# Aberrant expression prediction across human tissues

Florian R. Hölzlwimmer[1], Jonas Lindner[1], Nils Wagner[1,2], Francesco Paolo Casale[1,4,5], Vicente A. Yépez[1], Julien Gagneur[1,2,3,6]

[1] School of Computation, Information and Technology, Technical University of Munich, Garching, Germany
[2] Helmholtz Association - Munich School for Data Science (MUDS), Munich, Germany
[3] Institute of Human Genetics, School of Medicine, Technical University of Munich, Munich, Germany
[4] Institute of AI for Health, Helmholtz Munich, Neuherberg, Germany
[5] Helmholtz Pioneer Campus, Helmholtz Munich, Neuherberg, Germany
[6] Institute of Computational Biology, Helmholtz Center Munich, Neuherberg, Germany

## Abstract

Despite the frequent implication of aberrant gene expression in diseases, algorithms predicting aberrantly expressed genes of an individual are lacking. To address this need, we compiled an aberrant expression prediction benchmark covering 8.2 million rare variants from 633 individuals across 48 tissues. While not geared toward aberrant expression, the deleteriousness score CADD and the loss-of-function predictor LOFTEE showed mild predictive ability (1-1.5% average precision). Leveraging these and further variant annotations, we next trained AbExp, a model that yielded 10% average precision by combining in a tissue-specific fashion expression variability with variant effects on isoforms and on aberrant splicing. Integrating expression measurements from clinically accessible tissues led to another two-fold improvement. Furthermore, we show on UK Biobank blood traits that performing rare variant association testing using the continuous and tissue-specific AbExp variant scores instead of LOFTEE variant burden increases gene discovery sensitivity and enables improved phenotype predictions.

## Main

Aberrant gene expression, gene expression levels outside the physiological range, is a frequent cause of diseases. Aberrant underexpression of tumor suppressor genes and aberrant overexpression of oncogenes are hallmarks of oncogenesis[1,2]. Moreover, aberrant gene expression is a frequent cause of rare inheritable disorders[3–8] and contributes to risks for

common disease-associated traits[9].

Statistical methods to call expression outliers from RNA-seq data[10–13] applied to large cohorts have enabled investigating the genetic basis of aberrant expression. Rare variants have been found to be associated with expression outliers[14]. Specifically, rare variants including rare structural variants and rare variants likely triggering nonsense mediated decay such as premature stop codons, frameshift, and splice-disrupting variants are enriched among underexpression outliers[6,14,15]. Moreover, rare structural variants, notably gene duplication, were found enriched among overexpression outliers[14,15].

Building on these findings, algorithms have been developed to prioritize which variants may be the genetic cause of an expression outlier identified in an RNA-seq sample given the matched genome[14,15]. However, there is no algorithm predicting aberrant expression using only DNA sequence as input. A sequence-based predictor that can predict aberrant expression in multiple tissues and generalize to unseen variants could improve our ability to identify high impact rare variants in large genomic cohorts which in turn would aid in the identification of disease-associated genes and pinpointing disease-causal variants.

To address this unmet need, we here establish a benchmark for genotype-based prediction of aberrant gene expression across human tissues (Fig. 1). While underexpression outliers can be assumed to have strongly impaired or entirely lost function, the functional consequence of an overexpression outlier is less clear, as it could as well result in a gain of function. We therefore focussed in this study on underexpression outliers and developed AbExp, a machine learning model predicting aberrant underexpression in multiple tissues using DNA sequence as input. By integrating various variant annotations with tissue-specific isoform proportions and expression variability, AbExp significantly outperforms other loss-of-function annotators that were not explicitly developed for aberrant underexpression prediction. We demonstrate how AbExp can be used in rare variant gene association testing as well as phenotype prediction on the UK Biobank dataset using 40 blood traits. Finally, we show that when gene expression measurements from clinically accessible tissues are available, AbExp scores can be integrated to improve prediction of aberrant expression in non-accessible tissues.

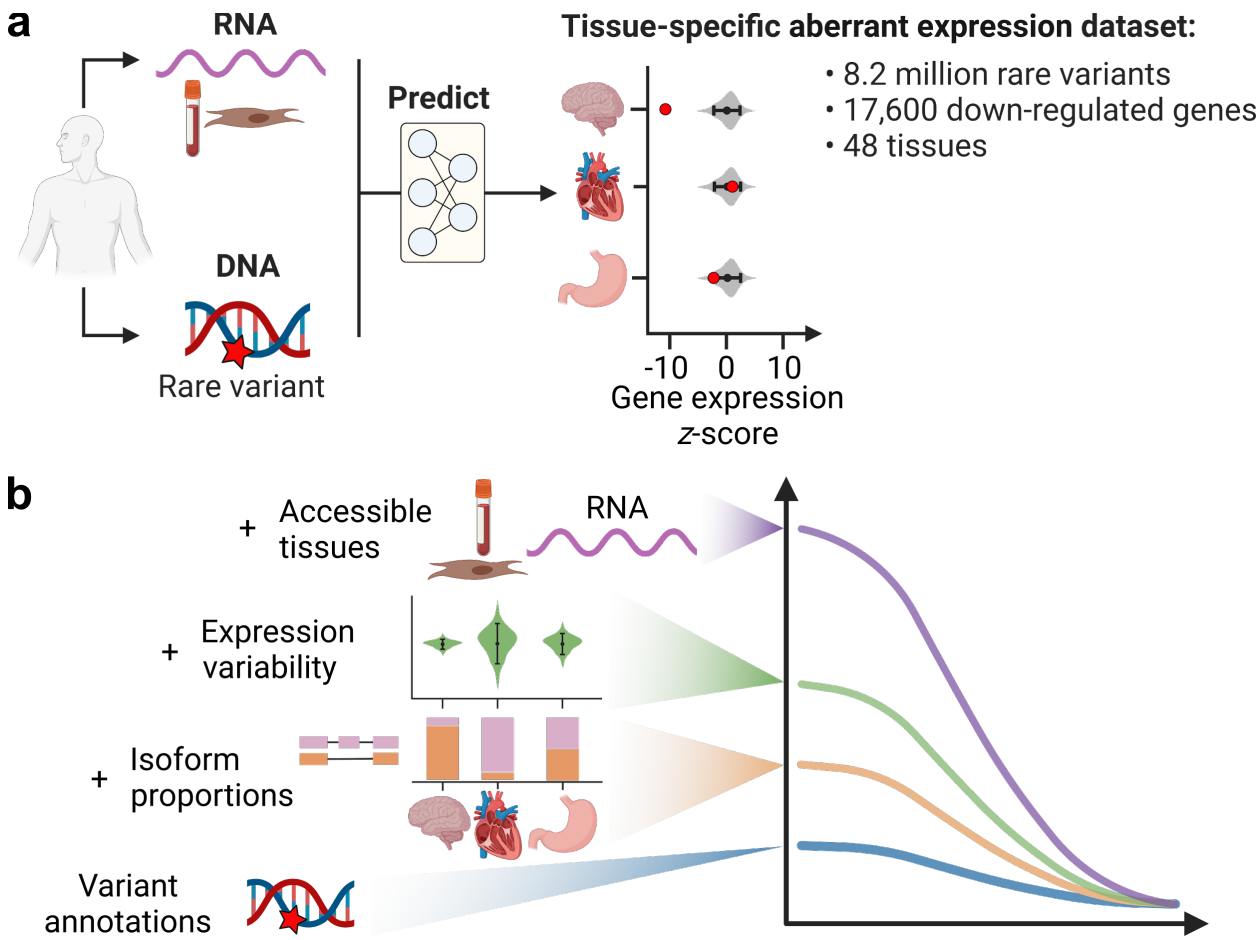



**Figure 1**: **Study design. (a)** We aimed to predict whether protein-coding genes are aberrantly underexpressed across 48 human tissues based on DNA and, optionally, RNA-seq data of clinically accessible tissues. Therefore, we created a benchmark for aberrant underexpression prediction by processing 11,096 RNA-seq samples from 633 individuals across 48 tissues from GTEx. **(b)** Assessing various variant and tissue annotations, we found that predictions could be significantly improved by weighting variant effects with tissue-specific isoform proportions and incorporating the expression variability of a gene. Further integration of expression measurements from clinically accessible tissues led to another two-fold improvement.

# Results

## A benchmark dataset of underexpression outliers across 48 human tissues

We first set out to predict which protein coding genes are aberrantly underexpressed in which tissues given an individual's genome. To this end, we created a benchmark dataset using the aberrant expression caller OUTRIDER[11] on 11,096 RNA-seq samples with paired whole-genome sequencing data of the Genotype-Tissue Expression dataset[16] (GTEx v8), spanning 48


3

tissues and 633 individuals. For the whole-genome sequencing data, we selected version 7 of the GTEx data since, unlike for the version 8 release, structural variant calls including transcript ablation were available. Although the choice against the version 8 genome release reduced the number of underexpression outliers by 20%, it allowed us to consider structural variants, which are important determinants of aberrant expression[15]. In each tissue, we restricted the analysis to all protein-coding genes with an average read-pair count of at least 450, an estimated minimal coverage required to detect 50%-reduction outliers[6]. Overall, 18,152 protein-coding genes out of 18,563 (97.8%) showed sufficient average coverage in at least one tissue. We defined a gene to be aberrantly underexpressed if OUTRIDER reported a false discovery rate (FDR) lower than 0.2 and the expression was lower than expected (Methods). We further removed samples with more than 50 outliers since such a high number of outliers could reflect poor sample quality, an unreliable fit of OUTRIDER, or trans-regulatory effects affecting a large number of genes in the sample. Across all analyzed 11,096 RNA-seq samples, the resulting benchmark dataset consisted of 17,637 underexpression outliers occurring in 5,428 unique genes and amounting to 1.6 outliers per sample on average. A detailed overview of how many samples, genes, etc. remained after each filtering step can be seen in suppl. table T1.

## Integrating rare variant annotations to predict underexpression outlier across tissues

We considered predicting aberrant expression for all 48 tissues for any protein-coding gene from rare variants, here defined as variants occurring in at most two GTEx individuals and with a gnomAD minor allele frequency of less than 0.1%. We reasoned that above 0.1% minor allele frequency, variants are unlikely to cause an expression outlier since the average frequency across samples of underexpression outliers among genes with sufficient RNA-seq coverage is about 0.01%. Moreover, we focussed on cis-regulatory variants by considering variants located within the gene and up to 5,000 bp around the gene to fully cover promoter and transcription termination regions.

We first investigated whether existing variant annotation tools, that were not developed for predicting aberrant underexpression, showed informative signals. Among the variant consequences annotated by Ensembl VEP we found frameshifts, variants affecting start and stop codons, splicing variants to be strongly enriched among underexpression outliers, consistent with previous reports[6,14,15] (Fig. 2a). Variants including frameshifts, splice-variants,

104 and stop-gains, which introduce premature stop codons are known to be strongly enriched
105 within gene underexpression outliers as these often trigger nonsense-mediated decay (NMD) of
106 the transcript[14]. LOFTEE is a tool that predicts a high-confidence subset of loss-of-function
107 associated variants, notably variants likely to trigger NMD, by implementing filters such as
108 removing stop-gained and frameshift variants that are within 50 bp of the end of the transcript as
109 these usually escape NMD, or variants that affect splicing only in UTRs[17]. In GTEx more than
110 17% of the aberrantly underexpressed genes had a LOFTEE variant, compared to non-outliers
111 that had a LOFTEE variant in less than 1% of the cases (Fig. 2a).

112 Moreover, we hypothesized that the deleteriousness score CADD[18] could also be predictive of
113 aberrant gene expression. The advantage of CADD, which was trained to distinguish between
114 simulated de novo variants and variants that have arisen and become fixed in human
115 populations, is that it provides a score for any variant. We found that CADD scores of rare GTEx
116 variants in underexpression outlier genes were in the median ~13 times higher than in non-
117 outliers (Fig. 2b).

118 Despite these enrichments, LOFTEE and CADD by themselves showed limited predictive value.
119 Using the sole LOFTEE-positive variants recalled 16.7% of the underexpression outliers at a
120 precision of 9.7% (Fig. 2c). While CADD scores showed some predictive value, at no single
121 score cutoff the same precision nor the same recall as LOFTEE could be reached (Fig. 2c).
122 Next, we trained a non-linear model that integrated all the above-mentioned features to
123 quantitatively predict the OUTRIDER *z*-score (Methods). Predicting the underlying quantitative
124 z-scores turned out to lead to better models than predicting the binary classes of outliers and
125 non-outliers. Ranking based on the predicted *z*-scores uniformly outperformed ranking based on
126 CADD scores on held-out data (Fig. 2c). Moreover, the integrative model reached the same
127 precision at the same recall as filtering for LOFTEE variants with the added value of providing a
128 continuous score, thereby allowing for applying more stringent cutoffs to yield a higher precision
129 (up to 13%, Fig. 2c). Lastly, the advantage of the integrative model over CADD and LOFTEE
130 was observed in aggregate (Fig. 2c) as well as across individual tissues according to the
131 average precision (measured by the area under the precision-recall curve here and elsewhere,
132 AUPRC, Fig. 2d).

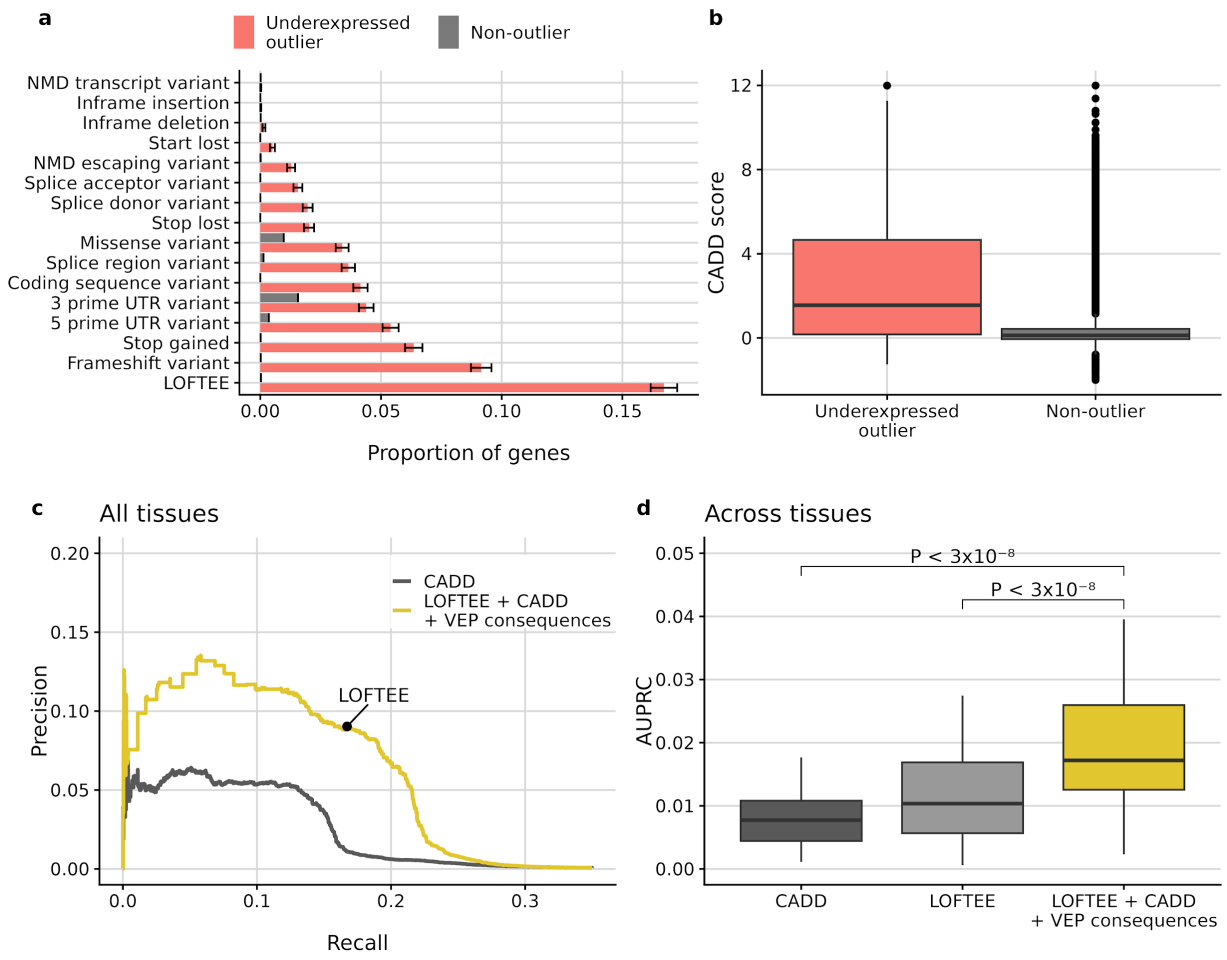


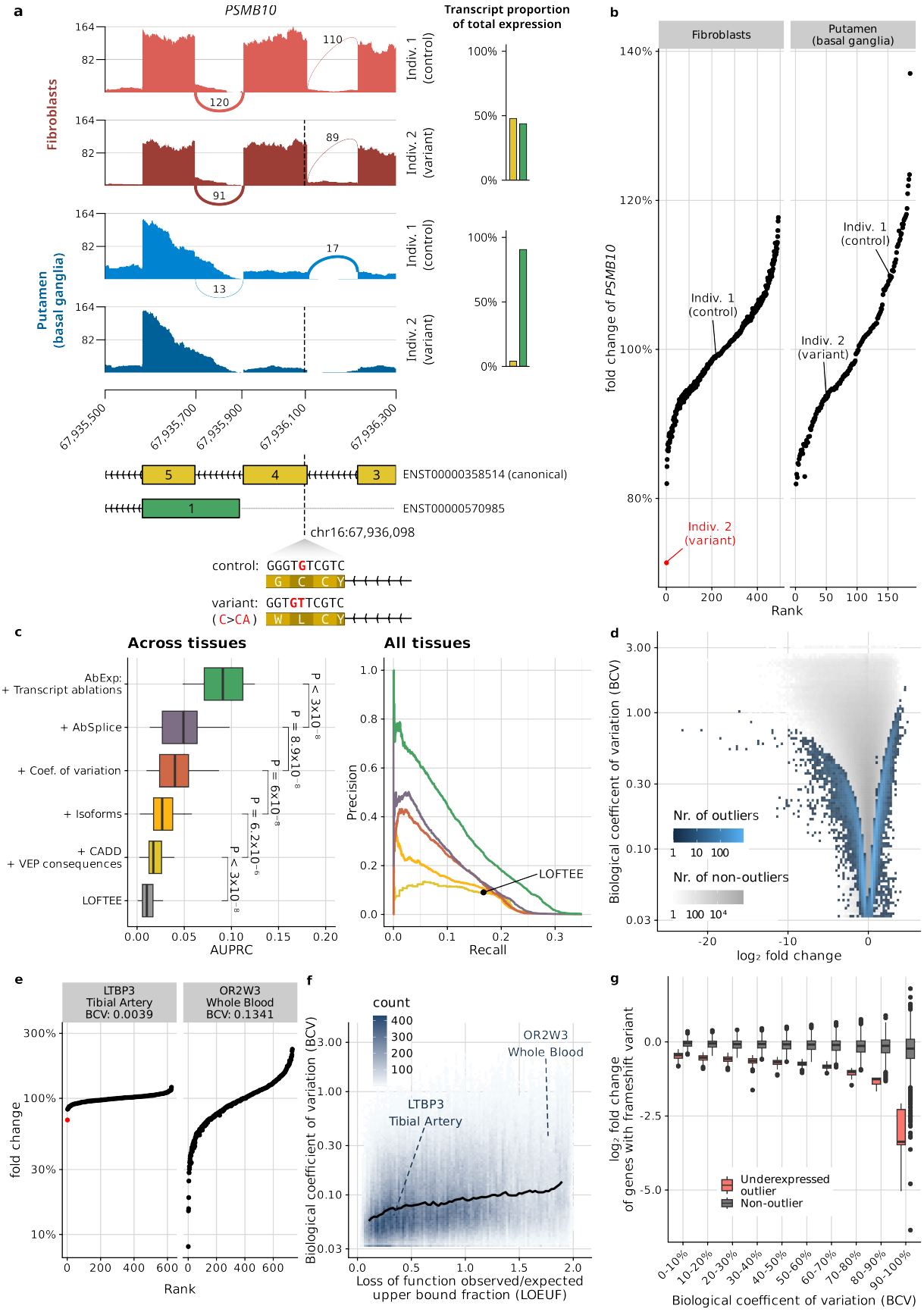

**Figure 2**: **Integrating rare variant annotations to predict underexpression outliers across tissues. (a)** Proportion of underexpressed outliers (red), and non-outliers (gray) with a rare variant of a given annotation (rows, Methods). Error bars mark 95% binomial confidence intervals. Notably, rare variants likely to trigger nonsense-mediated decay are strongly represented in underexpressed genes. **(b)** Distribution of CADD scores in different outlier classes. **(c)** Precision-recall curve for all tissues combined. LOFTEE shows up as a single point because it is a binary filter. **(d)** Distribution of average precision (AUPRC) across 26 GTEx tissue types (Methods). *P*-values were obtained using a paired Wilcoxon test. A non-linear model based on LOFTEE, CADD and VEP annotations significantly outperforms existing methods. For all boxplots: Center line, median; box limits, first and third quartiles; whiskers span all data within 1.5 interquartile ranges of the lower and upper quartiles.

# Accounting for tissue-specific isoform expression improves predictions

By construction, the predictions of this first model were independent of the tissue as neither the variant annotations nor the considered transcript isoforms were tissue-specific. However, since the transcript isoforms of a gene are often expressed at different proportions across tissues,


6

147 variants can have tissue-dependent effects[19]. For example, ENST00000358514, the canonical
148 transcript of *PSMB10* according to the MANE annotation[20], was estimated to generate only
149 about 4% of *PSMB10* total gene expression in putamen (Methods). The vast majority (91%) of
150 *PSMB10* gene expression in putamen was attributed to another transcript, ENST00000570985.
151 Conversely, in fibroblasts, the canonical transcript contributed to nearly 48% of the total gene
152 expression. Exon 4 is not included in the transcript ENST00000570985 but is included in the
153 canonical transcript ENST00000358514, explaining why a frameshift variant in exon 4 was
154 associated with a high impact on gene expression in cultured fibroblasts but showed a limited
155 effect in putamen (Fig. 3a,b).
156 Generally, we found that only 30% of the canonical transcripts contributed more than 90% of
157 their gene's total expression and that as much as 18% of the canonical transcripts contributed to
158 less than 10% of their gene's total expression (using MANE-select as the canonical transcript,
159 suppl. Fig. S1). Therefore, relevant information is lost when considering the variant
160 consequence assigned to a single isoform, even if it is annotated as the canonical one. To
161 address this issue, we calculated the isoform composition in every tissue and weighted the VEP
162 consequences and LOFTEE classification of each variant by the proportion of affected
163 transcripts per gene and tissue (Methods). Training the model using these tissue-specific
164 weighted annotations increased the average precision by 55% to reach 2.7% in median across
165 tissues (Fig. 3c).


a
PSMB10
Transcript proportion of total expression
Fibroblasts
Putamen (basal ganglia)
Indiv. 1 (control)
Indiv. 2 (variant)
110
120
89
91
17
13
67,935,500
67,935,700
67,935,900
67,936,100
67,936,300
ENST00000358514 (canonical)
ENST00000570985
chr16:67,936,098
control: GGGTGTCGTC
variant: GGTGTTCGTC
(C>CA)
b
Fibroblasts
Putamen (basal ganglia)
fold change of PSMB10
Rank
c
Across tissues
All tissues
AbExp: + Transcript ablations
+ AbSplice
+ Coef. of variation
+ Isoforms
+ CADD
+ VEP consequences
LOFTEE
AUPRC
Precision
Recall
d
Biological coefficient of variation (BCV)
Nr. of outliers
Nr. of non-outliers
log2 fold change
e
LTBP3 Tibial Artery BCV: 0.0039
OR2W3 Whole Blood BCV: 0.1341
fold change
Rank
f
count
OR2W3 Whole Blood
LTBP3 Tibial Artery
Biological coefficient of variation (BCV)
Loss of function observed/expected upper bound fraction (LOEUF)
g
log2 fold change of genes with frameshift variant
Underexpressed outlier
Non-outlier
Biological coefficient of variation (BCV)


8

**Figure 3**: **Tissue-specific features improve the model prediction. (a)** Sashimi plot of *PSMB10* for two individuals, one carrying no rare variant in this region (control, upper tracks), and one carrying a heterozygous frameshift variant (dashed line and lower tracks), in cultured fibroblasts (top) and putamen (bottom). The frameshift variant is located on exon 4 which is included on the canonical transcript (ENST00000358514) but not on transcript ENST00000570985. On the right, the barplots show the transcript expression proportions on each tissue on average across GTEx. **(b)** Fold change of gene expression against normalized gene expression rank for *PSMB10* in fibroblasts and putamen (basal ganglia) brain tissues. *PSMB10* is an expression outlier (red) in individual 2 in fibroblast but not in putamen, consistent with the rare variant triggering nonsense mediated decay and leading to a strong gene expression reduction in the tissue for which the exon 4-containing transcript is the major isoform. **(c)** Left: Distribution of average precision (AUPRC) across 26 GTEx tissue types. *P*-values were obtained using the paired Wilcoxon test. AbExp combines various variant and tissue annotations to predict aberrant gene expression and outperforms LOFTEE by about sevenfold. Right: Precision-recall curve for all tissues combined. **(d)** Biological coefficient of variation (BCV) against expression fold change across all genes and tissues. Highly variable genes require a larger fold change to be called an outlier. **(e)** Fold change of gene expression against normalized gene expression rank for *LTBP3* in tibial artery, an autosomal recessive gene whose defect can lead to dental anomalies and short stature [21], and for *OR2W3* in blood, an olfactory gene whose defect should not impair the viability of an individual [22]. Expression outliers are highlighted in red. *LTBP3* is tightly regulated with a fold change range of ± 20% among non-outliers. The individual marked in red carries a heterozygous frameshift variant that associates with 30% reduction and which is detected as an outlier. In contrast, *OR2W3* shows very large variations where individuals with 30% reductions are not outliers. **(f)** BCV versus loss of function observed/expected upper bound fraction [17] (LOEUF) across all genes and tissues. Genes with a high LOEUF are more tolerant to loss of function. The black line shows a running median between LOEUF and BCV. The autosomal recessive gene *LTBP3* has a low LOEUF, denoting a low loss of function tolerance. In contrast, the olfactory gene OR2W3 has a high LOEUF, denoting a large loss of function tolerance. **(g)** Distribution of gene expression fold changes among genes in different deciles of expression variability, given that the gene is affected by a frameshift variant. The higher the coefficient of variation, the larger the gene expression impact of the variants tends to be. For all boxplots: Center line, median; box limits, first and third quartiles; whiskers span all data within 1.5 interquartile ranges of the lower and upper quartiles.

## Incorporating the tissue-specific gene expression variability further improves predictions

Similar to other statistical models for RNA-seq data, OUTRIDER includes a measure of gene expression variability called the biological coefficient of variation[11,23]. In our benchmark dataset, the biological coefficient of variation captures the expression variability of genes per tissue across the GTEx population. We reasoned that the same expression fold changes could cause an expression outlier for a gene with low expression variability but not for a gene with high expression variability. Indeed, we observed that the minimal fold-change among expression outliers decreased with the biological coefficient variation (Fig. 3d). Therefore, a given relative reduction in gene expression can lead to aberrant expression in one gene or tissue but not


9

204 necessarily in another. For instance, a 30% reduction of the gene *LTBP3* in tibial artery
205 expression resulted in an outlier. In contrast, a 30% expression reduction in blood of *OR2W3*
206 would not lead to an outlier as *OR2W3* shows a very large gene expression variation in blood
207 ranging between 10% and 230% (Fig. 3e). It is not surprising that *OR2W3*, one of the over 800
208 human olfactory receptor genes[22] for which dysfunction is likely benign, exhibits more
209 expression variability than *LTBP3*, a gene whose dysfunction is associated with dental
210 anomalies and short stature[21]. Generally, we found that genes with lower expression variability
211 were more genetically constrained in the human population (i.e. harbored fewer loss-of-function
212 variants, Fig. 3f), in agreement with a previous report on primates[24].

213 Next, we aimed to improve our underexpression outlier predictor by adjusting for expression
214 variability. We first considered modeling expression fold-changes from variants and then
215 converting the predicted fold-changes into a *z*-score, under the assumption that variants affect
216 gene expression fold-changes independently of the expression variability. Using variants likely
217 triggering NMD to test this assumption, we noticed however that the same class of variants
218 associated with lower fold-changes among genes with lower expression variability (Fig. 3g,
219 suppl. Fig. S2), perhaps because genes with low expression variability are subject to regulatory
220 buffering mechanisms[25,26]. Therefore, we opted for a more general modeling approach in which
221 the biological coefficient of variation is provided as an input feature along with variant
222 annotations to a non-linear model predicting the *z*-score. This model increased the performance
223 by more than 50% to 4.0% average precision (median across tissues, Fig. 3c). Altogether, these
224 results show that gene expression variability must be taken into account when predicting
225 aberrantly expressed genes, and that predicting z-score rather than fold-change is more
226 relevant for variant interpretation.

## 227 Contribution of aberrant splicing prediction and transcript ablation

228 Aberrant splicing isoforms often contain a premature termination codon and as a consequence,
229 yet not always, are degraded by NMD[27]. Our model already included splice site information, as it
230 was based on LOFTEE, CADD, and tissue-specific isoform weighting of VEP annotations.
231 However, these features ignored splice sites that are not part of the genome annotations upon
232 which these tools are built. We have recently developed a model called AbSplice that predicts
233 aberrant splicing across tissues using a more comprehensive map of splice sites, including
234 unannotated weak splice sites, and their tissue-specific usage[28]. We found that AbSplice scores
235 were in median 10 times higher among underexpression outliers than among non-outliers

236 (suppl. Fig. S3). Integrating AbSplice scores significantly increased the performance to 4.9%
237 average precision in median across tissues (Fig. 3c).

238 We next considered transcript ablations, which could be inferred in GTEx from the structural
239 variant deletion calls. Overall, 33 out of the 43 GTEx individuals harboring transcript ablations
240 also showed an expression outlier in at least one tissue. Including transcript ablation variants
241 into the model resulted in a large gain in precision among the top-ranked predictions and
242 increased the average precision to 9.1% in median across tissues (Fig. 3c). In the following, we
243 refer to this model which integrates all features mentioned so far as AbExp. AbExp takes as
244 input a set of variants within 5,000 bp of any annotated transcript of a protein-coding gene and
245 returns a predicted z-score for each of the 48 tissues. For user convenience, we furthermore
246 suggest a high-confidence cutoff (AbExp < -3.4) corresponding to 50% precision and 8.5%
247 recall on our benchmark data, and a low-confidence cutoff (AbExp < -1.3) corresponding to 20%
248 precision and 19.1% recall.

## 249 AbExp replicates in independent datasets

250 We next assessed how AbExp performance replicated on two independent datasets. The first
251 dataset consisted of 299 individuals suspected to be affected by a mitochondrial disorder[6] with
252 whole-exome sequencing data paired with RNA-seq from fibroblasts. The second dataset
253 consisted of 290 whole-genome sequencing samples with RNA-seq from iPSC-derived motor
254 neurons from the AnswerALS research project[29]. Structural variant calls, and thus transcript
255 ablation calls, were not available on either dataset. Moreover, we observed that the recall for all
256 methods was twice as low on the ALS dataset than in GTEx and in the mitochondrial disorder
257 dataset (suppl. Fig. S4), perhaps because of poorer expression outlier calls, a stronger role of
258 epigenetic and trans-regulatory effects, or combinations thereof. Taking these differences into
259 account, our results on those two independent datasets were in agreement with the evaluation
260 on GTEx. We found that AbExp without transcript ablation annotation significantly outperformed
261 LOFTEE and CADD by two to three times larger average precision (suppl. Fig. S4). Here too,
262 AbExp without transcript ablation annotation allowed for slightly better precision at the same
263 recall than LOFTEE filtering, while offering a continuous score allowing for reaching much
264 higher precisions.

## AbExp improves rare variant association testing and phenotype prediction

The rise of large exome-sequencing and genome-sequencing cohorts empowers rare variant association testing (RVAT) which helps pinpoint causal genes for traits[30,31] and enables improved phenotype predictions, particularly among individuals showing extreme phenotypes[32]. RVAT consists of identifying genes within which the occurrence of likely high-impact variants associates with a trait[33,34]. To this end, accurate prediction of high-impact variants can be advantageous, suggesting that AbExp may have the potential to improve RVAT. To test this hypothesis, we considered 40 continuous blood traits including high-density lipoprotein cholesterol, glucose, and urate levels (suppl. table T2) from the UK Biobank 200k exome release[35]. We chose blood traits for this proof-of-concept investigation since they are well-studied and frequently measured to diagnose and monitor chronic disease conditions. Moreover, RVAT methods are typically better calibrated on continuous traits as opposed to binary traits[36,37].

To ease comparisons between variant annotations, we used linear regression as a common framework for rare variant association testing. As a realistic baseline, we considered RVAT based on LOFTEE variants, similar to the Genebass study[31]. For this baseline model, gene-trait association was tested by regressing the trait against the number of LOFTEE variants. The second model was designed to leverage that AbExp is both quantitative and tissue-specific. To this end, we considered the lowest AbExp *z*-score across all rare variants for each of the 48 tissues. The gene-trait association was tested by regressing the trait against the resulting 48 values. Two further models were considered by regressing against the minimum and against the median of the 48 values. To adjust for other relevant factors and effects due to common genetic variation, all four models included as covariates sex, age, the first 20 genetic principal components, a polygenic risk score predicting the trait, and common variants reported to be associated with the trait and located 250,000 bp around the gene (Methods). We fitted every model on two thirds of the dataset for gene-trait association discovery. Phenotype permutation analysis indicated that all models were calibrated (suppl. Fig. S5).

Using AbExp predictions for the 48 tissues, we identified in total 30% more gene-trait associations compared to the LOFTEE-based model (Fig. 4a), showing that AbExp can improve RVAT-based gene discovery. Notably, association testing using the tissue-specific predictions outperformed aggregated forms of the AbExp score in most cases by finding more gene-trait


12

297 associations (suppl. Fig. S6). In some instances, we could rationalize which tissues showed the
298 most significant associations. This was the case for the gene encoding Apolipoprotein B, a
299 major constituent of triglyceride-rich lipoproteins synthesized in the liver and whose predicted
300 aberrant underexpression in the liver was found to be negatively associated with blood
301 triglyceride levels (suppl. Table T3). However, the high correlation between the AbExp scores
302 per gene across tissues can make the estimation of each individual tissue-specific coefficient
303 unstable, and, therefore, their interpretation should be done with caution.

304 Having shown that AbExp can improve the gene-discovery sensitivity of RVAT, we next
305 assessed its utility in phenotype prediction. To this end, we used the remaining third of the
306 dataset which was not used for gene-trait association discovery. Specifically, we fitted gradient
307 boosted trees models predicting the traits given the AbExp scores on the one hand or the
308 number of LOFTEE variants on the other hand, of the genes discovered on the first two thirds of
309 the data. These models were controlled for sex, age, the first 20 genetic principal components,
310 and a polygenic risk score predicting the trait (Methods). The predictions rarely differed between
311 the common variant-based model and the model further including AbExp scores of rare variants,
312 as exemplified for the Alanine aminotransferase blood levels (Fig. 4b). For this trait, predictions
313 for less than 0.3% of all individuals differed between the two models by more than 1 standard
314 deviation of the population trait distribution (Fig. 4b). Remarkably, the trait values of those
315 differing individuals tended to deviate largely from the population average, suggesting that the
316 model integrating AbExp scores especially improves the predictions of individuals with extreme
317 phenotypes that common variants cannot explain (Fig. 4c).

318 On held-out data, the phenotype prediction model based on AbExp scores significantly
319 increased the amount of explained variation ($R^2$) over the model based on LOFTEE in 50% of
320 the traits and never significantly decreased $R^2$ (Fig. 4d). Considering the number of individuals
321 differing by more than 1 standard deviation to the common-variant based model, the AbExp
322 based model improved the prediction of 784 individuals across the 40 blood traits, while the
323 LOFTEE-based model only improved the prediction of 259 individuals (Fig. 4e). Moreover, the
324 advantage of using AbExp scores were similarly observed when predicting phenotypes with
325 regularized linear regression instead of gradient boosted trees (suppl. Fig. S7a,b), yet yielding
326 less accurate phenotype predictions (suppl. Fig. S7c).

327 Altogether, these results show that AbExp provides useful variant annotation for gene-trait

association discovery by rare variant association testing and for building improved genetic risk scores.

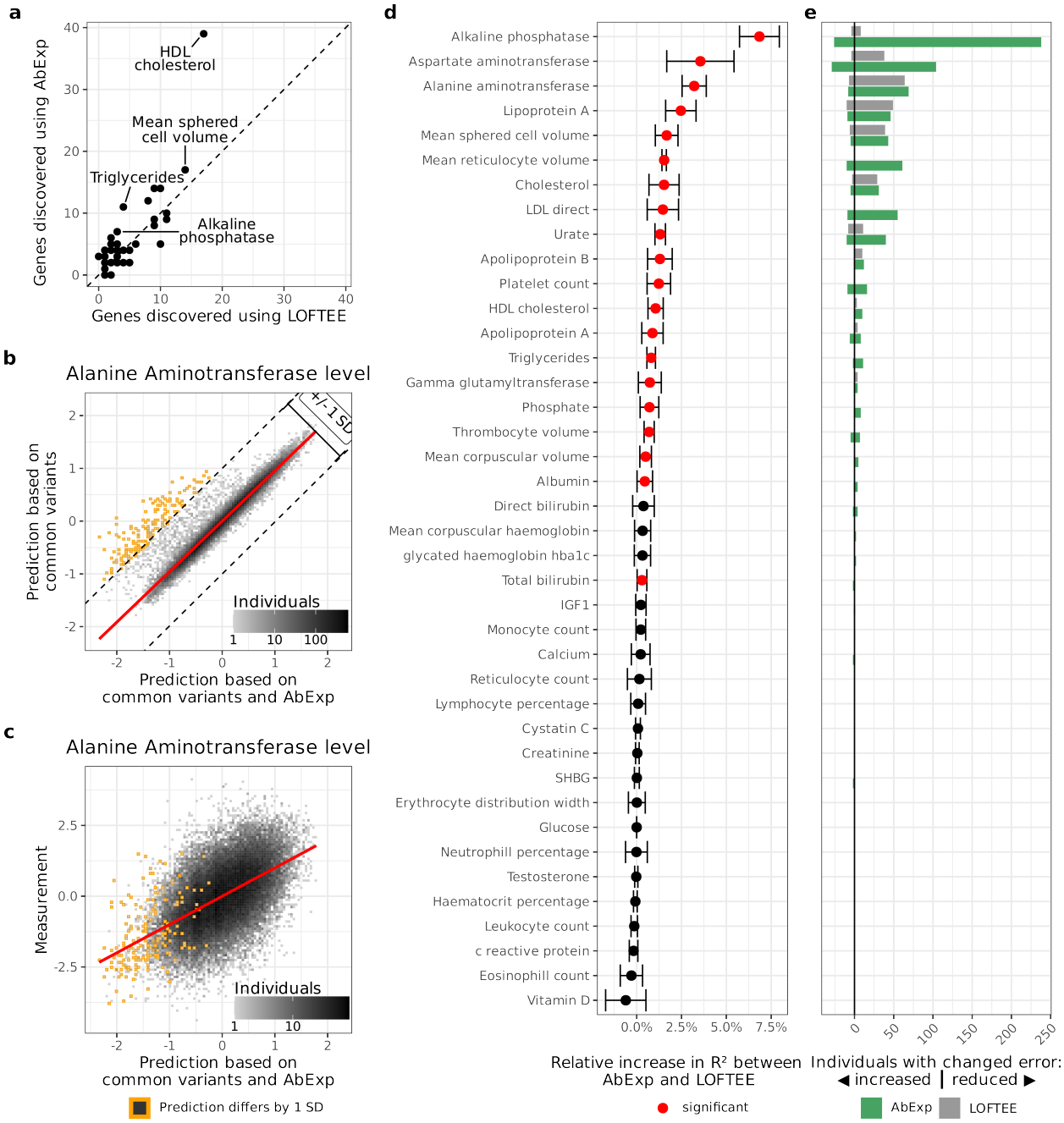



**Figure 4: AbExp improves rare variant association testing and phenotype prediction. (a)** Number of genes associating with different traits using a model based on LOFTEE or AbExp. **(b)** Alanine aminotransferase level predicted using a model solely based on common variants (y-axis) against predictions using a model based on common variants and AbExp scores (x-axis). Individuals whose predictions differed by more than 1 standard deviation of the population trait distribution are marked in orange. **(c)** Alanine aminotransferase measurements against predictions based on common variants and AbExp scores. Orange data points as in b). **(d)** Relative $R^2$ increase between AbExp-based and LOFTEE-based predictions across traits. Traits with a significant difference between both models are marked red (two-sided paired t-test, nominal $P < 0.05$). Error bars show the standard


14

deviation among 5 cross-validation folds. **(e)** Positive bars show the number of individuals with an error reduced by at least one standard deviation in the trait scale and therefore improved prediction, negative bars show the number of individuals with an error increased by at least one standard deviation in the trait scale and therefore worse prediction of the AbExp-based model (green) and the LOFTEE-based model (grey). All data presented in b-e) are computed on held-out folds of a 5-fold cross-validation within a third of the UKBB data not used for the gene discovery shown in a).

## Incorporating RNA-seq from clinically accessible tissues (CATs) boosts prediction performance

RNA sequencing is becoming increasingly popular for rare disease diagnostics as a complementary assay to genome or exome sequencing as it allows the direct measurement of aberrant gene regulation[4–6,38–40]. However, many rare disorders are suspected to originate from tissues that can only be very invasively sampled such as the brain or the heart. We and others [6,41] have shown that clinically accessible tissues (CATs), notably skin fibroblasts and to a lesser extent whole blood, share a substantial fraction of expressed genes with non-CATs and, therefore, are likely to capture aberrant expression occurring in non-CATs. The GTEx dataset – a dataset of post-mortem samples – offers a unique opportunity to test the validity of this assumption as it provides matched samples for CAT and non-CAT tissues. We found that the mere ranking of genes according to their OUTRIDER *z*-score in fibroblast RNA-seq samples led to an average precision in predicting underexpression outliers in non-CATs of 17.7% (median across tissues), significantly larger than the genome-based predictor AbExp (8.8%, $P = 4.2\text{x}10^{-7}$, Fig. 5a). Next, we developed a model taking as input AbExp, whether the gene is expressed in a CAT and, if so, its OUTRIDER z-score. Using RNA-seq from skin fibroblasts to predict aberrant underexpression in all other tissues, this model reached an average precision of 19.5% in median across tissues (Fig. 5a). Consistent with previous work based on shared expressed genes[6,41] and our work on aberrant splicing prediction [28], fibroblasts turned out be more informative than whole blood (median average precision 9.1% using RNA-seq only and 16.1% when integrating AbExp, Fig. 5b), owing to blood expressing less genes. Altogether, we have established a method integrating direct measurements of aberrant expression from RNA-seq data in a CAT along with genomic variant annotations to predict aberrant underexpression in other tissues. Doing so, we showed that integrating RNA-seq from fibroblasts yields a substantial improvement as it doubles the average precision over using genomic variants alone.


15

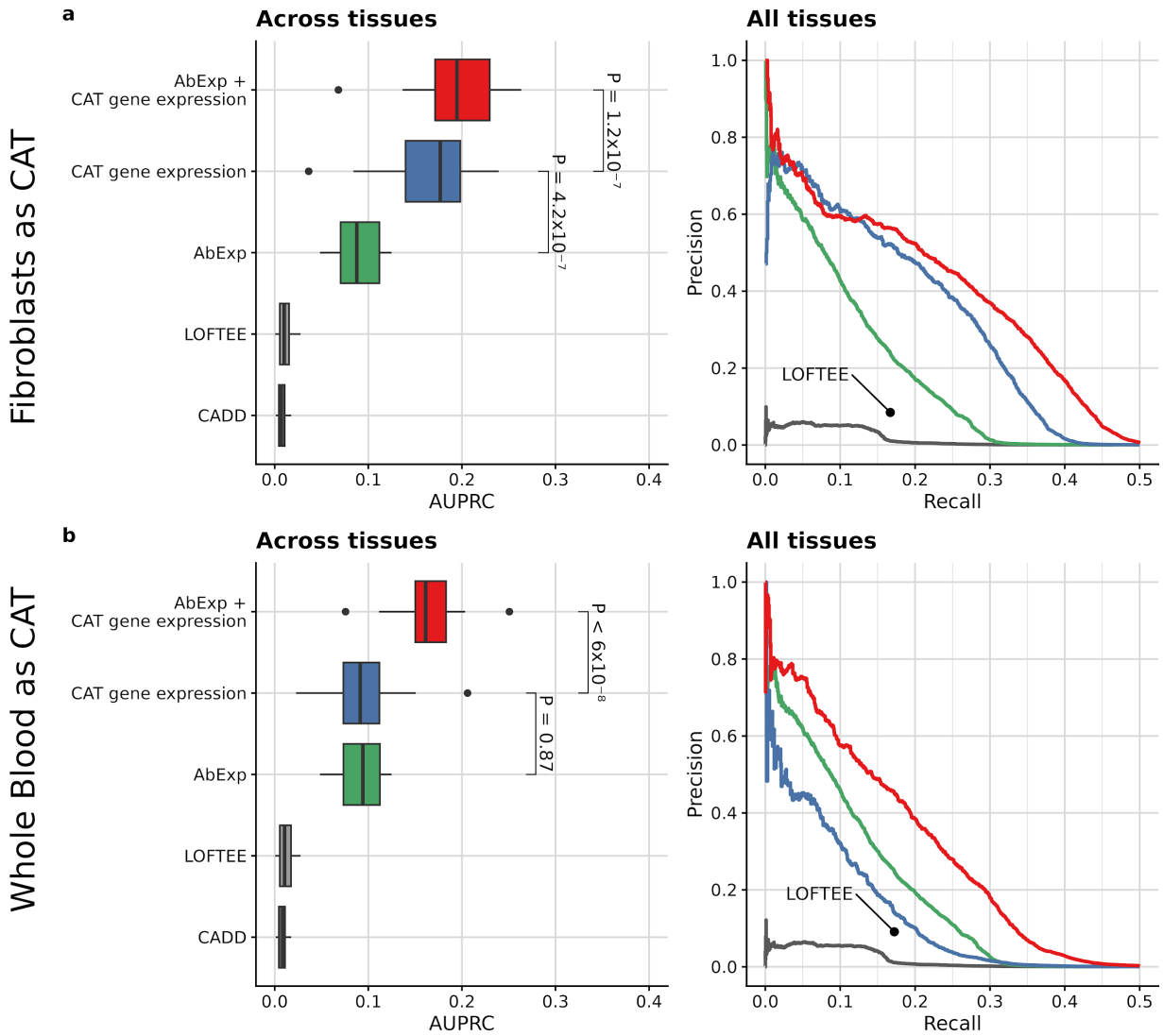



**Figure 5: Combining RNA-seq measurements from clinically accessible tissues with AbExp further improves the prediction performance. (a)** Left: Distribution per predictor (rows) of average precision (AUPRC) across 25 tissue types excluding skin tissues (Methods). Center line, median; box limits, first and third quartiles; whiskers span all data within 1.5 interquartile ranges of the lower and upper quartiles. *P*-values were obtained using the paired Wilcoxon test. The "Gene expression (CAT)" predictor ranks genes according to their OUTRIDER *z*-score in fibroblasts RNA-seq data. Right: Precision-recall curve aggregated across the same GTEx tissues as in the left panel. LOFTEE as a binary predictor is shown as a single point. **(b)** as in (a) using Whole blood as CAT and all other tissues as non-CAT.

# Discussion

Altogether, we established a benchmark dataset for aberrant gene underexpression prediction in 48 human tissues, addressing an unmet need in the area of high-impact variant effect prediction. We developed AbExp, a machine learning model predicting aberrant underexpression across tissues by integrating existing variant annotations with tissue-specific gene expression variability and transcript isoform composition. AbExp outperformed existing


16

variant annotation tools by up to 7-fold in average precision. Using UK Biobank dataset blood traits, we demonstrated that the continuous and tissue-specific AbExp scores provide added information over the state-of-the-art putative loss of function classifier LOFTEE for rare variant gene association testing as well as for phenotype prediction. Finally, we showed that AbExp scores can be combined with gene expression measurements from clinically accessible tissues to predict aberrant expression in other tissues yielding an increased prediction performance by 2-fold over AbExp.

Refining the predictor, while our primary objective, also shed light into the biology of underexpression outliers. We found that gene expression variability plays a dual role in this context. On the one hand, altering expression of a gene is more likely to result in an outlier if the expression of the gene varies little in the population than if it varies largely. On the other hand, each variant category was associated with milder fold-changes for genes with lower expression variability, indicating the involvement of regulatory mechanisms that confer robustness to genetic perturbations. We modeled the outcome of these two counteracting phenomena with a non-linear model trained from the data. Future biophysical investigations unraveling these buffering mechanisms could help improving the predictions and more generally improve variant interpretation. Our work also confirmed the importance of nonsense mediated decay, which underpins a substantial proportion of the outliers, and the need to take tissue-specific transcript isoform into account when interpreting splice-affecting and nonsense variants as pioneered by Cummings and colleagues[19].

This study has limitations. A basic assumption of AbExp is that an underexpression outlier is caused by a rare variant. However, it is possible that a rare combination of frequent genetic variants causes an expression outlier. Also, damage caused by one variant might be recovered by another variant, e.g. a second frameshift variant recovering the frame after a first frameshift variant. AbExp does not evaluate combinations of variants, which would require more complex modeling in particular by taking phasing into account. Moreover, AbExp covers the variants up to 5 kb away from transcript boundaries, missing middle-range and long-range enhancers. Future work could expand to such variants, for instance if sequence-based models of gene expression improved at this task[42]. Also, this work was focused on cis-acting regulation by considering only variants located within or near the genes. The effect of trans-acting gene regulation, which would need a very different modeling paradigm than investigating here in order to capture regulatory networks, remains to be addressed.


17

414 Predicting gene expression from sequence is a long-standing goal of computational biology that
415 is still far from completion. While existing sequence-based models of cis-regulation are trained
416 across the whole range of expression levels[43–45], we have proposed here to focus on extreme
417 expression variations. Extremes may not well be captured by models trained to globally predict
418 gene expression. Also, the biological mechanisms underpinning extreme expression variations
419 may differ from those governing moderate expression variations. However, the relevance of
420 extreme expression for clinical diagnostics and research is high. We hope that the benchmark
421 and algorithms we have developed will foster further research in this direction and aid in the
422 development and validation of methods predicting the impact of large-effect variants on the
423 human transcriptome.

# 424 Methods

## 425 Underexpression outlier benchmark dataset

### 426 GTEx dataset

427 We downloaded the GTEx RNA-seq read alignment files in the BAM format from dbGaP
428 (phs000424.v8.p2). We excluded tissues with less than 50 RNA-seq samples due to insufficient
429 statistical power[11]. This filter discarded the kidney cortex, leaving 48 tissues.
430 We obtained SNPs and small indels from the GTEx hg19 variant calls from the file
431 GTEx_Analysis_2016-01-
432 15_v7_WholeGenomeSeq_635Ind_PASS_AB02_GQ20_HETX_MISS15_PLINKQC.vcf.gz from
433 the dbGap entry phg000830.p1. Structural variants were obtained from Ferraro and
434 colleagues[15].

### 435 Expression outliers

436 Gene expression outlier analysis was performed following the aberrant expression module of
437 DROP v1.1.0[46] based on OUTRIDER[11]. To this end, we used as reference genome the GRCh37
438 primary assembly release 29 of the GENCODE project[47]. A fragment (reads pair) was assigned
439 to a gene if and only if both reads were entirely aligned within the gene, allowing for fragments
440 to be assigned to more than one gene. On each tissue separately, genes with an FPKM less
441 than 1 in 95% or more of the samples were considered to be not sufficiently expressed in the
442 tissue and filtered out, as previously described[11].

443 OUTRIDER is an expression outlier caller that uses an autoencoder to model RNA-seq
444 fragment count expectations with a negative binomial distribution. Specifically, OUTRIDER
445 models the probability of the observed fragment count $x_{s,g}$ for every gene *g* in a sample *s* as:

$$P(x_{s,g}|\mu_{s,g},\theta_{t(s),g}) = \mathrm{NB}(x_{s,g}|\mu_{s,g},\theta_{t(s),g}) \ (1)$$

446 where:

447 - $\mu_{s,g}$ is the expected fragment count
448 - $\theta_{t(s),g}$ is the dispersion parameter for the gene *g* in the tissue of sample *s t(s)*

449 OUTRIDER further outputs:

450 - the biological coefficient of variation: $\mathrm{BCV}_{t(s),g)} = \frac{1}{\sqrt{\theta_{t(s),g)}}}$
451 - the $\log_2$-transformed fold-change of the observed fragment count compared to the
452 expected fragment count: $\log_2 \mathrm{FC} = \log_2(x) - \log_2(\mu)$
453 - the nominal *p*-value
454 - the False Discovery Rate using the Benjamini-Yekutieli method[48]

455 The resulting table was subsetted to individuals with an available whole genome sequencing
456 and to protein-coding genes. Some of the data points could not be detected as outliers due to
457 lack of statistical power. To reduce the proportion of these insufficiently powered data points in
458 our benchmark, we discarded observations with an expected fragment count $\mu_{s,g}$ less than 450,
459 a minimal value that was empirically estimated to allow recovering half of the two-fold reduction
460 outliers transcriptome-wide upon a FDR cutoff of 5%[6]. We labeled as gene expression outliers
461 all observations with an FDR less than 20%. This relaxed FDR cutoff of 20% turned out to help
462 by leading to more robust evaluations and models. Lastly, RNA-seq samples that contained
463 more than 50 outliers were discarded because samples with numerous outliers may be samples
464 for which OUTRIDER could not adequately fit the data or for which gene expression is globally
465 affected, resulting in widespread expression aberrations throughout the genome that cannot be
466 predicted from local sequence variation.

## 467 OUTRIDER *z*-score computation

468 We quantile-mapped the OUTRIDER-fitted negative binomial distributions (Eq. 1) to the

standard normal distribution as follows:

$$z\text{-score} = \mathrm{CDF}^{-1}_{N(0,1)}(\mathrm{CDF}_{NB}(x|\mu,\theta)),$$

where $\mathrm{CDF}^{-1}_{N(0,1)}$ is the inverse cumulative distribution function of the standard normal distribution and $\mathrm{CDF}_{NB}(x|\mu,\theta)$ the negative-binomial cumulative distribution function.

### Precision-recall

We evaluated models using precision-recall curves due to the small proportion of outliers and summarized them with the area under the precision-recall curve[49] (AUPRC):

$$P = \frac{T_P}{T_P + F_P}$$
$$R = \frac{T_P}{T_P + F_N}$$
$$AP = \sum_n (R_n - R_{n-1}) P_n$$

, where $P_n$ and $R_n$ are the precision and recall at the $n^{th}$ top prediction. The AUPRC is the average precision for each threshold weighted by the recall difference.

The average precision was computed on all held-out data together on the one hand, and on held-out data for each of the 26 GTEx tissue types on the other hand. The GTEx tissue types group together highly similar tissues, notably many regions of the brain. This grouping by tissue type allows not reporting inflated performance driven by a set of highly similar tissues.

## Variant filtering, annotation, and gene-level aggregation

### Filtering for rare high-quality variants

We considered a variant to be rare if it had a minor allele frequency in the general population ≤0.001 based on the Genome Aggregation Database (gnomAD v2.1.1) and was found in at most 2 individuals within GTEx. Variants had to be supported by at least 10 reads and had to pass the conservative genotype-quality filter of GQ > 30. For structural variants, we only filtered for the number of occurrences in the GTEx dataset (less than in 2 individuals).

### Variant annotation with VEP and AbSplice

For all rare variants, we used Ensembl VEP[50] v108 to calculate consequences, LOFTEE loss-of-function annotation, as well as CADD v1.6 scores. Tissue-specific aberrant splicing predictions


20

were generated with AbSplice[28].

## Definition of gene-level features

For each combination of gene, individual, and tissue, we required a set of features to predict the underexpression outlier label. Therefore, we constructed a set of features starting from the annotations of the underlying rare variants. CADD predictions were max-aggregated per gene across the rare variants. For AbSplice, we kept the maximum absolute score in the corresponding gene and tissue.

Isoform-specific variant annotations, namely LOFTEE and VEP consequences, were aggregated differently depending on whether isoform proportions should be taken into account. When disregarding isoform proportions (i.e. the model “LOFTEE+CADD+consequences”), we only incorporated variants affecting the canonical transcript (canonical according to VEP). The gene was then assumed to have an annotation (LOFTEE classification and each VEP consequence, e.g. stop-gained) when any of the variants in the gene had this annotation. Since the VEP canonical transcript does not differ between tissues, all tissues end up with the same set of gene-level features for these annotations.

For all models incorporating isoform proportions, isoform-specific variant annotations were first weighted by the total proportion of isoforms *i* in tissue *t* that they affect:

$$w_v = \sum_{i}^{\text{isoforms}(g)} \text{proportion}(i,t) \cdot \delta_{v \text{ affects } i}$$

, where $\delta_{v \text{ affects } i}$ is 1 if the variant affects the isoform, otherwise 0, the proportion of an isoform *i* in a tissue *t* was estimated as the median TPM proportion across individuals among all isoform of the same gene *g*:

$$\text{proportion}(i,t) = \text{median}_{s \in \text{individuals}} \left( \frac{\text{tpm}(i,t,s)}{\sum_{x}^{\text{isoforms}(g)} \text{tpm}(x,t,s)} \right)$$

, and where tpm(*i, t, s*) is the transcript-level TPM obtained from GTEx v8 (dbGaP Accession phs000424.v8.p2). All resulting variant annotations were then max-aggregated per gene and tissue across variants.

# Model training

All 633 individuals were split into six cross-validation groups with approximately equal numbers of underexpression outliers and subtissues.


21

The DNA-based models were trained to predict the OUTRIDER *z*-scores. To this end, we used gradient-boosted trees[51] from the LightGBM[52] framework with default parameters: boosting_type: gbdt, learning_rate: 0.1, max_depth: -1, min_child_samples: 20, min_child_weight: 0.001, min_split_gain: 0, n_estimators: 100, num_leaves: 31, reg_alpha: 0, reg_lambda: 0, subsample: 1, subsample_for_bin: 200000, subsample_freq: 0.

For training the model integrating CAT RNA-seq data, the same cross-validation scheme was used as for the DNA-based models while excluding the CAT and highly related tissues thereof from the predicted tissues. When using fibroblasts as CAT, we excluded the non sun-exposed suprapubic skin, sun-exposed lower leg skin, and cultured fibroblasts. When using whole blood as CAT, we excluded whole blood and EBV-transformed lymphocytes from the predicted tissues. Unlike the DNA-based models, this model was trained to predict underexpressed outliers using a logistic regression taking as input i) a binary variable indicating whether the gene is expressed in the CAT, ii) the OUTRIDER *z*-score of the gene in the CAT, iii) the AbExp prediction for the target tissue, and all 3 interaction terms between those 3 variables.

## Replication in independent datasets

In the two replication datasets, variants were filtered for genotype quality ≥ 30 and read depth ≥ 10 reads. Moreover, rare variants were subsetted based on the gnomAD population with minor allele frequency ≤ 0.001. Those high-quality and rare variants were used as candidates for outlier prediction. Gene expression outliers were obtained with OUTRIDER and filtered for a sufficiently large expected number of fragments ($\mu > 450$). Samples with more than 50 outliers were removed.

The mitochondrial disease dataset[6] consisted of 311 whole-exome sequencing samples paired with RNA-seq from fibroblasts. After filtering, this dataset contained 501 underexpression outliers across 299 samples. A detailed overview of how many samples, genes, etc. remained after each filtering step can be seen in suppl. table T4.

For the amyotrophic lateral sclerosis (ALS) dataset, we downloaded 244 transcriptomes with matched whole-genome sequencing data from *dataportal.answerals.org*[29]. The data consisted of 205 cases diagnosed with amyotrophic lateral sclerosis and 39 control samples. RNA-seq measurements were obtained from iPSC-derived spinal motor neurons. After filtering, the dataset contained 739 underexpression outliers across 244 samples. The corresponding overview of how many samples, genes, etc. remained after each filtering step can be seen in suppl. table T5.


22

# UK Biobank rare variant association testing and phenotype prediction

We analyzed data from 200,593 caucasian unrelated individuals in the UK Biobank (fields 22006 and 22011), all of whom had genotypes available from exome-sequencing and microarrays as well as blood and urine measurements. A detailed list of used phenotypes can be found in suppl. table T2. For every trait, trait values were inverse rank normal transformed, as in the Genebass study[31].

## Identification of lead trait-associated variants

To control for common variants in the vicinity of each gene, trait-associated variants were obtained from the PanUKBB study[53]. Using plink v1.9[54,55], variants with a p-value ≤ 0.0001 were clumped in a 250 kbp window with an LD-threshold of $r^2 < 0.5$ to identify independent lead variants for every trait. The imputed genotypes were then subsetted for these lead variants in a 250 kbp window around each gene.

## Application of polygenic risk scores

Polygenic risk scores were selected from the study by Privé et al.[56] if available and otherwise from the study by Tanigawa et al.[57] (suppl. Table T2). Score files were obtained from the PGS catalog database[58] and applied to the imputed genotypes using plink v2.0[54,59].

## Calculation of AbExp and LOFTEE scores

The UKBB whole-exome sequencing data was subsetted for variants with a minor allele frequency ≤0.001 based on gnomAD v3.1.1 and filtered for genotype quality ≥ 30 and read depth ≥ 10 reads. The remaining variants were then annotated using Ensembl VEP[50] v108 with the LOFTEE plugin[17] and AbExp.

## Rare variant association testing

Gene association was tested using a likelihood ratio test between a restricted linear regression model containing only covariates and four linear regression models with additional predictor variables as described in Table 1.

| **Model** | **Predictor variables** | **Nr. of variables** |
|---|---|---|
| LOFTEE | Number of LOFTEE variants in the gene of | 1 |


23

| | | |
|---|---|---|
| | interest | |
| AbExp all tissues | The set of minimal AbExp scores across all rare variants in the gene (0 in absence of any rare variant) for each tissue | 48 |
| Minimum AbExp | The minimum AbExp score across all rare variants in the gene and across all tissues, 0 in absence of any rare variant | 1 |
| Median AbExp | The median AbExp score across all rare variants in the gene and across all tissues, 0 in absence of any rare variant | 1 |

576 **Table 1**: **Rare variant association test models.**

577 The following covariates were used for all models: sex, age, age², age times sex, age² times
578 sex, the 20 first genetic principal components (field 22009), lead associated variants in 250k bp
579 window around the gene of interest, and a polygenic risk score predicting the trait. The
580 polygenic risk scores and lead variants were based on the whole UK Biobank dataset. This mild
581 data leakage may have led to model overfitting. However, since these features were used as
582 covariates in both restricted and full models, the comparison between models remained valid. P-
583 value calibration of the models was assessed by permuting the phenotype once. Identification of
584 significantly trait-associated genes was performed on two thirds of the dataset.

## 585 Phenotype prediction

586 For the common-variant-based phenotype prediction model, the following features were used:
587 sex, age, age², age times sex, age² times sex, the 20 first genetic principal components (field
588 22009), and a polygenic risk score predicting the trait. In contrast to the rare variant association
589 testing, we did not include lead variants of the associating genes, as these would lead to a too
590 large number of predictor variables.
591 We compared two phenotype prediction models integrating rare and common variants. Both are
592 nonlinear models. To this end, we used gradient-boosted trees[51] from the LightGBM[52]
593 framework with default parameters: boosting_type: gbdt, learning_rate: 0.1, max_depth: -1,
594 min_child_samples: 20, min_child_weight: 0.001, min_split_gain: 0, n_estimators: 100,
595 num_leaves: 31, reg_alpha: 0, reg_lambda: 0, subsample: 1, subsample_for_bin: 200000,
596 subsample_freq: 0. Both models include the same features as the common-variant based

prediction models. The AbExp phenotype prediction model further included the AbExp scores of significantly trait-associated genes in 48 tissues, while the LOFTEE phenotype prediction model included the number of LOFTEE pLoF variants of significantly trait-associated genes. For each trait, the two models can have a different set of significantly trait-associated genes as identified by their corresponding RVATs. The phenotype prediction models were trained with 5-fold cross-validation on the remaining third of the dataset. All evaluations were performed on the hold-out folds.


25

# Declarations

## Ethics approval and consent to participate

No new data was generated for this study. The different ethics approval can be found in the corresponding publications (GTEx[60], AnswerALS[29], mitochondrial disease[6]). The UK Biobank was approved by the North West Multi-centre Research Ethics Committee (21/NW/0157). Our reference number approved by the UK Biobank is 25214. All UK Biobank study participants gave written informed consent.

The research conformed to the principles of the Declaration of Helsinki.

## Consent for publication

All individuals included or their legal guardians provided written consent to share pseudonymized patient data and analysis data, as described in the original publications.

## Competing interests

The authors declare that they have no competing interests.


## Acknowledgements

We thank Felix Brechtmann, Alexander Karollus, and Holger Prokisch for fruitful discussions and their valuable feedback on the manuscript. Figure 1 was created with BioRender.com.


## Authors' contributions

J.G. conceptualized the project. F.R.H., F.P.C. and J.G. designed the methodology. F.R.H., N.W. and V.A.Y. curated the data. F.R.H. and J.L. performed investigation and provided the software. F.R.H., J.L., and V.A.Y.. performed visualizations. F.R.H., V.A.Y. and J.G. wrote the original draft of the manuscript. All authors reviewed and edited the manuscript. J.G. supervised the project with the help of F.P.C. and V.A.Y.


## Funding

The German Bundesministerium für Bildung und Forschung (BMBF) supported the study through the Model Exchange for Regulatory Genomics project (MERGE; grant no. 031L0174A to F.R.H. and J.G.); the VALE (Entdeckung und Vorhersage der Wirkung von genetischen



26

Varianten durch Artifizielle Intelligenz für Leukämie Diagnose und Subtypidentifizierung) project (031L0203B to F.R.H., J.G); and the ERA PerMed project PerMiM (01KU2016B to J.L., V.A.Y. and J.G.). N.W. is supported by the Helmholtz Association under the joint research school 'Munich School for Data Science – MUDS'. F.P.C. was funded by the Free State of Bavaria's Hightech Agenda through the Institute of AI for Health (AIH).
The Genotype-Tissue Expression (GTEx) project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by the NCI, NHGRI, NHLBI, NIDA, NIMH and NINDS. This study was supported by data provided by the Answer ALS Consortium, administered by the Robert Packard Center for ALS at Johns Hopkins. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. UK Biobank was established by the Wellcome Trust medical charity, Medical Research Council, Department of Health, Scottish Government and the Northwest Regional Development Agency. It has also had funding from the Welsh Government, British Heart Foundation, Cancer Research UK and Diabetes UK. UK Biobank is supported by the National Health Service (NHS).

# Availability of data and materials

## Data availability

We provide the aberrant expression benchmark dataset, isoform proportions and the expected gene expression in GTEx v8 as open-access in the Zenodo repository[61] (doi: 10.5281/zenodo.8427312).

## Code availability

A Snakemake pipeline to calculate AbExp predictions can be found at https://github.com/gagneurlab/abexp.
The source code for the UK Biobank rare-variant association study and phenotype prediction can be found in the following repositories:

- Main analysis pipeline: https://github.com/gagneurlab/abexp-ukbb-trait-analysis


27

- Variant clumping: https://github.com/gagneurlab/abexp-ukbb-variant-clumping
- Polygenic risk score calculation: https://github.com/gagneurlab/abexp-ukbb-prs


28

658 # References


659 1. Wang, L.-H., Wu, C.-F., Rajasekaran, N. & Shin, Y. K. Loss of Tumor Suppressor Gene
660 Function in Human Cancer: An Overview. *Cell. Physiol. Biochem.* **51**, 2647–2693 (2018).
661 2. Kontomanolis, E. N. *et al.* Role of Oncogenes and Tumor-suppressor Genes in
662 Carcinogenesis: A Review. *Anticancer Res.* **40**, 6009–6015 (2020).
663 3. Gonorazky, H. D. *et al.* Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool
664 for Rare Mendelian Disease. *Am. J. Hum. Genet.* **104**, 466–483 (2019).
665 4. Frésard, L. *et al.* Identification of rare-disease genes using blood transcriptome sequencing
666 and large control cohorts. *Nat. Med.* **25**, 911–919 (2019).
667 5. Murdock, D. R. Enhancing Diagnosis Through RNA Sequencing. *Clin. Lab. Med.* **40**, 113–
668 119 (2020).
669 6. Yépez, V. A. *et al.* Clinical implementation of RNA sequencing for Mendelian disease
670 diagnostics. *Genome Med.* **14**, 38 (2022).
671 7. Dekker, J. *et al.* Web-accessible application for identifying pathogenic transcripts with RNA-
672 seq: Increased sensitivity in diagnosis of neurodevelopmental disorders. *Am. J. Hum.*
673 *Genet.* **110**, 251–272 (2023).
674 8. Deshwar, A. R. *et al.* Trio RNA sequencing in a cohort of medically complex children. *Am. J.*
675 *Hum. Genet.* **110**, 895–900 (2023).
676 9. Smail, C. *et al.* Integration of rare expression outlier-associated variants improves polygenic
677 risk prediction. *Am. J. Hum. Genet.* **109**, 1055–1064 (2022).
678 10. Li, B. *et al.* Evaluation of PrediXcan for prioritizing GWAS associations and predicting gene
679 expression. in *Biocomputing 2018* 448–459 (WORLD SCIENTIFIC, 2017).
680 doi:10.1142/9789813235533_0041.
681 11. Brechtmann, F. *et al.* OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed
682 Genes in RNA Sequencing Data. *Am. J. Hum. Genet.* (2018)

683 doi:10.1016/j.ajhg.2018.10.025.
684 12. Salkovic, E., Sadeghi, M. A., Baggag, A., Salem, A. G. R. & Bensmail, H. OutSingle: a novel
685 method of detecting and injecting outliers in RNA-Seq count data using the optimal hard
686 threshold for singular values. *Bioinformatics* **39**, btad142 (2023).
687 13. Segers, A., Gilis, J., Heetvelde, M. V., Baere, E. D. & Clement, L. Juggling offsets unlocks
688 RNA-seq tools for fast scalable differential usage, aberrant splicing and expression
689 analyses. 2023.06.29.547014 Preprint at https://doi.org/10.1101/2023.06.29.547014 (2023).
690 14. Li, X. *et al.* The impact of rare variation on gene expression across tissues. *Nature* **550**,
691 239–243 (2017).
692 15. Ferraro, N. M. *et al.* Transcriptomic signatures across human tissues identify functional rare
693 genetic variation. *Science* **369**, (2020).
694 16. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human
695 tissues. *Science* **369**, 1318–1330 (2020).
696 17. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in
697 141,456 humans. *Nature* **581**, 434–443 (2020).
698 18. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the
699 deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–
700 D894 (2019).
701 19. Cummings, B. B. *et al.* Transcript expression-aware annotation improves rare variant
702 interpretation. *Nature* **581**, 452–458 (2020).
703 20. Morales, J. *et al.* A joint NCBI and EMBL-EBI transcript set for clinical genomics and
704 research. *Nature* **604**, 310–315 (2022).
705 21. Dugan, S. L. *et al.* New recessive truncating mutation in LTBP3 in a family with oligodontia,
706 short stature, and mitral valve prolapse. *Am. J. Med. Genet. A.* **167**, 1396–1399 (2015).
707 22. Sharon, D., Kimchi, A. & Rivolta, C. OR2W3 sequence variants are unlikely to cause
708 inherited retinal diseases. *Ophthalmic Genet.* **37**, 366–368 (2016).

709 23. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for
710 differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140
711 (2010).
712 24. Fair, B. J. *et al.* Gene expression variability in human and chimpanzee populations share
713 common determinants. *eLife* **9**, e59929 (2020).
714 25. Bader, D. M. *et al.* Negative feedback buffers effects of regulatory variants. *Mol. Syst. Biol.*
715 **11**, 785 (2015).
716 26. Einarsson, H. *et al.* Promoter sequence and architecture determine expression variability
717 and confer robustness to genetic variants. *eLife* **11**, e80943 (2022).
718 27. Teran, N. A. *et al.* Nonsense-mediated decay is highly stable across individuals and tissues.
719 *bioRxiv* 2021.02.03.429654 (2021) doi:10.1101/2021.02.03.429654.
720 28. Wagner, N. *et al.* Aberrant splicing prediction across human tissues. *Nat. Genet.* **55**, 861–
721 870 (2023).
722 29. Baxi, E. G. *et al.* Answer ALS, a large-scale resource for sporadic and familial ALS
723 combining clinical and multi-omics data from induced pluripotent cell lines. *Nat. Neurosci.*
724 **25**, 226–237 (2022).
725 30. Wang, Q. *et al.* Rare variant contribution to human disease in 281,104 UK Biobank exomes.
726 *Nature* **597**, 527–532 (2021).
727 31. Karczewski, K. J. *et al.* Systematic single-variant and gene-based association testing of
728 thousands of phenotypes in 394,841 UK Biobank exomes. *Cell Genomics* **2**, 100168 (2022).
729 32. Fiziev, P. P. *et al.* Rare penetrant mutations confer severe risk of common diseases.
730 *Science* **380**, eabo1131 (2023).
731 33. Wu, M. C. *et al.* Rare-Variant Association Testing for Sequencing Data with the Sequence
732 Kernel Association Test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
733 34. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-Variant Association Analysis: Study
734 Designs and Statistical Tests. *Am. J. Hum. Genet.* **95**, 5–23 (2014).

735 35. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a
736 Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med.* **12**, e1001779
737 (2015).
738 36. Zhou, W. *et al.* SAIGE-GENE+ improves the efficiency and accuracy of set-based rare
739 variant association tests. *Nat. Genet.* **54**, 1466–1469 (2022).
740 37. Clarke, B. *et al.* Integration of variant annotations using deep set networks boosts rare
741 variant association genetics. 2023.07.12.548506 Preprint at
742 https://doi.org/10.1101/2023.07.12.548506 (2023).
743 38. Cummings, B. B. *et al.* Improving genetic diagnosis in Mendelian disease with transcriptome
744 sequencing. *Sci. Transl. Med.* (2017) doi:10.1126/scitranslmed.aal5209.
745 39. Kremer, L. S. *et al.* Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat.*
746 *Commun.* **8**, (2017).
747 40. Lunke, S. *et al.* Integrated multi-omics for rapid rare disease diagnosis on a national scale.
748 *Nat. Med.* **29**, 1681–1691 (2023).
749 41. Aicher, J. K., Jewell, P., Vaquero-Garcia, J., Barash, Y. & Bhoj, E. J. Mapping RNA splicing
750 variations in clinically-accessible and non-accessible tissues to facilitate Mendelian disease
751 diagnosis using RNA-seq. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **22**, 1181–1190 (2020).
752 42. Karollus, A., Mauermeier, T. & Gagneur, J. Current sequence-based models capture gene
753 expression determinants in promoters but mostly ignore distal enhancers. *Genome Biol.* **24**,
754 56 (2023).
755 43. Zhou, J. *et al.* Deep learning sequence-based ab initio prediction of variant effects on
756 expression and disease risk. *Nat. Genet.* **50**, 1171–1179 (2018).
757 44. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-
758 range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
759 45. Linder, J., Srivastava, D., Yuan, H., Agarwal, V. & Kelley, D. R. Predicting RNA-seq
760 coverage from DNA sequence as a unifying model of gene regulation. 2023.08.30.555582

761 Preprint at https://doi.org/10.1101/2023.08.30.555582 (2023).
762 46. Yépez, V. A. *et al.* Detection of aberrant gene expression events in RNA sequencing data.
763 *Nat. Protoc.* **16**, 1276–1296 (2021).
764 47. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes.
765 *Nucleic Acids Res.* **47**, D766–D773 (2019).
766 48. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under
767 dependency. *Ann. Stat.* **29**, 1165–1188 (2001).
768 49. Zhu, M. Recall, Precision and Average Precision. *Dep. Stat. Actuar. Sci. Univ. Waterloo*
769 *Waterloo* (2004).
770 50. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
771 51. Hastie, T., Tibshirani, R. & Friedman, J. Boosting and Additive Trees. in *The Elements of*
772 *Statistical Learning: Data Mining, Inference, and Prediction* (eds. Hastie, T., Tibshirani, R. &
773 Friedman, J.) 337–387 (Springer, 2009). doi:10.1007/978-0-387-84858-7_10.
774 52. Ke, G. *et al.* LightGBM: A Highly Efficient Gradient Boosting Decision Tree. in *Advances in*
775 *Neural Information Processing Systems* vol. 30 (Curran Associates, Inc., 2017).
776 53. Pan-UKB team. Pan-ancestry genetic analysis of the UK Biobank. (2020).
777 54. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer
778 datasets. *GigaScience* **4**, s13742-015-0047–8 (2015).
779 55. Shaun Purcell & Christopher Chang. PLINK 1.9. (2020).
780 56. Privé, F. *et al.* Portability of 245 polygenic scores when derived from the UK Biobank and
781 applied to 9 ancestry groups from the same cohort. *Am. J. Hum. Genet.* **109**, 12–23 (2022).
782 57. Tanigawa, Y. *et al.* Significant sparse polygenic risk scores across 813 traits in UK Biobank.
783 *PLOS Genet.* **18**, e1010105 (2022).
784 58. Lambert, S. A. *et al.* The Polygenic Score Catalog as an open database for reproducibility
785 and systematic evaluation. *Nat. Genet.* **53**, 420–425 (2021).
786 59. Shaun Purcell & Christopher Chang. PLINK 2.0. (2022).

60. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

61. Hölzlwimmer, F. R. Aberrant gene expression prediction benchmark based on GTEx v8. (2023) doi:10.5281/zenodo.8427311.


34