



FinnGen Project (2019). Exome sequencing of Finnish isolates enhances rare-variant association power. *Nature*, 572, 323-328. <https://doi.org/10.1038/s41586-019-1457-z>

Peer reviewed version

Link to published version (if available):  
[10.1038/s41586-019-1457-z](https://doi.org/10.1038/s41586-019-1457-z)

[Link to publication record in Explore Bristol Research](#)  
PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online via Springer Nature at <https://www.nature.com/articles/s41586-019-1457-z>. Please refer to any applicable terms of use of the publisher.

## University of Bristol - Explore Bristol Research

### General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: <http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/>

1 **Exome sequencing of Finnish isolates enhances rare-variant association power**

2 Locke, Adam E<sup>1,2,3,\*</sup>; Steinberg, Karyn Meltz<sup>2,4,\*</sup>; Chiang, Charleston WK<sup>5,6,7,\*</sup>; Service,  
3 Susan K<sup>5,\*</sup>; Havulinna, Aki S<sup>8,9</sup>; Stell, Laurel<sup>10</sup>; Pirinen, Matti<sup>8,11,12</sup>; Abel, Haley J<sup>2,13</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>2,14,15</sup>; Larson, David E<sup>2,13</sup>; Nelson, Joanne<sup>2</sup>; Nicholas,  
6 Thomas J<sup>2,16</sup>; Pietilä, Arto<sup>9</sup>; Ramensky, Vasily<sup>5,17</sup>; Ray, Debashree<sup>3,18</sup>; Scott, Laura J<sup>3</sup>;  
7 Stringham, Heather M<sup>3</sup>; Vangipurapu, Jagadish<sup>19</sup>; Welch, Ryan<sup>3</sup>; Yajnik, Pranav<sup>3</sup>; Yin,  
8 Xianyong<sup>3</sup>; Eriksson, Johan G<sup>20,21,22</sup>; Ala-Korpela, Mika<sup>23,24,25,26,27,28</sup>; Järvelin, Marjo-  
9 Riitta<sup>29,30,31,32,33</sup>; Männikkö, Minna<sup>30,34</sup>; Laivuori, Hannele<sup>8,35,36</sup>; FinnGen Project;  
10 Dutcher, Susan K<sup>2,13</sup>; Stitzel, Nathan O<sup>2,37</sup>; Wilson, Richard K<sup>2,14,15</sup>; Hall, Ira M<sup>1,2</sup>;  
11 Sabatti, Chiara<sup>10,38</sup>; Palotie, Aarno<sup>8,39,40</sup>; Salomaa, Veikko<sup>9</sup>; Laakso, Markku<sup>19,41</sup>; Ripatti,  
12 Samuli<sup>8,11,40</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>Quantitative and Computational Biology Section, Department of Biological Sciences,  
25 University of Southern California, Los Angeles, CA

26 <sup>8</sup>Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki,  
27 Finland

28 <sup>9</sup>National Institute for Health and Welfare, Helsinki, Finland

29 <sup>10</sup>Department of Biomedical Data Science, Stanford University, Stanford, CA

30 <sup>11</sup>Department of Public Health, University of Helsinki, Helsinki, Finland;

31 <sup>12</sup>Helsinki Institute for Information Technology HIIT and Department of Mathematics  
32 and Statistics, University of Helsinki, Helsinki, Finland

33 <sup>13</sup>Department of Genetics, Washington University School of Medicine, St. Louis, MO

34 <sup>14</sup>The Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH

35 <sup>15</sup>Department of Pediatrics, The Ohio State University College of Medicine, Columbus,  
36 OH

37 <sup>16</sup>USTAR Center for Genetic Discovery and Department of Human Genetics, University  
38 of Utah, Salt Lake City, UT

39 <sup>17</sup>Federal State Institution "National Medical Research Center for Preventive Medicine"  
40 of the Ministry of Healthcare of the Russian Federation, Moscow, Russia

41 <sup>18</sup>Departments of Epidemiology and Biostatistics, Bloomberg School of Public Health,  
42 Johns Hopkins University, Baltimore, MD

43 <sup>19</sup>Institute of Clinical Medicine, Internal Medicine, University of Eastern Finland,  
44 Kuopio, Finland

45 <sup>20</sup>Department of Public Health Solutions, National Institute for Health and Welfare,  
46 Helsinki, Finland  
47 <sup>21</sup>Folkhälsan Research Center, Helsinki, Finland  
48 <sup>22</sup>Department of General Practice and Primary Health Care, University of Helsinki,  
49 Helsinki and Helsinki University Hospital, Helsinki, Finland  
50 <sup>23</sup>Systems Epidemiology, Baker Heart and Diabetes Institute, Melbourne, Victoria,  
51 Australia  
52 <sup>24</sup>Computational Medicine, Faculty of Medicine, University of Oulu and Biocenter Oulu,  
53 University of Oulu, Oulu, Finland  
54 <sup>25</sup>NMR Metabolomics Laboratory, School of Pharmacy, University of Eastern Finland,  
55 Kuopio, Finland  
56 <sup>26</sup>Population Health Science, Bristol Medical School, University of Bristol, Bristol, UK  
57 <sup>27</sup>Medical Research Council Integrative Epidemiology Unit at the University of Bristol,  
58 Bristol, UK  
59 <sup>28</sup>Department of Epidemiology and Preventive Medicine, School of Public Health and  
60 Preventive Medicine, Faculty of Medicine, Nursing and Health Sciences, The Alfred  
61 Hospital, Monash University, Melbourne, Victoria, Australia  
62 <sup>29</sup>Biocenter Oulu, University of Oulu, Oulu, Finland  
63 <sup>30</sup>Center for Life Course Health Research, Faculty of Medicine, University of Oulu, Oulu,  
64 Finland  
65 <sup>31</sup>Unit of Primary Health Care, Oulu University Hospital, Oulu, Finland  
66 <sup>32</sup>Department of Epidemiology and Biostatistics, MRC-PHE Centre for Environment and  
67 Health, School of Public Health, Imperial College London, London, UK  
68 <sup>33</sup>Department of Life Sciences, College of Health and Life Sciences, Brunel University  
69 London, Uxbridge, UK  
70 <sup>34</sup>Northern Finland Birth Cohorts, Faculty of Medicine, University of Oulu, Oulu,  
71 Finland  
72 <sup>35</sup>Medical and Clinical Genetics, University of Helsinki and Helsinki University Hospital,  
73 Helsinki, Finland  
74 <sup>36</sup>Department of Obstetrics and Gynecology, Tampere University Hospital and University  
75 of Tampere, Faculty of Medicine and Life Sciences, Tampere, Finland  
76 <sup>37</sup>Cardiovascular Division, Department of Medicine, Washington University School of  
77 Medicine, St. Louis, MO  
78 <sup>38</sup>Department of Statistics, Stanford University, Stanford, CA  
79 <sup>39</sup>Analytical and Translational Genetics Unit (ATGU), Psychiatric &  
80 Neurodevelopmental Genetics Unit, Departments of Psychiatry and Neurology,  
81 Massachusetts General Hospital, Boston, MA  
82 <sup>40</sup>Broad Institute of MIT and Harvard, Cambridge, MA  
83 <sup>41</sup>Department of Medicine, Kuopio University Hospital, Kuopio, Finland  
84  
85 \*These authors contributed equally to this work.  
86 †These authors jointly supervised this work.

87 **ABSTRACT**

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

100

101 **INTRODUCTION**

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

110 populations that have expanded in isolation after recent bottlenecks, as alleles passing  
111 through the bottlenecks may rise to much higher frequencies than in other populations<sup>6-8</sup>.

112

113 Finland exemplifies such a history. Bottlenecks occurred at the founding of early-  
114 settlement regions (southern and western Finland) 2,000-4,000 years ago and again with  
115 internal migration to late-settlement regions (northern and eastern Finland) in the 15<sup>th</sup> and  
116 16<sup>th</sup> centuries<sup>9</sup>. Finland's subsequent population growth (to ~5.5 million) generated sizable  
117 geographic sub-isolates in late-settlement regions.

118

119 This unique population history has resulted in “the Finnish Disease Heritage”<sup>10</sup>, 36  
120 Mendelian diseases that are much more common in Finns than in other Europeans. These  
121 disorders concentrate in late-settlement regions of Finland<sup>10</sup>, and the genes responsible for  
122 them exhibit extreme enrichment of deleterious variants<sup>11-13</sup>. We created the FinMetSeq  
123 study to capitalize on the population history of late-settlement Finland to discover rare-  
124 variant associations with cardiovascular and metabolic disease-relevant quantitative traits  
125 through exome sequencing of two extensively phenotyped population cohorts, FINRISK  
126 and METSIM (Methods).

127

128 We successfully sequenced 19,292 FinMetSeq participants and tested the identified  
129 variants for association with 64 clinically relevant quantitative traits, discovering 43 novel  
130 associations with deleterious variants<sup>14,15</sup>: 19 associations (11 traits) in FinMetSeq alone  
131 and 24 associations (20 traits) in a combined analysis of FinMetSeq with 24,776 Finns  
132 from three cohorts with imputed genome-wide genotypes. Nineteen of the 26 variants

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

137

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

144

## 145 **RESULTS**

### 146 **Genetic variation**

147 In 19,292 successfully sequenced exomes, we identified 1,318,781 single nucleotide  
148 variants (SNVs) and 92,776 insertion/deletion (indel) variants (**Supplementary Tables 1-**  
149 **3, Supplementary Information**). Compared to NFE control exomes (gnomAD v2.1,  
150 **Extended Data Fig. 1A**), FinMetSeq exomes showed depletion of singletons and  
151 doubletons and excess variants with minor allele count (MAC) $\geq$ 5, particularly for  
152 predicted-deleterious alleles (**Extended Data Fig. 1B**).

153

### 154 **Association analyses**

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

157 transformation to normality for analyses (Methods, **Supplementary Tables 4 & 5**). Sixty-  
158 two of 64 traits exhibited significant heritability with common SNVs ( $P < 0.05$ ;  
159  $5\% < h^2 < 53\%$ ; **Extended Data Fig. 2A, Supplementary Table 6**), with substantial  
160 phenotypic and genetic correlations between traits (**Extended Data Fig. 2B**).

161

162 Single-variant association tests with genetic variants with  $MAC \geq 3$  among the 3,558 to  
163 19,291 individuals measured for each trait (**Supplementary Tables 4 & 5**) identified 1,249  
164 associations ( $P < 5 \times 10^{-7}$ ) at 531 variants (**Supplementary Table 7**); 53 traits associated  
165 with  $\geq 1$  variant (**Fig. 1A**). All 1,249 associations remained significant after multiple testing  
166 adjustment (exome-wide and across the 64 traits using a hierarchical procedure setting  
167 average FDR at 5%, Methods). Using this procedure on the 531 associated variants, we  
168 detected 287 more associations (**Supplementary Table 8**), most reflecting high correlation  
169 between lipid traits. Of the 531 variants, those at  $>10x$  frequency in FinMetSeq compared  
170 to NFE were more likely to be trait-associated ( $OR = 4.92$ ,  $P = 2.6 \times 10^{-5}$ ; **Extended Data Fig.**  
171 **1C**).

172

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

177

178 Most variants and loci (61%) associated to a single trait; 4% associated to  $\geq 10$  traits.  
179 Overlapping associations (**Extended Data Fig. 3A**) reflect both phenotypic and genetic

180 correlations and the estimated genetic correlation of trait pairs predicts shared loci between  
181 traits (**Extended Data Fig. 3B**). Gene-based association tests revealed 54 associations with  
182  $P < 3.88 \times 10^{-6}$  and multi-trait  $FDR < 0.05$  (Methods, **Supplementary Table 9**), including ten  
183 traits associated with *APOB* (**Extended Data Fig. 4**) and a novel association of *SECTMI*  
184 with HDL2-C (**Extended Data Fig. 5**).

185

186 To determine which of the 1,249 single-variant associations are distinct from previous  
187 GWAS findings, we repeated association analysis for each trait conditioning on published  
188 associated variants in the EBI GWAS Catalog (December 2016, Methods); 478  
189 associations at 126 loci remained significant ( $P < 5 \times 10^{-7}$ ), including at least one association  
190 for 48 traits (**Supplementary Table 10**). Conditionally-associated variants were more  
191 often rare (24% vs. 11%), more likely protein-altering (31% vs. 22%), and more frequently  
192  $>10x$  enriched in FinMetSeq relative to NFE (19% vs. 10%) than associated variants  
193 overall.

194

### 195 **Replication and follow-up**

196 We attempted to replicate the 478 single-variant associations (unconditional and  
197 conditional  $P \leq 5 \times 10^{-7}$ ) and follow up 2,120 sub-threshold associations from FinMetSeq  
198 (unconditional  $5 \times 10^{-7} < P \leq 5 \times 10^{-5}$  and conditional  $P \leq 5 \times 10^{-5}$ ) in 24,776 participants from  
199 three Finnish cohort studies: FINRISK<sup>16,17</sup> participants not in FinMetSeq (n=18,215),  
200 Northern Finland Birth Cohort 1966<sup>18</sup> (n=5,139), and Helsinki Birth Cohort<sup>19</sup> (n=1,412),  
201 all imputed using the Finnish SISu v2 reference panel ([www.sisuproject.fi](http://www.sisuproject.fi)). Following  
202 association analysis within each cohort, we conducted meta-analysis of the three



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

206

207 Of 448 significant variant-trait associations with replication data, 392 (87.5%) replicated  
208 at  $P < 0.05$  (**Supplementary Table 11**). Of the 1,417 sub-threshold associations, 431  
209 reached  $P < 5 \times 10^{-7}$  in the combined analysis (**Supplementary Table 12**);  $>60\%$  of variants  
210 we could not follow up were absent in the reference panel.

211

212 Among the significant associations from FinMetSeq or combined analysis, 43 were with  
213 26 predicted deleterious variants (six PTVs, 20 missense) that conditional analysis and  
214 literature review suggest are novel (**Table 1**). Nineteen associations (15 variants) were  
215 significant in FinMetSeq (**Table 1; Supplementary Table 11**); another 24 associations (16  
216 variants) reached significance in combined analysis (**Table 1; Supplementary Table 12**).  
217 Of these 43 associations, 34 were with 19 variants either seen only in Finland or enriched  
218  $>20$ -fold in FinMetSeq compared to NFE. Identifying associations for these 19 variants  
219 would have required much larger samples in NFE populations than in FinMetSeq (**Fig. 2A,**  
220 **B**). We provide brief summaries relating some of these associations to known biology and  
221 prior genetic evidence (**Table 1**, expanded version in **Supplementary Table 13,**  
222 **Supplementary Information**), highlighting here the most striking findings.

223

224 *Anthropometric traits.* A predicted damaging missense variant (p.Arg94Cys) in *THBS4*  
225 45X more frequent in FinMetSeq than in NFE was associated in the combined analysis

226 with a mean 5.9 kg decrease in body weight. *THBS4* encodes thrombospondin 4, a  
227 matricellular protein found in blood vessel walls and highly expressed in heart and adipose  
228 <sup>20</sup>. *THBS4* may regulate vascular inflammation<sup>21</sup> and has been implicated in heart disease  
229 risk<sup>22</sup>.

230

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

241

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

247

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

255

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

261

## 262 **Trait-associations and disease endpoints**

263 Genotype data from FinnGen<sup>33</sup> enabled us to test whether deleterious variants responsible  
264 for our novel trait associations contribute to related disease endpoints. We examined 22  
265 diseases for the 25 available variants in **Table 1**; three variants were associated with  
266 diseases in FinnGen at Bonferroni threshold  $P<0.05/(22\times 25)=9.0\times 10^{-5}$  (**Supplementary**  
267 **Table 14**).

268

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

271 pancreatitis risk. While this is the first disease association reported for *KRT40*, type I  
272 keratins regulate exocrine pancreas homeostasis<sup>34</sup>. A 29bp deletion causing a frameshift  
273 in *FAM151A* is associated in FinMetSeq with decreased total cholesterol in IDL and  
274 decreased IDL particle concentration, is 6.7X more frequent in FinMetSeq than NFE, and  
275 is associated in FinnGen with decreased risk of myocardial infarction. Interpretation of this  
276 association is complicated as the variant is also situated in an overlapping gene (*ACOT11*)  
277 involved in fatty acid metabolism and lies <1Mbp from a cardioprotective variant in  
278 *PCSK9*. Finally, a predicted damaging missense variant (p.Arg65Trp) in *DBH* associated  
279 with a mean 1.0 mmHg decrease in diastolic blood pressure in the combined analysis, is  
280 23.8X more frequent in FinMetSeq than in NFE, and is associated in FinnGen with  
281 decreased risk for hypertension. Distinct loci in this gene and gene-based tests are  
282 associated with mean arterial pressure<sup>35,36</sup>.

283

#### 284 **Replication outside Finland**

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

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

297

### 298 **Enriched variants cluster geographically**

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

307

308 To analyze the distribution of variants within late-settlement Finland, we delineated  
309 geographically distinct population clusters using haplotype sharing among 2,644 unrelated  
310 individuals with both parents born in the same municipality (Methods, **Extended Data Fig.**  
311 **8**). We compared variant counts across functional classes and frequencies between an  
312 early-settlement reference cluster and 12 clusters containing  $\geq 100$  individuals (**Extended**  
313 **Data Fig. 9, Supplementary Tables 17, 18**). Clusters representing the most heavily  
314 bottlenecked late-settlement regions (Lapland and Northern Ostrobothnia) displayed a  
315 deficit of singletons and enrichment of intermediate frequency variants compared to other  
316 clusters.

317

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

325

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

331

## 332 **DISCUSSION**

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

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

345

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

351

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

359

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

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



368

## References

- 369 1 Samocha, K. E. *et al.* Regional missense constraint improves variant  
370 deleteriousness prediction. *bioRxiv*, doi:https://doi.org/10.1101/148353 (2017).
- 371 2 Marouli, E. *et al.* Rare and low-frequency coding variants alter human adult  
372 height. *Nature* **542**, 186-190, doi:10.1038/nature21039 (2017).
- 373 3 Flannick, J. *et al.* Exome sequencing of 20,791 cases of type 2 diabetes and  
374 24,440 controls. *Nature* **570**, 71-76, doi:10.1038/s41586-019-1231-2 (2019).
- 375 4 Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J. & Richards, J.  
376 B. Genetic architecture: the shape of the genetic contribution to human traits and  
377 disease. *Nature reviews. Genetics* **19**, 110-124, doi:10.1038/nrg.2017.101 (2018).
- 378 5 Zuk, O. *et al.* Searching for missing heritability: designing rare variant association  
379 studies. *Proc Natl Acad Sci U S A* **111**, E455-464, doi:10.1073/pnas.1322563111  
380 (2014).
- 381 6 Xue, Y. *et al.* Enrichment of low-frequency functional variants revealed by  
382 whole-genome sequencing of multiple isolated European populations. *Nature*  
383 *communications* **8**, 15927, doi:10.1038/ncomms15927 (2017).
- 384 7 Southam, L. *et al.* Whole genome sequencing and imputation in isolated  
385 populations identify genetic associations with medically-relevant complex traits.  
386 *Nature communications* **8**, 15606, doi:10.1038/ncomms15606 (2017).
- 387 8 Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature*  
388 **461**, 747-753, doi:10.1038/nature08494 (2009).
- 389 9 Jakkula, E. *et al.* The genome-wide patterns of variation expose significant  
390 substructure in a founder population. *American journal of human genetics* **83**,  
391 787-794, doi:10.1016/j.ajhg.2008.11.005 (2008).
- 392 10 Polvi, A. *et al.* The Finnish disease heritage database (FinDis) update-a database  
393 for the genes mutated in the Finnish disease heritage brought to the next-  
394 generation sequencing era. *Hum Mutat* **34**, 1458-1466, doi:10.1002/humu.22389  
395 (2013).
- 396 11 Manning, A. *et al.* A Low-Frequency Inactivating AKT2 Variant Enriched in the  
397 Finnish Population Is Associated With Fasting Insulin Levels and Type 2  
398 Diabetes Risk. *Diabetes* **66**, 2019-2032, doi:10.2337/db16-1329 (2017).
- 399 12 Lim, E. T. *et al.* Distribution and medical impact of loss-of-function variants in  
400 the Finnish founder population. *PLoS genetics* **10**, e1004494,  
401 doi:10.1371/journal.pgen.1004494 (2014).
- 402 13 Service, S. K. *et al.* Re-sequencing expands our understanding of the phenotypic  
403 impact of variants at GWAS loci. *PLoS genetics* **10**, e1004147,  
404 doi:10.1371/journal.pgen.1004147 (2014).
- 405 14 Wurtz, P. *et al.* Quantitative Serum Nuclear Magnetic Resonance Metabolomics  
406 in Large-Scale Epidemiology: A Primer on -Omic Technologies. *American*  
407 *journal of epidemiology* **186**, 1084-1096, doi:10.1093/aje/kwx016 (2017).
- 408 15 Laakso, M. *et al.* The Metabolic Syndrome in Men study: a resource for studies of  
409 metabolic and cardiovascular diseases. *Journal of lipid research* **58**, 481-493,  
410 doi:10.1194/jlr.O072629 (2017).
- 411 16 Borodulin, K. *et al.* Forty-year trends in cardiovascular risk factors in Finland.  
412 *Eur J Public Health* **25**, 539-546, doi:10.1093/eurpub/cku174 (2015).

- 413 17 Abraham, G. *et al.* Genomic prediction of coronary heart disease. *Eur Heart J* **37**,  
414 3267-3278, doi:10.1093/eurheartj/ehw450 (2016).
- 415 18 Sabatti, C. *et al.* Genome-wide association analysis of metabolic traits in a birth  
416 cohort from a founder population. *Nature genetics* **41**, 35-46, doi:10.1038/ng.271  
417 (2009).
- 418 19 Pulizzi, N. *et al.* Interaction between prenatal growth and high-risk genotypes in  
419 the development of type 2 diabetes. *Diabetologia* **52**, 825-829,  
420 doi:10.1007/s00125-009-1291-1 (2009).
- 421 20 Fagerberg, L. *et al.* Analysis of the human tissue-specific expression by genome-  
422 wide integration of transcriptomics and antibody-based proteomics. *Mol Cell*  
423 *Proteomics* **13**, 397-406, doi:10.1074/mcp.M113.035600 (2014).
- 424 21 Corsetti, J. P. *et al.* Thrombospondin-4 polymorphism (A387P) predicts  
425 cardiovascular risk in postinfarction patients with high HDL cholesterol and C-  
426 reactive protein levels. *Thromb Haemost* **106**, 1170-1178, doi:10.1160/TH11-03-  
427 0206 (2011).
- 428 22 Zhang, X. J. *et al.* Association between single nucleotide polymorphisms in  
429 thrombospondins genes and coronary artery disease: A meta-analysis. *Thromb*  
430 *Res* **136**, 45-51, doi:10.1016/j.thromres.2015.04.019 (2015).
- 431 23 Beygo, J. *et al.* New insights into the imprinted MEG8-DMR in 14q32 and  
432 clinical and molecular description of novel patients with Temple syndrome. *Eur J*  
433 *Hum Genet* **25**, 935-945, doi:10.1038/ejhg.2017.91 (2017).
- 434 24 Wallace, C. *et al.* The imprinted DLK1-MEG3 gene region on chromosome  
435 14q32.2 alters susceptibility to type 1 diabetes. *Nature genetics* **42**, 68-71,  
436 doi:10.1038/ng.493 (2010).
- 437 25 Day, F. R. *et al.* Genomic analyses identify hundreds of variants associated with  
438 age at menarche and support a role for puberty timing in cancer risk. *Nature*  
439 *genetics* **49**, 834-841, doi:10.1038/ng.3841 (2017).
- 440 26 Perry, J. R. *et al.* Parent-of-origin-specific allelic associations among 106  
441 genomic loci for age at menarche. *Nature* **514**, 92-97, doi:10.1038/nature13545  
442 (2014).
- 443 27 Cleaton, M. A. *et al.* Fetus-derived DLK1 is required for maternal metabolic  
444 adaptations to pregnancy and is associated with fetal growth restriction. *Nature*  
445 *genetics* **48**, 1473-1480, doi:10.1038/ng.3699 (2016).
- 446 28 Chaves, J. A. *et al.* Genomic variation at the tips of the adaptive radiation of  
447 Darwin's finches. *Mol Ecol* **25**, 5282-5295, doi:10.1111/mec.13743 (2016).
- 448 29 Surakka, I. *et al.* The impact of low-frequency and rare variants on lipid levels.  
449 *Nature genetics* **47**, 589-597, doi:10.1038/ng.3300 (2015).
- 450 30 Ding, Y. *et al.* Plasma Glycine and Risk of Acute Myocardial Infarction in  
451 Patients With Suspected Stable Angina Pectoris. *J Am Heart Assoc* **5**,  
452 doi:10.1161/JAHA.115.002621 (2015).
- 453 31 Wittemans, L. B. L. *et al.* Assessing the causal association of glycine with risk of  
454 cardio-metabolic diseases. *Nature communications* **10**, 1060, doi:10.1038/s41467-  
455 019-08936-1 (2019).
- 456 32 Perry, R. J. *et al.* Acetate mediates a microbiome-brain-beta-cell axis to promote  
457 metabolic syndrome. *Nature* **534**, 213-217, doi:10.1038/nature18309 (2016).

- 458 33 Tabbassum, R. *et al.* Genetics of human plasma lipidome: Understanding lipid  
459 metabolism and its link to diseases beyond traditional lipids. *bioRxiv*,  
460 doi:10.1101/457960 (2018).
- 461 34 Casanova, M. L. *et al.* Exocrine pancreatic disorders in transgenic mice  
462 expressing human keratin 8. *J Clin Invest* **103**, 1587-1595, doi:10.1172/JCI5343  
463 (1999).
- 464 35 Surendran, P. *et al.* Trans-ancestry meta-analyses identify rare and common  
465 variants associated with blood pressure and hypertension. *Nature genetics* **48**,  
466 1151-1161, doi:10.1038/ng.3654 (2016).
- 467 36 Liu, C. *et al.* Meta-analysis identifies common and rare variants influencing blood  
468 pressure and overlapping with metabolic trait loci. *Nature genetics* **48**, 1162-  
469 1170, doi:10.1038/ng.3660 (2016).
- 470 37 Palmer, C. & Pe'er, I. Statistical correction of the Winner's Curse explains  
471 replication variability in quantitative trait genome-wide association studies. *PLoS*  
472 *genetics* **13**, e1006916, doi:10.1371/journal.pgen.1006916 (2017).
- 473 38 Norio, R. Finnish Disease Heritage I: characteristics, causes, background. *Hum*  
474 *Genet* **112**, 441-456, doi:10.1007/s00439-002-0875-3 (2003).
- 475 39 Service, S. *et al.* Magnitude and distribution of linkage disequilibrium in  
476 population isolates and implications for genome-wide association studies. *Nature*  
477 *genetics* **38**, 556-560, doi:10.1038/ng1770 (2006).
- 478 40 Chiang, C. W. K. *et al.* Genomic history of the Sardinian population. *Nature*  
479 *genetics*, doi:10.1038/s41588-018-0215-8 (2018).
- 480 41 Rivas, M. A. *et al.* Insights into the genetic epidemiology of Crohn's and rare  
481 diseases in the Ashkenazi Jewish population. *PLoS genetics* **14**, e1007329,  
482 doi:10.1371/journal.pgen.1007329 (2018).
- 483 42 Bastarache, L. *et al.* Phenotype risk scores identify patients with unrecognized  
484 Mendelian disease patterns. *Science* **359**, 1233-1239, doi:10.1126/science.aal4043  
485 (2018).
- 486 43 Niemi, M. E. K. *et al.* Common genetic variants contribute to risk of rare severe  
487 neurodevelopmental disorders. *Nature*, doi:10.1038/s41586-018-0566-4 (2018).
- 488 44 Surakka, I. S., A.-P.; Ruotsalainen, S.E.; Durbin, R.; Salomaa, V.; Daly, M.;  
489 Palotie, A.; Ripatti, S. The rate of false polymorphisms introduced when imputing  
490 genotypes from global imputation panels. *bioRxiv*, doi:10.1101/080770 (2016).
- 491 45 Collins, F. S. & Varmus, H. A new initiative on precision medicine. *N Engl J Med*  
492 **372**, 793-795, doi:10.1056/NEJMp1500523 (2015).
- 493

## 494 **Supplementary Information**

495 Supplementary Information is linked to the online version of the paper at  
496 [www.nature.com/nature](http://www.nature.com/nature).

497

## 498 **Acknowledgements**

499 Thanks to Terri Teshiba for coordinating ethical permissions and samples. Thanks to Sini  
500 Kerminen, Daniel Lawson, and George Busby for discussions and providing scripts to run  
501 fineSTRUCTURE. SR was supported by the Academy of Finland Center of Excellence in  
502 Complex Disease Genetics (Grant No 312062), Academy of Finland (No. 285380), the  
503 Finnish Foundation for Cardiovascular Research, the Sigrid Juselius Foundation,  
504 Biocentrum Helsinki and University of Helsinki HiLIFE Fellow grant. VR acknowledges  
505 support by RFBR, research project No. 18-04-00789 A. VS was supported by the Finnish  
506 Foundation for Cardiovascular Research. CS and LS received funding from HG006695,  
507 HL113315, and MH105578. MAK is supported by a Senior Research Fellowship from the  
508 National Health and Medical Research Council (NHMRC) of Australia (APP1158958). He  
509 also works in a unit that is supported by the University of Bristol and UK Medical Research  
510 Council (MC\_UU\_12013/1). The Baker Institute is supported in part by the Victorian  
511 Government's Operational Infrastructure Support Program. AUJ, DR, LJS, HMS, RW, PY,  
512 XY, and MB received funding from DK062370. SKS, CWKC, and NBF received funding  
513 from HL113315 and NS062691. The METSIM study was supported by grants from  
514 Academy of Finland (No. 321428), the Sigrid Juselius Foundation, the Finnish Foundation  
515 for Cardiovascular Research, Kuopio University Hospital, and Centre of Excellence of  
516 Cardiovascular and Metabolic Diseases supported by the Academy of Finland (ML).  
517 Sequencing was funded by 5U54HG003079, and AEL, KMS, HJB, CCC, CJK, KLK,  
518 DCK, DEL, JN, TJN, SKD, NOS, IMH, and RKW were funded by 5U54HG003079 and  
519 5UM1HG008853-03.

520

521 **Author Contributions**

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

531

#### 532 **Author Information**

533 Reprints and permission information is available at [www.nature.com/reprints](http://www.nature.com/reprints)

534

535 Competing interests statements:

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

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

541

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

544

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

## Figure Legends

### 551 **Figure 1. Characterization of associations.**

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

554

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

560

### 561 **Figure 2. Allelic enrichment in the Finnish population and its effect on genetic** 562 **discovery.**

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

568

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

571

### 572 **Figure 3. Geographical clustering of associated variants.**

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

576

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

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



## Figure Legends (Extended Data Figures)

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

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

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

591 C) Comparison of MAFs for trait-associated variants in FinMetSeq and NFE gnomAD.  
592 Plotted in gray background is a 2-D histogram of variants with non-zero allele frequencies  
593 in both gnomAD and FinMetSeq but no trait associations. Variants associated with at least  
594 one trait are colored and scaled inversely proportional to the logarithm of the association  
595 p-value. Variants >10x enriched in FinMetSeq compared to NFE are pink, those <10x  
596 enriched are in blue. The dashed line is the line of equal frequency. Two-sided uncorrected  
597 P-values are from a regression of trait on the count of alternative allele at each variant. The  
598 number of independent individuals used in each point is listed in **Supplementary Table 5**.  
599

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

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

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

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

605 **Extended Data Fig. 3. Properties of associations shared between traits.**

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

610

611 B) Relationship between estimated genetic correlation and extent of sharing of genetic  
612 associations. For each trait-pair, the extent of locus sharing is defined as the number of loci  
613 associated with both traits divided by the total number of loci associated with either trait.  
614 Analysis using the absolute value of the Pearson correlation of the residual series results in  
615 a very similar pattern. The number of trait pairs in each x-axis category are as follows: 0-  
616 1%: 819; 1-10%: 204, 11-20%: 102; 21-30%: 41; 31-40%: 29; 41-50%: 16, >50%: 13. The  
617 bar within each box is the median, the box represents the upper and lower quartiles,  
618 whiskers extend to 1.5x the interquartile range, and points represent outliers.

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

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

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

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

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

650 **Extended Data Fig. 8. Hierarchical clustering tree produced by fineSTRUCTURE.**  
651 We identified 16 subpopulations within the FinMetSeq dataset by applying a haplotype-  
652 based clustering algorithm, fineSTRUCTURE, on 2,644 unrelated individuals born by  
653 1955 whose parents were both born in the same municipality (Methods). Each  
654 subpopulation is named based on the most common parental birth location among its  
655 members, with the following abbreviations: NKa, North Karelia; NSv, North Savonia;  
656 SOs, South Ostrobothnia; NOs, North Ostrobothnia; Kai, Kainuu; Lap, Lapland; SuK,  
657 Surrendered Karelia. A map of Finland with regions labeled is supplied for reference. If  
658 multiple subpopulations share the same location label, the subpopulation is further

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

662 **Extended Data Fig. 9. Regional variation in allele frequencies by functional**  
663 **annotation.** Enrichment of variants by allelic class in regional sub-populations of late  
664 settlement Finland (defined in **Supplementary Table 17**). Each bin represents the ratio of  
665 variants in the subpopulation compared to the reference subpopulation (NSv3), after down-  
666 sampling the frequency spectra of all populations to 200 chromosomes. Pink cells represent  
667 an enrichment (ratio >1), blue cells represent a depletion (ratio <1). Sample sizes and  
668 confidence intervals on each enrichment ratios, and their P-values, are presented in  
669 **Supplementary Table 18**. The results are consistent with multiple bottlenecks in late  
670 settlement Finland, particularly for populations in Lapland and Northern Ostrobothnia.  
671

## **METHODS**

### **672 METSIM and FINRISK studies: designs, phenotypes, and sequenced participants**

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

677

678 **FINRISK** is a series of health examination surveys based on random population samples  
679 from five (six in 2002) geographical regions of Finland, carried out every five years  
680 beginning in 1972<sup>47</sup>. For exome sequencing, we chose 10,192 participants in the 1992-  
681 2007 FINRISK surveys from northeastern Finland (former provinces of North Karelia,  
682 Oulu, and Lapland).

683

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

688

### **689 Selection of traits, harmonization, exclusions, covariate adjustment, and 690 transformation**

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

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

701

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

710

### 711 **Exome sequencing**

712 We carried out exome sequencing in two phases.

713

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

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

719

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

727

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

731

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

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

737

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

741

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

747

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

754

755 Following initial variant filtering, we decomposed multi-allelic variants into bi-allelic  
756 variants, left-aligned indels, and dropped redundant variants using vt<sup>54</sup> (version 0.5). We  
757 filtered variants with >2% missing calls and/or Hardy-Weinberg  $p$ -value  $< 10^{-6}$ . We  
758 additionally removed variants with an overall allele balance (alternate AC/sum of total AC)  
759  $< 30\%$  in genotyped samples. We excluded 86 individuals with >2% missing variant calls  
760 yielding a final analysis set of 19,292 individuals.



761

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

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

771

772 **Annotation**

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

777

778 **Association testing**

779 *Single variants*

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

784 individuals. We used genotypes for sequenced variants with  $MAF \geq 1\%$  to construct the  
785 genetic relationship matrix (GRM).

786

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

788 To differentiate association signals identified here from known associations, for each trait  
789 we performed exome-wide association analysis conditioning on variants previously  
790 associated ( $P < 10^{-7}$ ) with that trait in the EBI GWAS catalog  
791 (<https://www.ebi.ac.uk/gwas/downloads>; December 4, 2016 version)<sup>64</sup>, publications, or  
792 manuscripts in preparation<sup>55,65-67</sup>. The keywords from the GWAS catalog we used to assign  
793 known variants to each trait are in **Supplementary Table 21**. We also manually curated  
794 published associations for specific metabolites<sup>65,68</sup>.

795

796 Using the combined HRC+exome panel, we pruned each trait-specific list of associated  
797 variants (“GWAS variants”) based on linkage disequilibrium (LD) ( $r^2 > 0.95$ ). Of 23 GWAS  
798 variants absent in the HRC+exome panel, we identified a proxy ( $r^2 > 0.80$ ) variant for 17;  
799 we excluded the remaining six variants from conditional analysis. The variants included in  
800 conditional analysis are listed in **Supplementary Table 22**. We extracted genotypes for  
801 variants used in conditional analysis from the HRC+exome panel and converted dosages  
802 to alternate allele counts by rounding to the nearest integer (0, 1, or 2). For conditional  
803 analyses, we imputed missing genotypes for the individuals without array data using the  
804 mean genotype. We then ran association analysis using the same linear mixed model  
805 approach as in unconditional analysis but including the complete set of pruned GWAS

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

808

### 809 *Defining loci*

810 To identify the number of distinct associations for each trait, we performed LD clumping  
811 using Swiss (<https://github.com/welchr/swiss>) of variants with (1) unconditional  $P < 5 \times 10^{-7}$   
812 or (2) both unconditional and conditional  $P < 5 \times 10^{-5}$  for at least one trait. For each variant  
813 in this subset, we provided Swiss with the minimum unconditional p-value across all traits.  
814 The clumping procedure starts with the variant with the smallest p-value, merges into one  
815 locus all variants within  $\pm 1$ Mbp that have  $r^2 > 0.5$  with the index variant, and iterates this  
816 process until no variants remain.

817

### 818 *Calculating effects and variance explained of individual variants*

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

824

### 825 *Gene-based testing*

826 We carried out gene-based association tests using the mixed model implementation of  
827 SKAT-O<sup>69</sup>, considering three different, but nested, sets of variants (variant “masks”):

828 (1) PTVs at any allele frequency with VEP annotations: frameshift\_variant,  
829 initiator\_codon\_variant, splice\_acceptor\_variant, splice\_donor\_variant, stop\_lost,  
830 stop\_gained;  
831 (2) PTVs included in (1) plus missense variants with MAF<0.1% scored as “damaging” or  
832 “deleterious” by all five functional prediction algorithms;  
833 (3) PTVs included in (1) plus missense variants with MAF<0.5% scored as “damaging” or  
834 “deleterious” by all five algorithms.

835

836 For each trait and mask, we only tested genes with at least two qualifying variants. Each  
837 mask contained a different number of genes with at least two qualifying variants: up to  
838 7,996, 12,795, and 12,890 for the three masks, respectively. The exact number of genes  
839 tested varied by trait due to sample size. We first used a Bonferroni-corrected exome-wide  
840 threshold for 12,890 genes, which corresponds to a threshold of  $P < 3.88 \times 10^{-6}$ . Analogous  
841 to single-variant association, we passed genes meeting this association threshold forward  
842 for additional consideration with hierarchical FDR correction, described below.

843

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

845 To control for multiple testing across 64 traits, we adopted an FDR controlling procedure<sup>70</sup>,  
846 using a two-stage hierarchical strategy (described in **Supplementary Information**). Stage  
847 1 identifies the set of R variants (or genes) associated with at least one trait ( $P < 5 \times 10^{-7}$  for  
848 single-variant unconditional results and  $P < 3.88 \times 10^{-6}$  for gene-based results), controlling  
849 genome-wide FDR across all variants at 0.05. Stage 2 identifies all traits associated with  
850 the discovered variants in a manner guaranteeing an average FDR<0.05.

851

852 **Genotype validation**

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

856

857 **Replication in additional Finnish cohorts**

858 We attempted to replicate significant single-variant associations ( $P < 5 \times 10^{-7}$ ) and follow-up  
859 suggestive single-variant associations ( $P < 5 \times 10^{-5}$ ) using imputed array data from up to  
860 24,776 individuals from three cohort studies: Northern Finland Birth Cohort 1966  
861 (NFBC1966)<sup>18</sup>, the Helsinki Birth Cohort Study (HBCS)<sup>19</sup>, and FINRISK study  
862 participants not included in FinMetSeq<sup>16,17</sup>.

863

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

871

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

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

877

### 878 **Association to disease endpoints**

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

882

### 883 **Association replication in UK Biobank**

884 For eight FinMetSeq anthropometric and blood pressure traits available in UKB (height,  
885 weight, BMI, hip circumference, waist circumference, fat percentage, systolic blood  
886 pressure, and diastolic blood pressure), we extracted, for variants reaching  $P < 5 \times 10^{-7}$  in our  
887 combined analysis, trait-variant association statistics from [http://www.nealelab.is/uk-](http://www.nealelab.is/uk-biobank)  
888 [biobank](http://www.nealelab.is/uk-biobank). Seven of the eight traits had at least one associated variant and 23 of the total of  
889 31 variants were available in UKBB. A comparison of association results is in  
890 **Supplementary Table 15.**

891

### 892 **Population genetic analyses**

#### 893 *Identifying unrelated individuals*

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

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

900

#### 901 *Enrichment of predicted-deleterious alleles in Finland*

902 We assessed enrichment of predicted-deleterious alleles in Finland by comparing the  
903 14,874 nearly unrelated FinMetSeq individuals to the 14,944 NFE control exomes in  
904 gnomAD (after removing NFE individuals from countries with substantial Finnish  
905 populations, Estonia and Sweden). We analyzed the two most common alleles at each site  
906 with base quality score >10, mapping quality score >20, and coverage equal to or greater  
907 than that found in  $\geq 80\%$  of variable sites (17.73X in FinMetSeq, 32.27X in gnomAD),  
908 resulting in ~38.6 Mbp for comparisons. We contrasted the proportional site frequency  
909 spectra for FinMetSeq and NFE for five functional variant categories (PTVs, missense,  
910 synonymous, UTR, and intronic variants) after down-sampling both datasets to 18,000  
911 chromosomes.

912

913 We also assessed the enrichment of deleterious alleles within subpopulations of the  
914 FinMetSeq dataset. We applied Chromopainter and fineSTRUCTURE on 2,644 unrelated  
915 FinMetSeq individuals whose parents were both born in the same municipality to identify  
916 16 sub-population clusters<sup>74</sup> (**Supplementary Information**). Of the 16 clusters, we used  
917 as the reference population a cluster for which the highest proportion of the parents of its  
918 members were from early-settlement Finland (NSv3, **Supplementary Table 17**). We used  
919 the twelve clusters with >100 members in subsequent analyses (**Supplementary Table**

920 17). We then compared the ratio of the site frequency spectra to the reference for PTVs,  
921 missense, and synonymous variants, down-sampling both datasets to 200 haploid  
922 chromosomes. For each comparison, we computed statistical evidence for enrichment or  
923 depletion at a given allele count bin by exact binomial test against a null of equal number  
924 of variants found in both the test and reference cluster.

925

### 926 *Geographical clustering of predicted functionally deleterious alleles*

927 We first generated a distance matrix tabulating the pairwise geographical distance between  
928 the birthplaces of all available parents of unrelated sequenced individuals. For each variant  
929 of interest, we computed for the minor allele carriers in FinMetSeq the mean distance  
930 among all parent pairs. We evaluated statistical significance of geographical clustering by  
931 comparing the observed mean distance to mean distances for up to 10,000,000 sets of  
932 randomly drawn non-carrier individuals matched by cohort status and number of parents  
933 with birthplace information available. Birthplaces of carrier and non-carrier individuals  
934 were plotted on a map of Finland, including regions that were ceded prior to WW2 (©  
935 Karttakeskus Oy, 2001).

936

937 To assess whether PTVs or missense variants may be more geographically clustered than  
938 synonymous variants, we first identified a set of near-independent variants ( $r^2 > 0.02$ ) with  
939  $MAC \geq 3$  and  $MAF \leq 5\%$  among the 14,874 unrelated individuals. For each variant, we  
940 computed the mean pairwise geographical distance between the birthplaces across all pairs  
941 of the available parents of carriers of the minor allele and regressed this mean distance on  
942 variant class (PTVs, missense, or synonymous) and  $MAC$ ,  $MAC^2$ , and  $MAC^3$



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

950

#### 951 *Heritability estimates and genetic correlations*

952 We used genome-wide array genotype data on the 13,326 unrelated individuals for whom  
953 both exome sequence and array data were available to estimate heritability and genetic  
954 correlations for the 64 traits. We constructed a GRM with PLINK<sup>75</sup> (v.1.90b,  
955 <https://www.cog-genomics.org/plink2>) by applying additional filters for  $MAF > 1\%$  and  
956 genotype missingness rate  $< 2\%$  to the set of previously-used genotyped SNVs, leaving  
957 205,149 SNVs for GRM calculation. We used the exact mixed model approach of biMM<sup>76</sup>  
958 (v.1.0.0, <http://www.helsinki.fi/~mjxpirin/download.html>) to estimate the heritability of  
959 our 64 traits and the genetic correlation of the 2,016 trait pairs.

960

#### 961 **Methods References**

- 962 46 Stancáková, A. *et al.* Changes in insulin sensitivity and insulin release in relation  
963 to glycemia and glucose tolerance in 6,414 Finnish men. *Diabetes* **58**, 1212-1221,  
964 doi:10.2337/db08-1607 (2009).
- 965 47 Borodulin, K. *et al.* Cohort Profile: The National FINRISK Study. *Int J*  
966 *Epidemiol*, doi:10.1093/ije/dyx239 (2017).
- 967 48 Wu, J. *et al.* A summary of the effects of antihypertensive medications on  
968 measured blood pressure. *Am J Hypertens* **18**, 935-942,  
969 doi:10.1016/j.amjhyper.2005.01.011 (2005).
- 970 49 Tobin, M. D., Sheehan, N. A., Scurrah, K. J. & Burton, P. R. Adjusting for  
971 treatment effects in studies of quantitative traits: antihypertensive therapy and

972 systolic blood pressure. *Statistics in medicine* **24**, 2911-2935,  
973 doi:10.1002/sim.2165 (2005).

974 50 Liu, D. J. *et al.* Exome-wide association study of plasma lipids in >300,000  
975 individuals. *Nature genetics*, doi:10.1038/ng.3977 (2017).

976 51 Friedewald, W. T., Levy, R. I. & Fredrickson, D. S. Estimation of the  
977 concentration of low-density lipoprotein cholesterol in plasma, without use of the  
978 preparative ultracentrifuge. *Clin Chem* **18**, 499-502 (1972).

979 52 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using  
980 next-generation DNA sequencing data. *Nat. Genet.* **43**, 491-498,  
981 doi:10.1038/ng.806 (2011).

982 53 Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in  
983 sequencing and array-based genotype data. *Am. J. Hum. Genet.* **91**, 839-848,  
984 doi:10.1016/j.ajhg.2012.09.004 (2012).

985 54 Tan, A., Abecasis, G. R. & Kang, H. M. Unified representation of genetic  
986 variants. *Bioinformatics* **31**, 2202-2204, doi:10.1093/bioinformatics/btv112  
987 (2015).

988 55 Davis, J. P. *et al.* Common, low-frequency, and rare genetic variants associated  
989 with lipoprotein subclasses and triglyceride measures in Finnish men from the  
990 METSIM study. *PLoS genetics* **13**, e1007079, doi:10.1371/journal.pgen.1007079  
991 (2017).

992 56 Das, S. *et al.* Next-generation genotype imputation service and methods. *Nature*  
993 *genetics* **48**, 1284-1287, doi:10.1038/ng.3656 (2016).

994 57 McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype  
995 imputation. *Nature genetics* **48**, 1279-1283, doi:10.1038/ng.3643 (2016).

996 58 McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122,  
997 doi:10.1186/s13059-016-0974-4 (2016).

998 59 Adzhubei, I. A. *et al.* A method and server for predicting damaging missense  
999 mutations. *Nature methods* **7**, 248-249, doi:10.1038/nmeth0410-248 (2010).

1000 60 Chun, S. & Fay, J. C. Identification of deleterious mutations within three human  
1001 genomes. *Genome research* **19**, 1553-1561, doi:10.1101/gr.092619.109 (2009).

1002 61 Schwarz, J. M., Cooper, D. N., Schuelke, M. & Seelow, D. MutationTaster2:  
1003 mutation prediction for the deep-sequencing age. *Nature methods* **11**, 361-362,  
1004 doi:10.1038/nmeth.2890 (2014).

1005 62 Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-  
1006 synonymous variants on protein function using the SIFT algorithm. *Nature*  
1007 *protocols* **4**, 1073-1081, doi:10.1038/nprot.2009.86 (2009).

1008 63 Kang, H. M. *et al.* Variance component model to account for sample structure in  
1009 genome-wide association studies. *Nature genetics* **42**, 348-354,  
1010 doi:10.1038/ng.548 (2010).

1011 64 Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide  
1012 association studies, targeted arrays and summary statistics 2019. *Nucleic Acids*  
1013 *Res* **47**, D1005-D1012, doi:10.1093/nar/gky1120 (2019).

1014 65 Kettunen, J. *et al.* Genome-wide study for circulating metabolites identifies 62  
1015 loci and reveals novel systemic effects of LPA. *Nature communications* **7**, 11122,  
1016 doi:10.1038/ncomms11122 (2016).

1017 66 Kettunen, J. *et al.* Genome-wide association study identifies multiple loci  
1018 influencing human serum metabolite levels. *Nature genetics* **44**, 269-276,  
1019 doi:10.1038/ng.1073 (2012).

1020 67 Teslovich, T. M. *et al.* Identification of seven novel loci associated with amino  
1021 acid levels using single-variant and gene-based tests in 8545 Finnish men from  
1022 the METSIM study. *Hum Mol Genet* **27**, 1664-1674, doi:10.1093/hmg/ddy067  
1023 (2018).

1024 68 Inouye, M. *et al.* Novel Loci for metabolic networks and multi-tissue expression  
1025 studies reveal genes for atherosclerosis. *PLoS Genet.* **8**, e1002907,  
1026 doi:10.1371/journal.pgen.1002907 (2012).

1027 69 Lee, S. *et al.* Optimal unified approach for rare-variant association testing with  
1028 application to small-sample case-control whole-exome sequencing studies.  
1029 *American journal of human genetics* **91**, 224-237, doi:10.1016/j.ajhg.2012.06.007  
1030 (2012).

1031 70 Peterson, C. B., Bogomolov, M., Benjamini, Y. & Sabatti, C. Many Phenotypes  
1032 Without Many False Discoveries: Error Controlling Strategies for Multitrait  
1033 Association Studies. *Genet. Epidemiol.* **40**, 45-56, doi:10.1002/gepi.21942 (2016).

1034 71 Loh, P. R. *et al.* Reference-based phasing using the Haplotype Reference  
1035 Consortium panel. *Nature genetics* **48**, 1443-1448, doi:10.1038/ng.3679 (2016).

1036 72 Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype  
1037 imputation method for the next generation of genome-wide association studies.  
1038 *PLoS genetics* **5**, e1000529, doi:10.1371/journal.pgen.1000529 (2009).

1039 73 Manichaikul, A. *et al.* Robust relationship inference in genome-wide association  
1040 studies. *Bioinformatics* **26**, 2867-2873, doi:10.1093/bioinformatics/btq559 (2010).

1041 74 Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population  
1042 structure using dense haplotype data. *PLoS genetics* **8**, e1002453,  
1043 doi:10.1371/journal.pgen.1002453 (2012).

1044 75 Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger  
1045 and richer datasets. *Gigascience* **4**, 7, doi:10.1186/s13742-015-0047-8 (2015).

1046 76 Pirinen, M. *et al.* biMM: efficient estimation of genetic variances and covariances  
1047 for cohorts with high-dimensional phenotype measurements. *Bioinformatics* **33**,  
1048 2405-2407, doi:10.1093/bioinformatics/btx166 (2017).