Transcript-aware analysis of rare predicted loss-of-function variants in the UK Biobank elucidate new isoform-trait associations

Rachel A. Hoffing, Aimee M. Deaton, Aaron M. Holleman, Lynne Krohn, Philip J. LoGerfo, Mollie E. Plekan, Sebastian Akle Serrano, Paul Nioi, Lucas D. Ward

Alnylam Pharmaceuticals
Cambridge, MA 02142, USA
Email: rhoffing@alnylam.com

A single gene can produce multiple transcripts with distinct molecular functions. Rare-variant association tests often aggregate all coding variants across individual genes, without accounting for the variants' presence or consequence in resulting transcript isoforms. To evaluate the utility of transcript-aware variant sets, rare predicted loss-of-function (pLOF) variants were aggregated for 17,035 protein-coding genes using 55,558 distinct transcript-specific variant sets. These sets were tested for their association with 728 circulating proteins and 188 quantitative phenotypes across 406,921 individuals in the UK Biobank. The transcript-specific approach resulted in larger estimated effects of pLOF variants decreasing serum *cis*-protein levels compared to the gene-based approach ($p_{binom} \leq 2x10^{-16}$). Additionally, 251 quantitative trait associations were identified as being significant using the transcript-specific approach but not the gene-based approach, including *PCSK5* transcript ENST00000376752 and standing height (transcript-specific statistic, $P = 1.3x10^{-16}$, effect = 0.7 SD decrease; gene-based statistic, P = 0.02, effect = 0.05 SD decrease) and *LDLR* transcript ENST00000252444 and apolipoprotein B (transcript-specific statistic, $P = 5.7x10^{-20}$, effect = 1.0 SD increase; gene-based statistic, $P = 3.0x10^{-4}$, effect = 0.2 SD increase). This approach demonstrates the importance of considering the effect of pLOFs on specific transcript isoforms when performing rare-variant association studies.

Keywords: UK Biobank; rare variant; transcriptome; quantitative traits

1. Introduction

Alternative splicing allows for one gene to produce many transcript isoforms. When these isoforms differ in their coding sequence content, they can result in proteins with distinct molecular functions. Over 95% of protein-coding genes are alternatively spliced[1] which contributes to the large diversity of the human transcriptome and proteome. This process is instrumental in creating the complex and coordinated gene expression patterns that underlie all biological processes.

Alterations to the transcriptome by genetic variation is instrumental in driving differences in phenotypic expression. Many of these disruptions have been identified through genome-wide association studies (GWAS), which test the impact of common single nucleotide variants (SNVs) on phenotypes on a population scale. Studies such as these are critical in drug-discovery efforts, as they can be used to identify new therapeutic targets for disease. Additionally, lack of genetic validation of therapeutic hypotheses has been shown to reduce the likelihood of a successful clinical trial[2,3], suggesting that genetically validated targets are essential in the development process.

A large amount of phenotype heritability is not well captured through common-variant GWAS alone[4]. Rare, coding SNVs can be exceptionally disruptive to the transcriptome and have dramatic

effects on phenotypes, even more so than common variants[5]. Rare variant association studies (RVAS) provide avenues to explain the "missing heritability" of traits[6], and provide a complementary approach to common-variant GWAS. Though, assessing genotype-phenotype associations with these low-frequency SNVs is difficult due to lack of sufficient sample size and statistical power.

One approach in ameliorating the statistical challenges of rare variant analysis is the aggregation of SNVs with similar predicted functional consequences, known as burden testing[7,8]. For example, an analysis may collect rare protein-truncating variants (PTVs), also known as predicted loss-of-function variants (pLOF), that are expected to result in non-functional gene products through nonsense-mediated decay[9]. These pLOF variants can then be aggregated and tested for their collective association with phenotypes of interest. This allows for an increase in statistical power and ability to detect genotype-phenotype associations which would otherwise be impossible at the level of single-variant tests.

However, burden testing assumes that all aggregated variants will have a similar effect on the function of the gene and, consequently, the associated phenotype. This assumption does not hold if a gene has multiple transcript isoforms with diverse downstream functions. For example, where a given SNV may encode a missense variant that is deleterious in some encoded protein isoforms but not others, or where an SNV may encode a variant of any function that overlaps some transcript isoforms but is not transcribed in others. The most common techniques for creating variant sets for burden testing consider the most deleterious consequence of an SNV across all documented isoforms. Subsequently, the expected impact of the SNV may be overestimated. Due to these challenges, we propose the inclusion of transcript-aware analyses when studying rare variants, in addition to the standard gene-based approach.

Our analysis uses whole-exome sequencing data from the UK Biobank to perform transcript-specific burden analyses on 406,921 individuals of European ancestry. Rare pLOFs were identified across 17,035 genes and aggregated by transcript, resulting in 55,558 unique, transcript-specific variant sets tested against the circulating levels of 728 *cis*-encoded proteins and 188 quantitative traits. The results of the transcript-specific burden tests were compared to the results from the maximally inclusive, standard, gene-based burden method.

2. Data

2.1. *UK Biobank*

The UK Biobank consists of approximately 500,000 volunteer participants, who were aged 40–69 years when recruited between 2006 and 2010[10,11]. Both array genotyping and whole-exome sequencing have been performed on most of these participants[12]. Data from genotyping, sequencing, questionnaires, primary care data, hospitalization data, cancer registry data, and death registry data were obtained through application number 26041. Proteomic profiling was also performed on a subset of participants through application number 65851[13]. Ethical oversight for the UK Biobank is provided by an Ethics and Governance Council which obtained informed consent from all participants to use these data for health-related research. Data management and analytics were performed using the REVEAL/SciDB translational analytics platform from Paradigm4.

2.2. *Variant calling and definition*

The source of genetic data for the main analysis was exome sequencing data. DNA from whole blood was extracted and sequenced by the Regeneron Genetics Center (RGC) using protocols previously described[14]. Of the variants called by RGC, additional quality-control filters were applied: Hardy-Weinberg equilibrium (among the European subpopulation, as defined by Pan-UKB[15]) $P > 1x10^{-10}$, and missingness across all individuals less than 2%. Variants were annotated using ENSEMBL Variant Effect Predictor (VEP)[16] version 109.3, using the LOFTEE plug-in version 1.0.4 to identify high-confidence predicted PTV variants[9] in protein-coding genes with minor allele frequency <1%. Bcftools was used to filter variants with genotype quality (GQ) > 20 and depth (DP) > 7 or 10 for SNPs and indels respectively. For the gene-based burden, variant effects were scored against all available transcripts in ENSEMBL, and the most severe predicted impact was retained. Variants were aggregated in each protein-coding gene as follows: pLOF variants were defined as "HC" (high confidence) from LOFTEE and their most severe consequence from VEP as "stop gained," "splice donor," "splice acceptor," or "frameshift." For the transcript-based burden, the consequence for each variant was assessed individually by transcript.

2.3. *Participant definition for overall analyses*

An initial round of quality control was performed by RGC, which removed subjects with evidence of contamination, discrepancies between chromosomal and reported sex, and high discordance between sequencing and genotyping array data. A European ancestry population was defined using data from the Pan-UKB Team[15], resulting in a set of 406,921 European ancestry individuals with exome sequencing data available. Two sets of genetic principal components (PCs) were defined, as described by Backman et al[17]: a set derived from common array variants, of which 10 were used, and a set derived from rare exome variants, of which 20 were used. Rare exome derived PCs were calculated by applying the following filters on variants on the autosomes: MAF > $2.6x10^{-5}$ and < 0.01, Hardy-Weinberg equilibrium P > $1x10^{-12}$, and genotype missingness < 2%. Regions of high LD were removed, and SNPs were pruned with PLINK's[18] indep-pairwise function, using a window-size of 1,000 base pairs, a step size of 100 base pairs, and an R^2 threshold of 0.1. Indels were removed, then R's Smart PCA was implemented to derive the PCs. Array derived PCs for the European subset were derived by imposing a MAF filter > 0.01 and INFO score = 1 before running Smart PCA.

2.4. *Phenotype sources*

The main source of phenotype data was from a release of structured data by the UK Biobank Data Showcase on December 22, 2022. We tested 188 quantitative phenotypes, including physical measures, blood counts, metabolomics, touchscreen questionnaire responses on family history, telomere length, and urine biochemistry. Quantitative traits were rank-based inverse normal transformed to have a mean of zero and a standard deviation of one.

2.5. *Tissue-expressed transcript isoforms*

GTEx version 8 bulk RNAseq data was aggregated across 54 tissue types from 948 donors. For each gene, expression was calculated across all tissues, identifying 145,219 transcripts with mean TPM expression > 0.

2.6. *Olink proteomics*

Characterization of 1,463 proteins across 54,306 individuals was undertaken by the UK Biobank Proteomics Project (UKB-PPP). Proteomic profiling was conducted across four panels utilizing the Olink Explore Assay. Sample collection, preparation, data pre-processing, and quality control is described in detail in Sun et al[13]. Quantified protein expression levels were rank-based inverse normal transformed to have a mean of zero and standard deviation of one.

3. Methods

3.1. *Transcript-specific variant set curation*

Rare, predicted loss-of-function (pLOF) variants sets (MAF < 1%) were created across 145,219 transcripts with mean TPM > 0 across all 53 GTEx tissue types. Overall, 72,769 transcripts had at least one overlapping rare pLOF variant. Identical variant sets that were representative of more than one transcript were combined into a single label, resulting in 55,558 unique transcript-specific variant sets across 17,035 genes.

3.2. *Whole-genome ridge regression analysis*

REGENIE v3.1.1[19] was used to perform a whole-genome ridge regression taking subject relatedness into account, while using a Firth approximation to estimate P values. For all quantitative traits, REGENIE was performed using an additive model across the entire European-ancestry population, including related individuals, controlling for age, sex, age^2, age x sex, age^2 x sex, 10 rare-variant derived principal components, and 20 common-variant derived principal components. For the Olink proteomics, batch numbers 1-7 were added as one-hot encoded covariates.

3.3. *Comparison of estimated effect sizes by approximating a binomial distribution*

The effect sizes across transcript and gene-based burden tests were compared in cases only where there was a significant association for a quantitative phenotype in both methods. Deviation from a binomial distribution was modeled using R's binom.test() to determine is the proportion of results with stronger associations in the transcript-based model differs from the null hypothesis.

3.4. *Binary case-control phenotype regression*

As a follow-up to the quantitative traits analysis, we tested a single binary phenotype, Alzheimer's disease, across multiple *TREM2* transcript-specific variant sets. Diagnoses were extracted from inpatient hospital diagnoses, the cancer and death registries, primary care, and self-reported data. We adjusted for age, sex, age^2, age x sex, age^2 x sex, 10 rare-variant derived principal components, 20 common-variant derived principal components, availability of primary care, and country of recruitment.

4. Results

4.1. *Transcript-specific variant sets show stronger associations with lower serum cis-protein levels*

To evaluate the validity of transcript-specific pLOF variant sets, they were first tested for their association with *cis*-encoded proteins. Variant sets with at least 10 carriers were tested across 728 circulating serum proteins in 47,297 individuals of European ancestry and compared to the gene-based approach. Several gene and transcript-specific variant sets were identical, and their removal resulted in 913 unique transcript variant sets tested across 432 serum protein levels. Among 580 results that were significant for both the transcript and gene-based burden approach, 75% (N = 437) had lower effect estimates on *cis*-serum proteins in the transcript-based burden (Figure 1), which is substantially greater than expected by chance ($p_{binom} \leq 2x10^{-16}$). Of the 437 transcript-based results with lower *cis*-protein effect estimates, 45 had non-overlapping 95% confidence intervals with the effect estimates of the gene-based approach.

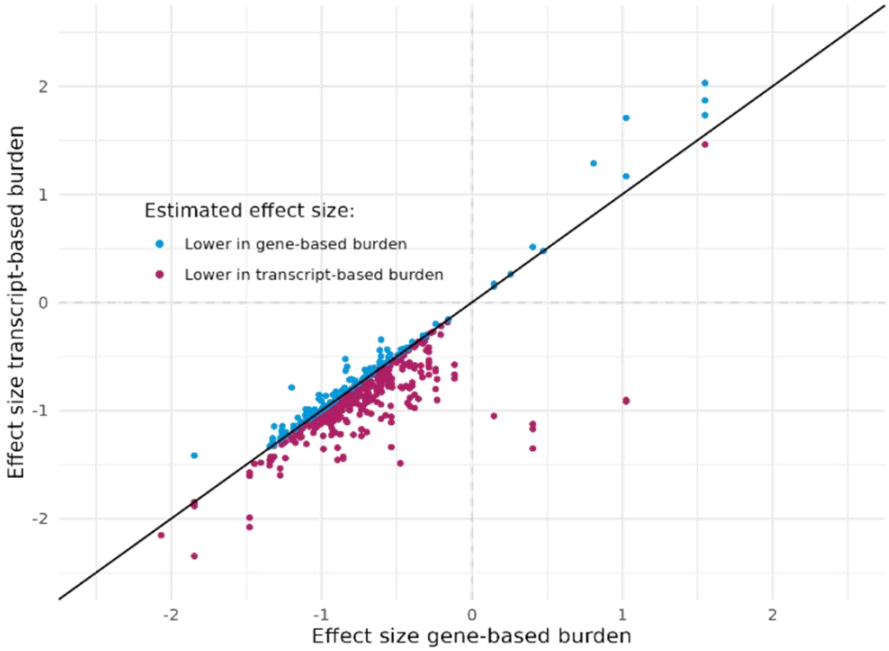

Figure 1. Comparison of estimated effect sizes on circulating serum proteins. Each dot represents an association of the transcript or gene-based burden with a *cis*-encoded protein.

4.2. *Some pLOF-cis protein associations are only detectable using transcript-specific variant sets*

The transcript-based burden on *cis*-proteins resulted in 35 associations across 21 loci that were non-significant in the gene-based burden. Of these associations only significant in the transcript-based burden, 22 associations across 12 loci had non-overlapping 95% confidence intervals with the gene-based approach (Table 1), and all of them had lower estimated effect sizes.

Gene/*cis*-protein	P value gene -based burden	Effect size gene -based burden	N carriers gene -based burden	95% CI gene -based burden	P value transcript -based burden	Effect size transcript -based burden	N carriers transcript-based burden	95% CI transcript-based burden	Transcripts
CD300LF	$5.3x10^{-1}$	-0.03	195	-0.1,0.1	$3.6x10^{-19}$	-1.0	38	-1.2,-0.8	ENST00000326165; ENST00000464910; ENST00000583937
CD84	$4.12x10^{-5}$	-0.2	51	-0.3,-0.1	$1.3x10^{-20}$	-1.0	10	-1.2,-0.8	ENST00000368048; ENST00000368051; ENST00000368054
CLEC10A	$3.0x10^{-4}$	-0.1	103	-0.2,-0.1	$7.1x10^{-18}$	-1.0	13	-1.2,-0.7	ENST00000416562; ENST00000571664
CPPED1	$4.6x10^{-2}$	-0.2	43	-0.5,0.01	$3.2x10^{-13}$	-1.6	13	-2.0,-1.1	ENST00000381774
MSR1	$1.4x10^{-2}$	-0.1	110	-0.2,-0.02	$5.0x10^{-13}$	-0.8	23	-1.0,-0.6	ENST00000262101
MSRA	$5.9x10^{-3}$	-0.2	93	-0.4,-0.1	$7.3x10^{-8}$	-1.3	13	-1.7,-0.8	ENST00000528246
NRP2	$1.8x10^{-1}$	0.1	32	-0.04,0.2	$2.6x10^{-14}$	-0.7	13	-0.9,-0.5	ENST00000357785
PLXNB2	$1.5x10^{-2}$	-0.1	85	-0.1,-0.01	$4.9x10^{-10}$	-0.4	23	-0.5,-0.3	ENST00000359337; ENST00000449103
SETMAR	$5.8x10^{-2}$	-0.1	133	-0.1,0.02	$1.2x10^{-09}$	-0.3	51	-0.4,-0.2	ENST00000425863
TREM2	$8.8x10^{-3}$	-0.2	66	-0.3,-0.04	$2.2x10^{-16}$	-1.2	17	-1.5,-0.9	ENST00000373113
TXNRD1	$5.4x10^{-5}$	-0.4	35	-0.6,-0.2	$1.3x10^{-9}$	-1.0	12	-1.3,-0.7	ENST00000503506; ENST00000524698; ENST00000526390; ENST00000526950; ENST00000529546
TYMP	$2.6x10^{-2}$	-0.2	54	-0.4,-0.02	$2.6x10^{-5}$	-0.9	13	-1.3,-0.5	ENST00000425169

Table 1. Transcript-specific results with significant association on circulating *cis*-proteins, and transcript-based burden 95% CI not-overlapping with gene-based burden 95% CI. Multiple transcripts listed when variant sets are identical.

From these data, we focused on *TREM2* as it has a known role in Alzheimer's disease (AD) risk. *TREM2* is primarily expressed in microglia, and rare loss-of-function mutations including the missense variant R47H have been shown to increase AD risk[20]. When testing *TREM2* transcript-specific pLOF variant sets (Figure 2), we observe more significant associations with larger reductions in serum TREM2 levels in the ENST00000338469 and ENST00000373113 models, compared to ENST00000373122 or the gene-based method (Table 2).

The primary variant that explains the difference in signal is rs538447052, a splice acceptor variant at the boundary of exon 4. The canonical transcript with the highest brain expression[21], ENST00000373113, and ENST00000338469, are both unaffected by rs538447052 as it functions there as an intron variant. By excluding rs538447052 from these variant sets, we see a much stronger association with decreasing serum TREM2.

Next, we tested the relationship between Alzheimer's disease and *TREM2* and its transcript isoforms. Our analysis is limited by a low number of affected carriers; however, we detect an enrichment of AD cases when using the more stringent *TREM2* transcript models, ENST00000373113 and ENST00000338469 (Table 2). This association is absent in the ENST00000373122 and gene-based models and is consistent with the weaker observed effects on serum TREM2.

Figure 2. *TREM2* transcript models and the gene-based, inclusive model.

Transcripts	P value AD	Odds ratio AD	95% CI AD	N carriers AD	N carriers with AD	P value serum TREM2	Effect size serum TREM2	95% CI serum TREM2	N carriers serum TREM2
ENST00000338469	$8.9x10^{-3}$	10.3	2.6,40.9	48	2	$7.2x10^{-16}$	-1.4	-1.0,-1.7	12
ENST00000373113	$1.2x10^{-2}$	9.1	2.3,35.9	55	2	$2.1x10^{-16}$	-1.2	-1.4,-0.9	17
Inclusive model	$9.6x10^{-2}$	1.0	0.3,3.1	435	3	$8.7x10^{-3}$	-0.2	-0.3,0.0	66
ENST00000373122	$9.8x10^{-2}$	01.0	0.30,3.1	428	3	$4.9x10^{-2}$	-0.2	-0.3,0.0	61

Table 2. *TREM2* transcript-specific associations with AD and circulating TREM2 levels.

4.3. *Some pLOF-cis protein associations have opposite directions of effect in the transcript and gene-based models*

Most effect size estimates maintain their direction of effect when comparing the gene and transcript-based methods. However, six associations from three loci resulted in opposing estimated effect sizes (Table 3). In all six cases, the transcript-variant set of pLOFs associates with lower serum *cis*-protein levels, as expected, while the gene-based method associates with higher serum *cis*-protein levels.

Gene/*cis*-protein	P value gene-based burden	Effect size gene-based burden	95% CI gene-based burden	N carriers gene-based burden	P value transcript-based burden	Effect size transcript-based burden	95% CI transcript-based burden	N carriers transcript-based burden	Transcripts
BST1	$8.8x10^{-69}$	0.4	0.4,0.5	543	$6.6x10^{-37}$	-1.1	-1.3,-01.0	37	ENST00000265016; ENST00000382346
BST1	$8.8x10^{-69}$	0.4	0.4,0.5	543	$1.1x10^{-36}$	-1.2	-1.4,-1.0	34	ENST00000505785
BST1	$8.8x10^{-69}$	0.4	0.4,0.5	543	$2.7x10^{-22}$	-1.4	-1.6,-1.1	15	ENST00000514445
GPNMB	$2.9x10^{-21}$	0.2	0.1,0.3	446	$1.1x10^{-212}$	-1.1	-1.1,-1.0	93	ENST00000409458
HMOX2	$2.2x10^{-21}$	1.0	0.8,1.2	26	$5.2x10^{-8}$	-0.9	-1.2,-0.6	11	ENST00000570445; ENST00000575051
HMOX2	$2.2x10^{-21}$	1.0	0.8,1.2	26	$1.4x10^{-7}$	-0.9	-1.3,-0.6	10	ENST00000574466; ENST00000575129; ENST00000576827

Table 3. Significant transcript-based burden results with opposing effect sizes compared to the gene-based burden. Multiple transcripts are listed when the variant sets are identical.

The difference in variants captured by the *BST1*, *GPNMB*, and *HMOX2* gene-based and transcript-based variant sets are primarily attributable to variants missing from the terminal exon (Figure 3). The most significantly associated transcript variant set for each locus mainly exclude a single, frequent variant from the last exon, rs144539516, rs11537976, and rs11537976, respectively.

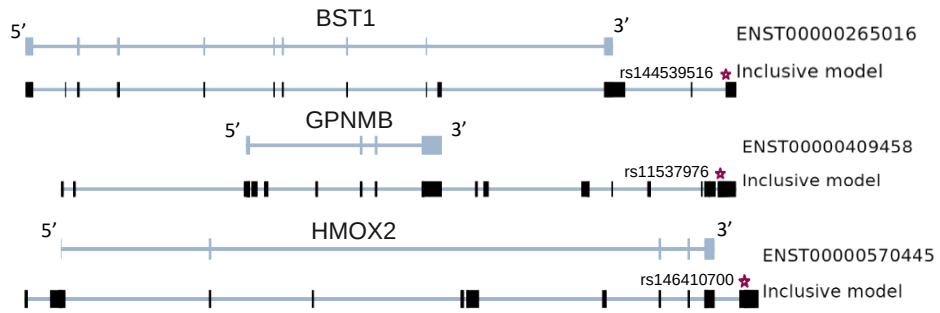

Figure 3: *BST1*, *GPNMB*, *HMOX2* inclusive gene-based model and representative transcript-based models.

Each of these excluded variants strongly associate with increased *cis*-serum protein levels when tested individually (Table 4). Rs144539516 and rs146410700 are 3' UTR variants in at least one transcript, which may affect the post-transcriptional stability of the RNA product. Rs146410700 also occasionally is identified as a missense variant in some transcripts and could influence protein stability, detectability, and post-translational regulation. Rs11537976 acts as a non-coding exon variant and may affect transcriptional regulation. In all instances, this provides an explanation for the unexpected gene-level association with increased protein.

Cis-protein	Rsid	P value	Effect size	95% CI	N carriers
GPNMB	rs11537976	$1.5x10^{-233}$	0.6	0.5, 0.6	318
BST1	rs144539516	$7.6x10^{-102}$	0.5	0.5, 0.6	506
HMOX2	rs146410700	$1.2x10^{-91}$	3.4	3.0, 3.7	11

Table 4: Single variant *cis*-protein association results for *BST1*, *GPNMB*, *HMOX2* variants rs11537976, rs144539516, and rs146410700

4.4. *Transcript-specific variant sets show stronger associations with quantitative traits*

Since the transcript-based variant sets show larger effects on circulating *cis*-proteins compared to the gene-based method, we next extended the analysis to quantitative traits. Transcript-specific pLOF variant sets with at least 10 carriers were tested for their association with 318 quantitative traits in 406,921 individuals of European ancestry and compared to the gene-based approach. After removing identical results between the transcript and gene-based approach, 6,981,491 transcript-trait and 2,740,011 gene-trait association tests were performed (Bonferroni corrected P value < $5.1x10^{-9}$). Among 1,010 associations that were significant in both the transcript and gene-based approach, 73% (N = 745) had more extreme effect sizes in the transcript-specific approach (Figure 4), which is substantially larger than expected by chance ($p_{binom} \leq 2x10^{-16}$). Of these, 75 had non overlapping 95% confidence intervals with the gene-based approach. Additionally, 46% of associations significant in both methods were more significant in the transcript-approach despite having a lower number of tested carriers in practically all instances.

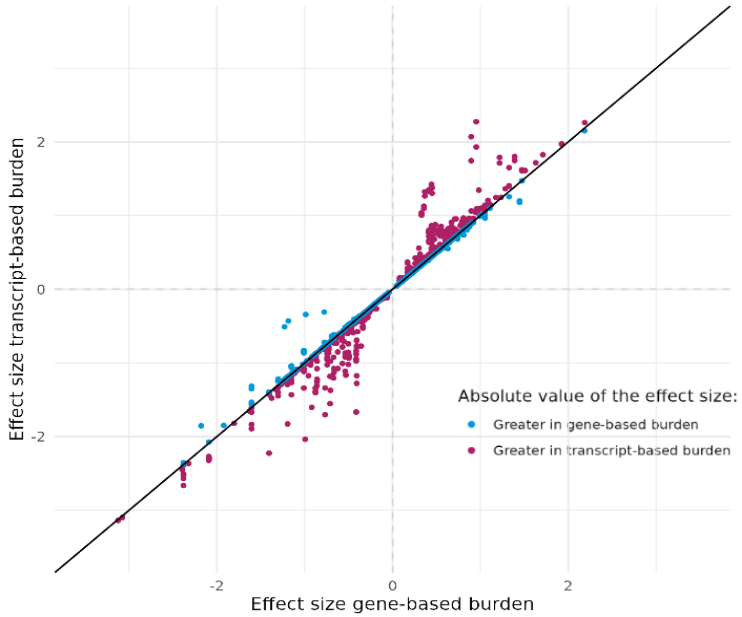

Figure 4. Comparison of estimated effect sizes on 188 quantitative traits for the transcript and gene-based burden. Each dot represents an association of the transcript or gene-based burden with a quantitative trait.

4.5. *Transcript-specific variant sets elucidate novel transcript-trait associations*

We identified 241 associations across 60 loci as being significant in the transcript-based approach but not in the gene-based burden. Of these, 56 transcript-trait associations had effect estimates with non-overlapping 95% confidence intervals with the gene-based burden (Table 5). These include *PCSK5* transcript ENST00000376752 and standing height (transcript-specific statistic, P = $1.3x10^{-16}$, effect = 0.72 SD decrease; gene-based statistic, P = 0.02, effect = 0.05 SD decrease) and *LDLR* transcript ENST00000252444 and apolipoprotein B (transcript-specific statistic, P = $5.7x10^{-20}$, effect = 1.0 SD increase; gene-based statistic, P = $3.0x10^{-4}$, effect = 0.2 SD increase). These data reflect genotype-phenotype associations that would have been otherwise undetected if testing only the standard, gene-based burden.

Gene	Phenotype	P value gene-based burden	Effect size gene-based burden	N carriers gene-based burden	95% CI gene-based burden	P value transcript-based burden	Effect size transcript-based burden	N carriers transcript-based burden	95% CI transcript-based burden	Transcripts
EPB41	Reticulocyte percentage	$8.5x10^{-9}$	0.33	282	0.2,0.4	$6.9x10^{-22}$	1.1	77	0.8,1.3	ENST00000373800
LDLR	Apolipoprotein B	$3.0x10^{-4}$	0.17	425	0.1,0.3	$5.8x10^{-20}$	1.0	78	0.8,1.2	ENST00000252444
SCUBE3	Standing height	$1.2x10^{-4}$	-0.12	373	-0.2,-0.1	$1.4x10^{-18}$	-0.6	71	-0.8,-0.5	ENST00000274938
EPB41	Total bilirubin	$1.5x10^{-4}$	0.19	275	0.1,0.3	$8.3x10^{-17}$	0.8	74	0.6,1.0	ENST00000373800
PCSK5	Standing height	$1.9x10^{-2}$	-0.05	829	-0.1,-0.01	$1.3x10^{-16}$	-0.7	50	-0.9,-0.5	ENST00000376752
UGT1A9	Total bilirubin	$6.9x10^{-3}$	0.17	227	0.04,0.3	$8.7x10^{-16}$	0.4	355	0.3,0.5	ENST00000354728
TINF2	Telomere Length	$1.3x10^{-5}$	0.44	92	0.2,0.6	$1.1x10^{-15}$	2.1	14	1.5,2.6	ENST00000557921
PFKM	HbA1c	$4.8x10^{-8}$	-0.25	391	-0.3,-0.2	$1.4x10^{-15}$	-0.5	201	-0.6,-0.4	ENST00000549941
TTN	Systolic blood pressure	$8.6x10^{-9}$	-0.08	4430	-0.1,-0.1	$1.8x10^{-14}$	-0.2	1807	-0.2,-0.1	ENST00000359218

CHD2	Lymphocyte percentage	$1.4x10^{-5}$	0.50	67	0.3,0.7	$1.8x10^{-14}$	2.1	12	1.5,2.6	ENST00000394196
NF1	Lymphocyte percentage	$4.4x10^{-8}$	-0.29	312	-0.4,-0.19	$3.5x10^{-13}$	-1.4	25	-1.8,-1.0	ENST00000431387
IGF2BP2	Standing height	$2.5x10^{-8}$	-0.47	53	-0.6,-0.3	$4.0x10^{-13}$	-0.9	25	-1.1,-0.7	ENST00000346192; ENST00000382199
PKD1	Urate	$3.3x10^{-6}$	0.21	313	0.1,0.3	$4.5x10^{-13}$	0.6	97	0.4,0.8	ENST00000423118
CREB3L3	Apolipoprotein A	$6.5x10^{-4}$	-0.07	1709	-0.1,-0.03	$2.2x10^{-12}$	-0.4	205	-0.6,-0.3	ENST00000078445; ENST00000595923
CLEC11A	Standing height	$3.7x10^{-5}$	-0.02	13452	-0.03,-0.01	$3.5x10^{-12}$	-0.1	5719	-0.1,-0.04	ENST00000250340
TPM4	Thrombocyte volume	$9.1x10^{-6}$	0.62	42	0.3,0.9	$8.4x10^{-12}$	1.8	12	1.3,2.3	ENST00000586833
COL18A1	Apolipoprotein A	$6.6x10^{-4}$	-0.07	1876	-0.1,-0.03	$1.1x10^{-11}$	-0.2	620	-0.3,-0.2	ENST00000355480
PFKM	Pyruvate	$8.7x10^{-9}$	0.55	105	0.4,0.7	$1.4x10^{-11}$	1.2	30	0.9,1.6	ENST00000546465
RNF10	Reticulocyte count	$1.7x10^{-1}$	0.04	1296	-0.02,0.1	$3.2x10^{-11}$	0.7	76	0.5,0.9	ENST00000413266
ANK1	HbA1c	$6.4x10^{-3}$	-0.23	118	-0.4,-0.1	$8.3x10^{-11}$	-0.9	47	-1.1,-0.6	ENST00000520299
NF1	Standing height	$1.0x10^{-6}$	-0.17	322	-0.2,-0.1	$9.6x10^{-11}$	-0.8	27	-1.0,-0.5	ENST00000431387
MARCHF8	Erythrocyte distribution width	$1.1x10^{-4}$	0.17	456	0.1,0.3	$9.7x10^{-11}$	0.5	165	0.3,0.6	ENST00000319836; ENST00000395769
PRC1	Platelet crit	$1.7x10^{-5}$	-0.22	304	-0.3,-0.1	$1.8x10^{-10}$	-0.5	130	-0.7,-0.4	ENST00000442656
LARP1	Mean corpuscular hemoglobin	$1.7x10^{-2}$	-0.23	87	-0.4,-0.04	$8.9x10^{-10}$	-1.0	30	-1.3,-0.7	ENST00000518297
MARCHF8	Immature reticulocyte fraction	$7.7x10^{-6}$	0.21	441	0.1,0.3	$1.0x10^{-09}$	0.5	157	0.3,0.6	ENST00000319836; ENST00000395769
UGT1A8	Total bilirubin	$4.8x10^{-2}$	-0.18	99	-0.4,0	$1.4x10^{-09}$	0.4	227	0.3,0.5	ENST00000373450
CREB3L3	Triglycerides	$2.2x10^{-3}$	0.06	1880	0.02,0.1	$1.7x10^{-9}$	0.4	223	0.3,0.5	ENST00000078445; ENST00000595923
PTCH1	Standing height	$5.2x10^{-4}$	-0.16	167	-0.3,-0.1	$2.7x10^{-9}$	-0.4	87	-0.5,-0.3	ENST00000468211

Table 5. Transcript-specific results with significant quantitative traits associations, and 95% CI of effect size not-overlapping with the gene-based burden 95% CI. For loci with multiple significant results, or multiple highly correlated phenotypes, the result with the lowest P value is shown. Multiple transcripts are listed when the variant sets are identical.

4.6. *Transcript-specific variant sets limit pLOF variants in low expression exonic regions*

One method by which the transcript-aware variant sets improve burden testing is by excluding variants within weakly expressed exonic region. An example of this improvement can be shown with LDL cholesterol and the low-density lipoprotein receptor (*LDLR*). We evaluated seven distinct transcript-isoforms variant sets for their association with apolipoprotein B, the main protein found in LDL. All seven tested *LDLR* transcript sets were more statistically significant and had larger effect sizes as compared to the gene-based inclusive method (Table 6).

The best performing *LDLR* transcript, ENST00000252444, compared to the worst performing *LDLR* transcript, ENST00000557933, and the gene-based model, lacks pLOF variants primarily in two critical regions: the first exon and part of the penultimate exon, highlighted in pink (Figure 5).

Figure 5. Two *LDLR* transcript models and the inclusive, gene-based model overlayed with pext = 0 regions in pink. No pLOF variants appear in the terminal exons of all three models.

In both regions, pext, or the proportion expressed across transcripts[22], is equal to 0, indicating that these regions have extremely low expression across all isoforms. All seven tested *LDLR* transcript-aware variant sets excluded some variants in the pext = 0 regions, and subsequently, resulted in an improved apolipoprotein B association compared to the gene-based method.

Gene	Phenotype	P value	Effect size	N carriers	95% CI	Transcripts	Median expression in all tissues (TPM)
LDLR	Apolipoprotein B	$5.8x10^{-20}$	1.0	78	0.8,1.2	ENST00000252444	7.8
LDLR	Apolipoprotein B	$6.7x10^{-18}$	0.9	84	0.7,1.1	ENST00000558518	0.5
LDLR	Apolipoprotein B	$3.2x10^{-16}$	1.0	67	0.7,1.2	ENST00000455727	0
LDLR	Apolipoprotein B	$1.0x10^{-14}$	0.9	65	0.7,1.2	ENST00000545707	0
LDLR	Apolipoprotein B	$4.4x10^{-13}$	0.9	56	0.7,1.2	ENST00000535915	0
LDLR	Apolipoprotein B	$1.6x10^{-12}$	0.6	118	0.5,0.8	ENST00000558013	0
LDLR	Apolipoprotein B	$1.7x10^{-9}$	0.5	124	0.4,0.7	ENST00000557933	0.1
LDLR	Apolipoprotein B	$3.8x10^{-4}$	0.2	287	0.1,0.3	Inclusive model	

Table 6. Comparison of *LDLR* transcript-based models and the inclusive, gene-based model on apolipoprotein B levels

4.7. *Transcript-specific variant sets exclude misannotated pLOF variants*

Additionally, the transcript-specific variant sets can improve association testing through the exclusion of misannotated variants. For example, polycysatin-1 (*PKD1)* is a well-characterized protein for its function in causing 85% of autosomal dominant polycystic kidney disease cases[23]. When damaged, the kidneys are unable to clear waste products like urea and creatinine which instead end up in high concentrations in the blood. Elevated serum urate is documented in rare-variant burden testing of *PKD1* pLOF variants[24]. Our results show an improved association of *PKD1* and urate using the transcript-based approach. When comparing the most significantly associated transcript variant set, ENST00000423118, and the gene-based burden, 12 variants are excluded. The most frequent among these is rs758337073, a *PKD1* variant labeled as "likely benign" by ClinVar[25]. Rs758337073 is a "stop gained" pLOF in ENST00000488185 and is subsequently designated as a pLOF in the gene-based method. However, rs758337073 is not considered a pLOF in 23/24 *PKD1* transcripts. ENST00000488185 has low overall expression, and zero expression in kidney cortex or medulla as shown by GTEx, indicating that this is likely a misannotated pLOF, and its inclusion in the gene-based method adds noise and dampens the *PKD1*-urate burden association (Figure 6).

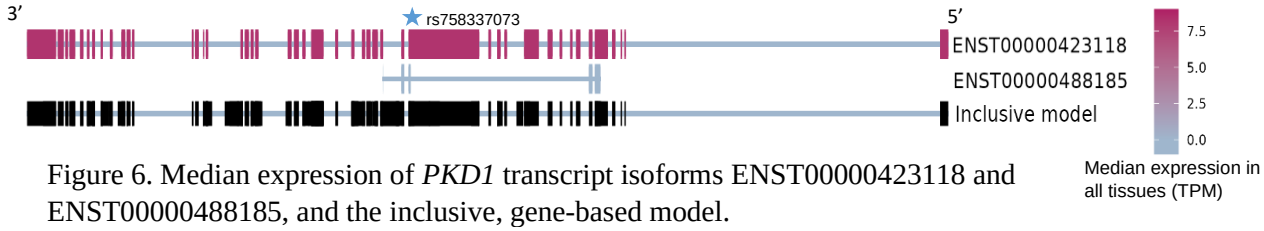

Figure 6. Median expression of *PKD1* transcript isoforms ENST00000423118 and ENST00000488185, and the inclusive, gene-based model.

5. Discussion

The drug discovery process is long, costly, and rarely ends in approval. Human genetic evidence provides an opportunity for novel target identification and validation for existing programs. Both

common and rare variant genetic analyses have been shown to improv the chances of a successful clinical trial and form the basis of rational drug discovery and development[26].

Our analysis highlights the importance of incorporating transcript-aware analyses into RVAS. We find that a transcript-aware approach broadly leads to lower circulating levels of *cis*-proteins as compared to the gene-based method. Since we expect pLOFs to lead to nonsense-mediated decay, and a reduction of functional RNA and protein products, this indicates that the included variants are more likely to be functioning as true LOFs. This is also evident in quantitative-trait testing, where we observe increased absolute value of effect sizes for the isoform-specific variant sets. The transcript-level approach also identifies novel isoform-trait associations, and in rare cases, identifies associations with an opposite direction of effect as compared to the gene-based method, as is the case with *GPNMB*, *HMOX2*, and *BST1* and their proteins encoded in *cis*. These data indicate the potential for a transcript-aware approach to elucidate new genetically validated drug targets, some of which may be isoform-specific.

Previously published literature has highlighted the importance of considering transcript data in RVAS. Cummings et al. described variants overlapping low confidence transcripts as a main contributor to the false annotation of pLOF variants. To counteract this, the authors developed the "proportion expressed across transcripts" (pext) score which quantifies the expression of transcript isoforms and exons. When testing pLOF variants in low pext-scored regions, the authors reported effect sizes comparable to the inclusion of synonymous variants. However, testing pLOF variants in high pext-scored regions resulted in substantially larger effect sizes[22]. This is consistent with our results showing that the transcript-level burden leads to larger effect sizes, in some cases, like for *LDLR*, by excluding variants in low expression exonic regions.

Our approach is limited in several ways. By only using transcript isoforms detected in at least one of 53 GTEx tissues, we exclude transcripts that may be expressed in other tissue and cell types. For example, several quantitative ocular phenotypes were tested, but we did not utilize data on ocular transcript isoform expression. Additionally, our analysis was only conducted on European-ancestry individuals due to limited sample size of other ancestral groups; RVAS in other populations may yield additional associations.

One drawback to the transcript-aware approach is the reduction in sample size, as all isoform-aware variant sets are smaller than their gene-based counterparts. Additionally, a given gene can have multiple alternatively spliced, biologically relevant isoforms, where a pLOF variant in any number of those isoforms may lead to the same deleterious effect on a phenotype. In that case, testing a single transcript would not be a sufficient representation, and instead it would be better to use a more inclusive multi-transcript or gene-based approach.

It is possible to test all transcript-variant sets alongside the gene-based method, as we have done here. However, this leads to an exceptionally stringent P value threshold and many highly related experiments. We suggest a curated implementation of the transcript-approach by testing only specific transcripts chosen *a priori*, for example, only canonical transcripts, MANE-select transcripts[27] which intend to choose the most biologically relevant, representative isoform for each gene, or highly expressed transcript isoforms in relevant tissue types.

References

1. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**, 1413–1415 (2008).
2. King, E. A., Wade Davis, J. & Degner, J. F. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS Genet* **15**, (2019).
3. Razuvayevskaya, O., Lopez, I., Dunham, I. & Ochoa, D. Why Clinical Trials Stop: The Role of Genetics. doi:10.1101/2023.02.07.23285407.
4. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* vol. 461 747–753 Preprint at https://doi.org/10.1038/nature08494 (2009).
5. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: Study designs and statistical tests. *American Journal of Human Genetics* vol. 95 5–23 Preprint at https://doi.org/10.1016/j.ajhg.2014.06.009 (2014).
6. Wainschtein, P. *et al.* Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nat Genet* **54**, 263–273 (2022).
7. Li, B. & Leal, S. M. Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. *Am J Hum Genet* **83**, 311–321 (2008).
8. Madsen, B. E. & Browning, S. R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* **5**, (2009).
9. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
10. 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**, (2015).
11. Bycroft, C. *et al.* Genome-wide genetic data on ~500,000 UK Biobank participants. doi:10.1101/166298.
12. Szustakowski, J. D. *et al.* Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. *Nature Genetics* vol. 53 942–948 Preprint at https://doi.org/10.1038/s41588-021-00885-0 (2021).
13. Sun, B. B. *et al.* Genetic regulation of the human plasma proteome in 54,306 UK Biobank participants 2 3. *Population Analytics of Janssen Data Sciences* **20**,.
14. Van Hout, C. V. *et al.* Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586**, 749–756 (2020).
15. Pan-UKB team. Pan UKB. https://pan.ukbb.broadinstitute.org (2020).
16. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, (2016).

17. Backman, J. D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).

18. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559–575 (2007).

19. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet* **53**, 1097–1103 (2021).

20. Cheng-Hathaway, P. J. *et al.* The Trem2 R47H variant confers loss-of-function-like phenotypes in Alzheimer's disease. *Mol Neurodegener* **13**, (2018).

21. Moutinho, M. *et al.* TREM2 splice isoforms generate soluble TREM2 species that disrupt long-term potentiation. *Genome Med* **15**, (2023).

22. Cummings, B. B. *et al.* Transcript expression-aware annotation improves rare variant interpretation. *Nature* **581**, 452–458 (2020).

23. Peters, D. J. M. & Sandkuijl, L. A. Genetic Heterogeneity of Polycystic Kidney Disease in Europe1. in 128–139 doi:10.1159/000421651.

24. Karczewski, K. J. *et al.* Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. *Cell Genomics* **2**, 100168 (2022).

25. Landrum, M. J. *et al.* ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* **46**, D1062–D1067 (2018).

26. Minikel, E. V., Painter, J. L., Dong, C. C. & Nelson, M. R. Refining the impact of genetic evidence on clinical success. doi:10.1101/2023.06.23.23291765.

27. Morales, J. *et al.* A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* (2022) doi:10.1038/s41586-022-04558-8.