

1 **Widespread signatures of negative selection in the genetic architecture of human complex traits**

2

3 Jian Zeng¹, Ronald de Vlaming^{2,3}, Yang Wu¹, Matthew R Robinson^{1,4}, Luke Lloyd-Jones¹, Loic
4 Yengo¹, Chloe Yap¹, Angli Xue¹, Julia Sidorenko¹, Allan F McRae¹, Joseph E Powell¹, Grant W
5 Montgomery¹, Andres Metspalu⁵, Tonu Esko⁵, Greg Gibson⁶, Naomi R Wray^{1,7}, Peter M
6 Visscher^{1,7}, Jian Yang^{1,7}

7

8 ¹ Institute for Molecular Bioscience, The University of Queensland, Brisbane, Queensland 4072,
9 Australia

10 ² Department of Complex Trait Genetics, VU University, Center for Neurogenomics and
11 Cognitive Research, Amsterdam, 1081 HV, The Netherlands

12 ³ Erasmus University Rotterdam Institute for Behavior and Biology, Rotterdam, 3062 PA, The
13 Netherlands

14 ⁴ Department of Computational Biology, University of Lausanne, 1011 Lausanne, Switzerland

15 ⁵ Estonian Genome Center, University of Tartu, Tartu, Estonia

16 ⁶ School of Biology and Center for Integrative Genomics, Georgia Institute of Technology, Atlanta,
17 GA 30332, USA

18 ⁷ Queensland Brain Institute, The University of Queensland, Brisbane, Queensland 4072,
19 Australia

20 Correspondence: Jian Yang <jian.yang@uq.edu.au>

21

22 **Abstract**

23 Estimation of the joint distribution of effect size and minor allele frequency (MAF) for genetic
24 variants is important for understanding the genetic basis of complex trait variation and can be
25 used to detect signature of natural selection. We develop a Bayesian mixed linear model that
26 simultaneously estimates SNP-based heritability, polygenicity (i.e. the proportion of SNPs with
27 nonzero effects) and the relationship between effect size and MAF for complex traits in
28 conventionally unrelated individuals using genome-wide SNP data. We apply the method to 28
29 complex traits in the UK Biobank data ($N = 126,752$), and show that on average across 28 traits,
30 6% of SNPs have nonzero effects, which in total explain 22% of phenotypic variance. We detect
31 significant ($p < 0.05/28 = 1.8 \times 10^{-3}$) signatures of natural selection for 23 out of 28 traits
32 including reproductive, cardiovascular, and anthropometric traits, as well as educational
33 attainment. We further apply the method to 27,869 gene expression traits ($N = 1,748$), and
34 identify 30 genes that show significant ($p < 2.3 \times 10^{-6}$) evidence of natural selection. All the
35 significant estimates of the relationship between effect size and MAF in either complex traits or
36 gene expression traits are consistent with a model of negative selection, as confirmed by

37 forward simulation. We conclude that natural selection acts pervasively on human complex
38 traits shaping genetic variation in the form of negative selection.

39

40 **Introduction**

41 Dissecting the genetic architecture of complex traits is important for understanding the genetic
42 basis of phenotypic variation and evolution. For a complex trait that influences fitness, natural
43 selection plays an important role in shaping its genetic architecture¹, which in turn provides
44 information to infer the action of natural selection. Given most traits are polygenic, natural
45 selection is likely to act simultaneously on many trait-associated variants that have pleiotropic
46 effects on fitness (known as polygenic selection²⁻⁴). Unlike a selective sweep model⁵ where there
47 are often a limited number of mutations under relatively strong selection, it is difficult to detect
48 the signals of polygenic selection due to the selection pressure being spread over many variants
49 of small effect. However, evidence for natural selection can be inferred from the relationship
50 between effect size and minor allele frequency (MAF) at the genome-wide variants. For example,
51 mutations that are deleterious to fitness are selected against and thus kept at low frequencies
52 by negative selection, resulting in a correlation between effect sizes and MAF^{6,7}. The estimation
53 of the joint distribution of effect size and MAF can be used to detect signature of natural
54 selection and thereby to infer the relationship between a complex trait and fitness.

55

56 Genome-wide association studies (GWAS) have detected thousands of SNPs associated with
57 complex traits, which have helped to characterize the genetic architecture of these traits⁸⁻¹³.
58 However, the genome-wide significant SNPs discovered in GWAS jointly tend to explain only a
59 fraction of the heritability as many SNPs with small effects yet to be detected¹⁴. Furthermore, a
60 proportion is missed due to the incomplete linkage disequilibrium (LD) between causal variants
61 and SNP markers¹⁴. To address the “missing heritability” problem^{14,15} in GWAS, mixed linear
62 model (MLM) approaches have been used to estimate the genetic variance explained by all SNPs
63 used in a GWAS. GREML is a prevailing class of MLM-based approaches where all SNP effects are
64 fitted together as random effects¹⁶. GREML analyses using common SNPs (MAF > 0.01) have
65 uncovered a large proportion of the “missing heritability” for height¹⁷, BMI¹⁷ and psychiatric
66 disorders¹⁸. The GREML method assumes that all SNPs have an effect on the trait¹⁶ and thus
67 does not allow us to estimate the degree of polygenicity (i.e. the proportion of SNPs with
68 nonzero effects). Bayesian multiple regression is another class of MLM-based methods that
69 enable us to make posterior inference about polygenicity by assuming SNP effects are drawn
70 from a mixture distribution of zero and nonzero components¹⁹⁻²¹. Bayesian MLM methods have
71 been widely used in livestock and plant breeding²² and have attracted increasing attention in
72 humans for characterizing the genetic architecture of complex traits and diseases^{20,23,24}.

73 However, neither GREML nor Bayesian MLM approaches explicitly model the relationship
74 between effect size and MAF, an important characteristic of the genetic architecture for complex
75 traits. This relationship can be used to detect signatures of natural selection^{7,25} and inform the
76 design of future genetic mapping studies.

77

78 In this study, we developed an MLM-based Bayesian method that can simultaneously estimate
79 SNP-based heritability, polygenicity and the joint distribution of effect size and MAF in
80 conventionally unrelated individuals using GWAS data. We applied the method to 28 complex
81 traits in the UK Biobank (UKB) data²⁶, and 27,869 gene expression traits in the Consortium for
82 the Architecture of Gene Expression (CAGE) dataset²⁷, and identified a number of complex traits
83 and gene expression traits for which there is significant evidence of natural selection on the
84 associated SNPs.

85

86 **Results**

87 **Method overview**

88 Under the Bayesian MLM framework, we propose to model the relationship between effect size
89 and MAF using the following mixture distribution as prior for each SNP effect

90

$$91 \quad \beta_j \sim N\left(0, [2p_j(1-p_j)]^S \sigma_\beta^2\right)\pi + \phi(1-\pi)$$

92

93 where β_j is the allelic substitution effect of a SNP j , p_j is the MAF of the SNP, σ_β^2 is a constant
94 factor (i.e. variance of SNP effects under a neutral model), ϕ is a point mass at zero, and π is the
95 proportion of SNPs with nonzero effects (polygenicity). The variance of the effect size of SNP j is
96 $\sigma_j^2 = [2p_j(1-p_j)]^S \sigma_\beta^2$, which is a function of MAF of the SNP. Thus, the parameter S measures
97 the relationship between effect size and MAF. If $S = 0$, the effect size is independent of MAF
98 (neutral model). If there is selection, the effect size can be positively ($S > 0$) or negatively ($S < 0$)
99 related to MAF. All these parameters are treated as unknown with appropriate priors (**Online**
100 **Methods**). Our model (referred to as BayesS) allows simultaneous estimation of multiple
101 characteristics of the genetic architecture: SNP-based heritability (h_{SNP}^2), polygenicity (π) and
102 the relationship between SNP effect and MAF (S). We use a gradient-based sampling algorithm,
103 Hamiltonian Monte Carlo²⁸, to sample S from the posterior distribution, and use Gibbs sampling
104 for other parameters in the model by assuming conjugate priors. Furthermore, we use a parallel
105 computing strategy following Fernando *et al.*²⁹ to allow the analysis to be scalable to very large
106 samples sizes ($N > 100,000$). Details of sampling scheme and parallel computing strategies are
107 given in the **Supplementary Note**.

108

109 In the hypothesis test against $S = 0$, we used two approaches to control false positives. The first
110 approach is to control the family-wise type I error rate (FWER) using the theory that the
111 posterior mode standardized by the posterior standard error (s.e.) asymptotically follows a
112 standard normal distribution under the null³⁰. The asymptotic normality of the posterior
113 distribution was justified by simulation with the UKB cohort (**Supplementary Fig. 1**). The
114 second approach is to control the proportion of false positives³¹ (PFP) among rejections (also
115 known as the marginal false discovery rate or mFDR³²) based on the posterior probability given
116 the data, *e.g.* $\Pr(S < 0|\mathcal{D})$ (**Supplementary Note**). We show by simulation that if the true
117 distribution of S is used as the prior, then rejecting $S = 0$ with $\Pr(S < 0|\mathcal{D}) \geq \gamma$ guarantees PFP
118 or mFDR to be less than $1 - \gamma$ (**Supplementary Fig. 2**). The former approach is more stringent
119 but the advantage of the latter approach is that the power is not inversely related to the number
120 of traits (tests)³¹.

121

122 **Assessing the robustness of parameter estimation through simulations based on real** 123 **genotype data**

124 We used simulations based on real GWAS genotype data from the Atherosclerosis Risk in
125 Communities (ARIC) study³³ to assess our method in estimating the parameters $\theta = [S, h_{SNP}^2, \pi]$.
126 The ARIC data consist of 12,942 unrelated individuals and 564,959 Affymetrix SNPs with MAF >
127 1% after quality control (**Online Methods**). In our simulation, 1,000 SNPs were chosen at
128 random to be causal variants, with their effects related to MAF through an S value ranging from
129 -1 to 1 in different scenarios (**Online Methods**). Since the number of causal variants was known,
130 polygenicity was assessed by the number of SNPs with nonzero effects (m_{NZ}). Based on the
131 Markov chain Monte Carlo (MCMC) samples, the point estimate ($\hat{\theta}$), standard error (s.e.) or
132 credible interval for each parameter was given by the mode, standard deviation (s.d.) or highest
133 probability density (HPD) of its posterior distribution, respectively.

134

135 Results (**Fig. 1**) show that when both causal variants and SNP markers were fitted in the
136 analysis, $\hat{\theta}$ from BayesS was unbiased with respect to the true parameters. When the causal
137 variants were not included in the analysis, both \hat{h}_{SNP}^2 and the absolute value of \hat{S} were slightly
138 underestimated, due to imperfect tagging, a similar issue as discussed in Yang *et al.*¹⁶. For
139 polygenicity, however, m_{NZ} estimate tended to be larger than the number of causal variants,
140 probably because some causal variants could be better tagged by multiple SNPs. Thus, in
141 practice, $\hat{\pi}$ should be interpreted as the proportion of non-null SNPs, which is likely to be larger
142 than the proportion of causal variants. Results also show that the s.e. for \hat{S} , \hat{h}_{SNP}^2 and $\hat{\pi}$ is

143 consistent with the s.d. of the estimates from 100 simulation replicates (**Supplementary Table**
144 **1**).

145

146 **Analysis of 28 complex traits in the UK Biobank data**

147 We applied the BayesS method to 36 complex traits on 126,545 unrelated individuals of
148 European ancestry in the UKB²⁶ with 483,634 Affymetrix SNPs (MAF > 1%) after quality
149 controls (**Online Methods**). Out of the 36 traits 21 have $N > 100,000$. Two commonly used
150 long-chain diagnostic tests were adopted to assess the convergence of the MCMC algorithm
151 (**Supplementary Note**). Traits with results that did not pass our convergence tests were those
152 with the smallest sample sizes, \hat{h}_{SNP}^2 close to zero, or both (**Supplementary Fig. 3**). We focus on
153 the results of 28 traits that passed both convergence tests for all of the three genetic
154 architecture parameters. These traits include 24 quantitative traits, 2 diseases: major
155 depressive disorder (MDD) and type 2 diabetes (T2D), and 2 categorical traits: male pattern
156 baldness (MPB) and years of schooling (educational attainment). **Supplementary Fig. 4** shows
157 the distributions of the estimates across these traits for the three genetic architecture
158 parameters.

159

160 **Comparison of the genetic architecture between Height and BMI**

161 The genetic architectures of height and BMI have been relatively well studied compared to
162 other complex traits³⁴⁻⁴⁰. Thus, it is interesting to compare our results for height and BMI (**Fig. 2**)
163 with the previous findings. Both traits have a large sample size in the UKB: $N = 126,545$ for
164 height and $N = 126,389$ for BMI. For both traits, a negative S was detected with extremely high
165 significance level ($P = 1.8 \times 10^{-106}$ for height and $P = 2.8 \times 10^{-14}$ for BMI), meaning that lower-
166 MAF variants tend to have larger effect size (absolute values). These results suggest that both
167 height- and BMI-associated SNPs have been under selection, in line with the conclusions drawn
168 from two recent studies^{36,39}. The posterior mode of S was -0.422 (s.e. = 0.019) for height,
169 remarkably lower than that of -0.295 (s.e. = 0.039) for BMI, suggesting that the proportion of
170 genetic variation attributable to SNPs with low MAF for height is larger than that for BMI,
171 consistent with the result from a previous study³⁶. These results also imply that overall height-
172 associated SNPs are under stronger selection than BMI-associated SNPs. The phenotypic
173 variance explained by common SNPs (MAF > 0.01) was 52.8% (s.e. = 0.3%) for height and 27.7%
174 (s.e. = 0.4%) for BMI, consistent with the estimates of h_{SNP}^2 for height and BMI based on
175 common SNPs reported previously³⁴⁻³⁶. The posterior distribution of π provides an estimate of
176 4.8% (s.e. = 0.1%) of SNPs having nonzero effects on height, significantly lower than that of 9.4%
177 (s.e. = 0.5%) for BMI. These results suggest that BMI is more polygenic but less heritable than
178 height, consistent with the results from a recent study using BayesR, a Bayesian multi-

179 component mixture model⁴¹. As a consequence, a BMI-associated SNP on average would explain
180 a relatively smaller proportion of h_{SNP}^2 , compared with a height-associated SNP, which may
181 explain the higher uncertainty in the estimates of the hyperparameters such as S and π for BMI.
182 These results also explain why the number of genome-wide significant SNPs (m_{GWS}) identified
183 from the GIANT meta-analysis for BMI ($m_{GWS} = 97$) was smaller than that for height ($m_{GWS} =$
184 697) despite the fact that the sample size for BMI ($N \approx 340,000$)³⁵ was considerably larger
185 than that for height ($N \approx 250,000$)³⁴.

186

187 **Inference on natural selection**

188 Of the 28 traits that passed our convergence tests, 23 traits (including reproductive,
189 cardiovascular and anthropometric traits and educational attainment) had significant negative S
190 estimates with $\Pr(S < 0 | \mathcal{D}) = 1$ and $P < 0.05/28$ (**Supplementary Table 2**), providing strong
191 evidence that these traits have been under selection. The estimates of S over traits ranged from
192 -0.601 (age at menopause) to 0.016 (fluid intelligence score) with mean -0.348, median -0.364
193 and s.d. 0.112. Interestingly, all the significant estimates of S were negative (see below for
194 forward simulation to infer the type of selection from the sign of S). The magnitudes of \hat{S} , *i.e.* $|\hat{S}|$,
195 reflects the strength of selection on the trait-associated SNPs. Traits with the largest $|\hat{S}|$ are
196 related to fertility and heart function (**Fig. 3**), including age at menopause ($\hat{S} = -0.601$, s.e. =
197 0.073), pulse rate ($\hat{S} = -0.481$, s.e. = 0.048), waist circumference adjusted for BMI (WCadjBMI,
198 $\hat{S} = -0.436$, s.e. = 0.036) and waist-hip ratio adjusted for BMI (WHRadjBMI, $\hat{S} = -0.436$, s.e. =
199 0.049). It has been reported that WCadjBMI and WHRadjBMI are associated with cardiovascular
200 events⁴², and WHRadjBMI is strongly correlated with pregnancy rate⁴³. Other reproductive and
201 cardiovascular traits, such as age at first live birth, age at menarche and blood pressure, had
202 relatively high $|\hat{S}|$ as well. Thus, our results suggest that reproductive and cardiovascular traits
203 are closely related to fitness and the SNPs that are associated with these traits have been under
204 relatively stronger selection than SNPs associated with other traits.

205

206 Height ($\hat{S} = -0.422$), handgrip strength (right: -0.404, left: -0.374), lung function related traits
207 (-0.405 - -0.362), heel bone mineral density (-0.394) and basal metabolic rate (-0.367) had a
208 moderate to high $|\hat{S}|$ (**Fig. 3** and **Supplementary Table 2**). Evidence of selection for height has
209 been reported from multiple studies using different approaches³⁶⁻⁴⁰. The two diseases, MDD and
210 T2D, had negative \hat{S} but the P -values did not reach FWER significance threshold, although the
211 posterior probability of $S < 0$ for T2D was as high as 0.983. However, the power to detect a
212 significant \hat{S} may not be comparable to those quantitative traits, given the number of cases was
213 less than 10,000 for each. A recent large-scale GWAS based on whole-genome sequencing data

214 also did not detect a signal of selection on T2D-associated variants⁸. Fluid intelligence score is
215 the only trait with \hat{S} at almost zero ($\hat{S} = 0.016$, s.e. = 0.096), which seems to suggest that fluid
216 intelligence (FI) is not pertinent to fitness. However, there is strong evidence of negative
217 selection on the SNPs associated with educational attainment (EA, $\hat{S} = -0.350$, s.e. = 0.055),
218 which is thought to be a proxy of intelligence. Indeed, the genetic correlation between EA and FI
219 was as high as 0.665 (s.e. = 0.052) estimated from a bivariate LD score regression⁴⁴. Thus, it may
220 be due to the limited statistical power that we did not detect the signal of selection for FI.

221
222 For traits with a significant estimate of S , we demonstrated the relationship between effect size
223 and MAF by a plot of the cumulative genetic variance explained by SNPs (V_g) against MAF (**Fig.**
224 **4**), where MCMC samples of SNP effects were used to compute V_g for SNPs with MAF smaller
225 than a threshold on the x-axis (**Supplementary Note**). Under an evolutionarily neutral model,
226 V_g is linearly proportional to MAF⁴⁵ (diagonal line), therefore the area under the curve (AUC) is
227 0.5. All traits with significant estimates of S had the curve of cumulative genetic variance above
228 the diagonal line, with $|\hat{S}|$ highly correlated with the AUC ($r = 0.902$), an alternative way of
229 illustrating the evidence of natural selection.

230

231 **Inference on SNP-based heritability**

232 The 28 traits had low to moderate estimates of h_{SNP}^2 with mean 0.221, median 0.212, and s.d.
233 0.093, and were all significantly above zero (**Supplementary Table 3**). Note that traits with
234 \hat{h}_{SNP}^2 close to zero had failed in MCMC convergence tests, therefore the mean h_{SNP}^2 estimate
235 across traits is likely to be inflated. For MDD ($\hat{h}_{SNP}^2 = 0.111$, s.e. = 0.021) and T2D ($\hat{h}_{SNP}^2 = 0.222$,
236 s.e. = 0.015), the estimates were on the liability scale and were converted from the observed
237 scale¹⁵, assuming a population prevalence of 15%⁴⁶ and 3%⁴⁷, respectively. The sorted
238 estimates across traits are shown in **Supplementary Fig. 5**. Besides height ($\hat{h}_{SNP}^2 = 0.528$),
239 traits with the highest \hat{h}_{SNP}^2 include basal metabolic rate (0.336), which has been reported to be
240 0.2–0.4 in model animals⁴⁸, and MPB (0.335), which has been reported to be a highly heritable
241 trait in both pedigree⁴⁹ and genomic⁵⁰ analysis. Traits with the lowest \hat{h}_{SNP}^2 include mean time
242 to correctly identify matches (0.081), MDD (0.111), birth weight (0.114) and neuroticism score
243 (0.125), in line with the low estimates of h_{SNP}^2 from previous studies in MDD⁵¹ and neuroticism
244 score⁵². Given that most published estimates were obtained using whole-genome imputed SNPs,
245 they are likely to be slightly higher than our estimates that are only based on the SNPs on
246 Affymetrix Axiom Genotyping Arrays. For example, a recent study⁵³ on educational attainment
247 in UKB gave an estimate of 0.21 (s.e. = 0.006), slightly higher than our estimate of 0.182 (s.e. =
248 0.004). Our estimate of 0.528 (s.e. = 0.003) for height is slightly but not significantly lower than

249 that of 0.56 (s.e. = 0.023) in Yang *et al.*³⁶. For BMI, our estimate of 0.277 (s.e. = 0.004) is highly
250 consistent with that of 27% (s.e. = 2.5%) in Yang *et al.*³⁶. Across traits, \hat{h}_{SNP}^2 seems to be
251 independent of either \hat{S} or $\hat{\pi}$ but the s.e. of \hat{S} and $\hat{\pi}$ decrease as \hat{h}_{SNP}^2 increases (**Supplementary**
252 **Fig. 6**).

253

254 **Inference on polygenicity**

255 The distribution of $\hat{\pi}$ had mean 5.9%, median 5.5% and s.d. 3.6% across traits, and ranged from
256 0.6% (s.e. = 0.1%) to 13.6% (s.e. = 1.3%) (**Supplementary Table 4**). This suggests that all the
257 28 complex traits are polygenic with ~30,000 common SNPs with nonzero effects on average.
258 Note that our simulation above suggests that this is likely to be an overestimation of the number
259 of causal variants (**Fig. 1**). Interestingly, age at menopause, the trait with highest magnitude of \hat{S}
260 (-0.601), had the lowest estimate of polygenicity $\hat{\pi}$ (0.6%, s.e. = 0.1%) (**Supplementary Fig. 5**).
261 Educational attainment had the highest $\hat{\pi}$ (13.6%, s.e. = 1.3%), which is reasonable because it is
262 a compound trait of several sub-phenotypes so that many SNPs have an effect. It is followed by
263 age at first live birth ($\hat{\pi}$ = 13.3%, s.e. = 2.5%), body fat percentage ($\hat{\pi}$ = 11.1%, s.e. = 0.8%) and
264 BMI ($\hat{\pi}$ = 9.4%, s.e. = 0.5%). On the contrary, these traits had low to moderate magnitude of \hat{S} .

265

266 **Analysis of gene expression traits in the CAGE data**

267 Analysing expression levels of all probes in the CAGE dataset²⁷ (1,748 unrelated individuals of
268 European ancestry) using the standard BayesS approach would be computationally challenging,
269 as it would require us to perform 36,778 distinct BayesS analyses. However, given that many
270 probes have a very limited number of associated SNPs, we developed a nested version of the
271 BayesS model. This nested approach speeds up the analyses by considering SNPs in proximity
272 collectively as a window, which allows for fast “jumping” over windows with zero effect (**Online**
273 **Methods**). We showed by simulation that the nested model produces similar results as the
274 standard BayesS approach in the analyses of both simulated (**Supplementary Fig. 7**) and UKB
275 data (**Supplementary Fig. 8**) while being six times as fast as the standard BayesS approach for
276 traits with low polygenicity (**Supplementary Fig. 9**). Using the nested BayesS model, we were
277 able to fit 1,066,738 imputed SNPs (MAF > 1% and in common with those on HapMap³⁵⁴) for
278 the gene expression traits by partitioning the genome into 12,937 non-overlapping 200-Kb
279 segments. Thus, the polygenicity (π) is interpreted as the proportion of segments with nonzero
280 effects in nested BayesS.

281

282 After convergence tests, 27,869 probes remained, most of which had low \hat{h}_{SNP}^2 (mean = 0.147,
283 median = 0.122 and s.d. = 0.088) and polygenic architecture (mean π = median π = 5.2% and s.d.
284 = 2.3%) (**Supplementary Fig. 10**). With unrelated individuals only, our \hat{h}_{SNP}^2 were moderately

285 correlated with the GREML estimates ($r = 0.568$, **Supplementary Fig. 11**) despite the relatively
286 small sample size. The estimates of polygenicity $\hat{\pi}$ suggest widespread trans-regulatory effects
287 on gene expression in humans. To identify genes under selection, we mapped 21,303 out of the
288 27,869 probes to the genome with at least “good” probe annotation quality⁵⁵, which tagged
289 15,615 genes. Applying a Bonferroni correction for the number of probes mapped to the
290 genome, we identified 32 probes that had \hat{S} significantly different from zero ($P < 0.05/$
291 $21,303 = 2.3 \times 10^{-6}$; **Fig. 5** and **Supplementary Table 5**). These probes were mapped to 30
292 unique genes (**Fig. 6**) and all had negative \hat{S} (mean = -1.259, s.d. = 0.185), moderate \hat{h}_{SNP}^2 (mean
293 = 0.412, s.d. = 0.075) and small $\hat{\pi}$ (mean = 0.0268, s.d. = 0.011). The alternative approach to
294 control false positives is to limit PFP, which is less stringent but more powerful compared with
295 limiting FWER. With this approach, a number of additional probes were identified with
296 $\Pr(S < 0 | \mathcal{D}) \geq 0.95$, giving a significant set of 266 probes for which 67 probes had
297 $\Pr(S < 0 | \mathcal{D}) = 1$ (**Fig. 5**). After mapping these probes to genes, a total of 252 genes were
298 identified with the proportion of false positives $< 5\%$. The results of gene ontology (GO) over-
299 representation tests showed that these genes were enriched in the molecular function of IgG
300 binding ($P = 0.032$ after Bonferroni correction). Moreover, we detected 45 genes that had
301 $\Pr(S > 0 | \mathcal{D}) \geq 0.95$ (**Fig. 5**), which were enriched in the molecular function of α_1 -adrenergic
302 receptor activity ($P = 0.048$) and potassium channel activity ($P = 0.016$). These results are
303 consistent with a previous review⁵⁶ that a proportion of genes showing evidence of selection
304 were significantly enriched in the function of immunity, receptor and potassium channel
305 activity.

306

307 **The directions of parameter S under different types of natural selection**

308 Besides detecting selection and quantifying its strength on the trait-associated SNPs, the sign of
309 S allows us to further infer the type of selection. To demonstrate this, we used forward
310 simulations (**Online Methods**) to simulate common types of natural selection for a quantitative
311 trait by relating the normally distributed phenotype to fitness through a hypothetical function
312 (**Fig. 7**, top row). In the last generation of selection, the relationship between the variance σ_j^2 in
313 the effect of coded allele and its frequency showed different patterns across different types of
314 selection (**Fig. 7**, bottom row). As expected, when all the variants were selectively neutral, σ_j^2
315 was uniformly distributed across MAF ($S = 0$). Under stabilizing selection, σ_j^2 was negatively
316 related to MAF ($S < 0$), a result of purifying trait-associated variants with large effect size which
317 was deleterious to fitness through pleiotropy (also known as negative selection). Both
318 directional (in either direction) and disruptive selection led to a positive relationship between
319 σ_j^2 and MAF ($S > 0$). This is because in both cases, alleles with favourable effects increased in

320 frequency due to positive selection, so that high MAF bins were enriched with derived alleles of
321 large effect. The difference is that disruptive selection kept the alleles with large effects at the
322 intermediate frequencies, while directional selection persistently drove them toward fixation,
323 resulting in a sigmoidal or convex shape of the relationship between σ_j^2 and MAF
324 (**Supplementary Fig. 12**). In conclusion, estimate of S is informative to detect the signature of
325 natural selection and is able to distinguish stabilising selection from directional and disruptive
326 selection for a trait. At the level of genetic variants, a negative (positive) value of S is indicative
327 of negative (positive) selection on the variants associated with the trait.

328

329 **Discussion**

330 We infer the action of natural selection on a complex trait from the signature left in the genetic
331 architecture – the relationship between effect size and MAF. We introduced a method to
332 simultaneously estimate the SNP-based heritability, polygenicity and the relationship between
333 effect size and MAF using all genome-wide SNPs. In contrast to the contemporary methods that
334 use independent SNPs that are significantly associated with traits^{3,37,57,58}, our method accounts
335 for genome-wide SNP effects jointly and therefore has more statistical power to detect the
336 signature of selection for polygenic traits. Results of the simulations using real genotypes
337 showed that our estimate of the relationship (S) is unbiased when the causal variants are
338 observed; otherwise, the estimate tends to be conservative depending on the LD between SNPs
339 and the causal variants (**Fig. 1**). We detected significant signatures of natural selection ($S \neq 0$)
340 for 23 out of 28 complex traits in the UKB data, with the strongest selection signals from the
341 reproductive and cardiovascular trait-associated SNPs, followed by those associated with height,
342 handgrip strength, lung function and other anthropometric traits as well as educational
343 attainment (**Fig. 3**). Our findings are in line with an increasing body of literature supporting the
344 hypothesis of widespread polygenic selection on standing variants in complex traits^{4,39,40,59,60}.
345 Together with the high prevalence of selection signals across traits (23/28 = 82%), our
346 observation of high degree of polygenicity (~6% on average) underlines the role of pleiotropy
347 in the action of natural selection.

348

349 In the analysis of the UKB data, all the significant estimates of S for 23 traits were negative (**Fig.**
350 **3**), consistent with a model of negative selection (**Fig. 7**). The evidence of negative selection
351 against the trait-associated variants has been previously reported in some of these traits, such
352 as height and BMI³⁶. A recent study on ~110,000 Icelanders also detected negative selection on
353 EA-increasing variants over recent generations, as a result of delayed reproduction and fewer
354 children for the people with higher EA⁶¹. To support our results on some of the other traits, we

355 used the imputed SNPs based on a reference panel constructed by Haplotype Reference
356 Consortium⁶² to estimate the genetic variance across SNPs that are stratified by MAF and LD
357 scores (GREML-LDMS³⁶). We found that the genetic contribution of rare SNPs (MAF < 0.01) is
358 disproportional to that of common SNPs (MAF > 0.01) in height, BMI, WHR, and diastolic blood
359 pressure (**Supplementary Table 6**). These results also suggest that negative selection is
360 pervasive across traits, in line with the conclusion drawn from the BayesS analysis.

361

362 In the analyses of CAGE data, we identified 30 genes showing significant signatures, all of which
363 had negative \hat{S} (**Fig. 6**). With a less stringent criterion for hypothesis testing, we identified
364 additional 267 genes but only 45 of them had positive \hat{S} (**Fig. 5**). These results again suggest the
365 predominant role of negative selection in the human genome^{59,63,64} and support the hypothesis
366 that gene expression evolves primarily under stabilizing selection^{65,66}. The genes that showed
367 evidence of negative selection in our analysis may be functionally important and may deserve a
368 downstream study. There are gene-level metrics, such as Residual Variation Intolerance Score
369 (RVIS)⁶⁷ and Gene Damage Index (GDI)⁶⁸, aiming to prioritize genes for disease involvement
370 based on the functional variation within a gene, which can be used to infer the strength of
371 natural selection on the (coding) sequence of the gene. In contrast, our method interrogates
372 genome-wide SNPs to detect signals of selection on the SNPs associated with the expression
373 level of a gene, which largely depend on the trans-effects for polygenic transcripts. We found
374 that \hat{S} is not correlated with either RVIS or GDI (**Supplementary Fig. 13**). However, some genes
375 indeed showed strong evidence of selection in both lines. For example, *HERC2* ($\hat{S} = -1.16$, RVIS =
376 -5.99) is a pigmentation-related gene which has been suggested a target of recent selection^{69,70}.
377 In addition, genes with significantly negative \hat{S} generally also had low GDI, which is considered
378 to be an indicator of the relative biological indispensability⁶⁸. We do not expect to detect all
379 genes that are known to be under selection, such as the lactase gene *LCT* ($\hat{S} = 0.221$, s.e. = 0.317).
380 One possible reason is that the signatures of selection for these genes are concentrated in the
381 cis-regions and therefore might be diluted when we use all genome-wide SNPs to estimate S .

382

383 We conclude with several caveats. First, the polygenicity estimate ($\hat{\pi}$) only partially reflects the
384 actual fraction of causal variants since SNPs can possess nonzero effects through LD with
385 unobserved causal variants. Nevertheless, $\hat{\pi}$ can be used to compare the levels of polygenicity
386 across traits. Second, the power of detecting a signal of natural selection (i.e. testing against $S =$
387 0) may improve if whole-genome sequence (WGS) or imputed sequence data, which include a
388 large number of rare variants, are used in the analysis. However, it is computationally
389 challenging to run BayesS on all the WGS variants in a large cohort like UKB, a common problem
390 in the analysis of individual-level data with Bayesian methods. Depending on the sparsity of the

391 genetic signals, the nested BayesS provides a possibility to run the analysis in a manageable
392 amount of time but would still require a huge amount of memory to store the genotype matrix
393 for the WGS variants. A more practical approach is to model the genetic architecture using
394 summary statistics. Finally, in the simulation we observed a small inflation in estimating S using
395 the nested BayesS model, when all causal variants were genotyped and the true S is positive
396 (**Supplementary Fig. 7**). This suggests that the positive \hat{S} may be slightly overestimated in the
397 CAGE data analysis, but it would not change our conclusions since there was no significant
398 positive \hat{S} . Given that most complex traits have negative estimates of the relationship between
399 effect size and MAF, we expect to discover more rare variants of large effect in future GWAS
400 using WGS or imputed data.

401

402 **Online Methods**

403 **The BayesS model.** BayesS is a Bayesian multiple regression that simultaneously fits all the
404 SNP effects as random:

$$405 \quad \mathbf{y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

406 where \mathbf{y} is the vector of phenotypes, μ is the fixed effect, \mathbf{X} is the matrix of SNP genotype scores
407 centred by the column means, $\boldsymbol{\beta}$ is the vector of SNP effects, and \mathbf{e} is the residuals. The fixed
408 effect has a flat prior: $\mu \propto \text{constant}$. In practice, we fitted principal components and other
409 covariates as well in the model as fixed effects. It is common to standardize the SNP genotypes
410 such that each column of \mathbf{X} has variance one. But we do not standardize the SNP genotypes, as
411 the standardization implicitly assumes a strong negative relationship between SNP effect size
412 and MAF ($S = -1$)^{36,71-73}. As shown in Method Overview, we assume that the SNP effect β_j has a
413 hierarchical mixture prior

$$414 \quad \beta_j \sim N\left(0, [2p_j(1-p_j)]^S \sigma_\beta^2\right) \pi + \phi(1-\pi)$$

415 where ϕ is a point mass at zero and π , the proportion of SNPs with nonzero effects, is the
416 polygenicity. We allow data to dominate the inference on polygenicity by assuming a uniform
417 prior

$$418 \quad \pi \sim U(0, 1).$$

419 The variance of SNP effects, which quantifies our prior belief on the effect size, is modelled to be
420 related to MAF p_j through S , which is assumed to have a normal prior

$$421 \quad S \sim N(0, \sigma_S^2).$$

422 Namely, we *a priori* believe a selectively neutral model with some certainty (quantified by the
423 given variance) to allow the detection of selection to be driven by the data. We set $\sigma_S^2 = 1$ as the
424 prior in the analysis of UK Biobank traits, but a more informative prior $\sigma_S^2 = 0.1$ was used in the

425 analysis of CAGE gene expressions to shrink noise heavier toward zero given the much smaller
426 sample size. The prior for the common variance factor is

$$427 \quad \sigma_{\beta}^2 \sim \nu_{\beta} \tau_{\beta}^2 \chi_{\nu_{\beta}}^{-2}$$

428 where $\nu_{\beta} = 4$ and τ_{β}^2 is computed utilizing the characteristic of the distribution: if $\sigma^2 \sim \nu \tau^2 \chi_{\nu}^{-2}$,
429 then $E(\sigma^2) = \nu \tau^2 / (\nu - 2)$. Rearranging the equation gives

$$430 \quad \tau_{\beta}^2 = \frac{\nu_{\beta} - 2}{\nu_{\beta}} E(\sigma_{\beta}^2)$$

431 where

$$432 \quad E(\sigma_{\beta}^2) = \frac{V_g}{\tilde{\pi} \sum_j [2p_j(1 - p_j)]^{1+S}}$$

433 with V_g , $\tilde{\pi}$ and \tilde{S} are the prior knowledge of the genetic variance, π and S . To remove the
434 dependence of the hyperparameter τ_{β}^2 on the prior values of the genetic variance, π and S , we
435 compute τ_{β}^2 deterministically using the sampled values of these parameters for the first 2,000
436 MCMC cycles, and then set τ_{β}^2 to the average value across these cycles. Likewise, the prior for the
437 residual variance is

$$438 \quad \sigma_e^2 \sim \nu_e \tau_e^2 \chi_{\nu_e}^{-2}$$

439 where $\nu_e = 4$ and $\tau_e^2 = \frac{\nu_e - 2}{\nu_e} V_e$ with V_e a prior knowledge of the residual variance. Note that
440 when $S = 0$, our model becomes BayesC π ¹⁹, a method that has been widely used for genomic
441 prediction in agriculture, or Bayesian Variable Selection Regression (BVSR) in statistics
442 literature⁷⁴. The sampling scheme of the parameters is given in the **Supplementary Note**.

443
444 The nested BayesS model is developed based on a previously published method, BayesN⁷⁵, to
445 speed up computation when a large number of SNPs is included in the analysis. In the nested
446 BayesS, the genome is partitioned into W -Kb non-overlapping segments. Each window *a priori*
447 has k SNPs with nonzero effects, where W and k are some given numbers. SNPs in the same
448 window are individually modelled as in BayesS as well as collectively considered as a window
449 effect with a normal-zero mixture prior. Remarkable speedups are obtained by “jumping” fast
450 over the windows with zero effect, focusing solely on the windows that harbour genetic signals.
451 Thus, the reduction in computing time is inversely related to the polygenicity, which is defined
452 here as the proportion of segments with nonzero effects. When the causal variants are not
453 observed, choosing $k > 1$ may lead to better performance in parameter estimation than BayesS,
454 as it refines the signal of causal variant by allowing the flanking SNPs to jointly capture its effect.
455 Details on the nested BayesS and the comparison with the standard BayesS are given in the
456 **Supplementary Note**.

457

458 **Estimation of heritability.** We estimate the SNP-based heritability using the sampled values of
459 SNP effects in MCMC. By definition, the genetic variance is the variance of the genetic values
460 across individuals. In each MCMC cycle, we calculate the genetic values for each individual (\tilde{g}_i)
461 using SNPs with sampled nonzero effects ($\tilde{\beta}_j$):

$$462 \quad \tilde{g}_i = \sum_j X_{ij} \tilde{\beta}_j$$

463 Then, the genetic variance in the current cycle is

$$464 \quad \tilde{\sigma}_g^2 = \frac{\sum_i \tilde{g}_i^2}{N} - \left(\frac{\sum_i \tilde{g}_i}{N} \right)^2$$

465 Conditional on the sampled value of residual variance ($\tilde{\sigma}_e^2$), the SNP-based heritability is

$$466 \quad \tilde{h}_{SNP}^2 = \frac{\tilde{\sigma}_g^2}{\tilde{\sigma}_g^2 + \tilde{\sigma}_e^2}$$

467 The mean over all cycles after burn-in is the estimate of heritability

$$468 \quad \hat{h}_{SNP}^2 = E(\tilde{h}_{SNP}^2)$$

469 The standard deviation of the MCMC samples gives the standard error of the estimate and the
470 highest probability density gives the credible interval for posterior inference.

471

472 **ARIC simulation analysis.** The simulation based on Atherosclerosis Risk in Communities (ARIC)
473 cohort³³ was used for testing the methods. We used PLINK 1.9⁷⁶ to carry out standard quality
474 control (QC) procedures on the dataset, including removal of SNPs with missingness > 5%,
475 Hardy-Weinberg equilibrium test $P < 10^{-6}$, or MAF < 1%, and removal of individuals with
476 missing genotypes < 1% and genetic relationship < 0.05 estimated from all SNPs after QC using
477 GCTA-GRM⁷⁷. After all the QC steps, a total of 12,942 unrelated individuals and 564,959 SNPs
478 remained. A quantitative trait was simulated by choosing 1,000 SNPs at random as causal
479 variants with their effects sampled from a standard normal distribution. To simulate a spectrum
480 of relationships between MAF and effect size, the effect size was multiplied by $[2p_j(1-p_j)]^S$
481 where $S = -2, -1, 0, 1, 2$, representing negative to positive relationship between effect size
482 and MAF including the case of independence when $S = 0$. An individual phenotype specific to a
483 given value of S was generated by adding a random normal residual with the variance identical
484 to the genetic variance, giving each simulated trait a heritability of 0.5. The simulation process
485 was repeated for 100 times. BayesS and the nested BayesS were used to analyse the simulated
486 data with and without the causal variants in the model. To evaluate the robustness of our
487 method to the starting values of parameters, we used an arbitrary value of 0 for S , 0.2 for
488 heritability, and 0.05 for π , respectively to start the MCMC. In the nested BayesS, the length of
489 window was set to be 200-Kb with 2 SNPs *a priori* fitted in the model. It is noteworthy that the
490 distribution of genetic variance explained by each causal variant was not identical for different

491 scenarios of S in the true model. Under HWE, the genetic variance at locus j is $[2p_j(1 -$
492 $p_j)]^{S+1} \beta_j^2$ with $\beta_j \sim N(0, 1)$ in the simulation. Compared to a trait with $S < 0$, a trait with $S > 0$
493 has a larger proportion of loci each explaining a small proportion of variance, given an identical
494 distribution of MAF at the causal variants between traits (**Supplementary Fig. 14**). Thus, in the
495 scenario of $S > 0$, it is more difficult to capture the causal variants by SNP markers if causal
496 variants are not observed.

497

498 **Analysis of the UK Biobank data.** We have access to 46 complex traits in the UK Biobank²⁶,
499 where the phenotype data were collected from over 500,000 individuals aged between 40 and
500 69 across the United Kingdom. The interim release contains genotypes for 152,736 samples at
501 806,466 SNPs on a customized Affymetrix Axiom array after QC procedures⁷⁸. We selected a
502 subset of 140,408 individuals that had a self-reported gender identical to the genetically
503 inferred gender and a European ethnicity derived from a principal component analysis together
504 with self-reported ethnicity. Furthermore, we removed individuals with genomic relatedness $>$
505 0.05 estimated from all SNPs using GCTA-GRM⁷⁷ and SNPs with genotype missing rate $>$ 5%,
506 Hardy-Weinberg equilibrium test $P < 10^{-6}$, or MAF $<$ 1%. The final data set consisted of
507 126,752 individuals of European ancestry with 483,634 common SNPs (MAF $>$ 1%). After
508 removal of 5 duplicated traits and 5 traits with sample size (N) $<$ 20,000, we had 36 traits
509 remained for analysis, including 32 quantitative traits (anthropometric, cardiovascular and
510 reproductive), 2 categorical traits – male pattern baldness (MPB) and years of schooling
511 (educational attainment) and 2 diseases – type 2 diabetes (T2D) and major depressive disease
512 (MDD). The sample sizes of the traits are shown in **Supplementary Table 2**, where most traits
513 had $N >$ 100,000. The prevalence of T2D and MDD in the sample was 5.35% and 6.70%,
514 respectively. The phenotypes of quantitative traits were standardized within each sex group
515 after regressing out the age effect. For educational attainment, the years of schooling are pre-
516 adjusted by sex, a third order polynomial of year-of-birth and year-of-birth by sex interactions.
517 We used BayesS for the analysis, where the first 20 principal components (PC) of GRM were
518 fitted as fixed effects to account for the effects due to population stratification. For the disease
519 traits, sex and age were fitted as covariates in addition to PCs, and for MPB, only age was fitted
520 as the additional covariate.

521

522 **Consortium for the Architecture of Gene Expression (CAGE) data set.** We analyzed the
523 mRNA levels for 36,778 transcript expression traits (probes) from the Consortium for the
524 Architecture of Gene Expression (CAGE)²⁷ data set using the nested BayesS method. The CAGE
525 data comprised of measurements from 36,778 gene expression probes in peripheral blood, with
526 a subset of 1,748 unrelated (genomic relatedness $>$ 0.05) European individuals from the total

527 2,765 individuals used for this analysis. Full details of the cohorts contributing to CAGE, and
528 their sample preparation, normalization and genotype imputation are outlined in Lloyd-Jones *et*
529 *al.*²⁷. Briefly, the quantification of gene expression for each cohort was performed using the
530 Illumina Whole-Genome Expression BeadChips. The gene expression levels in each cohort were
531 initially normalized, followed by a quantile adjustment to standardize the distribution of
532 expression levels across samples. We corrected for age, gender, cell counts and batch effects as
533 well as hidden heterogeneous sources of variability. The rank-based inverse-normal
534 transformation was used to normalize the measurements for each probe to be normally-
535 distributed with mean 0 and variance 1. Probes measuring expression levels of genes located on
536 chromosomes X and Y were removed from the analysis. The initial CAGE dataset consisted of
537 seven unique cohorts that were genotyped on different SNP arrays. Therefore, genotype data
538 were imputed to the 1000 Genomes Phase 1 Version 3 reference panel⁷⁹, resulting in 7,763,174
539 SNPs passing quality control of which 1,066,738 SNPs overlapped with HapMap3 and were used
540 for analysis.

541

542 **Forward simulation for different types of natural selection.** We ran forward simulations
543 using SLiM⁸⁰ to confirm that the relationship between effect size and MAF is subject to different
544 types of natural selection. We simulated a 10-Mb region where new mutations occurred with
545 probability of 0.95 to be neutral and of 0.05 to be a causal variant with an effect sampled from a
546 standard normal distribution. The mutation rate was set to be 1.65×10^{-8} ⁸¹. The phenotype of
547 an individual was simulated based on the genotypic values at all segregating causal variants in
548 the current generation with a heritability of 0.1. We simulated the evolution of a population of
549 1,000 individuals over 10,000 generations (this is equivalent to 100,000 generations in a
550 population of 10,000 individuals⁸²). The first 5,000 generations were used a burn-in period,
551 where the phenotype did not affect fitness and all variants (including the causal variants) were
552 under neutral variation. From generation 5,001, we related the standardized phenotype with
553 mean zero and variance one to fitness through a hypothetical function that represents different
554 types of selection (**Fig. 7**, top row). For directional selection, the phenotype was positively or
555 negatively correlated to fitness through a simple linear function. For stabilizing selection, we
556 used a normal curve to model that fitness achieved optimum at intermediate phenotype value.
557 For disruptive selection, a reversed normal curve was used to model that the phenotypes at the
558 two tails produced highest fitness. In the last generation of selection, we investigated the joint
559 distribution of effects and frequencies of the derived alleles, the joint distribution of effects and
560 frequencies of the coded alleles (arbitrarily chosen as in reality where derived alleles are
561 unknown), and the relationship between the variance of the coded-allele effects and MAF. We
562 collected results from 200 simulation replicates.

563

564

565 **Acknowledgments**

566 This research was supported by the Australian Research Council (DP160101343), the
567 Australian National Health and Medical Research Council (1107258, 1078901, 1078037, and
568 1113400), and the Sylvia & Charles Viertel Charitable Foundation (Senior Medical Research
569 Fellowship). R.d.V. acknowledges funding from an ERC consolidator grant (647648 EdGe,
570 awarded to Philipp Koellinger). This study makes use of data from dbGaP (accession number:
571 phs000090.v3.p1) and UK Biobank Resource (Application number: 12514). A full list of
572 acknowledgements to these data sets can be found in the **Supplementary Note**. At the end, we
573 wish to acknowledge The University of Queensland's Research Computing Centre (RCC) for its
574 support in this research.

575

576 **Author contributions**

577 J.Y., P.M.V. and R.d.V. conceived the study. J.Y., J.Z. and P.M.V. designed the experiment. J.Z.
578 derived the analytical methods, conducted all analyses and developed the software with
579 assistance and guidance from J.Y., Y.W., M.R.R., L.L.J., L.Y.D., C.Y. A.X. and J.S.. L.L.J., A.F.M., J.E.P.,
580 G.W.M., A.M., T.E., G.G. and P.M.V. provided the CAGE data. J.Z. and J.Y. wrote the manuscript with
581 the participation of all authors. All authors reviewed and approved the final manuscript.

582

583 **References**

- 584 1. Hansen, T.F., Alvarez-Castro, J.M., Carter, A.J., Hermisson, J. & Wagner, G.P. Evolution of
585 genetic architecture under directional selection. *Evolution* **60**, 1523-36 (2006).
- 586 2. Pritchard, J.K. & Di Rienzo, A. Adaptation - not by sweeps alone. *Nat Rev Genet* **11**, 665-7
587 (2010).
- 588 3. Hancock, A.M. *et al.* Colloquium paper: human adaptations to diet, subsistence, and
589 ecoregion are due to subtle shifts in allele frequency. *Proc Natl Acad Sci U S A* **107 Suppl**
590 **2**, 8924-30 (2010).
- 591 4. Pritchard, J.K., Pickrell, J.K. & Coop, G. The genetics of human adaptation: hard sweeps,
592 soft sweeps, and polygenic adaptation. *Curr Biol* **20**, R208-15 (2010).
- 593 5. Smith, J.M. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genet Res* **23**, 23-35
594 (1974).
- 595 6. Wright, S. The Distribution of Gene Frequencies in Populations of Polyploids. *Proc Natl*
596 *Acad Sci U S A* **24**, 372-7 (1938).
- 597 7. Eyre-Walker, A. Evolution in health and medicine Sackler colloquium: Genetic
598 architecture of a complex trait and its implications for fitness and genome-wide
599 association studies. *Proc Natl Acad Sci U S A* **107 Suppl 1**, 1752-6 (2010).
- 600 8. Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* **536**, 41-7
601 (2016).
- 602 9. Park, J.H. *et al.* Distribution of allele frequencies and effect sizes and their
603 interrelationships for common genetic susceptibility variants. *Proc Natl Acad Sci U S A*
604 **108**, 18026-31 (2011).

- 605 10. Berndt, S.I. *et al.* Genome-wide meta-analysis identifies 11 new loci for anthropometric
606 traits and provides insights into genetic architecture. *Nature Genetics* **45**, 501-U69
607 (2013).
- 608 11. Slatkin, M. Genotype-specific recurrence risks as indicators of the genetic architecture of
609 complex diseases. *American Journal of Human Genetics* **83**, 120-126 (2008).
- 610 12. Liu, C.T., Raghavan, S. & Maruthur, N. Trans-ethnic meta-analysis and functional
611 annotation illuminates the genetic architecture of fasting glucose and insulin. *The*
612 *American Journal of ...* (2016).
- 613 13. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological
614 insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-7 (2014).
- 615 14. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**,
616 747-53 (2009).
- 617 15. Lee, S.H., Wray, N.R., Goddard, M.E. & Visscher, P.M. Estimating missing heritability for
618 disease from genome-wide association studies. *Am J Hum Genet* **88**, 294-305 (2011).
- 619 16. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human
620 height. *Nat Genet* **42**, 565-9 (2010).
- 621 17. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common
622 SNPs. *Nature Genetics* **43**, 519-U44 (2011).
- 623 18. Gratten, J., Wray, N.R., Keller, M.C. & Visscher, P.M. Large-scale genomics unveils the
624 genetic architecture of psychiatric disorders. *Nat Neurosci* **17**, 782-90 (2014).
- 625 19. Habier, D., Fernando, R.L., Kizilkaya, K. & Garrick, D.J. Extension of the bayesian alphabet
626 for genomic selection. *BMC Bioinformatics* **12**, 186 (2011).
- 627 20. Moser, G. *et al.* Simultaneous discovery, estimation and prediction analysis of complex
628 traits using a bayesian mixture model. *PLoS Genet* **11**, e1004969 (2015).
- 629 21. Meuwissen, T.H.E., Hayes, B.J. & Goddard, M.E. Prediction of total genetic value using
630 genome-wide dense marker maps. *Genetics* **157**, 1819-1829 (2001).
- 631 22. de Los Campos, G., Hickey, J.M., Pong-Wong, R., Daetwyler, H.D. & Calus, M.P. Whole-
632 genome regression and prediction methods applied to plant and animal breeding.
633 *Genetics* **193**, 327-45 (2013).
- 634 23. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with bayesian sparse linear
635 mixed models. *PLoS Genet* **9**, e1003264 (2013).
- 636 24. Lloyd-Jones, L.R. *et al.* Inference on the Genetic Basis of Eye and Skin Colour in an
637 Admixed Population via Bayesian Linear Mixed Models. *Genetics* (2017).
- 638 25. Mancuso, N. *et al.* The contribution of rare variation to prostate cancer heritability. *Nat*
639 *Genet* **48**, 30-5 (2016).
- 640 26. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide
641 range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015).
- 642 27. Lloyd-Jones, L.R., Holloway, A., McRae, A. & Yang, J. The genetic architecture of gene
643 expression in peripheral blood. *The American Journal of ...* (2017).
- 644 28. Neal, R.M. MCMC Using Hamiltonian Dynamics. *Handbook of Markov Chain Monte Carlo*,
645 113-162 (2011).
- 646 29. Fernando, R.L., Dekkers, J.C. & Garrick, D.J. A class of Bayesian methods to combine large
647 numbers of genotyped and non-genotyped animals for whole-genome analyses. *Genet*
648 *Sel Evol* **46**, 50 (2014).
- 649 30. Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. Book Reviews. *Journal of the American*
650 *Statistical Association* **109**, 1325-1337 (2014).
- 651 31. Fernando, R.L. *et al.* Controlling the proportion of false positives in multiple dependent
652 tests. *Genetics* **166**, 611-9 (2004).
- 653 32. Storey, J.D. The optimal discovery procedure: a new approach to simultaneous
654 significance testing. *Journal of the Royal Statistical Society Series B-Statistical*
655 *Methodology* **69**, 347-368 (2007).
- 656 33. Psaty, B.M. *et al.* Cohorts for Heart and Aging Research in Genomic Epidemiology
657 (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide
658 association studies from 5 cohorts. *Circ Cardiovasc Genet* **2**, 73-80 (2009).

- 659 34. Wood, A.R. *et al.* Defining the role of common variation in the genomic and biological
660 architecture of adult human height. *Nat Genet* **46**, 1173-86 (2014).
- 661 35. Locke, A.E. *et al.* Genetic studies of body mass index yield new insights for obesity
662 biology. *Nature* **518**, 197-206 (2015).
- 663 36. Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing
664 heritability for human height and body mass index. *Nat Genet* **47**, 1114-20 (2015).
- 665 37. Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*
666 **528**, 499-503 (2015).
- 667 38. Marouli, E. *et al.* Rare and low-frequency coding variants alter human adult height.
668 *Nature* **542**, 186-190 (2017).
- 669 39. Robinson, M.R. *et al.* Population genetic differentiation of height and body mass index
670 across Europe. *Nat Genet* **47**, 1357-62 (2015).
- 671 40. Field, Y. *et al.* Detection of human adaptation during the past 2000 years. *Science* **354**,
672 760-764 (2016).
- 673 41. Matthew R. Robinson, G.E., Gerhard Moser, Luke R. Lloyd-Jones, Marcus A. Triplett,
674 Zhihong Zhu, Ilja M. Nolte, Jana V. van Vliet-Ostaptchouk, Harold Snieder, The LifeLines
675 Cohort Study, Tonu Esko, Lili Milani, Reedik Mägi, Andres Metspalu, Patrik K. E.
676 Magnusson, Nancy L. Pedersen, Erik Ingelsson, Magnus Johannesson, Jian Yang, David
677 Cesarini and Peter M. Visscher. Genotype-covariate interaction effects and the
678 heritability of adult body mass index. *Nature Genetics* (In press).
- 679 42. Koning, D.L., Merchant, A.T. & Pogue, J. Waist circumference and waist-to-hip ratio as
680 predictors of cardiovascular events: meta-regression analysis of prospective studies.
681 *European heart journal* **28.7**, 850-856 (2007).
- 682 43. Wass, P., Waldenstrom, U., Rossner, S. & Hellberg, D. An android body fat distribution in
683 females impairs the pregnancy rate of in-vitro fertilization-embryo transfer. *Hum*
684 *Reprod* **12**, 2057-60 (1997).
- 685 44. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits.
686 *Nat Genet* **47**, 1236-41 (2015).
- 687 45. Hill, W.G., Goddard, M.E. & Visscher, P.M. Data and theory point to mainly additive
688 genetic variance for complex traits. *PLoS Genet* **4**, e1000008 (2008).
- 689 46. Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci
690 with shared effects on five major psychiatric disorders: a genome-wide analysis. *The*
691 *Lancet* **381**, 1371-1379 (2013).
- 692 47. Das, S.K. & Elbein, S.C. The Genetic Basis of Type 2 Diabetes. *Cellscience* **2**, 100-131
693 (2006).
- 694 48. Konarzewski, M. & Ksiazek, A. Determinants of intra-specific variation in basal metabolic
695 rate. *J Comp Physiol B* **183**, 27-41 (2013).
- 696 49. Nyholt, D.R., Gillespie, N.A., Heath, A.C. & Martin, N.G. Genetic basis of male pattern
697 baldness. *J Invest Dermatol* **121**, 1561-4 (2003).
- 698 50. Liu, F. *et al.* Prediction of male-pattern baldness from genotypes. *Eur J Hum Genet* **24**,
699 895-902 (2016).
- 700 51. Hyde, C.L. *et al.* Identification of 15 genetic loci associated with risk of major depression
701 in individuals of European descent. *Nat Genet* **48**, 1031-6 (2016).
- 702 52. de Moor, M.H. *et al.* Meta-analysis of Genome-wide Association Studies for Neuroticism,
703 and the Polygenic Association With Major Depressive Disorder. *JAMA Psychiatry* **72**,
704 642-50 (2015).
- 705 53. Davies, G. *et al.* Genome-wide association study of cognitive functions and educational
706 attainment in UK Biobank (N=112 151). *Mol Psychiatry* **21**, 758-67 (2016).
- 707 54. International HapMap 3 Consortium. Integrating common and rare genetic variation in
708 diverse human populations. *Nature* **467**, 52-8 (2010).
- 709 55. Barbosa-Morais, N.L. *et al.* A re-annotation pipeline for Illumina BeadArrays: improving
710 the interpretation of gene expression data. *Nucleic Acids Res* **38**, e17 (2010).
- 711 56. Bustamante, C.D. *et al.* Natural selection on protein-coding genes in the human genome.
712 *Nature* **437**, 1153-7 (2005).

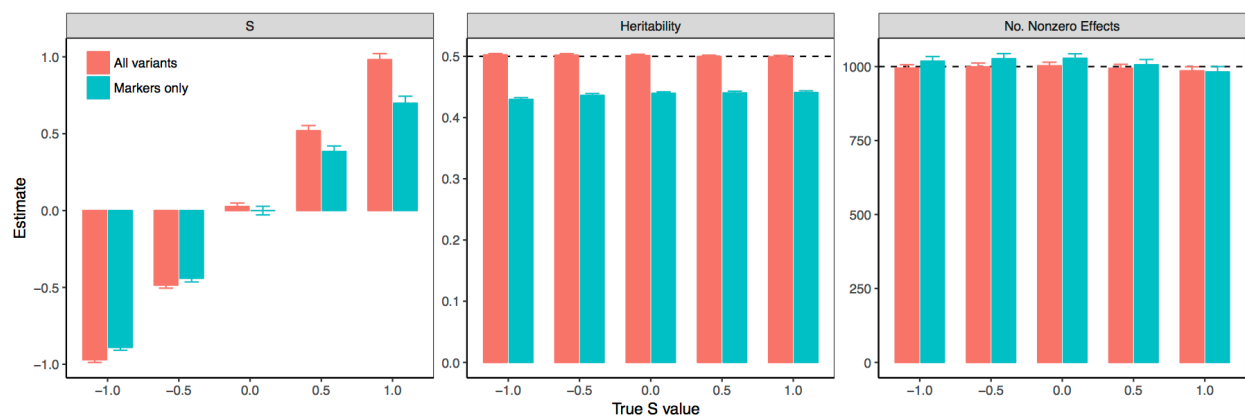
- 713 57. Casto, A.M. & Feldman, M.W. Genome-wide association study SNPs in the human genome
714 diversity project populations: does selection affect unlinked SNPs with shared trait
715 associations? *PLoS Genet* **7**, e1001266 (2011).
- 716 58. Berg, J.J. & Coop, G. A population genetic signal of polygenic adaptation. *PLoS Genet* **10**,
717 e1004412 (2014).
- 718 59. Hernandez, R.D. *et al.* Classic Selective Sweeps Were Rare in Recent Human Evolution.
719 *Science* **331**, 920-924 (2011).
- 720 60. Turchin, M.C. *et al.* Evidence of widespread selection on standing variation in Europe at
721 height-associated SNPs. *Nature Genetics* **44**, 1015-+ (2012).
- 722 61. Kong, A. *et al.* Selection against variants in the genome associated with educational
723 attainment. *Proc Natl Acad Sci U S A* **114**, E727-E732 (2017).
- 724 62. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat*
725 *Genet* **48**, 1279-83 (2016).
- 726 63. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural
727 selection in hominid evolution. *PLoS Genet* **5**, e1000471 (2009).
- 728 64. Sohail, M. *et al.* Negative selection in humans and fruit flies involves synergistic epistasis.
729 *Science* **356**, 539-542 (2017).
- 730 65. Romero, I.G., Ruvinsky, I. & Gilad, Y. Comparative studies of gene expression and the
731 evolution of gene regulation. *Nat Rev Genet* **13**, 505-16 (2012).
- 732 66. Gilad, Y., Oshlack, A. & Rifkin, S.A. Natural selection on gene expression. *Trends Genet* **22**,
733 456-61 (2006).
- 734 67. Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S. & Goldstein, D.B. Genic intolerance to
735 functional variation and the interpretation of personal genomes. *PLoS Genet* **9**,
736 e1003709 (2013).
- 737 68. Itan, Y. *et al.* The human gene damage index as a gene-level approach to prioritizing
738 exome variants. *Proc Natl Acad Sci U S A* **112**, 13615-20 (2015).
- 739 69. Grossman, S.R. *et al.* A composite of multiple signals distinguishes causal variants in
740 regions of positive selection. *Science* **327**, 883-6 (2010).
- 741 70. Chen, G.B., Lee, S.H., Zhu, Z.X., Benyamin, B. & Robinson, M.R. EigenGWAS: finding loci
742 under selection through genome-wide association studies of eigenvectors in structured
743 populations. *Heredity (Edinb)* **117**, 51-61 (2016).
- 744 71. Lee, S.H. *et al.* Estimation of SNP heritability from dense genotype data. *Am J Hum Genet*
745 **93**, 1151-5 (2013).
- 746 72. Speed, D., Hemani, G., Johnson, M.R. & Balding, D.J. Improved heritability estimation from
747 genome-wide SNPs. *Am J Hum Genet* **91**, 1011-21 (2012).
- 748 73. Speed, D. *et al.* Reevaluation of SNP heritability in complex human traits. *Nature Genetics*
749 (2017).
- 750 74. Guan, Y. & Stephens, M. Bayesian variable selection regression for genome-wide
751 association studies and other large-scale problems. *The Annals of Applied Statistics*,
752 1780-1815 (2011).
- 753 75. Zeng, J. Whole genome analyses accounting for structures in genotype data. *PhD diss.,*
754 *IOWA STATE UNIVERSITY* (2015).
- 755 76. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer
756 datasets. *Gigascience* **4**, 7 (2015).
- 757 77. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex
758 trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).
- 759 78. UkbioBank, U.K. Genotyping and quality control of UK Biobank, a large-scale, extensively
760 phenotyped prospective resource. Available at *biobank.ctsu.ox.ac.uk/crystal/docs/genotyping_qc.pdf*. Accessed April 1 (2015).
- 761
762 79. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092
763 human genomes. *Nature* **491**, 56-65 (2012).
- 764 80. Messer, P.W. SLiM: simulating evolution with selection and linkage. *Genetics* **194**, 1037-
765 9 (2013).

- 766 81. Palamara, P.F. *et al.* Leveraging Distant Relatedness to Quantify Human Mutation and
767 Gene-Conversion Rates. *Am J Hum Genet* **97**, 775-89 (2015).
768 82. Enard, D., Messer, P.W. & Petrov, D.A. Genome-wide signals of positive selection in
769 human evolution. *Genome Res* **24**, 885-95 (2014).
770

771 **Figures**

772 Figure 1: Estimation of the genetic architecture parameters, e.g. S , heritability and polygenicity, for a
773 simulated trait using the ARIC data. Results are the mean estimates with s.e.m. (cap) over 100
774 simulation replicates for a spectrum of S parameter values. Colour indicates the results of BayesS
775 with both causal variants and SNP markers (red) or with SNP markers only (blue). The heritability at
776 the 1,000 randomly selected causal variants was 0.5. The number of nonzero effects is the number
777 of SNPs with nonzero effects averaged over MCMC iterations.

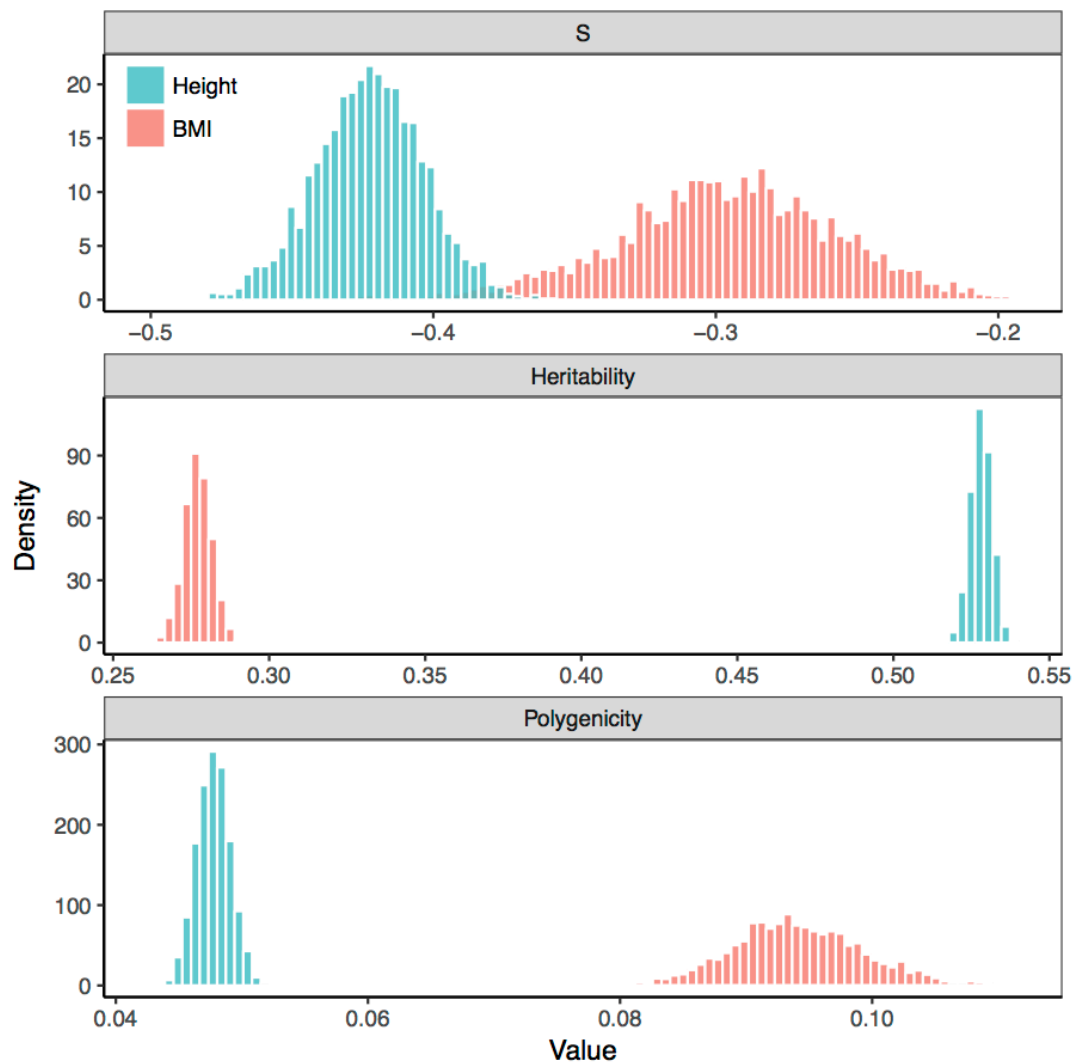
778



779

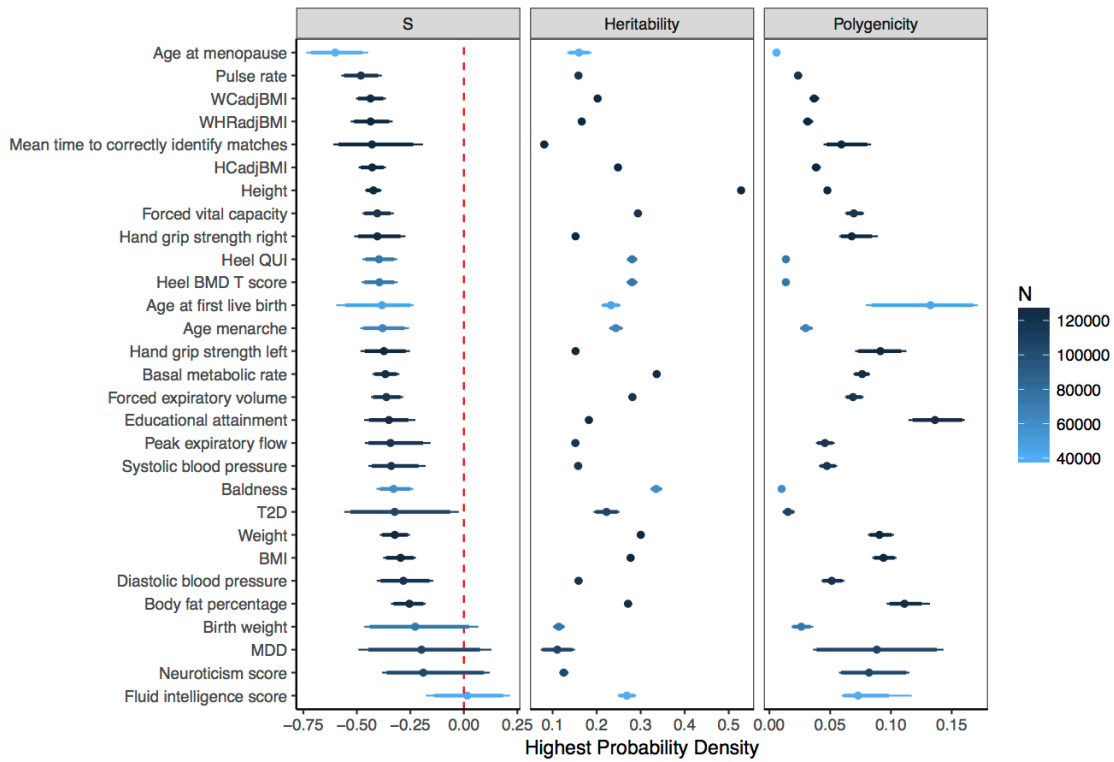
780

781 Figure 2: Posterior distributions of the genetic architecture parameters for height versus BMI using
782 data from UKB. S measures the relationship between SNP effect size and MAF. Polygenicity is
783 defined as the proportion of SNPs with nonzero effects.
784



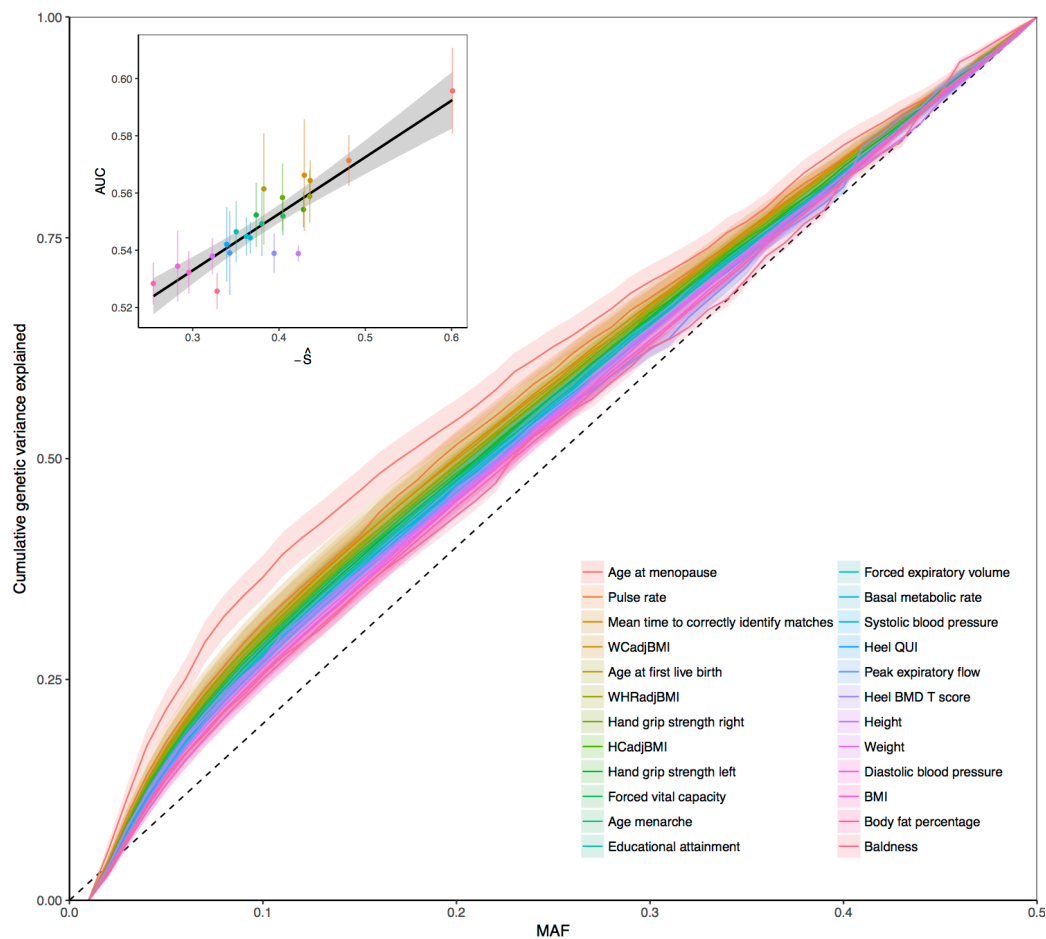
785

786 Figure 3: Posterior modes with credible intervals for the genetic architecture parameters using
787 BayesS. Results are for the 28 UKB complex traits that had passed convergence tests on the MCMC
788 chain. The bold line represents 95% credible interval (highest posterior density, HPD) and the thin
789 line represents 90% credible interval. Sample size N for each trait is shown by the colour gradient.
790 Polygenicity is defined as the proportion of genome-wide SNPs with nonzero effects on the trait.
791



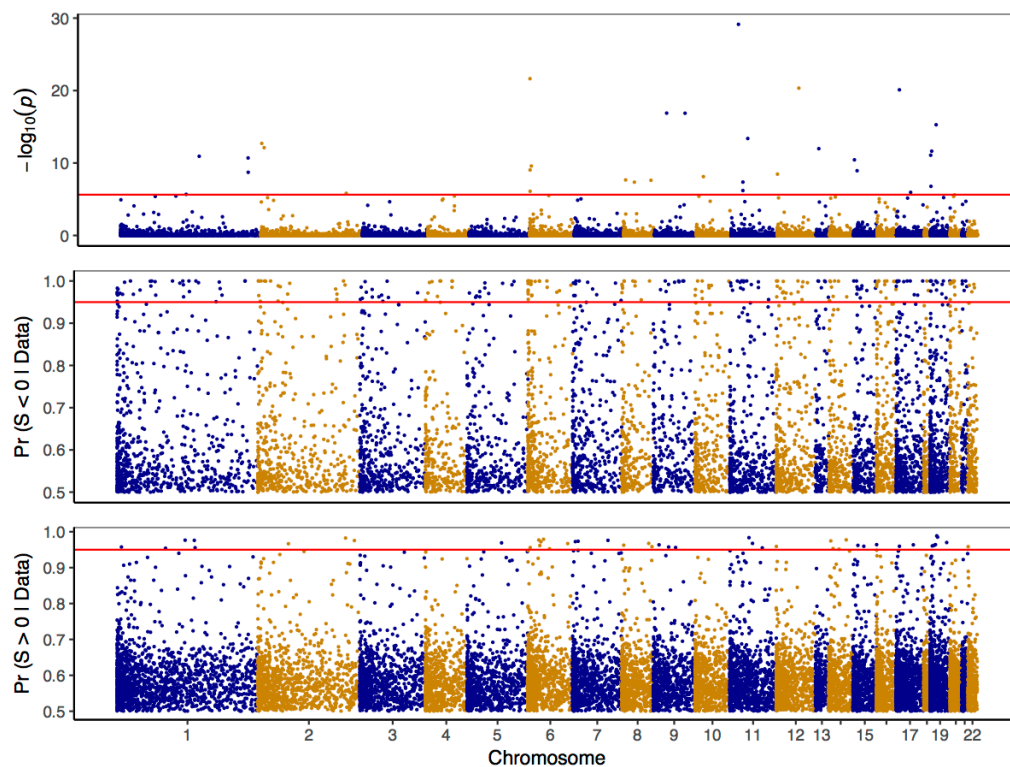
792

793 Figure 4: Cumulative genetic variance explained by SNPs with MAF smaller than a threshold on the x-
794 axis. The lines are the posterior means for the 23 UKB complex traits from UKB for which the
795 estimates of S were significantly different from zero. Shadow shows the posterior standard error.
796 The inner graph shows the relationship between the area under the curve (AUC) of the cumulative
797 genetic variance and negative \hat{S} (bar shows s.e.) across traits.
798



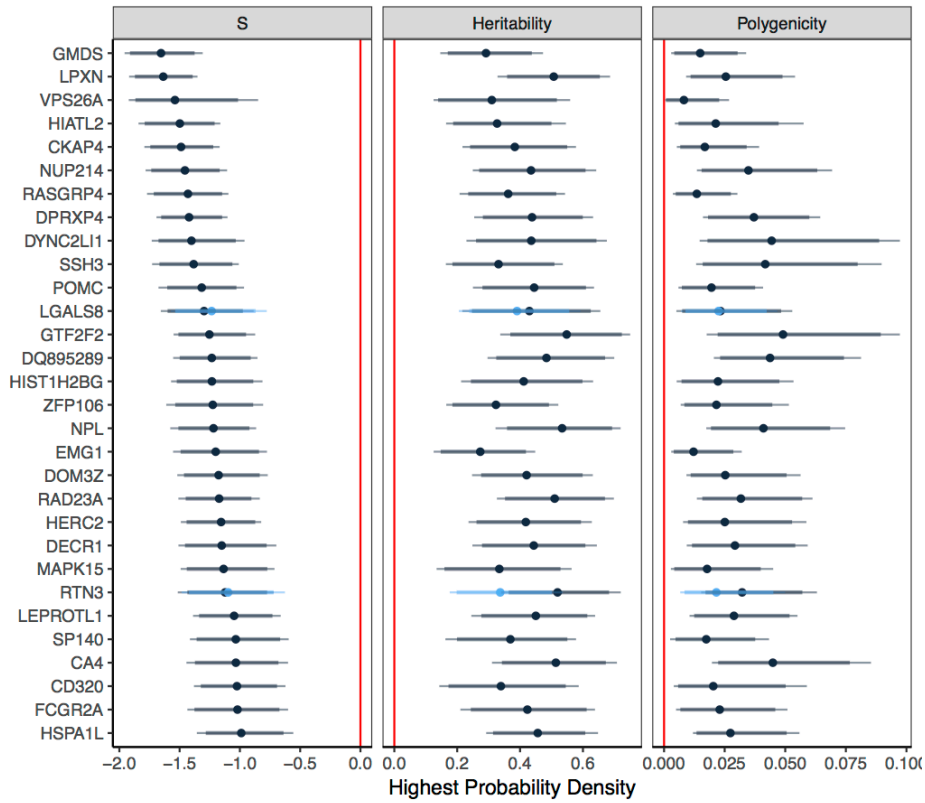
799

800 Figure 5: Genome-wide evidence of selection from the p-values to test against $S = 0$ and the
801 posterior probability of $S < 0$ or $S > 0$ for 21,303 probes in the CAGE data after QC. The red line
802 shows the significant threshold of 0.05 after Bonferroni correction ($p\text{-value} = 2.3 \times 10^{-6}$) or 0.95 for
803 the posterior probabilities.
804



805

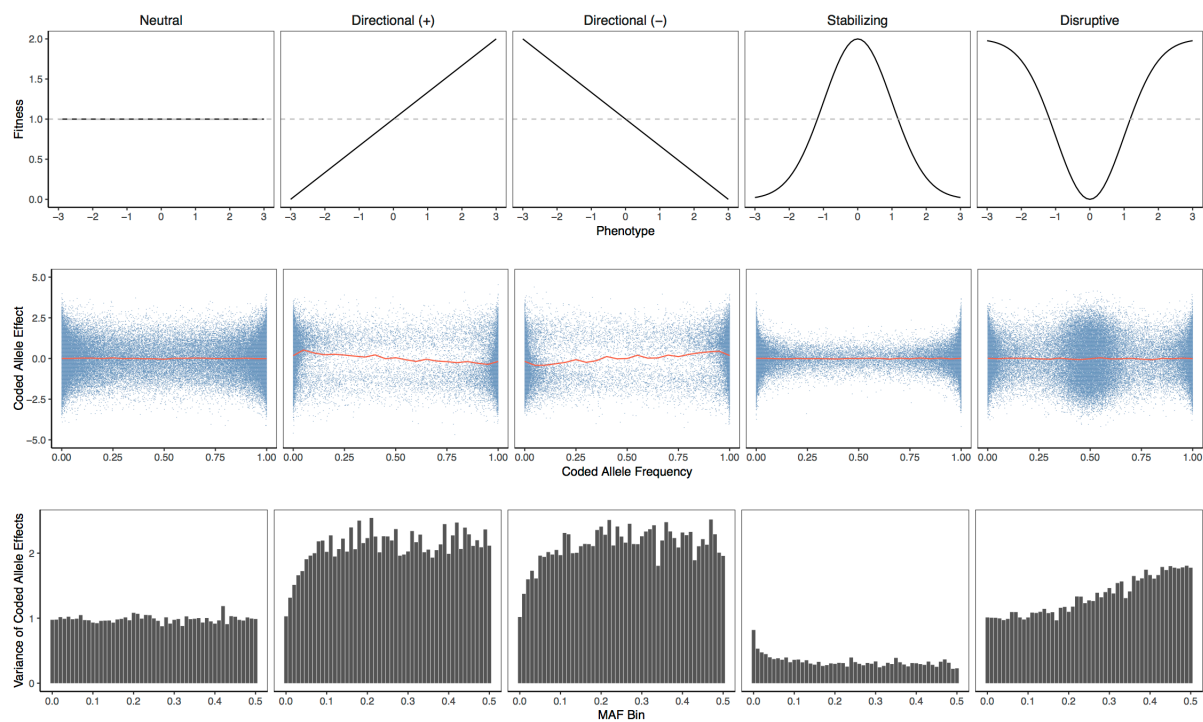
806 Figure 6: Estimation of the genetic architecture parameters for 30 genes (corresponding to 32
807 probes) with significant \hat{S} (p-value < 0.05/21,303) in the analysis of CAGE data. Results are the
808 posterior modes with credible intervals obtained from the nested BayesS model. The bold line
809 represents 95% credible interval (highest posterior density, HPD) and the thin line represents 90%
810 credible interval. Polygenicity is defined as the proportion of 200-Kb windows with nonzero effects
811 in the genome. The light colour shows the results of the second probe tagging one gene.
812



813

814

815 Figure 7: Forward simulations with different types of selection. A quantitative trait was generated
816 based on a simulated chromosome segment of 10Mb (5% causal and 95% neutral mutations in each
817 generation). The trait heritability was 0.1. The top row shows the functions used to relate the
818 phenotype (normally distributed) to fitness in different modes of selection: neutral variation,
819 directional selection with the phenotype positively (+) or negatively (-) correlated to fitness,
820 stabilizing selection and disruptive selection. The 2nd row shows the joint distributions of the coded
821 allele effects and frequencies, where the coded allele at each causal variant was chosen at random
822 from the derived and ancestral alleles, and the red line shows the means of coded allele effects in
823 allele frequency intervals of 0.05. The bottom row shows the relationships between the variance of
824 coded allele effects and MAF. Results were collected from 200 replicates of simulation.
825



826