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

Remo Monti[1,2], Pia Rautenstrauch[2,3], Mahsa Ghanbari[2], Alva Rani James[1], Uwe Ohler[2,3,a], Stefan Konigorski[1,4,a], and Christoph Lippert[1,4,a,*]

[1] *Digital Health - Machine Learning, Hasso Plattner Institute, University of Potsdam, 14482, Potsdam, Germany*
[2] *Berlin Institute for Medical Systems Biology, Max Delbrueck Center for Molecular Medicine, 10115, Berlin, Germany*
[3] *Department of Biology, Humboldt University, 10115, Berlin, Germany*
[4] *Hasso Plattner Institute for Digital Health, Icahn School of Medicine at Mount Sinai, New York 10029-6574, NY, USA*
[*] *corresponding author*
[a] *these authors contributed equally*

Abstract

Here we present an exome-wide rare genetic variant association study for 30 biomarkers in 191,640 individuals in the UK Biobank. We perform gene-based association tests for separate functional variant categories to increase interpretability and identify 201 significant gene-biomarker associations, which include novel associations such as *GIGYF1* with diabetes markers. In addition to performing 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 powerful and computationally efficient combination of the likelihood-ratio and score tests that found 32% more associations than the score test alone. Kernel-based tests identified 12-31% more associations than their gene-based collapsing counterparts with large overlaps, 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 this approach to identify potential novel gain of function variants in *PIEZO1*, and interpret a position-specific association of *ABCA1*-variants with inflammation marker CRP. Our results show the benefits of separately investigating different functional mechanisms when performing rare-variant association tests, and highlight the strengths of biomarker panels for large biobanks.

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 lie in non-coding regions of the genome [2]. Furthermore, effect sizes of common variants tend to be small, as

variants with large detrimental effects are selected against, which limits their frequency [3, 4].

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

While gene-based variant collapsing performs best in the presence of many causal variants with effect sizes that point in the same direction (e.g. increasing risk for disease), kernel-based tests have advantages in cases of opposing effects and fewer causal variants [7]. To increase the fraction of causal variants, exome-wide RVAS that use variant collapsing have defined qualifying variants based on annotations such as allele frequencies or variant effect predictions, and excluded all other observed variants from the association tests [6, 12, 13, 14]. These studies have mostly focused on non-synonymous variants, where software tools identify protein truncating variants and distinguish between benign and potentially deleterious missense variants [15, 16, 17, 18].

Here, we perform an extensive RVAS using exome sequencing data from the UK Biobank [19]. For approximately 190,000 individuals, 30 quantitative biomarkers provide objectively quantifiable measures related to the health status of individuals [20], making them attractive phenotypes for genome-wide association studies [21]. We go beyond the collapsing tests for coding variants described above and explore the use of kernel-based association tests and deep-learning-derived effect predictions for gene regulatory variants, namely for splicing [22] and the binding of RNA-binding proteins (RBPs)[23].

Specifically, we use quantitative functional variant effect predictions to group and weigh variants in gene-based association tests and increase interpretability. The greater flexibility of kernel-based tests allowed us to design variant-category-specific tests and combine collapsing and non-collapsing approaches in the same test. For kernel-based tests, we show that a computationally efficient combination of the score test and the LRT identifies 32% more significant associations on average compared to the score test alone. We find 201 significant gene-biomarker associations in total, of which 40% have not been previously reported to GWAS databases [24, 25]. Finally, we interpret associations that were only found for specific variant categories, or associations for which gene-based variant collapsing and kernel-based tests gave vastly different results.

2 Results

2.1 Data description and workflow

We performed an RVAS of 30 quantitative serum biomarkers in UK Biobank 200k WES release [19]. These biomarkers contain established disease risk factors, diagnostic markers, and markers for phenotypes otherwise not well assessed in the UK Biobank cohort. We roughly categorized these markers into cardiovascular, bone and joint, liver, renal, hormonal and diabetes markers (Supplementary Table S1). After removing related individuals and restricting the analysis to those with no missing covariates 192,352 participants remained. 16,737,187 rare (MAF $< 0.1\%$) variants were observed in this subset and passed basic quality criteria (Methods). The median sample size for the biomarkers was 182,144 and ranged from 16,022 (Rheumatoid factor) to 183,105 (Alkaline phosphatase) (Supplementary Table S1). 191,640 participants had at least one measured biomarker.

We used functional variant effect predictions to group variants into categories and perform functionally informed gene-based association tests. Specifically we chose to investigate strict protein loss of function (pLOF, e.g. frame shift or protein truncating variants) variants, missense variants, splice-altering variants, and variants predicted to change the binding of RNA-binding proteins (Figure 1, Methods). We treated these categories separately during association testing to increase interpretability, resulting in multiple tests per gene, which we refer to as separate *models*. Specifically, we adapted either kernel-based tests, gene-based variant collapsing, or both types of association test depending on the variant effect category (Methods). We make variant effect predictions for all variants in the UK Biobank 200k exome release available[1], as well as our analysis pipeline[2] and software used to perform association tests[3].

2.2 Functionally informed association tests

Protein loss of function We predicted the effects of genetic variants on protein coding genes using the Ensembl variant effect predictor [15] and found 475,732 pLOF variants with a median of 20 pLOF variants per gene (Figure 2, Methods). For pLOF variants, we assumed a large fraction of potentially causal variants, and that variants within the same gene should by and large affect the phenotype in the same direction. For these reasons, we performed tests using gene-based variant collapsing tests (Methods), and found 92 significant associations originating from 53 distinct genes across the genome.

Missense We defined 1,836,348 high-impact missense variants based on PolyPhen-2 [16] and SIFT [17] (Methods). 18,420 genes contained at least one high-impact missense variant, with a median of 73 high-impact missense variants observed per gene.

We hypothesized that missense variants in the same gene might have both trait-increasing and trait-decreasing effects, and that there might be fewer causal variants. For these reasons we performed not only a gene-based variant collapsing test, but also a kernel-based association test. We designed a missense-specific

[1] https://github.com/HealthML/ukb-200k-wes-vep
[2] https://github.com/HealthML/faatpipe
[3] https://github.com/HealthML/seak

kernel that collapses variants locally by amino acid position, which affected 20% of variants (Methods). We identified 101 significant associations using gene-based variant collapsing, and 128 using kernel-based association tests, with an overlap of 88. The total of 141 associations identified by either model originated from 78 distinct genes.

Combining pLOF and missense As we expected many missense variants to effectively lead to a loss of function, we combined pLOF and missense variants in additional models. We performed both a joint variant collapsing test, and a kernel-based test combining gene-based collapsing of pLOF variants and local collapsing of missense variants (Supplementary Methods). We identified 148 significant associations with gene-based variant collapsing, and 167 with kernel-based association tests, with an overlap of 125 in 102 genes. Combining missense variants and protein LOF variants provided 23 unique associations which were not found when testing variants in either category alone, and which where also not found by other models.

Splicing We located 775,349 potentially splice-altering rare single nucleotide variants in 17,168 genes by cross referencing against published SpliceAI variant effect predictions ([22]; Methods). The median number of variants per gene was 32. We hypothesized that these splice variants could have complex downstream consequences and decided to compare both gene-based variant collapsing and kernel-based association tests. For kernel-based tests, we used the weighted linear kernel [8].

We identified 38 significant associations with gene-based variant collapsing in 29 distinct genes. As our definition of pLOF variants included variants that directly hit annotated splice donor/acceptor sites, there was a considerable overlap of 98,017 variants between these annotations (21% of all pLOF variants). We therefore expected (and found) large overlaps (32, 84%) in the significant associations for pLOF and splice variants using gene-based variant collapsing. Kernel-based tests identified 50 significant associations in 35 genes. These included 32 hits already identified by pLOF gene-based collapsing, but a larger number of hits not identified by other models.

Combining pLOF and splicing We expected many predicted splice variants to lead to a loss of function, and therefore explored joint tests with pLOF variants. We applied both a joint variant collapsing test, and a kernel-based test that only collapses pLOF variants (Supplementary Methods). Joint gene-based collapsing of splice variants with pLOF variants identified 78 associations, of which 68 (87%) had already been identified by our protein LOF analysis. On the other hand, kernel-based association tests identified 94 significant associations of which 74 (79%) had already been found using pLOF variants. While combining splice and pLOF variants did yield the lowest p-values across all models for 14 significant associations, it did not uniquely identify additional associations which hadn't been found by other models already.

RBP-binding Splicing is only one of several eukaryotic post-transcriptional regulatory mechanisms mediated by interactions of RNA-binding proteins (RBPs) with their target RNAs. As the UK Biobank WES data also contain variants in non-protein-coding parts of mRNAs, namely in introns (41.4%) and UTRs (5.2%), we

reasoned that we may be able to identify variants with effects on gene regulation such as those mediated by differential binding of RBPs. Specifically, we investigated if changes in the binding of RBPs predicted by DeepRiPe [23] could be associated with biomarker levels. The six RBPs QKI, MBNL1, TARDBP, ELAVL1, KHDRBS1 and HNRNPD were selected based on their binding preferences (introns, exons)[26], the high performance of the model to predict genuine target sites for these RBPs, and the reported presence of clear binding sequence motifs. We predicted variant effects for these RBPs and identified 395,462 variants with large predicted effects in 17,459 genes, with a median of 13 variants per gene (Methods).

As we expected a low number of causal variants and potentially opposing effect sizes, we only performed kernel-based association tests and identified 10 significant associations in 9 distinct genes.

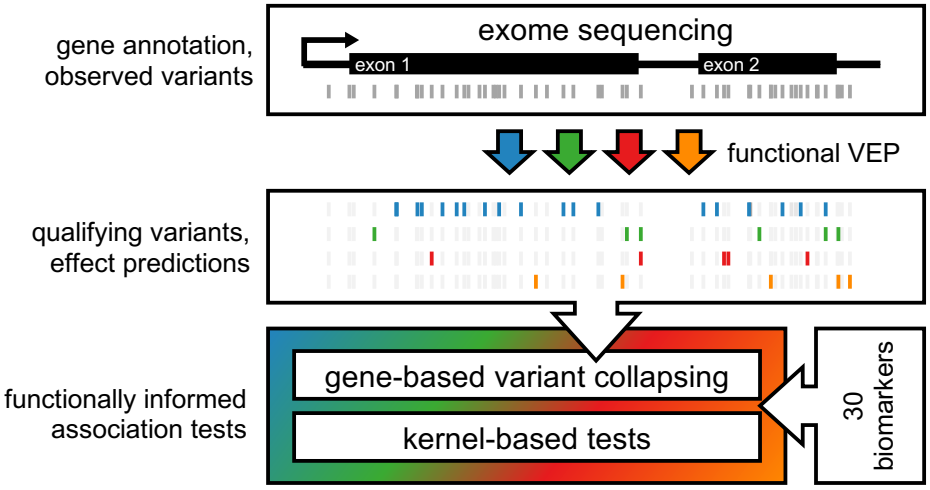


Figure 1: **Rare-variant association testing pipeline.** Exome sequencing measures exon-proximal genetic variants. All variants are subjected to functional variant effect prediction (VEP). Qualifying variants are determined based on the variant effect predictions and minor allele frequencies (MAF<0.1%) 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.

2.3 Integrative analysis overview

Merging the results from all models yielded a total of 201 gene-biomarker associations originating from 120 distinct genes (Supplementary Table 2). We found at least one significant association for all but three biomarkers (urea, oestradiol and rheumatoid factor) (Figure 2, Supplementary Figures S1-S3). For the majority of associations (119, 60%), combining missense and protein LOF variants produced the smallest p-values. We calculated the genomic inflation factor λ across all tests that were performed genome-wide, and did not find evidence of inflated type I error levels (Figure S7, Supplementary Data). Of the 120 distinct loci, 46 (38%) were associated with more than one biomarker, and a few genes had five or more significant associations: *ANGPTL3*, *APOB*, *JAK2*, *GIGYF1* and *G6PC*. Many of the genes we found to be associated with specific biomarkers had either been implicated in diseases related to these biomarkers (e.g. *LRP2* with renal markers [27]) or are mechanistically related to the biomarkers themselves (e.g. cystatin C with its own gene, *CST*). According to

the NHGRI-EBI GWAS Catalog [2, 28] and PhenoScanner ($p < 10^{-7}$) [24, 25] the majority of significant loci (60%) contained (primarily common, modest effect-size) variants already reported to be associated with the respective biomarkers, as shown for the most significant associations in Table 1.

We recovered 136 (83%) of the 163 significant associations reported by [14] for the same phenotypes in their gene-based variant collapsing analysis on the UK Biobank 200k WES release (Supplementary Table 3), although the sets of participants, preprocessing, covariates, selection criteria for qualifying variants, thresholds for genome-wide significance, and statistical tests differed. 4 (4.8%) out of the 83 associations only found by our analysis could be explained by the significance cutoff, which was lower in their study (5×10^{-9}). Of the 79 remaining associations, 19 (24%) could be explained by the use of DeepRiPe- and SpliceAI predictions (their collapsing models considered only non-synonymous variants). We further found associations for which the kernel-based missense model produced the lowest p-values to be overrepresented in this set (20, 25%, 1.7 fold). 52 out of 79 (65%) were detectable by the single-variant tests reported in the same study [14] which included both rare and common variants.

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

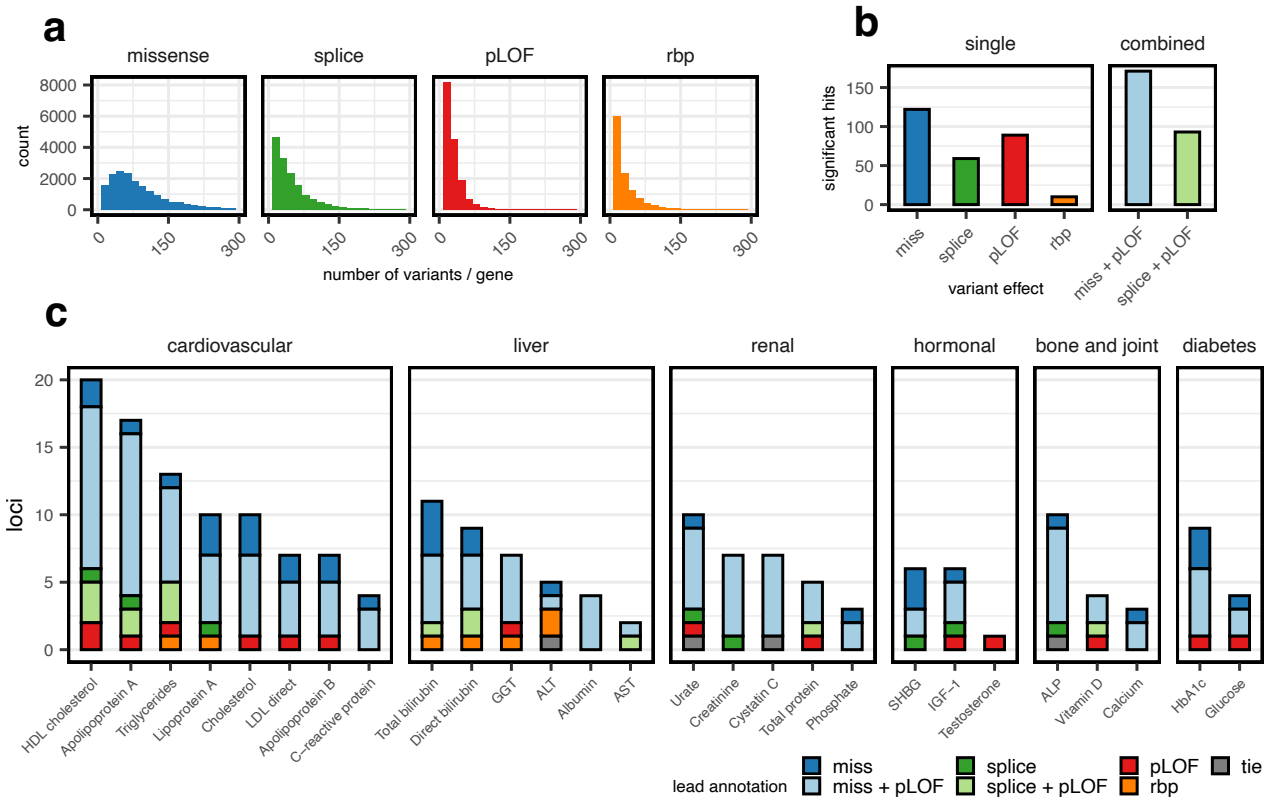


Figure 2: **Association tests overview.** (a) Histograms of the number of qualifying variants per tested gene for the different variant categories. Ranges are truncated at 300 variants, which affected 731 genes for missene, 84 for splice, 7 for pLOF and 15 for rbp (b) Bar plot of the number of significant loci found by testing qualifying variants in the different categories separately (single) or in combination (combined). (c) Bar plot showing 201 significant gene-biomarker associations for 27 biomarkers (x-axis), colored by the variant effect which gave the lowest p-value (lead annotation). ALP: alkaline phosphatase; ALT: alanine aminotransferase; AST: Aspartate aminotransferase; GGT: Gamma glutamyltransferase

.

category	lead gene	phenotype	p-value	variant effect	test	N_{var}	$N_{carrier}$	prev. reported
bone and joint	*GPLD1*	Alkaline phosphatase	2.7×10^{-191}	miss + pLOF	gbvc	222	1645	yes
bone and joint	*ASGR1*	Alkaline phosphatase	7.1×10^{-106}	miss + pLOF	K	97	517	yes
bone and joint	*CASR*	Calcium	6.3×10^{-50}	miss + pLOF	K	140	400	yes
bone and joint	*ALB*	Calcium	3.6×10^{-28}	miss + pLOF	K	117	995	no
bone and joint	*APOB*	Vitamin D	5×10^{-21}	splice + pLOF	gbvc	171	999	yes
bone and joint	*HSPG2*	Alkaline phosphatase	9.5×10^{-21}	miss	K	1363	12120	no
cardiovascular	*ABCA1*	Apolipoprotein A (+3)	3.7×10^{-202}	miss + pLOF	K	396	2184	yes
cardiovascular	*APOB*	LDL direct (+3)	3.7×10^{-161}	miss + pLOF	K	893	5041	yes
cardiovascular	*PCSK9*	LDL direct (+2)	3.3×10^{-110}	miss + pLOF	K	203	1405	yes
cardiovascular	*LCAT*	HDL cholesterol (+1)	1.7×10^{-84}	miss + pLOF	gbvc	107	345	yes
cardiovascular	*CETP*	HDL cholesterol (+1)	2.2×10^{-84}	miss + pLOF	K	101	795	yes
cardiovascular	*APOC3*	Triglycerides (+2)	7.1×10^{-75}	miss + pLOF	gbvc	28	297	yes
diabetes	*PIEZO1*	HbA1c	7.3×10^{-150}	miss	K	908	9087	yes
diabetes	*GCK*	HbA1c (+1)	3×10^{-41}	miss + pLOF	K	61	175	yes
diabetes	*RHAG*	HbA1c	1.6×10^{-35}	miss + pLOF	K	87	793	no
diabetes	*SPTA1*	HbA1c	1.1×10^{-28}	miss + pLOF	K	679	5499	yes
diabetes	*PFKM*	HbA1c	4.4×10^{-23}	miss + pLOF	gbvc	190	947	yes
diabetes	*JAK2*	HbA1c	7.3×10^{-18}	miss	K	223	1053	no
hormonal	*IGFALS*	IGF-1	1.6×10^{-72}	miss + pLOF	gbvc	252	2135	no
hormonal	*HNF4A*	SHBG	3.3×10^{-25}	miss	K	72	582	yes
hormonal	*CLEC10A*	SHBG	6.4×10^{-23}	miss	K	39	682	no
hormonal	*IGFBP3*	IGF-1	8×10^{-21}	miss	K	51	566	no
hormonal	*DNAH2*	SHBG	4.5×10^{-16}	miss	K	875	6926	no
hormonal	*KDM6B*	SHBG	2.7×10^{-12}	splice	K	65	255	no
liver	*UGT1A10*	Total bilirubin (+1)	8.6×10^{-74}	miss + pLOF	K	169	2162	yes
liver	*MROH2A*	Total bilirubin (+1)	1.2×10^{-44}	miss	K	369	3420	yes
liver	*SLCO1B3*	Direct bilirubin (+1)	7.5×10^{-31}	miss + pLOF	gbvc	194	1734	yes
liver	*FCGRT*	Albumin	2.5×10^{-30}	miss + pLOF	K	90	480	yes
liver	*SLCO1B1*	Total bilirubin (+1)	7.9×10^{-28}	miss + pLOF	gbvc	197	1868	yes
liver	*PLEC*	ALT	3.3×10^{-24}	miss	K	1721	16444	no
renal	*SLC22A12*	Urate	0	miss	gbvc	124	1109	yes
renal	*SLC2A9*	Urate	9×10^{-83}	miss + pLOF	K	156	722	yes
renal	*ALPL*	Phosphate	1.7×10^{-81}	miss	gbvc	129	1071	.
renal	*FCGRT*	Total protein	7.2×10^{-26}	miss + pLOF	K	90	479	no
renal	*PDZK1*	Urate	2.9×10^{-22}	pLOF	gbvc	23	180	yes
renal	*SLC22A2*	Creatinine	5.5×10^{-19}	miss + pLOF	gbvc	171	1192	yes

Table 1: **Top hits for every biomarker category.** The six most significant associations for each category are shown, excluding associations of biomarker genes with themselves, and collapsing repeated genes. The number in brackets denotes the number of other biomarkers in the same category the lead gene was also associated with. p-value: smallest p-value over all variant effect categories (column: variant effect) and tests for the lead gene. All p-values shown are derived from the LRT except *PDZK1*-Urate which comes from the score test. test: kernel-based (K) or gene-based variant collapsing (gbvc); N_{var}: number of variants; $N_{carrier}$: number of carriers; prev. reported: whether a GWAS hit for the same biomarker was previously reported within the lead gene (GWAS catalog or PhenoScanner).

2.4 GIGYF1 is associated with diabetes markers

Overall, we identified 31 significant associations in 22 loci exclusively in association tests incorporating pLOF variants alone, or pLOF combined with missense or splice variants. For the majority of these associations, combined testing with other variant categories led to smaller p-values. A notable exception were the five significant associations of *GIGYF1* with biomarkers for diabetes [29] and cardiovascular disease risk, which were only significant for gene-based variant collapsing with pLOF variants. Specifically, we found positive associations with glucose ($p = 3 \times 10^{-9}$) and and glycated haemoglobin (HbA1c, $p = 2 \times 10^{-10}$), and negative associations with LDL direct ($p = 1.6 \times 10^{-9}$), Cholesterol ($p = 6.5 \times 10^{-10}$) and Apolipoprotein B ($p = 3.8 \times 10^{-9}$), making it one of the genes with the most associations in the data set. In total we found 69 carriers of *GIGYF1* pLOF variants. Given its role in IGF-1 signalling together with *GRB10* [30, 31], *GIGYF1* pLOF variants could be a plausible albeit rare mechanism for diabetes. At the time of writing, two studies using either the UK Biobank 200k exome sequencing release [32], or a larger tranche of UK Biobank exome sequencing data yet unreleased to the public [33], have confirmed the association of *GIGYF1* with Type II diabetes (T2D). The association with T2D was not reported in in [14], however, associations with HbA1c and Cholesterol also reached genome-wide significance in their analysis (Supplementary Table 3).

2.5 Combining variant annotations yields six associations for G6PC

We found six associations for *G6PC*, four of which were only found when combining protein LOF and missense annotations with gene-based variant collapsing tests (alkaline phosphatase, triglycerides, SHBG, urate). Variants in *G6PC* cause Glucose-6-phosphatase deficiency type Ia (also called glycogen storage disease Ia) [34, 35], an autosomal recessive disease categorized by growth retardation, enlarged kidneys and liver, low blood glucose, and high blood lipid and uric acid levels. Consistent with signs of inflammation and impaired kidney and liver function, we found elevated levels of alkaline phosphatase ($p = 8.4 \times 10^{-9}$), gamma glutamyltransferase ($p = 1.7 \times 10^{-11}$), urate ($p = 2.6 \times 10^{-9}$), triglycerides ($p = 8.9 \times 10^{-9}$), and C-reactive protein ($p = 1.96 \times 10^{-13}$) in individuals with predicted high-impact missense or pLOF mutations in *G6PC*. We further identified a significant association with decreased levels of sex hormone-binding globulin (SHBG, $p = 6.1 \times 10^{-10}$), which is primarily produced in the liver [36]. All p-values above are those given by gene-based variant collapsing tests combining missense and pLOF variants, the model that gave the lowest p-values for all these associations. While Glucose-6-phosphatase deficiency type Ia is a rare recessive disease, our findings show that altered biomarker levels indicative of mild symptoms are detectable in heterozygous carriers of missense and LOF variants in the *G6PC* gene.

2.6 Novel potential gain of function variants in *PIEZO1*

Our testing strategy allowed us to identify genes in which specific variant categories might play an important role. One such example was *PIEZO1*, a mechanosensitive cation channel [37], which we found associated with diabetes marker HbA1c.

For gene-based variant collapsing tests, we found a significant negative association of missense variants with HbA1c ($p = 2.8 \times 10^{-39}$), while the test for pLOF variants was not significant ($p = 0.862$, 623 carriers). By far the lowest p-value for this gene was given by the kernel-based test for missense variants ($p = 7.296 \times 10^{-150}$). Combining pLOF variants and missense variants did not lead to smaller p-values, but was still highly significant ($p = 3.5 \times 10^{-148}$, kernel-based LRT). The large differences between variant categories and type of association tests lead us to closer investigate the 908 predicted high-impact missense variants in 9,352 individuals for this gene.

We performed single-variant score tests and identified multiple missense variants with strong negative associations with HbA1c (Table 2). One these variants, 16:88719665:G:A or T2127M (rs587776991) is a gain of function variant that slows down inactivation kinetics of PIEZO1 in patients with dehydrated hereditary stomatocytosis (a disorder of red blood cells), together with other gain of function variants [38, 39]. Decreased levels of HbA1c had previously been observed in individuals with red blood cell disorders [40, 41, 42].

We therefore hypothesised that the other highly significant variants could also potentially be gain of function variants. We grouped the missense variants within *PIEZO1* by the amino acid positions they affected and performed local variant collapsing. This allowed us to identify other positions in *PIEZO1* (e.g. 2110R or 2474V) that are potentially sensitive to gain of function mutations (Table 2, Figure 3).

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 heriditary stomatocytosis [43], while *SPTA1* mutations cause hereditary elliptocytosis [44].

While it has been suggested that *PIEZO1* stimulates insulin release [45], 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 [46].

variant id	position	weight	variant	$N_{carrier}$	variant p-val.	$\beta_{variant}$	$\pm$	position p-value	$\beta_{position}$	$\pm$
16:88736318:C:T	463	0.983	A/T	109	1.9×10^{-10}	-0.5591	0.088	2.5×10^{-10}	-0.56	0.088
16:88736317:G:A	463	0.998	A/V	1	0.92	0.0938	0.92	2.5×10^{-10}	-0.56	0.088
16:88731880:C:G	1008	0.99	G/R	65	7.9×10^{-14}	-0.8494	0.11	5×10^{-15}	-0.87	0.11
16:88731880:C:T	1008	0.99	G/R	3	0.013	-1.3151	0.53	5×10^{-15}	-0.87	0.11
16:88726891:A:C	1175	0.998	F/V	4	4.5×10^{-8}	-2.5068	0.46	4.5×10^{-8}	-2.5	0.46
16:88726565:C:T	1260	0.706	V/I	78	6.2×10^{-8}	-0.5619	0.1	2×10^{-8}	-0.69	0.12
16:88726565:C:A	1260	1	V/F	1	0.062	-1.7112	0.92	2×10^{-8}	-0.69	0.12
16:88721626:C:G	1772	0.988	R/P	13	4.5×10^{-6}	-1.1661	0.25	2.4×10^{-9}	-1.1	0.18
16:88721627:G:C	1772	0.976	R/G	10	0.0008	-0.9712	0.29	2.4×10^{-9}	-1.1	0.18
16:88721626:C:A	1772	0.982	R/L	1	0.042	-1.8650	0.92	2.4×10^{-9}	-1.1	0.18
16:88721627:G:A	1772	0.998	R/C	1	0.5	-0.6168	0.92	2.4×10^{-9}	-1.1	0.18
16:88721423:C:G	1804	0.867	G/A	31	8.1×10^{-11}	-1.0700	0.16	8.1×10^{-11}	-1.1	0.18
16:88719717:G:A	2110	0.998	R/W	9	5.5×10^{-18}	-2.6395	0.31	6×10^{-33}	-2.7	0.22
16:88719716:C:T	2110	0.894	R/Q	9	1.4×10^{-16}	-2.5241	0.31	6×10^{-33}	-2.7	0.22
16:88719665:G:A	2127	1	T/M	22	1×10^{-31}	-2.2895	0.2	1×10^{-31}	-2.3	0.2
16:88716656:G:T	2277	0.956	L/M	314	2.9×10^{-39}	-0.6885	0.052	2.9×10^{-39}	-0.7	0.054
16:88716649:C:T	2279	0.829	R/H	16	9.8×10^{-9}	-1.3146	0.23	3.5×10^{-9}	-0.99	0.17
16:88716650:G:A	2279	0.976	R/C	16	0.0063	-0.6260	0.23	3.5×10^{-9}	-0.99	0.17
16:88716649:C:G	2279	0.763	R/P	1	0.58	-0.5004	0.92	3.5×10^{-9}	-0.99	0.17
16:88716234:C:T	2365	0.812	G/R	9	3.3×10^{-16}	-2.4926	0.31	3.3×10^{-16}	-2.8	0.34
16:88715751:C:T	2474	0.987	V/M	101	4.7×10^{-8}	-0.4983	0.091	1.2×10^{-8}	-0.52	0.091
16:88715751:C:G	2474	0.879	V/L	2	0.029	-1.4117	0.65	1.2×10^{-8}	-0.52	0.091
16:88715751:C:A	2474	0.879	V/L	1	0.82	-0.2111	0.92	1.2×10^{-8}	-0.52	0.091

Table 2: **Potential *PIEZO1* gain of function variants** Variants are grouped and ordered by the 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 (score test); $\beta_{position}$: position effect size ($\pm$ standard error). Positions relate to the ENST00000301015 transcript

2.7 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 [47]. 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 these biomarkers [48].

Yet, one additional association with inflammation marker C-reactive protein (CRP) was only identified by the kernel-based association test ($p = 3.997 \times 10^{-27}$). This prompted us to further investigate the 344 high-impact missense variants observed in *ABCA1*. Single-variant score tests and collapsing by amino-acid position identified two missense variants (9:104831048:C:A, 9:104831048:C:G) in one of the extracellular domains affecting the same amino acid (W590) which were associated with strongly decreased levels of CRP (Figure 3). The two variants

carried most of the signal in this gene with single-variant p-values of 2.08×10^{-31} for W590L (A allele, 54 carriers) and 8.41×10^{-8} for W590S (G allele, 12 carriers, score test).

The W590S-variant leads to reduced cholesterol and phospholipid efflux, while retaining expression and ability to bind APOA1[49, 50]. The other and more common variant, W590L, has been observed [51, 52], but to our knowledge not experimentally evaluated.

The binding of APOA1 to ABCA1 activates anti-inflammatory pathways via JAK2 and STAT3 in macrophages [53]. Because W590S has been shown to slow dissociation of bound APOA1[50], this provides a plausible causal mechanism for the reduced levels of CRP we observe in carriers of the W590S-variant. We hypothesize that W590L might act through the same mechanism. This property could set these variants apart from other missense variants in *ABCA1*, which have been reported to abolish binding of APOA1[49].

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

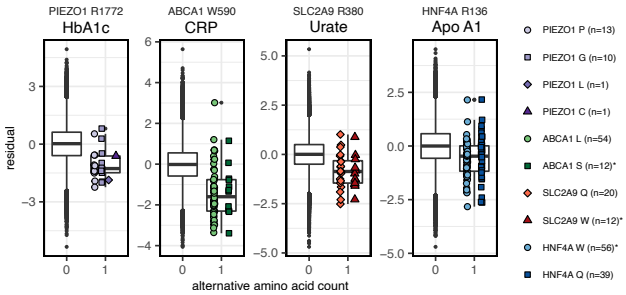


Figure 3: **Local collapsing of missense variants.** Dosage box plots showing the alternative amino acid counts (x-axis) against the covariate-adjusted quantile transformed phenotypes (y-axis). Collapsing variants by amino acid position identified negative associations of *PIEZO1* R1772 with HbA1c ($p = 2.36 \times 10^{-9}$), *ABCA1* W590 with C-reactive protein ($p = 1.13 \times 10^{-37}$), *SLC2A9* R380 with Urate ($p = 6 \times 10^{-10}$) and *HNF4A* R136 with Apolipoprotein A ($p = 1.3 \times 10^{-10}$, score test). Collapsing together with ClinVar variants with reported conditions (*) helps place novel variants into disease context. For all 4 associations, collapsed p-values were lower than those of the single variants.

2.8 A pathogenic JAK2 gain of function variant is associated with multiple biomarkers

We found seven significant associations of variants in *JAK2* with biomarker levels, making it the gene with the highest number of significant associations. While its negative association with IGF-1 was only significant for protein LOF variants, six additional associations were found exclusively with kernel-based association tests incorporating missense variants. Upon closer investigation, we found a single missense variant (9:5073770:G:T) which was largely responsible for the significant associations with biomarkers for lipid metabolism (Apolipoprotein A, HDL cholesterol, cholesterol, LDL direct; all negative), kidney function (Cystatin C; postive) and diabetes (HbA1c; negative). This variant, also known as V617F or rs77375493 leads to constitutive phosphorylation activity and is a known prognostic marker in myeloproliferative neoplasms [54, 55]. *JAK2* is therefore an example of a gene in which we observe a single gain of function variant with potential effects on multiple biomarkers.

2.9 Unique associations identified by splice-predictions

In total, we identified 7 associations exclusively when incorporating SpliceAI variant effect predictions of which 6 were only found using kernel-based association tests. Specifically, we found associations of variants in *SLC9A5* with Apolipoprotein A and HDL cholesterol, *NDUFB8* with Aspartate aminotransferase, *ATL3* with Urate, *GH1* 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 *GH1* (Growth Hormone 1) with IGF-1 [56], the only hit in this subset also found by gene-based variant collapsing. The interpretation of single highly significant variants driving 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 4a), was also a missense variant (with low to moderate predicted impact [52]). *ECE1* and *KDM6B* lie in proximity to the genes coding for the biomarkers they were found associated with (*ALPL*, *SHBG*), therefore we couldn't exclude transcriptional cis-regulatory effects as the cause of these associations.

2.10 Associations identified by RBP-binding predictions

Out of the 10 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 4), 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* with triglycerides we had independently found with other variant categories and the close proximity of an *ANGPTL3* exon.

We further investigated this variant by assessing binding probabilities of RBPs beyond the six RBPs in focus here that are represented in DeepRiPe. We found that the variant was also predicted to decrease the binding probability of BCLAF1, a factor related to mRNA processing [59] and increase binding of HNRNPL (Figure 4). Using attribution maps [23], we found that instead of strengthening or inserting a new QKI binding motif, this variant weakens a splice donor signal in the presence of upstream binding motifs for the splicing regulators QKI and HNRNPL (Supplementary Figure S4). SpliceAI predicted only a weak upstream donor loss (0.02) for this variant, which was well below the threshold of 0.1 we used to identify splice-altering variants, but indicative of the same trend.

The remaining five associations exclusively identified by association tests incorporating RBP-binding predictions were those of *ATG16L1* with Total bilirubin and direct bilirubin, *UPB1* with Gamma glutamyltransferase, *SLC39A4* with Alanine aminotransferase and *SHARPIN* with Alanine aminotransferase.

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

Common intronic variants of *ATG16L1* were previously found to be associated with Crohn's disease and inflammatory bowel disease [60, 61] and increased bilirubin levels [62, 63]. To our knowledge, rare variants with associations to bilirubin levels within *AT16L1*, especially those potentially affecting RBP-binding, have not been identified by previous studies.

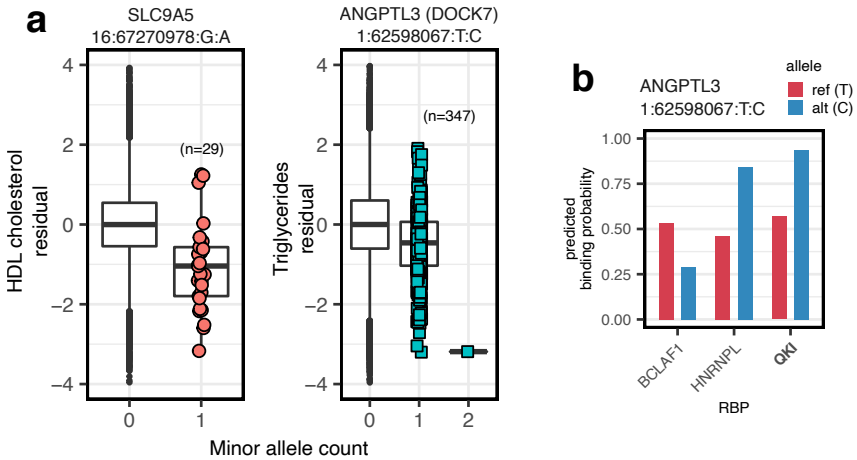


Figure 4: **Variants identified by deep learning** (a) Dosage box plots showing covariate-adjusted quantile transformed phenotypes against minor allele counts for variants in *SLC9A5* and *ANGPTL3/DOCK7*. A predicted splice-variant 16:67270978:G:A is negatively associated with HDL cholesterol ($p = 8.36 \times 10^{-12}$, score test), whereas intronic 1:62598067:T:C is negatively associated with Triglycerides ($p = 1.25 \times 10^{-25}$, score test). (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.

2.11 Combined likelihood ratio and score tests (sLRT)

In order to benefit from both the speed of the score tes [8] and higher power of the LRT [11] we investigated the use of a combination of these tests, which we call the sLRT (score-LRT). The sLRT is a likelihood ratio test that is performed only when initial score tests reach nominal significance at a given cutoff t. If this threshold is not reached, it returns the score test p-value. Throughout our analysis, we used $t = 0.1$, and found that it was unlikely that a larger threshold would have identified many more associations (Supplementary Figure S5). As the run time is dominated by the cost of computing the LRT, this test can achieve a computational speedup factor of roughly $1/t = 10$ over the LRT under the null hypothesis.

The sLRT identified more significant associations than the score test particularly when performing kernel-based association tests (Supplementary Figure S6). For missense and splice variants, the kernel-based sLRT captured all associations that would have been identified by the kernel-based score test alone. In both cases, the sLRT identified additional associations. However, a large fraction of these additional associations (40% for splice variants, and 61% for missense variants) were also identified by gene-based variant collapsing for which

the sLRT and score test gave almost identical results. Nevertheless, we found the majority of the remaining additional associations to be plausible and/or previously reported in other association studies and therefore used the sLRT throughout our analysis, except for pLOF variants, where we only performed only a gene-based variant collapsing test.

3 Discussion

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

Overall, our approach was successful at identifying disease genes without explicitly using disease diagnoses themselves, even for recessive diseases (*G6PC*, glycogen storage disease Ia [34]), or diseases with mixed inheritance patterns (*LRP2*, Fanconi syndrom [27]), while keeping the number of tests low compared to phenome-wide association studies.

While some of the changes in biomarker levels we detected might be be sub-clinical, they could interfere with the diagnosis of common conditions which rely on biomarkers. To prevent misdiagnoses and enable preventative care, our results could aid the design of targeted sequencing panels that focus on the genes with the highest impacts. For example, we found 8% of participants to harbor at least one protein LOF variant in any of the significant loci for that variant category. In other words, it's fairly common to have at least one uncommon LOF variant with potential effects on biomarker levels.

One major contribution of our study was the design and successful application of kernel-based tests that incorporate quantitative functional variant effect predictions for large exome sequencing data. A previous study had tried to apply kernel-based score tests using SKAT [8] on the 50k WES release, but found mostly unreliable results [12]. In contrast to their approach we did not re-weigh variants according to allele frequencies. Furthermore, we showed that a computationally efficient combination of the LRT and score test has potentially higher power than the score test alone, and identified more associations than gene-based variant collapsing.

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

Although kernel-based tests should also provide benefits if genes contain variants with strong opposing effect sizes, we did not find this to be a widespread phenomenon for the associations we identified (with some exceptions such as *APOB*). More such cases could appear for non-coding regions in whole-genome sequencing studies.

While we found a large overlap between our associations and those found in the variant collapsing analysis presented in [14], the differences highlight the sensitivity of gene-based tests to qualifying criteria for rare variants, which can make them harder to reproduce. By making our analysis pipeline public, we hope to

increase reproducibility and enable others to explore different qualifying criteria more easily. By performing kernel-based tests and including variants potentially acting through splicing and the binding of RBPs, we identified additional associations without the need for single-variant tests.

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

We explored the use of deep-learning-derived variant effect predictions for splicing and the binding of RBPs. The restriction to exon-proximal regions meant we only observed a fraction of the variants potentially acting through these mechanisms. Associations found by incorporating splice-predictions largely overlapped with those identified with pLOF variants (which included simple splice donor/acceptor variants). While we found some associations exclusively with splice-predictions, these were mostly due to single variants and would need further validation (e.g. *SLC9A5*). Similar reasoning holds for the associations found with predictions for RBP-binding. We anticipate that deep-learning-based predictions will become more valuable for non-coding regions in whole-genome sequencing studies, for which the approaches we developed will also be applicable.

Deep-learning-derived functional annotations have been considered in other studies in the context of association testing. Proposed methods include signed LD-score regression [64], or the association tests presented in DeepWAS [65]. However, these methods have not been designed for rare variants. Other statistical methods that combine multiple functional annotations could potentially further reduce the number of tests [66].

In future studies, methods like AlphaFold [67] could allow specific testing of effects on protein folding. Methods that allow predicting residue-residue interactions within proteins could enable the mostly unsupervised identification of protein domains and their separate testing [68].

The methodological advances and practices we applied in this association study also apply to those situations, and serve as potential baselines for functionally informed kernel-based association tests with rare variants.

4 Methods

4.1 UK Biobank Data processing

All 30 blood biochemistry biomarkers (category 17518) from the UK Biobank were quantile-transformed to match a normal distribution with mean 0 and unit standard deviation using scikit-learn (v0.22.2) [69]. For testosterone, which showed a clear bimodal distribution based on sex, quantile transformation was performed separately for both sexes. Sex, BMI, age at recruitment, smoking status and the first 10 genetic principal components were used as covariates (Supplementary Table S1). Smoking status (never, previous, current) was encoded in three separate binary variables. Participants with any missing covariates were excluded. We used the ukb_gen_samples_to_remove function of the ukbtools package (v0.11.3) [70] together with pre-computed relatedness scores (ukbA_rel_sP.txt, see UK Biobank Resource 531) to remove closely related individuals, keeping

only one representative of groups that are related to the 3rd degree or less. After removing 6,293 related individuals and restricting to those with no missing covariates, 192,352 participants remained. This sample was 55% female (45% male) and the average age at recruitment was 56.46 years ($\sigma = 8.08$). Furthermore, the average BMI was 27.37 ($\sigma = 4.77$) and our subset contained 18,562 current and 67,109 previous smokers.

In our analysis we made use of the PLINK-formatted exome sequencing genotype data. The final results presented in this manuscript were derived from the 200k WES release produced by the OQFE pipeline [19]. The UK Biobank pipeline already implements quality filters [19, 71]. Additionally, we removed all variants that violated the Hardy-Weinberg equilibrium (HWE) assumption (HWE exact test p-value below the threshold of 10^{-5}) and variants genotyped in less than 90% of participants. Furthermore, we calculated minor allele frequencies within all unrelated participants with complete covariates (see above), and excluded variants with minor allele frequencies above 0.1% from the rare-variant association tests. We did not analyze variants on sex chromosomes. 16,737,187 variants passed these filters, of which 43.64% were singletons. We directly use UK Biobank variant identifiers (which include chromosome and 1-based hg38 positions) to name variants in order to facilitate comparisons.

4.2 Variant effect prediction and annotation

Protein loss of function and missense We predicted effects for all genetic variants that passed basic filtering using the Ensembl Variant Effect Predictor [15] (VEP, v101; cache version 97), including scores from Polyphen-2 [16] (v2.2.2) and SIFT [17] (v5.2.2). All variants marked as splice_acceptor_variant, splice_donor_variant, frameshift_variant, stop_gained, stop_lost or start_lost were considered protein loss of function (pLOF) variants as in [12]. We further annotated missense variants by calculating impact scores (averages between deleterious-probabilities given by PolyPhen-2 and SIFT), which were used to filter and weigh variants in the association tests. Specifically, Missense variants were included if their impact score was at least 0.8, or if they affected amino acid positions for which another variant with impact score of at least 0.8 was observed.

Splicing We retrieved published pre-computed variant effect predictions produced by the SpliceAI deep learning model [22] for single nucleotide variants. SpliceAI predicts consequences of genetic variants for nearby splice sites, specifically splice donor loss/gain or splice acceptor loss/gain. We used the splice-site-proximal masked delta scores (v1.3). In the masked files, scores corresponding to the strengthening of annotated splice sites and weakening of non-annotated splice sites are set to 0, as these are generally less pathogenic. We included splice-variants in the association tests if at least one of the four SpliceAI delta scores was greater or equal to 0.1. The maxima over the different delta scores for every variant were used to weigh variants in the association tests (Supplementary Methods).

RBP-binding We predicted the effects of all genetic variants on the binding of 6 RNA-binding proteins (RBPs) using a modified version of the DeepRiPe deep neural network [23], in which predictions are purely sequence-based. We predicted the differences in binding by subtracting the predictions for the reference alleles

from those for the alternative allele [72], and used these variant effect predictions to filter and weigh variants during the association tests (Supplementary Methods). Variants were included into the association tests if at least one predicted effect on any of the RBPs had an absolute value greater or equal to 0.25.

4.3 Statistical models and tests

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

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

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

Any valid variance-covariance matrix can be substituted for $\boldsymbol{K}_g$. In order to use efficient algorithms for estimating the parameters σ_e^2 and σ_g^2 and performing association tests, we require $\boldsymbol{K}_g$ to be factored as a quadratic form [9, 11]:

$$\boldsymbol{K}_g = \phi(\boldsymbol{G})\phi(\boldsymbol{G})^T, \tag{2}$$

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

Regardless of the choice of kernel (and hence ϕ) the statistical test is defined by the null hypothesis $H_0 : \sigma_g^2 = 0$, and the alternative hypothesis $H_1 : \sigma_g^2 \geq 0$. Both a score test and likelihood ratio test (LRT) have been described for this application. While the score test is often chosen in statistical genetics applications due to its speed and software availability, the LRT has been shown to have higher power but is computationally more demanding [8, 11, 73].

In order to avoid computing the LRT for all genes but still profit from potentially higher power, we performed score tests genome-wide and only performed the LRT if score tests (within the specific variant category) reached nominal significance, an approach which we call the score-LRT (sLRT, Supplementary Methods). The sLRT

returns the p-value for the score test if nominal significance was not reached, otherwise it returns the p-value for the likelihood ratio test.

We applied the statistical framework above to perform both gene-based variant collapsing tests and kernel-based association tests, corresponding to different functions ϕ (Supplementary Methods). We adjust p-values for the total number of tests performed using Bonferroni correction (FWER = 0.05), which lead to a cutoff of 1.4435×10^{-8}.

4.4 Gene-based testing procedure

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

For missense variants, we performed both gene-based variant collapsing and kernel-based association tests using the sLRT. For the kernel-based tests with missense variants, we designed a kernel that collapses variants by amino acid position (local collapsing), and weighs them by their impact score. Additionally, in cases where either missense-variant score test used in the sLRT was nominally significant ($p < 0.1$), we combined missense and protein LOF variants for joint tests. For these joint tests, we investigated both the use of joint gene-based collapsing test and a kernel-based test that combines collapsing of pLOF variants with local collapsing of missense variants by concatenation (Supplementary Methods).

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

Because some of the genes in the Ensembl 97 release share exons, we encountered cases in which these genes shared associations caused by the same variants. We do not report these as distinct gene-biomaker associations in the main text (except when comparing to [14], who reported all such associations), but include the full list in Supplementary Table 2.

4.5 Data availability

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

4.6 Code availability

Code that allows reproducing results from this study is available on github (https://github.com/HealthML/faatpipe).

4.7 Contributions

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

Acknowledgements

The authors wish to thank Wolfgang Kopp for valuable comments on the manuscript. This research has been conducted using the UK Biobank Resource under Application Number 40502. This research has received funding by the German Federal Ministry of Education and Research (BMBF) in the project KI-LAB-ITSE (project number 01—S19066) and the European Commission in the Horizon 2020 project INTERVENE (Grant agreement ID: 101016775).

References

[1] Sudlow, C. *et al.* Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *Plos med* **12**, e1001779 (2015).

[2] Buniello, A. *et al.* The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research* **47**, D1005–D1012 (2019).

[3] Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).

[4] Hernandez, R. D. *et al.* Ultrarare variants drive substantial cis heritability of human gene expression. *Nature genetics* **51**, 1349–1355 (2019).

[5] Zhu, Q. *et al.* A genome-wide comparison of the functional properties of rare and common genetic variants in humans. *The American Journal of Human Genetics* **88**, 458–468 (2011).

[6] Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics* **83**, 311–321 (2008).

[7] Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics* **95**, 5–23 (2014).

[8] Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* **89**, 82–93 (2011).

[9] Listgarten, J. *et al.* A powerful and efficient set test for genetic markers that handles confounders. *Bioinformatics* **29**, 1526–1533 (2013).

[10] Povysil, G. *et al.* Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nature Reviews Genetics* **20**, 747–759 (2019).

[11] Lippert, C. *et al.* Greater power and computational efficiency for kernel-based association testing of sets of genetic variants. *Bioinformatics* **30**, 3206–3214 (2014).

[12] Cirulli, E. T. *et al.* Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. *Nature Communications* **11**, 542 (2020).

[13] Van Hout, C. V. *et al.* Exome sequencing and characterization of 49,960 individuals in the uk biobank. *Nature* **586**, 749–756 (2020).

[14] Wang, Q. *et al.* Surveying the contribution of rare variants to the genetic architecture of human disease through exome sequencing of 177,882 uk biobank participants. *bioRxiv* (2020).

[15] McLaren, W. *et al.* The ensembl variant effect predictor. *Genome biology* **17**, 1–14 (2016).

[16] Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using polyphen-2. *Current protocols in human genetics* **76**, 7–20 (2013).

[17] Ng, P. C. & Henikoff, S. Sift: Predicting amino acid changes that affect protein function. *Nucleic acids research* **31**, 3812–3814 (2003).

[18] Sundaram, L. *et al.* Predicting the clinical impact of human mutation with deep neural networks. *Nature genetics* **50**, 1161–1170 (2018).

[19] Szustakowski, J. D. *et al.* Advancing human genetics research and drug discovery through exome sequencing of the uk biobank. *medRxiv* (2020).

[20] Strimbu, K. & Tavel, J. A. What are biomarkers? *Current Opinion in HIV and AIDS* **5**, 463 (2010).

[21] Sinnott-Armstrong, N. *et al.* Genetics of 35 blood and urine biomarkers in the uk biobank. *Nature genetics* 1–10 (2021).

[22] Jaganathan, K. *et al.* Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548 (2019).

[23] Ghanbari, M. & Ohler, U. Deep neural networks for interpreting rna-binding protein target preferences. *Genome research* **30**, 214–226 (2020).

[24] Staley, J. R. *et al.* Phenoscanner: a database of human genotype–phenotype associations. *Bioinformatics* **32**, 3207–3209 (2016).

[25] Kamat, M. A. *et al.* Phenoscanner v2: an expanded tool for searching human genotype–phenotype associations. *Bioinformatics* **35**, 4851–4853 (2019).

[26] Mukherjee, N. *et al.* Deciphering human ribonucleoprotein regulatory networks. *Nucleic acids research* **47**, 570–581 (2019).

[27] Willnow, T. E. & Christ, A. Endocytic receptor lrp2/megalin—of holoprosencephaly and renal fanconi syndrome. *Pflügers Archiv-European Journal of Physiology* **469**, 907–916 (2017).

[28] Magno, R. & Maia, A.-T. gwasrapidd: an r package to query, download and wrangle gwas catalog data. *Bioinformatics* **36**, 649–650 (2020).

[29] Rohlfing, C. L. *et al.* Defining the relationship between plasma glucose and hba1c: analysis of glucose profiles and hba1c in the diabetes control and complications trial. *Diabetes care* **25**, 275–278 (2002).

[30] Giovannone, B. *et al.* Two novel proteins that are linked to insulin-like growth factor (igf-i) receptors by the grb10 adapter and modulate igf-i signaling. *Journal of Biological Chemistry* **278**, 31564–31573 (2003).

[31] Wang, L. *et al.* Peripheral disruption of the grb10 gene enhances insulin signaling and sensitivity in vivo. *Molecular and cellular biology* **27**, 6497–6505 (2007).

[32] Curtis, D. Weighted burden analysis in 200,000 exome-sequenced subjects characterises rare variant effects on risk of type 2 diabetes. *medRxiv* (2021).

[33] Deaton, A. M. *et al.* Gene-level analysis of rare variants in 363,977 whole exome sequences reveals an association of gigyf1 loss of function with diabetes. *medRxiv* (2021).

[34] Chou, J. Y. & Mansfield, B. C. Mutations in the glucose-6-phosphatase-α (g6pc) gene that cause type ia glycogen storage disease. *Human mutation* **29**, 921–930 (2008).

[35] Froissart, R. *et al.* Glucose-6-phosphatase deficiency. *Orphanet journal of rare diseases* **6**, 1–12 (2011).

[36] Nakhla, A. M. *et al.* Human sex hormone-binding globulin gene expression-multiple promoters and complex alternative splicing. *BMC Molecular Biology* **10**, 1–18 (2009).

[37] Coste, B. *et al.* Piezo1 and piezo2 are essential components of distinct mechanically activated cation channels. *Science* **330**, 55–60 (2010).

[38] Albuisson, J. *et al.* Dehydrated hereditary stomatocytosis linked to gain-of-function mutations in mechanically activated piezo1 ion channels. *Nature communications* **4**, 1–9 (2013).

[39] Andolfo, I. *et al.* Multiple clinical forms of dehydrated hereditary stomatocytosis arise from mutations in piezo1. *Blood* **121**, 3925–3935 (2013).

[40] Song, A. *et al.* Low hba1c with normal hemoglobin in a diabetes patient caused by piezo1 gene variant: A case report. *Frontiers in Endocrinology* **11**, 356 (2020).

[41] Nakatani, R. *et al.* Importance of the average glucose level and estimated glycated hemoglobin in a diabetic patient with hereditary hemolytic anemia and liver cirrhosis. *Internal Medicine* **57**, 537–543 (2018).

[42] Finan, E. & Joseph, J. Glycosylated haemoglobin: a false sense of security. *BMJ Case Reports CP* **11**, e227668 (2018).

[43] Bruce, L. J. *et al.* The monovalent cation leak in overhydrated stomatocytic red blood cells results from amino acid substitutions in the rh-associated glycoprotein. *Blood* **113**, 1350–1357 (2009).

[44] Sahr, K. *et al.* Sequence and exon-intron organization of the dna encoding the alpha i domain of human spectrin. application to the study of mutations causing hereditary elliptocytosis. *The Journal of clinical investigation* **84**, 1243–1252 (1989).

[45] Deivasikamani, V. *et al.* Piezo1 channel activation mimics high glucose as a stimulator of insulin release. *Scientific reports* **9**, 1–10 (2019).

[46] Andolfo, I., Russo, R., Gambale, A. & Iolascon, A. Hereditary stomatocytosis: an underdiagnosed condition. *American journal of hematology* **93**, 107–121 (2018).

[47] Tang, C.-K. *et al.* Effect of apolipoprotein ai on atp binding cassette transporter a1 degradation and cholesterol efflux in thp-1 macrophage-derived foam cells. *Acta biochimica et biophysica Sinica* **36**, 218–226 (2004).

[48] Marcil, M. *et al.* Mutations in the abc 1 gene in familial hdl deficiency with defective cholesterol efflux. *The Lancet* **354**, 1341–1346 (1999).

[49] Vaughan, A. M., Tang, C. & Oram, J. F. Abca1 mutants reveal an interdependency between lipid export function, apoa-i binding activity, and janus kinase 2 activation. *Journal of lipid research* **50**, 285–292 (2009).

[50] Nagao, K., Zhao, Y., Takahashi, K., Kimura, Y. & Ueda, K. Sodium taurocholate-dependent lipid efflux by abca1: effects of w590s mutation on lipid translocation and apolipoprotein ai dissociation. *Journal of lipid research* **50**, 1165–1172 (2009).

[51] Probst, M. C. *Development and evaluation of multiplex and high-throughput SNP analysis for the ABCA1 gene.* Ph.D. thesis (2004).

[52] Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

[53] Tang, C., Liu, Y., Kessler, P. S., Vaughan, A. M. & Oram, J. F. The macrophage cholesterol exporter abca1 functions as an anti-inflammatory receptor. *Journal of Biological Chemistry* **284**, 32336–32343 (2009).

[54] James, C. *et al.* A unique clonal jak2 mutation leading to constitutive signalling causes polycythaemia vera. *nature* **434**, 1144–1148 (2005).

[55] Rumi, E. *et al.* Clinical effect of driver mutations of jak2, calr, or mpl in primary myelofibrosis. *Blood, The Journal of the American Society of Hematology* **124**, 1062–1069 (2014).

[56] Goddard, A. D. *et al.* Mutations of the growth hormone receptor in children with idiopathic short stature. *New England Journal of Medicine* **333**, 1093–1098 (1995).

[57] Enkhmaa, B., Anuurad, E., Zhang, W., Tran, T. & Berglund, L. Lipoprotein (a): genotype–phenotype relationship and impact on atherogenic risk. *Metabolic syndrome and related disorders* **9**, 411–418 (2011).

[58] Shadrina, A. S. *et al.* Prioritization of causal genes for coronary artery disease based on cumulative evidence from experimental and in silico studies. *Scientific reports* **10**, 1–15 (2020).

[59] Sarras, H., Alizadeh Azami, S. & McPherson, J. P. In search of a function for bclaf1. *TheScientificWorld-Journal* **10**, 1450–1461 (2010).

[60] Jostins, L. *et al.* Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).

[61] Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature genetics* **47**, 979–986 (2015).

[62] Choi, Y. *et al.* Causal associations between serum bilirubin levels and decreased stroke risk: a two-sample mendelian randomization study. *Arteriosclerosis, thrombosis, and vascular biology* **40**, 437–445 (2020).

[63] Seo, J. Y. *et al.* A genome-wide association study on liver enzymes in korean population. *Plos one* **15**, e0229374 (2020).

[64] Reshef, Y. A. *et al.* Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk. *Nature genetics* **50**, 1483–1493 (2018).

[65] Arloth, J. *et al.* Deepwas: Multivariate genotype-phenotype associations by directly integrating regulatory information using deep learning. *PLoS computational biology* **16**, e1007616 (2020).

[66] Li, X. *et al.* Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nature genetics* **52**, 969–983 (2020).

[67] Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).

[68] Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118** (2021).

[69] Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *the Journal of machine Learning research* **12**, 2825–2830 (2011).

[70] Hanscombe, K. B., Coleman, J. R., Traylor, M. & Lewis, C. M. ukbtools: An r package to manage and query uk biobank data. *PLoS One* **14**, e0214311 (2019).

[71] Van Hout, C. V. *et al.* Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586**, 749–756 (2020).

[72] Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods* **12**, 931–934 (2015).

[73] Zhou, W. *et al.* Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *Nature genetics* **52**, 634–639 (2020).

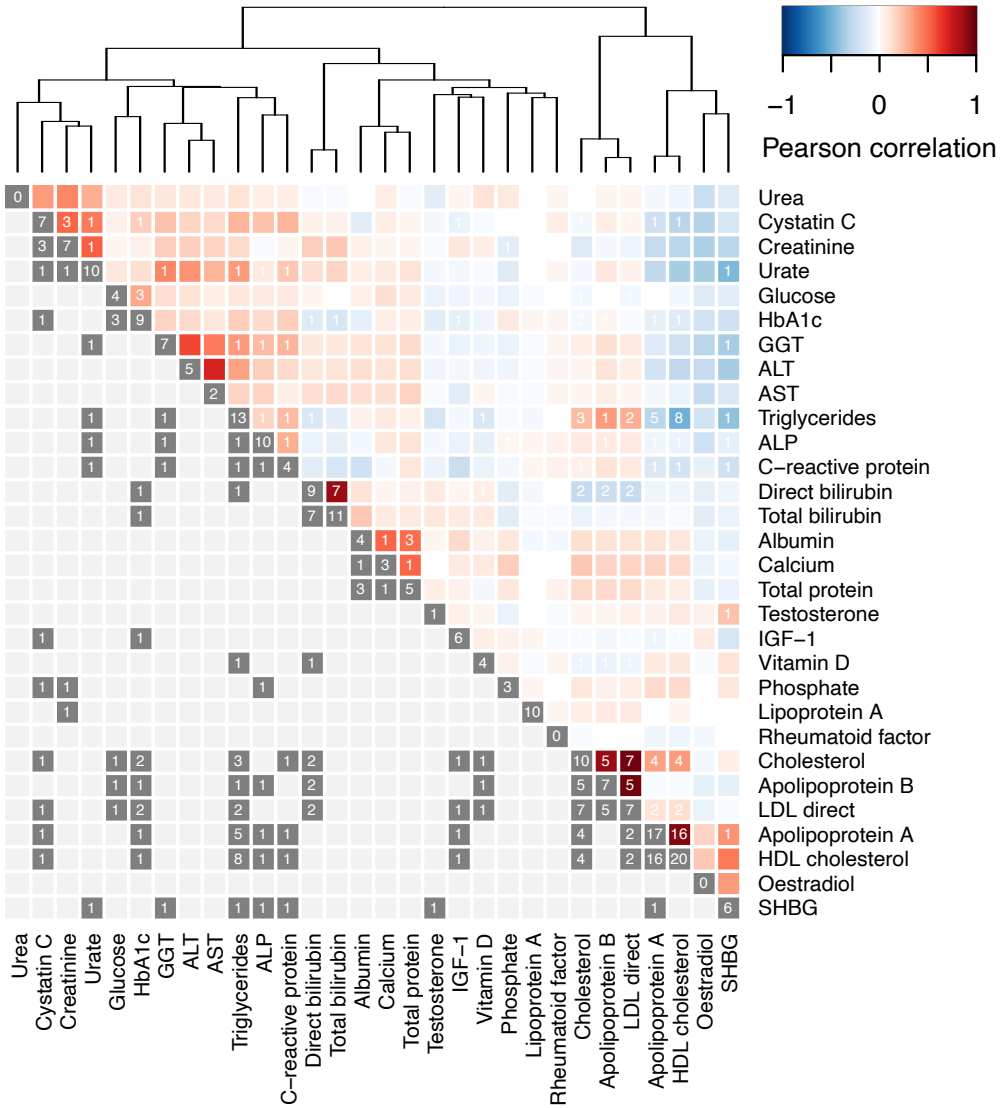
[74] Davies, R. B. The distribution of a linear combination of $\chi 2$ random variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **29**, 323–333 (1980).

[75] Kuonen, D. Miscellanea. saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika* **86**, 929–935 (1999).

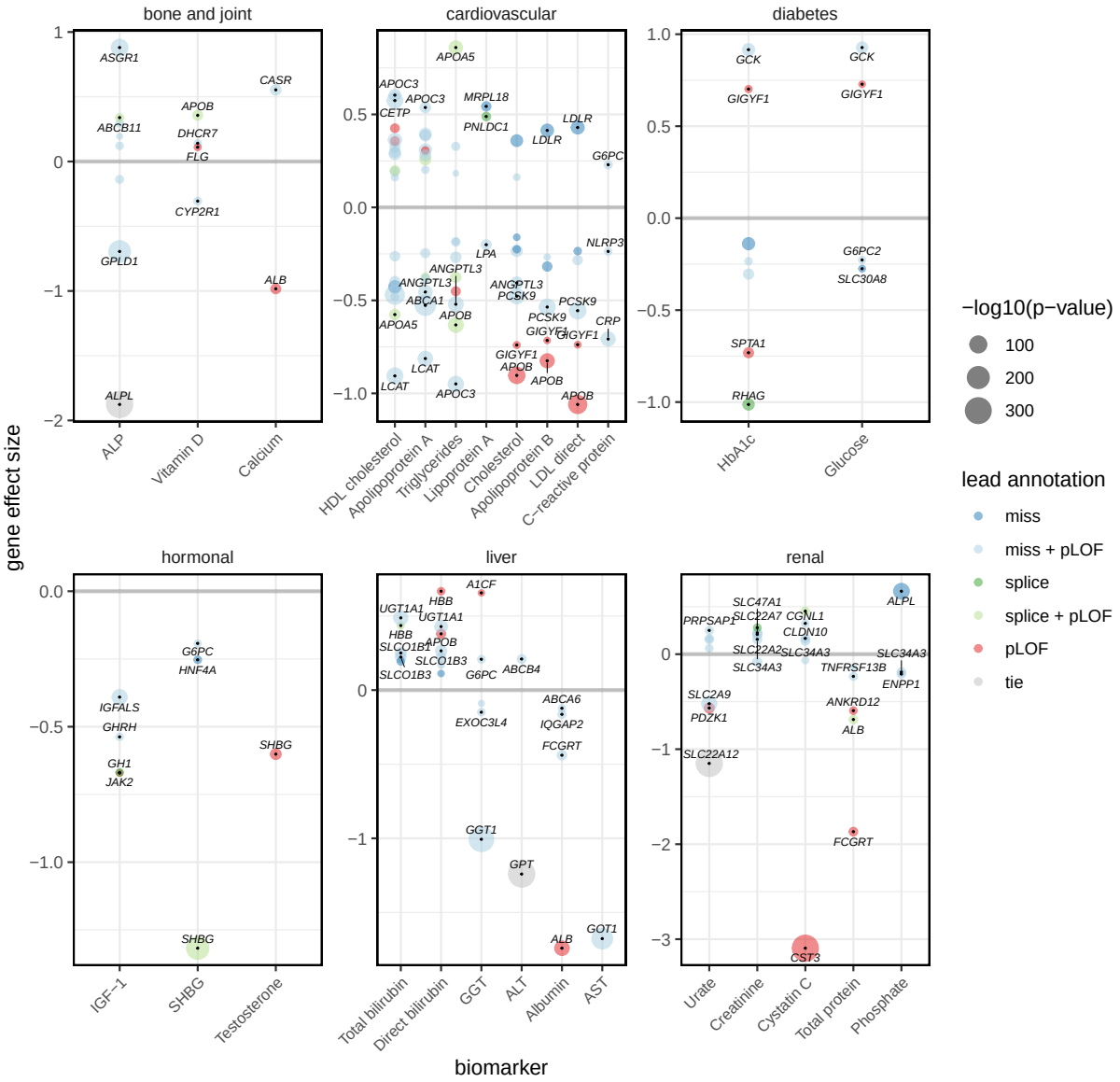
[76] Lippert, C. *et al.* Fast linear mixed models for genome-wide association studies. *Nature methods* **8**, 833–835 (2011).

[77] Scheipl, F., Greven, S. & Kuechenhoff, H. Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational statistics & data analysis* **52**, 3283–3299 (2008).

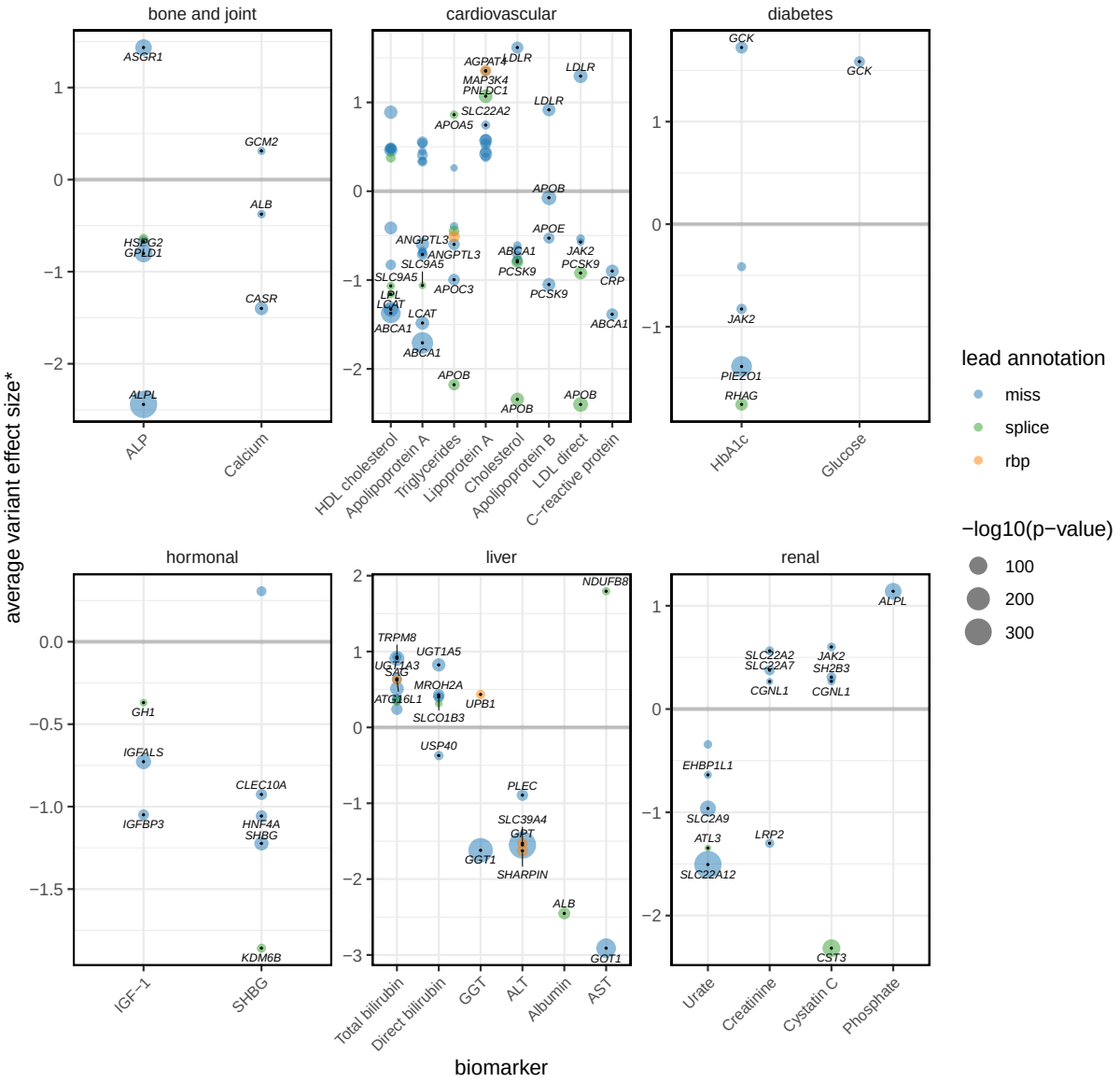
[78] Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13**, 762–775 (2012).



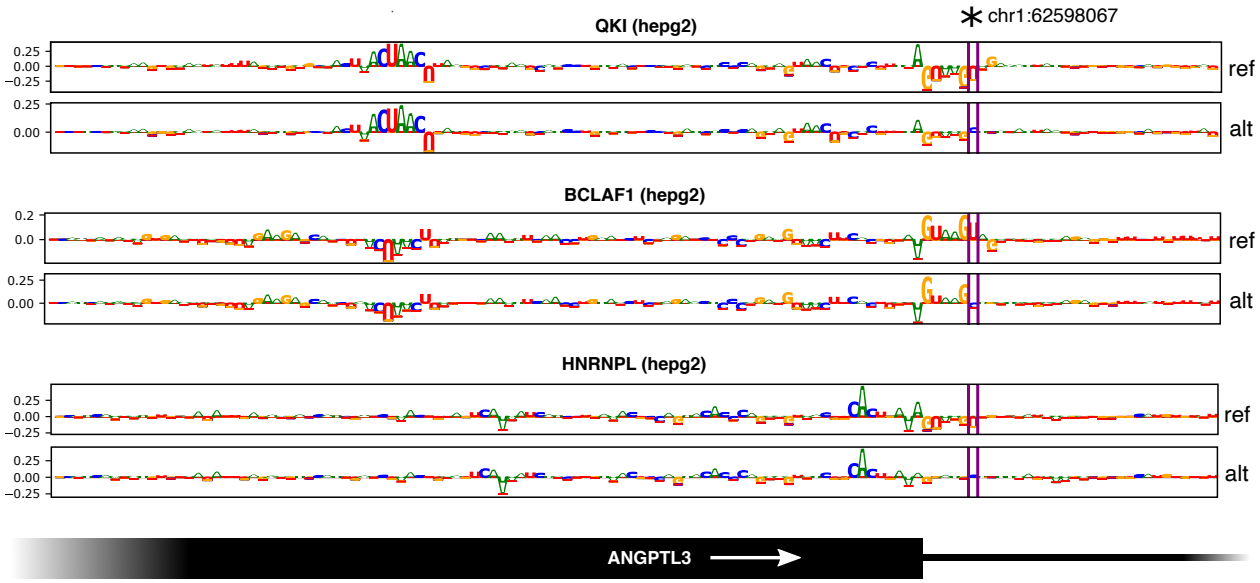
Supplementary Figure S1: **Biomarker correlation and number of hits.** Heatmap showing Pearson correlation between pre-processed biomarkers (upper triangle) and number of significant associations (cell notes). Rows and columns are clustered using complete linkage on the Euclidean distances of the correlation 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). ALP: alkaline phosphatase; ALT: alanine aminotransferase; AST: Aspartate aminotransferase; GGT: Gamma glutamyltransferase.



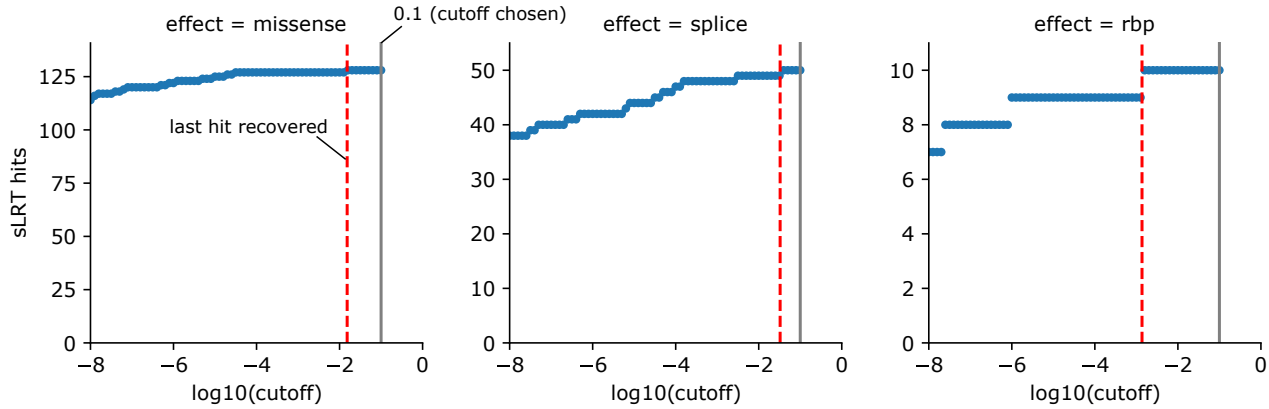
Supplementary Figure S2: **Gene-based variant collapsing results overview.** Collapsing variants allows defining gene effect sizes. Bubble plots showing the gene effect sizes (y-axis) of significant associations for each biomarker (x-axis). The four genes with largest absolute effect sizes are labeled 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. ALP: alkaline phosphatase; ALT: alanine aminotransferase; AST: Aspartate aminotransferase; GGT: Gamma glutamyltransferase.



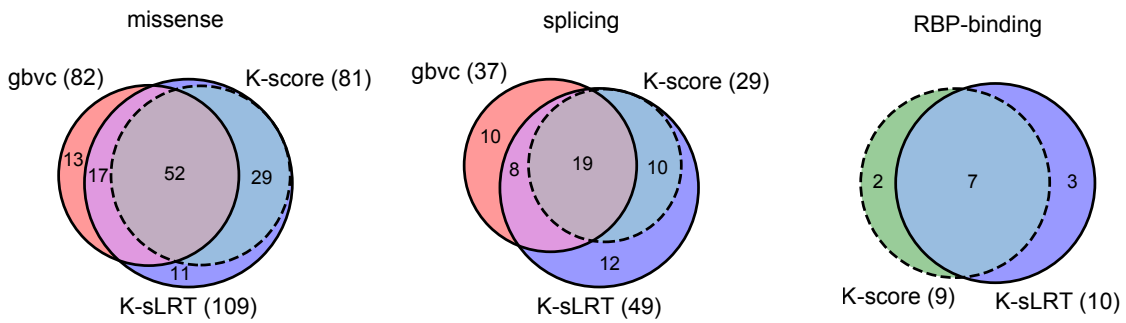
Supplementary Figure S3: **Kernel-based tests results overview.** We calculated the average effect size for *variants with single-variant p-values below 10^{-5} (score test) within significant genes found by kernel-based tests if the cumulative minor allele count across these variants was at least 5. Bubble plots showing these average effect sizes (y-axis) for each biomaker (x-axis). The four genes with largest average effect sizes are labeled for each biomaker. 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. ALP: alkaline phosphatase; ALT: alanine aminotransferase; AST: Aspartate aminotransferase; GGT: Gamma glutamyltransferase.



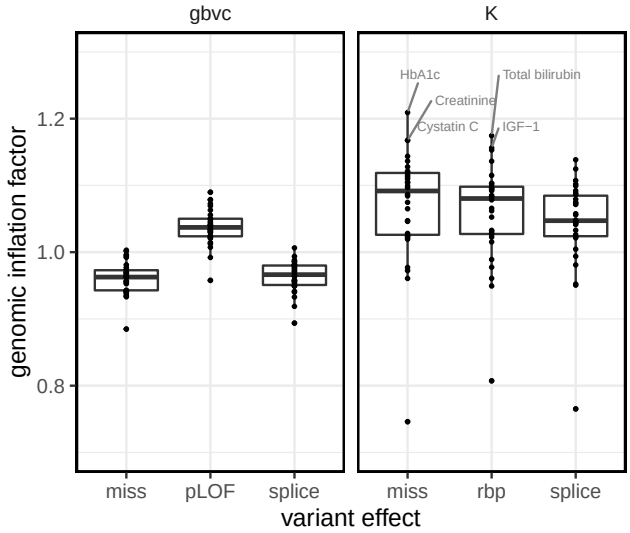
Supplementary Figure S4: **Attribution maps.** The variant 1:62598067:T:C at one-based position chr1:62598067 makes DeepRiPe predict increased binding probabilities for HNRNPL and QKI, and decreased probability for BCLAF1. Attribution maps for reference (ref) and alternative (alt) sequences as described in [23] 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 S5: **sLRT number of significant gene-biomarker associations vs score test cutoff for different variant effect categories.** Plots showing the number of significant associations found by the LRT depending on the nominal significance cutoff chosen for the score test (which determines whether the LRT is performed). We used the cutoff of 0.1 in our analysis. We found that a smaller nominal significance cutoff of 0.033 would still have recovered all genome-wide significant associations we reported, and a cutoff of 0.001 would have still recovered all but four (97.9%). Of those four associations, three would still have been found by association tests for other variant effect categories.



Supplementary Figure S6: **Kernel-based sLRT vs score test comparison.** 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, where performed, sLRT). For missense and splice variants, the hits identified by the kernel-based score test (dashed circle) were a subset of those identified by the kernel-based sLRT. The sLRT identified additional associations, of which a large fraction was also found by gbvc.



Supplementary Figure S7: **Genomic inflation factor across models.** We calculated λ for all tests that were performed exome-wide. Boxplots showing λ 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.

variable type	variable name	biomarker category	UKB datafield	N	GWAS catalog EFO	PhenoScanner EFO	comment
phenotype	Alanine aminotransferase	liver	30620	183037	EFO_0004735	EFO_0004735	abbreviation: ALT
phenotype	Albumin	liver	30600	168467	EFO_0004535	EFO_0004535	
phenotype	Alkaline phosphatase	bone and joint	30610	183105	EFO_0004533	EFO_0004533	abbreviation: ALP
phenotype	Apolipoprotein A	cardiovascular	30630	167375	EFO_0004614	EFO_0004614	
phenotype	Apolipoprotein B	cardiovascular	30640	182168	EFO_0004615	-	PhenoScanner queried by trait names
phenotype	Aspartate aminotransferase	liver	30650	182456	EFO_0004736	EFO_0004736	abbreviation: AST
phenotype	C-reactive protein	cardiovascular	30710	182681	EFO_0004458	EFO_0004458	abbreviation: CRP
phenotype	Calcium	bone and joint	30680	168380	EFO_0004838	EFO_0004838	
phenotype	Cholesterol	cardiovascular	30690	183098	EFO_0004574	EFO_0004574	
phenotype	Creatinine	renal	30700	183001	EFO_0004518	EFO_0004518	
phenotype	Cystatin C	renal	30720	183083	EFO_0004617	-	PhenoScanner queried by trait names
phenotype	Direct bilirubin	liver	30660	155280	EFO_0004570	EFO_0004570	
phenotype	Gamma glutamyltransferase	liver	30730	183018	EFO_0004532	EFO_0004532	abbreviation: GGT
phenotype	Glucose	diabetes	30740	168244	EFO_0004468	EFO_0004465	
phenotype	Glycated haemoglobin (HbA1c)	diabetes	30750	182851	EFO_0004541	EFO_0004541	
phenotype	HDL cholesterol	cardiovascular	30760	168371	EFO_0004612	EFO_0004612	
phenotype	IGF-1	hormonal	30770	182119	EFO_0004627	EFO_0004627	Insulin-like Growth Factor 1
phenotype	LDL direct	cardiovascular	30780	182783	EFO_0004611	EFO_0004611	
phenotype	Lipoprotein A	cardiovascular	30790	146534	EFO_0006925	EFO_0006925	
phenotype	Oestradiol	hormonal	30800	30409	-	-	no sign. associations
phenotype	Phosphate	renal	30810	168136	-	-	
phenotype	Rheumatoid factor	bone and joint	30820	16022	-	-	no sign. associations
phenotype	SHBG	hormonal	30830	166849	EFO_0004696	EFO_0004696	Sex Hormone Binding Globulin
phenotype	Testosterone	hormonal	30850	165537	EFO_0004908	EFO_0004908	
phenotype	Total bilirubin	liver	30840	182357	EFO_0004570	EFO_0004570	
phenotype	Total protein	renal	30860	168259	EFO_0004536	EFO_0004536	
phenotype	Triglycerides	cardiovascular	30870	182948	EFO_0004530	EFO_0004530	
phenotype	Urate	renal	30880	182900	EFO_0004531	EFO_0004531	
phenotype	Urea	renal	30670	182974	-	-	no sign. associations
phenotype	Vitamin D	bone and joint	30890	174820	EFO_0004631	EFO_0004631	
covariate	Age at recruitment	-	21022	192352	-	-	
covariate	BMI	-	21001	192352	-	-	Body Mass Index
covariate	Genetic principal components	-	22009	192352	-	-	PC1 - PC10
covariate	Sex	-	31	192352	-	-	female, male
covariate	Smoking status	-	20116	192352	-	-	never, previous, current
genotype	Exome OQFE variants, PLINK format	-	23155	192352	-	-	

Supplementary Table S1: **UK Biobank data** Variables, UK Biobank datafields and samples sizes (N). EFO terms were used to match result with those reported in the NHGRI-EBI Catalog of human genome-wide association studies (GWAS catalog) and PhenoScanner.

5 Supplementary Methods

5.1 Variant weight calculation

All association tests we performed incorporated variant weights, which were derived from the 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:

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

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 alleles, while applying 4bp shifts around the center. We averaged four predictions across these small shifts to reduce variability. Finally, we calculated variant effect predictions v_{ij} for each variant i and RBP-replicate j by subtracting the prediction for the reference allele ($p_{ij,ref}$) from the prediction for the alternative allele ($p_{ij,alt}$) [72]:

$$v_{ij} = p_{ij,alt} - p_{ij,ref} \tag{4}$$

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

5.2 Score- and likelihood ratio test implementation

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

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

Where $\boldsymbol{X}$ is the $N \times q$ covariate design matrix (fixed effect) and $\boldsymbol{\alpha}$ is the vector of fixed-effect parameters.

The variance-covariance matrix of $\boldsymbol{y}$ is composed of the independently distributed residual variance ($\boldsymbol{I}_N$ scaled by σ_e^2) and the kernel-matrix $\boldsymbol{K}_g$ (scaled by σ_g^2), which captures the genetic similarity between individuals. $\boldsymbol{K}_g$ is a function of the $N \times m$ matrix of mean-centered minor allele counts $\boldsymbol{G}$ (random effect) of the genetic variants we wish to test.

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

$$\boldsymbol{K}_g = \phi(\boldsymbol{G})\phi(\boldsymbol{G})^T \tag{6}$$

Where the function ϕ transforms $\boldsymbol{G}$ into intermediate variables before performing the test.

The test statistic of the score test approximates the change of the log likelihood of a model when including $\boldsymbol{K}_g$ over the null model, which does not include $\boldsymbol{K}_g$ ($\sigma_g^2 = 0$)[8]. We calculated test statistics using fast algorithms described in [11] and applied Davies's method for the calculation of p-values [74] with accuracy of 10^{-7} and 10^6 iterations. Where Davie's method returned p-values of 0, or in the rare cases where Davies method returned invalid (negative) p-values, we used saddle point approximation instead [75].

The test statistic of the likelihood ratio test is twice the difference of the log restricted likelihood of the alternative model and the null mode [9]. We used FaST-LMM's LMM class [76] to fit the null and alternative models using restricted maximum likelihood and then calculated test statistics. To generate a null distribution we sampled 100 test statistics for every LR test, using our own port of RLRsim [77] in Python. Finally, we fit a parametric null distribution $\pi\chi_0^2 + (1-\pi)a\chi_d^2$ with free parameters π, a and d to the pooled simulated test statistics using log-quantile regression on the 10% of largest test statistics, and used this distribution to calculate p-values as described in [9].

5.3 Gene-based variant collapsing tests

In gene-based variant collapsing, all qualifying variants overlapping a specific gene are collapsed into a single variable prior to association testing, i.e. $\phi(\boldsymbol{G})$ in Equation 6 returns an $N \times 1$-vector. We modified the approach in [12] by incorporating variant effect predictions as weights. Within a specific gene, any participant could carry 0 or more qualifying variants, where each variant i has a weight w_i (derived from variant effect prediction, see above). Specifically, the collapsed score is the largest weight of any of the variants observed for a specific participant, or 0 if no qualifying variants were observed for that participant. This score makes three assumptions: additive effects are negligible (or unrealistic), variants with larger weights dominate over those with smaller weights and all variants affect the quantitative trait in the same direction.

5.4 Functionally informed kernel-based tests

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

$$\boldsymbol{K}_g = \boldsymbol{GWSWG^T}, \tag{7}$$

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

While a shared regression coefficient ($\boldsymbol{S} = \mathbf{1}_m \mathbf{1}_m^T$) might be a poor assumption in some cases, so can completely independent regression coefficients ($\boldsymbol{S} = \boldsymbol{I}_m$). The former, when substituted into (7), has been referred to as the weighted counting burden test, whereas the latter is commonly called the weighted linear kernel [78]. In our analysis, we define $\boldsymbol{S}$ based on available prior knowledge and type of variant effect prediction.

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

$$\boldsymbol{K}_g = \boldsymbol{GWCC^TWG^T} \tag{8}$$

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

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

$$\boldsymbol{S} = \boldsymbol{LL}^T = \boldsymbol{Q} \circ \boldsymbol{R} \tag{9}$$

Where $\boldsymbol{Q}$ captures the similarity of variants based on their variant effect predictions and $\boldsymbol{R}$ captures the similarity of variants based on their positions. Specifically, let $\boldsymbol{v}_i$ be the vector of variant effect predictions for variant i. Then the element q_{ij} of $\boldsymbol{Q}$ is the cosine similarity between $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$. We chose to model

the position-dependent similarity with a Gaussian kernel. If x_i is the chromosomal position of variant i, $r_{i,j} = \exp(-\gamma(x_i - x_j)^2)$, where we set $\gamma = -\frac{\log(0.5)}{50^2}$. At this value of γ two variants that are 50bp apart have a similarity of 0.5, which decays rapidly as the distance increases. As both $\boldsymbol{Q}$ and $\boldsymbol{R}$ are positive definite matrices, so is $\boldsymbol{Q} \circ \boldsymbol{R}$. This kernel makes the assumption that variants that are in close proximity and have aligned variant effect predictions should affect the phenotype in the same direction.

5.5 sLRT detailed description

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

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

Splicing For splice-variants we performed score tests using gene-based variant collapsing and the linear weighted kernel for all genes. Again, if either of the two score tests were nominally significant ($p < 0.1$), we performed likelihood ratio tests (sLRT). As we did for missense variants, we then also performed combined association tests with protein loss of function variants using both gene-based variant collapsing and a kernel-based LRT. For the kernel-based test, we concatenated the protein LOF indicator variable to the matrix of weighted minor allele counts $\boldsymbol{GW}$ (Equation 7, where $\boldsymbol{S} = I_m$). In the cases where a variant was annotated both as a splice-variant and pLOF variant, we treated it as a pLOF variant in the joint tests.

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

5.6 Cross-referencing against GWAS databases

We queried the NHGRI-EBI GWAS Catalog [2] and PhenoScanner [24, 25] in order to see if single variants within the genes we found significantly associated with a specific biomarker had already been reported to be associated with that biomarker. For each gene, we submitted region queries using the gene boundaries with the gwasrapidd [28] and phenoscanner R-packages. For PhenoScanner, we set the p-value threshold to 10^{-7}. Matching our results to those contained in these databases required us to define a mapping of UK Biobank biomarkers to the Experimental Factor Ontology (EFO) terms used in those databases. This mapping is provided in Supplementary Table S1. Additionally, as EFO terms for PhenoScanner were not always defined, we performed the following matching: "Apolipoprotein B" (UKB phenotype) to "APOB apolipoprotein B" (PhenoScanner trait) and "Cystatin C" to PhenoScanner traits "log eGFR cystatin C", "Serum cystatin c estimated glomerular filtration rate eGFR" and "Cystatin C in serum". The UK Biobank biomarker "Phosphate" was not defined in either database and could therefore not be queried.