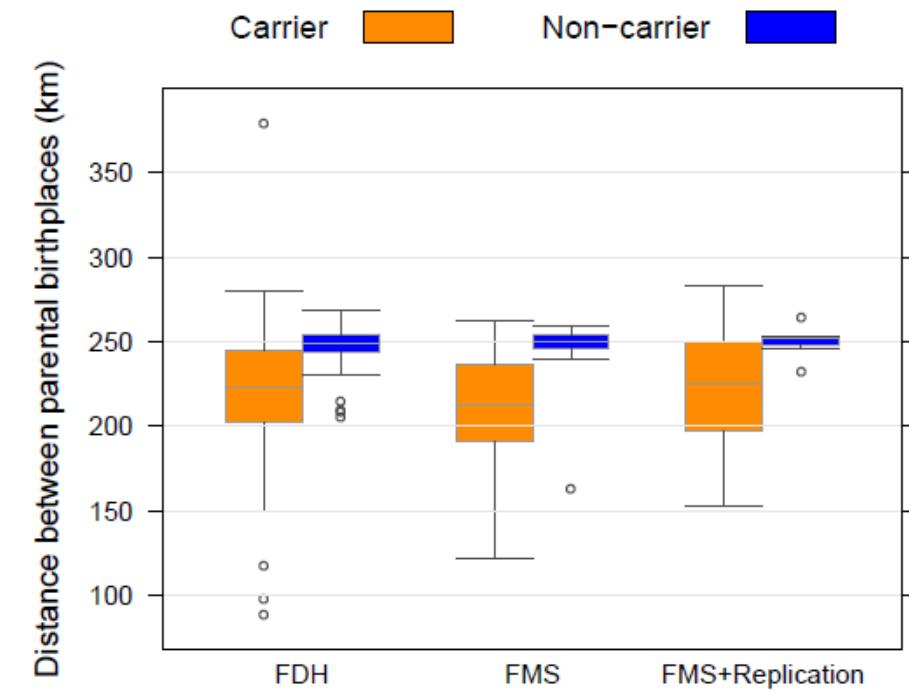


Figure 5. Geographical clustering of associated variants.

A) Geographical clustering of PTVs as a function of MAC and frequency enrichment over NFE from gnomAD. For each PTV ( $r^2 \leq 0.02$ ,  $MAC \geq 3$ ,  $MAF \leq 0.05$ ) we computed the mean distance between birth places of available parents of all carriers of the minor allele. We compared the frequency of the minor allele in FinMetSeq to gnomAD NFE. Blue and pink colors denote the frequency is lower or higher in FinMetSeq than in gnomAD NFE, respectively. The size of the point is proportional to the logarithm of the frequency ratio difference. In general, we observe that variants enriched in FinMetSeq are more geographically clustered.

B) Example of geographical clustering for a trait associated variant. The birth locations of all parents of carriers (orange) and a matching number of parents of non-carriers (blue) of the minor allele for variant chr3:125831672 (rs780671030, p.Arg722X) in *ALDH1L1* are displayed on a map of Finland. This variant is associated with serum glycine levels in FinMetSeq and has a frequency of 0 in NFE samples from gnomAD. The parents of carriers are born on average 135 km apart, the parents of non-carriers on average 250 km apart ( $P < 10^{-7}$  by permutation).

C) Comparison of geographical clustering between Finnish Disease Heritage (FDH) mutations and trait-associated variants that are  $>10$  times more frequent in FinMetSeq than in NFE. The degree of geographical clustering (based on parental birthplace) is comparable between carriers of those variants that showed significant associations in FinMetSeq alone (FMS) and carriers of FDH mutations, and greater than that seen in carriers of variants that showed significant association only in the combined analysis (FMS+Replication). For all variants, carriers of the minor allele displayed greater clustering than non-carriers. The bar within each box is the median, the box represents the inter-quartile range, whiskers extend up to 1.5x the interquartile range, and outliers are presented as individual points.