

1 **Rare variant effect estimation and polygenic risk prediction**

2

3 **Kisung Nam¹, Minjung Kho¹, Wei Zhou^{2,3,4}, Bhramar Mukherjee⁵, Seunggeun Lee¹**

4

5 **Abstract**

6

7 Due to their low frequency, estimating the effect of rare variants is challenging. Here, we
8 propose RareEffect, a method that first estimates gene or region-based heritability and then
9 each variant effect size using an empirical Bayesian approach. Our method uses a variance
10 component model, popular in rare variant tests, and is designed to provide two levels of effect
11 sizes, gene/region-level and variant-level, which can provide better interpretation. To adjust
12 for the case-control imbalance in phenotypes, our approach uses a fast implementation of the
13 Firth bias correction. We demonstrate the accuracy and computational efficiency of our
14 method through extensive simulations and the analysis of UK Biobank whole exome
15 sequencing data for five continuous traits and five binary disease phenotypes. Additionally,
16 we show that the effect sizes obtained from our model can be leveraged to improve the
17 performance of polygenic scores.

18

¹ Graduate School of Data Science, Seoul National University, Seoul, South Korea

² Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA

³ Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA

⁴ Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

⁵ Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI, USA

19 Introduction

20 With the availability of extensive sequencing data in biobanks¹, the study of rare variants has
21 become more feasible than ever before. Rare variants have been identified as potential
22 causative factors for numerous complex diseases²⁻⁸, and their exploration is crucial in
23 unraveling the genetic risk factors of complex traits⁹. To identify the association between rare
24 variants and complex traits, gene or region-based tests^{10,11}, including the Burden test, the
25 sequence kernel association test (SKAT)¹² and its adaptive optimized version (SKAT-O)¹³, have
26 been commonly used. Recently, several methods, including STAAR¹⁴, SAIGE-GENE¹⁵, and
27 SAIGE-GENE+¹⁶ are developed to run region-based tests in biobank scale data.

28

29 To elucidate the effect of rare variants on complex diseases and traits and utilize them for risk
30 prediction, in addition to calculating p-values for association tests, effect size estimation is
31 required. However, estimating the effect size of rare variants remains a challenge. The low
32 minor allele frequencies make single variant-based estimations unstable. Popular association
33 tests like SKAT and SKAT-O are score tests, so only provide p-values. Although the Burden test
34 approach provides a gene-burden effect size, this may not accurately reflect the true effect of
35 rare variants in the presence of null variants and variants with opposing association directions.

36

37 To address the challenges, we introduce RareEffect, a novel method that estimates gene or
38 region-based heritability and subsequently calculates each variant's effect size using an
39 empirical Bayesian approach. Utilizing a variance component model, similar to that in the SKAT
40 test, our method offers dual-level effect size estimation, region-level and variant-level, for
41 enhanced interpretability. Unlike the Burden approach, our model flexibly estimates effect
42 sizes in a data-driven manner. To reduce the computational burden for estimating variance
43 components, we implemented the Factored Spectrally Transformed Linear Mixed Models
44 (FaST-LMM)¹⁷, which leverages the low rank of the genetic relatedness matrix. We also utilize
45 the strategy to collapse ultra-rare variants, as used in SAIGE-GENE+, to reduce the sparsity of
46 the genotype matrix and improve power of estimating the effect of ultra-rare variants. For
47 binary traits, we additionally apply a fast implementation of the Firth bias reduction method
48 to stably estimate the effect sizes.

49

50 From simulation studies, we showed that the proposed method is computationally fast and
51 reliably estimates each gene heritability. We also showed that the proposed approach
52 outperformed linear regression or ridge regression in terms of root mean squared error (RMSE)
53 for estimating the individual-variant level effect sizes. From the UK Biobank (UKB) Whole
54 Exome Sequencing (WES) data analysis of 5 quantitative traits and 5 binary disease traits, we
55 demonstrate that exonic rare variants can explain substantial phenotypic variabilities, but the
56 degree differs by phenotypes. We also showed that our approach has higher explanatory
57 power in explaining the phenotype variability compared to models based on burden tests.
58 These findings provide strong evidence for the practical utility of our method in leveraging
59 rare variant data for risk prediction and heritability estimation.

60

61 **Result**

62 *Overview of methods*

63 Our proposed method encompasses three steps. The overview of the method is outlined in
64 **Fig. 1** and is described in detail in the Methods section.

65

66 In step 1, we fit a null linear or logistic mixed-effect model without genotypes to estimate the
67 covariate effects. This step involves fitting the model using the average information restricted
68 maximum likelihood (AI-REML)¹⁸ approach, which is utilized in the SAIGE¹⁹ and GMMAT²⁰
69 framework. Residuals will be used as covariate-adjusted phenotypes in the subsequent steps.

70

71 In step 2, we model the effect of rare variants ($MAF \leq 0.01$) as random effects and estimate the
72 variance components, and hence heritability. To mitigate the computational demands, we
73 adopt the Factored Spectrally Transformed Linear Mixed Models (FaST-LMM)¹⁷. Given $n \gg k$,
74 FaST-LMM reduces the computation cost from $O(n^3)$ of conventional mixed model algorithm
75 to $O(nk^2)$ where n is the number of samples in the dataset and k is the number of genetic
76 variants in a single group (See **Methods**). Additionally, we leveraged the sparsity in genotypes.
77 To incorporate the fact that genetic effects vary by functional annotations, we fit the model
78 separately for distinct categories, and then combine them to calculate gene or region-level
79 heritability. For whole exome analysis, we include three functional categories: (1) Loss-of-
80 function (LoF) variants; (2) missense variants; and (3) synonymous variants. Within each

81 category, the ultra-rare variants, defined as those with a minor allele count (MAC) of lower
82 than 10, are collapsed into a single variant, as employed in SAIGE-GENE+¹⁶. As estimating
83 multiple variance components due to distinct functional groups is not feasible with FaST-LMM,
84 we marginally calculate variance component for each functional group separately, and then
85 adjust them using method of moments (MoM) approach (See **Methods**).

86
87 In step 3, following the estimation of variance components, we calculate the effect size of each
88 variant using the Best Linear Unbiased Predictor (BLUP) estimates²¹⁻²⁴. We further estimate
89 the prediction error variance (PEV) of the effect size estimates to assess the reliability of the
90 variant-level effect sizes. Additionally, using the estimated PEV, we can obtain confidence
91 intervals for each variant-level effect size. For binary traits, we implemented Firth bias
92 correction as a subsequent step. This correction mechanism mitigates bias and rectifies
93 abnormal estimates, especially in scenarios where the case-control ratio is imbalanced.
94 Recognizing the imperative of scalability in large-scale biobank data analyses, we developed a
95 fast implementation of Firth correction, which reduces computation complexity from
96 $O(Mnk_F)$ to $O(n_{nz}k_F)$ where M is the average number of iterations for convergence of Firth
97 corrected beta, n_{nz} is the number of individuals with non-zero genotype, and k_F denotes the
98 number of variants that needs to be corrected (See **Methods**).

99 100 *UK Biobank WES data analysis*

101 We applied our method to five quantitative traits (HDL cholesterol, LDL cholesterol,
102 triglycerides, height, body mass index (BMI)) and five binary traits (breast cancer, prostate
103 cancer, lymphoid leukemia, type 2 diabetes, and coronary atherosclerosis) in UKB. For LDL
104 cholesterol (LDL-C) level, we adjusted the pre-medication levels by dividing the raw LDL-C level
105 by 0.7 for individuals on cholesterol-lowering medication²⁵.

106
107 We computed gene-level effect sizes by leveraging the estimated heritability derived from the
108 mixed effects model. Recognizing the inherent unsigned nature of heritability, we assign a sign
109 by incorporating variant-level effect sizes of loss-of-function (LoF) variants within each gene.
110 This allows us to discern the direction of the effect of the gene on the trait (See **Methods**). As
111 expected, the gene-based association test p-values and the magnitude of the gene-level effect
112 size showed a substantial correlation (**Fig. 2** and **Supplementary Figure 1**). Incorporating gene-

113 level effect size and direction on top of the gene-based association tests can add significant
114 value to genetic analyses. For example, the signed heritability clearly shows that the
115 impairment of APOC3 function increases HDL cholesterol (HDL-C) level but decreases
116 triglycerides (**Fig. 2(a)** and **2(c)**).

117
118 We estimated the variant-level effect sizes of exonic variants and presented two genes, *APOC3*
119 and *SLC12A3*, on HDL-C levels as examples (**Fig. 3**). As expected, variants demonstrating
120 significant associations, in terms of p-values, also exhibited larger effect sizes. Using RareEffect,
121 we observed that the majority of variants in *APOC3* displayed positive effect sizes. The Burden
122 and SKAT-O p-values from SAIGE-GENE+ were both highly significant (Burden p-value =
123 1×10^{-298} and SKAT-O p-value = 1×10^{-300}). In contrast, variants in *SLC12A3* exhibited both
124 positive and negative effect sizes. Consequently, Burden p-value was not significant (Burden
125 p-value = 0.01), but the SKAT-O p-value was significant (Burden p-value = 7×10^{-12}). This
126 distinction in the directionality of effect sizes cannot be discerned through the burden
127 approach.

128
129 We extended our analyses to estimate polygenic risk scores using rare variants with effect
130 sizes estimated from RareEffect (PRS_{RE}). The UK Biobank data were randomly split into
131 training and test sets (ratio = 8:2), and PRS_{RE} was constructed using the genes with p-values
132 $< 2.5 \times 10^{-6}$ from SAIGE-GENE+ in the training set (See **Methods**). We included top 10 genes
133 when the number of genes with p-values $< 2.5 \times 10^{-6}$ is smaller than 10. We further
134 integrated these PRS_{RE} with PRS derived from common variants (PRS_{common}) to evaluate the
135 practical utility of our approach. When combining the PRS from common and rare variants,
136 we constructed a composite score, a linear combination of PRS from common and rare
137 variants. We evaluated the predictive performance in terms of R^2 . Our methods consistently
138 exhibited superior prediction accuracy for all tested quantitative traits, compared to a
139 comparative approach that relied on per-allele effect sizes derived from burden tests (**Fig. 4**,
140 **Supplementary Table 1**, and **Supplementary Figure 2**).

141
142 The improvement became particularly pronounced when predicting lipid phenotypes among
143 individuals deemed at high risk. For instance, when predicting HDL and LDL cholesterol levels,
144 PRS_{RE} achieved R^2 of 0.4737 and 0.6287 when restricting individuals with top/bottom 0.5%

145 in terms of PRS_{RE} . When the composite score was used for risk prediction, R^2 were 0.6237
146 and 0.6645 when restricted top/bottom 0.5% individuals. In contrast, common variants only
147 PRS model had lower R^2 (0.5417 for HDL and 0.5823 for LDL) for top/bottom 0.5% individuals.
148 Notably, the sub-groups identified as high-risk by PRS_{common} and PRS_{RE} were substantially
149 distinct (**Supplementary Figure 3**), underscoring the complementary nature of rare variants in
150 detecting individuals at elevated disease risk. Additionally, we observed that our model
151 showed higher predictability compared to the Burden approach which showed R^2 of 0.3683
152 and 0.2262 for HDL and LDL with top/bottom 0.5% in terms of PRS using burden score
153 (PRS_{burden}), respectively.

154

155 Our method exhibited marginally lower predictive performance for binary traits, as measured
156 by AUC, compared to the burden approach. We observed that for chosen binary traits, there
157 were fewer genes associated with the trait, and their signals appeared weaker when
158 contrasted with tested continuous traits. Nonetheless, our method offers potential benefits,
159 as it can enhance predictability by combining PRS_{RE} with PRS_{common} and PRS_{burden} , which
160 yielded better results. For instance, in the case of lymphoid leukemia, when evaluating
161 individuals in the top/bottom 0.5% based on common PRS, the $PRS_{common} + PRS_{RE} +$
162 PRS_{burden} approach exhibited an AUC of 0.8649, whereas the $PRS_{common} + PRS_{burden}$
163 approach demonstrated an AUC of 0.8559, with the common-only approach yielding an AUC
164 of 0.8559.

165

166 We further examined the relationship between phenotype outliers and the PRS in identifying
167 individuals at high risk for common diseases²⁶. We first defined the phenotype outliers as
168 individuals with phenotype value exceeding a certain z-score cutoff and calculated the
169 proportion of individuals with high PRS among phenotype outliers. Specifically, for LDL
170 cholesterol levels, PRS_{RE} successfully pinpointed individuals at phenotypic extremes, who
171 exhibited a tenfold higher likelihood of possessing a PRS_{RE} falling within 0.1st percentile
172 compared to the baseline population (**Supplementary Figure 4**). PRS_{RE} and PRS_{common}
173 utilize distinct set of variants and show minimal correlation (**Supplementary Table 2**).
174 Therefore, integrating these models into a unified framework enables the identification of a
175 significantly larger cohort at high risk than achievable through PRS_{common} alone.

176

177 *Simulation study*

178 To assess the predictive accuracy of our method, we conducted extensive simulations under
179 diverse scenarios (see **Methods**) for both binary and quantitative traits. To mimic real data,
180 we utilized actual genotypes from the UKB dataset, specifically the array-genotyped data for
181 common variants ($MAF \geq 0.01$) and the UKB WES data for rare variants ($MAF \leq 0.01$).

182

183 We compared the performance of our method against other existing approaches such as
184 linear regression, which is used for standard single-variant association test, or ridge regression
185 in terms of RMSE. For quantitative traits, our method consistently demonstrated a lower RMSE
186 of 0.1703 on average, outperforming the comparative methods which showed RMSE of 0.1770
187 (ridge regression) to 0.1881 (linear regression) (**Supplementary Table 3**) when estimating the
188 effect size. For binary traits, our method also exhibited lower predictive error particularly in
189 scenarios of low disease prevalence compared to ridge regression (**Supplementary Table 4**).
190 Conversely, ridge regression showed marginally reduced error compared to our method in
191 instances of high disease prevalence; however, the difference in accuracy remains modest.

192

193 *Computation and memory cost*

194 Analyzing 166,960 samples from the UKB WES data to estimate the effect size of the *DOCK6*
195 gene, which contains 4,114 rare variants, we observed that the computation time for
196 RareEffect with a simulated phenotype was approximately 90% lower than that of ridge
197 regression. Specifically, RareEffect completed the analysis in 4.2 seconds, compared to 44.6
198 seconds for ridge regression (**Supplementary Figure 5**). The memory usage for RareEffect
199 during the analysis was 1.14GB for *DOCK6* gene (**Supplementary Figure 6**). For binary traits,
200 an additional step of performing Firth bias correction is required. We observed that the
201 normal Firth bias correction process took 708 seconds to analyze a gene with 250 variants
202 (after collapsing) across 342,409 individuals. However, by implementing our fast version of
203 Firth correction, the computation time was dramatically reduced to 2.9 seconds
204 (**Supplementary Figure 7**).

205

206 **Discussion**

207 Our study introduces a novel method aimed at estimating the effect size of rare variants and
208 can be extended to estimate gene-level effect size by employing a two-stage framework of
209 generalized linear mixed models. By leveraging the variant-level effect size estimates obtained
210 through our approach, we can examine the collective impact of rare variants within a gene
211 and quantify their contribution to the overall heritability of complex traits.

212

213 In order to obtain accurate estimates for effect sizes while optimizing computational efficiency,
214 we employed several techniques in our analysis. First, we utilized the FaST-LMM¹⁷ algorithm
215 to expedite computation and reduce memory usage. FaST-LMM leverages the spectral
216 decomposition of the genetic relatedness matrix, allowing for efficient calculation of the
217 variance component in mixed models. Second, we implemented the optimized version of Firth
218 bias correction by utilizing the sparsity of genotype matrix and skipping the computation of
219 hat matrix at every iteration. Third, we employed a collapsing strategy that reduces the
220 sparsity of the data, similar to the approach employed in SAIGE-GENE+. These algorithmic
221 approaches significantly accelerate the estimation process and enhance computational
222 scalability, particularly when dealing with large-scale datasets and complex genetic analyses
223 involving rare variants.

224

225 Beyond its immediate applications in effect size estimation, our proposed method offers
226 significant potential for enhancing polygenic risk prediction models. Traditionally, these
227 models have relied on common variants, often neglecting the valuable insights provided by
228 rare variants. Our analysis reveals that the correlation between PRS_{RE} and phenotype values
229 in the general population is not substantial (**Supplementary Figure 8**). However, we
230 demonstrate that the RareEffect method effectively identifies individuals with high genetic
231 risk. By incorporating our approach, we can substantially improve the predictive accuracy and
232 precision of polygenic risk scores.

233

234 AI-based methods^{27,28} have been developed to enhance the pathogenicity prediction of rare
235 variants. Although these approaches help to identify effect sizes of pathogenic variants and
236 can be used for risk prediction, they may not be as effective for identifying beneficial or gain-
237 of-function variants. Additionally, the performance of these methods can be limited when
238 applied to non-protein-altering variants. Our approach can accommodate the predictively

239 pathogenic variants identified by AI-based models by forming them into a separate category,
240 thereby enhancing performance.

241

242 Our study, however, is not without limitations. While RareEffect demonstrates comparable or
243 superior performance in estimating the effect size for simulated binary phenotypes, our
244 evaluation revealed only marginal enhancements in predictive performance, as measured by
245 the area under the curve (AUC), compared to the traditional common PRS across tested
246 disease phenotypes. This could be attributed to the trait's reliance on a limited number of
247 ultra-rare variants, which our method collapses into super-variants, thereby complicating the
248 estimation of variant-level effect sizes for true causal variants. Despite its efficacy in effect size
249 estimation, RareEffect's ability to improve prediction accuracy remains constrained,
250 highlighting a potential area for future refinement and investigation for risk prediction of
251 binary traits. Additionally, it is important to note that RareEffect is based on BLUP estimate,
252 characterized by shrinkage properties, leading to biased estimates of effect sizes. However,
253 despite this bias, RareEffect provides more stable estimates compared to unbiased methods
254 like linear regression, which tend to be unstable for rare variant analysis due to the low allele
255 frequency inherent in such variants.

256

257 In summary, our results demonstrate that incorporating information from rare variants
258 enables the accurate estimation of gene-level and variant-level effect sizes, as well as the
259 identification of high-risk individuals who might remain undetected by conventional polygenic
260 risk scores (PRS) methods relying on common variants.

261

262

263 **Method**

264 *Generalized linear mixed model*

265 We denote the phenotype of the i th individual using y_i for both quantitative and binary traits
266 in a study with sample size n . \mathbf{X} represents the $n \times (p + 1)$ vector with p covariates including
267 the intercept and \mathbf{G}_j is the $n \times k$ matrix representing the minor allele counts for k rare
268 variants in gene or region j . The generalized linear mixed model (GLMM) can be expressed as:

269

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\alpha} + \mathbf{G}_j\boldsymbol{\beta}_j + \mathbf{b}$$

270 where $\boldsymbol{\mu}$ is the mean phenotype, $\mathbf{b} \sim MVN(0, \sigma_g^2 \mathbf{K})$ is the random effect, and \mathbf{K} is an $n \times n$
271 genetic relatedness matrix (GRM). And g is the link function which is an identity function for
272 quantitative traits with error term $\boldsymbol{\epsilon} \sim MVN(0, \sigma_e^2 \mathbf{I}_n)$ and a logit function for binary traits. The
273 parameter $\boldsymbol{\alpha}$ is a $(p + 1) \times 1$ vector of fixed effect coefficients and $\boldsymbol{\beta}_j$ is a $k \times 1$ vector of the
274 random genetic effect.

275

276 *Fitting the null generalized mixed model (step 1)*

277 We used the average information restricted maximum likelihood (AI-REML) algorithm to fit
278 the null GLMM (i.e., $H_0: \boldsymbol{\beta}_j = \mathbf{0}$) as in SAIGE step 1.

279

280 *Estimation of the gene-level (region-level) heritability (step 2)*

281 We estimate the effect size of rare variants using the following linear mixed model:

282

$$\tilde{\mathbf{y}} = \mathbf{G}_j \boldsymbol{\beta}_j + \boldsymbol{\epsilon}$$

283 where $\tilde{\mathbf{y}} = (\tilde{y}_1, \dots, \tilde{y}_n)^T = \mathbf{y} - \hat{\mathbf{y}}$ for quantitative traits, $\tilde{y}_i = \frac{d\eta_i}{d\mu_i} (y_i - \hat{\mu}_i) = \frac{1}{\hat{\mu}_i(1-\hat{\mu}_i)} (y_i -$
284 $\hat{\mu}_i)$, a working residual from iteratively reweighted least squares (IRWLS) for binary traits, and
285 $\hat{\mu}_i$ is the mean phenotype for individual i , which can be obtained from step 1. When obtaining
286 \mathbf{G}_j , as in SAIGE-GENE+, we collapsed ultra-rare variants with minor allele count (MAC) ≤ 10 by
287 each gene and functional group to reduce the sparsity. We further implemented an option to
288 collapse all loss-of-function (LoF) rare variants into a single column, irrespective of their minor
289 allele count, adopting the burden approach. This approach is predicated on the assumption
290 that all rare LoF variants share a common effect size and direction. For binary traits, when we
291 use working residuals, the variance of \tilde{y}_i differs by individual, so we cannot apply the
292 optimization technique of FaST-LMM. Therefore, we divide both sides by the square root of
293 variance of \tilde{y}_i to make the variance be the same across individuals. We estimate the effect
294 size using the modified model, for binary traits:

295

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-\frac{1}{2}} \tilde{\mathbf{y}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-\frac{1}{2}} \mathbf{G}_j \boldsymbol{\beta}_j + \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-\frac{1}{2}} \boldsymbol{\epsilon}$$

296 where $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \text{diag}\left(\frac{1}{\hat{\mu}_i(1-\hat{\mu}_i)}\right)$.

297

298 In this model, the prior distribution of $\boldsymbol{\beta}_j$ are assumed to follow $MVN(0, \tau\boldsymbol{\Sigma})$, and the noise $\boldsymbol{\epsilon}$

299 is assumed to follow $MVN(0, \psi\mathbf{I}_n)$ for quantitative traits, while we assume $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-\frac{1}{2}} \boldsymbol{\epsilon}$ follows

300 $MVN(0, \psi \mathbf{I}_n)$ for binary traits. When there is no prior knowledge of the correlation within $\boldsymbol{\beta}$,
 301 $\boldsymbol{\Sigma}$ is set to be an identity matrix. But in general, $\boldsymbol{\Sigma}$ does not have to be an identity or a diagonal
 302 matrix.

303

304 To estimate the variance component parameters (τ, ψ) , we use factored spectrally
 305 transformed linear mixed models (FaST-LMM) algorithm. Let $\tilde{\mathbf{G}} = [\tilde{\mathbf{G}}_1, \dots, \tilde{\mathbf{G}}_k]$ be an $n \times k$
 306 genotype matrix of the region with $\tilde{\mathbf{G}}_j = \mathbf{G}_j$ for quantitative traits and $\tilde{\mathbf{G}}_j = \hat{\boldsymbol{\Sigma}}_\epsilon^{-\frac{1}{2}} \mathbf{G}_j$ for binary
 307 traits. The variance of $\tilde{\mathbf{y}}$ (quantitative traits) or $\hat{\boldsymbol{\Sigma}}_\epsilon^{-\frac{1}{2}} \tilde{\mathbf{y}}$ (binary traits) can be written as

308

$$\tau \tilde{\mathbf{G}} \boldsymbol{\Sigma} \tilde{\mathbf{G}}^T + \psi \mathbf{I}_n$$

309

310 Traditional approaches to estimate the variance components require either calculating inverse
 311 matrix or conducting spectral decomposition of the $n \times n$ matrix $\tilde{\mathbf{G}} \boldsymbol{\Sigma} \tilde{\mathbf{G}}^T$, so the time
 312 complexity is of $O(n^3)$. In contrast, FaST-LMM algorithm uses the fact that $\tilde{\mathbf{G}} \boldsymbol{\Sigma} \tilde{\mathbf{G}}^T$ has rank at
 313 most k , so to reduce the computation complexity. Suppose $\mathbf{Z} = \tilde{\mathbf{G}} \mathbf{L}$ is an $n \times k$ matrix where
 314 \mathbf{L} is a Cholesky factor of $\boldsymbol{\Sigma}$. Then $\tilde{\mathbf{G}} \boldsymbol{\Sigma} \tilde{\mathbf{G}}^T = \mathbf{Z} \mathbf{Z}^T$. FaST-LMM carries out singular value
 315 decomposition on \mathbf{Z} and calculate likelihood for (τ, ψ) . With given $n \gg k$, calculation of
 316 \mathbf{Z} and its singular value decomposition requires only $O(nk^2)$ of time complexity. And we
 317 further improved the computation efficiency utilizing the sparsity of \mathbf{Z}^{29} . In biobank-scale data,
 318 n is in the hundreds of thousands, and k is in the tens to hundreds on average which means
 319 $n \gg k$.

320

321 Using the above optimization technique, we estimate the variance components by each group.
 322 Consider one group (LoF, missense or synonymous) in a single gene j in the model:

323

$$\tilde{\mathbf{y}} = \mathbf{G}_{\cdot, j} \boldsymbol{\beta}_{\cdot, j} + \boldsymbol{\epsilon}$$

324

325 We first marginally estimate the maximum-likelihood estimator (MLE) of variance
 326 components $\tau_{LoF, j}$, $\tau_{mis, j}$, and $\tau_{syn, j}$. As the marginal estimates do not account for LD among
 327 variants in different groups, we adjust the estimates using method of moments (MoM)
 approaches³⁰. The MoM estimator can be obtained by solving the following linear system:

328

$$\begin{bmatrix} \mathbf{T} & \mathbf{b} \\ \mathbf{b}^T & n \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\tau}} \\ \hat{\psi} \end{bmatrix} = \begin{bmatrix} \mathbf{c} \\ \tilde{\mathbf{y}}^T \tilde{\mathbf{y}} \end{bmatrix}$$

329 where \mathbf{T} is a 3×3 matrix for joint estimation or a scalar (1×1 matrix) for marginal
 330 estimation with entries $T_{k,l} = \text{tr}(\tilde{\mathbf{G}}_k \boldsymbol{\Sigma}_k \tilde{\mathbf{G}}_k^T \tilde{\mathbf{G}}_l \boldsymbol{\Sigma}_l \tilde{\mathbf{G}}_l^T)$ where $k, l \in \{1, 2, 3\}$ for joint estimation.
 331 \mathbf{b} is a 3-vector for joint estimation or a scalar for marginal estimation with entries $b_k =$
 332 $\text{tr}(\tilde{\mathbf{G}}_k \boldsymbol{\Sigma}_k \tilde{\mathbf{G}}_k^T)$, \mathbf{c} is a 3-vector for joint estimation or a scalar for marginal estimation with entries
 333 $c_k = \tilde{\mathbf{y}}^T \tilde{\mathbf{G}}_k \boldsymbol{\Sigma}_k \tilde{\mathbf{G}}_k^T \tilde{\mathbf{y}}$, and $\hat{\boldsymbol{\tau}} = \begin{bmatrix} \hat{\tau}_1 \\ \hat{\tau}_2 \\ \hat{\tau}_3 \end{bmatrix}$ where (τ_1, τ_2, τ_3) denotes $(\tau_{LoF,j}, \tau_{mis,j}, \tau_{syn,j})$ for a single
 334 gene j , respectively.

335

336 After estimating the marginal and joint MoM estimator of variance components, we adjust
 337 the MLE of the marginal variance components by:

$$338 \quad \hat{\tau}_{i,j} = \hat{\tau}_{MLE,mar,i,j} \times \frac{\hat{\tau}_{MoM,joint,i,j}}{\hat{\tau}_{MoM,mar,i,j}}$$

339 where $i \in \{LoF, mis, syn\}$. The MoM estimator of the variance component was not directly
 340 utilized in our study since the MoM estimator could yield negative variance components. In
 341 cases of the variance component estimated by the MoM is negative, we used marginal
 342 variance component without adjustment. Both MoM and likelihood-based approaches
 343 exhibited the similar trends of variance components (**Supplementary Figure 9**). Additionally,
 344 we assume that the variance explained by rare variants in a single gene is negligibly small
 345 compared to the total variance of $\tilde{\mathbf{y}}$, therefore, $\hat{\boldsymbol{\psi}} \approx \text{Var}(\tilde{\mathbf{y}})$.

346

347 We estimated the heritability from rare variants of gene j using these adjusted variance
 348 components. In a joint model,

$$349 \quad \text{Var}(\tilde{\mathbf{y}}) = \mathbf{G}_j \boldsymbol{\Sigma}_j \mathbf{G}_j^T + \psi \mathbf{I}_n$$

350 where $\mathbf{G}_j = [\mathbf{G}_{LoF,j} \quad \mathbf{G}_{mis,j} \quad \mathbf{G}_{syn,j}]$ and $\boldsymbol{\Sigma}_j = \begin{bmatrix} \hat{\tau}_{LoF,j} \boldsymbol{\Sigma}_{LoF,j} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \hat{\tau}_{mis,j} \boldsymbol{\Sigma}_{mis,j} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \hat{\tau}_{syn,j} \boldsymbol{\Sigma}_{syn,j} \end{bmatrix}$.

351 Therefore,

$$352 \quad \sum_{i=1}^n \text{Var}(\tilde{\mathbf{y}}_i) = \text{tr}(\mathbf{G}_j \boldsymbol{\Sigma}_j \mathbf{G}_j^T) + n\psi \quad (\text{for quantitative traits})$$

$$353 \quad \sum_{i=1}^n \text{Var}(\tilde{\mathbf{y}}_i) = \text{tr}(\mathbf{G}_j \boldsymbol{\Sigma}_j \mathbf{G}_j^T) + n \text{tr}(\hat{\boldsymbol{\Sigma}}_\epsilon) \quad (\text{for binary traits})$$

354

355 Subsequently, the heritability from rare variants of gene j can be denoted as:

356
$$h_j^2 = \frac{\text{tr}(\mathbf{G}_j \boldsymbol{\Sigma}_j \mathbf{G}_j^T)}{\text{tr}(\mathbf{G}_j \boldsymbol{\Sigma}_j \mathbf{G}_j^T) + n\psi} \quad (\text{for quantitative traits})$$

357
$$h_j^2 = \frac{\text{tr}(\mathbf{G}_j \boldsymbol{\Sigma}_j \mathbf{G}_j^T)}{\text{tr}(\mathbf{G}_j \boldsymbol{\Sigma}_j \mathbf{G}_j^T) + \psi \text{tr}(\tilde{\boldsymbol{\Sigma}}_\epsilon)} \quad (\text{for binary traits})$$

358 Additionally, to determine the direction of the gene-level effect, we obtained the sign of the
359 linear combination of the effect sizes of loss-of-function variants in a gene, weighted by their
360 MAFs:

361
$$(\text{signed heritability of gene } j) = h_j^2 \times \text{sgn}(\sum_{j \in \text{LoF}} \beta_j \text{MAF}_j)$$

362 This measure gives the information of the magnitude of genetic effects from rare variants in a
363 single gene and its direction of effects.

364

365 *Estimation of the variant-level effect size (step 3)*

366 The effect sizes at variant-level resolution are estimated using the adjusted variance
367 components in the previous step by:

368
$$\hat{\boldsymbol{\beta}}_j = (\tilde{\mathbf{G}}_j^T \tilde{\mathbf{G}}_j + \hat{\psi} \boldsymbol{\Sigma}_j^{-1})^{-1} \tilde{\mathbf{G}}_j^T \tilde{\mathbf{y}}$$

369 for each gene or region.

370

371 We further estimated the prediction error variance (PEV) by:

372
$$PEV(\hat{\boldsymbol{\beta}}_j) = \left(\tilde{\mathbf{G}}_j^T \tilde{\mathbf{G}}_j + \frac{\hat{\psi}}{\hat{\tau}} \mathbf{I} \right)^{-1}$$

373 for each gene or region. Using this PEV, we can obtain confidence intervals for effect sizes.

374

375 For binary traits, the Firth bias correction³¹ is a more accurate method to estimate SNP effect
376 sizes³²⁻³⁴, particularly in cases marked by a significant case-control imbalance. We incorporate
377 this correction into our analytical framework. Additionally, we introduce an L2 penalty term
378 to account for the prior distribution of $\boldsymbol{\beta}$. The Firth corrected effect estimates can be
379 calculated numerically by optimizing the following objective function:

380
$$\hat{\boldsymbol{\beta}}_* = \text{argmax}_{\boldsymbol{\beta}} [\log L(\boldsymbol{\beta}) + \frac{1}{2} \log |I(\boldsymbol{\beta})| - \frac{1}{2\hat{\tau}} \|\boldsymbol{\beta}\|_2^2]$$

381 where L denotes the likelihood function and I is the Fisher information.

382

383 To improve computational efficiency, we developed the fast implementation of Firth bias
384 correction. First, we compute the hat matrix, $\mathbf{H} = \mathbf{W}^{\frac{1}{2}} \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{\frac{1}{2}}$, $\mathbf{W} = \text{diag}(\hat{\mu}_i(1 -$
385 $\hat{\mu}_i))$, only once rather than recalculating it at each iteration, assuming that the iterative
386 weights (\mathbf{W}) change slowly as a function of the mean μ_i ^{20,35}. We compared, using both
387 simulation and real data, the difference in effect size estimates between computing the hat
388 matrix once and computing it at each iteration. Our findings indicate minimal differences in
389 the effect size estimates (**Supplementary Figure 10**). Given that Firth correction is applied to
390 each variant, we iterate the hat matrix calculation for k_F times, where k_F represents the
391 number of variants that need to be corrected. This strategy results in a reduction in the
392 computational complexity with Firth correction from $O(Mnk_F)$ to $O(nk_F)$, where M is the
393 average number of iterations needed for Firth corrected effect size convergence. Second, we
394 extend Firth correction to accommodate sparse genotype matrices. Specifically, we restrict
395 our computations to individuals with non-zero genotypes when determining the score and
396 Fisher information. This further reduces the time complexity of the Firth correction, now at
397 $O(n_{nz}k_F)$, where n_{nz} denotes the number of individuals with non-zero genotype.
398 Considering that we are estimating the effect size of rare variants with $\text{MAF} < 0.01$, we
399 observed that $n_{nz} < 0.01n$ which implies that leveraging sparsity makes the estimation of
400 effect sizes more than 100 times faster compared to a non-sparse approach.

401

402 We note that Firth correction is used when the absolute value of estimated effect size
403 ($\hat{\beta}$) surpasses a predefined threshold. In this study, we used threshold of $\log 2 \approx 0.693$ for
404 simulation studies and real data analysis.

405

406 *Rare-variant PRS calculation*

407 We further calculated polygenic risk scores using the variant-level effect sizes of rare variants
408 (PRS_{RE}) in genes with gene-level p-value from SAIGE-GENE+ lower than 2.5×10^{-6} . PRS_{RE}
409 of individual i can be calculated as:

$$410 \quad PRS_{RE,i} = \sum_{j \in J} \mathbf{G}_{ij} \hat{\beta}_j$$

411 where J denotes a set of genes with gene-level p-value lower than 2.5×10^{-6} . Additionally,
412 we combined these PRS_{RE} with PRS_{common} only. We applied PRS-CS³⁶ to obtain the variant-

413 level weights for the calculation the PRS_{common} . We compared the predictive performance
414 of the PRS_{RE} to PRS_{burden} . PRS_{burden} were obtained in two different ways: (1) collapsing all
415 rare variants into one super-variant regardless of the functionality of the variants and (2)
416 collapsing rare variants by functionality of the variants (LoF, missense, and synonymous). After
417 collapsing, we fitted the following linear model to estimate the per-allele effect sizes:

$$418 \quad \tilde{y} = \mathbf{G}_{burden,j} \boldsymbol{\beta}_{burden,j}$$

419 The burden PRS are also calculated as a linear combination of per-allele effect sizes and the
420 collapsed genotypes of each individual.

421

422 We compared the predictive performance in terms of R^2 for quantitative traits, and the area
423 under receiver operating characteristic curve (AUROC) for binary traits of the following linear
424 models:

$$425 \quad \tilde{y} \sim PRS_{composite}$$

$$426 \quad \tilde{y} \sim PRS_{common} + PRS_{RE}$$

$$427 \quad \tilde{y} \sim PRS_{common} + PRS_{burden}$$

428 where $PRS_{composite}$ is a linear combination of PRS_{common} and PRS_{RE} with weights trained
429 from the training set.

430

431 *UK Biobank data analysis*

432 In this study, we used WES data of 393,247 White British participants in the UK Biobank. The
433 UK Biobank is a UK-based prospective cohort of ~500,000 individuals aged 40 to 69 at
434 enrollment. We split the train and test data 8:2 randomly for the PRS evaluation. We applied
435 quality control (QC) procedures prior to the analysis. We first removed redundant samples
436 and individuals with sex mismatch or sex chromosome aneuploidy. Additionally, we further
437 removed variants with a missingness rate across individuals > 0.1 , HWE p-value $< 10^{-15}$, and
438 monomorphic variants. We generated group files, which define the list of variants in genes
439 and its functional annotation, by using the loss-of-function transcript effect estimator
440 (LOFTEE)³⁷. We regarded a variant as loss-of-function (LoF) only in case of it is labeled as a
441 high-confidence (HC) LoF variant, and variants with low-confidence (LC) were regarded as
442 missense variants.

443

444 Using the data after QC, we applied our method to five quantitative traits (HDL cholesterol,
445 LDL cholesterol, triglycerides, height, and body mass index) and five binary traits (breast
446 cancer, prostate cancer, lymphoid leukemia, type 2 diabetes, and coronary atherosclerosis).
447 We defined the disease by mapping ICD-10 codes to Phecodes using the PheWAS R package³⁸.

448

449 *Simulation study*

450 To generate outcome phenotypes, we used the following model for quantitative and binary
451 traits:

$$452 \quad y_i = X_{i1} + X_{i2} + G_{i,common}\beta_{common} + G_{i,rare}\beta_{rare} + \epsilon$$

$$453 \quad \text{logit}(P(Y_i = 1)) = \alpha + X_{i1} + X_{i2} + G_{i,common}\beta_{common} + G_{i,rare}\beta_{rare}$$

454 where X_{i1} and X_{i2} are covariates, and $G_{i,common}$ and $G_{i,rare}$ are genotype vectors of common
455 variants and rare variants of i th individual, respectively. The intercept α for binary traits is
456 determined by the disease prevalence. The covariates X_{i1} and X_{i2} were simulated from
457 Bernoulli(0.5) and $N(0, 1)$, respectively. For common variant effect, we randomly selected
458 $L = 30,000$ LD-pruned common variants with $MAF > 1\%$, and assumed that the effect size of
459 single common variant follows $N(0, \frac{1}{L})$. We selected 10 causal genes in UKB WES 200k data for
460 generation of phenotypes. We used eight different scenarios regarding rare variants: (1)
461 proportion of causal variants, (2) effect size of causal variants, and (3) direction of effect within
462 a single gene. For the proportion of causal variants, we assumed (1) 20% of LoF, 10% of
463 missense, and 2% of synonymous variants, or (2) 30% of LoF, 10% of missense, and 2% of
464 synonymous variants among rare variants that are not ultra-rare are causal. For ultra-rare
465 variants, we assumed that the proportion of causal variants are three times higher than the
466 non-ultra-rare variants, that is, (1) 60% of LoF, 30% of missense, and 6% of synonymous ultra-
467 rare variants, or (2) 90% of LoF, 30% of missense, and 6% of synonymous ultra-rare variants
468 are causal. Regarding the effect size of causal variants, we assumed that the absolute effect
469 sizes of causal variants are (1) $|0.5 \log_{10} MAF|$ for LoF variants, and $|0.25 \log_{10} MAF|$ for
470 missense and synonymous variants, or (2) $|0.3 \log_{10} MAF|$ for LoF variants, and
471 $|0.15 \log_{10} MAF|$ for missense and synonymous variants. We further assumed that the effect
472 directions are (1) the same among all causal variants in a single gene, or (2) same for 100% of
473 LoF, 80% of missense, and 50% of synonymous variants, while remaining variants have the

474 opposite direction of effect. For eight combinations of scenarios, we repeated the simulation
475 for 100 times.

476

477 *Computation cost evaluation*

478 We evaluated the computation time and memory usage using simulated data as described
479 above, comprising 166,960 individuals of White British ancestry from the UKB WES 200k
480 dataset. Additionally, we examined computation time and memory usage across subsets with
481 sample sizes of 10k, 30k, 50k, and 100k. For each generative scenario, we reported the mean
482 of 5 attempts for computation times and memory usage, comparing them with multiple linear
483 regression, simple linear regression (as in GWAS), and ridge regression. The evaluation for
484 linear regression and ridge regression was done using `lm` and `glm` functions in R, respectively.

485

486 **Data availability**

487 The analysis results for 5 quantitative and 5 binary phenotypes of UKB WES data analysis
488 results are available at: [https://storage.googleapis.com/leelabsg/RareEffect/RareEffect_](https://storage.googleapis.com/leelabsg/RareEffect/RareEffect_effect_size.zip)
489 [effect_size.zip](https://storage.googleapis.com/leelabsg/RareEffect/RareEffect_effect_size.zip) (variant-level effect size) and [https://storage.googleapis.com/leelabsg/](https://storage.googleapis.com/leelabsg/RareEffect/RareEffect_h2.zip)
490 [RareEffect/RareEffect_h2.zip](https://storage.googleapis.com/leelabsg/RareEffect/RareEffect_h2.zip) (gene-level signed heritability).

491

492 **Code availability**

493 RareEffect is implemented as a part of SAIGE software, which is an open-source R package,
494 available at <https://github.com/saigegit/SAIGE>. RareEffect is available in SAIGE version 1.3.7
495 or higher.

496

497 **Author Contribution**

498 K.N. and S.L. designed experiments. K.N. performed experiments and analyzed the UKB WES
499 data. K.N. and S.L. implemented the software with input from W.Z.. M.K. and B.M. provided
500 helpful advice. K.N. and S.L. wrote the manuscript with input from all co-authors.

501

502 **Acknowledgements**

503 This research was supported by the Brain Pool Plus (BP+) Program through the National
504 Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT
505 (2020H1D3A2A03100666) and the grants funded by the Ministry of Food and Drug Safety,
506 Republic of Korea (Grant Number: 23212MFDS202). This research was conducted using the
507 UK Biobank Resource under application number 45227.

508

509

510 **References**

- 511 1. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data.
512 *Nature* **562**, 203-209 (2018).
- 513 2. Gibson, G. Rare and common variants: twenty arguments. *Nat Rev Genet* **13**, 135-45
514 (2012).
- 515 3. Kryukov, G.V., Shpunt, A., Stamatoyannopoulos, J.A. & Sunyaev, S.R. Power of deep, all-
516 exon resequencing for discovery of human trait genes. *Proc Natl Acad Sci U S A* **106**,
517 3871-6 (2009).
- 518 4. MacArthur, D.G. *et al.* A systematic survey of loss-of-function variants in human
519 protein-coding genes. *Science* **335**, 823-8 (2012).

- 520 5. Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human
521 genomes. *Nature* **491**, 56-65 (2012).
- 522 6. Rivas, M.A. *et al.* Deep resequencing of GWAS loci identifies independent rare variants
523 associated with inflammatory bowel disease. *Nat Genet* **43**, 1066-73 (2011).
- 524 7. Saint Pierre, A. & Génin, E. How important are rare variants in common disease?
525 *Briefings in Functional Genomics* **13**, 353-361 (2014).
- 526 8. Perrone, F., Cacace, R., van der Zee, J. & Van Broeckhoven, C. Emerging genetic
527 complexity and rare genetic variants in neurodegenerative brain diseases. *Genome*
528 *Med* **13**, 59 (2021).
- 529 9. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**,
530 747-53 (2009).
- 531 10. Lee, S., Abecasis, G.R., Boehnke, M. & Lin, X. Rare-variant association analysis: study
532 designs and statistical tests. *Am J Hum Genet* **95**, 5-23 (2014).
- 533 11. Chen, W., Coombes, B.J. & Larson, N.B. Recent advances and challenges of rare variant
534 association analysis in the biobank sequencing era. *Front Genet* **13**, 1014947 (2022).
- 535 12. Wu, M.C. *et al.* Rare-variant association testing for sequencing data with the sequence
536 kernel association test. *Am J Hum Genet* **89**, 82-93 (2011).
- 537 13. Lee, S. *et al.* Optimal unified approach for rare-variant association testing with
538 application to small-sample case-control whole-exome sequencing studies. *Am J Hum*
539 *Genet* **91**, 224-37 (2012).
- 540 14. Li, X. *et al.* Dynamic incorporation of multiple in silico functional annotations
541 empowers rare variant association analysis of large whole-genome sequencing studies
542 at scale. *Nat Genet* **52**, 969-983 (2020).
- 543 15. Zhou, W. *et al.* Scalable generalized linear mixed model for region-based association
544 tests in large biobanks and cohorts. *Nat Genet* **52**, 634-639 (2020).
- 545 16. Zhou, W. *et al.* SAIGE-GENE+ improves the efficiency and accuracy of set-based rare
546 variant association tests. *Nat Genet* **54**, 1466-1469 (2022).
- 547 17. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nat*
548 *Methods* **8**, 833-5 (2011).
- 549 18. Gilmour, A.R., Thompson, R. & Cullis, B.R. Average information REML: An efficient
550 algorithm for variance parameter estimation in linear mixed models. *Biometrics* **51**,
551 1440-1450 (1995).
- 552 19. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample
553 relatedness in large-scale genetic association studies. *Nat Genet* **50**, 1335-1341 (2018).
- 554 20. Chen, H. *et al.* Control for Population Structure and Relatedness for Binary Traits in
555 Genetic Association Studies via Logistic Mixed Models. *Am J Hum Genet* **98**, 653-66
556 (2016).
- 557 21. Robinson, G.K. That BLUP is a Good Thing: The Estimation of Random Effects. *Statistical*
558 *Science* **6**, 15-32, 18 (1991).
- 559 22. Henderson, C.R. Best Linear Unbiased Estimation and Prediction under a Selection
560 Model. *Biometrics* **31**, 423-447 (1975).
- 561 23. Maier, R.M. *et al.* Improving genetic prediction by leveraging genetic correlations
562 among human diseases and traits. *Nat Commun* **9**, 989 (2018).
- 563 24. Chen, C.Y., Han, J., Hunter, D.J., Kraft, P. & Price, A.L. Explicit Modeling of Ancestry
564 Improves Polygenic Risk Scores and BLUP Prediction. *Genet Epidemiol* **39**, 427-38
565 (2015).

- 566 25. Graham, S.E. *et al.* The power of genetic diversity in genome-wide association studies
567 of lipids. *Nature* **600**, 675-679 (2021).
- 568 26. Fiziev, P.P. *et al.* Rare penetrant mutations confer severe risk of common diseases.
569 *Science* **380**, eabo1131 (2023).
- 570 27. Sundaram, L. *et al.* Predicting the clinical impact of human mutation with deep neural
571 networks. *Nat Genet* **50**, 1161-1170 (2018).
- 572 28. Gao, H. *et al.* The landscape of tolerated genetic variation in humans and primates.
573 *Science* **380**, eabn8153 (2023).
- 574 29. Berry, M.W. Large-Scale Sparse Singular Value Computations. *International Journal of*
575 *Supercomputer Applications and High Performance Computing* **6**, 13-49 (1992).
- 576 30. Pazokitoroudi, A. *et al.* Efficient variance components analysis across millions of
577 genomes. *Nat Commun* **11**, 4020 (2020).
- 578 31. Firth, D. Bias Reduction of Maximum Likelihood Estimates. *Biometrika* **80**, 27-38 (1993).
- 579 32. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for
580 quantitative and binary traits. *Nat Genet* **53**, 1097-1103 (2021).
- 581 33. Wang, X. Firth logistic regression for rare variant association tests. *Front Genet* **5**, 187
582 (2014).
- 583 34. Dey, R., Schmidt, E.M., Abecasis, G.R. & Lee, S. A Fast and Accurate Algorithm to Test
584 for Binary Phenotypes and Its Application to PheWAS. *Am J Hum Genet* **101**, 37-49
585 (2017).
- 586 35. Breslow, N.E. & Clayton, D.G. Approximate Inference in Generalized Linear Mixed
587 Models. *Journal of the American Statistical Association* **88**, 9-25 (1993).
- 588 36. Ge, T., Chen, C.Y., Ni, Y., Feng, Y.A. & Smoller, J.W. Polygenic prediction via Bayesian
589 regression and continuous shrinkage priors. *Nat Commun* **10**, 1776 (2019).
- 590 37. Karczewski, K.J. *et al.* The mutational constraint spectrum quantified from variation in
591 141,456 humans. *Nature* **581**, 434-443 (2020).
- 592 38. Wu, P. *et al.* Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow
593 Development and Initial Evaluation. *JMIR Med Inform* **7**, e14325 (2019).
- 594
- 595

596 **Figures Legends**

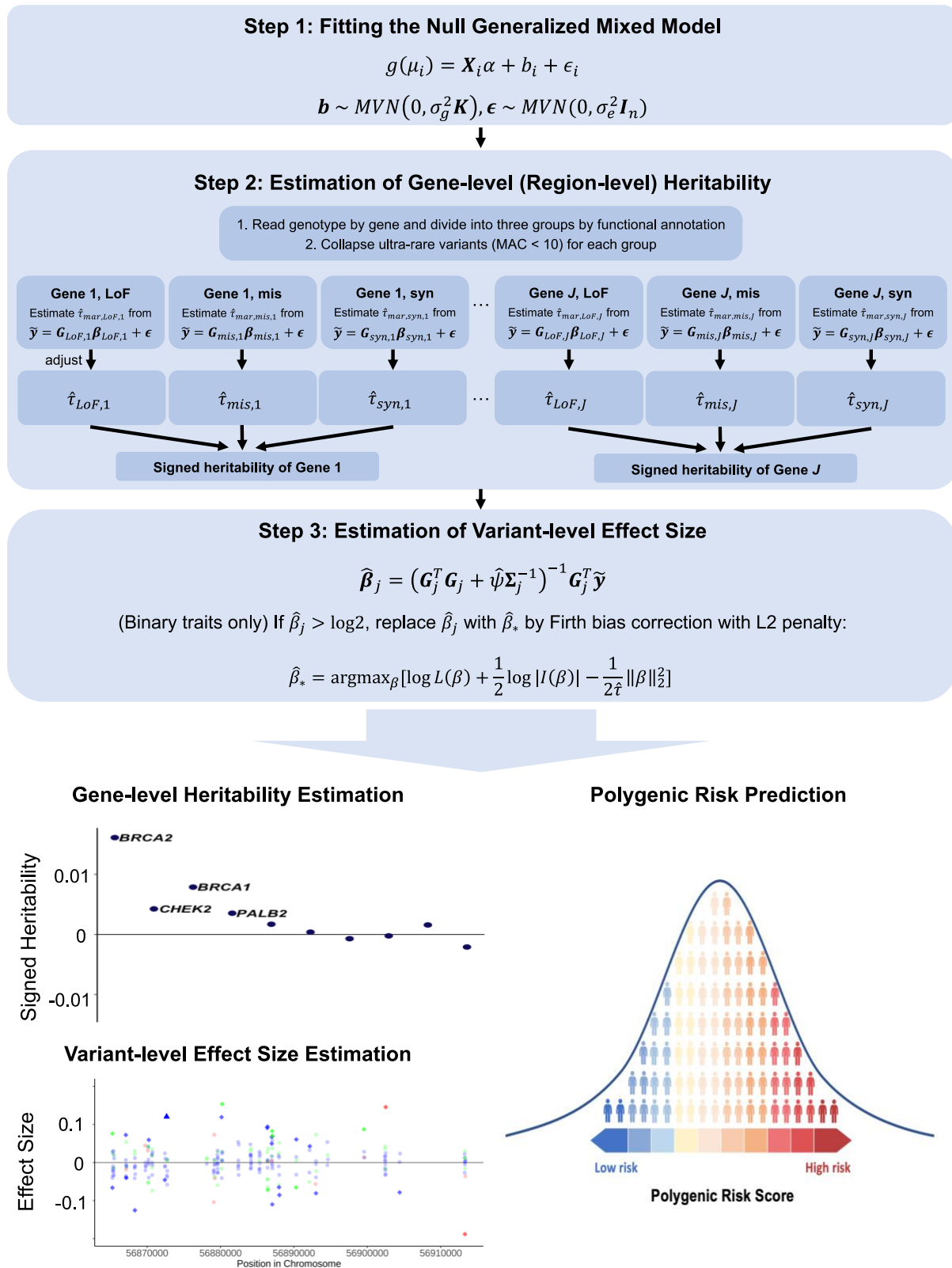
597 **Fig. 1. Overview of RareEffect.**

598 **Fig. 2. Estimated signed heritability for lipid phenotypes using 392,748 White British**
599 **samples in UK Biobank whole exome sequencing data**

600 **Fig. 3. Variant-level effect size on HDL cholesterol for variants in *APOC3* gene (chromosome**
601 **11) and *SLC12A3* gene (chromosome 16)**

602 **Fig. 4. Comparison of performance of risk prediction models for lipid phenotypes using**
603 **314,198 White British samples (80% of the whole White British samples) in UK Biobank**
604 **whole exome sequencing data**

605 **Fig. 1. Overview of RareEffect.**



606

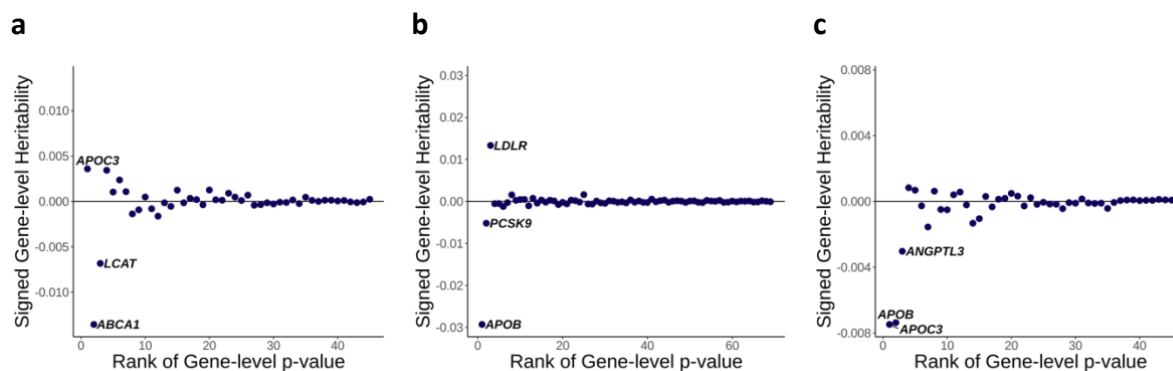
607 RareEffect encompasses three steps. In step 1, we fit a null GLMM using AI-REML approach,

608 and obtain residuals for the subsequent steps. In step 2, we divide variants by gene and its

609 functional annotation (LoF, missense, and synonymous). We first estimate the variance

610 component of each group and adjust them using MoM approach. In step 3, we calculate the
611 variant-level effect size using BLUP estimates. For binary traits, Firth bias correction is
612 additionally applied to adjust the case-control imbalance. Through RareEffect, we provide
613 region-level and variant-level effect sizes for enhanced interpretability and improved risk
614 prediction performance.
615

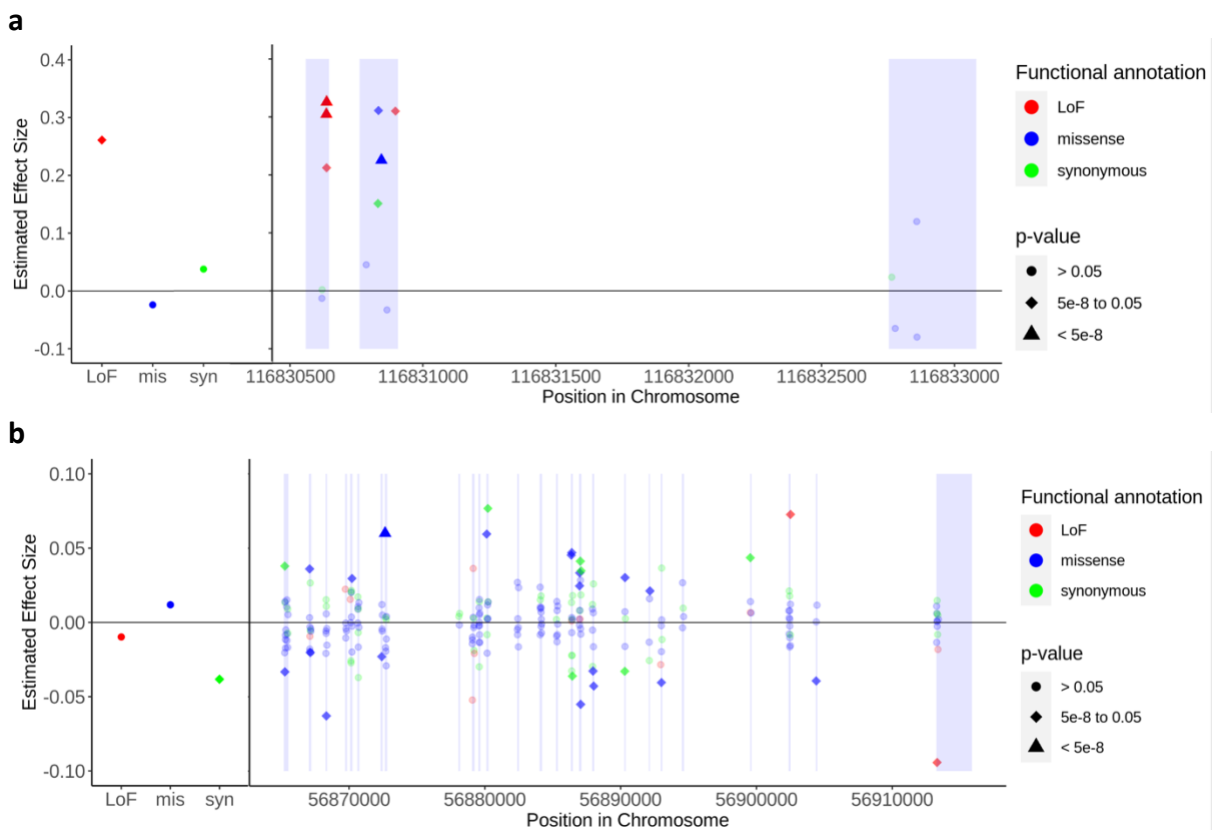
616 **Fig. 2. Estimated signed heritability for lipid phenotypes using 392,748 White British**
617 **samples in UK Biobank whole exome sequencing data**
618



619 Signed gene-level heritability from RareEffect for (a) HDL cholesterol level, (b) LDL
620 cholesterol level, and (c) triglycerides level. Gene-level p-values were obtained from SAIGE-
621 GENE+, and we included genes with p-values $< 2.5 \times 10^{-6}$. The x-axis represents the rank
622 order of genes based on their gene-level p-values. Lower ranks correspond to genes with
623 more significant p-values. The y-axis shows the signed gene-level heritability for each gene.
624 Signed heritability indicates the direction (positive or negative) and magnitude of the genetic
625 contribution of the gene to the trait.

626 **Fig. 3. Variant-level effect size on HDL cholesterol for variants in *APOC3* gene (chromosome**
627 **11) and *SLC12A3* gene (chromosome 16).**

628
629



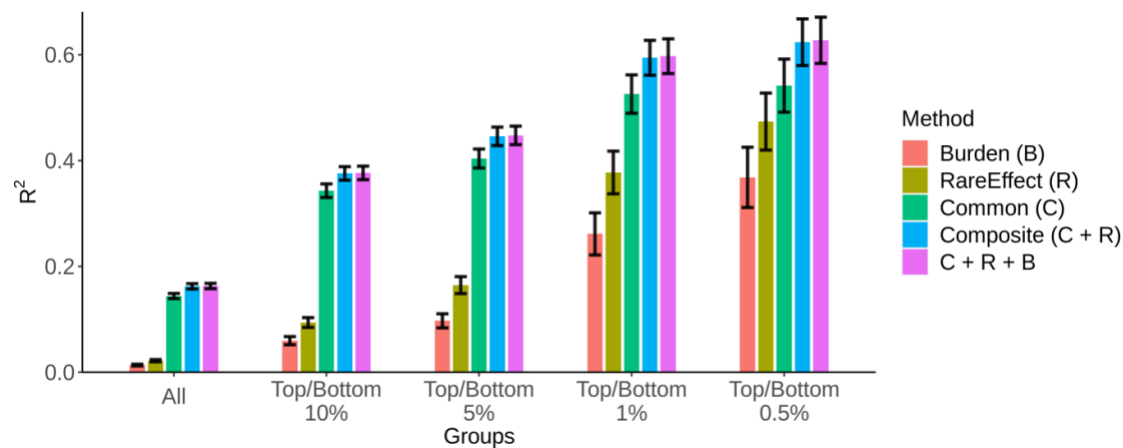
632

633 Variant-level effect size on HDL cholesterol level for (a) *APOC3* and (b) *SLC12A3* genes. Gene-
634 level p-values were obtained from SAIGE-GENE+, and we included genes with p-values <
635 2.5×10^{-6} . The left panel shows the effect size of collapsed ultra-rare variants categorized by
636 their functional annotations: loss-of-function (LoF), missense, and synonymous, respectively.
637 The right panel displays the variant-level effect size of rare variants. Variants are color-coded
638 based on their functional annotation: red for LoF, blue for missense, and green for
639 synonymous. The shapes of the points indicate the significance of the variants: circles
640 represent p-values > 0.05 , diamonds represent p-values between 5×10^{-8} and 0.05 , and
641 triangles represent p-values $< 5 \times 10^{-8}$. Single-variant p-values were obtained from SAIGE,
642 while the p-values of collapsed variants were derived from linear regression by regressing \tilde{y}
643 on each collapsed variant. The exon region is shaded.

644 **Fig. 4. Comparison of performance of risk prediction models for lipid phenotypes using**
645 **314,198 White British samples (80% of the whole White British samples) in UK Biobank**
646 **whole exome sequencing data.**

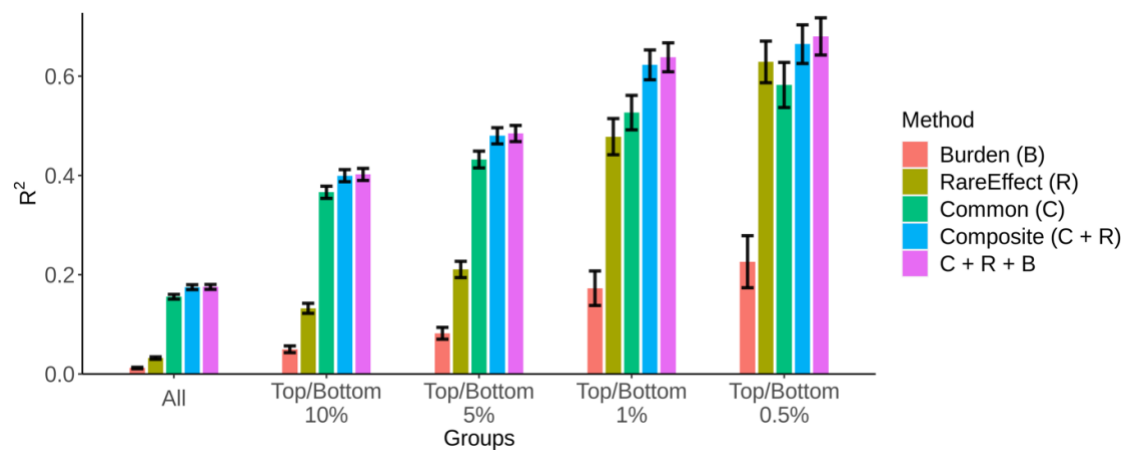
647

648 **a**



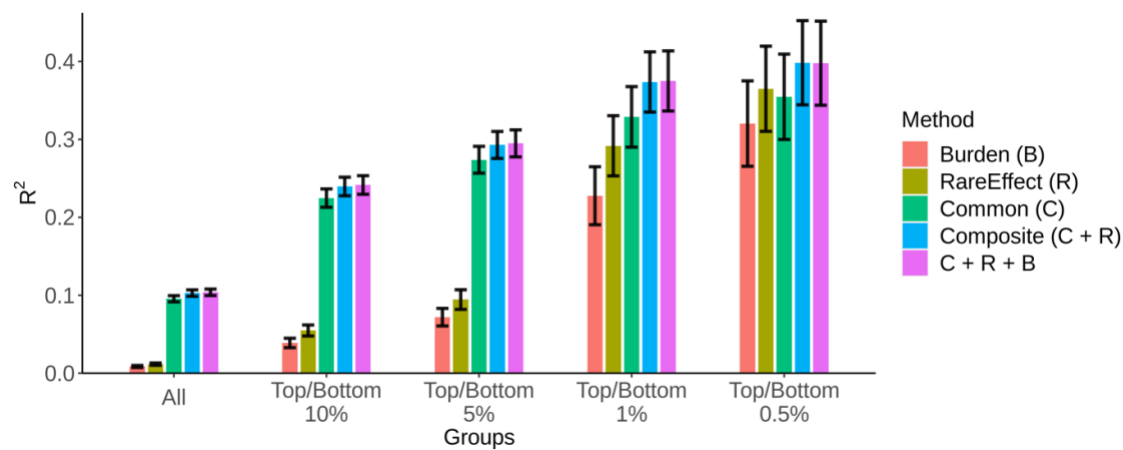
649

650 **b**



651

652 **c**



653

654 Coefficient of determination (R^2) of the risk prediction models for (a) HDL cholesterol level,
655 (b) LDL cholesterol level, and (c) Triglycerides level. We evaluated the R^2 of five models using:
656 (1) PRS_{burden} only, (2) PRS_{RE} only, (3) PRS_{common} only, (4) composite score, (5)
657 $PRS_{common} + PRS_{RE} + PRS_{burden}$ by five subgroups: (1) all samples, (2) samples with
658 top/bottom 10% PRS, (3) samples with top/bottom 5% PRS, (4) samples with top/bottom 1%
659 PRS, and (5) samples with top/bottom 0.5% PRS. The black vertical lines represent the 95%
660 confidence interval of the R^2 estimates.

661

662

663 **Supplementary Figures**

- 664 1. Estimated signed heritability for 10 tested traits
- 665 2. Predictive performance for 10 tested traits
- 666 3. Relationship between common PRS (PRS_{common}) and RareEffect PRS (PRS_{RE})
- 667 4. Enrichment of high-risk individuals in terms of RareEffect PRS (PRS_{RE}), common
- 668 variant PRS (PRS_{common}) and composite score in phenotype outliers
- 669 5. Computation time of RareEffect, linear regression, and ridge regression
- 670 6. Memory Usage of RareEffect by number of variants
- 671 7. Computation time of the fast implementation Firth bias correction and the normal
- 672 Firth correction
- 673 8. Relationship between RareEffect PRS (PRS_{RE}) and phenotype values
- 674 9. Comparison of estimated variance components between RareEffect and MoM
- 675 estimator
- 676 10. Comparison of estimated effect size between computing the hat matrix at every
- 677 iteration and only once in Firth bias correction

678

679 **Supplementary Tables**

- 680 1. Predictive performance for 10 tested phenotypes in UK Biobank
- 681 2. Pearson correlation between PRS_{common} and PRS_{RE}
- 682 3. Predictive performance (RMSE) for simulated continuous data by scenario
- 683 4. Predictive performance (RMSE) for simulated binary data by scenario