

## 1 Deep-coverage whole genome sequences and blood lipids among 16,324 individuals

2  
3 Pradeep Natarajan\*(1,2,3), Gina M. Peloso\*(4), S. Maryam Zekavat\*(3,5,6), May Montasser(7),  
4 Andrea Ganna(3,8), Mark Chaffin(3), Amit V. Khera(1,2,3), Wei Zhao(9), Jonathan M.  
5 Bloom(3,8), Jesse M. Engreitz(3,10), Jason Ernst(11), Jeffrey R. O'Connell(7), Sanni E.  
6 Ruotsalainen(12), Maris Alver(13), Ani Manichaikul(14), W. Craig Johnson(15), James A.  
7 Perry(7), Timothy Poterba(3,8), Cotton Seed(3,8), Ida L. Surakka(12), Tonu Esko(13), Samuli  
8 Ripatti(12), Veikko Salomaa(12), Adolfo Correa(16), Ramachandran S. Vasani(17,18,19),  
9 Manolis Kellis (3,20), Benjamin M. Neale(1,2,3,8), Eric S. Lander(3), Goncalo Abecasis(21),  
10 Braxton Mitchell(7), Stephen S. Rich(14), James G. Wilson(16,22), L. Adrienne Cupples(4,19),  
11 Jerome I. Rotter(23), NHLBI TOPMed Lipids Working Group, Cristen J. Willer\*(24), Sekar  
12 Kathiresan\*(1,2,3)

13  
14 1 Center for Genomic Medicine and Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA  
15 02114  
16 2 Department of Medicine, Harvard Medical School, Boston, MA 02115  
17 3 Broad Institute of Harvard & MIT, Cambridge, MA 02142  
18 4 Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118  
19 5 Yale School of Medicine, New Haven, CT 06510  
20 6 Department of Computational Biology & Bioinformatics, Yale University, New Haven, CT 06520  
21 7 School of Medicine, University of Maryland, Baltimore, MD 21201  
22 8 Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114  
23 9 Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI 48109  
24 10 Society of Fellows, Harvard University, Cambridge, MA 02138  
25 11 Department of Biological Chemistry, University of California, Los Angeles, Los Angeles, CA 90095  
26 12 Institute for Molecular Medicine Finland, Helsinki Finland  
27 13 Estonian Genome Center, University of Tartu, Tartu, Estonia  
28 14 School of Medicine, University of Virginia, Charlottesville, VA 22908  
29 15 Department of Biostatistics, University of Washington, Seattle, WA 98195  
30 16 Department of Medicine, University of Mississippi Medical Center, Jackson, MS 39216  
31 17 Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS 39216  
32 18 Department of Epidemiology, Boston University School of Public Health, Boston, MA 02118  
33 19 Framingham Heart Study, Framingham, MA 01702  
34 20 Computer Science and Artificial Intelligence Lab (CSAIL), Massachusetts Institute of Technology, Cambridge,  
35 MA 02139  
36 21 Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109  
37 22 Sections of Preventive Medicine and Epidemiology and Cardiology, Department of Medicine, Boston University  
38 School of Medicine, Boston, MA 02118  
39 23 Institute for Translational Genomics and Population Sciences, LABioMed and Departments of Pediatrics and  
40 Medicine, Harbor-UCLA Medical Center, Torrance, CA 90502  
41 24 Departments of Human Genetics, Internal Medicine, and Computational Medicine and Bioinformatics,  
42 University of Michigan, Ann Arbor, MI 48109

43  
44 Summary Paragraph: 203 words  
45 Main Text: 2,864  
46 References: 44

47           **Deep-coverage whole genome sequencing at the population level is now feasible and**  
48 **offers potential advantages for locus discovery, particularly in the analysis rare mutations**  
49 **in non-coding regions. Here, we performed whole genome sequencing in 16,324**  
50 **participants from four ancestries at mean depth >29X and analyzed correlations of**  
51 **genotypes with four quantitative traits – plasma levels of total cholesterol, low-density**  
52 **lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol, and triglycerides. We**  
53 **conducted a discovery analysis including common or rare variants in coding as well as non-**  
54 **coding regions and developed a framework to interpret genome sequence for dyslipidemia**  
55 **risk. Common variant association yielded loci previously described with the exception of a**  
56 **few variants not captured earlier by arrays or imputation. In coding sequence, rare variant**  
57 **association yielded known Mendelian dyslipidemia genes and, in non-coding sequence, we**  
58 **detected no rare variant association signals after application of four approaches to**  
59 **aggregate variants in non-coding regions. We developed a new, genome-wide polygenic**  
60 **score for LDL-C and observed that a high polygenic score conferred similar effect size to a**  
61 **monogenic mutation (~30 mg/dl higher LDL-C for each); however, among those with**  
62 **extremely high LDL-C, a high polygenic score was considerably more prevalent than a**  
63 **monogenic mutation (23% versus 2% of participants, respectively).**

64 Plasma lipids, including total cholesterol, low-density lipoprotein cholesterol (LDL-C),  
65 high-density lipoprotein cholesterol (HDL-C), and triglycerides, are heritable risk factors for  
66 atherosclerotic cardiovascular disease.<sup>1,2</sup> Understanding the inherited basis for plasma lipid  
67 levels has led to new treatments and to tests to identify individuals at risk for disease. Advances  
68 in technologies to characterize DNA sequence variants (i.e., Sanger sequencing, genotyping  
69 arrays, exome sequencing) have progressively allowed us to solve monogenic forms of  
70 dyslipidemia and to uncover common DNA sequence variants as well as rare mutations that  
71 contribute to plasma lipid levels in the population. However, due to the inherent limitations of  
72 genotyping arrays and exome sequencing, the non-coding regions of the genome remains  
73 incompletely characterized, particularly for rare mutations. In addition, the relative contributions  
74 of common DNA sequence variants and rare coding mutations to extreme lipid values has not  
75 been delineated.

76 It is now possible to directly enumerate the whole genome sequences (WGS) of large  
77 number of individuals. When performed at sufficient depth of coverage (>20-fold coverage per  
78 base), WGS can detect single nucleotide polymorphisms, insertions, and deletions across the  
79 allele frequency spectrum in both non-coding and coding regions. These advances allow us to  
80 test the incremental value of WGS as a tool for locus discovery and also develop a framework to  
81 understand why a specific individual might have an extreme lipid value. Towards these two  
82 goals, we studied whole genome sequences in 16,324 participants of European, African, East  
83 Asian, and Hispanic ancestries with available plasma lipids phenotypes.

84

85 **Deep-coverage whole genome sequencing of 16,324 participants**

86

87 Participants of the Framingham Heart Study (FHS), Old Order Amish (OOA), Jackson Heart  
88 Study (JHS), Multi-Ethnic Study of Atherosclerosis (MESA), FINRISK Study (FIN), and  
89 Estonian Biobank (EST) underwent WGS (**Fig. 1**). Following quality control (**Supplementary**  
90 **Table 1**), 16,324 with plasma lipids available were included in analysis (**Supplementary Table**  
91 **2**). The mean (standard deviation) age was 51 (15) years and 8,669 (53%) were women. 5,911  
92 (36%) of participants were of non-European ancestry (**Supplementary Table 2, Supplementary**  
93 **Fig. 1A-C**). The proportion of individuals on lipid-lowering medications was low (9%).

94 WGS target coverage was >30X for FHS, OOA, JHS, and MESA (as a part of the  
95 NIH/NHLBI Trans-Omics for Precision Medicine (TOPMed) research program) and was >20X  
96 for EST and FIN (**Supplementary Fig. 2**). The mean (standard deviation, SD) attained coverage  
97 for >30X target samples was 37.1(5.4)X and for >20X target was 29.8(5.4)X.

98 After performing quality control, a total of 189 million unique variants were discovered  
99 across all datasets. Total variant count characteristics varied by cohort due to sample sizes,  
100 relatedness, ethnicity, and population history (**Fig. 2**). As expected, the MESA cohort, of largely  
101 unrelated individuals of four diverse ethnicities, had the most variants per individual while the  
102 OOA cohort, a founder population of European ancestry, had the fewest variants per individual  
103 (**Supplementary Table 3**). The median number of variants per individual was 3,391,000, of  
104 which on average 4,878 were observed in only a single individual.

105

## 106 **Common variant association study**

107 We first analyzed common variants, i.e., those that occur often enough that it is practical to test  
108 each variant individually. We considered variants that had MAF > 0.1% within at least one of the  
109 three WGS variant callsets (minor allele count >16 for a callset that included FHS/OOA/JHS, >9

110 for MESA, and >6 for FIN and EST) (**Fig. 1**). We associated such variants with each of the four  
111 plasma lipids within each callset and performed inverse-variance weighted meta-analysis.  
112 Overall, 32,086,348 variants were included in this analysis. Systemic inflation was not observed  
113 (**Supplementary Table 4** and **Supplementary Fig. 3A-D**). We used a conventional statistical  
114 threshold for genome-wide significance ( $\alpha = 5 \times 10^{-8}$ )<sup>3</sup> (**Supplementary Fig. 3E-H**). We  
115 observed 592, 697, 447, and 522 variants associated with total cholesterol, LDL-C, HDL-C, and  
116 triglycerides, respectively (**Supplementary Table 5**).  
117 These variants were distributed at 10, 7, 13, and 9 loci previously associated with total  
118 cholesterol, LDL-C, HDL-C, and triglycerides, respectively, and 5 at putative novel lipid loci  
119 (**Supplementary Table 6**).<sup>4-7</sup> Of the variants at known loci, 12 (38.7%) were lead variants in  
120 prior associations, 8 (25.8%) new lead variants were in high linkage disequilibrium (LD) ( $r^2 >$   
121 0.8) with prior lead variants, and the remaining 11 (35.5%) new lead variants were in low LD  
122 with prior lead variants.

123 At a conventional alpha threshold of  $5 \times 10^{-8}$ , we discovered 5 associations at putative  
124 novel lipid loci (**Supplementary Table 6**). For example, rs3215707 (MAF 2.0%), a 1-bp  
125 deletion at 9p24.1, was associated with HDL-C (+3.3 mg/dL,  $P = 1.3 \times 10^{-8}$ ). rs3215707 occurs  
126 within an intron of *PLGRKT* and overlies active promoter and strong enhancer histone  
127 modification signals for hepG2 cells (**Supplementary Fig. 4**). The deletion is not in LD with  
128 SNPs and thus the association was not detectable by prior genome-wide association analyses.  
129 Within each callset, estimated effects were consistent (heterogeneity  $P = 0.53$ ) and all  
130 demonstrated at least nominal association ( $P < 0.05$ ) (**Supplementary Table 7**).

131 We performed iterative conditional analyses to identify distinct independent associations  
132 among 16 loci jointly reaching  $P < 5 \times 10^{-8}$  for LDL-C, HDL-C, and triglycerides. While only 4

133 (25%) loci displayed evidence of allelic heterogeneity at  $P < 5 \times 10^{-8}$ , 13 (81.3%) had at least  
134 moderate evidence ( $P < 1 \times 10^{-4}$ ) of allelic heterogeneity (**Supplementary Table 8**). Through  
135 conditional analyses for LDL-C, we identified a low-frequency haplotype specific to African  
136 Americans (MAF 0.1% FHS, 0% OOA, 1.0% JHS,) including variants in LD ( $r^2 > 0.8$ ) at a  
137 transcriptional transition region within the first intron of *LDLR* (rs17242843), *LDLR* promoter  
138 (rs17249141), and enhancer 4kb upstream from the *LDLR* transcription start site (rs114197570)  
139 (**Supplementary Fig. 5, Supplementary Fig. 6**). Presence of these variants resulted in a 28  
140 mg/dL lowering of LDL cholesterol ( $P = 2 \times 10^{-11}$ ), indicating a gain-of-function effect on *LDLR*  
141 (**Supplementary Fig. 7**).

142

#### 143 **Rare variant association study of coding variants**

144 Rare variants occur too infrequently to allow association tests of individual variants and thus,  
145 require aggregating rare variants into sets and testing quantitative trait distribution among  
146 carriers of a set versus non-carriers.<sup>8</sup> We aggregated coding sequence variants within each gene  
147 that were predicted to lead to loss of function (e.g. nonsense, canonical splice-site, or frameshift)  
148 or annotated as “disruptive” by the ensemble MetaSVM<sup>9</sup> *in silico* approach. The median  
149 [interquartile range] combined MAF per gene was 0.25% [0.090-0.69%] (**Supplementary Fig.**  
150 **8**). To account for known bidirectional effects of disruptive mutations in some Mendelian  
151 dyslipidemia genes, we accordingly used a mixed model sequence kernel association test  
152 (SKAT).<sup>10,11</sup> Five genes associated with lipids at an exome-wide level ( $\alpha = 0.05$  / ~20,000  
153 protein-coding genes =  $2.5 \times 10^{-6}$ ) (*LDLR*, *PCSK9*, and *APOE* for LDL-C, *LCAT* for HDL-C,  
154 and *APOC3* for triglycerides). Each has been previously established as a cause of Mendelian  
155 forms of dyslipidemia (**Supplementary Table 9**).

156

### 157 **Rare variant association study of non-coding variants**

158 Next, we sought to determine whether rare variants in non-coding regions associate with plasma  
159 lipids. We used four approaches to aggregate rare, non-coding variants. (**Fig. 3**). First, we  
160 aggregated variants within “sliding windows” of 3kb in length.<sup>12,13</sup> Second, we connected a non-  
161 coding variant to a gene if it resided in a segment annotated as an enhancer (and within 20kb of a  
162 gene) or a region annotated as a promoter (and within 5kb of the transcription start site of a gene,  
163 TSS). Third, using gene expression information we connected a non-coding variant to a gene if it  
164 resided in a region annotated as an enhancer. Finally, we connected a non-coding variant to a  
165 gene based on a model which predicted gene-enhancer pairs using a chromatin-state model that  
166 we previously developed.<sup>14</sup> Regulatory annotations were derived from the ENCODE and NIH  
167 Roadmap projects for two cell types - HepG2 and adipose nuclei - relevant to lipoprotein  
168 metabolism. For these analyses, we consider a  $P < 0.05 / 254,032$  groups =  $2.0 \times 10^{-7}$  as significant  
169 (**Supplementary Table 10, Supplementary Table 11**).

170 Using the sliding window approach, we observed suggestive associations for 3kb  
171 windows at the *CETP* (start chr16:56667000) locus (minimum  $P = 4 \times 10^{-6}$ ) and at the *APOA1-*  
172 *APOC3* (start chr11:117094500) locus (minimum  $P = 8 \times 10^{-6}$ ) with HDL-C. 17.6% of non-  
173 coding sliding windows occurring within 1Mb of known lead lipid variants were at least  
174 nominally ( $P < 0.05$ ) associated with lipids versus 4.4% in other regions of the genome across all  
175 traits ( $P$  difference =  $8 \times 10^{-272}$ ).

176 An aggregation of rare non-coding variants at only two genes - *LDLR* and *APOE* -  
177 associated with LDL-C and total cholesterol ( $P < 2 \times 10^{-7}$ ) (**Supplementary Fig. 9**)  
178 (**Supplementary Table 12**). The strongest *LDLR* signal ( $P = 9.7 \times 10^{-11}$ ) was seen for an analysis

179 that connected enhancers and promoters to a gene based on physical proximity (approach #2  
180 above). Closer inspection of the specific variants shows that this signal is driven by the low-  
181 frequency haplotype specific to African Americans that was described earlier (**Supplementary**  
182 **Fig. 10**) (**Supplementary Table 12**). The strongest *APOE* signal ( $P = 8.1 \times 10^{-26}$ ) was observed in  
183 the model connecting enhancers to a gene by gene expression (approach #3 above). Accounting  
184 for the strongest common variant association at the locus (rs7412, the *APOE*  $\epsilon 2$  isoform allele),  
185 this signal attenuates to non-significance ( $P = 1.8 \times 10^{-2}$ ), suggesting that the non-coding variants  
186 are driven by the *APOE*  $\epsilon 2$  isoform. Beyond these two results, we found no additional signals for  
187 a burden of non-coding variants.

188

### 189 **Contribution of monogenic and polygenic models to extreme LDL-C values**

190 With the availability of sequence in both coding and non-coding regions in the same samples, we  
191 estimated the simultaneous contribution of monogenic and polygenic determinants to extreme  
192 LDL-C in individuals of European (EA) and African (AA) ancestry. We defined ‘extreme’ as the  
193 top or bottom 5<sup>th</sup> ancestry-specific percentile of LDL-C. Analyses were conducted in FHS and  
194 MESA-EA subjects (extremes defined as LDL-C >183 mg/dl or LDL-C <72.9 mg/dl) and JHS  
195 and MESA-AA subjects (extremes defined as LDL-C >198.6 mg/dl or LDL-C <71 mg/dl),  
196 separately.

197 Among participants with extremely high LDL-C, we searched for mutations in any of six  
198 Mendelian genes previously linked to high LDL-C (*LDLR*, *APOB*, *PCSK9*, *ABCG5*, *ABCG8*, and  
199 *LDLRAP1*) (**Supplementary Table 13**).

200 To determine polygenic contribution, we implemented a systematic approach to derive,  
201 test, and validate a new ‘genome-wide’ polygenic score for LDL-C using mutually independent



202 datasets. A polygenic score provides a quantitative assessment of the cumulative risk associated  
203 with multiple common risk alleles for each individual. Scores for each individual participant are  
204 created by adding up the number of risk alleles at each variant and then multiplying the sum by  
205 the literature-based effect size.

206 We derived a genome-wide polygenic score based on the association statistics of all  
207 available common ( $MAF \geq 0.01$ ) SNPs with LDL-C, as determined by our previously published  
208 genome-wide association study.<sup>7</sup> The correlation between the variants was assessed using the  
209 European reference population from the 1000 Genomes study.<sup>17</sup> The LDpred computational  
210 algorithm was then used to construct genome-wide polygenic scores.<sup>15</sup> This Bayesian approach  
211 calculates a posterior mean effect size for each variant based on a prior (association with LDL-C  
212 in a previously published study) and subsequent shrinkage based on the extent to which this  
213 variant is correlated with similarly associated variants in a reference population. The underlying  
214 Gaussian distribution additionally considers the fraction of causal (e.g. non-zero effect sizes)  
215 markers. Because this fraction is unknown for any given disease, LDpred uses a range of  
216 plausible values to construct different polygenic scores. We also applied various  $r^2$  and p-value  
217 thresholds to the previously published results. As a comparison to the expanded polygenic  
218 scores, we generated an additional polygenic score restricted to lead variants ( $P < 5 \times 10^{-8}$ ) at  
219 distinct genomic loci, weighted by discovery beta (restricted score). The best score was  
220 determined based on maximal model fit ( $R^2$ ) from a linear regression models in a health-care  
221 biobank of 25,534 unrelated individuals (Nord-Trøndelag Health Study, HUNT)<sup>16</sup>  
222 **(Supplementary Table 14).**

223 For LDL-C, a genome-wide polygenic score incorporating 2 million single nucleotide  
224 polymorphisms with LDpred provided the best model fit **(Supplementary Table 15)**. We

225 applied this polygenic score separately within the WGS samples in FHS, JHS, and MESA. We  
226 labelled individuals as high polygenic score if they fell in the top 5<sup>th</sup> percentile of race-specific  
227 scores (**Table 1**).

228 Among EA participants, a monogenic mutation was associated with an odds ratio of  
229 10.92 (95% CI 3.71-32.14) for extremely high LDL-C whereas a high polygenic score associated  
230 with an odds ratio of 7.65 (95% CI 5.56-10.52). In EA individuals, those who carried a  
231 monogenic mutation had 30 mg/dl higher LDL-C (when compared with non-carriers;  $P = 2.1 \times 10^{-4}$ )  
232 and those who had a high polygenic score had 33 mg/dl greater LDL-C (when compared with  
233 all others;  $P = 1.7 \times 10^{-57}$ ). Of 287 EA participants with extremely high LDL-C, 2% carried a  
234 monogenic mutation and 23% had a high polygenic score.

235 Among AA participants, a monogenic mutation was associated with an odds ratio of 7.43  
236 (95% CI 3.01-18.35) for extremely high LDL-C whereas a high polygenic score associated with  
237 an odds ratio of 3.2 (95% CI 2.1-4.89). In AA individuals, those who carried a monogenic  
238 mutation had 41 mg/dl higher LDL-C (when compared with non-carriers;  $p = 2.3 \times 10^{-7}$ ) and those  
239 who had a high polygenic score had 17 mg/dl greater LDL-C (when compared with all others;  $P$   
240  $= 6.4 \times 10^{-10}$ ). Of 217 AA participants with extremely high LDL-C, 3% carried a monogenic  
241 mutation and 13% had a high polygenic score.

242 We replicated the association between a high polygenic score and extremely high LDL-C  
243 in an independent sample, the ARIC cohort. Among ARIC-EA individuals, a high polygenic  
244 score was associated with an odds ratio of 7.35 (95% CI 5.95-9.10;  $P < 2 \times 10^{-16}$ ) for extremely  
245 high LDL-C and 42.8 mg/dl (95% CI 40.0-47.5;  $P < 2 \times 10^{-16}$ ) higher LDL-C compared with  
246 individuals without a high polygenic score. Among ARIC-AA participants, a high polygenic  
247 score was associated with an odds ratio of 2.7 (95% CI 1.77-4.09;  $P < 3.3 \times 10^{-6}$ ) for extremely

248 high LDL-C and a 23.2 mg/dl (95% CI 15.0-31.5;  $P = 3.8 \times 10^{-8}$ ) higher LDL-C compared with  
249 individuals without a high polygenic score.

250 We analyzed the monogenic and polygenic contribution to extremely low LDL-C in EA  
251 and AA participants and found similar patterns where monogenic mutations as well as a  
252 polygenic score conferred similar effect sizes (**Table 1**).

253

## 254 **Discussion**

255 Here, we performed whole genome sequencing in 16,342 ethnically diverse individuals  
256 and analyzed the incremental value of WGS for locus discovery and for clinical interpretation.  
257 We replicated associations for 28 common variant loci previously associated with lipids in much  
258 larger genome-wide association analyses. We identified an association for a low frequency 1-bp  
259 deletion at 9p24.1 with HDL-C. While we replicated burden associations of rare coding  
260 mutations at known Mendelian lipid genes, we did not detect any burden associations of rare  
261 non-coding mutations through four different approaches. Lastly, we developed a new genome-  
262 wide polygenic score and showed that such a score confers an effect size on LDL-C similar to  
263 carrying a monogenic mutation, but is also notable for a much greater frequency.

264 These results permit several conclusions. Using whole genome sequencing as a discovery  
265 tool, the incremental yield of new loci was modest. In particular, despite a genome-wide search  
266 using four different aggregation approaches and regulatory annotations from two relevant tissues,  
267 we identified no burden-of-rare-variant signals in non-coding regions. Mutation target size and  
268 natural selection pressure is smaller in non-coding regions when compared with coding regions;  
269 based on these considerations, power calculations have suggested that sample sizes may need to  
270 considerably larger to identify rare variant burden associations in non-coding regions.<sup>8</sup>

271           Of great interest, we observed that the relative contribution of polygenic score to  
272 extremely high LDL-C is considerably greater than monogenic mutations. For example, in EA  
273 individuals, both high polygenic score and a monogenic mutation confer similar effects (~30  
274 mg/dl higher LDL-C) but a high polygenic score is present in 23% of participants with  
275 extremely high LDL-C whereas a monogenic mutation is present in only 2%. In most individuals  
276 who carry diagnosis of familial hypercholesterolemia, no monogenic mutation is identified with  
277 clinical exome sequencing<sup>17,18</sup>; for a large fraction of these ‘mutation-negative’ familial  
278 hypercholesterolemia, high polygenic scores may be operative.

279           Important caveats and limitations should be considered. First, appropriate definitions of  
280 statistical significance for WGS association analyses have not been harmonized in the field. The  
281 convention of  $\alpha=5 \times 10^{-8}$  comes from the assumption of performing 1,000,000 independent  
282 tests. Based on our findings and simulations from others<sup>3</sup>,  $10^{-9}$  may be more appropriate for  
283 analyses across diverse ethnicities to allele frequency 0.1%. Second, power is somewhat  
284 diminished with our rare variant meta-analysis approach to combine *P* values with Fisher’s  
285 method. Given known diverse coding mutations within Mendelian genes with bidirectional  
286 effects and the inability to assume unidirectional effects within the non-coding space, we  
287 employed a SKAT statistical framework. Prior approaches leveraging covariance matrices for  
288 SKAT meta-analysis were computationally inefficient for the dataset and multiple grouping  
289 strategies.<sup>19,20</sup> Thus, our approach is conservative. Third, the polygenic scores described here  
290 were derived from genome-wide association studies performed largely in EA ancestry  
291 participants.<sup>7</sup> Because allele frequencies, linkage disequilibrium patterns, and effect sizes of  
292 common polymorphisms vary by ancestry, the predictive capacity of polygenic score was  
293 attenuated in non-European ancestry individuals.<sup>21</sup> This is an important limitation for the field

294 that requires efforts to characterize common genomic variation influencing complex traits among  
295 non-Europeans.

296 In summary, we present a large-scale WGS analysis of plasma lipids in 16,324 ethnically  
297 diverse participants. Common, non-coding variants and rare, coding variants contribute to  
298 plasma lipid variation; however, association signals for rare, non-coding mutations were not  
299 detectable.

300 TABLES/FIGURES

301 **Fig 1. Schematic of genomic variant discovery and analyses.**

302 Variants were jointly discovered in three distinct sets: 1) FHS, JHS, and OOA; 2) MESA; and 3)  
303 EST and FIN. Cohorts included in analyses are denoted by color-coded icons. Allele frequency  
304 space assessed are indicated for analyses.

305 EST = Estonia, FHS = Framingham Heart Study, FIN = Finland, JHS = Jackson Heart Study,  
306 MESA = Multi-Ethnic Study of Atherosclerosis, OOA = Old Order Amish

307

308 **Fig 2. Deep-coverage whole genome sequencing identifies genomic variation across the**  
309 **allelic spectrum.**

310 Variant counts by allele count/frequency bin within each of the cohorts. Variants were jointly  
311 discovered in three distinct sets: 1) FHS, JHS, and OOA; 2) MESA; and 3) EST and FIN.

312 AC = allele count, EST = Estonia, FHS = Framingham Heart Study, FIN = Finland, JHS =  
313 Jackson Heart Study, MAF = minor allele frequency, MESA = Multi-Ethnic Study of  
314 Atherosclerosis, OOA = Old Order Amish

315

316 **Fig 3. Schematic of non-coding rare variant analyses.**

317 Four grouping schematics of rare non-coding variants (MAF < 1%). The sliding window  
318 approach tiles across the genome at fixed widths, only including variants overlying annotations  
319 consistent with enhancers, promoters, and DHS in non-exonic regions. All other approaches  
320 attempt to map non-coding putative functional genomic regions with discrete genes as the  
321 analytical unit. Overall, they are based on: 1) promoter, enhancer, and DHS annotations near a

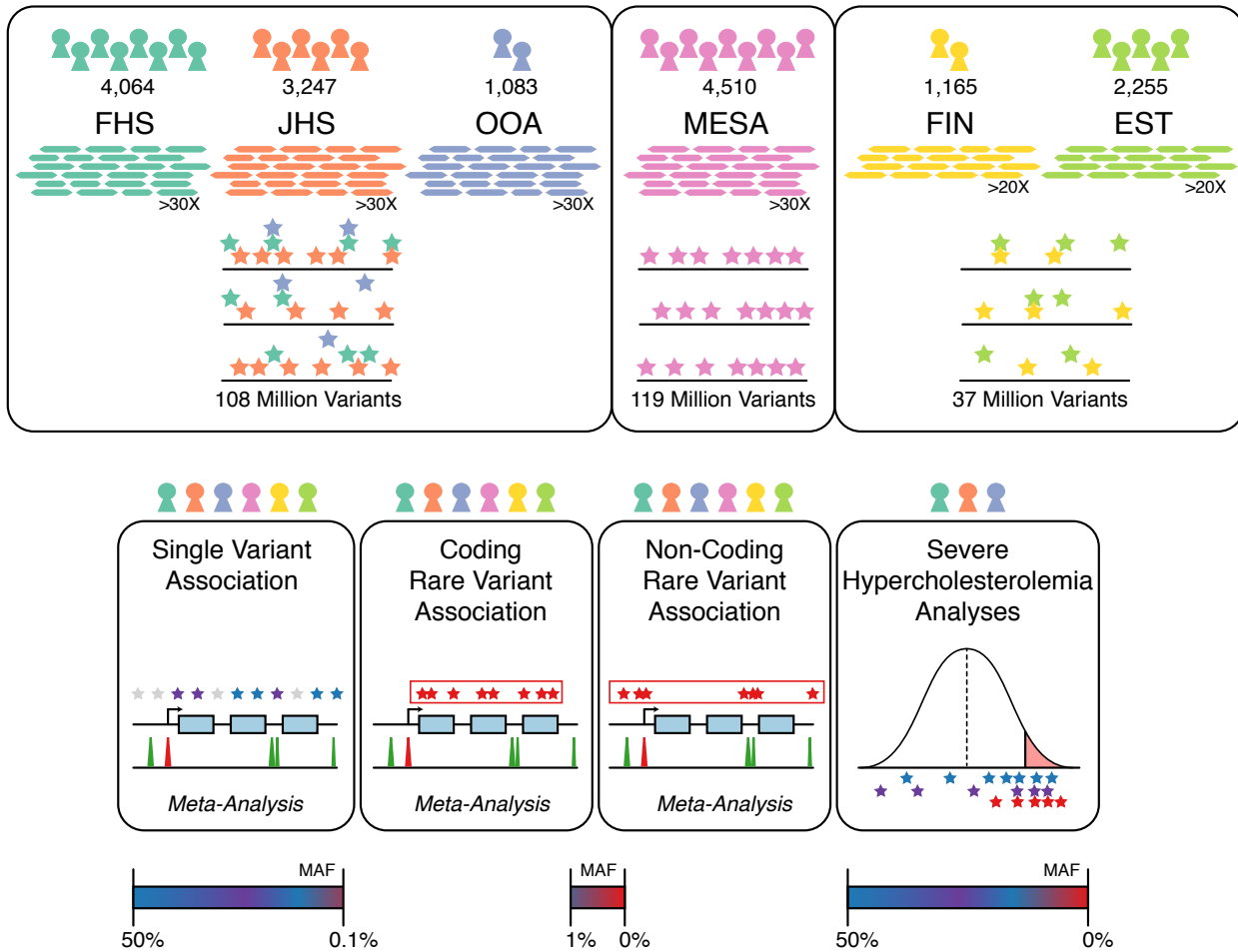
322 gene's transcription start site, 2) co-occurrence of enhancer and DHS annotations with HepG2  
323 gene expression, and 3) H3K27ac marks within Hi-C contact regions mapped to genes.  
324 DHS = DNase hypersensitivity site, MAF = minor allele frequency

325

326 **Table 1. Delineating the monogenic and polygenic contributions to extremely high or low**  
327 **LDL cholesterol concentrations.**

328 (a). Effect of monogenic mutation or polygenic score on odds for extremely high or low LDL-C.  
329 Values are represented as OR [95% CI] for association with given trait. (b). Effect of monogenic  
330 mutation or polygenic score on LDL-C in mg/dl. Values are represented as beta [95% CI] in  
331 mg/dl for LDL-C. Multi-variable associations were performed with sex + age + age<sup>2</sup> (effects not  
332 listed) with monogenic carrier status + high polygenic score using logistic (a) and linear  
333 regression (b). Polygenic risk score was derived from 2 million variants using LDpred. High  
334 polygenic score was defined as membership in the top 5<sup>th</sup> percentile of the ancestry-specific  
335 score distribution. EA, European American; AA, African American.

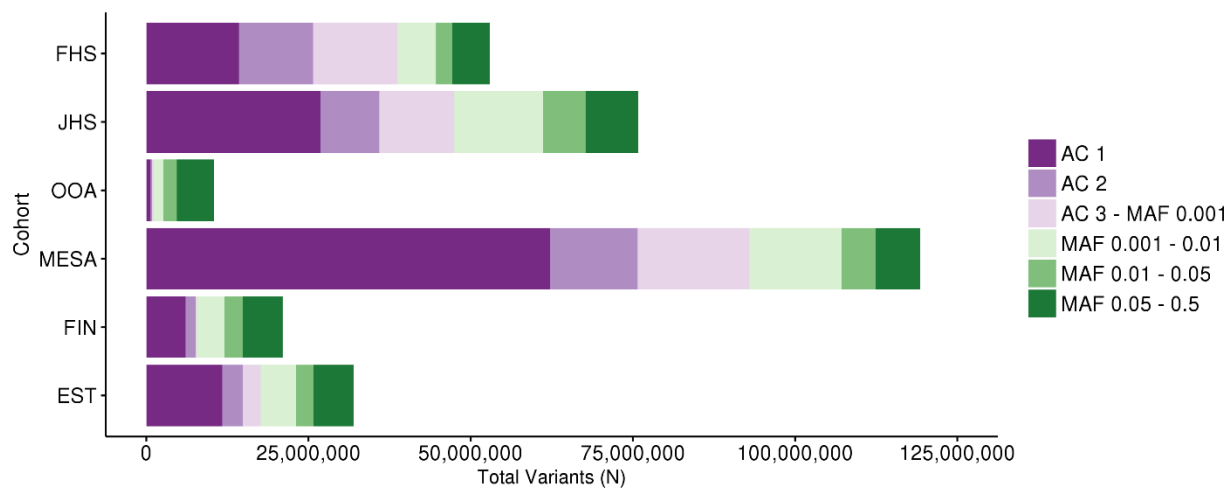
336 **Fig 1. Schematic of genomic variant discovery and analyses.**



337

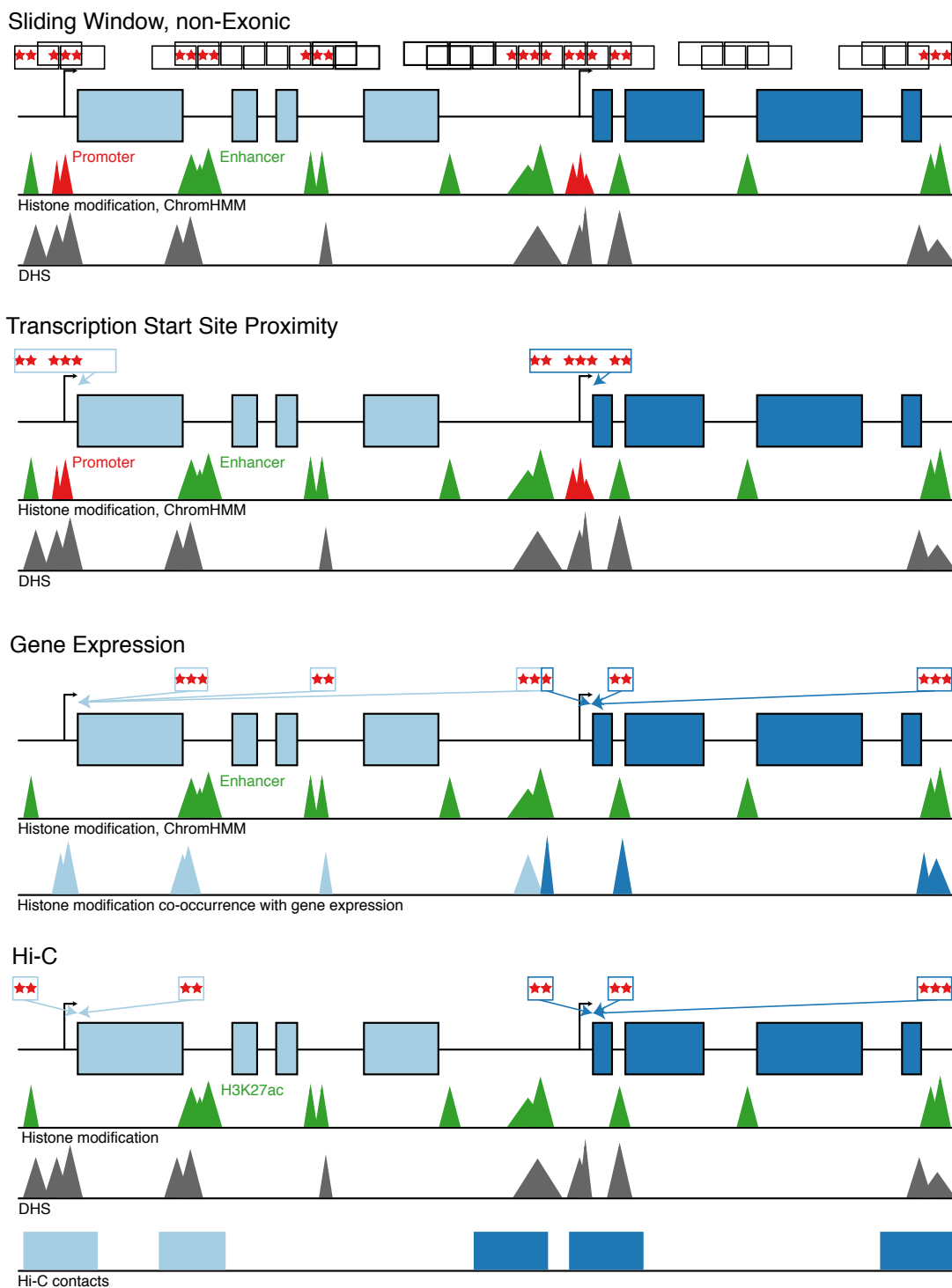


338 **Fig 2. Deep-coverage whole genome sequencing identifies genomic variation across the**  
339 **allelic spectrum.**



340

341 **Fig 3. Schematic of non-coding rare variant analyses.**



342

343 **Table 1. Delineating the monogenic and polygenic contributions to extreme LDL**

344 **cholesterol concentrations.**

345 (a). Effect of monogenic mutation or polygenic score on odds for extremely high or low LDL-C

<b>Extremely High LDL-C</b>										
Ancestry	N <sub>total</sub>	N <sub>extreme</sub>	Monogenic carrier (N <sub>extreme</sub> )	High polygenic score (N <sub>extreme</sub> )	Monogenic carrier OR (95% CI)	Monogenic carrier p-value	Monogenic carrier PAF	High polygenic score OR (95% CI)	High polygenic score p-value	Top 5 <sup>th</sup> percentil of Polygenic score PAF
EA	5910	284	5	64	10.92 (3.71,32.14)	1.4x10 <sup>-5</sup>	1.60	7.65 (5.56,10.52)	5.7x10 <sup>-36</sup>	19.6
AA	4380	217	7	29	7.43 (3.01,18.35)	1.4x10 <sup>-5</sup>	2.79	3.2 (2.1,4.89)	6.7x10 <sup>-8</sup>	9.2
<b>Extremely Low LDL-C</b>										
Ancestry	N <sub>total</sub>	N <sub>extreme</sub>	Monogenic carrier (N <sub>extreme</sub> )	Low polygenic score (N <sub>extreme</sub> )	Monogenic carrier OR (95% CI)	Monogenic carrier p-value	Monogenic carrier PAF	Low polygenic score OR (95% CI)	Low polygenic score p-value	Bottom 5 <sup>th</sup> percentil of Polygenic score PAF
EA	5910	286	6	82	21.73 (6.2,76.15)	1.5x10 <sup>-6</sup>	2.00	10.38 (7.69,14.02)	1.5x10 <sup>-52</sup>	25.9
AA	4380	218	11	32	13.83 (6.25,30.62)	9.4x10 <sup>-11</sup>	4.68	3.7 (2.46,5.58)	3.9x10 <sup>-10</sup>	10.7

346

347 (b). Effect of monogenic mutation or polygenic score on LDL-C in mg/dl

348

**Monogenic mutation or high polygenic score**

Ancestry	N <sub>total</sub>	Monogenic carrier (N)	High Polygenic score (N)	Monogenic carrier Beta mg/dl	Monogenic carrier SE	Monogenic carrier p-value	High Polygenic score Beta mg/dl	High Polygenic score SE	High Polygenic score p-value
EA	5910	18	297	29.98	8.07	2.1x10 <sup>-4</sup>	33.07	2.05	1.7x10 <sup>-57</sup>
AA	4380	25	220	41.05	7.93	2.3x10 <sup>-7</sup>	16.96	2.74	6.4x10 <sup>-10</sup>

**Monogenic mutation or low polygenic score**

Ancestry	N <sub>total</sub>	Monogenic carrier (N)	Low Polygenic score (N)	Monogenic carrier Beta mg/dl	Monogenic carrier SE	Monogenic carrier p-value	Low polygenic score Beta mg/dl	Low polygenic score SE	Low polygenic score p-value
EA	5910	12	297	-47.25	9.55	7.7x10 <sup>-7</sup>	-35.00	2.00	7.9x10 <sup>-67</sup>
AA	4380	28	220	-41.41	7.47	3.1x10 <sup>-8</sup>	-20.41	2.74	1.1x10 <sup>-13</sup>

349

350

351 ACKNOWLEDGEMENTS

352 Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed)  
353 program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for  
354 “NHLBI TOPMed: Whole Genome Sequencing and Related Phenotypes in the Framingham  
355 Heart Study” (phs000974.v1.p1) was performed at the Broad Institute of MIT and Harvard  
356 (HHSN268201500014C). WGS for “NHLBI TOPMed: The Jackson Heart Study”  
357 (phs000964.v1.p1) was performed at the University of Washington Northwest Genomics Center  
358 (HHSN268201100037C). WGS for “NHLBI TOPMed: Genetics of Cardiometabolic Health in  
359 the Amish” (phs000956.v1.p1) was performed at the Broad Institute of MIT and Harvard  
360 (3R01HL121007-01S1). WGS for “NHLBI TOPMed: Multi-Ethnic Study of Atherosclerosis  
361 (MESA)” (phs001416.v1.p1) was performed at the Broad Institute of MIT and Harvard  
362 (3U54HG003067-13S1). Centralized read mapping and genotype calling, along with variant  
363 quality metrics and filtering were provided by the TOPMed Informatics Research Center  
364 (3R01HL-117626-02S1). Phenotype harmonization, data management, sample-identity QC, and  
365 general study coordination, were provided by the TOPMed Data Coordinating Center (3R01HL-  
366 120393-02S1). We gratefully acknowledge the studies and participants who provided biological  
367 samples and data for TOPMed.

368

369 The Framingham Heart Study has been supported by contracts N01-HC-25195 and  
370 HHSN268201500001I and grant R01 HL092577. The Framingham Heart Study thanks the study  
371 participants and the multitude of investigators who over its 70 year history continue to contribute  
372 so much to further our knowledge of heart, lung, blood and sleep disorders and associated traits.  
373 The Jackson Heart Study (JHS) is supported and conducted in collaboration with Jackson State

374 University (HHSN268201300049C and HHSN268201300050C), Tougaloo College  
375 (HHSN268201300048C), and the University of Mississippi Medical Center  
376 (HHSN268201300046C and HHSN268201300047C) contracts from the National Heart, Lung,  
377 and Blood Institute (NHLBI) and the National Institute for Minority Health and Health  
378 Disparities (NIMHD). The authors also wish to thank the staffs and participants of the JHS. Dr.  
379 Wilson is supported by U54GM115428 from the National Institute of General Medical Sciences.

380  
381 This analysis was supported by the American Heart Association 17SDG33680041 (P.N.), the  
382 National, Heart, Lung, and Blood Institute of the US National Institutes of Health grants K01  
383 HL125751 (G.M.P.), R01 HL127564 (C.W., S.K.), and TOPMed analysis support grant (G.M.P.,  
384 P.N.), the Ofer and Shelly Nemirovsky Research Scholar award from Massachusetts General  
385 Hospital (S.K.), and the Donovan Family Foundation (S.K.). The sponsors had no role in the  
386 design and conduct of the study; collection, management, analysis, and interpretation of the data;  
387 preparation, review, or approval of the manuscript; and decision to submit the manuscript for  
388 publication.

389

#### 390 AUTHOR CONTRIBUTIONS

391 Concept and design: P.N., G.M.P., S.M.Z., G.A., J.G.W., L.A.C., C.J.W., S.K.. Acquisition,  
392 analysis, or interpretation of data: P.N., G.M.P., S.M.Z., M.M., A.G., M.C., A.V.K., W.Z., J.B.,  
393 J.R.O., S.E.R., M.A., J.A.P., I.L.S., T.E., S.R., A.C., B.N., G.A., B.M., S.S.R., J.G.W., L.A.C.,  
394 J.I.R., C.J.W., S.K.. Drafting of the manuscript: P.N., G.M.P., S.M.Z., M.M., A.G., S.R., B.M.,  
395 S.S.R., J.G.W., L.A.C., J.I.R., C.J.W., S.K. Administrative, technical, or material support: P.N.,  
396 G.M.P., S.K.

## 397 METHODS

### 398 **Study participants**

399 Please refer to **Supplementary Text** for study participant details.

### 400 **Whole genome sequencing, variant calling, and genotyping**

401 Sequencing was performed at one of four sequencing centers, with all members within a  
402 cohort sequenced at the same center. For the TOPMED phase 1 data, 4,148 FHS individuals and  
403 1,095 OOA individuals were sequenced at the Broad Institute of Harvard and MIT (Cambridge,  
404 MA), while 3,266 JHS individuals were sequenced at University of Washington Northwest  
405 Genomics Center (Seattle, WA). 4,601 MESA individuals were additionally sequenced at the  
406 Broad Institute of Harvard and MIT as part of TOPMED Phase 2. 1,180 Finnish FINRISK  
407 individuals and 2,281 Estonian Biobank participants were sequenced at the Broad Institute of  
408 Harvard and MIT (Cambridge, MA).

409 TOPMED phase 1 BAM files provided by the sequencing centers were harmonized by  
410 the TOPMed Informatics Research Center (IRC) before joint variant discovery and genotype  
411 calling across studies. In brief, sequence data were received from each sequencing center in the  
412 form of bam files mapped to the 1000 Genomes hs37d5 build 37 decoy reference sequence.  
413 Processing was coordinated and managed by the ‘GotCloud’ processing pipeline.<sup>22</sup>

414 The two sequence quality criteria were used in order to pass sequence data on for joint  
415 variant discovery and genotyping are: estimated DNA sample contamination below 3%, and  
416 fraction of the genome covered at least 10x 95% or above. DNA sample contamination was  
417 estimated from the sequencing center read mapping using software verifyBamId.<sup>23</sup>

418 The genotype call sets used for analysis are from “freeze 3a” of the variant calling  
419 pipeline performed by the TOPMed Informatics Research Center (Center for Statistical Genetics,

420 University of Michigan, Hyun Min Kang, Tom Blackwell and Goncalo Abecasis). The software  
421 tools used in this version of the pipeline are available in the following repository:  
422 [https://github.com/statgen/topmed\\_freeze3\\_calling](https://github.com/statgen/topmed_freeze3_calling). Variant detection from each sequenced (and  
423 aligned) genome is performed by vt discover2 software tool.<sup>24</sup> The variant calling software tools  
424 are under active development; updated versions can be accessed at <http://github.com/atks/vt> or  
425 <http://github.com/hyunminkang/apigenome>.

426 WGS for MESA, FINRISK, and the Estonian Biobank was performed using the Illumina  
427 HiSeqX platform at the Broad Institute of Harvard and MIT (Cambridge, MA). DNA samples  
428 are informatically received into the Genomics Platform's Laboratory Information Management  
429 System via a scan of the tube barcodes using a Biosero flatbed scanner. All samples are then  
430 weighed on a BioMicro Lab's XL20 to determine the volume of DNA present in sample tubes.  
431 Following this the samples are quantified in a process that uses PICO-green fluorescent dye.  
432 Once volumes and concentrations are determined the samples are then handed off to the Sample  
433 Retrieval and Storage Team for storage in a locked and monitored -20C<sup>0</sup> walk-in freezer.

434 Libraries were constructed and sequenced on the Illumina HiSeqX with the use of 151-bp  
435 paired-end reads for WGS and output was processed by Picard to generate aligned BAM files (to  
436 hg19).<sup>25,26</sup> Samples were tracked by automated LIMS messaging. Samples were fragmented with  
437 acoustic shearing and libraries were prepared with a KAPA Biosystems kit. Libraries were  
438 normalized to 1.7 nM. Cluster amplication was performed using Illumina cBot and flowcells  
439 were sequenced in HiSeq X. Variants were discovered using the Geome Analysis Toolkit  
440 (GATK) v3 HaplotypeCaller according to Best Practices.<sup>27</sup> Variants from MESA samples were  
441 were generated in one callset. Finland and Estonia samples were jointly called in a separate  
442 callset.

## 443 **Whole genome sequence quality control**

444           The following three approaches were used by the TOPMed Genetic Analysis Center to  
445 identify and resolve sample identity issues: (1) concordance between annotated sex and  
446 biological sex inferred from the WGS data, (2) concordance between prior SNP array genotypes  
447 and WGS-derived genotypes, and (3) comparisons of observed and expected relatedness from  
448 pedigrees.

449           The variant filtering in TOPMed Freeze 3 were performed by (1) first calculating  
450 Mendelian consistency scores using known familial relatedness and duplicates, and (2) training  
451 SVM classifier between the known variant sites (positive labels) and the Mendelian inconsistent  
452 variants (negative labels). Two additional hard filters were applied: (1) Excess heterozygosity  
453 filter (EXHET), if the Hardy-Weinberg disequilibrium p-value was less than  $1 \times 10^{-6}$  in the  
454 direction of excess heterozygosity. An additional  $\sim 3,900$  variants were filtered out by this filter,  
455 and (2) Mendelian discordance filter (DISC), with 3 or more Mendelian inconsistencies or  
456 duplicate discordances observed from the samples. An additional  $\sim 370,000$  variants were filtered  
457 out by this filter. Functional annotation for each variant was provided in the INFO field using  
458 snpEff 4.1 with a GRCh37.75 database.<sup>28</sup> Analysis used hard-call genotypes, without genotype  
459 likelihoods. Genotypes with a depth  $< 10$  were excluded.

460           Additional measures for quality control of TOPMed Phase I Freeze 3 and quality control  
461 for MESA, Finland, and Estonia were performed using the Hail software package.<sup>29</sup> Samples  
462 were filtered by contamination ( $>3.0\%$  for all, except  $>5.0\%$  for Finland and Estonia), chimeras  
463  $>5\%$ , GC dropout  $>4$ , raw coverage ( $<30X$  for all, except  $<19X$  for Finland and Estonia),  
464 indeterminant genotypic sex or genotypic/phenotypic sex mismatch.



465 Variants for MESA, Finland, and Estonia were initially filtered by GATK Variant  
466 Quality Score Recalibration. Additionally, genotypes with  $GQ < 20$ ,  $DP < 10$  or  $> 200$ , and poor  
467 allele balance (homozygous with  $< 0.90$  supportive reads or heterozygous with  $< 0.20$  supportive  
468 reads) were removed. And variants within low complexity regions were removed across all  
469 samples.<sup>30</sup> Variants with  $> 5\%$  missing calls, quality by depth  $< 2$  (SNPs) or  $< 3$  (indels),  
470 InbreedingCoeff  $< -0.3$ , and pHWE  $< 1 \times 10^{-9}$  (within each cohort) were filtered out.

#### 471 **Annotation**

472 Variants were annotated with Hail using annotations from Ensembl's Variant Effect  
473 Predictor<sup>31</sup> for protein-coding annotations and Reg2Map HoneyBadger2-intersect for regulatory  
474 annotations at DNase I regions  $-\log_{10}(P) \geq 10$   
475 ([https://personal.broadinstitute.org/meuleman/reg2map/HoneyBadger2-intersect\\_release/](https://personal.broadinstitute.org/meuleman/reg2map/HoneyBadger2-intersect_release/)).

#### 476 **Traits**

477 Conventionally measured plasma lipids, including total cholesterol, LDL-C, HDL-C, and  
478 triglycerides, were included for analysis. LDL-C was either calculated by the Friedwald equation  
479 when triglycerides were  $< 400$  mg/dl or directly measured. Given the average effect of statins,  
480 when statins were present, total cholesterol was adjusted by dividing by 0.8 and LDL-C by  
481 dividing by 0.7, as previously done.<sup>32</sup> Triglycerides were natural log transformed for analysis.

#### 482 **Single variant association analysis**

483 Single variant analysis was performed in EPACTS with Efficient Mixed-Model  
484 Association eXpedited (EMMAX) for associating each variant site with each lipid outcome  
485 within each jointly called variant call file (VCF).<sup>10</sup> The results were meta-analyzed for each trait  
486 using the METAL software. All analyses were adjusted for age, age<sup>2</sup>, sex, cohort, and an  
487 empirically derived kinship matrix to account for both familial and more distant relatedness. For

488 the TOPMed Phase I VCF, which included OOA, LDL-C and total cholesterol analyses were  
489 also adjusted for *APOB* p.R3527Q and triglycerides and HDL-C analyses were also adjusted for  
490 *APOC3* p.R19Ter. To ensure robust results, we only performed single variant analysis for  
491 variants with a MAF > 0.1%. Variants were meta-analyzed across all three VCFs using  
492 METAL.<sup>33</sup> Summary statistics only for variants with MAF > 0.1% for the given VCF were  
493 included in the meta-analysis. Statistical significance alpha of  $5 \times 10^{-8}$  was used for these  
494 analyses.

495 For loci with at least one variant with  $P < 5 \times 10^{-8}$  within the TOPMed Phase I VCF,  
496 iterative conditional association analysis was performed. Iterative conditioning was performed  
497 until  $P > 1 \times 10^{-4}$  was attained.

#### 498 **Rare variant association analyses**

##### 499 Coding

500 We first identified rare (MAF < 1%) mutations for each VCF within coding sequences.  
501 After Variant Effect Predictor<sup>31</sup> annotation, we identified loss-of-function (e.g. nonsense,  
502 canonical splice-site, and frameshift) and disruptive missense (by MetaSVM<sup>9</sup>).

##### 503 Sliding window

504 We also performed rare variant association tests within the non-coding space  
505 (**Supplementary Figure 7**). As before, we performed a “sliding window” approach aggregating  
506 3kb (overlapping by 1.5kb) windows and considering rare variants occurring within enhancer or  
507 promoter elements at DNase I hypersensitivity sites.

##### 508 Proximity to transcription start sites

509 For non-coding tests, we next attempted to link rare non-coding variants with genes for  
510 association testing using regulatory annotations for HepG2 and adipose nuclei from ENCODE

511 and NIH Roadmap. Given prior observations showing enrichment of functional promoter  
512 variants at *LIPG* with HDL-C extremes<sup>34</sup>, we similarly aggregated variants near transcription  
513 start sites (TSS). Prior studies have shown that approximately 80% of cis-eQTLs fall within  
514 100kb of TSS.<sup>35</sup> To increase likelihood that of mapping regulatory variants to the nearest gene,  
515 we were more restrictive and included variants overlapping promoter sequences +/- 5kb and  
516 enhancer sequences +/- 20kb of TSS at DNase I hypersensitivity sites.

#### 517 Linking enhancers to genes by gene expression

518 We also linked chromatin state defined enhancers with genes using data from the  
519 Roadmap Epigenomics project<sup>36</sup> and the method presented previously<sup>37</sup> with a few small  
520 modifications.<sup>38</sup> The method predicts links using chromatin state information, position of the  
521 enhancer relative to the TSS, and the correlation of multiple chromatin marks with gene  
522 expression across cell types. Here we used the correlation with gene expression of the signal of  
523 five chromatin marks: H3K27ac, H3K9ac, H3K4me1, H3K4me2, and DNaseI hypersensitivity.  
524 The gene expression data was the RPKM expression data for protein coding exons across 56  
525 reference epigenomes from the Roadmap Epigenomics project (available in the file  
526 57epigenomes.RPKM.pc from <http://compbio.mit.edu/roadmap>; Universal Human Reference  
527 was excluded). The chromatin mark signal was the  $-\log_{10}(P)$  tracks averaged to a 200-bp  
528 resolution. As input to our code we used the version of those tracks first averaged at 25-bp  
529 resolution using the ‘Convert’ command of ChromImpute.<sup>39</sup> In computing correlation between a  
530 specific chromatin mark signal and gene expression we used the Pearson correlation and omitted  
531 from the calculation samples lacking both chromatin mark signal and gene expression data. We  
532 made predictions separately for each of the 127 reference epigenomes and locations assigned to  
533 chromatin states, 6\_EnhG, 7\_Enh, and 12\_EnhBiv, of the 15-state core 5-marks ChromHMM

534 model.<sup>36,40</sup> We restricted our predictions to chromatin state assignments on chr1-22 and chrX.  
535 We considered linking 200-bp bins within 1MB of a TSS of each gene as annotated in the file  
536 Ensembl\_v65.Gencode\_v10.ENSG.gene\_info available from <http://compbio.mit.edu/roadmap>.<sup>36</sup>  
537 If a gene had multiple TSS, then we only used the outermost TSS.

538         The method for linking is based on determining for each combination of cell type,  
539 chromatin state, and position relative to the TSS the estimated probability the set of correlations  
540 we observed would come from the actual data compared to randomized data. To this end we  
541 created a training set of actual observed correlations (positive examples) and correlations  
542 computed after randomizing which gene expression values were assigned to which genes  
543 (negative examples) separately for each combination of cell type, chromatin state, and position  
544 relative to the TSS. Each entry in the training set has five features corresponding to correlations  
545 for each of the considered chromatin marks. There is a positive and a corresponding negative  
546 entry for each instance of the specified chromatin state in the specified cell type at the specified  
547 position relative to the TSS or within 5kb of it (for smoothing purposes). We trained a logistic  
548 regression classifier to discriminate actual correlations with randomized correlations. We used  
549 the logistic regression library implemented in the Weka package version 3.7.3 with the  
550 regularization parameter set to 1.<sup>41</sup> For considering linking a specific instance of a chromatin  
551 state assignment in a specific cell type and position relative to the TSS of a gene we applied the  
552 corresponding classifier. Let  $p$  denote the probability the classifier gives of being in the positive  
553 class of the actual observed correlations. We retained those links for which  $p/(1-p)$  was greater  
554 than or equal to 2.5. The method we used here is implemented in the code LinkingRM.java. For  
555 the analyses presented here we used those links for the primary enhancer state, 7\_Enh.  
556 HiC/enhancer links to genes

557 To connect noncoding variants with putative target genes, we predicted functional gene-  
558 enhancer pairs using a chromatin state-based model we previously developed.<sup>14</sup> This model  
559 assumes that the impact of an enhancer on gene expression is determined by the product of its  
560 intrinsic “Activity” (for which we use quantitative DNase-Seq and H3K27ac ChIP-Seq levels as  
561 a proxy) and the “Contact Frequency” at which the enhancer physically encounters its target  
562 promoter in the nucleus (for which we use Hi-C data as a proxy). We previously found such an  
563 Activity by Contact (ABC) model accurately identifies enhancers whose perturbation leads to  
564 changes in gene expression in the human MYC locus<sup>14</sup>, and we have since found that the same  
565 model can identify enhancers across other gene loci and cell types (C. Fulco, E. Lander, and J.  
566 Engreitz, *in preparation*). We extended our previously published model to predict enhancer-gene  
567 connections in the liver, using DNase-Seq and H3K27ac ChIP-Seq data from a hepatocarcinoma  
568 cell line (HepG2) previously generated by the ENCODE project.<sup>42</sup> To define putative regulatory  
569 elements, we expanded DNase-Seq peak calls from ENCODE by 500 bp on either side and  
570 merged overlapping peaks.<sup>14</sup> For each element, we calculated Activity as a function of the  
571 normalized read count of H3K27ac and DNase-Seq. Because high-resolution Hi-C data is not  
572 available for HepG2 cells, we estimated the Contact probability between putative regulatory  
573 elements and genes using the average profile across deeply sequenced Hi-C libraries from 7  
574 different cell types<sup>43</sup> as previously described<sup>14</sup>. For each putative enhancer-gene pair, we  
575 calculated an “ABC score” equal to the Activity  $\times$  Contact of the putative enhancer normalized  
576 by the sum of Activity  $\times$  Contact across all other putative elements within 5 Mb of the target  
577 gene. We tuned free parameters in this model (such as the relative weight of DNase-Seq and  
578 H3K27ac data and a pseudocount to add to Hi-C data) and chose a threshold cutoff using a set of  
579 experimentally measured enhancer-promoter connections in two cell types (C. Fulco, E. Lander,

580 and J. Engreitz, *in preparation*). This analysis defined, for each expressed gene, a set of elements  
581 predicted to regulate that gene in HepG2 cells. These sets of elements were used for gene-level  
582 variant burden tests.

### 583 Statistical test

584 We tested the association of the aggregate of the aforementioned groupings with each  
585 lipid trait using the mixed-model Sequence Kernel Association Test (SKAT) implementation in  
586 EPACTS to account for bidirectional effects.<sup>10</sup> Analyses were adjusted for age, age<sup>2</sup>, sex, cohort,  
587 and empiric kinship. *P* values were meta-analyzed using Fisher's method. Statistical significance  
588 for each gene-based test was  $0.05 / 20,000 \text{ tests} = 2.5 \times 10^{-6}$ .

### 589 **Lipid extremes analysis**

#### 590 Traits

591 We first defined LDL-C extremes as the top and bottom ancestry-specific 5<sup>th</sup> percentiles  
592 from the data (LDL-C > 183 mg/dl or > 198.6 mg/dl for EA and AA, respectively; LDL-C < 72.9  
593 mg/dl or <71 mg/dl for EA and AA, respectively).

#### 594 Monogenic determinants

595 We catalogued mutations in Mendelian genes previously linked to extreme LDL-C  
596 (**Supplementary Table 13**). We included variants that were previously linked to Mendelian  
597 dyslipidemia in ClinVar (“pathogenic” or “likely pathogenic” with no “benign”) or loss-of-  
598 function, and had an allele frequency <1% (autosomal dominant) or <10% (autosomal recessive).  
599 Genotypes were only considered based on expected inheritance pattern (autosomal dominant or  
600 autosomal recessive).

#### 601 Polygenic score derivation

602 We evaluated three distinct approaches to generate weighted polygenic scores using prior  
603 genome-wide association analysis summary statistics<sup>7</sup>: 1) only lead variants at genome-wide  
604 significant loci, 2) varying  $P$  and linkage disequilibrium  $r^2$  thresholds (defined by 1000G CEU)  
605 using PLINK<sup>44</sup>, and 3) all variants but adjusting weights according to  $P$  and  $r^2$  (by 1000G CEU)  
606 with LDpred varying  $\rho$ <sup>15</sup>. To minimize errors from strand flips, A/T and C/G SNPs were  
607 excluded. The scores were calculated as additive sums of risk allele counts for included SNPs  
608 multiplied by weights (discovery effect estimates for 1) and 2), or adjusted by LDpred for 3)).

609 Polygenic scores were generated within the HUNT cohort, the training set.<sup>16</sup> Lipid values  
610 were extracted from the electronic health record; absence of lipid-lowering therapy was  
611 prioritized. For each trait, the model with the best fit, as measured by  $R^2$ , was chosen to applying  
612 to the testing set, TOPMed samples.

### 613 Statistical analysis

614 In a multivariable model, we associated likelihood of membership within the extreme tail  
615 of a trait with monogenic mutation carrier status, high (top 5<sup>th</sup> percentile) or low (bottom 5<sup>th</sup>  
616 percentile) polygenic score, age, age<sup>2</sup>, and sex, separately in European American (EA from FHS  
617 and MESA-EA) and African American (AA from JHS and MESA-AA) samples. We also ran  
618 linear regression models with continuous LDL-C and the independent variables listed above.

619

620

621

622 REFERENCES

623

- 624 1. Emerging Risk Factors, C. *et al.* Major lipids, apolipoproteins, and risk of vascular  
625 disease. *JAMA* **302**, 1993-2000 (2009).
- 626 2. Kathiresan, S. *et al.* A genome-wide association study for blood lipid phenotypes in the  
627 Framingham Heart Study. *BMC Med Genet* **8 Suppl 1**, S17 (2007).
- 628 3. Fadista, J., Manning, A.K., Florez, J.C. & Groop, L. The (in)famous GWAS P-value  
629 threshold revisited and updated for low-frequency variants. *Eur J Hum Genet* **24**, 1202-5  
630 (2016).
- 631 4. Kathiresan, S. *et al.* Common variants at 30 loci contribute to polygenic dyslipidemia.  
632 *Nat Genet* **41**, 56-65 (2009).
- 633 5. Surakka, I. *et al.* The impact of low-frequency and rare variants on lipid levels. *Nat Genet*  
634 **47**, 589-97 (2015).
- 635 6. Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood  
636 lipids. *Nature* **466**, 707-13 (2010).
- 637 7. Willer, C.J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat Genet*  
638 (2013).
- 639 8. Zuk, O. *et al.* Searching for missing heritability: designing rare variant association  
640 studies. *Proc Natl Acad Sci U S A* **111**, E455-64 (2014).
- 641 9. Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for  
642 nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet* **24**, 2125-37  
643 (2015).
- 644 10. Kang, H.M. *et al.* Variance component model to account for sample structure in  
645 genome-wide association studies. *Nat Genet* **42**, 348-54 (2010).
- 646 11. Wu, M.C. *et al.* Rare-variant association testing for sequencing data with the sequence  
647 kernel association test. *Am J Hum Genet* **89**, 82-93 (2011).
- 648 12. Morrison, A.C. *et al.* Practical Approaches for Whole-Genome Sequence Analysis of  
649 Heart- and Blood-Related Traits. *Am J Hum Genet* **100**, 205-215 (2017).
- 650 13. Morrison, A.C. *et al.* Whole-genome sequence-based analysis of high-density lipoprotein  
651 cholesterol. *Nat Genet* **45**, 899-901 (2013).
- 652 14. Fulco, C.P. *et al.* Systematic mapping of functional enhancer-promoter connections with  
653 CRISPR interference. *Science* **354**, 769-773 (2016).
- 654 15. Vilhjalmsson, B.J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic  
655 Risk Scores. *Am J Hum Genet* **97**, 576-92 (2015).
- 656 16. Holmen, O.L. *et al.* Systematic evaluation of coding variation identifies a candidate  
657 causal variant in TM6SF2 influencing total cholesterol and myocardial infarction risk. *Nat*  
658 *Genet* **46**, 345-51 (2014).
- 659 17. Abul-Husn, N.S. *et al.* Genetic identification of familial hypercholesterolemia within a  
660 single U.S. health care system. *Science* **354**(2016).
- 661 18. Khera, A.V. *et al.* Diagnostic Yield and Clinical Utility of Sequencing Familial  
662 Hypercholesterolemia Genes in Patients With Severe Hypercholesterolemia. *J Am Coll*  
663 *Cardiol* **67**, 2578-89 (2016).



- 664 19. Zhan, X., Hu, Y., Li, B., Abecasis, G.R. & Liu, D.J. RVTESTS: an efficient and comprehensive  
665 tool for rare variant association analysis using sequence data. *Bioinformatics* **32**, 1423-6  
666 (2016).
- 667 20. Feng, S., Liu, D., Zhan, X., Wing, M.K. & Abecasis, G.R. RAREMETAL: fast and powerful  
668 meta-analysis for rare variants. *Bioinformatics* **30**, 2828-9 (2014).
- 669 21. Martin, A.R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across  
670 Diverse Populations. *Am J Hum Genet* **100**, 635-649 (2017).
- 671 22. Jun, G., Wing, M.K., Abecasis, G.R. & Kang, H.M. An efficient and scalable analysis  
672 framework for variant extraction and refinement from population-scale DNA sequence  
673 data. *Genome Res* **25**, 918-25 (2015).
- 674 23. Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in  
675 sequencing and array-based genotype data. *Am J Hum Genet* **91**, 839-48 (2012).
- 676 24. Tan, A., Abecasis, G.R. & Kang, H.M. Unified representation of genetic variants.  
677 *Bioinformatics* **31**, 2202-4 (2015).
- 678 25. Li, B. & Leal, S.M. Methods for detecting associations with rare variants for common  
679 diseases: application to analysis of sequence data. *Am J Hum Genet* **83**, 311-21 (2008).
- 680 26. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler  
681 transform. *Bioinformatics* **25**, 1754-60 (2009).
- 682 27. Van der Auwera, G.A. *et al.* From FastQ data to high confidence variant calls: the  
683 Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **11**, 11 10 1-  
684 11 10 33 (2013).
- 685 28. Cingolani, P. *et al.* A program for annotating and predicting the effects of single  
686 nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster*  
687 strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80-92 (2012).
- 688 29. Ganna, A. *et al.* Ultra-rare disruptive and damaging mutations influence educational  
689 attainment in the general population. *Nat Neurosci* **19**, 1563-1565 (2016).
- 690 30. Li, H. Toward better understanding of artifacts in variant calling from high-coverage  
691 samples. *Bioinformatics* **30**, 2843-51 (2014).
- 692 31. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API  
693 and SNP Effect Predictor. *Bioinformatics* **26**, 2069-70 (2010).
- 694 32. Peloso, G.M. *et al.* Association of low-frequency and rare coding-sequence variants with  
695 blood lipids and coronary heart disease in 56,000 whites and blacks. *Am J Hum Genet*  
696 **94**, 223-32 (2014).
- 697 33. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of  
698 genomewide association scans. *Bioinformatics* **26**, 2190-1 (2010).
- 699 34. Khetarpal, S.A. *et al.* Mining the LIPG allelic spectrum reveals the contribution of rare  
700 and common regulatory variants to HDL cholesterol. *PLoS Genet* **7**, e1002393 (2011).
- 701 35. Consortium, G.T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot  
702 analysis: multitissue gene regulation in humans. *Science* **348**, 648-60 (2015).
- 703 36. Zhou, X. *et al.* Epigenomic annotation of genetic variants using the Roadmap Epigenome  
704 Browser. *Nat Biotechnol* **33**, 345-6 (2015).
- 705 37. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell  
706 types. *Nature* **473**, 43-9 (2011).

- 707 38. Liu, Y., Sarkar, A., Kheradpour, P., Ernst, J. & Kellis, M. Evidence of reduced  
708 recombination rate in human regulatory domains. *Genome Biol* **18**, 193 (2017).
- 709 39. Ernst, J. & Kellis, M. Large-scale imputation of epigenomic datasets for systematic  
710 annotation of diverse human tissues. *Nat Biotechnol* **33**, 364-76 (2015).
- 711 40. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and  
712 characterization. *Nat Methods* **9**, 215-6 (2012).
- 713 41. Frank, E., Hall, M., Trigg, L., Holmes, G. & Witten, I.H. Data mining in bioinformatics  
714 using Weka. *Bioinformatics* **20**, 2479-81 (2004).
- 715 42. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome.  
716 *Nature* **489**, 57-74 (2012).
- 717 43. Rao, S.S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles  
718 of chromatin looping. *Cell* **159**, 1665-80 (2014).
- 719 44. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based  
720 linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
- 721