

# Identifying interpretable gene-biomarker associations with functionally informed kernel-based tests in 190,000 exomes

Remo Monti<sup>1,2</sup>, Pia Rautenstrauch<sup>2,3</sup>, Mahsa Ghanbari<sup>2</sup>, Alva Rani James<sup>1</sup>, Uwe Ohler<sup>2,4,a</sup>, Stefan Konigorski<sup>1,5,a</sup>, and Christoph Lippert<sup>1,5,a,\*</sup>

<sup>1</sup> *Digital Health - Machine Learning, Hasso Plattner Institute, University of Potsdam, 14482 Potsdam, Germany*

<sup>2</sup> *Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), 13125 Berlin, Germany*

<sup>3</sup> *Humboldt Universität zu Berlin, Department of Computer Science, 10099 Berlin, Germany*

<sup>4</sup> *Humboldt Universität zu Berlin, Department of Biology, 10099 Berlin, Germany*

<sup>5</sup> *Hasso Plattner Institute for Digital Health, Icahn School of Medicine at Mount Sinai, New York 10029-6574, NY, USA*

*\* corresponding author*

*<sup>a</sup> these authors contributed equally*

## Abstract

Here we present an exome-wide rare genetic variant association study for 30 blood biomarkers in 191,971 individuals in the UK Biobank. We compare gene-based association tests for separate functional variant categories to increase interpretability and identify 193 significant gene-biomarker associations. Genes associated with biomarkers were  $\sim 4.5$ -fold enriched for conferring Mendelian disorders. In addition to performing weighted gene-based variant collapsing tests, we design and apply variant-category-specific kernel-based tests that integrate quantitative functional variant effect predictions for missense variants, splicing and the binding of RNA-binding proteins. For these tests, we present a statistically powerful and computationally efficient combination of the likelihood-ratio and score tests that found 36% more associations than the score test alone. Kernel-based tests identified 13% more associations than their gene-based collapsing counterparts and had advantages in the presence of gain of function missense variants. We introduce local collapsing by amino acid position for missense variants and use it to interpret associations and identify potential novel gain of function variants in *PIEZO1*. Our results show the benefits of investigating different functional mechanisms when performing rare-variant association tests, and demonstrate pervasive rare-variant contribution to biomarker variability.

## 1 Introduction

Large biobanks that combine in-depth phenotyping with exome sequencing for hundreds of thousands of individuals promise new insights into the genetic architecture of health and disease [1]. Whilst common-variant association studies have detected tens of thousands of loci associated with heritable traits, the underlying functional mechanisms remain largely unknown due to linkage disequilibrium and the fact that the majority of loci

6 lie in non-coding regions of the genome [2]. Furthermore, effect sizes of common variants tend to be small, as  
7 variants with large detrimental effects are selected against, which limits their frequency [3, 4].

8 Association studies using whole exome-sequencing (WES) do not face these issues to the same extent, as  
9 they largely contain more interpretable loci where an enrichment for large effect sizes is expected [5]. However,  
10 the majority of genetic variants identified by WES are extremely rare, and the vast number of these variants  
11 poses challenges for rare variant association studies (RVAS), given the burden of multiple testing and low  
12 statistical power due to low allele frequencies [6, 7]. For these reasons, variants in RVAS are typically grouped  
13 into sets that correspond to functional units, such as genes, before association testing [6, 8, 9, 7]. Not only does  
14 this strategy aggregate signal and thereby increase statistical power, but it also lessens the burden of multiple  
15 testing. Burden tests, for example, collapse variants within genes into a single variable prior to association  
16 testing, i.e., perform *gene-based variant collapsing* [6, 10]. Alternatively, *kernel-based tests* aggregate groups of  
17 variants into a so-called kernel matrix that is tested using a score test [8] or the likelihood ratio test (LRT) [9]  
18 without the need for collapsing. Among these, the LRT has higher statistical power when effect sizes are large,  
19 but is computationally more expensive [11, 12].

20 While gene-based variant collapsing performs best in the presence of many causal variants with effect sizes  
21 that point in the same direction (e.g., increasing risk for disease), kernel-based tests have advantages in cases  
22 of opposing effects and fewer causal variants [7]. To increase the fraction of causal variants, exome-wide RVAS  
23 that use variant collapsing have defined qualifying variants based on annotations such as allele frequencies or  
24 variant effect predictions, and excluded all other observed variants from the association tests [6, 13, 14, 15, 16].  
25 These studies have mostly focused on non-synonymous variants, where software tools identify protein truncating  
26 variants and distinguish between benign and potentially deleterious missense variants [17, 18, 19, 20].

27 Here, we perform an extensive RVAS using exome sequencing data from the UK Biobank [21]. For approx-  
28 imately 190,000 individuals, 30 quantitative biomarkers provide objectively quantifiable measures related to  
29 the health status of individuals [22], making them attractive phenotypes for genome-wide association studies  
30 [23]. We go beyond the collapsing tests for coding variants described above and explore the use of kernel-based  
31 association tests and deep-learning-derived variant effect predictions for gene-regulatory variants, namely for  
32 splicing [24] and the binding of RNA-binding proteins (RBPs) [25].

33 Specifically, we use quantitative functional variant effect predictions to group and weigh variants in gene-  
34 based association tests and increase interpretability. The greater flexibility of kernel-based tests allowed us to  
35 design variant-category-specific tests and combine collapsing and non-collapsing approaches in the same tests.  
36 For kernel-based tests, we show that a computationally efficient combination of the score test and the restricted  
37 likelihood ratio test (RLRT) can identify 36% more significant associations compared to the score test alone. We  
38 find 193 significant gene-biomarker associations in total, the majority of which confirm associations previously  
39 reported to GWAS databases (87%) [26, 27].

40 We find that including participants from diverse ancestries identifies 14% more associations than limiting to  
41 participants with strict inferred European genetic ancestry, while increasing sample size by 17%. Associations  
42 with biomarkers frequently occurred in genes linked to Mendelian disorders, and comparisons to other association

43 studies confirmed their plausible disease relevance. Finally, we interpret associations that were only found for  
44 specific functional variant categories, or associations for which weighted gene-based variant collapsing and kernel-  
45 based tests gave vastly different results. This provided novel biological insights and highlighted the benefits of  
46 a diverse testing strategy for RVAS.

## 47 **2 Results**

### 48 **2.1 Data description and workflow**

49 We performed an RVAS of 30 quantitative serum biomarkers in UK Biobank 200k WES release [21]. These  
50 biomarkers contain established disease risk factors, diagnostic markers, and markers for phenotypes otherwise  
51 not well assessed in the UK Biobank cohort. They can roughly be grouped into cardiovascular, bone and  
52 joint, liver, renal, hormonal and diabetes markers (Table 1). After removing related individuals and restricting  
53 the analysis to those with no missing covariates 191,971 participants with diverse genetic ancestry remained.  
54 15,702,718 rare ( $MAF < 0.1\%$ ) variants were observed in this subset and passed quality criteria, including  
55 pruning variants with large deviations from the observed European MAF in other ancestries (Supplementary  
56 Methods). The median sample size for the biomarkers was 181,784 and ranged from 15,997 (Rheumatoid factor)  
57 to 182,742 (Alkaline phosphatase). 191,260 participants had at least one measured biomarker.

58 We used functional variant effect predictions to group and weigh variants and performed functionally in-  
59 formed gene-based association tests. Specifically we chose to investigate strict protein loss of function variants  
60 (pLOF, e.g., frame shift or protein truncating variants), missense variants, splice-altering variants, and variants  
61 predicted to change the binding of RNA-binding proteins (Figure 1, Methods), the latter two originating from  
62 deep learning models [28, 25].

63 We treated these categories separately during association testing to increase interpretability, resulting in  
64 multiple tests per gene, which we refer to as separate *models*. Specifically, we adapted either kernel-based tests,  
65 weighted gene-based variant collapsing, or both types of association test depending on the variant effect category  
66 (Methods). Tests were performed using a combination of score tests and restricted likelihood ratio tests. We  
67 make variant effect predictions for all variants in the UK Biobank 200k exome release available<sup>1</sup>, as well as our  
68 analysis pipeline<sup>2</sup> and software used to perform association tests<sup>3</sup>.

### 69 **2.2 Functionally informed association tests**

70 **Protein loss of function** We predicted the effects of genetic variants on protein coding genes using the  
71 Ensembl variant effect predictor [17] and found 463,479 pLOF variants with a median of 20 pLOF variants  
72 per gene (Figure 2, Methods). For pLOF variants, we assumed a large fraction of potentially causal variants,  
73 and that variants within the same gene should, by and large, affect the phenotype in the same direction. For

---

<sup>1</sup><https://github.com/HealthML/ukb-200k-wes-vep>

<sup>2</sup><https://github.com/HealthML/faatpipe>

<sup>3</sup><https://github.com/HealthML/seak>

category	biomarker	short	N	sign. genes
bone and joint	Alkaline phosphatase	ALP	182,742	10
bone and joint	Calcium		168,054	3
bone and joint	Rheumatoid factor		15,997	0
bone and joint	Vitamin D		174,473	4
cardiovascular	Apolipoprotein A	ApoA	167,050	18
cardiovascular	Apolipoprotein B	ApoB	181,808	6
cardiovascular	C-reactive protein	CRP	182,320	3
cardiovascular	Cholesterol		182,735	9
cardiovascular	HDL cholesterol	HDL	168,043	20
cardiovascular	LDL direct	LDL	182,420	6
cardiovascular	Lipoprotein A		146,249	13
cardiovascular	Triglycerides		182,587	12
diabetes	Glucose		167,916	3
diabetes	Glycated haemoglobin	HbA1c	182,494	9
hormonal	IGF-1		181,759	7
hormonal	Oestradiol		30,343	0
hormonal	SHBG		166,521	6
hormonal	Testosterone		165,211	1
liver	Alanine aminotransferase	ALT	182,675	5
liver	Albumin		168,139	4
liver	Aspartate aminotransferase	AST	182,096	2
liver	Direct bilirubin		154,963	6
liver	Gamma glutamyltransferase	GGT	182,655	7
liver	Total bilirubin		181,998	8
renal	Creatinine		182,639	7
renal	Cystatin C		182,720	7
renal	Phosphate		167,808	3
renal	Total protein		167,931	5
renal	Urate		182,538	8
renal	Urea		182,612	1

Table 1: **UK Biobank blood biomarkers analyzed in this study.** Biomarker categories, biomarker names, their abbreviations used in the text (short), sample size (i.e. non-missing with complete covariates, N) and number of distinct genes significantly associated with each biomarker (sign. genes) after multiple testing correction.

74 these reasons, we performed gene-based variant collapsing tests (Methods), and found 88 significant associations  
75 originating from 53 distinct genes.

76 **Missense** We defined 1,834,082 high-impact missense variants based on PolyPhen-2 [18] and SIFT [19] (Meth-  
77 ods). 18,420 genes contained at least one high-impact missense variant, with a median of 73 high-impact missense  
78 variants observed per gene.

79 We hypothesized that missense variants in the same gene might have both trait-increasing and trait-  
80 decreasing effects, and that there might be fewer causal variants. Therefore, we did not only perform weighted  
81 variant collapsing tests for these variants, but also kernel-based association tests. We designed a missense-  
82 specific kernel that collapses variants locally by amino acid position, which affected 20% of variants. We further  
83 used the Cauchy Combination Test (CCT) [29] to dynamically incorporate pLOF variants in these tests (Meth-  
84 ods, Supplementary Methods). Combining missense and pLOF variants has been shown to increase the number  
85 of discoveries in RVAS, due to many missense variants leading to a loss of function [13, 15, 16]. We identified  
86 146 significant associations using gene-based variant collapsing, and 160 using kernel-based association tests,

87 with an overlap of 121. The total of 185 associations identified by either model originated from 101 distinct  
88 genes.

89 **Splicing** We located 737,795 potentially splice-altering rare single nucleotide variants in 17,158 genes by cross  
90 referencing against published SpliceAI variant effect predictions ([24]; Methods). The median number of variants  
91 per gene was 30. We hypothesized that these splice variants could have complex downstream consequences and  
92 decided to compare both weighted gene-based variant collapsing and kernel-based association tests. For kernel-  
93 based tests, we used the weighted linear kernel [8]. As for missense variants, we dynamically incorporated pLOF  
94 variants in these tests to increase power.

95 We identified 75 significant associations with gene-based variant collapsing in 44 distinct genes, whereas  
96 kernel-based tests identified 88 significant associations in 51 genes. As our definition of pLOF variants included  
97 variants that directly hit annotated splice donor/acceptor sites, there was a considerable overlap of 95,842  
98 variants between these annotations (21% of all pLOF variants). We therefore expected (and found) large  
99 overlaps (69, 74%) in the significant associations for pLOF and splice variants.

100 **RBP-binding** Splicing is only one of several eukaryotic post-transcriptional regulatory mechanisms mediated  
101 by interactions of RNA-binding proteins (RBPs) with their target RNAs. As the UK Biobank WES data also  
102 contain variants in non-protein-coding parts of mRNAs, namely in introns (41.43%) and UTRs (11.32%), we  
103 reasoned that we may be able to identify variants with regulatory effects mediated by differential binding of  
104 RBPs. Specifically, we investigated if changes in the binding of RBPs predicted by DeepRiPe [25] could be  
105 associated with biomarker levels. The six RBPs QKI, MBNL1, TARDBP, ELAVL1, KHDRBS1 and HNRNPD  
106 were selected based on their binding preferences (introns, exons) [30], the high performance of the model to  
107 predict genuine target sites for these RBPs, and the reported presence of clear binding sequence motifs. We  
108 predicted variant effects for these RBPs and identified 370,880 variants with large predicted effects in 17,394  
109 genes, with a median of 12 variants per gene (Methods).

110 As we expected a low number of causal variants and potentially opposing effect sizes, we only performed  
111 kernel-based association tests and identified 9 significant associations in 9 distinct genes.

## 112 **2.3 Integrative analysis overview**

113 Merging the results from all models yielded a total of 193 associations originating from 117 distinct genes  
114 (Supplementary Table S1, 212 associations if counting genes with shared exons separately). We found at least one  
115 significant association for all but two biomarkers (oestradiol and rheumatoid factor) (Figure 2, Supplementary  
116 Figures S1-S4). For the majority of associations (174; 82%), tests combining missense and protein LOF variants  
117 gave the smallest p-values.

118 We calculated the genomic inflation factor  $\lambda_{GC}$  across all tests that were performed genome-wide, and did  
119 not find evidence of inflated type I error levels (Figure S8, Supplementary Data).  $\lambda_{GC}$  ranged from 0.84 to 1.06  
120 for gene-based variant collapsing tests (median = 0.97) and from 0.92 to 1.08 for kernel-based tests (median

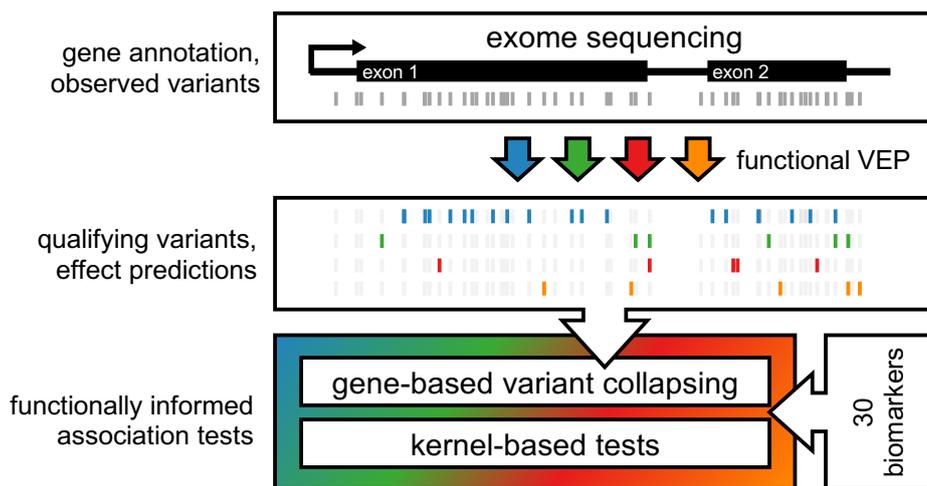


Figure 1: **Rare-variant association testing pipeline.** Exome sequencing measures exon-proximal genetic variants. All variants are subjected to functional variant effect prediction (VEP). Qualifying variants are determined based on the variant effect predictions and minor allele frequencies (MAF <0.1%) and categorized based on their predicted functional impacts (protein loss of function, missense, splicing, RBP-binding). Finally, we test the different categories of qualifying variants in gene-based association tests against 30 biomarkers using gene-based variant collapsing and kernel-based tests.

121 = 1.01). Of the 117 distinct genes, 43 (37%) were associated with more than one biomarker, and a few genes  
122 had five or more significant associations: *ANGPTL3*, *APOB*, *JAK2*, *GIGYF1* and *G6PC*. Many of the genes  
123 we found to be associated with biomarkers had either been implicated in diseases related to those biomarkers  
124 (e.g., *LRP2* with renal markers [31]) or are mechanistically related to the biomarkers themselves (e.g., cystatin  
125 C with its own gene, *CST*).

126 80% of genes ( $\pm 5\text{kb}$ ) contained single variants associated with the same biomarkers identified by a study using  
127 imputed genotypes and a larger sample size on the same data [23]. Yet, 97% of gene-phenotype associations  
128 remained significant after conditioning on the most significant variants identified by that study (considering  
129 variants  $\pm 500\text{kb}$  around the gene start positions). We therefore conclude that most associations we observed  
130 are largely independent of the signals captured by array-based genotypes of (mostly) common variants.

## 131 2.4 Including all ancestries increases power for biomarker traits

132 We repeated our analysis restricting to 164,148 individuals of strict inferred European ancestry (Supplementary  
133 Methods). While 175 associations were significant in both analyses, 37 were only identified in the analysis with  
134 all ancestries (AA) and 11 were only significant in the analysis with EUR individuals (Figure 3, Supplementary  
135 Table S3). The majority of hits significant in one analysis but not in the other were close to the significance  
136 threshold. Outliers such as the associations of *ALD16A1* with Urate could all be explained by hard thresholds  
137 for variant inclusion and strong signals from single variants. Both the type of association test and variant  
138 category that produced the lowest gene-trait p-values were largely consistent between both analyses (Figure 3).

## 139 2.5 Clinical relevance of biomarker associations

140 Genes identified by biomarker associations were highly enriched for those involved in Mendelian disorders (odds  
141 ratio 4.47, p-value  $3.08 \times 10^{-16}$ , one-sided Fisher's exact test), as determined by querying the Online Mendelian  
142 Inheritance in Man database (OMIM, <https://omim.org/>). 74 out of 125 (59.2%) distinct (i.e., non-overlapping)  
143 genes identified had at least one OMIM entry. Our results show that changes at the biomarker level are detectable  
144 even in the 27 genes for which only recessive disorders were listed.

145 We further examined whether rare variants in these genes were associated with binary or continuous traits  
146 at larger sample sizes by comparing with two studies published at the time of writing [15, 16] (Supplementary  
147 Table 4). 21 out of 125 genes associated with biomarkers were also associated with one or more binary traits,  
148 whereas 14 genes were associated with non-hematological quantitative traits in either study (Table 2).

149 Often such cases could be placed in a plausible causal context. For example, we found variants in *SLC22A12*,  
150 *ABCG2* and *ALDH16A1* to be associated with increased levels of urate, a causal factor for gout, which they  
151 were also associated with. Another example are genes associated with growth hormone IGF-1 (*IGFALS*, *GHI*,  
152 *GHRH*, *PAPPA2*) which are also associated with anthropometric traits (e.g., sitting height). We identified  
153 patterns of biomarker associations for *GIGYF1* consistent with Type II diabetes (T2D), and an association  
154 with T2D was confirmed at larger sample sizes [15, 16] and other independent studies [32, 33]. Six out of seven  
155 biomarker associations for *JAK2* were only identified by kernel-based tests, and the gain of function variant  
156 driving these associations (V617F or rs77375493 [34]) was also associated with myeloproliferative neoplasms.  
157 In other cases, trait-biomarker relationships were not readily apparent, such as the association of variants in  
158 *SYNJ2* with lower levels of  $\gamma$ -glutamyltransferase and increased risk of hearing loss.

159 Separating variant effect categories during association testing and comparing kernel-based to gene-based  
160 collapsing tests allowed us to further interpret our results, as illustrated with several examples below.

## 161 2.6 Combining variant annotations yields six associations for G6PC

162 We found five associations for *G6PC*, three of which were only found when combining protein LOF and missense  
163 variants with gene-based variant collapsing tests (alkaline phosphatase, SHBG, urate). Variants in *G6PC* cause  
164 Glucose-6-phosphatase deficiency type Ia [35, 36], an autosomal recessive disease categorized by growth retar-  
165 dation, enlarged kidneys and liver, low blood glucose, and high blood lipid and uric acid levels. Consistent with  
166 signs of inflammation and impaired kidney and liver function, we found elevated levels of alkaline phosphatase  
167 ( $p = 1.15 \times 10^{-9}$ ),  $\gamma$ -glutamyltransferase ( $p = 2.64 \times 10^{-11}$ ), and C-reactive protein ( $p = 1.62 \times 10^{-13}$ ) in in-  
168 dividuals with predicted high-impact missense or pLOF mutations in *G6PC*. We further identified a significant  
169 association with decreased levels of sex hormone-binding globulin (SHBG,  $p = 2.88 \times 10^{-9}$ ), which is primar-  
170 ily produced in the liver [37]. All p-values above are those given by gene-based weighted variant collapsing  
171 tests combining missense and pLOF variants, which gave the lowest p-values for all these associations. While  
172 Glucose-6-phosphatase deficiency type Ia is a rare recessive disease, our findings show that altered biomarker  
173 levels indicative of mild symptoms are detectable in heterozygous carriers of missense and LOF variants in the

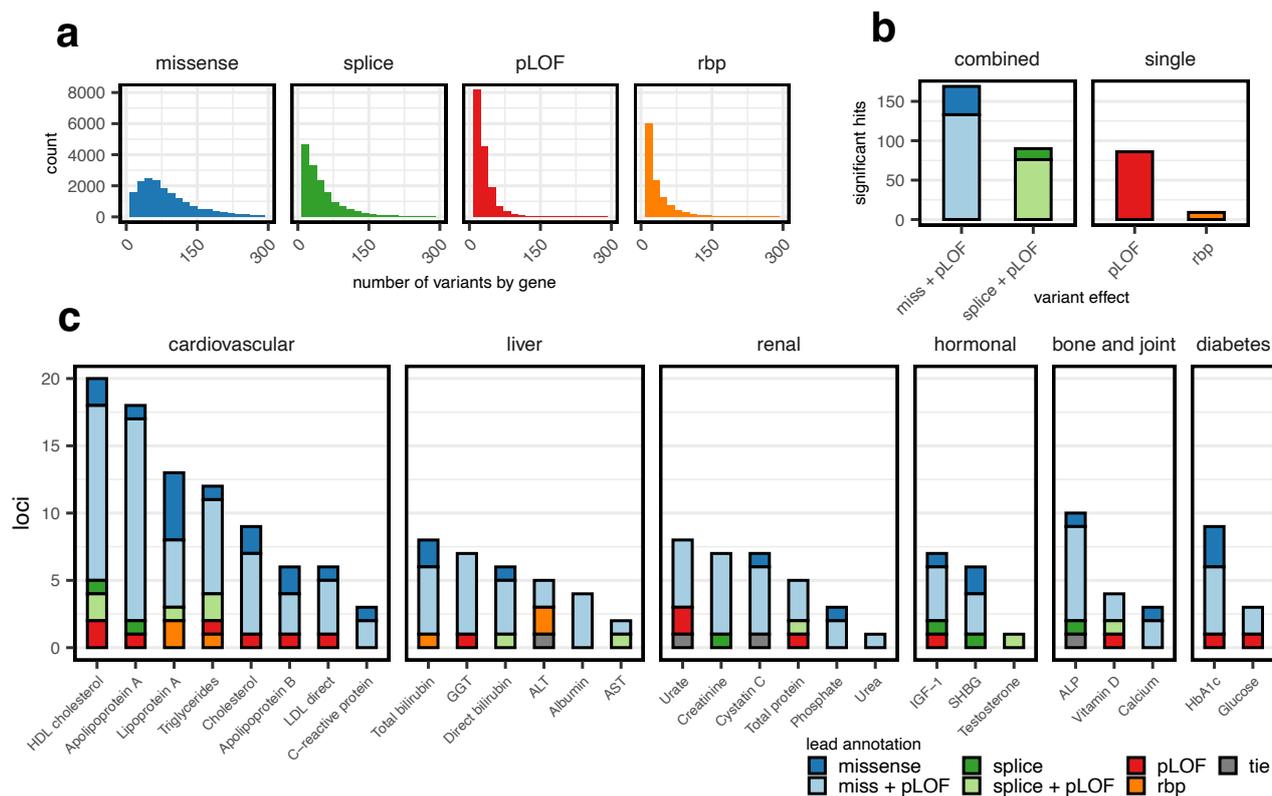


Figure 2: **Association tests overview.** (a) Histograms of the number of qualifying variants per tested gene for the different variant categories. Ranges are truncated at 300 variants, which affected 730 genes for missense, 84 for splice, 7 for pLOF and 15 for rbp (b) Bar plot of the number of significant genes found by testing qualifying variants in the different categories. Tests for splice and missense variants (left) were dynamically combined with pLOF variants, i.e., two p-values, one arising from a test including only missense/splice variants, and one combining those variants with pLOF variants were combined using the Cauchy combination test. (c) Bar plot showing 193 significant gene-biomarker associations for 28 biomarkers (x-axis). Bars in panels (a) and (b) are colored by the variant type (or combination thereof) which gave the lowest p-value (lead annotation).

174 *G6PC* gene, a pattern which was consistent for many of the identified genes.

## 175 2.7 Novel potential gain of function variants in *PIEZO1*

176 Our testing strategy allowed us to identify genes in which specific variant categories might play an important  
 177 role. One such example was *PIEZO1*, a mechanosensitive cation channel [38], which we found associated with  
 178 the diabetes marker HbA1c.

179 For weighted gene-based variant collapsing tests, we found a significant negative association of missense  
 180 variants with HbA1c ( $p = 7.47 \times 10^{-29}$ ), while the test for pLOF variants was not significant ( $p = 0.89$ , 609  
 181 carriers). By far the lowest p-value for this gene was given by the kernel-based test for missense variants  
 182 ( $p = 2.8 \times 10^{-108}$ ). The large differences between variant categories and types of association test lead us to  
 183 closer investigate the 858 predicted high-impact missense variants in 7,412 individuals for this gene.

184 We performed single-variant score tests and identified multiple missense variants with strong negative as-  
 185 sociations with HbA1c (Table 3). One of these variants, 16:88719665:G:A or T2127M (rs587776991) is a gain  
 186 of function variant that slows down inactivation kinetics of *PIEZO1* in patients with dehydrated hereditary

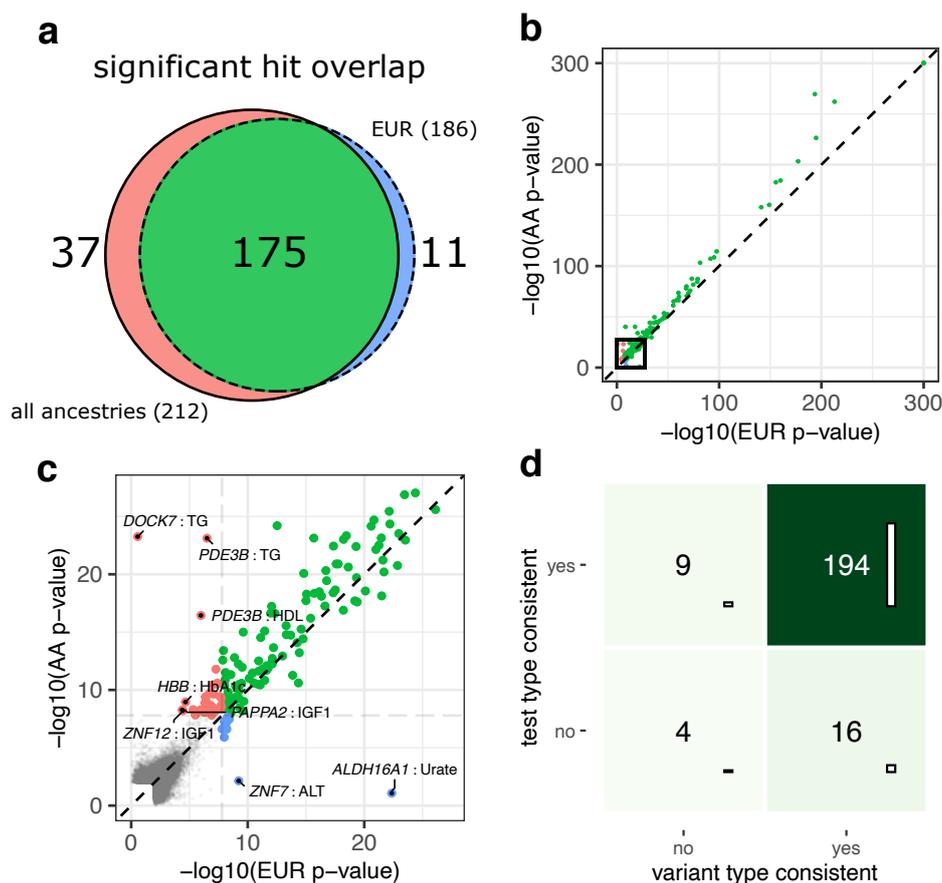


Figure 3: **Comparison of EUR vs. all-ancestry (AA) analyses.** (a) Venn diagram of the number of significant associations ( $p < 1.61 \times 10^{-8}$ ) identified in either analysis. (b) Scatter plot showing the smallest p-value for each gene-biomarker association in the EUR analysis (x-axis) vs. in the analysis with all ancestries (y-axis) on the negative log<sub>10</sub>-scale. The thick black box denotes the area highlighted in panel (c). Associations significant in either analysis are color-coded according to (a) and drawn thicker. The significance threshold is drawn as a light gray dashed horizontal/vertical line. Non-significant associations are drawn in gray. Associations with  $p > 0.1$  in both analyses are not shown. (d) For significant gene-biomarker associations in either analysis, we show if the variant effect types and the test-types (gbvc or kernel-based) which gave the smallest p-values were consistent between the two analyses.

187 stomatocytosis (a disorder of red blood cells), together with other gain of function variants [39, 40]. Decreased  
188 levels of HbA1c had previously been observed in individuals with red blood cell disorders [41, 42, 43].

189 We therefore hypothesized that the other highly significant variants could also potentially be gain of function  
190 variants. We grouped the missense variants within *PIEZO1* by the amino acid positions they affected and  
191 performed local variant collapsing. This allowed us to identify other positions in *PIEZO1* (e.g., 2110R or  
192 2474V) that are potentially sensitive to gain of function mutations (Table 3, Figure 4).

193 When comparing the results of our all-ancestry analysis which includes ancestry-based variant pruning to a  
194 version which did not, we came across another missense variant (L2277M, 16:88716656:G:T) which was much  
195 more common in individuals with inferred South Asian ancestry (SAS,  $MAF_{SAS} = 3.45\%$ ) than European  
196 ancestry ( $MAF_{EUR} = 0.006\%$ ,  $p = 0$ , two-sided Fisher's exact test), and strongly associated with decreased  
197 levels of HbA1c ( $\beta = -0.71 \pm 0.06$  s.d.,  $p = 1.16 \times 10^{-29}$ , score-test restricted to individuals of SAS ancestry).  
198 This association remained significant when conditioning on a recently identified nearby non-coding variant

category	gene name	biomarker	analysis	p-value	effect	test	N <sub>variant</sub>	N <sub>carrier</sub>	OMIM	traits (examples)[15, 16]
bone and joint	HSPG2	ALP	AA	2.27e-19	m	K	1,424	10,275	r	repair of conjunctiva
bone and joint	B4GALNT3	ALP	AA	3.91e-14	m	gbvc	246	2,499		height
bone and joint	FLG	Vitamin D	AA	1.54e-10	p	gbvc	350	2,448	d,r	asthma, eczema
cardiovascular	APOB	LDL direct (+5)	AA	9.69e-159	m	K	855	3,948	d,r	high cholesterol (diag.)
cardiovascular	PCSK9	LDL direct (+2)	AA	4.58e-88	m	gbvc	186	968	d	high cholesterol (diag.)
cardiovascular	APOC3	Triglycerides (+2)	AA	2.18e-74	m	K	28	297	m	lipid-lowering med
cardiovascular	LDLR	LDL direct (+3)	AA	2.29e-50	m	gbvc	214	1,303	d,r	high cholesterol (diag.)
cardiovascular	PDE3B	Triglycerides (+1)	AA	7.66e-24	m	K	225	1,510		height, mass
cardiovascular	JAK2	HDL (+6)	AA	5.22e-19	m	K	254	1,022	m,s	neoplasms
cardiovascular	G6PC	CRP (+4)	AA	1.62e-13	m	gbvc	85	828	r	arm fat free mass
cardiovascular	TM6SF2	Triglycerides (+2)	AA	3.10e-13	m	gbvc	85	1,392		liver fat percentage
cardiovascular	SRSF2	HDL	EUR*	6.19e-09	m	K	10	51		myeloid leukaemia
diabetes	PIEZO1	HbA1c	AA	5.59e-108	m	K	1,011	8,021	d,r	varicose veins, height, mass
diabetes	GCK	HbA1c (+1)	EUR	1.36e-40	m	gbvc	53	144	d,r	T2D
diabetes	GIGYF1	HbA1c (+4)	AA	1.14e-10	p	gbvc	41	68		T2D, education score Engl.
diabetes	HHB	HbA1c	AA	1.10e-09	m	K	33	92	d,r	thalassaemia
hormonal	SHBG	SHBG (+1)	AA	4.64e-227	m	K	105	669		heel bone mineral density
hormonal	IGFALS	IGF-1	AA	1.14e-78	m	K	242	1,900	r	height, mass
hormonal	GH1	IGF-1	AA	1.85e-10	s	gbvc	49	470	d,r	height, mass
hormonal	GHRH	IGF-1	AA	2.07e-09	m	gbvc	22	117	m	height, mass
hormonal	PAPPA2	IGF-1	AA	5.38e-09	m	gbvc	288	1,208	r	height, mass
liver	UGT1A1	Total bilirubin (+1)	AA	2.90e-67	m	K	148	864	m,r	Gilbert's syndrome
liver	ABCB4	ALT	AA	2.30e-10	m	gbvc	226	1,192	d,r	cholelithiasis
liver	SYNJ2	GGT	AA	1.55e-08	m	gbvc	434	3,020		hearing loss
renal	SLC22A12	Urate (+1)	AA	0	m	gbvc	151	1,044	r	gout
renal	ALDH16A1	Urate	EUR*	4.40e-23	m	K	266	2,060		gout
renal	ABCG2	Urate	AA	2.19e-14	m	gbvc	170	1,575	m	gout
renal	TNFRSF13B	Total protein	AA	1.62e-12	m	gbvc	76	898	d,r	interpol. age when op.
renal	ANKRD12	Total protein	AA	3.18e-11	p	gbvc	53	121		walking pace: slow
renal	PKD1	Creatinine	EUR*	2.77e-10	m	K	1,255	10,695	d	cystic kidney disease

Table 2: **Significant genes with non-haematological trait associations in other RVAS.** Table showing the biomarker associations with the lowest p-values for all genes that were also significantly associated with non-haematological traits (e.g., disease diagnoses or anthropometric traits) in [15] or [16] in either the all-ancestry analysis (AA) or the EUR-ancestry analysis (EUR). Only example traits are shown in column *traits* (full list in Supplementary Table S4). The numbers in brackets indicates how many other biomarkers were associated with the same gene in our analysis. *effect*: the lead variant effect type (m: missense+pLOF, p: pLOF, s: splice+pLOF). *test*: the test-type which gave the lowest p-value. *OMIM*: If disorders are linked to the gene in OMIM, their inheritance patterns (d: dominant, m: missing, r: recessive, s: somatic). Associations marked with (\*) were not significant in the AA-analysis.

(16:88784993:C:G,  $p = 9.1 \times 10^{13}$ )[44], yet, not the other way around ( $p = 0.55$ ). 16:88784993:C:G could therefore be tagging signal from the potentially causal gain of function variant L2277M. We were able to replicate this association in individuals with extended EUR ancestry ( $p = 2.73 \times 10^{-6}$ , Supplementary Methods).

Consistent with a role of red blood cell disorders, we also found associations of *RHAG* (Rh Associated Glycoprotein) and *SPTA1* with decreased levels of HbA1c. Mutations in *RHAG* cause overhydrated hereditary stomatocytosis [45], while *SPTA1* mutations cause hereditary elliptocytosis [46].

While it has been suggested that *PIEZO1* stimulates insulin release [47], the decreased levels of HbA1c we observed in individuals with *PIEZO1*-variants are more likely explained by (perhaps subclinical) forms of stomatocytosis or other abnormalities in red blood cells resulting from increased membrane permeability, i.e., a gain of function [48].

## 2.8 Position-specific association of *ABCA1* variants with inflammation marker CRP

We found four significant associations of *ABCA1* with biomarker levels. Three of these, namely the associations with Apolipoprotein A, HDL cholesterol, and cholesterol, are directly related to its role as an ATP-dependent transporter of cholesterol [49]. In line with previous findings, in our gene-based variant collapsing analysis we found both pLOF and high-impact missense variants to be strongly associated with decreased serum levels of

variant id	position	weight	variant	$N_{carrier}$	variant p-val.	$\beta_{variant}$	$\pm$	position p-value	$\beta_{position}$	$\pm$
16:88736318:C:T	463	0.98	A/T	109	1.52e-10	-0.56	0.09	2.01e-10	-0.56	0.09
16:88736317:G:A	463	1.00	A/V	1	8.97e-01	0.12	0.92	2.01e-10	-0.56	0.09
16:88734895:G:A	610	0.68	L/F	63	4.82e-08	-0.63	0.12	9.55e-08	-0.71	0.13
16:88734894:A:C	610	0.99	L/R	4	7.00e-01	-0.18	0.46	9.55e-08	-0.71	0.13
16:88731880:C:G	1,008	0.99	G/R	65	7.20e-14	-0.85	0.11	4.30e-15	-0.88	0.11
16:88731880:C:T	1,008	0.99	G/R	3	1.18e-02	-1.33	0.53	4.30e-15	-0.88	0.11
16:88726891:A:C	1,175	1.00	F/V	4	4.71e-08	-2.50	0.46	4.71e-08	-2.50	0.46
16:88726565:C:T	1,260	0.71	V/I	78	8.03e-08	-0.56	0.10	2.65e-08	-0.68	0.12
16:88726565:C:A	1,260	1.00	V/F	1	6.67e-02	-1.68	0.92	2.65e-08	-0.68	0.12
16:88721626:C:G	1,772	0.99	R/P	12	3.11e-05	-1.10	0.26	1.61e-08	-1.07	0.19
16:88721627:G:C	1,772	0.98	R/G	10	9.12e-04	-0.96	0.29	1.61e-08	-1.07	0.19
16:88721626:C:A	1,772	0.98	R/L	1	3.90e-02	-1.89	0.92	1.61e-08	-1.07	0.19
16:88721627:G:A	1,772	1.00	R/C	1	4.89e-01	-0.63	0.92	1.61e-08	-1.07	0.19
16:88721423:C:G	1,804	0.87	G/A	31	7.92e-11	-1.07	0.16	7.92e-11	-1.15	0.18
16:88719717:G:A	2,110	1.00	R/W	9	6.68e-18	-2.63	0.31	5.28e-33	-2.66	0.22
16:88719716:C:T	2,110	0.89	R/Q	9	1.03e-16	-2.53	0.31	5.28e-33	-2.66	0.22
16:88719665:G:A	2,127	1.00	T/M	22	1.11e-31	-2.29	0.20	1.11e-31	-2.29	0.20
16:88716234:C:T	2,365	0.81	G/R	9	3.74e-16	-2.49	0.31	3.74e-16	-2.76	0.34
16:88715751:C:T	2,474	0.99	V/M	101	3.77e-08	-0.50	0.09	9.58e-09	-0.52	0.09
16:88715751:C:G	2,474	0.88	V/L	2	2.99e-02	-1.41	0.65	9.58e-09	-0.52	0.09
16:88715751:C:A	2,474	0.88	V/L	1	8.00e-01	-0.23	0.92	9.58e-09	-0.52	0.09

Table 3: **Potential *PIEZO1* gain of function variants** Variants are grouped and ordered by amino acid position and single-variant p-values. All variants with position p-values below  $10^{-7}$  are shown. weight: impact score;  $N_{carrier}$ : number of carriers; variant p-val.: single-variant p-value (score test);  $\beta_{variant}$ : variant effect size ( $\pm$  standard error); position p-value: p-value when collapsing variants by position after weighting (score test);  $\beta_{position}$ : position effect size ( $\pm$  standard error). Positions relate to the ENST00000301015 transcript.

215 these biomarkers [50].

216 Yet, one additional association with the inflammation marker C-reactive protein (CRP) was only identified  
 217 by kernel-based association tests ( $p = 1.33 \times 10^{-27}$ ). This prompted us to further investigate the 336 high-impact  
 218 missense variants observed in *ABCA1*. Single-variant score tests and collapsing by amino-acid position identified  
 219 two missense variants (9:104831048:C:A, 9:104831048:C:G) in one of the extracellular domains affecting the same  
 220 amino acid (W590) which were associated with strongly decreased levels of CRP (Figure 4). The two variants  
 221 carried most of the signal in this gene with single-variant p-values of  $6.3 \times 10^{-32}$  for W590L (A allele, 54 carriers)  
 222 and  $8.09 \times 10^{-8}$  for W590S (G allele, 12 carriers, score test).

223 The W590S-variant leads to reduced cholesterol and phospholipid efflux, while retaining expression and  
 224 ability to bind APOA1 [51, 52]. The other and more common variant, W590L, has been observed [53, 54], but  
 225 to our knowledge not experimentally evaluated.

226 The binding of APOA1 to *ABCA1* activates anti-inflammatory pathways via JAK2 and STAT3 in macrophages  
 227 [55]. Because W590S has been shown to slow down dissociation of bound APOA1 [52], this provides a plausible  
 228 causal mechanism for the reduced levels of CRP we observe in carriers of the W590S-variant. We hypothesize  
 229 that W590L might act through the same mechanism. This property could set these variants apart from other  
 230 missense variants in *ABCA1*, which have been reported to abolish binding of APOA1 [51].

231 *ABCA1* could therefore be a gene in which some variants elicit both a gain of function (slower dissociation  
 232 of APOA1) and a loss of function (decreased cholesterol efflux) with distinct effects on different biomarkers.

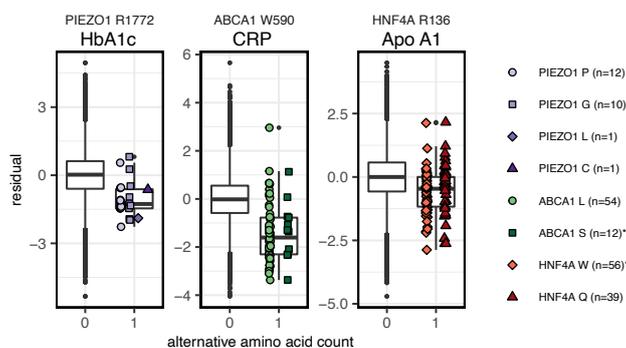


Figure 4: **Local collapsing of missense variants.** Dosage box plots showing the alternative amino acid counts (x-axis) against the covariate-adjusted quantile transformed phenotypes (y-axis). Collapsing variants by amino acid position identified negative associations of *PIEZO1* R1772 with HbA1c ( $p = 1.61 \times 10^{-8}$ ), *ABCA1* W590 with C-reactive protein ( $p = 3.43 \times 10^{-38}$ ), and *HNF4A* R136 with Apolipoprotein A ( $p = 1.2 \times 10^{-10}$ , score test). Collapsing together with ClinVar variants with reported conditions (marked with “\*”) helps place novel variants into disease context. For all 3 associations, collapsed p-values were lower than those of the single variants. The lower and upper hinges indicate first and third quartiles, whiskers extend to the largest/lowest values no further than  $1.5 \times \text{IQR}$  away from the upper/lower hinges and black points denote outliers.

## 2.9 Unique associations identified by splice-predictions

In total, we identified 6 gene-biomarkere associations exclusively when incorporating SpliceAI variant effect predictions of which 4 were only found using kernel-based association tests. Specifically, we found associations of variants in *SLC9A5* with ApoA and HDL cholesterol, *NDUFB8* with Aspartate aminotransferase, *GHI* with IGF-1, *ECE1* with with Alkaline phosphatase, and *KDM6B* with SHBG.

Most of these associations were mainly caused by single variants. An exception was the known association of *GHI* (Growth Hormone 1) with IGF-1 [56], one of the two hits in this subset also found by gene-based variant collapsing. The interpretation of single highly significant variants driving these associations could not necessarily be narrowed down to a single mechanism. For example, the predicted splice variant in the last exon responsible for the two significant associations of *SLC9A5* (16:67270978:G:A, 29 carriers, Figure 5a), was also a missense variant (albeit with low to moderate predicted impact [54]). *ECE1* and *KDM6B* lie in proximity to the genes coding for the biomarkers they were found associated with (*ALPL*, *SHBG*), therefore we could not exclude transcriptional cis-regulatory effects as the cause of these associations.

## 2.10 Associations identified by RBP-binding predictions

Out of the 9 significant associations we identified using DeepRiPe variant effect predictions, the associations of *ANGPTL3* with Triglycerides, and *AGPAT4*, *PLG* and *LPA* with Lipoprotein A had also been found using other models. The extreme heritability of Lipoprotein A, which is largely due to variation in the *LPA* gene [57, 58], makes it hard to interpret the associations for the lead genes *LPA*, *AGPAT4* and *PLG*, that all lie within a megabase distance to *LPA*. The association we observed for *ANGPTL3* was largely driven by a single intronic variant (1:62598067:T:C, Figure 5), which was predicted to increase the binding probability of QKI. The same variant on the opposite strand was also responsible for the association of *DOCK7* with triglycerides. However, we determined that *ANGPTL3* was more likely the causal gene, given the associations of *ANGPTL3*

255 with triglycerides we had independently found with other variant categories and the close proximity of an  
256 *ANGPTL3* exon.

257 We further investigated this variant by assessing binding probabilities of RBPs beyond the six RBPs in focus  
258 here that are represented in DeepRiPe. We found that the variant was also predicted to decrease the binding  
259 probability of BCLAF1, a factor related to mRNA processing [59] and increase binding of splice-regulator  
260 HNRNPL (Figure 5). Using attribution maps [25], we found that instead of strengthening or inserting a QKI  
261 binding motif, this variant weakens a splice donor signal in the presence of upstream binding motifs for the  
262 splicing regulators QKI and HNRNPL (Supplementary Figure S5). SpliceAI predicted only a weak upstream  
263 donor loss (0.02) for this variant, which was well below the threshold of 0.1 we used to identify splice-altering  
264 variants, but indicative of the same trend.

265 The remaining five associations exclusively identified by association tests incorporating RBP-binding pre-  
266 dictions were those of *ATG16L1* with total bilirubin, *SOD2* with Lipoprotein A, *SLC39A4* with Alanine  
267 aminotransferase and *SHARPIN* with Alanine aminotransferase.

268 Both *SHARPIN* and *SLC39A4* lie within half a megabase of the Alanine aminotranferase gene (*GPT*), there-  
269 fore we could not exclude potential transcriptional cis-regulatory effects as the true cause for these associations.  
270 Furthermore, single variants carried most of the signal for both genes.

271 Common intronic variants of *ATG16L1* were previously found to be associated with Crohn's disease, in-  
272 flammatory bowel disease [60, 61], and increased bilirubin levels [62, 63, 23]. However, the association remained  
273 significant after conditioning on the intronic top variant from a previous study [23].

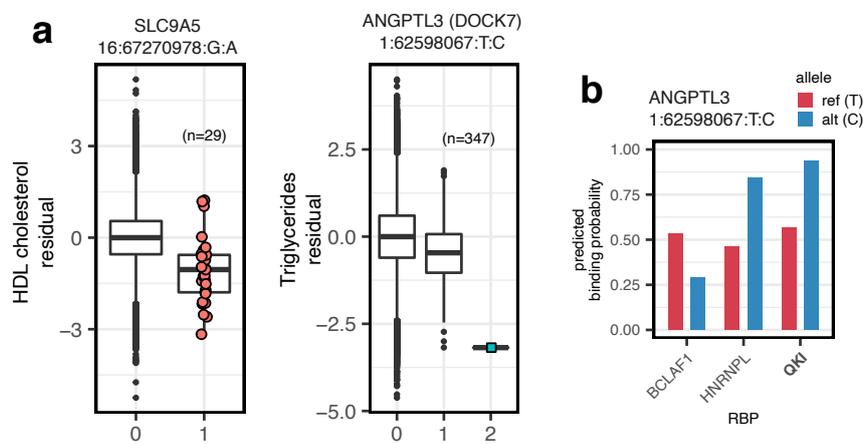


Figure 5: **Variants prioritized by deep learning models** (a) Dosage box plots showing covariate-adjusted quantile transformed phenotypes against minor allele counts for variants in *SLC9A5* and *ANGPTL3/DOCK7*. A predicted splice-variant 16:67270978:G:A is negatively associated with HDL cholesterol ( $p = 7.83 \times 10^{-12}$ , score test), whereas intronic 1:62598067:T:C is negatively associated with Triglycerides ( $p = 1.37 \times 10^{-25}$ , score test). The lower and upper hinges indicate first and third quartiles, whiskers extend to the largest/lowest values no further than  $1.5 \times \text{IQR}$  away from the upper/lower hinges and black points denote outliers. (b) DeepRiPe binding probabilities for 1:62598067:T:C for three RBPs in HepG2 cells. While predicted probabilities for the reference sequence are ambiguous, the alternative allele shifts binding probabilities in favor of QKI and HNRNPL. All RBPs with absolute predicted variant effects above 0.2 and binding probabilities greater than 0.5 for either reference or alternative alleles are shown.

## 274 **2.11 Combined likelihood ratio and score tests (sLRT)**

275 In order to benefit from both the speed of the score test [8] and higher power of the LRT in the presence of  
276 large effect sizes [11, 12] we investigated the use of a combination of these tests, we call the sLRT (score-LRT).  
277 The sLRT is a likelihood ratio test that is performed only when initial score tests reach nominal significance at  
278 a given cutoff  $t$ . If this threshold is not reached, it returns the score test p-value. Throughout our analysis, we  
279 used  $t = 0.1$ , and found that it was unlikely that a larger threshold would have identified many more associations  
280 (Supplementary Figure S6). As the run time is dominated by the cost of computing the LRT, this test can  
281 achieve a computational speedup factor of roughly  $1/t = 10$  over the LRT under the null hypothesis.

282 We found that the sLRT was able to identify more significant associations than the score test for missense  
283 and splice-variants when performing kernel-based association tests (Supplementary Figure S7). However, a  
284 large fraction of these additional associations (31% for splice variants, and 48% for missense variants) could also  
285 identified by gene-based variant collapsing for which the sLRT and score test gave almost identical results.

286 Nevertheless, we found the majority of the remaining additional associations to be plausible and/or previ-  
287 ously reported in other association studies and therefore used the sLRT throughout our analysis, except for  
288 pLOF variants, where we only performed a gene-based variant collapsing test.

289 While the score-test is the locally most powerful test (i.e., for alternatives that are close to the null hypothesis  
290 [64]), our empirical results suggest that the restricted LRT finds more associations. Based on simulations, [11]  
291 found that the LRT has higher power if variant effect sizes are large, which is true for many rare variants  
292 observed in exome sequencing studies.

## 293 **3 Discussion**

294 In our analysis, we combined gene-based variant collapsing and kernel-based tests under a common framework  
295 and performed functionally informed gene-based tests for rare variants with 30 biomarkers.

296 Overall, our approach was successful at identifying disease genes without explicitly using disease diagnoses  
297 themselves, even for recessive diseases (e.g., *G6PC* and glycogen storage disease Ia [35] or *ABCA1* and Tangier  
298 disease [65]), or diseases with mixed inheritance patterns, while keeping the number of tests low compared to  
299 phenome-wide association studies.

300 While some of the changes in biomarker levels we detected might be sub-clinical, they could interfere with  
301 the diagnosis of common conditions which rely on biomarkers. To prevent misdiagnoses and enable preventative  
302 care, our results could aid the design of targeted sequencing panels that focus on the genes with the highest  
303 impacts. For example, we found 8% of the participants to harbor at least one protein LOF variant in any of  
304 the significant loci for that variant category. In other words, it is fairly common to have at least one uncommon  
305 LOF variant with potential effects on biomarker levels. Properly imputing and including the newly discovered  
306 *PIEZO1*-L2277M variant in individuals with South Asian ancestry could improve polygenic risk scores for  
307 HbA1c and related diagnostics in that population.

308 One major contribution of our study was the design and successful application of kernel-based tests that

309 incorporate quantitative functional variant effect predictions for large exome sequencing data. A previous study  
310 had tried to apply kernel-based score tests using SKAT [8] on the 50k WES release, but concluded that results  
311 were difficult to reproduce [13]. In contrast to their approach, we did not re-weight variants according to allele  
312 frequencies. Furthermore, we showed that a computationally efficient combination of the restricted LRT and  
313 score test has potentially higher power than the score test alone (possibly due to the large effect sizes), and  
314 identified more associations than gene-based variant collapsing.

315 When comparing gene-based collapsing and kernel-based tests for missense variants, we found kernel-based  
316 tests to have advantages in the presence of gain of function variants (*PIEZO1*, *ABCA1*, *JAK2*), where they  
317 identified plausible causal associations missed by gene-based collapsing tests. These genes likely are examples  
318 of a low fraction of causal variants, a regime in which kernel-based tests are statistically more powerful than  
319 gene-based collapsing [7].

320 While we found a large overlap between our associations and those found in other studies [66, 23, 16],  
321 the differences highlight the sensitivity of gene-based tests to qualifying criteria for rare variants (i.e., allele  
322 frequencies or variant impact predictions), which can make them harder to reproduce (Supplementary Table  
323 S4). By making our analysis pipeline public, we hope to increase reproducibility and enable others to explore  
324 different qualifying criteria more easily.

325 We demonstrated how local collapsing of missense variants by amino acid position aids interpretation and  
326 causal reasoning in the presence of previously validated variants. Local collapsing was directly built into the  
327 kernel-based tests we performed for missense variants, where it affected 20% of variants, a number which will  
328 increase with larger and datasets.

329 We explored the use of deep-learning-derived variant effect predictions for splicing and the binding of RBPs.  
330 The restriction to exon-proximal regions meant we only observed a fraction of the variants potentially acting  
331 through these mechanisms. Associations found by incorporating splice-predictions largely overlapped with those  
332 identified with pLOF variants (which included simple splice donor/acceptor variants). While we found some  
333 associations exclusively with splice-predictions, these were mostly due to single variants and would need further  
334 validation (e.g., *SLC9A5*). Similar reasoning holds for the associations found with predictions for RBP-binding.

335 We anticipate that deep-learning-based predictions will become more valuable for non-coding regions in  
336 whole-genome sequencing studies, for which the approaches we developed will also be applicable.

337 Deep-learning-derived functional annotations have been considered in other studies in the context of asso-  
338 ciation testing. Proposed methods include signed LD-score regression [67], or the association tests presented in  
339 DeepWAS [68]. However, these methods have not been designed for very rare variants.

340 In future studies, methods like AlphaFold [69] could allow specific testing of effects on protein folding.  
341 Methods that allow predicting residue-residue interactions within proteins could enable the mostly unsupervised  
342 identification of protein domains and their separate testing [70]. The methodological advances and practices in  
343 this association study also apply to those situations, and serve as potential baselines for functionally informed  
344 kernel-based association tests with rare variants.

## 345 4 Methods

### 346 4.1 UK Biobank Data processing

347 All 30 blood biochemistry biomarkers (category 17518) from the UK Biobank were quantile-transformed to  
348 match a normal distribution with mean 0 and unit standard deviation using scikit-learn (v0.22.2) [71]. For  
349 testosterone, which showed a clear bimodal distribution based on sex, quantile transformation was performed  
350 separately for both sexes. We performed ancestry scoring as described in [72] based on the 1000 Genomes  
351 super populations [73], which was used to prune variants and select participants with similar genetic ancestry  
352 (Supplementary Methods).

353 Sex, BMI, age at recruitment, smoking status, genetic principal components and continuous ancestry pre-  
354 dictions were used as covariates (Supplementary Table S2). Smoking status (never, previous, current) was  
355 encoded in three separate binary variables. Participants with any missing covariates were excluded. We used  
356 the `ukb_gen_samples_to_remove` function of the `ukbtools` package (v0.11.3) [74] together with pre-computed re-  
357 latedness scores (`ukbA_rel_sP.txt`, see UK Biobank Resource 531) to remove closely related individuals, keeping  
358 only one representative of groups that are related to the 3rd degree or less. After removing 6,293 related in-  
359 dividuals and restricting to those with no missing covariates, 191,971 participants remained. This sample was  
360 55% female (45% male) and the average age at recruitment was 56.5 years ( $\sigma = 8$ ). Furthermore, the average  
361 BMI was 27.37 ( $\sigma = 4.76$ ) and our subset contained 18,529 current and 66,988 previous smokers.

362 In our analysis we made use of the PLINK-formatted exome sequencing genotype data. The final results  
363 presented in this manuscript were derived from the 200k WES release produced by the OQFE pipeline [21].  
364 The UK Biobank pipeline already implements quality filters [21, 75]. Additionally, we removed all variants that  
365 violated the Hardy-Weinberg equilibrium (HWE) assumption (HWE exact test p-value below the threshold of  
366  $10^{-5}$ ) and variants genotyped in less than 90% of participants.

367 We calculated minor allele frequencies within all unrelated participants with complete covariates (see above),  
368 and excluded variants with a study-wide allele frequency above 0.1%. In the analysis including all ancestries,  
369 we additionally performed ancestry-based pruning to remove variants with large allele frequency differences  
370 between ancestries (Supplementary Methods). This led to the exclusion of 1,033,382 variants. We found this  
371 step to be critical for preventing test statistic inflation for kernel-based tests with certain phenotypes. We did  
372 not analyze variants on sex chromosomes.

373 In the all-ancestry analysis 15,701,695 variants passed these filters, of which 45.87% were singletons. 12,793,493  
374 were considered for the EUR ancestry analysis (42.2% singletons). We directly use UK Biobank variant identi-  
375 fiers (which include chromosome and 1-based hg38 positions) to name variants in order to facilitate comparisons.

### 376 4.2 Variant effect prediction and annotation

377 **Protein loss of function and missense** We predicted effects for all genetic variants that passed basic fil-  
378 tering using the Ensembl Variant Effect Predictor [17] (VEP, v101; cache version 97), including scores from

379 Polyphen-2 [18] (v2.2.2) and SIFT [19] (v5.2.2). All variants marked as splice\_acceptor\_variant, splice\_donor\_variant,  
380 frameshift\_variant, stop\_gained, stop\_lost or start\_lost were considered protein loss of function (pLOF) variants  
381 as in [13]. We further annotated missense variants by calculating impact scores (averages between deleterious-  
382 probabilities given by PolyPhen-2 and SIFT), which were used to filter and weigh variants in the association  
383 tests. Specifically, Missense variants were included if their impact score was at least 0.8, or if they affected  
384 amino acid positions for which another variant with impact score of at least 0.8 was observed.

385 **Splicing** We retrieved published pre-computed variant effect predictions produced by the SpliceAI deep learn-  
386 ing model [24] for single nucleotide polymorphisms. SpliceAI predicts consequences of genetic variants for nearby  
387 splice sites, specifically splice donor loss/gain or splice acceptor loss/gain. We used the splice-site-proximal  
388 masked delta scores (v1.3). In the masked files, scores corresponding to the strengthening of annotated splice  
389 sites and weakening of non-annotated splice sites are set to 0, as these are generally less pathogenic. We included  
390 splice-variants in the association tests if at least one of the four SpliceAI delta scores was greater or equal to  
391 0.1. The maxima over the different delta scores for every variant were used to weigh variants in the association  
392 tests (Supplementary Methods).

393 **RBP-binding** We predicted the effects of all genetic variants on the binding of 6 RNA-binding proteins  
394 (RBPs) using a modified version of the DeepRiPe deep neural network [25], in which predictions are purely  
395 sequence-based. We predicted the differences in binding by subtracting the predictions for the reference alleles  
396 from those for the alternative allele [76], and used these variant effect predictions to filter and weigh variants  
397 during the association tests (Supplementary Methods). Variants were included into the association tests if at  
398 least one predicted effect on any of the RBPs had an absolute value greater or equal to 0.25.

### 399 4.3 Statistical models and tests

400 Let  $N(\boldsymbol{\mu}; \boldsymbol{\Sigma})$  denote a multivariate Normal distribution with means  $\boldsymbol{\mu}$  and a variance-covariance matrix  $\boldsymbol{\Sigma}$ .  
401 We wish to jointly test the association of  $m$  genetic variants with a quantitative trait  $\mathbf{y}$  for a sample of  $N$   
402 observations (i.e., participants), while controlling for  $q$  covariates. Within the linear mixed model framework,  
403  $\mathbf{y}$  can be modelled as follows [8, 9]:

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\alpha}; \sigma_e^2 \mathbf{I}_N + \sigma_g^2 \mathbf{K}_g), \quad (1)$$

404 where  $\mathbf{X}$  is the  $N \times q$  covariate design matrix (fixed effect) and  $\boldsymbol{\alpha}$  is the vector of fixed-effect parameters,  
405 which together determine the mean values of  $\mathbf{y}$ . The variance-covariance matrix of  $\mathbf{y}$  is composed of the  
406 independently distributed residual variance ( $\mathbf{I}_N$  scaled by  $\sigma_e^2$ ) and the kernel-matrix  $\mathbf{K}_g$  (scaled by  $\sigma_g^2$ ), which  
407 captures the genetic similarity between individuals.  $\mathbf{K}_g$  is a function of the  $N \times m$  matrix of mean-centered  
408 minor allele counts  $\mathbf{G}$  (random effect) of the genetic variants we wish to test.

409 Any valid variance-covariance matrix can be substituted for  $\mathbf{K}_g$ . In order to use efficient algorithms for  
410 estimating the parameters  $\sigma_e^2$  and  $\sigma_g^2$  and performing association tests, we require  $\mathbf{K}_g$  to be factored as a

411 quadratic form [9, 11]:

$$\mathbf{K}_g = \phi(\mathbf{G})\phi(\mathbf{G})^T, \quad (2)$$

412 where the function  $\phi$  transforms  $\mathbf{G}$  into intermediate variables before performing the test. Finding an  
413 appropriate function  $\phi$  depends on the underlying biological assumptions, and the available prior information.  
414 Gene-based variant collapsing approaches are a special case, in which the function  $\phi$  returns an  $N \times 1$  vector (a  
415 single variable) as output. Therefore kernel-based tests and variant collapsing methods can be treated under the  
416 same statistical framework. In our analysis,  $\phi$  is a function that transforms  $\mathbf{G}$  taking variant effect predictions  
417 and, for missense and RBP-variants, variant positions into account (Supplementary Methods).

418 Regardless of the choice of kernel (and hence  $\phi$ ) the statistical test is defined by the null hypothesis  $H_0 :$   
419  $\sigma_g^2 = 0$ , and the alternative hypothesis  $H_1 : \sigma_g^2 \geq 0$ . Both a score test and likelihood ratio test (LRT) have been  
420 described for this application. While the score test is often chosen in statistical genetics applications due to its  
421 speed and software availability, the LRT has been shown to have higher power when effect-sizes are large but is  
422 computationally more demanding [8, 11, 12].

423 In order to avoid computing the LRT for all genes but still profit from potentially higher power, we performed  
424 score tests genome-wide and only performed the restricted LRT if score tests (within the specific variant cate-  
425 gory) reached nominal significance, an approach which we call the score-LRT (sLRT, Supplementary Methods).  
426 The sLRT returns the p-value for the score test if nominal significance was not reached, otherwise it returns  
427 the p-value for the likelihood ratio test.

428 We sampled RLRT test statistics using fast exact algorithms described in [77], and fit parametric null distri-  
429 butions to the pooled test statistics across genes in order to calculate p-values [11] (Supplementary Methods).  
430 We did this separately for different variant-effect and test types. This method gave highly similar p-values  
431 to using gene-specific null distributions (Supplementary Figures S9-S12), but is faster as it requires fewer test  
432 statistics to be simulated per gene and fewer distributions to be fit.

433 We applied the statistical framework above to perform both gene-based variant collapsing tests and kernel-  
434 based association tests, corresponding to different functions  $\phi$  (Supplementary Methods). Additionally, for splice  
435 and missense variants, tests using only those variant categories and tests combining these variant categories  
436 with pLOF variants were integrated into single tests using the Cauchy combination test [29]. We adjust p-values  
437 for the total number of 3,091,910 tests in the all-ancestry analysis using Bonferroni correction ( $\text{FWER} = 0.05$ ),  
438 which lead to a cutoff of  $1.6171 \times 10^{-8}$ .

#### 439 **4.4 Gene-based testing procedure**

440 We performed gene-based tests for all protein coding genes in the Ensembl 97 release. For all pLOF variants  
441 we performed gene-based variant collapsing using the score test genome-wide.

442 For missense variants, we performed both a weighted gene-based variant collapsing and kernel-based as-  
443 sociation tests using the sLRT. For the kernel-based tests with missense variants, we designed a kernel that

444 collapses variants by amino acid position (local collapsing), and weighs them by their impact score. Addition-  
445 ally, in cases where either missense-variant score test used in the sLRT was nominally significant ( $p < 0.1$ ),  
446 we combined missense and protein LOF variants for joint tests. For these joint tests, we investigated both the  
447 use of joint weighted gene-based variant collapsing and a kernel-based test that combines collapsing of pLOF  
448 variants with local collapsing of missense variants by concatenation (Supplementary Methods). The p-values of  
449 the combined tests were integrated using the Cauchy combination method[29] (individual p-values are reported  
450 in Supplementary Table S1).

451 For predicted splice-variants we followed the same strategy as for missense variants, however, we used the  
452 weighted linear kernel [8] without local collapsing instead. Finally, in the association tests including variants  
453 predicted to change the binding of RBPs, we only performed kernel-based association tests using the sLRT. For  
454 this purpose we designed a kernel that can take into account both variant positions and direction of variant  
455 effects (Supplementary Methods).

456 Because some of the genes in the Ensembl 97 release share exons, we encountered cases in which these genes  
457 shared associations caused by the same variants. We do not report these as distinct genes in the main text or  
458 abstract, but include the full list of 212 associations in Supplementary Table S1.

## 459 4.5 Data availability

460 Variant effect predictions for all variants in the 200k exome sequencing release are made available on github  
461 (<https://github.com/HealthML/ukb-200k-wes-vep>).

## 462 4.6 Code availability

463 A snakemake pipeline that allows reproducing results from this study is available on github  
464 (<https://github.com/HealthML/faatpipe>).

## 465 4.7 Contributions

466 R.M., S.K. and C.L. conceived and designed the study. R.M., P.R., A.R. and S.K. performed initial prototyping.  
467 R.M. and P.R. wrote software to perform the statistical tests with guidance from U.O., S.K. and C.L.. R.M.  
468 wrote the analysis pipeline with guidance from A.R., S.K., U.O. and C.L.. R.M. carried out the statistical  
469 analyses with guidance from S.K., U.O. and C.L.. M.G. supplied the weights for the DeepRiPe model and  
470 produced attribution maps. P.R. and R.M. queried GWAS databases and compared results with other studies.  
471 R.M., P.R, S.K. and C.L. wrote the manuscript. All authors revised the manuscript.

## 472 Acknowledgements

473 The authors wish to thank Wolfgang Kopp for valuable comments on the manuscript. This research has  
474 been conducted using the UK Biobank Resource under Application Number 40502. This research has received

475 funding by the German Federal Ministry of Education and Research (BMBF) in the project KI-LAB-ITSE  
476 (project number 01—S19066) and the European Commission in the Horizon 2020 project INTERVENE (Grant  
477 agreement ID: 101016775).

## 478 References

- 479 [1] Sudlow, C. *et al.* Uk biobank: an open access resource for identifying the causes of a wide range of complex  
480 diseases of middle and old age. *Plos med* **12**, e1001779 (2015).
- 481 [2] Buniello, A. *et al.* The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays  
482 and summary statistics 2019. *Nucleic acids research* **47**, D1005–D1012 (2019).
- 483 [3] Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
- 484 [4] Hernandez, R. D. *et al.* Ultrarare variants drive substantial cis heritability of human gene expression.  
485 *Nature genetics* **51**, 1349–1355 (2019).
- 486 [5] Zhu, Q. *et al.* A genome-wide comparison of the functional properties of rare and common genetic variants  
487 in humans. *The American Journal of Human Genetics* **88**, 458–468 (2011).
- 488 [6] Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application  
489 to analysis of sequence data. *The American Journal of Human Genetics* **83**, 311–321 (2008).
- 490 [7] Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and  
491 statistical tests. *The American Journal of Human Genetics* **95**, 5–23 (2014).
- 492 [8] Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association  
493 test. *The American Journal of Human Genetics* **89**, 82–93 (2011).
- 494 [9] Listgarten, J. *et al.* A powerful and efficient set test for genetic markers that handles confounders. *Bioin-*  
495 *formatics* **29**, 1526–1533 (2013).
- 496 [10] Povysil, G. *et al.* Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nature*  
497 *Reviews Genetics* **20**, 747–759 (2019).
- 498 [11] Lippert, C. *et al.* Greater power and computational efficiency for kernel-based association testing of sets  
499 of genetic variants. *Bioinformatics* **30**, 3206–3214 (2014).
- 500 [12] Zhou, J. J., Hu, T., Qiao, D., Cho, M. H. & Zhou, H. Boosting gene mapping power and efficiency with  
501 efficient exact variance component tests of single nucleotide polymorphism sets. *Genetics* **204**, 921–931  
502 (2016).
- 503 [13] Cirulli, E. T. *et al.* Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes  
504 from two cohorts. *Nature Communications* **11**, 542 (2020).

- 505 [14] Van Hout, C. V. *et al.* Exome sequencing and characterization of 49,960 individuals in the uk biobank.  
506 *Nature* **586**, 749–756 (2020).
- 507 [15] Wang, Q. *et al.* Rare variant contribution to human disease in 281,104 uk biobank exomes. *Nature* **597**,  
508 527–532 (2021).
- 509 [16] Backman, J. D. *et al.* Exome sequencing and analysis of 454,787 uk biobank participants. *Nature* **599**,  
510 628–634 (2021).
- 511 [17] McLaren, W. *et al.* The ensembl variant effect predictor. *Genome biology* **17**, 1–14 (2016).
- 512 [18] Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations  
513 using polyphen-2. *Current protocols in human genetics* **76**, 7–20 (2013).
- 514 [19] Ng, P. C. & Henikoff, S. Sift: Predicting amino acid changes that affect protein function. *Nucleic acids*  
515 *research* **31**, 3812–3814 (2003).
- 516 [20] Sundaram, L. *et al.* Predicting the clinical impact of human mutation with deep neural networks. *Nature*  
517 *genetics* **50**, 1161–1170 (2018).
- 518 [21] Szustakowski, J. D. *et al.* Advancing human genetics research and drug discovery through exome sequencing  
519 of the uk biobank. *Nature genetics* **53**, 942–948 (2021).
- 520 [22] Strimbu, K. & Tavel, J. A. What are biomarkers? *Current Opinion in HIV and AIDS* **5**, 463 (2010).
- 521 [23] Sinnott-Armstrong, N. *et al.* Genetics of 35 blood and urine biomarkers in the uk biobank. *Nature genetics*  
522 1–10 (2021).
- 523 [24] Jaganathan, K. *et al.* Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548  
524 (2019).
- 525 [25] Ghanbari, M. & Ohler, U. Deep neural networks for interpreting rna-binding protein target preferences.  
526 *Genome research* **30**, 214–226 (2020).
- 527 [26] Staley, J. R. *et al.* Phenoscanner: a database of human genotype–phenotype associations. *Bioinformatics*  
528 **32**, 3207–3209 (2016).
- 529 [27] Kamat, M. A. *et al.* Phenoscanner v2: an expanded tool for searching human genotype–phenotype associ-  
530 ations. *Bioinformatics* **35**, 4851–4853 (2019).
- 531 [28] Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535–  
532 548.e24 (2019).
- 533 [29] Liu, Y. & Xie, J. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary  
534 dependency structures. *Journal of the American Statistical Association* **115**, 393–402 (2020).

- 535 [30] Mukherjee, N. *et al.* Deciphering human ribonucleoprotein regulatory networks. *Nucleic acids research* **47**,  
536 570–581 (2019).
- 537 [31] Willnow, T. E. & Christ, A. Endocytic receptor lrp2/megalin—of holoprosencephaly and renal fanconi  
538 syndrome. *Pflügers Archiv-European Journal of Physiology* **469**, 907–916 (2017).
- 539 [32] Zhao, Y. *et al.* Gigyf1 loss of function is associated with clonal mosaicism and adverse metabolic health.  
540 *Nature Communications* **12**, 1–6 (2021).
- 541 [33] Deaton, A. M. *et al.* Gene-level analysis of rare variants in 379,066 whole exome sequences identifies an  
542 association of gigyf1 loss of function with type 2 diabetes. *Scientific reports* **11**, 1–16 (2021).
- 543 [34] James, C. *et al.* A unique clonal jak2 mutation leading to constitutive signalling causes polycythaemia  
544 vera. *nature* **434**, 1144–1148 (2005).
- 545 [35] Chou, J. Y. & Mansfield, B. C. Mutations in the glucose-6-phosphatase- $\alpha$  (g6pc) gene that cause type ia  
546 glycogen storage disease. *Human mutation* **29**, 921–930 (2008).
- 547 [36] Froissart, R. *et al.* Glucose-6-phosphatase deficiency. *Orphanet journal of rare diseases* **6**, 1–12 (2011).
- 548 [37] Nakhla, A. M. *et al.* Human sex hormone-binding globulin gene expression-multiple promoters and complex  
549 alternative splicing. *BMC Molecular Biology* **10**, 1–18 (2009).
- 550 [38] Coste, B. *et al.* Piezo1 and piezo2 are essential components of distinct mechanically activated cation  
551 channels. *Science* **330**, 55–60 (2010).
- 552 [39] Albuissou, J. *et al.* Dehydrated hereditary stomatocytosis linked to gain-of-function mutations in mechan-  
553 ically activated piezo1 ion channels. *Nature communications* **4**, 1–9 (2013).
- 554 [40] Andolfo, I. *et al.* Multiple clinical forms of dehydrated hereditary stomatocytosis arise from mutations in  
555 piezo1. *Blood* **121**, 3925–3935 (2013).
- 556 [41] Song, A. *et al.* Low hba1c with normal hemoglobin in a diabetes patient caused by piezo1 gene variant: A  
557 case report. *Frontiers in Endocrinology* **11**, 356 (2020).
- 558 [42] Nakatani, R. *et al.* Importance of the average glucose level and estimated glycated hemoglobin in a diabetic  
559 patient with hereditary hemolytic anemia and liver cirrhosis. *Internal Medicine* **57**, 537–543 (2018).
- 560 [43] Finan, E. & Joseph, J. Glycosylated haemoglobin: a false sense of security. *BMJ Case Reports CP* **11**,  
561 e227668 (2018).
- 562 [44] Sun, Q. *et al.* Analyses of biomarker traits in diverse uk biobank participants identify associations missed  
563 by european-centric analysis strategies. *Journal of human genetics* 1–7 (2021).
- 564 [45] Bruce, L. J. *et al.* The monovalent cation leak in overhydrated stomatocytic red blood cells results from  
565 amino acid substitutions in the rh-associated glycoprotein. *Blood* **113**, 1350–1357 (2009).

- 566 [46] Sahr, K. *et al.* Sequence and exon-intron organization of the dna encoding the alpha i domain of human  
567 spectrin. application to the study of mutations causing hereditary elliptocytosis. *The Journal of clinical*  
568 *investigation* **84**, 1243–1252 (1989).
- 569 [47] Deivasikamani, V. *et al.* Piezo1 channel activation mimics high glucose as a stimulator of insulin release.  
570 *Scientific reports* **9**, 1–10 (2019).
- 571 [48] Andolfo, I., Russo, R., Gambale, A. & Iolascon, A. Hereditary stomatocytosis: an underdiagnosed condi-  
572 tion. *American journal of hematology* **93**, 107–121 (2018).
- 573 [49] Tang, C.-K. *et al.* Effect of apolipoprotein ai on atp binding cassette transporter a1 degradation and  
574 cholesterol efflux in thp-1 macrophage-derived foam cells. *Acta biochimica et biophysica Sinica* **36**, 218–  
575 226 (2004).
- 576 [50] Marcil, M. *et al.* Mutations in the abc 1 gene in familial hdl deficiency with defective cholesterol efflux.  
577 *The Lancet* **354**, 1341–1346 (1999).
- 578 [51] Vaughan, A. M., Tang, C. & Oram, J. F. Abca1 mutants reveal an interdependency between lipid export  
579 function, apoa-i binding activity, and janus kinase 2 activation. *Journal of lipid research* **50**, 285–292  
580 (2009).
- 581 [52] Nagao, K., Zhao, Y., Takahashi, K., Kimura, Y. & Ueda, K. Sodium taurocholate-dependent lipid efflux  
582 by abca1: effects of w590s mutation on lipid translocation and apolipoprotein ai dissociation. *Journal of*  
583 *lipid research* **50**, 1165–1172 (2009).
- 584 [53] Probst, M. C. *Development and evaluation of multiplex and high-throughput SNP analysis for the ABCA1*  
585 *gene*. Ph.D. thesis (2004).
- 586 [54] Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans.  
587 *Nature* **581**, 434–443 (2020).
- 588 [55] Tang, C., Liu, Y., Kessler, P. S., Vaughan, A. M. & Oram, J. F. The macrophage cholesterol exporter abca1  
589 functions as an anti-inflammatory receptor. *Journal of Biological Chemistry* **284**, 32336–32343 (2009).
- 590 [56] Goddard, A. D. *et al.* Mutations of the growth hormone receptor in children with idiopathic short stature.  
591 *New England Journal of Medicine* **333**, 1093–1098 (1995).
- 592 [57] Enkhmaa, B., Anuurad, E., Zhang, W., Tran, T. & Berglund, L. Lipoprotein (a): genotype–phenotype  
593 relationship and impact on atherogenic risk. *Metabolic syndrome and related disorders* **9**, 411–418 (2011).
- 594 [58] Shadrina, A. S. *et al.* Prioritization of causal genes for coronary artery disease based on cumulative evidence  
595 from experimental and in silico studies. *Scientific reports* **10**, 1–15 (2020).
- 596 [59] Sarras, H., Alizadeh Azami, S. & McPherson, J. P. In search of a function for bclaf1. *TheScientificWorld-*  
597 *Journal* **10**, 1450–1461 (2010).

- 598 [60] Jostins, L. *et al.* Host–microbe interactions have shaped the genetic architecture of inflammatory bowel  
599 disease. *Nature* **491**, 119–124 (2012).
- 600 [61] Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and  
601 highlight shared genetic risk across populations. *Nature genetics* **47**, 979–986 (2015).
- 602 [62] Choi, Y. *et al.* Causal associations between serum bilirubin levels and decreased stroke risk: a two-sample  
603 mendelian randomization study. *Arteriosclerosis, thrombosis, and vascular biology* **40**, 437–445 (2020).
- 604 [63] Seo, J. Y. *et al.* A genome-wide association study on liver enzymes in korean population. *Plos one* **15**,  
605 e0229374 (2020).
- 606 [64] Lin, X. Variance component testing in generalised linear models with random effects. *Biometrika* **84**,  
607 309–326 (1997).
- 608 [65] Rust, S. *et al.* Tangier disease is caused by mutations in the gene encoding atp-binding cassette transporter  
609 1. *Nature genetics* **22**, 352–355 (1999).
- 610 [66] Wang, Q. *et al.* Surveying the contribution of rare variants to the genetic architecture of human disease  
611 through exome sequencing of 177,882 uk biobank participants. *bioRxiv* (2020).
- 612 [67] Reshef, Y. A. *et al.* Detecting genome-wide directional effects of transcription factor binding on polygenic  
613 disease risk. *Nature genetics* **50**, 1483–1493 (2018).
- 614 [68] Arloth, J. *et al.* Deepwas: Multivariate genotype-phenotype associations by directly integrating regulatory  
615 information using deep learning. *PLoS computational biology* **16**, e1007616 (2020).
- 616 [69] Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature*  
617 **577**, 706–710 (2020).
- 618 [70] Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million  
619 protein sequences. *Proceedings of the National Academy of Sciences* **118** (2021).
- 620 [71] Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *the Journal of machine Learning research*  
621 **12**, 2825–2830 (2011).
- 622 [72] Pain, O. *et al.* Evaluation of polygenic prediction methodology within a reference-standardized framework.  
623 *PLoS genetics* **17**, e1009021 (2021).
- 624 [73] Fairley, S., Lowy-Gallego, E., Perry, E. & Flicek, P. The international genome sample resource (igsr)  
625 collection of open human genomic variation resources. *Nucleic Acids Research* **48**, D941–D947 (2020).
- 626 [74] Hanscombe, K. B., Coleman, J. R., Traylor, M. & Lewis, C. M. ukbtools: An r package to manage and  
627 query uk biobank data. *PLoS One* **14**, e0214311 (2019).

- 628 [75] Van Hout, C. V. *et al.* Exome sequencing and characterization of 49,960 individuals in the UK Biobank.  
629 *Nature* **586**, 749–756 (2020).
- 630 [76] Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence  
631 model. *Nature methods* **12**, 931–934 (2015).
- 632 [77] Scheipl, F., Greven, S. & Kuechenhoff, H. Size and power of tests for a zero random effect variance or  
633 polynomial regression in additive and linear mixed models. *Computational statistics & data analysis* **52**,  
634 3283–3299 (2008).
- 635 [78] Consortium, . G. P. *et al.* A global reference for human genetic variation. *Nature* **526**, 68 (2015).
- 636 [79] Davies, R. B. The distribution of a linear combination of  $\chi^2$  random variables. *Journal of the Royal*  
637 *Statistical Society: Series C (Applied Statistics)* **29**, 323–333 (1980).
- 638 [80] Kuonen, D. Miscellanea. saddlepoint approximations for distributions of quadratic forms in normal vari-  
639 ables. *Biometrika* **86**, 929–935 (1999).
- 640 [81] Lippert, C. *et al.* Fast linear mixed models for genome-wide association studies. *Nature methods* **8**, 833–835  
641 (2011).
- 642 [82] Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies.  
643 *Biostatistics* **13**, 762–775 (2012).
- 644 [83] Magno, R. & Maia, A.-T. gwasrapid: an r package to query, download and wrangle gwas catalog data.  
645 *Bioinformatics* **36**, 649–650 (2020).

## 5 Supplementary Methods

### 5.1 Ancestry scoring and ancestry-based variant pruning

We used GenoPred [72] to perform ancestry scoring of all UK Biobank participants. GenoPred uses an elastic net model based on the first 100 genetic principal components of the 1000 Genomes genetic data and super population assignments [78] to predict the genetic ancestry of individuals.

In the analysis including all ancestries we performed ancestry-based variant pruning, as follows. We limited the analysis to participants with complete covariates, and to variants with a study-wide MAF below 0.1% within those individuals. We then defined groups of individuals based on the ancestry prediction model. An individual was assigned to one of the 5 ancestry groups defined in the 1000 Genomes reference if the predicted probability for that ancestry was greater than 0.5. 182,288 participants were identified as having predominantly European ancestry (EUR), 4,302 South Asian ancestry (SAS), 3,775 African ancestry (AFR), 1,126 East Asian ancestry (EAS) and 308 admixed American ancestry (AMR). 172 individuals could not be placed in any of these groups, as all probabilities were below 0.5. These predictions were used to group individuals with similar genetic ancestry according to the 1000 Genomes and do not reflect ethnicity.

We then performed two-sided Fisher's exact tests to identify variants with large deviations from the study-wide EUR allele frequency in any of the other super populations, and excluded variants with  $p < 10^{-5}$  from the analysis. Variants selectively missing in any of the super populations were excluded.

The predictions from the ancestry prediction model were also used to define the group of participants which were used for the EUR-only analysis, where we applied a more stringent cutoff of  $p(\text{EUR}) > 0.995$ .

### 5.2 Variant weight calculation

All association tests we performed incorporated variant weights, which were derived from variant effect predictions. All variant weights we used are numbers between 0 and 1. For protein LOF variants all weights were set to 1. For missense variants, we calculated the weights as follows:

$$w_i = \frac{(1 - s_{i,SIFT}) + s_{i,Polyphen}}{2} \quad (3)$$

where  $w_i$  is the weight for variant  $i$ .  $s_{i,SIFT}$  and  $s_{i,Polyphen}$  denote the SIFT and Polyphen scores for variant  $i$ , respectively (potentially averaged across different transcript variants). This score can be interpreted as the average of the predicted probability of the variant being deleterious predicted by the two methods.

For splice variants, the weight  $w_i$  for a specific variant  $i$ , was set to the maximum of its four SpliceAI delta scores.

Regarding the predictions for the binding of RBPs, we proceeded as follows: While the experiments for the RBP QKI had been replicated in three cell lines, those for the other 5 RBPs had only been performed in a single cell line. As every replicate is a separate model output, this resulted in a total of 8 predictions for every genetic variant. We predicted the binding probability of each RBP to sequences centered on the major and minor

678 alleles, while applying 4bp shifts around the center. We averaged four predictions across these small shifts to  
679 reduce variability. Finally, we calculated variant effect predictions  $v_{ij}$  for each variant  $i$  and RBP-replicate  $j$  by  
680 subtracting the prediction for the reference allele ( $p_{ij,ref}$ ) from the prediction for the alternative allele ( $p_{ij,alt}$ )  
681 [76]:

$$v_{ij} = p_{ij,alt} - p_{ij,ref} \quad (4)$$

682 These variant effect predictions are numbers between  $-1$  and  $1$ , where the sign denotes a gain of binding ( $+$ )  
683 or a loss of binding ( $-$ ). They were used to determine variant weights and variant similarities during association  
684 testing (see below), where we set the weight  $w_i$  of variant  $i$  to the largest absolute value of  $v_i$ .

### 685 5.3 Score- and likelihood ratio test implementation

686 As stated in the main text, let  $N(\boldsymbol{\mu}; \boldsymbol{\Sigma})$  denote a multivariate Normal distribution with means  $\boldsymbol{\mu}$  and a variance-  
687 covariance matrix  $\boldsymbol{\Sigma}$ . We wish to jointly test the association of  $m$  genetic variants with a quantitative trait  $\mathbf{y}$   
688 for a sample of  $N$  observations (i.e., participants), while controlling for  $q$  covariates. Within the linear mixed  
689 model framework,  $\mathbf{y}$  can be modelled as follows [8, 9]:

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\alpha}; \sigma_e^2 \mathbf{I}_N + \sigma_g^2 \mathbf{K}_g) \quad (5)$$

690 Where  $\mathbf{X}$  is the  $N \times q$  covariate design matrix (fixed effect) and  $\boldsymbol{\alpha}$  is the vector of fixed-effect parameters.  
691 The variance-covariance matrix of  $\mathbf{y}$  is composed of the independently distributed residual variance ( $\mathbf{I}_N$  scaled  
692 by  $\sigma_e^2$ ) and the kernel-matrix  $\mathbf{K}_g$  (scaled by  $\sigma_g^2$ ), which captures the genetic similarity between individuals.  $\mathbf{K}_g$   
693 is a function of the  $N \times m$  matrix of mean-centered minor allele counts  $\mathbf{G}$  (random effect) of the genetic variants  
694 we wish to test.

695 In order to use efficient algorithms for estimating the parameters  $\sigma_e^2$  and  $\sigma_g^2$  and performing association tests,  
696 we require  $\mathbf{K}_g$  to be factored as a quadratic form [9, 11]:

$$\mathbf{K}_g = \phi(\mathbf{G})\phi(\mathbf{G})^T \quad (6)$$

697 Where the function  $\phi$  transforms  $\mathbf{G}$  into intermediate variables before performing the test.

698 The test statistic of the score test approximates the change of the log likelihood of a model when including  $\mathbf{K}_g$   
699 over the null model, which does not include  $\mathbf{K}_g$  ( $\sigma_g^2 = 0$ )[8]. We calculated test statistics using fast algorithms  
700 described in [11] and applied Davies's method for the calculation of p-values [79] with accuracy of  $10^{-7}$  and  $10^6$   
701 iterations. Where Davies's method returned p-values of 0, or in the rare cases where Davies method returned  
702 invalid (negative) p-values, we used saddle point approximation instead [80].

703 The test statistic of the restricted ratio test is twice the difference of the log restricted likelihood of the  
704 alternative model and the null mode [9]. We used FaST-LMM's LMM class [81] to fit the null and alternative  
705 models using restricted maximum likelihood and then calculated test statistics. To generate a null distribution

706 we sampled 100 test statistics for every LR test, using our own port of RLRsim [77] in Python. Finally, we fit  
707 a parametric null distribution  $\pi\chi_0^2 + (1 - \pi)a\chi_d^2$  with free parameters  $\pi$ ,  $a$  and  $d$  to the pooled simulated test  
708 statistics using log-quantile regression on the 10% of largest test statistics [11], and used this distribution to  
709 calculate p-values as described in [9]. We used separate null distributions for all variant-type to test-type and  
710 phenotype combinations (see details below).

711 The pooled null distribution is a uniform mixture of gene-specific null distributions. If we were able to obtain  
712 more than a single test statistic per gene, this distribution would not be guaranteed to control the gene-specific  
713 type-1 error rate. However, as it represents an average mixture across all tests, the average type-1 error rate  
714 across all genes remains controlled.

715 We compared this approach to using gene-specific null distributions (i.e., fitting a separate null distribution  
716 for every test and gene, similar to the method described in [12]), and found that they produced highly similar  
717 results (Supplementary Figures S9-S11).

718 To this end, we performed two analyses: First, we looked at genes close to or below the genome-wide  
719 significance threshold. We compared the Pearson correlation of the log10-p-values derived from the genome-  
720 wide pooled or gene-specific null distributions for genes with association p-values below  $10^{-6.5}$  in any of the  
721 initially performed tests. We did this separately for the different variant categories and test-types based on  
722 250,000 gene-specific samples from the null distribution. Second, for every phenotype and variant category,  
723 we randomly sampled 100 genes with p-values *above*  $10^{-6.5}$  (in any of the previously performed tests for that  
724 variant category) and cumulative minor allele counts of at least 5, and repeated the comparison (again based  
725 on 250,000 samples per test).

726 For associations close to or below the significance threshold, the average  $r^2$  was 0.999 for kernel-based tests,  
727 and 0.999 for gbvc tests. For non-significant associations, average  $r^2$  values were 0.9897 for kernel-based tests,  
728 and 0.999 for gbvc tests. We conclude that the pooled null distribution is a good approximation of the individual  
729 gene-level null distributions. An example illustrating this approach is shown in Supplementary Figure S12.

## 730 5.4 Gene-based variant collapsing tests

731 In gene-based variant collapsing, all qualifying variants overlapping a specific gene are collapsed into a single  
732 variable prior to association testing, i.e.,  $\phi(\mathbf{G})$  in Equation 6 returns an  $N \times 1$ -vector. We modified the  
733 approach in [13] by incorporating variant effect predictions as weights. Within a specific gene, any participant  
734 could carry 0 or more qualifying variants, where each variant  $i$  has a weight  $w_i$  (derived from variant effect  
735 prediction, see above). Specifically, the collapsed score is the largest weight of any of the variants observed for  
736 a specific participant, or 0 if no qualifying variants were observed for that participant. This score makes three  
737 assumptions: additive effects are negligible (or unrealistic), variants with larger weights dominate over those  
738 with smaller weights and all variants affect the quantitative trait in the same direction.

## 739 5.5 Functionally informed kernel-based tests

740 The kernels we used in this analysis follow the general form:

$$K_g = \mathbf{G}\mathbf{W}\mathbf{S}\mathbf{W}\mathbf{G}^T, \quad (7)$$

741 where  $\mathbf{W}$  is an  $m \times m$  diagonal matrix containing the square roots of variant weights on the diagonal and  
742 the  $m \times m$  matrix  $\mathbf{S}$  captures similarities between the genetic variants.  $\mathbf{G}$  is the  $n \times m$  matrix of mean-centered  
743 minor allele counts of the qualifying variants within the gene to be tested.  $\mathbf{S}$  can be interpreted as the variance-  
744 covariance matrix of regression coefficients of intermediate variables  $\mathbf{G}\mathbf{W}$ . We use  $\mathbf{W}$  and  $\mathbf{S}$  to incorporate  
745 variant effect predictions (and other variant annotations) into the association tests.

746 While a shared regression coefficient ( $\mathbf{S} = \mathbf{1}_m\mathbf{1}_m^T$ ) might be a poor assumption in some cases, so can  
747 completely independent regression coefficients ( $\mathbf{S} = \mathbf{I}_m$ ). The former, when substituted into (7), has been  
748 referred to as the weighted counting burden test, whereas the latter is commonly called the weighted linear  
749 kernel [82]. In our analysis, we define  $\mathbf{S}$  based on available prior knowledge and type of variant effect prediction.

750 **Missense** For the analysis of missense variants, we introduce the locally collapsing kernel. Local collapsing  
751 aggregates groups of variants into single variables before performing the association test. “Local” refers to the  
752 fact that the groups are defined by the proximity of variants in the DNA-, RNA- or amino acid sequence. We  
753 grouped variants if they affect the same exact amino acid position of a specific gene. Once the groups are  
754 defined, local collapsing can be expressed as a matrix multiplication:  $\mathbf{S} = \mathbf{C}\mathbf{C}^T$  and the kernel (7) becomes:

$$K_g = \mathbf{G}\mathbf{W}\mathbf{C}\mathbf{C}^T\mathbf{W}\mathbf{G}^T \quad (8)$$

755 Here  $\mathbf{C}$  is the  $m$ -variants by  $g$ -groups collapsing matrix. Therefore  $\mathbf{G}\mathbf{W}\mathbf{C}$  is the  $n \times g$  weighted locally  
756 collapsed genotype matrix (where the columns now represent amino-acid positions instead of single genetic  
757 variants). The columns of  $\mathbf{C}$  define the group assignments and directionality of variant effects. For every  
758 variant  $i$  from 1 to  $m$  with (potentially signed) variant effect  $v_i$  and group  $j$  from 1 to  $g$ ,  $c_{ig} = \text{sgn } v_i$  if  
759 variant  $i$  belongs to group  $j$ , else  $c_{ig} = 0$ . In our case, variant effect predictions were unsigned (all positive).  
760 The assumptions of the locally collapsing kernel are that variants within groups share a common regression  
761 coefficient once they have been scaled by and aligned with the direction of their variant effect predictions.

762 **RBP-binding** Sometimes there are no clearly defined groups of variants or multiple (potentially directional)  
763 variant effect predictions need to be accounted for at once and therefore variants can’t easily be collapsed. Given  
764 what we know about the location of variants and their predicted effects, we might still make assumptions about  
765  $\mathbf{S}$ . As long as  $\mathbf{S}$  is positive definite, we can find a suitable square root  $\mathbf{L}$  so that  $\mathbf{L}\mathbf{L}^T = \mathbf{S}$  using the Cholesky  
766 decomposition. In the association tests involving directional predictions for the binding of RNA-binding proteins  
767 we calculated  $\mathbf{S}$  by forming the element-wise product of two  $m \times m$  matrices:

$$S = LL^T = Q \circ R \quad (9)$$

768 Where  $Q$  captures the similarity of variants based on their variant effect predictions and  $R$  captures the  
769 similarity of variants based on their positions. Specifically, let  $v_i$  be the vector of variant effect predictions  
770 for variant  $i$ . Then the element  $q_{ij}$  of  $Q$  is the cosine similarity between  $v_i$  and  $v_j$ . We chose to model  
771 the position-dependent similarity with a Gaussian kernel. If  $x_i$  is the chromosomal position of variant  $i$ ,  
772  $r_{i,j} = \exp(-\gamma(x_i - x_j)^2)$ , where we set  $\gamma = -\frac{\log(0.5)}{50^2}$ . At this value of  $\gamma$  two variants that are 50bp apart  
773 have a similarity of 0.5, which decays rapidly as the distance increases. As both  $Q$  and  $R$  are positive definite  
774 matrices, so is  $Q \circ R$ . This kernel makes the assumption that variants that are in close proximity and have  
775 aligned variant effect predictions should affect the phenotype in the same direction.

## 776 5.6 sLRT detailed description

777 **Missense** For missense variants, we iterated over all genes and performed score tests using gene-based variant  
778 collapsing and kernel-based tests (locally collapsing kernel), i.e., the diagonal elements  $w_{ii}$  of  $W$  in Equation 7  
779 contained the square roots of the impact scores of variants. If either score test p-value was nominally significant  
780 ( $p < 0.1$ ) we also performed the following steps: 1. Calculation of restricted likelihood ratio test statistics  
781 (sLRT), 2. gene-based variant collapsing combining both missense and loss of function variants in a joint test,  
782 3. concatenation of the collapsed pLOF variable to the locally collapsed weighted matrix of missense variant  
783 minor allele counts ( $GWC$ , Equation 8) and a joint kernel-based RLRT. For all likelihood ratio tests, we  
784 simulate 100 gene-specific test statistics from the null distribution each.

785 Once all genes were processed for a specific phenotype, we fit separate null distributions to the pooled  
786 simulated test statistics for each of the 4 groups of likelihood ratio tests: collapsed missense variants (gbvc),  
787 jointly collapsed pLOF and missense variants (gbvc), locally collapsed missense variants (kernel-based), locally  
788 collapsed missense variants concatenated with collapsed pLOF variants (kernel-based). P-values were then  
789 calculated for all tests based on those distributions. Finally, the p-values for the two kernel-based tests, and  
790 the two gbvc tests were combined using the Cauchy combination test, resulting in a single kernel-based and a  
791 single gbvc test per gene.

792 We used the locally collapsing kernel in the kernel-based association tests for missense variants, as it had given  
793 more unique associations and overall slightly lower p-values for the most significant genes in initial experiments  
794 on the 50k WES release, and was more interpretable compared to other approaches.

795 **Splicing** For splice-variants we performed score tests using gene-based variant collapsing and the linear  
796 weighted kernel for all genes. Again, if either of the two score tests were nominally significant ( $p < 0.1$ ),  
797 we performed likelihood ratio tests (sLRT). As we did for missense variants, we then also performed combined  
798 association tests with protein loss of function variants using both gene-based variant collapsing and a kernel-  
799 based LRT. For the kernel-based test, we concatenated the protein LOF indicator variable to the matrix of

800 weighted minor allele counts  $\mathbf{GW}$  (Equation 7, where  $\mathbf{S} = I_m$ ). In the cases where a variant was annotated  
801 both as a splice-variant and pLOF variant, we treated it as a pLOF variant in the joint tests. As we did for  
802 missense variants, after calculating p-values using 4 separate null distributions for every phenotype, we com-  
803 bined the two kernel-based tests and the two gene-based collapsing tests into single tests using the Cauchy  
804 combination test.

805 **RBP-binding** For variants predicted to alter the binding of RBPs we only performed kernel-based association  
806 tests using the kernel in Equation 7, where we used the largest absolute value of the variant effect predictions as  
807 the weights, and calculated  $\mathbf{S}$  as described above in Equation 9. We iterated over all genes and performed gene-  
808 based score tests. Because the DeepRiPe variant effect predictions are strand-specific, we did this independently  
809 for genes on the forward or reverse strands. If the score test for a specific gene was nominally significant  
810 ( $p < 0.1$ ), we performed the likelihood ratio test for that gene (sLRT). If the variants tested also included  
811 variants annotated as protein loss of function variants, we removed them and repeated the tests to avoid false  
812 positives.

## 813 5.7 Conditional association tests

814 For significant associations after multiple testing correction, we performed conditional association tests. For  
815 every significant gene-biomarker association, we identified single variants significantly associated with the same  
816 biomarker within  $\pm 500\text{kb}$  of the gene start position, based on the summary statistics provided by [23]. We  
817 then conditioned on the single variant with the lowest p-value (if any) by incorporating it as a covariate in a  
818 gene-specific null model (in case of ties, the closest variant to the gene start position was chosen). We then  
819 fit the alternative model, calculated the model likelihoods and RLRT test statistics, and simulated 250,000  
820 gene-specific RLRT test statistics for every alternative model (i.e., combinations of variant- and test-types).  
821 We fit parametric null distributions to these test statistics using the 10% of largest test statistics (as described  
822 above), and calculated p-values based on these null distributions. We then combined the p-values using the  
823 CCT (if combined tests with pLOF variants were performed). Conditional p-values and the variants that were  
824 conditioned on are reported in in Supplementary Table S1.

## 825 5.8 Cross-referencing against GWAS databases

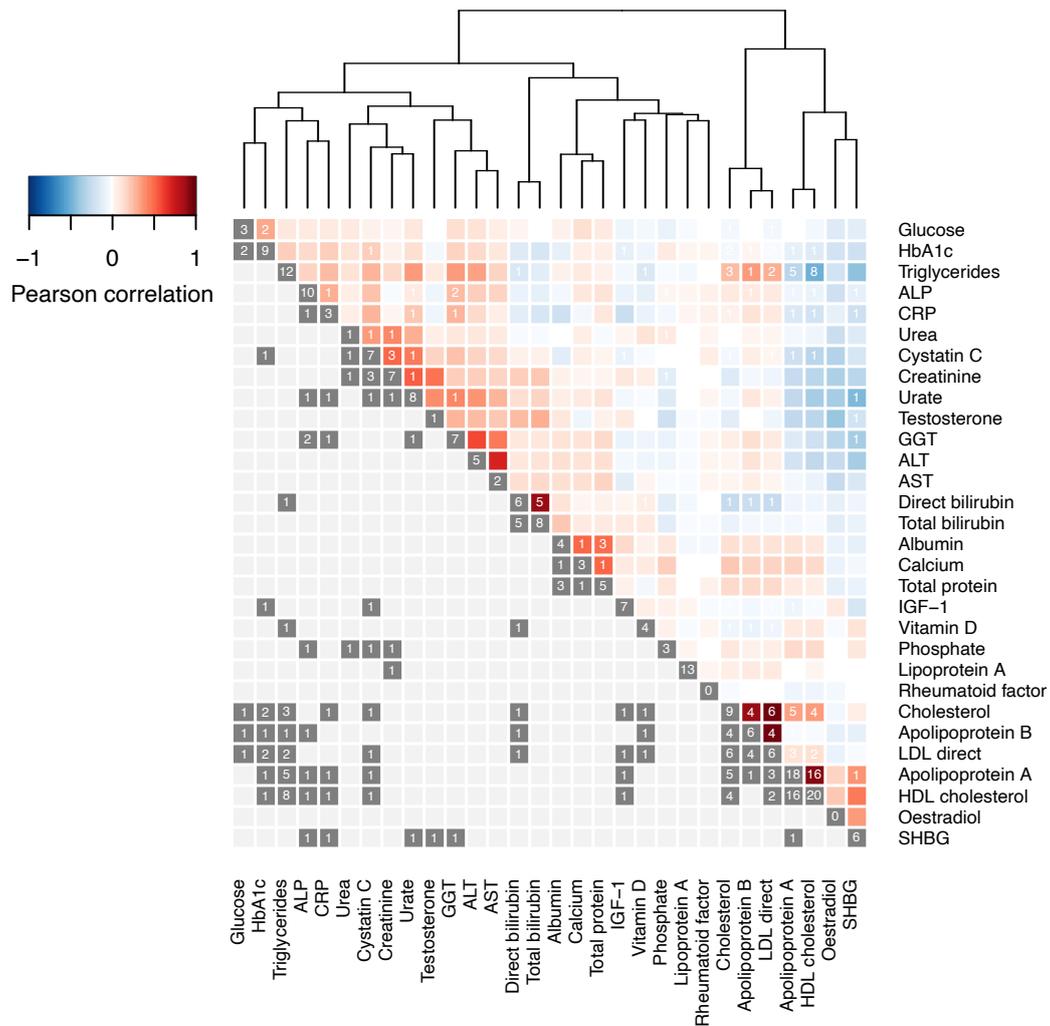
826 We queried the NHGRI-EBI GWAS Catalog [2] and PhenoScanner [26, 27] in order to see if single variants  
827 within the genes we found significantly associated with a specific biomarker had already been reported to be  
828 associated with that biomarker. For each gene, we submitted region queries using the gene boundaries with  
829 the `gwasrapidd` [83] (v0.99.11) and `phenoscanner` (v1.0) R-packages. For PhenoScanner, we set the p-value  
830 threshold to  $10^{-7}$ . Matching our results to those contained in these databases required us to define a mapping  
831 of UK Biobank biomarkers to the Experimental Factor Ontology (EFO) terms used in those databases. This  
832 mapping is provided in Supplementary Table S2. Additionally, as EFO terms for PhenoScanner were not always  
833 defined, we performed the following matching: "Apolipoprotein B" (UKB phenotype) to "APOB apolipoprotein

834 B" (PhenoScanner trait), "Cystatin C" to PhenoScanner traits "log eGFR cystatin C", "Serum cystatin c  
835 estimated glomerular filtration rate eGFR" and "Cystatin C in serum", and "Urea" to "Renal function related  
836 traits urea".

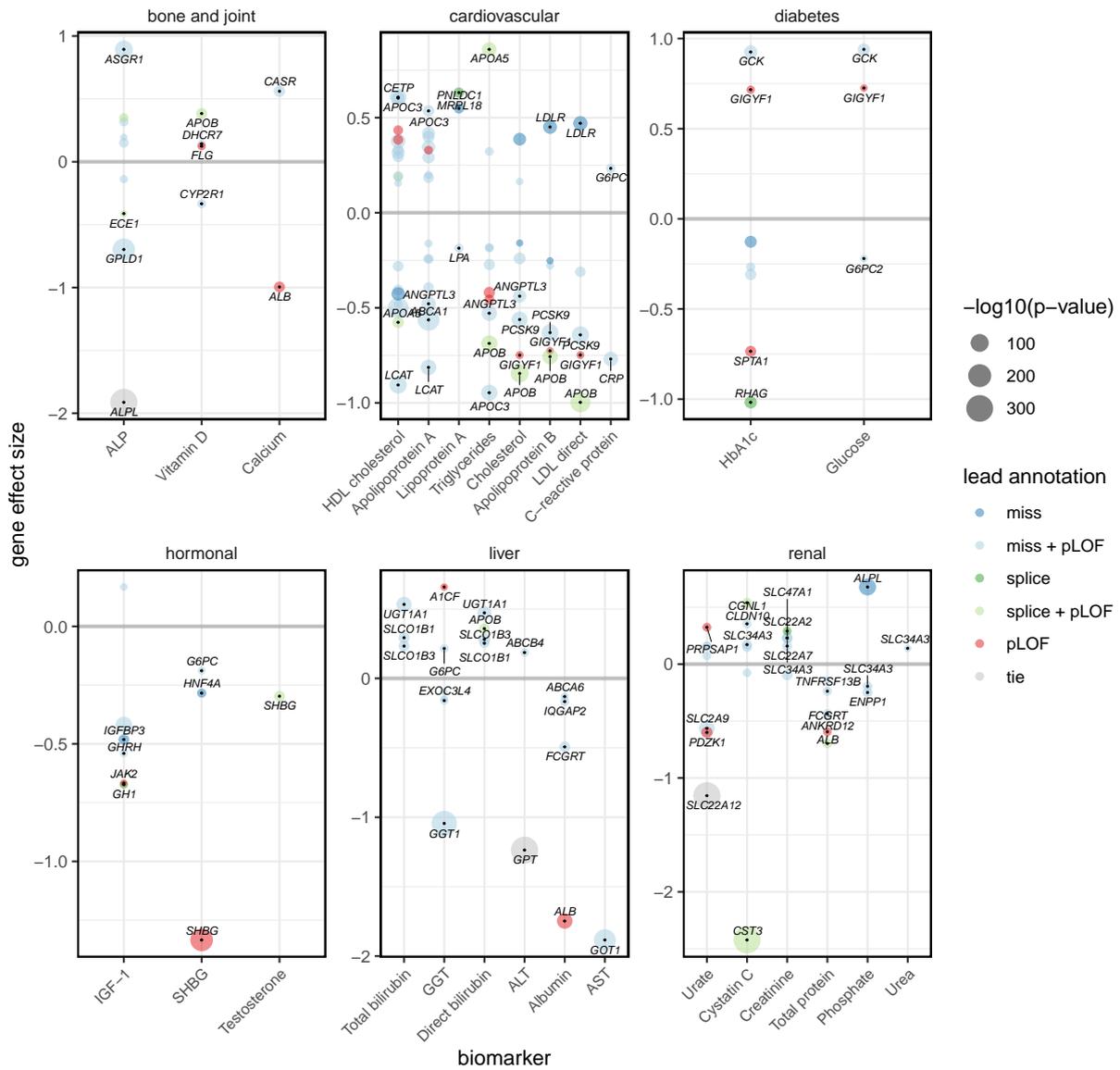
## 837 **5.9 PIEZO1 L2277M association tests**

838 We used the ancestry classifications described above to define a group of individuals of SAS ancestry, and one  
839 of extended EUR ancestry (both using a cutoff of  $> 0.5$  in the ancestry classification model). This is a less  
840 stringent cutoff than that used in the EUR-analysis for all biomarkers and increased the number of observed  
841 carriers in the EUR-ancestry group from 5 to 21, all heterozygous. We used the same covariates as in the  
842 all-ancestry analysis.

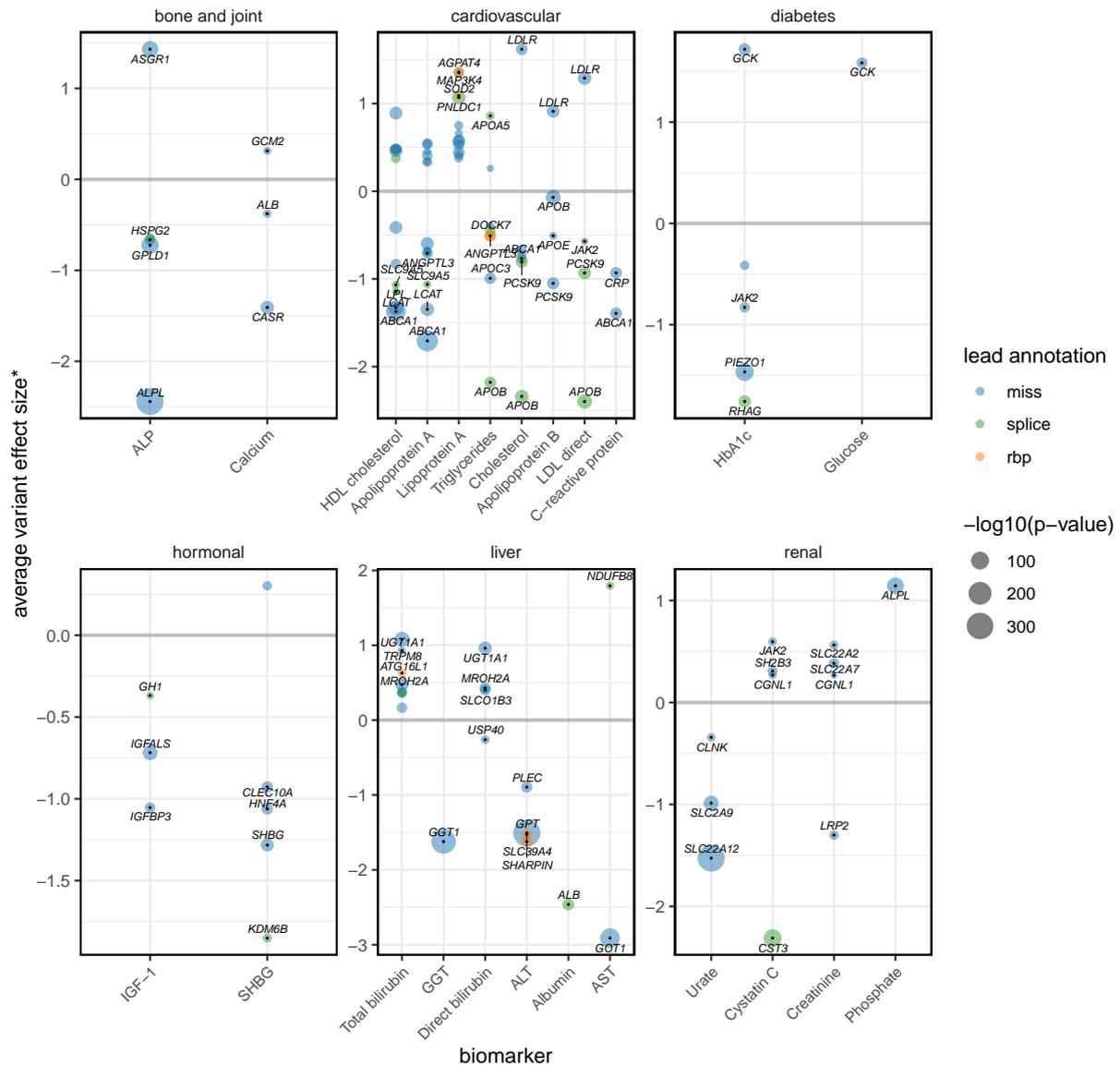
843 Genotypes were derived from exome sequencing and we performed association tests using the score test with  
844 ancestry-specific null-models. For the association tests in the SAS group, we performed conditional tests by  
845 conditioning the test statistic on the genotypes of the 16:88784993:C:G variant, which were also derived from  
846 exome sequencing.



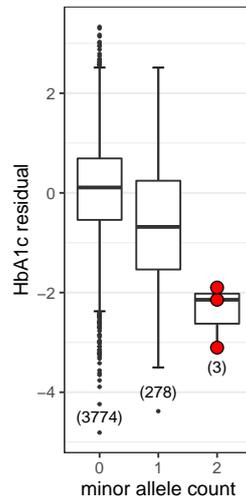
Supplementary Figure S1: **Biomarker correlation and number of hits.** Heatmap showing Pearson correlation between pre-processed biomarkers (upper triangle) and number of significant associations (cell notes). Rows and columns are clustered using complete linkage on the Euclidean distances of the correlation matrix between phenotypes (dendrogram). While some even weakly (anti-)correlated biomarkers share significant associations (e.g., Cholesterol and Glucose, gene: *GIGYF1*), other highly correlated markers do not share significant associations (e.g., GGT, ALT, AST).



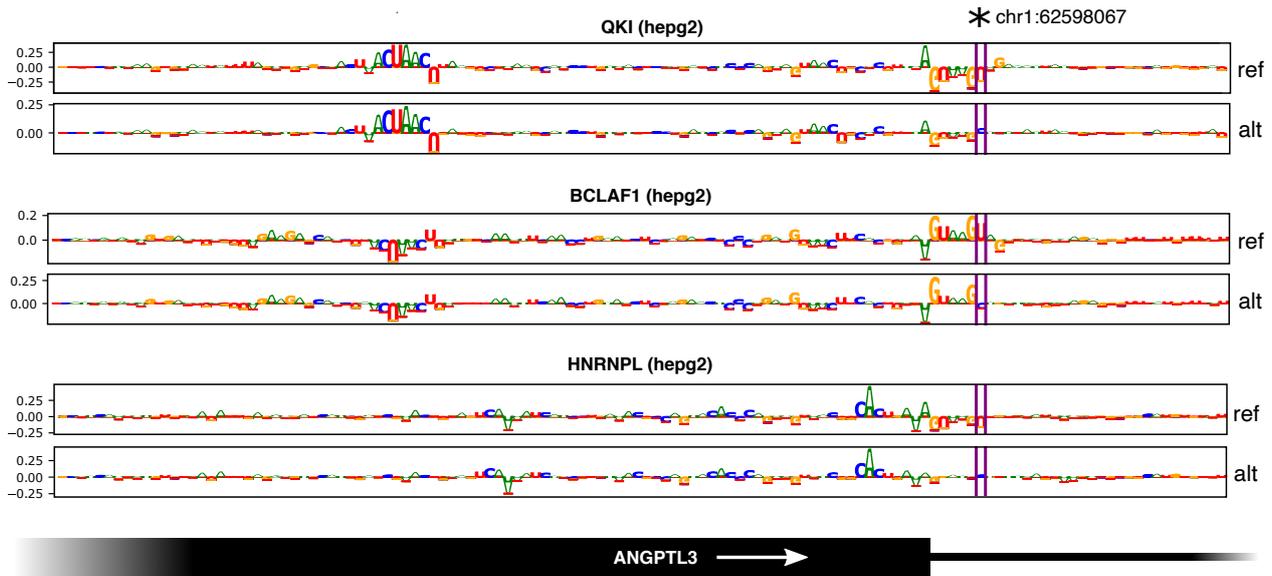
Supplementary Figure S2: **Gene-based variant collapsing results overview.** Collapsing variants allows defining gene effect sizes. Bubble plots showing the gene effect sizes (y-axis) of significant associations for each biomarker (x-axis). The four genes with largest absolute effect sizes are labeled for each biomarker. Larger bubble size indicates higher significance. P-values and effect sizes are those given by the most significant variant effect category (lead annotation). In case of ties ( $p = 0$ , gray) the average effect size across annotations is shown. Effect sizes are calculated on covariate-corrected quantile transformed phenotypes.



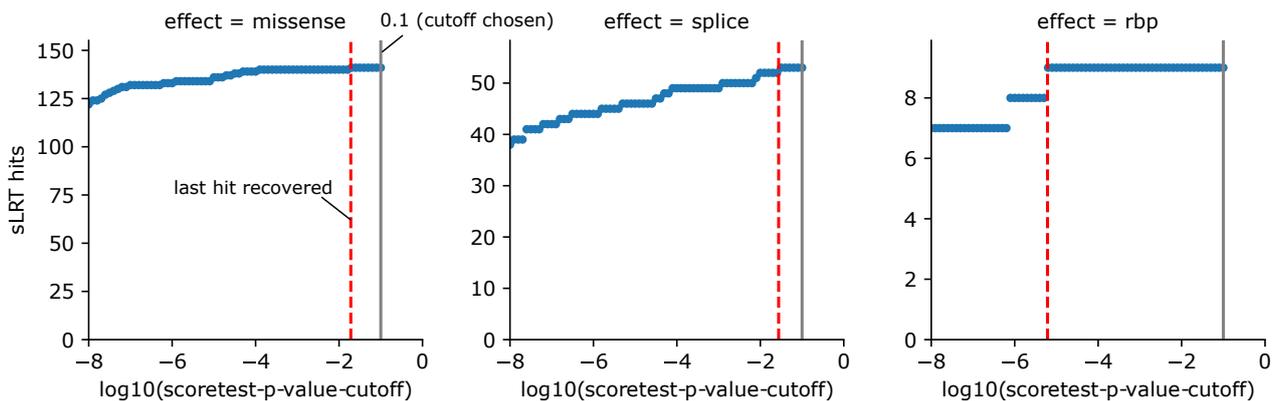
Supplementary Figure S3: **Kernel-based tests results overview.** We calculated the average effect size for \*variants with single-variant p-values below  $10^{-5}$  (score test) within significant genes found by kernel-based tests if the cumulative minor allele count across these variants was at least 5. Bubble plots showing these average effect sizes (y-axis) for each biomarker (x-axis). The four genes with largest average effect sizes are labeled for each biomarker. Larger bubble size indicates higher significance of the gene-based test. P-values and average effect sizes are those given by the most significant variant effect category (lead annotation). Effect sizes are calculated on covariate-corrected quantile transformed phenotypes.



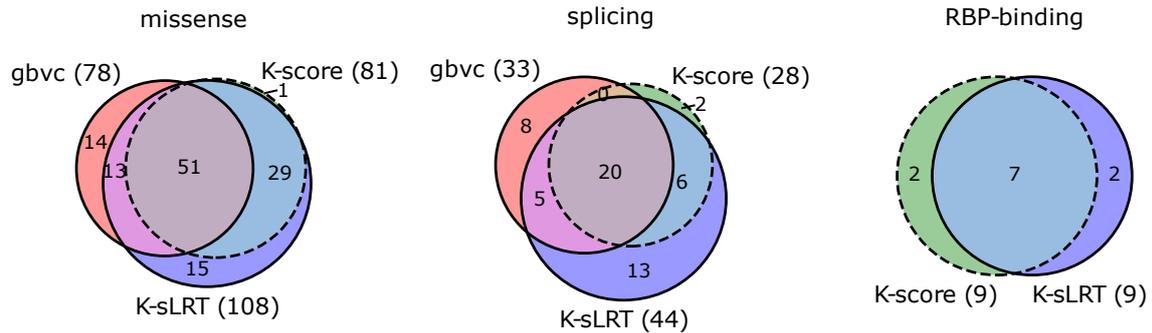
Supplementary Figure S4: ***PIEZO1* L2277M, (16:88716656:G:T)**. Dosage plot of *PIEZO1* L2277M in individuals of inferred South Asian ancestry. Numbers in brackets denote the number of carriers of 0, 1 or 2 minor alleles (x-axis). The three only homozygous carriers are shown in red. The y-axis contains the covariate-adjusted quantile transformed values for HbA1c. The lower and upper hinges indicate first and third quartiles, whiskers extend to the largest/lowest values no further than  $1.5 \times \text{IQR}$  away from the upper/lower hinges and black points denote outliers.



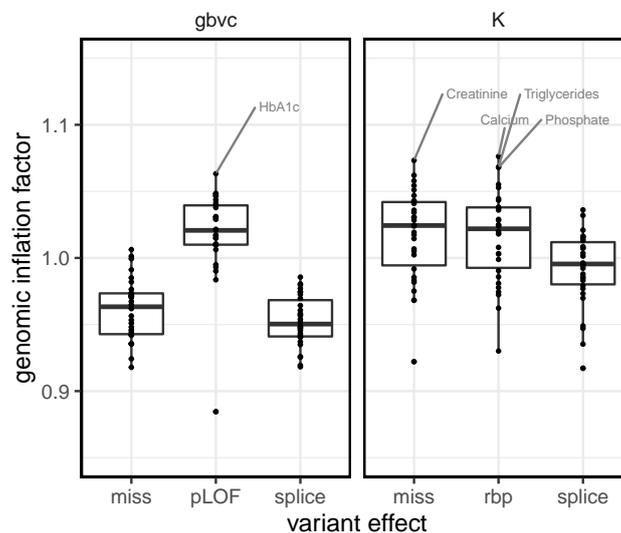
Supplementary Figure S5: **Attribution maps.** The variant 1:62598067:T:C at one-based position chr1:62598067 makes DeepRiPe predict increased binding probabilities for HNRNPL and QKI, and decreased probability for BCLAF1. Attribution maps for reference (ref) and alternative (alt) sequences as described in [25] highlight important nucleotides proximal to an ANGPTL3 exon boundary. The predictions for QKI depend positively on an upstream QKI binding motif (ACUAAC), and negatively on the splice donor signal (GUAAGU). The pattern is inverted for BCLAF1. Weakening of the splice signal by the alternative variant increases predicted binding probabilities for QKI and HNRNPL.



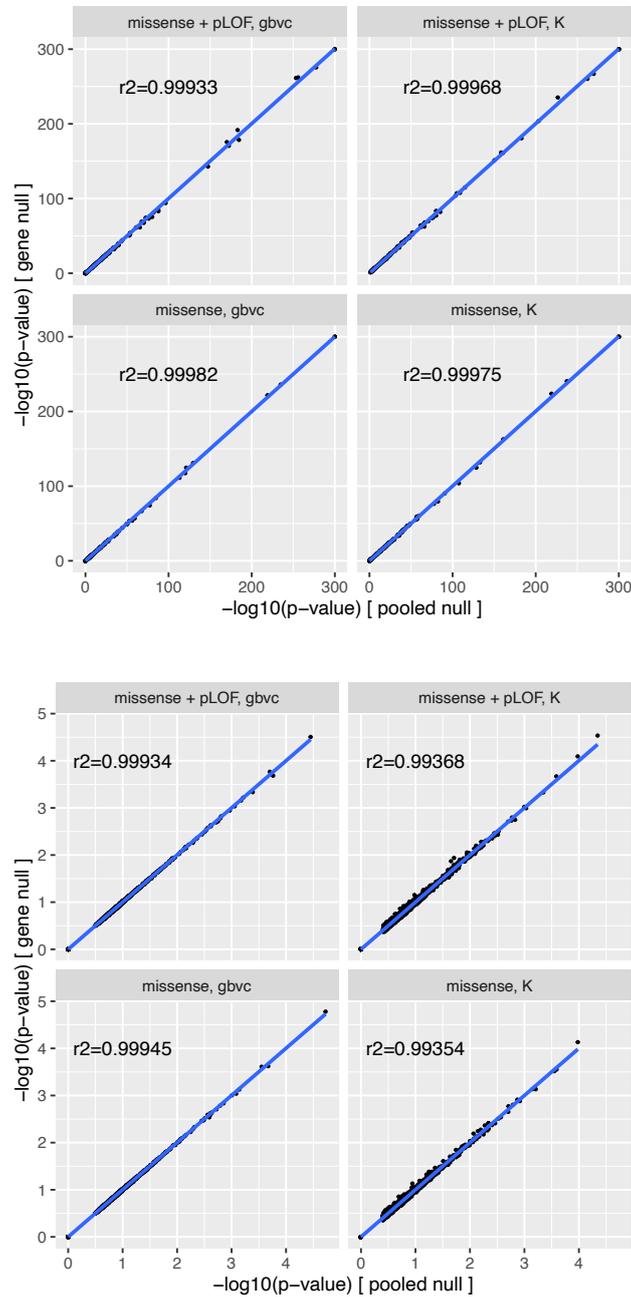
Supplementary Figure S6: **sLRT number of significant gene-biomarker associations vs. score test cutoff.** We consider the scenario of performing five score tests per gene and biomarker (one collapsing and one kernel-based test for missense and splice variants, and one kernel-based test for RBP-variants), resulting in 2,540,128 tests genome-wide and a Bonferroni corrected p-value threshold of  $1.96 \times 10^{-8}$  ( $\alpha = 0.05$ ). The plots show the number of significant associations found by the sLRT depending on the nominal significance cutoff chosen for the score tests (which determine whether the LRT is performed). We used the cutoff of 0.1 (gray vertical line) in our analysis.



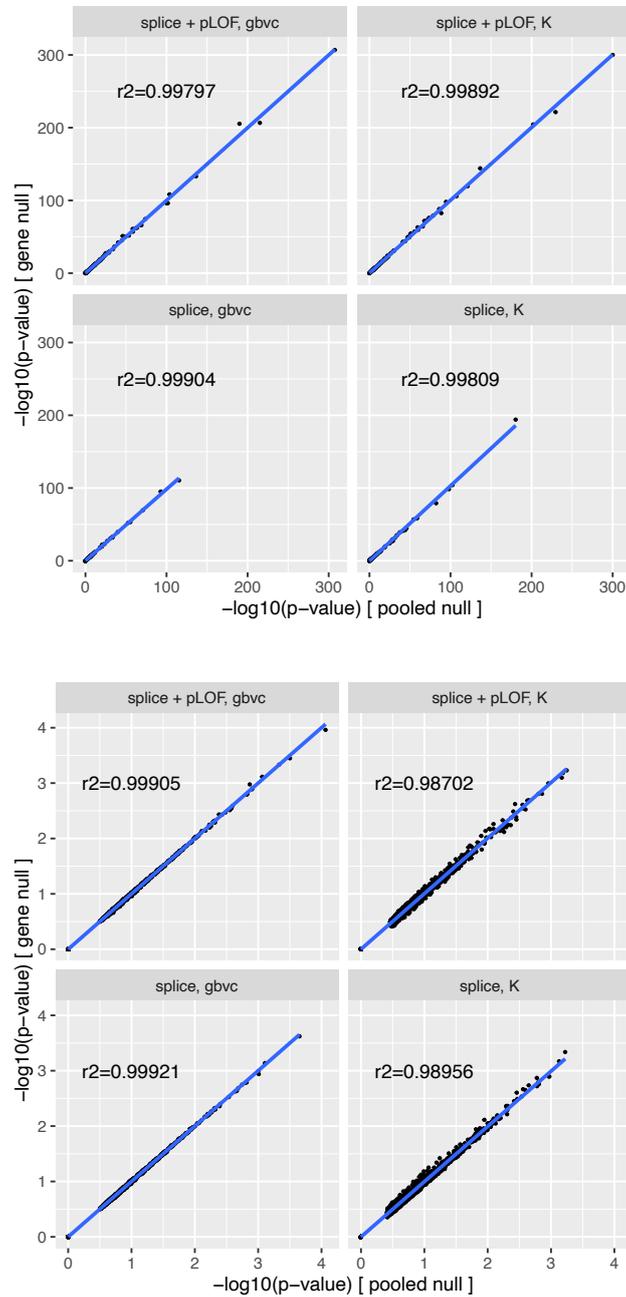
Supplementary Figure S7: **Kernel-based sLRT vs. score test comparison.** We consider the scenario of performing five score tests per gene and biomarker (one collapsing and one kernel-based test for missense and splice variants, and one kernel-based test for RBP-variants), resulting in 2,540,128 tests genome-wide and a Bonferroni corrected p-value threshold of  $1.96 \times 10^{-8}$  ( $\alpha = 0.05$ ). Venn diagrams showing the significant locus-biomarker associations identified by the kernel-based score test (K-score), kernel-based sLRT (K-sLRT) and gene-based variant collapsing (gbvc). For missense and splice variants, the kernel-based sLRT identified more significant associations than the kernel-based score test. A large fraction of these additional associations was also found by gbvc tests.



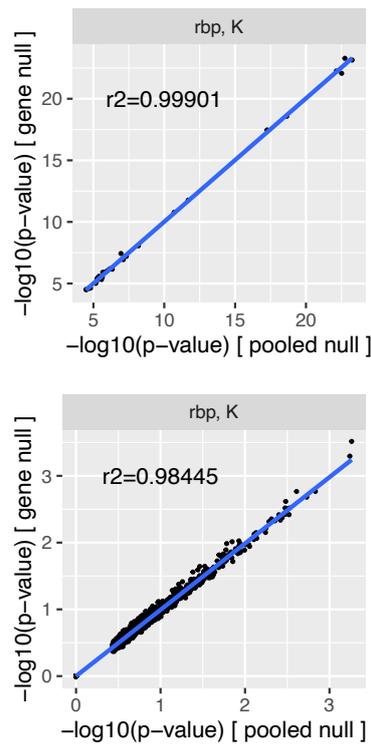
Supplementary Figure S8: **Genomic inflation factor across models.** We calculated  $\lambda_{GC}$  for all tests that were performed exome-wide in the all-ancestry analysis. Boxplots showing  $\lambda_{GC}$  across all 30 phenotypes (y-axis) against the different variant categories and types of association tests. All values refer to the sLRT, except for gbvc-pLOF, where we only performed the score test. Left: gene based variant collapsing (gbvc); Right: kernel-based tests (K). QQ-plots for all models that resulted in at least one significant association are given in the Supplementary Data.



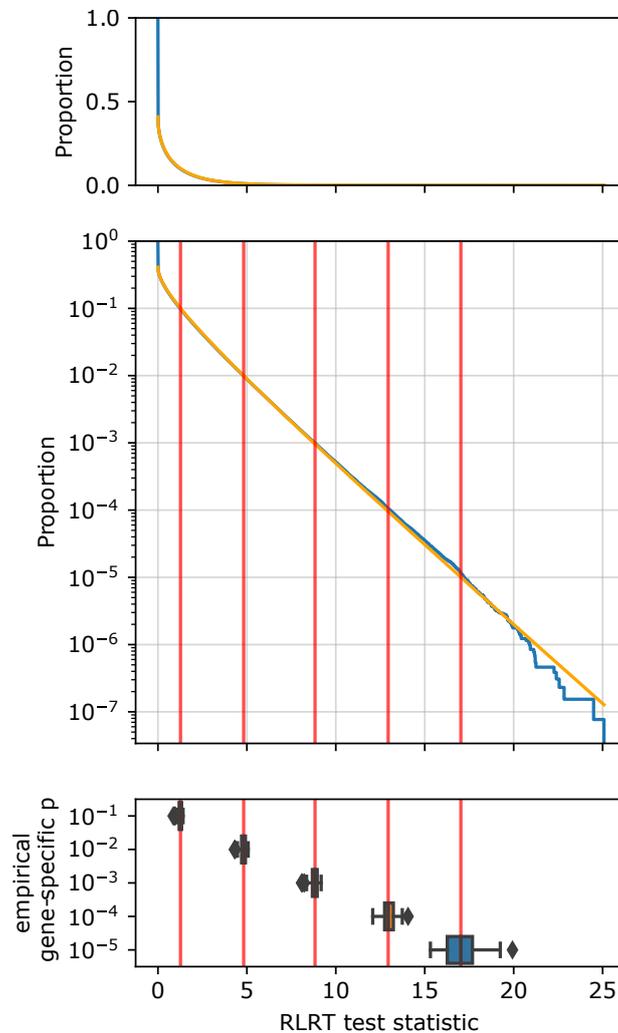
Supplementary Figure S9: **RLRT p-value comparison (gene-specific vs. pooled null distribution) for tests with missense variants.** Scatter plots showing the negative log<sub>10</sub>-p-values produced by pooled null distributions vs. those from gene-specific null distributions for associations close to or below the significance threshold (top) and randomly selected genes (bottom). Blue lines indicate the linear regression fit.  $r^2$ : r-squared values calculated for the points that do not lie in the origin.



Supplementary Figure S10: **RLRT p-value comparison (gene-specific vs. pooled null distribution) for tests with splice variants.** Scatter plots showing the negative log<sub>10</sub>-p-values produced by pooled null distributions vs. those from gene-specific null distributions for associations close to or below the significance threshold (top) and randomly selected genes (bottom). Blue lines indicate the linear regression fit.  $r^2$ : r-squared values calculated for the points that do not lie in the origin.



Supplementary Figure S11: **RLRT p-value comparison (gene-specific vs. pooled null distribution) for tests with predicted RBP-binding altering variants.** Scatter plots showing the negative log<sub>10</sub>-p-values produced by pooled null distributions vs. those from gene-specific null distributions for associations close to or below the significance threshold (top) and randomly selected genes (bottom). Blue lines indicate the linear regression fit. *r*<sup>2</sup>: r-squared values calculated for the points that do not lie in the origin.



Supplementary Figure S12: **RLRT pooled null distribution vs empirical gene-specific quantiles example.** For the kernel-based RLRT with missense variants and phenotype triglycerides, we sampled 250,000 test statistics each for 52 genes and fit a  $\chi^2$ -mixture distribution (parametric null) to the pooled test statistics ( $0.59 \times \chi_0^2 + (1 - 0.59) \times 0.96 \times \chi_{0.98}^2$ ). The top two panels show the empirical inverse cumulative distribution function of the pooled test statistics (y; blue line), against the value of the test statistic (x). The survival function of the non-zero mixture component of the parametric null is overlaid in orange. In the lower panel, the box plot shows the gene-specific quantiles of test statistics (x) corresponding to gene-specific empirical p-value thresholds of  $10^{-1}$ ,  $10^{-2}$ ,  $10^{-3}$ ,  $10^{-4}$  and  $10^{-5}$  (y) for the 52 genes. Red vertical lines denote the median values, and are extended to the plot above, showing that the mixture distribution accurately captures the median gene-specific quantiles in the tail.