

# Disease-specific prioritization of non-coding GWAS variants based on chromatin accessibility

Qianqian Liang<sup>1,2</sup>, Abin Abraham<sup>3</sup>, John A Capra<sup>4</sup> and Dennis Kostka<sup>1,5,\*</sup>

<sup>1</sup> Department of Developmental Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

<sup>2</sup> Department of Human Genetics, University of Pittsburgh School of Public Health, Pittsburgh, PA, USA

<sup>3</sup> Children's Hospital of Philadelphia, Philadelphia, PA, USA

<sup>4</sup> Department of Epidemiology & Biostatistics and Bakar Computational Health Sciences Institute, University of California, San Francisco, CA, USA

<sup>5</sup> Department of Computational & Systems Biology and Center for Evolutionary Biology and Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

\* Correspondence: [kostka@pitt.edu](mailto:kostka@pitt.edu)

## Abstract

Non-protein-coding genetic variants are a major driver of the genetic risk for human disease; however, identifying which non-coding variants contribute to which diseases, and their mechanisms, remains challenging. In-silico variant prioritization methods quantify a variant's severity in the context of having a phenotypic effect; but for most methods the specific phenotype and disease context of the prediction are poorly defined. For example, many commonly used methods provide a single organism-wide score for each variant, while other methods summarize a variant's impact specifically in certain tissues and/or cell-types. Here we propose a complementary disease-specific variant prioritization scheme, which is motivated by the observation that the variants contributing to different diseases often operate through different biological mechanisms.

We combine tissue/cell-type specific scores into disease-specific scores with a logistic regression approach and apply it to 25,000 non-coding variants spanning 111 diseases. We show that disease-specific aggregation of tissue/cell-type specific scores (GenoSkyline, Fit-Cons2, DNA accessibility) significantly improves the association of common non-coding genetic variants with disease (average precision: 0.151, baseline=0.09), compared with organism-wide scores (GenoCanyon, LINSIGHT, GWAVA, eigen, CADD; average precision: 0.129, baseline=0.09). Calculating disease similarities based on data-driven aggregation weights highlights meaningful disease groups (e.g., immune system related diseases and mental/behavioral disorders), and it provides information about tissues and cell-types that drive these similarities (e.g., lymphoblastoid T-cells for immune-system diseases). We also show that so-learned similarities are complementary to genetic similarities as quantified by genetic correlation. Overall, our aggregation approach demonstrates the strengths of disease-specific variant prioritization, leads to improvement in non-coding variant prioritization, and it enables interpretable models that link variants to disease via specific tissues and/or cell-types.

## 1 Introduction

Characterizing non-coding genetic variants in the human genome is essential for making progress toward better understanding the genetic components of disease, because ~90% of disease-associated variants discovered by genome-wide association studies (GWAS) are located in non-protein-coding

5 regions [1]. Further on, whole-genome sequencing (WGS) discovers disease-associated variants genome-  
6 wide [2, 3] and is increasingly becoming an assay of choice. Therefore, approaches for characterizing  
7 and prioritizing non-coding variants can be expected to play an increasingly important role, especially  
8 when assessing discovered variants in the context of functional follow-up experimental studies.

9 Efforts to computationally characterize and better understand non-coding variants take advantage  
10 of sequence, functional genomics, comparative genomics, and epigenomics data [4, 5, 6], and more.  
11 These data are combined and used to train and develop supervised and/or unsupervised models that  
12 attempt to quantify a variant’s impact [7]. We find it conceptually useful to distinguish between  
13 variant scores that model overall impact (that is on the level of the whole organism, *organism-level*  
14 scores) and scores that quantify impact in a specific context, like a tissue or a cell-type (i.e., *tissue-level*  
15 scores). Examples for organism-level scores are CADD [8], Eigen [9], or LINSIGHT [10], while scores  
16 from methods like GenoSkyline [11], Fitcons2 [12], or FUN-LDA [13] are tissue-specific.

17 Often interest in a set of variants is from the perspective of studying a specific disease. In that case,  
18 organism-level scores are likely to be overly general. That is, a variant’s impact might be considered  
19 high because it disrupts the functional role of a sequence element. However, that functional role may  
20 be unrelated to the disease of interest. In one study, for instance, organism-level scores like CADD  
21 and DANN were unable to discover an enrichment signal for brain-related traits, while context-specific  
22 variant scores focusing on relevant tissues were successful [14]. This demonstrates that tissue-specific  
23 scores can address the issue of disease specificity to some extent. However, aspects of disease-relevant  
24 tissues typically remain unknown, and often more than one tissue is implicated with a specific trait  
25 (termed “multifactorial” and “polyfactorial” traits) [15]. This suggests the use of *disease-specific*  
26 variant scores that characterize variants in the context of a specific disease phenotype of interest.

27 Computational methods for disease-specific variant prioritization do exist. Some approaches are  
28 geared towards one disease (e.g, congenital heart disease [16], amyotrophic lateral sclerosis [17]) or  
29 towards a specific class of diseases (e.g., autoimmune diseases [18]). This focus prevents them from  
30 being readily adapted to other disease types. Others, like DIVAN [19], PINES [20], and ARVIN [21],  
31 cover a broader range of disease types. Of these, ARVIN requires a priori knowledge of disease-relevant  
32 tissues, whereas DIVAN and PINES do not. PINES uses an enrichment-based method to predict and  
33 up-weight disease-relevant tissues/cell-types, whereas DIVAN uses a more complex machine learning  
34 algorithm. The PINES approach has been evaluated on a relatively small set of traits (~10 different  
35 contexts), while DIVAN’s more complex model renders understanding the relationship between different  
36 tissues and diseases difficult.

37 In this work, we derive disease-specific variant scores by combining published tissue-specific scores.  
38 We use a carefully regularized logistic regression approach to derive data-driven disease-specific com-  
39 bination weights, which allow us to better associate variants with disease. In addition, they enable us  
40 to quantify a similarity between different disease phenotypes. Using the NHGRI-EBI GWAS catalog  
41 [1] we compiled a benchmark dataset containing about 63k phenotype-associated non-protein-coding  
42 single nucleotide variants across 111 disease phenotypes (together with matched random controls). We  
43 then demonstrate that using disease-specific combination weights outperforms conventional organism-  
44 level approaches, that our interpretable model has competitive performance, and that it enables a  
45 disease similarity measure that captures information complementary to established measures like ge-  
46 netic correlation.

## 2 Results

### 2.1 Non-coding GWAS variants associated with disease phenotypes, and matched controls

In order to study variant prioritization methods, we created a dataset of “positive” (i.e., disease associated) non-coding variants, matched with a random set of “negative” or “control” variants. This setup allowed us to quantitatively assess prioritization methods based on their performance in discriminating positive from control variants.

#### 2.1.1 Disease-associated non-coding SNVs

We used a subset of single nucleotide variants (SNVs) reported in the EBI/NIH GWAS catalog [1] to compile an inventory of disease-associated non-coding variants. Specifically, we focused in reported variants that (a) do not overlap protein-coding sequence (see **Methods**) and (b) that are associated with a disease phenotype as noted in the Experimental Factor Ontology (EFO) trait description, which is provided within the catalog. We define disease phenotypes as descendants of the EFO term “disease” (EFO:0000408). Focusing on disease terms with at least 100 annotated SNVs resulted in 26,080 associations involving 20,656 SNVs and 67 disease phenotypes. The EFO provides parent-child relations between disease terms (parent = more general, child = more specific), and propagating SNVs from child-terms to parent-terms increased the number of disease phenotypes with at least 100 SNVs, resulting in 77,028 association between 25,516 SNVs and 111 diseases. We find that most of the SNVs we recover are located in intronic (60.5%) and intergenic (25.8%) sequence (**Fig. 1A**), and that a majority of SNVs are directly annotated to a single disease phenotype (**Fig. 1B**). After propagating annotated SNVs from child to parent terms, SNV-to-disease annotations become predominantly many:many (**Fig. 1B**). **Suppl. Data SD1** lists disease terms and corresponding numbers of disease-associated SNVs.

#### 2.1.2 Control SNVs

For each disease-associated SNV we selected ~10 matched control-SNVs using a re-implementation of the SNPsnap approach [22], while avoiding duplicate control-SNV across the overall dataset (see **Methods**). This yielded 255,137 control SNVs (for some disease associated SNVs we could not retrieve the full ten control SNVs). With these results we have access to data for 111 disease terms, containing disease-associated SNVs together with matched controls. **Suppl. Data SD2** and **SD3** contain information about all disease and control SNVs used in this study, respectively.

### 2.2 Disease-specific non-coding variant prioritization with organism-level variant scores is only moderately successful

We assessed how well current commonly-used organism-level variant scores are able to prioritize disease-associated vs. control-SNVs for the 111 disease terms we studied. **Fig. 2** summarizes results, where boxplots of two performance measures (area under the ROC curve and average precision (= area under the precision recall curve)) are shown for CADD [8], eigen [9], GenoCanyon [11], GWAVA [23], and LINSIGHT [10] scores. We find that organism-level scores, while improving upon random guessing, are only moderately successful in correctly prioritizing disease-associated non-coding variants. Comparing variant scores with each other we find that relative performance differences appear overall robust with

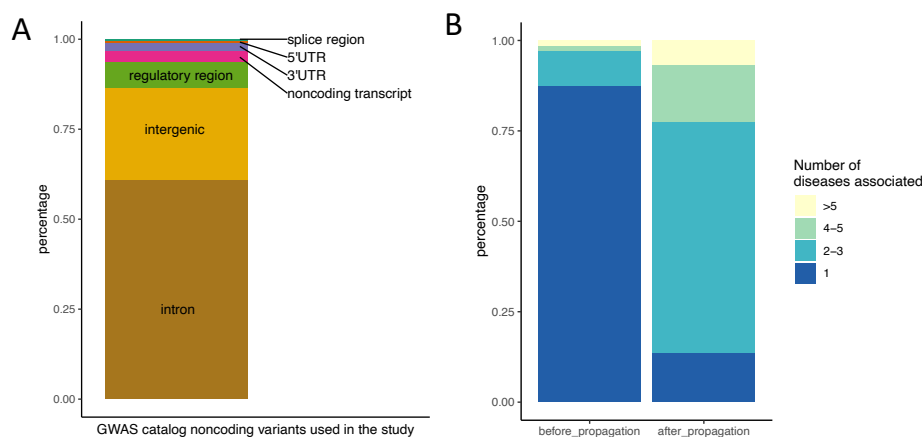


Figure 1: **Disease-associated non-coding SNVs.** (A) Genomic context of non-coding SNVs used in this study. (B) Percentage of the SNVs used that are annotated to 1, 2-3, 4-5 or more than 5 disease phenotypes, before and after propagating SNV-phenotype associations according to EFO parent-child annotations. Genomic context annotation is adapted from the CONTEXT column from the GWAS catalog, where we combine splice donor, splice region and splice acceptor variants into splice variants and I combine TF binding variants and regulatory regions variants into regulatory region variants.

86 respect to the metric employed (area under the ROC curve vs. average precision). It is qualitatively  
87 visible that CADD performs less favorably than other methods, but also that there are differences  
88 between these. We therefore compared performance between different scores in more detail.

89 We studied the performance of different scores at two levels of resolution: In aggregate across all  
90 disease terms, and for each disease term separately. For both approaches we used Wilcoxon signed-  
91 ranks tests to decide whether one score significantly outperforms another score (= significant p-value)  
92 or whether performance is tied (= non-significant p-value); see **Methods** section. Results are sum-  
93 marized in **Tab. 1**. We find that GenoCanyon has better performance compared with other variant  
94 scores, followed by LINSIGHT, GWAVA and eigen, while CADD is consistently outperformed by other  
95 methods. Performance differences between LINSIGHT, GWAVA and eigen are not significant when ag-  
96 gregating across disease terms (last three columns in **Tab.1**); however, when counting individual terms  
97 LINSIGHT has most wins and fewest losses, while eigen has most losses and fewest wins, leading to the  
98 ordering displayed in **Tab.1. Suppl. Data SD4 and SD5** contain results for all comparisons. Overall  
99 these quantitative results are in-line with the visual impression from **Fig. 2**. Next, we investigated  
100 if the performance of organism-level variant scores could be improved by using tissue-specific scoring  
101 approaches.

## 102 2.3 Disease-specific scores improve non-coding variant prioritization

### 103 2.3.1 Disease-specific aggregation weights for tissue-specific variant scores

104 We studied three tissue-specific scores for variant prioritization to explore if their usage can improve  
105 the performance of organism-level scores. Specifically, we used Genoskyline [11] and Fitcons2 [12] as  
106 scores designed to prioritize variants, and we also evaluated DNase I hypersensitivity (DHS) profiles  
107 from the ENCODE project [6]. All of these scores are available for 127 contexts [5] spanning a diverse  
108 set of cell and tissue types, including heart, brain, immune cells, and more.

109 For each tissue-specific score we assess two approaches to prioritize variants. First, as a baseline

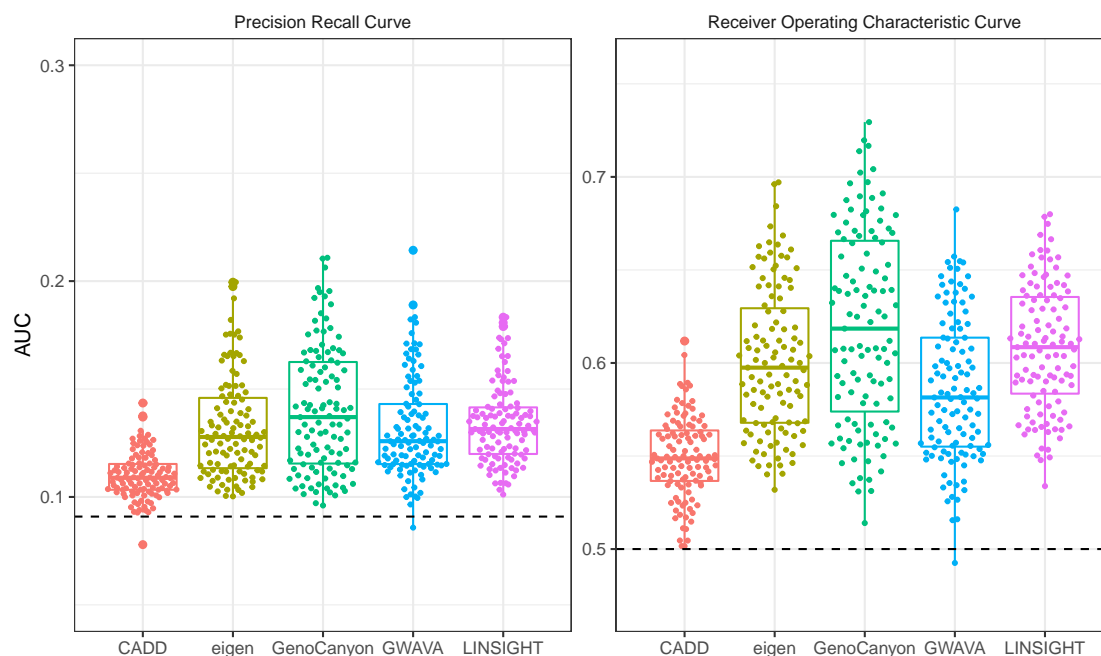


Figure 2: *Organism-level variant scores are moderately successful in prioritizing non-coding disease-associated variants.* Different organism-level variant prioritization scores are shown on the x-axis, the y-axis displays performance in terms of average precision (area under the precision recall curve, left panel) and area under the receiver-operator curve (right panel). Each point represents a specific disease term from the experimental factor ontology. Horizontal lines spanning data sets show expectations under random guessing.

110 approach we aggregate scores across tissues in a *disease-agnostic* way. That is, for a specific variant we  
111 average scores at the variant position across all tissues (termed *tissue-mean*), essentially producing a  
112 organism-level type score, independent of the disease term under consideration. Second, we aggregate  
113 scores across tissues in a *disease-specific* way. Briefly, we train a regularized logistic regression model  
114 for each disease term that learns disease-specific tissue aggregation weights. In a nested cross-validation  
115 setup learned weights are then applied to held-out variants, allowing for a fair performance assessment  
116 of this approach (termed *tissue-weighted*), see **Methods**. **Fig. 3** summarizes our findings.

117 In **Fig. 3A** we show tissue-mean performance (measured by average precision) for the three scores  
118 we study on the left, and tissue-weighted performance on the right. For all three scores tissue-weighted  
119 significantly outperforms tissue-mean (Wilcoxon signed-ranks test, p-values < 0.0001). **Fig. 3B** shows  
120 tissue-mean vs. tissue-weighted comparisons for each score, and we see that in almost all disease terms  
121 tissue-weighted outperforms tissue-mean. See **Suppl. Data SD6** and **SD7** for tissue-mean vs. tissue-  
122 weighted performances for each disease term, and for aggregated performances across all disease terms.  
123 The improvement remains evident if we limit disease-associated SNVs to one variant per LD block, and  
124 also when we insure that the SNVs in the training and test datasets are not on the same chromosome  
125 (See **Suppl. Fig. S17 - S20** and the Supplemental material for more details).

126 While the performance-gain for tissue-weighted is broadly consistent across diseases, for some it  
127 is more pronounced than for others. To illustrate this observation, we selected four disease terms  
128 with a high performance gain, four terms with a medium gain, and four terms where we observed  
129 the least gain (Best improvement, ranking 1-4; middle improvement, ranking 20-23; least improve-

Score/Method	By disease term			Aggregated		
	Wins	Losses	Ties	Wins	Losses	Ties
GenoCanyon	307	106	31	4	0	0
LINSIGHT	281	146	17	1	1	2
GWAVA	221	196	27	1	1	2
eigen	219	201	24	1	1	2
CADD	24	403	17	0	4	0

Table 1: **Relative performance of organism-level variant scores.** Wins, Losses, Ties refers to significantly better (or worse, or tied) performance across all possible pairings (see **Methods**). The first three columns summarize separate comparisons for each disease term (for each row there are four other methods and 111 terms, i.e. 444 comparisons), while the last three columns represent results of comparisons between scores aggregated across terms. Average precision was used as the performance metric, and Wilcoxon signed-ranks tests to determine wins and losses (p-values less than 0.05 are reported as ties).

130 ment, ranking 108-111). **Fig. 4** shows our findings, where variability in tissue-weighted performance  
131 induced by varying train-test-fold splits during cross-validation is also displayed. We see that for  
132 Celiac Disease (EFO:0001060), Systemic Scleroderma (EFO:0000717), Chronic Lymphocytic Leukemia  
133 (EFO:0000095) and Sclerosing Cholangitis (EFO:0004268) performance is consistently improved for  
134 all three tissue-weighted scores, while for Retinopathy (EFO:0003839), Endometriosis (EFO:0001065),  
135 Diabetic Nephropathy (EFO:0000401) and HIV-1 Infection(EFO:0000180) we find no improvement.  
136 We also note that disease terms with pronounced improvement appear to have better baseline (i.e.,  
137 tissue-mean) performance than disease terms where we find little or no benefit of the tissue-weighted  
138 approach. Improvement for diseases shown in **Fig. 4** is largest for DHS, but, consistent with **Fig. 3**,  
139 we see improvement for Fitcons2 and Genoskyline as well.

### 140 **2.3.2 DNase I hypersensitivity (DHS) scoring outperforms other tissue specific scores**

141 To quantify relative performance of the three different tissue-specific scores, we proceed similarly  
142 to organism-level scores. Focusing on pairwise comparisons we find that DHS scores outperform  
143 Genoskyline and Fitcons2 for most disease terms, and on average (see **Tab. 2**). This observation  
144 is consistent with **Fig. 3** and 4, which often show higher average precision values for DHS than  
145 for the other two scores. Notably, baseline (i.e., tissue-mean) performance of DHS does not appear  
146 significantly better than that of Genoskyline (**Fig. 3**). **Suppl. Data SD8** and **SD9** contain details for  
147 comparisons between DHS, Fitcons2 and Genoskyline for all disease terms. Next, we explored whether  
148 disease-specific tissue weights outperform organism-level scores.

### 149 **2.3.3 DNase I hypersensitivity (DHS) tissue-weighted scoring outperforms organism-level variant scores**

151 To compare the DHS tissue-weighted score with organism-level scores, we directly contrasted their  
152 performance. Similar to before, **Tab. 3** summarizes DHS “wins” (= significantly better performance  
153 of DHS tissue-weighted,  $p\text{-value} \leq 0.05$ ), losses, and ties, compared with five organism-level variant  
154 scores, individually (i.e., per disease term) and aggregated across disease terms. In addition, **Tab. ST4**  
155 summarizes pair-wise comparisons between tissue-weighted DHS and each organism-level score. We  
156 find that DHS tissue-weighted outperforms all organism-level scores in the aggregated analyses, and  
157 that it outperforms all other scores on the majority of disease terms (it only performs significantly

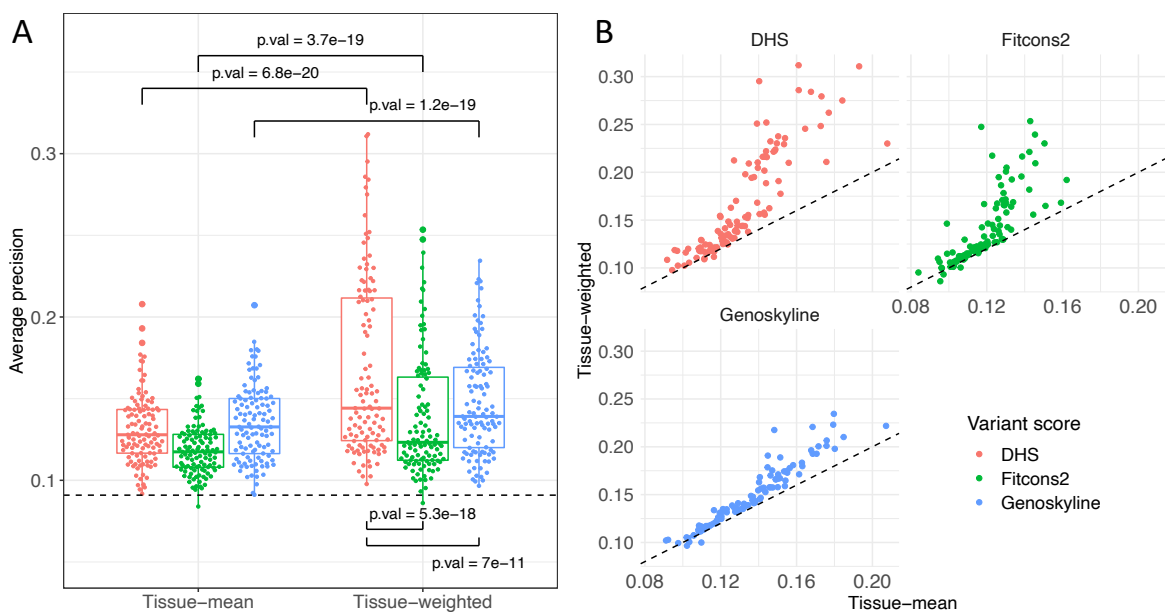


Figure 3: ***Disease-specific tissue weights improve variant prioritization.*** Performance of three tissue-specific variant scores (DHS, Fitcons2, Genoskyline) is used to prioritize non-coding disease-associated variants for disease terms using two approaches: *tissue-mean* (i.e., disease-agnostic, baseline) on the left side and *tissue-weighted* (i.e., disease specific) on the right side. P-values were calculated using a Wilcoxon signed-ranks test (A). Scatter plot of tissue-mean vs. tissue-weighted performance (average precision) for each tissue-specific score; dashed line denotes the diagonal (B).

158 worse than any other score in 44 out of 550 comparisons).

159 GenoCanyon is the most competitive organism-level score, where DHS is significantly better for 92  
 160 terms out of 111 (~83%). Interestingly, LINSIGHT performs better against DHS than GenoCanyon,  
 161 which is the best overall performing organism-level score (see **Tab. ST4**). **Suppl. Data SD10** contains  
 162 detailed results for each comparison. We also find that DHS outperforms organism-level scores when  
 163 aggregating over disease terms (also see **Suppl. Data SD11**).

164 To illustrate the gain in performance, we selected four example disease terms where disease-  
 165 specific variant prioritization yielded high improvements, medium improvements, comparable per-  
 166 formance, and worse performance, respectively. Selection was based on ranking differences between  
 167 DHS and GenoCanyon: best improvement, ranks 1-4; medium improvements, ranks 25-28; compa-  
 168 rable performance, ranks 64-67; GenoCanyon better, ranks 108-111. Results are summarized in  
 169 **Fig. 5**, where we find substantial improvements using tissue-weighted scoring for Systemic Sclero-  
 170 derma (EFO:0000717), Celiac Disease (EFO:0001060), Sclerosing Cholangitis (EFO:0004268) and Mul-  
 171 tiple Sclerosis (EFO:0003885), for which we have already noticed substantial improvement of DHS  
 172 tissue-weighted over DHS tissue-mean. Disease terms where GenoCanyon is performing better include  
 173 Venous Thromboembolism (EFO:0004286), Diverticular Disease (EFO:0009959), Non-small Cell Lung  
 174 Carcinoma (EFO:0003060), and Lung Adenocarcinoma (EFO:0000571).

175 To make DHS tissue-weighted scores available, we generated pre-computed scores for 111 diseases  
 176 at every base across the genome (for chromosomes 1-22, available at <https://doi.org/10.7910/DVN/AUAJ7K>).  
 177 Scores were calculated at 25 bp resolution, the same as DHS scores.

Score/Method	By disease term			Aggregated		
	Wins	Losses	Ties	Wins	Losses	Ties
DHS	180	22	20	2	0	0
Genoskyline	96	94	32	1	1	0
Fitcons2	19	179	24	0	2	0

Table 2: *DHS outperforms other tissue-specific scores*. Wins, Losses, Ties refer to significantly better (or worse, or tied) performance across all possible score pairings (see **Methods**). The first three columns summarize separate comparisons for each disease term (for each row there are two other methods and 111 terms, i.e., 222 comparisons), while the last three columns represent results of comparisons aggregated over disease terms. Average precision was used as the performance metric, and the Wilcoxon signed-ranks test to determine wins and losses (p-values less than 0.05 are reported as ties).

Score/Method	By disease term			Aggregated		
	Wins	Losses	Ties	Wins	Losses	Ties
DHS	474	44	37	5	0	0
GenoCanyon	314	198	43	4	1	0
LINSIGHT	298	230	27	1	2	2
GWAVA	233	289	33	1	2	2
eigen	223	299	33	1	2	2
CADD	28	510	17	0	5	0

Table 3: *DHS outperforms organism-level variant scores*. Wins, Losses, Ties refer to significantly better (or worse, or tied) performance across all possible score pairings (see **Methods**). The first three columns summarize separate comparisons for each disease term (for each row there are two other methods and 111 terms, i.e., 555 comparisons), while the last three columns represent results of comparisons aggregated over terms. Average precision was used as the performance metric, and the Wilcoxon signed-ranks test to determine wins and losses (p-values less than 0.05 were reported as ties).



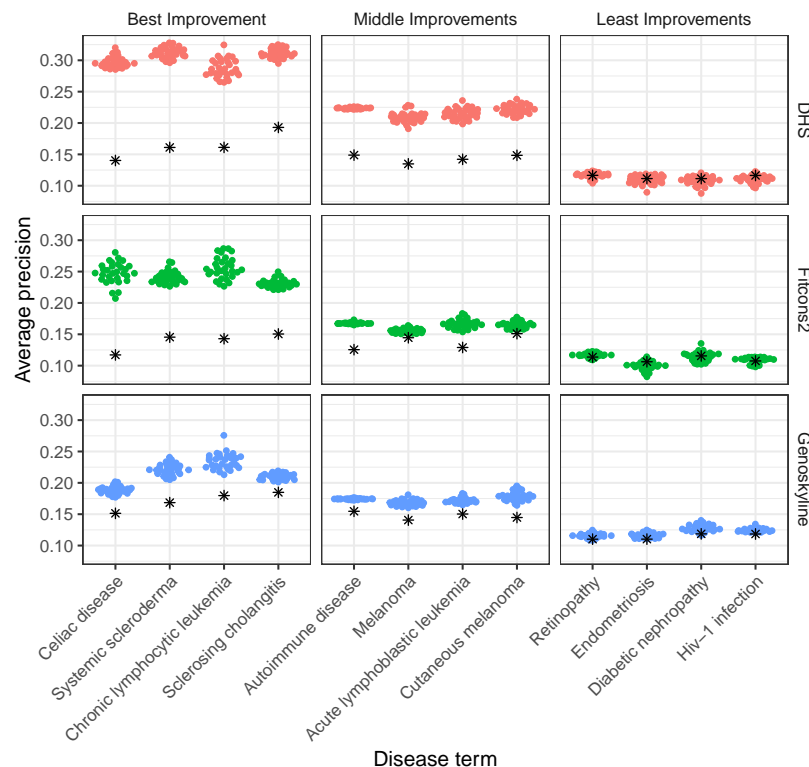


Figure 4: *Improvement through disease-specific tissue weights is consistent across scores but varies with disease term.* Shown is the performance of tissue-weighted variant scores (colored points) vs. tissue-mean (black asterisks) as a baseline, for three tissue scores (rows) and four diseases, stratified by improvement observed: best improvement for the first column, moderate improvement for the middle column, and least improvement for the right column. X-axes denote disease terms, the y-axis average precision. Different points for tissue-weighted scores represent different data-splits in the nested cross validation procedure.

## 178 2.4 DNase I hypersensitivity (DHS) scoring performs well compared with 179 DIVAN

180 Here we compare the performance of tissue-weighted DHS scoring with DIVAN [19], a disease-specific  
181 variant score for 45 diseases. DIVAN is based on a more complicated feature-selection and ensemble-  
182 learning framework, and it uses a variety of other functional genomics features, in addition to DNase I  
183 hypersensitivity. To compare our method with DIVAN, we mapped EFO disease terms to MeSH terms  
184 (as used by DIVAN) and use MeSH terms for this section (See **Suppl. Data SD12**). Because DIVAN  
185 uses a supervised learning approach, and because the published model was trained using GWAS SNVs,  
186 it was necessary to create specific train and test datasets to ensure a meaningful comparison between  
187 tissue-weighted DHS and DIVAN.

188 Therefore, to assess performance of both DIVAN and DHS, we created a test set of disease-  
189 associated variants (and their matched controls) that were published later than 2016 (DIVAN's pub-  
190 lication date). That is, these variants are unlikely to have been a part of DIVAN's training data. We  
191 also created a training set for DHS tissue-weighted containing only SNVs published prior to 2016. This  
192 resulted in training data that (a) is distinct from the test set and (b) draws on similar information that  
193 was available for DIVAN's training. Further on, we only selected disease terms for this training/test



Figure 5: *DHS disease-specific tissue weights improve variant prioritization compared with organism-level scores.* For four strata (best improvement, middle improvement, comparable performance, and worse performance) we selected four disease terms and compared performance results. GenoCanyon (best organism-level score) performance is denoted in black, DHS tissue-weighted in red. Different performances of DHS tissue-weighted represent variation different data splits during nested cross validation (see **Methods**).

194 data combination where at least 20 term-associated SNVs were present in the training data, and where  
 195 at least 50 SNVs were present in the test data. This approach yielded 29 disease terms for this analysis.  
 196 We then re-trained tissue-weighted DHS on this training data and compared with DIVAN on the test  
 197 data. In addition, we added the organism-level GenoCanyon score as a reference.

198 To assess performance, we performed all pairwise comparisons for each disease term, and evaluated  
 199 performance based on average precision. **Tab. 4** summarizes observations, where we find that DHS per-  
 200 forms significantly better than GenoCanyon and DIVAN in a majority of comparisons; however, there  
 201 is a substantial number of comparisons (22 out of 58) where either GenoCanyon or DIVAN outperform  
 202 DHS. **Fig. 6** further illustrates these comparisons. In panel **A** we show performance across disease  
 203 terms, grouped by the best-performing method. We see that tissue-weighted DHS outperforms DIVAN  
 204 and GenoCanyon substantially on Multiple Sclerosis (MeSH:D009103), Psoriasis (MeSH:D011565) and  
 205 Inflammatory Bowel Disease (MeSH:D015212); DIVAN outperforms GenoCanyon and DHS on Arthri-  
 206 tis, rheumatoid (MeSH:D001172) and Heart failure (MeSH:D006333); GenoCanyon outperforms DHS  
 207 and DIVAN on Stroke (MeSH:D020521) and Alzheimer disease (MeSH:D000544). In panels **B-D**  
 208 we directly summarize comparison results; we observe that the DHS tissue-weighted score often has  
 209 an advantage in terms where prioritization efforts are overall more successful (upper right quadrants).  
 210 Finding overall good performance for our approach, we next more closely examined the disease-specific  
 211 tissue aggregation weights we derive with our approach.

## 212 2.5 Disease-specific tissue weights reflect biomedical relevance

213 In addition to prioritizing SNPs, we can interpret the disease-specific tissue weights that our model  
 214 learns in the context of disease mechanisms. Specifically, large tissue weights implicate tissues with a  
 215 prominent role in associating SNVs with a disease in our model; therefore, one may hypothesize that  
 216 such tissues or cell-types have a function in the etiology of that disease. To investigate this hypothesis,  
 217 we analyzed tissue weights of the top-performing models we derived, where each model represents a

Score	Wins	Losses	Ties	Winning percent
DHS	34	22	2	61
GenoCanyon	26	31	1	46
DIVAN	25	32	1	44

Table 4: *DHS tissue-weighted disease-specific scoring outperforms DIVAN*. Across 29 disease terms, this table summarizes all pairwise comparison for DHS tissue-weighted, GenoCanyon and DIVAN using a specifically created test dataset. Wins, losses, and ties refer to significantly better (or worse, or tied) performance. Average precision was used as the performance metric, and the Wilcoxon signed-ranks test to determine wins and losses (p-values less than 0.05 were ties). Winning percent =  $\#Wins/(\#Wins+\#Losses)$

218 different disease.

219 Results are summarized in **Tab. 5**; they include the two top-performing models, Systemic sclero-  
220 derma (rank 1) and Sclerosing cholangitis (rank 2). In order to report a diverse range of diseases,  
221 we next excluded any diseases that are descendants of immune system disease (EFO:0000540) or lym-  
222 phoma (EFO:0000574). From the remaining diseases, we identify the next three highest-ranked models:  
223 Colorectal adenoma (rank 15), Atrial fibrillation (rank 20), and Cutaneous melanoma (rank 21). For  
224 each diseases, we list the five tissues with the largest tissue-weights, and their tissue group.

225 The tissues we associate with disease, overall, appear reasonable and generally are in-line with  
226 existing knowledge about disease mechanisms. Systemic scleroderma is an autoimmune disorder that  
227 can affect skin and internal organs [24]. We find that GM12878 lymphoblastoid cells (a type of B cell)  
228 are among highest-weighted tissues, as were other types of B cells (primary B cell and B cell lymphoma,  
229 respectively). This in-line with previous studies that have shown that B cells play a role in system  
230 scleroderma [25, 26]. Sclerosing cholangitis is an inflammatory condition that leads to scarring and  
231 narrowing of the bile ducts [27]. We highlight various inflammation-related types of blood cells, such  
232 as T cells and monocytes, which were previously suggested to play a role in the disease [28]. Colorectal  
233 adenoma is a benign tumor that develops in the lining of the colon or rectum. Our model identified  
234 rectal mucosa and stomach mucosa as the most-highly weighted tissues, and the function of rectal  
235 mucosa in colorectal cancer has been previously studied [29]. While the direct relationship between  
236 other gastrointestinal tissues and the development of colorectal adenoma has not been established,  
237 the association between gastrointestinal microbiome and colorectal adenomas has been discovered [30].  
238 Regarding atrial fibrillation, our approach highlights fetal heart and lung tissues. In addition, we  
239 identified skeletal muscle cells. In the case of cutaneous melanoma, a type of skin cancer, our approach  
240 emphasizes foreskin melanocyte cells and a specific type of T cell. Apart from these, we highlight  
241 cervical carcinoma cell lines and endothelial primary cells.

242 Overall, we conclude that the tissue weights we derive carry biomedically meaningful information  
243 and are able to highlight tissue contexts that may play a role in disease etiology. To further explore  
244 this finding, we used a resource of the epimap consortium [15], where disease-tissue associations are  
245 reported that derived differently from the one we obtained in two key ways: First, epimap uses their  
246 enhancer definitions based on a much larger set of genome annotations. Second, epimap’s enrichment  
247 test contrasts disease-associated SNP enrichment in a specific tissue’s enhancer set compared to all  
248 enhancers, whereas our method effectively compares open chromatin harboring disease-associated SNPs  
249 vs control SNPs tissue-by-tissue. Nevertheless, results are summarized in **Suppl. Data ST7**, and we  
250 find that out of the 25 tissues we associate with disease terms 14 have an estimated false discovery  
251 rate of less than 4% in the epimap analysis as well. Notably, a ground truth for these association is

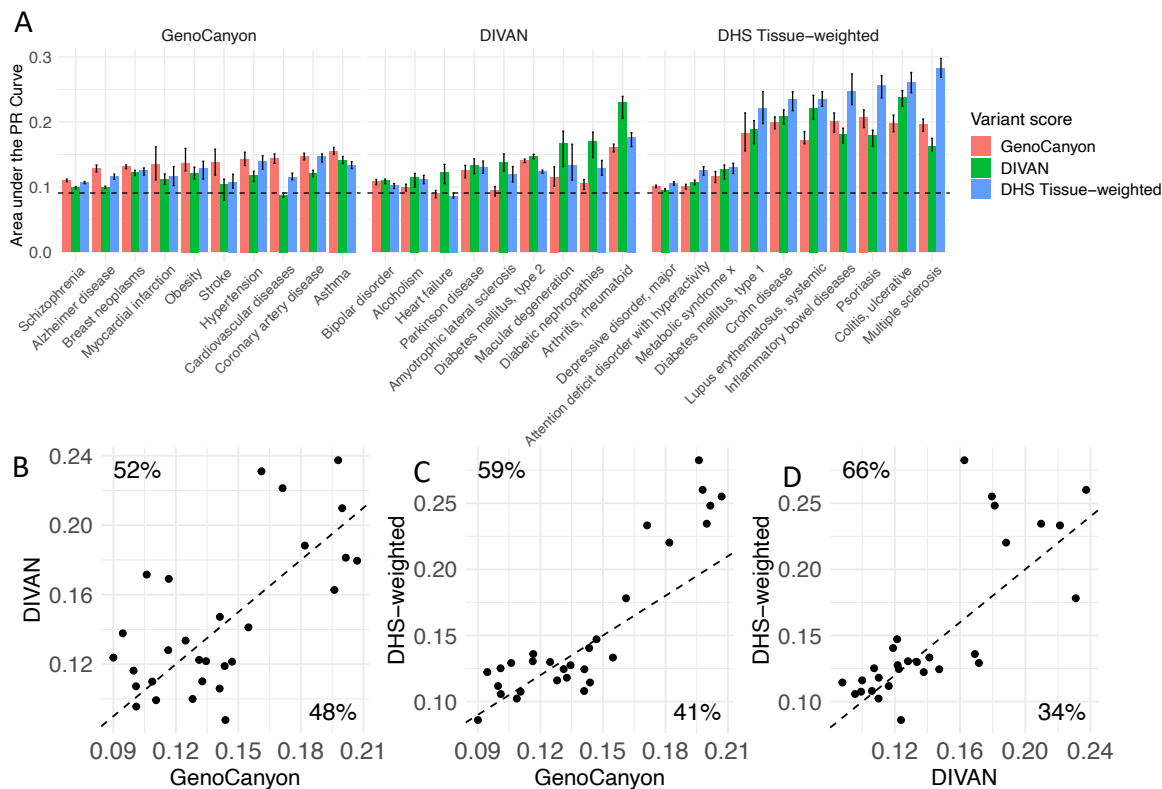


Figure 6: *DHS tissue-weighted scoring outperforms DIVAN*. Performance of DIVAN, GenoCanyon, and DHS tissue-weighted across a test set, with disease terms grouped by the best-performing method. Vertical striped indicates the minimum and maximum performance of 30 bootstrap samples (A). Performance scatter plots of GenoCanyon vs. DIVAN performance (B); GenoCanyon vs. DHS-weighted (C); DIVAN vs. DHS-weighted performance (D). Average precision was used for these plots; dashed lines denote equal performance. Percentages denote the fraction of points above and below the diagonal, respectively.

252 generally unknown; but we interpret the overlap in associations as encouraging, while complementary  
 253 associations are expected, given the differences in methodology. Based on this overall finding of  
 254 meaningful disease-tissue associations, we next further explored the use of tissue-weights in disease  
 255 characterization.

## 256 2.6 Disease-term similarity based on DHS tissue-weighted modeling reveals 257 meaningful groups

258 Disease-specific tissue weights for aggregating DHS scores, which are learned by our approach, can  
 259 highlight tissues and cell-types with a role in the disease (see previous section). Therefore, we derived  
 260 and explored a measure for disease similarity based on these weights.

### 261 2.6.1 Disease similarities based on disease-specific tissue weights for non-coding variant 262 prioritization

263 In our DHS tissue-weighted approach, for each disease term DNA accessibility across the same set of  
 264 tissue and cell-type contexts is used to predict whether a certain SNV is disease-associated, or not.  
 265 This results in disease-specific tissue aggregation weights (that is, coefficients in our logistic regression

Rank	ID	Tissue name	Group
Systemic scleroderma			
1	E116	GM12878 Lymphoblastoid Cells	blood
2	E032	Primary B cells from peripheral blood	blood
3	E041	Primary T helper cells PMA-I stimulated	blood
4	E123	K562 Leukemia Cells	blood
5	E030	Primary neutrophils from peripheral blood	blood
Sclerosing cholangitis			
1	E116	GM12878 Lymphoblastoid Cells	blood
2	E061	Foreskin Melanocyte Primary Cells skin03	skin
3	E102	Rectal Mucosa Donor 31	gi_rectum
4	E041	Primary T helper cells PMA-I stimulated	blood
5	E029	Primary monocytes from peripheral blood	blood
Colorectal adenoma			
1	E102	Rectal Mucosa Donor 31	gi_rectum
2	E110	Stomach Mucosa	gi_stomach
3	E057	Foreskin Keratinocyte Primary Cells skin02	skin
4	E101	Rectal Mucosa Donor 29	gi_rectum
5	E028	Breast variant Human Mammary Epithelial Cells (vHMEC)	breast
Atrial fibrillation			
1	E083	Fetal Heart	heart
2	E108	Skeletal Muscle Female	muscle
3	E107	Skeletal Muscle Male	muscle
4	E088	Fetal Lung	lung
5	E120	HSMM Skeletal Muscle Myoblasts Cells	muscle
Cutaneous melanoma			
1	E061	Foreskin Melanocyte Primary Cells skin03	skin
2	E059	Foreskin Melanocyte Primary Cells skin01	skin
3	E117	HeLa-S3 Cervical Carcinoma Cell Line	cervix
4	E041	Primary T helper cells PMA-I stimulated	blood
5	E122	HUVEC Umbilical Vein Endothelial Primary Cells	vascular

Table 5: *Top-ranked tissues for five diseases.* For five diseases when show the top-five tissues with the largest tissue weights in the corresponding model we derive. The first column is the tissue rank, the second the tissue’s roadmap ID, the third the tissue name, the fourth the tissue group, and the fifth listst the adjusted p-value in an enrichment analysis performed by epimap [15].

model)  $\{\beta^{(i)} \in \mathbb{R}^d\}_{i=1}^n$ , where  $i$  is indexing disease terms,  $n$  is the number of disease terms studied, and  $d$  denotes the number of tissues/cell-types with DHS scores. For our similarity measure between two diseases, say  $i$  and  $j$ , we then use a version of the Pearson correlation between  $\beta^{(i)}$  and  $\beta^{(j)}$  that takes uncertainty in the estimated aggregation weights into account (see **Methods**). That is, if an overlapping set of tissues/cell-types drive the prioritization of SNVs for two diseases, similarity is high; if different tissues are used, similarity is low.

Using this approach we calculated disease similarities for the 111 disease terms we study. Resulting similarities are visualized in (**Fig. 7**), where we show a similarity-based two-dimensional UMAP projection of disease terms. We observe that disease terms segregate into separate groups, with a coarse grouping between immune related diseases (lower left inlay, black) and others (lower left inlay, gray). A higher-resolution group structure was obtained by sub-clustering, where we grouped disease terms into seven groups (main panel, **Fig. 7**). Clusters names are based on EFO disease terms that include a large amount of cluster members as child-terms (see **Methods** and **Suppl. Fig. S10-S16**); **Tab. 6** lists disease terms per cluster. In addition to the clear separation of immune-related diseases from others, we also find a very homogeneous group consisting of mental and behavioural disorders, containing terms like schizophrenia (EFO:0000692) and anxiety disorder (EFO:0006788), and a group of skin cancers. The remaining three groups are more heterogeneous, but with two of them containing several terms related to cardiovascular disease (EFO:0000319) and digestive system disorders (EFO:1000218), respectively. By design similar tissues in each group drive SNP-disease associations, and we next examined which tissues play a role in each of the clusters.

In order to find group-specific tissues, we examined for each cluster the top five tissues that (a)

heterogenous	digest/cancer	immune	cardiovascular/others
<ul style="list-style-type: none"> <li>■ heterogenous</li> <li>adolescent idiopathic scoliosis</li> <li>age-related macular degeneration</li> <li>■ alcohol dependence</li> <li>amyotrophic lateral sclerosis</li> <li>chronic obstructive pulmonary disease</li> <li>■ dental caries</li> <li>diabetic nephropathy</li> <li>■ drug dependence</li> <li>□ endometriosis</li> <li>epilepsy</li> <li>gout</li> <li>hiv infection</li> <li>hiv-1 infection</li> <li>■ lung adenocarcinoma</li> <li>■ lung carcinoma</li> <li>neuropathy</li> <li>■ non-alcoholic fatty liver disease</li> <li>■ non-small cell lung carcinoma</li> <li>obesity</li> <li>■ periodontitis</li> <li>peripheral neuropathy</li> <li>scoliosis</li> <li>■ squamous cell lung carcinoma</li> <li>■ venous thromboembolism</li> </ul>	<ul style="list-style-type: none"> <li>■ digest/cancer</li> <li>■ autoimmune thyroid disease</li> <li>■ breast carcinoma</li> <li>■ cancer</li> <li>■ cardiovascular disease</li> <li>■ colorectal adenoma</li> <li>■ colorectal cancer</li> <li>■ coronary artery disease</li> <li>■ diabetes mellitus</li> <li>■ digestive system carcinoma</li> <li>■ digestive system disease</li> <li>■ female reproductive system disease</li> <li>■ hypertension</li> <li>■ multiple myeloma</li> <li>■ neurotic disorder</li> <li>■ pancreatic carcinoma</li> <li>■ prostate carcinoma</li> <li>■ respiratory system disease</li> <li>■ squamous cell carcinoma.</li> <li>■ type i diabetes mellitus</li> <li>■ type ii diabetes mellitus</li> </ul>	<ul style="list-style-type: none"> <li>■ immune</li> <li>■ acute lymphoblastic leukemia</li> <li>adult onset asthma</li> <li>■ allergic rhinitis</li> <li>■ allergy</li> <li>■ atopic asthma</li> <li>■ celiac disease</li> <li>■ childhood onset asthma</li> <li>■ chronic lymphocytic leukemia</li> <li>■ cirrhosis of liver</li> <li>■ hypothyroidism</li> <li>■ juvenile idiopathic arthritis</li> <li>■ lymphoid leukemia</li> <li>■ lymphoma</li> <li>■ neoplasm of mature b-cells</li> <li>■ non-hodgkins lymphoma</li> <li>■ systemic lupus erythematosus</li> <li>■ systemic sclerosis</li> </ul>	<ul style="list-style-type: none"> <li>■ cardiovascular/others</li> <li>■ alzheimer's disease</li> <li>■ atherosclerosis</li> <li>■ atrial fibrillation</li> <li>■ cardiac arrhythmia</li> <li>■ chronic kidney disease</li> <li>■ diverticular disease</li> <li>■ glaucoma</li> <li>■ heart failure</li> <li>■ metabolic syndrome</li> <li>■ migraine disorder</li> <li>■ osteoarthritis</li> <li>■ ovarian carcinoma</li> <li>■ parkinson's disease</li> <li>■ peripheral arterial disease</li> <li>■ retinopathy</li> <li>■ stroke</li> <li>■ uterine fibroid</li> </ul>
<ul style="list-style-type: none"> <li>■ immune/autoimmune</li> <li>■ ankylosing spondylitis</li> <li>■ asthma</li> <li>■ autoimmune disease</li> <li>■ crohn's disease</li> <li>■ hypersensitivity reaction disease</li> <li>■ immune system disease</li> <li>■ inflammatory bowel disease</li> <li>■ kidney disease</li> <li>■ liver disease</li> <li>■ multiple sclerosis</li> <li>■ psoriasis</li> <li>■ rheumatoid arthritis</li> <li>■ sclerosing cholangitis</li> <li>■ skin disease</li> <li>■ ulcerative colitis</li> </ul>	<ul style="list-style-type: none"> <li>■ mental</li> <li>■ anorexia nervosa</li> <li>■ anxiety disorder</li> <li>■ attention deficit hyperactivity disorder</li> <li>■ autism spectrum disorder</li> <li>■ bipolar disorder</li> <li>■ eating disorder</li> <li>■ mental or behavioural disorder</li> <li>■ mood disorder</li> <li>■ movement disorder</li> <li>■ obsessive-compulsive disorder</li> <li>■ psychosis</li> <li>■ schizophrenia</li> <li>■ tourette syndrome</li> <li>■ unipolar depression</li> </ul>	<ul style="list-style-type: none"> <li>■ skin cancer</li> <li>■ cutaneous melanoma</li> <li>■ keratinocyte carcinoma</li> <li>■ melanoma</li> <li>■ non-melanoma skin carcinoma</li> </ul>	<ul style="list-style-type: none"> <li>■ legend</li> <li>■ digestive system disease</li> <li>■ immune system disease</li> <li>■ autoimmune disease</li> <li>■ cardiovascular</li> <li>■ mental or behavioural disorder</li> <li>■ skin cancer</li> <li>■ cancer</li> </ul>

Table 6: *Disease groups based on DHS tissue-weights*. For each disease group disease terms are shown. The colored squares denote the disease groups in the EFO ontology.

287 contribute most to disease association and (b) are cluster specific (see **Methods**). Results are summa-  
288 rized in **Fig. 8**; we note that both disease groups related to the immune system highlight blood tissues  
289 (such as E043: Primary T helper cells from peripheral blood and E116: GM12878 Lymphoblastoid  
290 Cells, see **Suppl. Data** SD23 for all names of standard epigenomes), with the group containing in-  
291 flammatory bowel disease, Crohn's disease, and ulcerative colitis also containing rectum tissues (such  
292 as E101: Rectal Mucosa Donor 29). Brain tissues contribute to disease associations for mental and  
293 behavioral disorders, skin tissues to skin cancer, and gastro-intestinal / stomach tissue to the cluster  
294 with digestive system diseases. We also note that a clear association of specific tissues with a dis-  
295 ease group correlates with better classification performance of our model for SNP-disease association  
296 (**Fig. 8**; for example, see the immune and immune/autoimmune clusters). We note, though, that  
297 not for all clusters the corresponding tissue associations are equally compelling, as illustrated in the  
298 same figure. While the clusters we derive resemble broader disease groups, for each disease a specific  
299 combination of tissues is used to derive whether a variant might be associated, and some tissues con-  
300 tribute to several clusters. For instance, one blood cell type (E116, GM12878 Lymphoblastoid Cells)  
301 contributes to both immune clusters, but also to diseases in the digestive/cancer, heterogeneous and  
302 skin cancer clusters. Another blood cell type (E043, Primary T helper cells from peripheral blood)  
303 displays a similar pattern. **Suppl. Fig. S9** shows the same heatmap as **Fig. 8**, but for all tissues.  
304 Overall, these results suggest that our modeling approach successfully identifies tissues with a role  
305 in disease etiology. Finally, we explore how our disease similarities relate to genetic similarities as  
306 measured by genetic correlation between diseases.

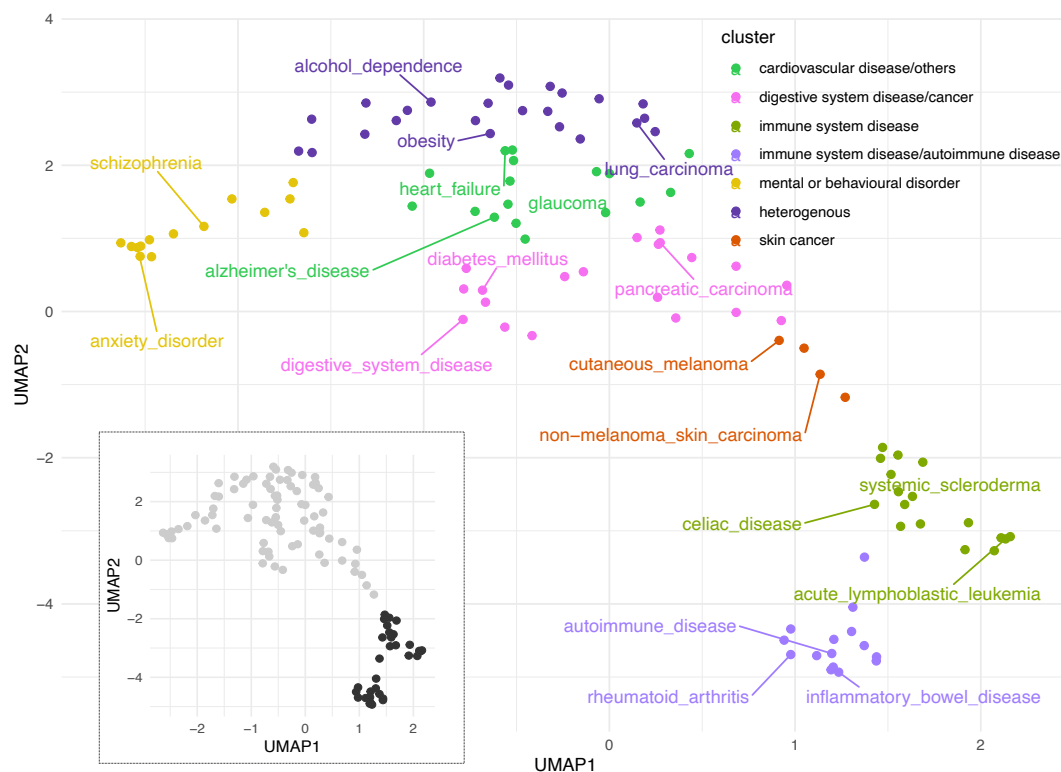


Figure 7: *Similarity-based two-dimensional projection visualizes 111 diseases.* Two dominant disease groups emerge in this visualization (immune system related disease terms (black) and others (gray), in the inlay). Hierarchical clustering was used to group diseases into seven clusters, with colors indicating broad disease types (see **Tab. 6** for details).

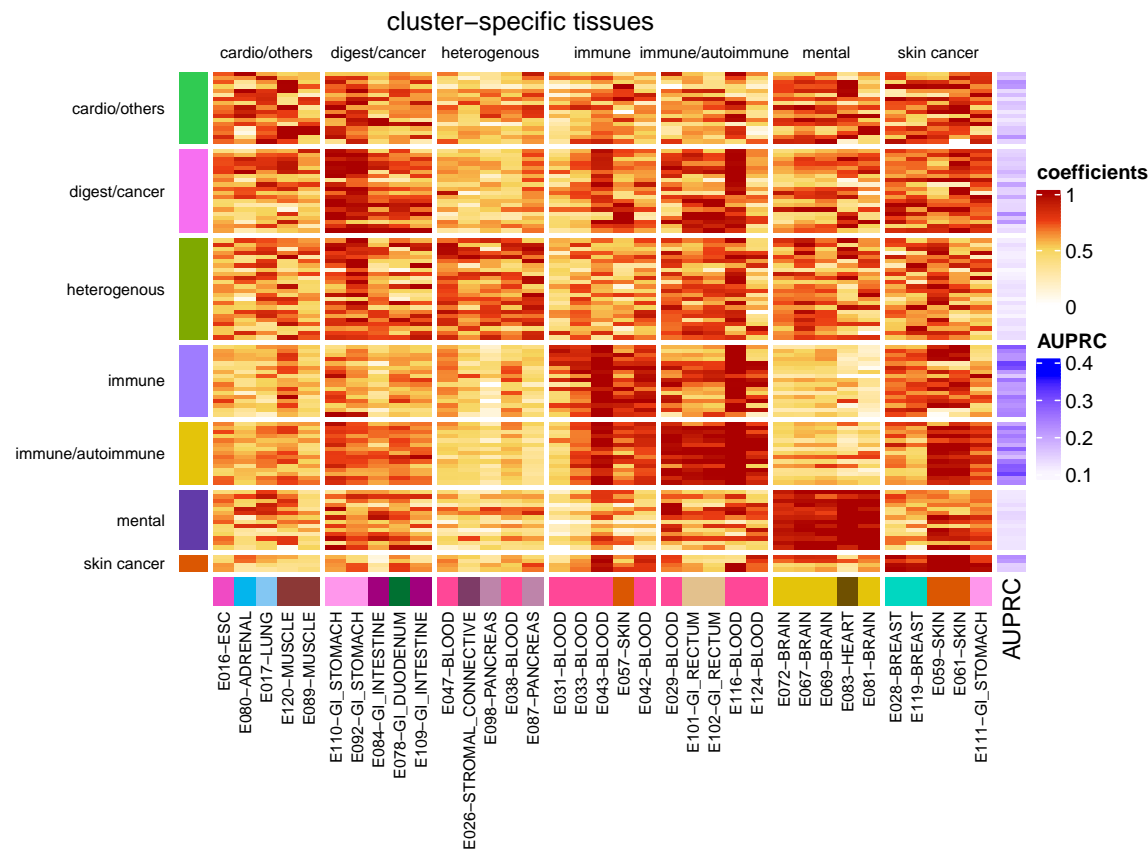


Figure 8: **Heatmap of top-five tissue-weights for 111 diseases.** Regularized model coefficients (i.e., tissue weights) of five disease-cluster-specific tissues (columns) are shown for 111 diseases (rows). Coefficients are scaled by disease, and rows are grouped into sets of cluster-specific tissues (see **Methods** section). Bottom annotation shows tissue names of cluster-specific tissues (names are shown in the format of ‘Tissue name’ - ‘Tissue group’); annotation on the left side shows disease cluster, and annotating on the right side shows model performance in terms of AUPRC).

## 307 2.6.2 Model-based similarities are complementary to genetic correlation.

308 Here we compare the disease-disease similarities we derived ( $s_m$ ) with genetic correlations from the  
 309 GWAS Atlas ( $s_g$ ), where genetic correlation measures shared genetic causes between two traits [31].  
 310 For 6,105 possible disease pairs of the 111 diseases terms we study, estimates of genetic correlation for  
 311 595 pairs were available from the GWAS Atlas (see **Methods**). Overall, for these 595 disease pairs  
 312 we observe only weak (but statistically significant) correlation between model similarities and genetic  
 313 correlations ( $r = 0.32$ ,  $p \text{ value} = 2.4E - 15$ ), where the scatter plot is shown in **Fig. 9** panel A.

314 We also see that most disease pairs are not annotated with substantial genetic correlations, or  
 315 with high model-based similarities (90% of disease pairs have  $s_m < 0.25$ , and  $s_g < 0.2$ ). Therefore, we  
 316 explored three different regimes: Disease pairs where both similarity measures are high ( $s_m \geq 0.25$  and  
 317  $s_g \geq 0.20$ ), pairs with high genetic correlations and low model similarity ( $s_m < 0.25$  and  $s_g \geq 0.20$ )  
 318 vice versa (quadrants indicated in **Fig. 9A**, named quadrants B, C and D). The top eight most  
 319 extreme examples from each regime are summarized in **Tab. 7**. In the following we discuss some  
 320 examples in more detail. Specifically we explore two immune system diseases for quadrant B; two  
 321 mental or behavioral disorders for quadrant C; and one immune system disease and one mental or



322 behavioral disorder for quadrant D. We note that the pairs we examine have no annotated parent-  
323 child relationships in the EFO.

324 - Ulcerative colitis (UC, EFO:0000729) and Crohn’s disease (CD, EFO:0000384) have both high  
325 genetic correlation ( $s_g = 0.53$ ) and model similarity ( $s_m = 0.84$ ), see **Fig. 9A**. This suggests  
326 that they share genetic causes, and that the same tissues are informative for SNP-disease as-  
327 sociation. While shared genetic causes for UC and CD have been pointed out (e.g., [32]), our  
328 model for SNP-disease association allows us to explore relevant tissue contexts. In **Fig. 9B** we  
329 show a scatter plot of tissue weights for both diseases, where color indicates the importance of  
330 each tissue to model similarity (see **Methods**). We observe that open chromatin in blood (E116,  
331 GM12878 Lymphoblastoid Cells; E124, Monocytes-CD14+ RO01746 Primary Cells; E041, Pri-  
332 mary T helper cells PMA-I stimulated) and rectum (E102, Rectal Mucosa Donor 31) is positively  
333 associated with SNP-disease association in both diseases; this is consistent with a previous study  
334 where blood cell types are found to be relevant in many autoimmune diseases, including UC and  
335 CD [33]. In addition, symptoms or complications in rectum is also observed in UC and CD [34].  
336 Interestingly, open chromatin in GI-intestine (E085, fetal intestine small) is negatively associated  
337 with SNP-disease association, along with other intestine tissues (E084, fetal intestine large and  
338 E109, small intestine, with the 61th and 86th smallest tissue weight, respectively, amongst 127  
339 contexts). This indicates fetal intestine or small intestine might be less involved in UC and CD  
340 etiology, compared to their juvenile and adult counterparts.

341 - Autism spectrum disorder (ASD, EFO:0003756) and anorexia nervosa(AN, EFO:0004215)] is an  
342 example where we observe a low genetic correlation ( $s_g = -0.05$ ) and a moderate high model  
343 similarity ( $s_m = 0.34$ ); a scatter plot of their tissue weights is shown in **Fig. 9C**. Note that we did  
344 not choose one of the highlighted pairs in **Tab. 7** for this quadrant, because we already discussed  
345 a immunessystem realted disease pair. We observe that both disease models give heart and brain  
346 tissue (E083, fetal heart and E081, fetal brain male) high tissue weights. This is consistent with  
347 the observation of brain abnormalities in ASD and AN [35, 36]. While the presence of fetal  
348 heart is less intuitive, we note that children with abnormal heart development are more likely  
349 to develop ASD, suggesting a connection between the disease and the fetal heart [37]. We also  
350 note that while genetic correlation between ASD and AN is low, a link between the two diseases  
351 on the phenotypic level is being suggested [38, 39]; the tissue context we identified could provide  
352 information about shared molecular aspects of disease etiology as well.

353 - For obsessive compulsive disorder (EFO:0004242) and celiac disease (EFO:0001060) we observe  
354 low model similarities ( $s_m = -0.26$ ) and moderately high genetic correlation( $s_g = 0.36$ ); **Fig. 9**  
355 **D** shows the scatter plot of tissue weights. Several studies have shown that nervous system disease  
356 and immune related diseases have shared genetic background [40, 41]. However, in contrast to the  
357 other two examples, there is little relation between tissue weights in these two diseases. Blood  
358 cell types are highlighted in celiac disease, while brain and fetal heart tissues are highlighted  
359 in obsessive compulsive disorder. For celiac disease, the top six tissue contexts are blood cells,  
360 including different types of T cells (E041, Primary T helper cells PMA-I stimulated; E043,  
361 Primary T helper cells from peripheral blood and E034, Primary T cells from peripheral blood)  
362 and lymphoblasts (E116, GM12878 Lymphoblastoid Cells), which is consistent with findings that  
363 alterations in T cells and lymphoblasts can lead to celiac disease [42, 43].

364 Overall, these examples illustrate that the disease similarities we derive are complementary to  
365 genetic correlation. In addition, tissue contexts highlighted by our tissue-weights allow for biomedical

366 interpretations of observed similarities (i.e., which are the relevant tissue contexts) and can be used to  
 367 generate molecular hypotheses about disease etiology.

368 In summary, our results show that disease-specific variant prioritization performs well for non-coding  
 369 GWAS variants, compared with organism-level approaches. We also demonstrate that disease-specific  
 370 tissue-weights are biomedically meaningful and can be used to generate hypotheses about disease  
 371 mechanism. Therefore, we believe this type of variant characterization is a useful tool for researchers  
 372 studying the molecular and genetic causes of disease.

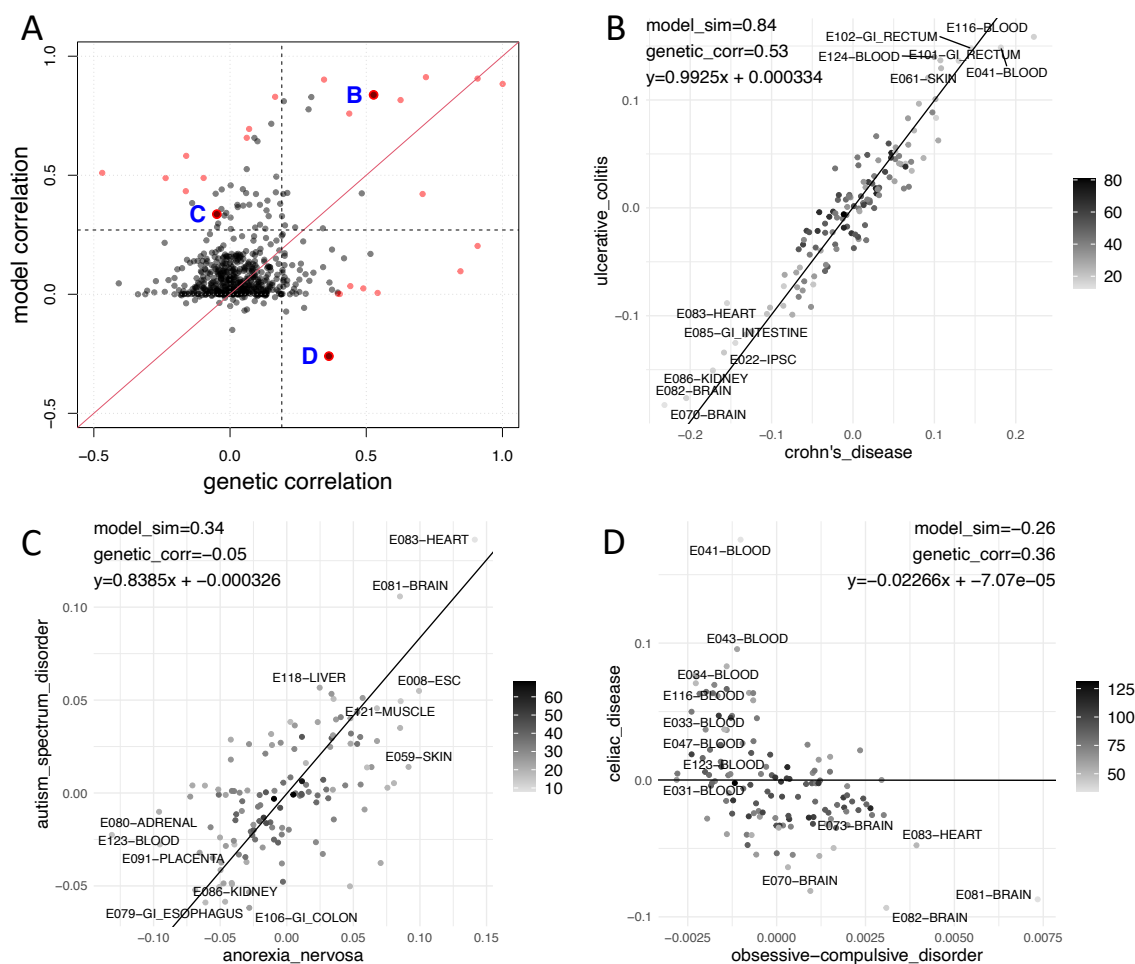


Figure 9: **Genetic correlation and model similarity.** (A) Genetic correlation vs. model similarity for 595 disease pairs. Each point is a disease pair, where the x-axis denotes the genetic correlation and y-axis is the disease model similarity. For three quadrants we highlight disease pairs, denoted by B, C, and D). (B-D) Scatter plot of tissue coefficients in three example disease pairs, where (B) shows Crohn's disease vs inflammatory bowel disease; (C) shows anorexia nervosa vs autism spectrum disorder and (D) shows celiac disease vs obsessive compulsive disorder. Lines denote a weighted linear regression line underlying our disease similarities. Color codes for the weight for each tissue when conducting weighted regression analysis.

Disease 1	Disease 2	$s_g$	$s_m$	Quadrant
Inflammatory bowel disease	Ulcerative colitis	1.00	0.88	B
Diabetes mellitus	Type ii diabetes mellitus	0.91	0.91	B
Crohn's disease	Inflammatory bowel disease	0.72	0.91	B
Sclerosing cholangitis	Ulcerative colitis	0.63	0.82	B
Crohn's disease	Ulcerative colitis	0.53	0.84	B
Ankylosing spondylitis	Sclerosing cholangitis	0.35	0.90	B
Inflammatory bowel disease	Sclerosing cholangitis	0.44	0.76	B
Bipolar disorder	Schizophrenia	0.71	0.42	B
Rheumatoid arthritis	Systemic lupus erythematosus	-0.47	0.51	C
Celiac disease	Systemic lupus erythematosus	-0.16	0.58	C
Sclerosing cholangitis	Systemic lupus erythematosus	-0.24	0.49	C
Crohn's disease	Sclerosing cholangitis	0.17	0.83	C
Rheumatoid arthritis	Sclerosing cholangitis	0.07	0.69	C
Crohn's disease	Rheumatoid arthritis	0.06	0.66	C
Systemic lupus erythematosus	Ulcerative colitis	-0.16	0.43	C
Crohn's disease	Systemic lupus erythematosus	-0.10	0.49	C
Type i diabetes mellitus	Type ii diabetes mellitus	0.85	0.10	D
Diabetes mellitus	Type i diabetes mellitus	0.91	0.20	D
Celiac disease	Obsessive-compulsive disorder	0.36	-0.26	D
Diabetes mellitus	Obesity	0.54	0.01	D
Obesity	Osteoarthritis	0.49	0.02	D
Attention deficit hyperactivity disorder	Obesity	0.44	0.03	D
Attention deficit hyperactivity disorder	Osteoarthritis	0.40	0.00	D
Obesity	Type i diabetes mellitus	0.40	0.00	D

Table 7: *Example disease pairs of genetic correlation and model similarities.* This table shows the genetic correlation and model similarity for some disease pairs as we selected.  $s_g$ : genetic correlation;  $s_m$ : model similarity. For quadrant B, C, D we pick 8 disease pairs, where  $s_g + s_m$ ,  $s_g - s_m$  and  $s_m - s_g$  are the highest, respectively.

### 3 Discussion

373

374 Most variant scores prioritize non-coding variants either at the level of the whole organism (e.g, CADD  
 375 [8], GenoCanyon [44]), or they provide tissue-specific scores (e.g, GenoSkyline [11], Fitcons2 [12]). Here  
 376 we present a straightforward strategy to combine tissue-specific variant scores in a disease-specific  
 377 manner. We show that for common genetic variants in the GWAS catalog [1] our approach leads to  
 378 better performance than organism-level or tissue-specific scores (see **Fig. 5**). Pre-computed disease-  
 379 specific prioritization scores are available at <https://doi.org/10.7910/DVN/AUAJ7K>.

380 Comparing different variant prioritization methods we note that we use area under the precision-recall  
 381 curve as an evaluation metric, and that the performance of all methods is modest. We believe that  
 382 is because our analysis (**a**) focuses explicitly on non-coding variants, (**b**) stratifies SNVs by disease-  
 383 phenotype, and (**c**) utilizes unbiased matching of control-SNVs (SNPsnap-matching, see Section 4.1.2).  
 384 Each of these points affects the SNV sets we use for our analysis, and therefore the performance metrics  
 385 we report. For transparency we provide all disease-associated variants we use (with matched negatives)  
 386 in our supplemental data. As a more general point we also note that associations reported in the  
 387 GWAS catalog contain causal as well as non-causal SNPs, which will also contribute to sub-optimal  
 388 performance measures of all the variant scores we assess.

389 We included a comparison with the DIVAN method in our evaluation, which also includes compar-  
 390 ing GenoCanyon with DIVAN. Part of this comparison is analogous to results reported in Chen et al.  
 391 [19]; however, the performances we observed do not agree perfectly, as detailed in **Suppl. Data SD15**.  
 392 Broadly, looking at overlapping/matching disease terms, our results appear more favorable for Geno-  
 393 Canyon. These differences are likely due to different test sets used in the two evaluations (i.e., the  
 394 GWAS catalog (this study) vs. GRASP).

395 We also note that there is other research associating variants with disease terms in a similar  
 396 setting, notably PINES [20] and LSMM [45]. We did not compare directly with PINES, because no

397 pre-computed scores are available; also, we note that while performance reported in this publication  
398 in terms of AUROC is higher than our results, a less stringent un-matched test set of random/control  
399 variants was used in these analyses. For LSMM we note that we leverage variants associated with  
400 EFO disease terms across studies, while LSMM uses summary statistics on a per-study basis. Using  
401 aggregate data from different studies allows our approach to consider parent-child relationships of the  
402 EFO ontology using variant aggregation (see Section 2.1).

403 We demonstrate that our approach can be used to calculate similarities between disease terms, see  
404 Section 2.6.1. Since this similarity measure is derived from non-coding SNVs associated with disease,  
405 one could expect it is largely congruent with genetic correlation between disease traits. However, that  
406 is not the case (see **Fig. 9**), most likely because we focus on a small subset of disease-associated SNVs  
407 reported in the GWAS catalog. For example, obsessive-compulsive disorder and celiac disease have  
408 a high genetic correlation ( $s_g = 0.36$ ) but do not share noncoding SNPs in the GWAS catalog (and  
409 low model similarity  $s_m = -0.26$ ); on the other hand, autism spectrum disorder and anorexia nervosa  
410 have a low genetic correlation ( $s_g = -0.05$ ) but share a number of significant SNPs in the GWAS  
411 catalog (and relative high model similarity  $s_m = 0.34$ ). In addition, interpretation of model similarity  
412 between disease terms is different from genetic correlation; high model similarity implies that disease-  
413 associated SNVs reside in DNA-accessible regions in an overlapping set of tissues, but the identity of  
414 individual SNVs (and whether they overlap) is inconsequential. For example, asthma and rheumatoid  
415 arthritis have only 15 shared SNPs (out of 732 and 1283 SNPs in rheumatoid arthritis and asthma,  
416 respectively), but exhibit high model similarity ( $s_m = 0.53$ ). This shows that model similarity between  
417 two diseases can involve similar tissues even if they do not share a genetic background. Further on, we  
418 note that estimates of genetic correlation also may depend on the study used. For example, systemic  
419 lupus erythematosus (SLE) has a negative genetic correlation ( $s_g = -0.47$ ) with rheumatoid arthritis  
420 (RA) (and other inflammatory diseases) when using the SLE summary statistics from Julia et al. [46]  
421 (as retrieved from the GWAS Atlas [31]), whereas another study (Lu et al., [47]) found SLE to have a  
422 positive genetic correlation ( $s_g = 0.41$ ) with RA when using the SLE summary statistics from Bentham  
423 et al. [48].

424 We note that in our analyses we used the EFO ontology to aggregate variants annotated in the  
425 NIH/EBI GWAS catalog. That is, for each disease term directly-annotated variants were used, and,  
426 in addition, variants annotated to descendant terms in the ontology were also included. This approach  
427 allowed us to compile a more exhaustive set of variants per term. However, some amount of caution  
428 should be exercised when using disease models with more general terms, such as "cardiovascular  
429 disease" for example, as they may encompass heterogeneous diseases.

430 Our approach is expected to improve as more variants are associated with disease, and as disease-  
431 associations get more refined. In addition, increasing amounts of epigenomics data, such as epimaps  
432 [15] and ENCODE5 [6], could be incorporated and they have the potential to improve the disease  
433 associations we learn.

434 In summary, we have provided a straightforward method to leverage tissue-specific variant scores  
435 for disease-specific variant prioritization. We show that this approach performs well compared with  
436 current methods, and we show that the resulting association models are interpretable and lead to useful  
437 characterization of disease terms. Overall, our contributions are useful for the following two reasons:  
438 Conceptually, because they highlight the value of disease-specific variant prioritization. In addition,  
439 we provide pre-computed association scores for 111 disease terms that researchers can use in practice  
440 to interpret their variant data.

## 4 Methods

### 4.1 Data sources and processing

#### 4.1.1 Disease-associated variants

Disease-associated non-coding single nucleotide variants were retrieved from the NHGRI-EBI Catalog of human genome-wide association studies database (GWAS catalog, version 2020-12-02, downloaded from <https://www.ebi.ac.uk/gwas/docs/file-downloads>). These data contained 122,396 unique non-coding SNPs spanning 2,782 phenotypes, where non-coding was defined as variants not overlapping protein-coding sequence (GENCODEv36); we also excluded variants annotated as protein coding sequence variants (e.g. missense variants, frameshift variants) as a SNP's "functional class" in the GWAS Catalog. Further, variants in the GWAS Catalog are annotated with phenotypes using the Experimental Factor Ontology (EFO, <https://www.ebi.ac.uk/efo>) [49]. We focused on variants with phenotype terms annotated in disease domain of the EFO (i.e., all terms/traits/phenotypes we consider are descendants of the term "disease" (EFO:0000408, EFO version 3.24.0, accessed 2020-11-17). Further on, SNPs in the HLA region, and SNPs with minor allele frequency (MAF) less than 1% in the European population as reported by the International Genome Sample Resource were excluded (as they cannot be matched to control SNPs with the SNPsnap approach, see below). Out of 31103 SNVs, a total of 5225 SNVs were removed. Finally, in our analyses we restricted ourselves to phenotypes with at least 100 annotated non-coding SNPs. **Suppl. Data** SD1 and SD2 contain 111 phenotypes and 77,028 phenotype-associated SNPs we used in this study. We also grouped SNPs in LD blocks (SNPsnap,  $r^2 \geq 0.5$ ) and identify SNPs with the minimum p-value per block ("representative SNP"); we provide this information, which we use in some of the analyses described below, in **Suppl. Data** SD2.

#### 4.1.2 Control variants

For each disease-associated SNP we generated matched control non-coding variants (MAF  $\geq 1\%$ ) using four different strategies, where the non-coding is again defined as discussed above (**Section** 4.1.1). The four strategies are:

**Random** For each disease-associated SNP, we selected ten SNPs from common variants in 1000G EUR at random (i.e., equal probability for all SNPs) as controls.

**TSS-matching** We processed common non-coding SNVs and selected a subset of these variants as controls, where the distribution of distances to the nearest protein-coding gene's transcription start site (TSS) are matched between control set and disease-associated SNPs (similar to GWAVA, [23]). Specifically, we sorted all common non-coding SNPs by the distance to the nearest TSS and divided them into 50 bins, where each bin contains the same number of SNVs. Then, for each disease-associated SNP, we randomly selected ten control SNPs from the bin containing the disease-associated SNP's distance to the nearest gene.

**SNPsnap-matching** Using SNPsnap [22], we matched control SNPs to disease-associated variants in terms of minor allele frequency, gene density (distance cutoff 1d0.8), distance to the nearest gene TSS, and number of SNPs in LD. Our parameters for maximum allowable deviation were: 5%, 50%, 20% and 50%, respectively. We randomly selected ten control SNPs per disease-associated SNP from SNPsnap's results, and we ensured there are no duplicated control SNPs for different disease-associated SNPs. If there were less than 10 control SNPs returned by SNPsnap, we kept

481 all of the control SNPs. If no control SNPs were matched, we removed the disease-associated  
482 SNVs (a total of 311 SNVs) from our analyses.

483 **SNPsnap-TSS-matching** Essentially the same as in SNPsnap-matching, but controlling **only** for  
484 the distance to the nearest genes (maximum allowable deviation: 20%); for three other attributes  
485 “maximum allowable deviation” is set to 10,000%. We note that in both SNPsnap-matching and  
486 SNPsnap-TSS-matching, distance is measured by distance to the nearest gene, whereas for TSS-  
487 matching only protein-coding genes are considered.

488 In all four matching strategies we excluded variants annotated in the GWAS catalog as control SNPs.  
489 **Suppl. Data SD3** contains the four sets of control variants.

#### 490 4.1.3 Additional data sources, variant scores

491 We used pre-computed SNP annotations from the following sources:

- 492 - CADD v.1.3: [http://krishna.gs.washington.edu/download/CADD/v1.3/1000G\\_phase3.tsv.](http://krishna.gs.washington.edu/download/CADD/v1.3/1000G_phase3.tsv.gz)  
493 [gz](http://krishna.gs.washington.edu/download/CADD/v1.3/1000G_phase3.tsv.gz)
- 494 - EigenPC v.1.1: <https://xioniti01.u.hpc.mssm.edu/v1.1>
- 495 - Fitcons2: <http://compgen.cshl.edu/fitCons2/hg19>
- 496 - GenoCanyon: [http://genocanyon.med.yale.edu/GenoCanyon\\_Downloads.html](http://genocanyon.med.yale.edu/GenoCanyon_Downloads.html)
- 497 - GenoSkylinePlus: [http://genocanyon.med.yale.edu/GenoSkylineFiles/GenoSkylinePlus/](http://genocanyon.med.yale.edu/GenoSkylineFiles/GenoSkylinePlus/GenoSkylinePlus_bed.tar.gz)  
498 [GenoSkylinePlus\\_bed.tar.gz](http://genocanyon.med.yale.edu/GenoSkylineFiles/GenoSkylinePlus/GenoSkylinePlus_bed.tar.gz)
- 499 - GWAVA v.1.0: [ftp://ftp.sanger.ac.uk/pub/resources/software/gwava/v1.0/VEP\\_plugin/](ftp://ftp.sanger.ac.uk/pub/resources/software/gwava/v1.0/VEP_plugin/gwava_scores.bed.gz)  
500 [gwava\\_scores.bed.gz](ftp://ftp.sanger.ac.uk/pub/resources/software/gwava/v1.0/VEP_plugin/gwava_scores.bed.gz)
- 501 - LINSIGHT: <http://compgen.cshl.edu/%7Eyihuang/tracks/LINSIGHT.bw>
- 502 - DIVAN: <https://sites.google.com/site/emorydivan>
- 503 - DHS accessibility: We downloaded Avocado-imputed [50] DNase1 hypersensitive sites (DHS)  
504 signal for 127 ENCODE biological contexts (tissues / cell types) from [https://noble.gs.](https://noble.gs.washington.edu/proj/avocado/data/avocado_full/DNase/)  
505 [washington.edu/proj/avocado/data/avocado\\_full/DNase/](https://noble.gs.washington.edu/proj/avocado/data/avocado_full/DNase/).

## 506 4.2 Tissue-weighted variant prioritization based on DNase1 hypersensitiv- 507 ity

### 508 4.2.1 A penalized logistic regression model for context-weighted score averaging

509 For predicting SNP’s associations with a disease term, we consider SNPs as observations, and each  
510 SNP is described as a vector  $\mathbf{x} \in \mathbb{R}^d$  of variant scores in  $d$  tissues/contexts; we arrange vectors  $\{\mathbf{x}^i\}_{i=1}^n$   
511 for  $n$  observations in a matrix  $X \in \mathbb{R}^{n \times d}$ , together with a vector  $y$  of  $n$  binary entries, indicating for  
512 each SNP association with a specific disease term (no=0/yes=1). In addition, we denote the average  
513 score (across contexts) for a SNP  $i$  by  $\bar{x}^i$ , which is also a baseline score because it aggregates across  
514 contexts.

515 We use a logistic regression model of the form

$$\log \frac{p_i}{1 - p_i} = \alpha_0 + \alpha \bar{x}^i + \beta' \mathbf{x}^i \quad \text{s.t.} \quad \alpha \geq 0 \quad (1)$$

516 where  $\alpha_0 \in \mathbb{R}$ ,  $\alpha \in \mathbb{R}_+$  and  $\beta \in \mathbb{R}^d$  are regression coefficients, and  $p_i$  is the probability that SNP  $i$  is  
517 associated with a disease that is studied. We fit a regularized version of the negative log likelihood

$$\arg \min_{\alpha_0, \alpha, \beta} -\frac{1}{n} \sum_{i=1}^n \left[ \log(1 - p_i) + y_i \log \frac{p_i}{1 - p_i} \right] + \lambda \|\beta\|_2 \quad (2)$$

518 where the dependence on  $\alpha, \beta$  of the first term is through Equation (1). For large regularization  
519 parameters  $\lambda$  this will yield small  $\beta \rightarrow \mathbf{0}$  and recover the baseline ( $\bar{x}$ ) of unweighted averaging of  
520 context scores (scaled by a non-negative factor  $\alpha$ ). We implemented this approach using the R package  
521 glmnet (version 2.0-18, [51]) and determined the regularization parameter via 5-fold cross validation  
522 (cv.glmnet function) through maximizing the area under the (cross-validated) ROC curve. Class  
523 weights were employed to balance skewed class sizes.

## 524 4.2.2 Disease similarities from context-weighted score averaging

525 Context-weighted score averaging, as described above, results in disease-specific coefficient vectors  
526 ( $\{\beta^{(i)}\}$ , with  $i$  indexing disease terms), together with bootstrap estimates for the standard deviation  
527 of each coefficient (that can be arranged in corresponding vectors  $\{\gamma^{(i)}\}$ ). Specifically, we use 5-fold  
528 cross-validation repeated 10 times, yielding 50 coefficient vectors for each disease. We use their mean  
529 for our estimate of  $\beta^{(i)}$ , and their standard deviation as an estimate of  $\gamma^{(i)}$ .

For a pair of diseases ( $d_i, d_j$ ) we then define a disease similarity through similarity of associated  
coefficient vectors  $\beta^{(i)}$  and  $\beta^{(j)}$ , taking into account our estimates of coefficient variability. Specifically,  
we fit a weighted linear regression model (i.e., regressing  $\beta^{(i)}$  on  $\beta^{(j)}$ ), with regression weights taking  
into account coefficient variability as follows:

$$w_k^{(i,j)} = 1 / \sqrt{s_k^i s_k^j} \quad \text{and} \quad s_k^\circ = \alpha \gamma_k^{(\circ)} + (1 - \alpha)m \quad \text{for} \quad \circ \in \{i, j\},$$

530 where we chose  $m$  to be the 25% quantile of all (estimated) standard deviations observed, and  $\alpha = 3/4$ .  
531 Therefore,  $s_j^i$  and  $s_k^j$  are shrunken versions of the standard deviations for the regression coefficients  
532 of disease  $i$  and disease  $j$  in tissue/context  $k$ , respectively. Finally, for disease pairs with a positive  
533 coefficient from the weighted linear regression we take the coefficient of determination ( $R^2$ ) as a simi-  
534 larity measure; for disease pairs with a negative coefficient, we take  $-R^2$ . We note that for constant  
535 regression weights  $\{w_k^{(i,j)}\}$  this is equal to the Pearson correlation between the coefficient vectors we  
536 obtain from context-weighted score averaging (i.e.,  $\text{cor}(\beta^{(i)}, \beta^{(j)})$ ).

## 537 4.3 Variant prioritization performance

### 538 4.3.1 Tissue-weighted cross-validation performance

539 To measure the cross-validation performance of Tissue-weighted, we use repeated cross-validation [52]  
540 to reduce the variance (due to the random partitioning of data into 5 folds). Here, we repeated 5-  
541 fold cross-validation 30 times, and record the performance of each repeat. We later use the mean  
542 performance of the 30 repeats as the performance of that method and we also show the variance in  
543 figures such as Fig. 4.

### 544 4.3.2 Comparing organism level scores

545 For each disease we have disease-associated and control SNVs, and corresponding pre-computed  
546 organism-level scores. With this setup we calculate performance metrics of interest (area under the

547 receiver operator characteristic curve (AUROC) and average precision (AUPR)), and obtain disease-  
548 specific performance metrics for each scoring approach. To compare performance between organism-  
549 level scores on the same disease we use performance measures computed on 30 bootstrap samples  
550 (each bootstrap sample randomly contains 90% of disease and control variants) and then employ the  
551 Wilcoxon signed-ranks test to test to assess differences in performance. This yields p-values as reported  
552 in **Suppl. Data SD4**.

553 With respect to aggregating comparisons across diseases, we note that disease terms can (and do)  
554 share SNVs, so performance metrics in different terms are not necessarily independent. Also, disease  
555 terms can vary substantially in the number of annotated SNPs. We again use Wilcoxon signed-ranks  
556 test [53] on performance metrics (computed using all disease-associated- and control-SNVs for each  
557 disease term) to compare two organism-level variant scores aggregate across diseases. This approach  
558 yields p-values, as reported in **Suppl. Data SD5**.

### 559 4.3.3 Comparing tissue-weighted scores

560 Tissue-weighted baseline scores (see above) are calculated in the same way as organism-level scores.  
561 For tissue-weighted scores with data-driven tissue-specific weighting (see above), we use cross-validated  
562 performance measure for each bootstrap sample and the same 30 bootstrap samples as when we com-  
563 pared between organism-level scores. And then we use the same Wilcoxon signed-ranks tests to mea-  
564 sure the difference. For comparing scores aggregated across diseases we again proceed analogous to  
565 organism-level scores and use a Wilcoxon signed-ranks test on cross-validated disease-specific perfor-  
566 mance measures. Results are summarized in **Suppl. Data SD8** and **SD9**.

### 567 4.3.4 Comparing organism-level and tissue-weighted scores

568 For comparisons between organism-level and tissue-combined scores we again use a bootstrap approach:  
569 for a specific disease term we use the Wilcoxon signed-ranks tests as discussed above to compare per-  
570 formance measures from organism-level scores with tissue-weighted scores. We note that this approach  
571 does not take into account: *(a)* Variability in the organism-level scores originating from variability of  
572 the data they are derived from, and *(b)* The possibility that organism-level scores may have already  
573 used SNPs in their score derivation process, and we use them again for evaluation in their score  
574 derivation process. However, we don't expect these issue to substantially confound or results, and we  
575 note that incurred bias in our comparisons would expected to be in favor of organism-level scores.  
576 Results are summarized in **Suppl. Data SD6**, **SD7**, **SD10** and **SD11**.

### 577 4.3.5 DIVAN performance assessment and comparison.

578 To assess and compare our performance with DIVAN [19], we generated a test set of SNPs from the  
579 GWAS catalog that were *i)* added after DIVAN had been published (i.e., after 05/28/2016) and *ii)* not  
580 present in the database used to train DIVAN (Association Result Browser [https://www.ncbi.nlm.nih.gov/projects/gapplus/sgap\\_plus.htm](https://www.ncbi.nlm.nih.gov/projects/gapplus/sgap_plus.htm)) and *iii)* not within 1kb distance around SNPs used to  
582 train DIVAN and *iv)* were annotated to a disease phenotype addressed by DIVAN.

583 Control SNPs were generated using SNPsnap matching, as described above. To be able to satisfy  
584 criterion *iv)*, we mapped our disease terms (EFO terms) to disease terms used by DIVAN (MeSH terms)  
585 using the EMBL-EBI Ontology Xref Service (OxO, <https://www.ebi.ac.uk/spot/oxo/>, retrieved  
586 on April 19, 2020) and were able to resolve 41 out of 45 terms (**Suppl. Data SD12**). Of these, we  
587 keep terms with 20 or more disease associated SNPs in the test set and 50 or more SNPs in a training



588 set that we also construct (see below), yielding 29 overall disease phenotypes we use in our analysis.  
589 In order to fairly compare DIVAN with our logistic regression approach we constructed a training  
590 set using disease-associated SNPs from the GWAS catalog and the Phenotype-Genotype Integrator  
591 (PheGenI, <https://www.ncbi.nlm.nih.gov/gap/phegeni>) [54], excluding SNPs in the test dataset  
592 describe above, or SNPs within 1kb around test SNPs. **Suppl. Data** SD13 summarizes test and  
593 training data used for this analysis. Results are summarized in **Suppl. Data** SD14.

#### 594 4.3.6 Performance assessment using chromosome hold-out

595 To assess the performance of our DHS tissue-weighted score we also used a chromosome hold-out  
596 strategy, with test SNPs on different chromosomes from training data. Specifically, for each disease,  
597 we choose a set of chromosomes that contains approximately 20% SNVs with a 1/10 positive to negative  
598 ratio (the same as the cross-validation setting) as a test set. Selection of test chromosomes is performed  
599 for each disease term separately, as disease-associated SNPs differ. To automate the procedure, we  
600 deployed (binary) linear programming to pick out chromosomes in test set for each disease.

601 Specifically, for each disease term we solve the optimization problem

$$\begin{aligned} & \operatorname{argmax}_{\{x_i\}_{i=1}^{22}} \sum_{i=1}^{22} c_i x_i \\ & \text{subject to } \sum_{i=1}^{22} w_i^+ x_i \leq 0.2 \text{ and } x_i \in \{0, 1\}, \end{aligned}$$

602 where  $\{x_i\}$  are binary indicator variables whether a chromosome is included in the test/hold-out set;  
603  $w_i^-$  and  $w_i^+$  are the fraction of disease-associated ( $w_i^+$ ) and control SNPs ( $w_i^-$ ) on chromosome  $i$  and  
604 weights in the objective function are defined as  $c_i = w_i^+ - |w_i^+ - w_i^-|$ . This approach selects, for each  
605 disease term, a set of chromosomes to hold out that contain about 20% of disease-associated SNPs  
606 and that approximately reflects the overall imbalance between disease-associated and control SNPs.  
607 **Suppl. Fig.** S17 and S18 contain performance evaluations on chromosome hold-out sets.

#### 608 4.3.7 Performance assessment using one SNP per LD block

609 To assess the effect of SNP correlation on our results we also performed analyses using only a single  
610 representative SNP per LD block (defined by  $r^2 \geq 0.5$ , see **Section** 4.1.1). Results are shown in  
611 **Suppl. Fig.** S19 and S20.

## 612 4.4 Comparison with genetic correlation

613 We retrieved genetic correlation values from the GWAS atlas [31]. To be able to use these data we  
614 mapped EFO disease terms (used in the NIH-NCBI GWAS Catalog and in our study) to terms used  
615 in the GWAS atlas study. To do so, we extracted synonyms of each EFO term (as listed on EFO  
616 ontology) and compared each synonym to the "trait" and "uniqtrait" column in the GWAS atlas data.  
617 All matches (with one tolerated letter substitution) will be used.

In this approach a single EFO term can map to multiple GWAS atlas traits and studies. To estimate the genetic correlation between two EFO terms (say  $d_i$  and  $d_j$ ), we use a weighted combination of genetic correlation values:

$$r_g(d_i, d_j) = \sum_{l,m} w_{lm} r_g(s(d_i)_l, s(d_j)_m)$$

618 where  $r_g(\cdot, \cdot)$  is the genetic correlation of two diseases,  $\{s(d_i)\}_{i=1}^r$  and  $\{s(d_j)\}_{j=1}^s$  are the GWAS atlas

619 studies that are mapped to EFO term  $d_i$  and  $d_j$ , respectively;  $w_{lm}$  is a weight for each combination of  
620 the GWAS atlas studies accounting for the sample sizes of different studies used to estimate genetic  
621 correlation values. We choose

$$w_{lm} = \tilde{w}(s(d_i)_l) \cdot \tilde{w}(s(d_j)_m)$$

622 where

$$\tilde{w}(s(d_i)_l) = \text{size}(s(d_i)_l) / \sum_k \text{size}(s(d_i)_k)$$

623 where "size" denotes the sample size of a study. This scheme puts higher weights on studies with large  
624 sample sizes and smaller weights to studies with smaller sample sizes.

## 625 4.5 Notes about epimap comparison, cluster annotation and display

### 626 4.5.1 Epimap trait-tissue association for Table 5

627 We obtained the latest snp-centric GWAS enrichments table from the EpiMap Repository at <http://compbio.mit.edu/epimap/>. We retrieve tissues with adjusted p-values for each disease. We map the  
628 tissue names used in our study (Standard Roadmap Epigenomes, as labeled by EID) to tissue names used  
629 in epimap (biosamples, as labeled as BSS biosample id) by adapting the scripts from [https://github.com/cboix/EPIMAP\\_ANALYSIS/blob/master/metadata\\_scripts/get\\_roadmap\\_mapping.R](https://github.com/cboix/EPIMAP_ANALYSIS/blob/master/metadata_scripts/get_roadmap_mapping.R). If there  
630 are more than one biosamples tissues mapped to roadmap tissues, we reported the p value of the tissue  
631 with the most significant results.  
632  
633

### 634 4.5.2 Cluster names in Table 6

635 To name each cluster/group of diseases/EFO terms we choose the EFO term that contains most of  
636 the cluster/group members. In **Suppl. Data SD21** we summarize the terms with high term frequency  
637 in each cluster, where term frequency is the fraction of *descendant* terms present. For example, the  
638 EFO term "immune system disease" (EFO:0000540) has a term frequency of 0.588 in the "immune-1  
639 cluster"; this means that 58.8% of EFO terms in that cluster are descendants of EFO:0000540. We  
640 exclude the terms that are overly broad such as the term "disease" or "experimental factor ontology".  
641 For each cluster, we rank the cluster member EFO terms using term frequency and select as name a  
642 meaningful term with the high term frequency. For one cluster where no term had high frequency we  
643 chose the name "heterogeneous".

644 We also show a diagrams of EFO disease term relationships in each cluster in **Suppl. Fig. S10-S16**.  
645 Occasionally we include ancestor EFO terms not present in the cluster in a diagram, which are marked  
646 by asterisks.

### 647 4.5.3 Dimension reduction and coefficient heatmap

648 **UMAP plot** The two-dimensional UMAP plot of 111 EFO disease terms in **Fig. 7** is based on  
649 disease similarities based on context-weighted score averaging (see section 4.2.2). The `umap`  
650 function of the `uwot` R package was used with parameters `n_neighbors = 15`, `ret_model = TRUE`,  
651 `PCA_center = FALSE`.

**Coefficient heatmap** The heatmap in **Fig. 8** displays coefficient vectors of models for disease asso-  
ciation (see section 4.2.1), normalized for each disease. Specifically, for each disease and tissue

coefficient  $x_i$

$$\tilde{x}_i = \begin{cases} (x_i - x_{\min})/x_{95} & x_i \leq x_{95} \\ 1 & x_i > x_{95} \end{cases}$$

652 where  $x_{\min}$  is the minimum coefficient for a disease, and  $x_{95}$  is the 95% quantile.

653 **Cluster-associated tissues** For each cluster, we show the top-five tissues that are most associated  
654 with the cluster (**Fig. 8**). To identify these tissues we conduct a two-sample Wilcoxon test (one-  
655 sided) on every tissue, where we compare normalized tissue coefficients for this cluster to the the  
656 other with the highest coefficients on average. The five tissues with the smallest p-value are then  
657 selected as top-five tissues.

658 **Tissue-associated clusters** For the heatmap with all tissues in **Suppl. Fig. S9**, we assigned a cluster  
659 to each tissue. For each tissue, we calculated the median (across disease terms of a cluster) of  
660 the normalized coefficients for all clusters; the cluster with the highest median was assigned.

## 661 Data and code availability

662 Public data repositories were used as detailed in the Methods section, and data underlying tables and  
663 figures is available as supplemental information online. 25bp-resolution tissue-weighted DHS scores  
664 are available for download at <https://doi.org/10.7910/DVN/AUAJ7K>, and computer code used to  
665 generate analyses presented is available at [link-to-github](#).

## References

- 666
- 667 [1] A. Buniello et al. “The NHGRI-EBI GWAS Catalog of published genome-wide association  
668 studies, targeted arrays and summary statistics 2019”. In: *Nucleic Acids Res* 47.D1 (2019),  
669 pp. D1005–d1012. ISSN: 0305-1048. DOI: [10.1093/nar/gky1120](https://doi.org/10.1093/nar/gky1120).
- 670 [2] Joon-Yong An et al. “Genome-wide de novo risk score implicates promoter variation in autism  
671 spectrum disorder”. In: *Science (New York, N.Y.)* 362.6420 (2018), eaat6576. ISSN: 1095-9203  
672 0036-8075. DOI: [10.1126/science.aat6576](https://doi.org/10.1126/science.aat6576). URL: <https://pubmed.ncbi.nlm.nih.gov/30545852/>  
673 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6432922/>.
- 674 [3] Ernest Turro et al. “Whole-genome sequencing of patients with rare diseases in a national health  
675 system”. In: *Nature* 583.7814 (2020), pp. 96–102. ISSN: 1476-4687 0028-0836. DOI: [10.1038/](https://doi.org/10.1038/s41586-020-2434-2)  
676 [s41586-020-2434-2](https://doi.org/10.1038/s41586-020-2434-2). URL: <https://pubmed.ncbi.nlm.nih.gov/32581362/>  
677 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7610553/>.
- 678 [4] Kerstin Lindblad-Toh et al. “A high-resolution map of human evolutionary constraint using 29  
679 mammals”. In: *Nature* 478.7370 (2011), pp. 476–482. ISSN: 1476-4687 0028-0836. DOI: [10.1038/](https://doi.org/10.1038/nature10530)  
680 [nature10530](https://doi.org/10.1038/nature10530). URL: <https://pubmed.ncbi.nlm.nih.gov/21993624/>  
681 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3207357/>.
- 682 [5] A. Kundaje et al. “Integrative analysis of 111 reference human epigenomes”. In: *Nature* 518.7539  
683 (2015), pp. 317–30. ISSN: 0028-0836 (Print) 0028-0836. DOI: [10.1038/nature14248](https://doi.org/10.1038/nature14248).
- 684 [6] Jill E. Moore et al. “Expanded encyclopaedias of DNA elements in the human and mouse  
685 genomes”. In: *Nature* 583.7818 (2020), pp. 699–710. ISSN: 1476-4687. DOI: [10.1038/s41586-](https://doi.org/10.1038/s41586-020-2493-4)  
686 [020-2493-4](https://doi.org/10.1038/s41586-020-2493-4). URL: <https://doi.org/10.1038/s41586-020-2493-4>.
- 687 [7] Phil H. Lee et al. “Principles and methods of in-silico prioritization of non-coding regulatory  
688 variants”. In: *Human genetics* 137.1 (2018), pp. 15–30. ISSN: 1432-1203 0340-6717. DOI: [10.](https://doi.org/10.1007/s00439-017-1861-0)  
689 [1007/s00439-017-1861-0](https://doi.org/10.1007/s00439-017-1861-0). URL: <https://pubmed.ncbi.nlm.nih.gov/29288389/>  
690 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5892192/>.
- 691 [8] M. Kircher et al. “A general framework for estimating the relative pathogenicity of human genetic  
692 variants”. In: *Nat Genet* 46.3 (2014), pp. 310–5. ISSN: 1061-4036. DOI: [10.1038/ng.2892](https://doi.org/10.1038/ng.2892).
- 693 [9] I. Ionita-Laza et al. “A spectral approach integrating functional genomic annotations for coding  
694 and noncoding variants”. In: *Nat Genet* 48.2 (2016), pp. 214–20. ISSN: 1061-4036. DOI: [10.1038/](https://doi.org/10.1038/ng.3477)  
695 [ng.3477](https://doi.org/10.1038/ng.3477).
- 696 [10] Y. F. Huang, B. Gulko, and A. Siepel. “Fast, scalable prediction of deleterious noncoding variants  
697 from functional and population genomic data”. In: *Nat Genet* 49.4 (2017), pp. 618–624. ISSN:  
698 1061-4036. DOI: [10.1038/ng.3810](https://doi.org/10.1038/ng.3810).
- 699 [11] Q. Lu et al. “Integrative Tissue-Specific Functional Annotations in the Human Genome Provide  
700 Novel Insights on Many Complex Traits and Improve Signal Prioritization in Genome Wide  
701 Association Studies”. In: *PLoS Genet* 12.4 (2016), e1005947. ISSN: 1553-7390. DOI: [10.1371/](https://doi.org/10.1371/journal.pgen.1005947)  
702 [journal.pgen.1005947](https://doi.org/10.1371/journal.pgen.1005947).
- 703 [12] Brad Gulko and Adam Siepel. “An evolutionary framework for measuring epigenomic information  
704 and estimating cell-type-specific fitness consequences”. In: *Nature Genetics* 51.2 (2019), pp. 335–  
705 342. ISSN: 1546-1718. DOI: [10.1038/s41588-018-0300-z](https://doi.org/10.1038/s41588-018-0300-z). URL: [https://doi.org/10.1038/](https://doi.org/10.1038/s41588-018-0300-z)  
706 [s41588-018-0300-z](https://doi.org/10.1038/s41588-018-0300-z).

- 707 [13] D. Backenroth et al. “FUN-LDA: A Latent Dirichlet Allocation Model for Predicting Tissue-  
708 Specific Functional Effects of Noncoding Variation: Methods and Applications”. In: *Am J Hum*  
709 *Genet* 102.5 (2018), pp. 920–942. ISSN: 0002-9297 (Print) 0002-9297. DOI: [10.1016/j.ajhg.](https://doi.org/10.1016/j.ajhg.2018.03.026)  
710 [2018.03.026](https://doi.org/10.1016/j.ajhg.2018.03.026). URL: <https://pubmed.ncbi.nlm.nih.gov/29727691/>.
- 711 [14] Kévin Vervier and Jacob J. Michaelson. “TiSAn: estimating tissue-specific effects of coding and  
712 non-coding variants”. In: *Bioinformatics* 34.18 (2018), pp. 3061–3068. ISSN: 1367-4803. DOI: [10.](https://doi.org/10.1093/bioinformatics/bty301)  
713 [1093/bioinformatics/bty301](https://doi.org/10.1093/bioinformatics/bty301). URL: <https://doi.org/10.1093/bioinformatics/bty301>.
- 714 [15] Carles A. Boix et al. “Regulatory genomic circuitry of human disease loci by integrative epige-  
715 nomics”. In: *Nature* 590.7845 (2021), pp. 300–307. ISSN: 1476-4687. DOI: [10.1038/s41586-020-](https://doi.org/10.1038/s41586-020-03145-z)  
716 [03145-z](https://doi.org/10.1038/s41586-020-03145-z). URL: <https://doi.org/10.1038/s41586-020-03145-z>.
- 717 [16] Felix Richter et al. “Genomic analyses implicate noncoding de novo variants in congenital heart  
718 disease”. In: *Nature Genetics* 52.8 (2020), pp. 769–777. ISSN: 1546-1718. DOI: [10.1038/s41588-](https://doi.org/10.1038/s41588-020-0652-z)  
719 [020-0652-z](https://doi.org/10.1038/s41588-020-0652-z). URL: <https://doi.org/10.1038/s41588-020-0652-z>.
- 720 [17] Ali Yousefian-Jazi et al. “Functional fine-mapping of noncoding risk variants in amyotrophic lat-  
721 eral sclerosis utilizing convolutional neural network”. In: *Scientific Reports* 10.1 (2020), p. 12872.  
722 ISSN: 2045-2322. DOI: [10.1038/s41598-020-69790-6](https://doi.org/10.1038/s41598-020-69790-6). URL: [https://doi.org/10.1038/](https://doi.org/10.1038/s41598-020-69790-6)  
723 [s41598-020-69790-6](https://doi.org/10.1038/s41598-020-69790-6).
- 724 [18] Ali Yousefian-Jazi et al. “Functional annotation of noncoding causal variants in autoimmune  
725 diseases”. In: *Genomics* 112.2 (2020), pp. 1208–1213. ISSN: 0888-7543. DOI: [https://doi.org/](https://doi.org/10.1016/j.ygeno.2019.07.006)  
726 [10.1016/j.ygeno.2019.07.006](https://doi.org/10.1016/j.ygeno.2019.07.006). URL: [https://www.sciencedirect.com/science/article/](https://www.sciencedirect.com/science/article/pii/S0888754319301272)  
727 [pii/S0888754319301272](https://www.sciencedirect.com/science/article/pii/S0888754319301272).
- 728 [19] L. Chen, P. Jin, and Z. S. Qin. “DIVAN: accurate identification of non-coding disease-specific  
729 risk variants using multi-omics profiles”. In: *Genome Biol* 17.1 (2016), p. 252. ISSN: 1474-7596  
730 (Print) 1474-7596. DOI: [10.1186/s13059-016-1112-z](https://doi.org/10.1186/s13059-016-1112-z). URL: [https://www.ncbi.nlm.nih.gov/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5139035/)  
731 [pmc/articles/PMC5139035/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5139035/).
- 732 [20] Corneliu A. Bodea et al. “PINES: phenotype-informed tissue weighting improves prediction of  
733 pathogenic noncoding variants”. In: *Genome Biology* 19.1 (2018), p. 173. ISSN: 1474-760X. DOI:  
734 [10.1186/s13059-018-1546-6](https://doi.org/10.1186/s13059-018-1546-6). URL: <https://doi.org/10.1186/s13059-018-1546-6>.
- 735 [21] Long Gao et al. “Identifying noncoding risk variants using disease-relevant gene regulatory net-  
736 works”. In: *Nature Communications* 9.1 (2018), p. 702. ISSN: 2041-1723. DOI: [10.1038/s41467-](https://doi.org/10.1038/s41467-018-03133-y)  
737 [018-03133-y](https://doi.org/10.1038/s41467-018-03133-y). URL: <https://doi.org/10.1038/s41467-018-03133-y>.
- 738 [22] Tune H. Pers, Pascal Timshel, and Joel N. Hirschhorn. “SNPsnap: a Web-based tool for identifi-  
739 cation and annotation of matched SNPs”. In: *Bioinformatics* 31.3 (2014), pp. 418–420. ISSN:  
740 1367-4803. DOI: [10.1093/bioinformatics/](https://doi.org/10.1093/bioinformatics/btu655)  
741 [btu655](https://doi.org/10.1093/bioinformatics/btu655). URL: [https://doi.org/10.1093/](https://doi.org/10.1093/bioinformatics/btu655)  
[bioinformatics/btu655](https://doi.org/10.1093/bioinformatics/btu655).
- 742 [23] G. R. Ritchie et al. “Functional annotation of noncoding sequence variants”. In: *Nat Methods*  
743 11.3 (2014), pp. 294–6. ISSN: 1548-7091. DOI: [10.1038/nmeth.2832](https://doi.org/10.1038/nmeth.2832).
- 744 [24] MedlinePlus [Internet]. Bethesda (MD): National Library of Medicine (US). *Systemic Sclero-*  
745 *derma*. URL: <https://medlineplus.gov/genetics/condition/systemic-scleroderma/>.
- 746 [25] Marina D Kraaij and Jacob M van Laar. “The role of B cells in systemic sclerosis”. In: *Biologics:*  
747 *Targets and Therapy* 2.3 (2008), pp. 389–395. ISSN: 1177-5491.
- 748 [26] Benjamin Thoreau, Benjamin Chaigne, and Luc Mouthon. “Role of B-cell in the pathogenesis of  
749 systemic sclerosis”. In: *Frontiers in Immunology* 13 (2022), p. 933468. ISSN: 1664-3224.

- 750 [27] MedlinePlus [Internet]. Bethesda (MD): National Library of Medicine (US). *Primary sclerosing*  
751 *cholangitis*. URL: [https://medlineplus.gov/genetics/condition/primary-sclerosing-](https://medlineplus.gov/genetics/condition/primary-sclerosing-cholangitis/)  
752 [cholangitis/](https://medlineplus.gov/genetics/condition/primary-sclerosing-cholangitis/).
- 753 [28] Lilly Kristin Kunzmann et al. “Monocytes as potential mediators of pathogen-induced T-helper  
754 17 differentiation in patients with primary sclerosing cholangitis (PSC)”. In: *Hepatology* 72.4  
755 (2020), pp. 1310–1326. ISSN: 0270-9139.
- 756 [29] Temitope O Keku et al. “Rectal mucosal proliferation, dietary factors, and the risk of colorectal  
757 adenomas”. In: *Cancer epidemiology, biomarkers & prevention: a publication of the American*  
758 *Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*  
759 7.11 (1998), pp. 993–999. ISSN: 1055-9965.
- 760 [30] Santosh Dulal and Temitope O Keku. “Gut microbiome and colorectal adenomas”. In: *Cancer*  
761 *journal (Sudbury, Mass.)* 20.3 (2014), p. 225.
- 762 [31] Kyoko Watanabe et al. “A global overview of pleiotropy and genetic architecture in complex  
763 traits”. In: *Nature Genetics* 51.9 (2019), pp. 1339–1348. ISSN: 1546-1718. DOI: [10.1038/s41588-](https://doi.org/10.1038/s41588-019-0481-0)  
764 [019-0481-0](https://doi.org/10.1038/s41588-019-0481-0). URL: <https://doi.org/10.1038/s41588-019-0481-0>.
- 765 [32] Yuanhao Yang et al. “Investigating the shared genetic architecture between multiple sclerosis and  
766 inflammatory bowel diseases”. In: *Nature Communications* 12.1 (2021), p. 5641. ISSN: 2041-1723.  
767 DOI: [10.1038/s41467-021-25768-0](https://doi.org/10.1038/s41467-021-25768-0). URL: <https://doi.org/10.1038/s41467-021-25768-0>.
- 768 [33] K. K. Farh et al. “Genetic and epigenetic fine mapping of causal autoimmune disease variants”.  
769 In: *Nature* 518.7539 (2015), pp. 337–43. ISSN: 0028-0836 (Print) 0028-0836. DOI: [10.1038/](https://doi.org/10.1038/nature13835)  
770 [nature13835](https://doi.org/10.1038/nature13835).
- 771 [34] C. McDowell, U. Farooq, and M. Haseeb. “Inflammatory Bowel Disease”. In: *StatPearls*. Treasure  
772 Island (FL): StatPearls Publishing Copyright © 2022, StatPearls Publishing LLC., 2022.
- 773 [35] C. Lord et al. “Autism spectrum disorder”. In: *Lancet* 392.10146 (2018), pp. 508–520. ISSN:  
774 0140-6736 (Print) 0140-6736. DOI: [10.1016/s0140-6736\(18\)31129-2](https://doi.org/10.1016/s0140-6736(18)31129-2).
- 775 [36] G. Olivo, S. Gaudio, and H. B. Schiöth. “Brain and Cognitive Development in Adolescents with  
776 Anorexia Nervosa: A Systematic Review of fMRI Studies”. In: *Nutrients* 11.8 (2019). ISSN: 2072-  
777 6643. DOI: [10.3390/nu11081907](https://doi.org/10.3390/nu11081907).
- 778 [37] E. R. Sigmon et al. “Congenital Heart Disease and Autism: A Case-Control Study”. In: *Pediatrics*  
779 144.5 (2019). ISSN: 0031-4005. DOI: [10.1542/peds.2018-4114](https://doi.org/10.1542/peds.2018-4114).
- 780 [38] Z. C. Zhou, D. B. McAdam, and D. R. Donnelly. “Endophenotypes: A conceptual link between  
781 anorexia nervosa and autism spectrum disorder”. In: *Research in Developmental Disabilities* 82  
782 (2018), pp. 153–165. ISSN: 0891-4222. DOI: <https://doi.org/10.1016/j.ridd.2017.11.008>.  
783 URL: <https://www.sciencedirect.com/science/article/pii/S0891422217303025>.
- 784 [39] Margherita Boltri and Walter Sapuppo. “Anorexia Nervosa and Autism Spectrum Disorder:  
785 A Systematic Review”. In: *Psychiatry Research* 306 (2021), p. 114271. ISSN: 0165-1781. DOI:  
786 <https://doi.org/10.1016/j.psychres.2021.114271>. URL: [https://www.sciencedirect.](https://www.sciencedirect.com/science/article/pii/S0165178121005667)  
787 [com/science/article/pii/S0165178121005667](https://www.sciencedirect.com/science/article/pii/S0165178121005667).
- 788 [40] D. S. Tylee et al. “Genetic correlations among psychiatric and immune-related phenotypes based  
789 on genome-wide association data”. In: *Am J Med Genet B Neuropsychiatr Genet* 177.7 (2018),  
790 pp. 641–657. ISSN: 1552-4841 (Print) 1552-4841. DOI: [10.1002/ajmg.b.32652](https://doi.org/10.1002/ajmg.b.32652).

- 791 [41] C. Y. Li et al. “Genome-wide genetic links between amyotrophic lateral sclerosis and autoimmune  
792 diseases”. In: *BMC Med* 19.1 (2021), p. 27. ISSN: 1741-7015. DOI: [10.1186/s12916-021-01903-y](https://doi.org/10.1186/s12916-021-01903-y).
- 793 [42] X. Yu et al. “Innate Lymphoid Cells and Celiac Disease: Current Perspective”. In: *Cell Mol  
794 Gastroenterol Hepatol* 11.3 (2021), pp. 803–814. ISSN: 2352-345x. DOI: [10.1016/j.jcmgh.2020.  
795 12.002](https://doi.org/10.1016/j.jcmgh.2020.12.002).
- 796 [43] B. Jabri and L. M. Sollid. “T Cells in Celiac Disease”. In: *J Immunol* 198.8 (2017), pp. 3005–  
797 3014. ISSN: 0022-1767 (Print) 0022-1767. DOI: [10.4049/jimmunol.1601693](https://doi.org/10.4049/jimmunol.1601693).
- 798 [44] Q. Lu et al. “A statistical framework to predict functional non-coding regions in the human  
799 genome through integrated analysis of annotation data”. In: *Sci Rep* 5 (2015), p. 10576. ISSN:  
800 2045-2322. DOI: [10.1038/srep10576](https://doi.org/10.1038/srep10576).
- 801 [45] Jingsi Ming et al. “LSMM: a statistical approach to integrating functional annotations with  
802 genome-wide association studies”. In: *Bioinformatics* 34.16 (2018), pp. 2788–2796. ISSN: 1367-  
803 4803. DOI: [10.1093/bioinformatics/bty187](https://doi.org/10.1093/bioinformatics/bty187). URL: [https://doi.org/10.1093/bioinformatics/  
804 bty187](https://doi.org/10.1093/bioinformatics/bty187).
- 805 [46] A. Julià et al. “Genome-wide association study meta-analysis identifies five new loci for systemic  
806 lupus erythematosus”. In: *Arthritis Res Ther* 20.1 (2018), p. 100. ISSN: 1478-6354 (Print) 1478-  
807 6354. DOI: [10.1186/s13075-018-1604-1](https://doi.org/10.1186/s13075-018-1604-1).
- 808 [47] H. Lu et al. “Detection of Genetic Overlap Between Rheumatoid Arthritis and Systemic Lupus  
809 Erythematosus Using GWAS Summary Statistics”. In: *Front Genet* 12 (2021), p. 656545. ISSN:  
810 1664-8021 (Print) 1664-8021. DOI: [10.3389/fgene.2021.656545](https://doi.org/10.3389/fgene.2021.656545).
- 811 [48] J. Bentham et al. “Genetic association analyses implicate aberrant regulation of innate and  
812 adaptive immunity genes in the pathogenesis of systemic lupus erythematosus”. In: *Nat Genet*  
813 47.12 (2015), pp. 1457–1464. ISSN: 1061-4036 (Print) 1061-4036. DOI: [10.1038/ng.3434](https://doi.org/10.1038/ng.3434).
- 814 [49] James Malone et al. “Modeling sample variables with an Experimental Factor Ontology”. In:  
815 *Bioinformatics* 26.8 (Mar. 2010), pp. 1112–1118. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/  
816 btq099](https://doi.org/10.1093/bioinformatics/btq099). eprint: [https://academic.oup.com/bioinformatics/article-pdf/26/8/1112/  
817 13848104/btq099.pdf](https://academic.oup.com/bioinformatics/article-pdf/26/8/1112/13848104/btq099.pdf). URL: <https://doi.org/10.1093/bioinformatics/btq099>.
- 818 [50] J Schreiber, J Bilmes, and WS Noble. “Completing the ENCODE3 compendium yields accurate  
819 imputations across a variety of assays and human biosamples.” In: *Genome Biol* 21 (Mar. 2020),  
820 p. 82.
- 821 [51] J Friedman, T Hastie, and R Tibshirani. “Regularization Paths for Generalized Linear Models  
822 via Coordinate Descent.” In: *J Stat Softw* 33 (2010), pp. 1–22.
- 823 [52] Gitte Vanwinckelen and Hendrik Blockeel. “On estimating model accuracy with repeated cross-  
824 validation”. In: *BeneLearn 2012: Proceedings of the 21st Belgian-Dutch Conference on Machine  
825 Learning* (2012), pp. 39–44. ISSN: 978-94-6197-044-2.
- 826 [53] Janez Demsar. “Statistical Comparisons of Classifiers over Multiple Data Sets”. In: *Journal of  
827 Machine Learning Research* 7 (2006), pp. 1–30. URL: [http://www.jmlr.org/papers/v7/  
828 demsar06a.html](http://www.jmlr.org/papers/v7/demsar06a.html).
- 829 [54] EM Ramos et al. “Phenotype-Genotype Integrator (PheGenI): synthesizing genome-wide asso-  
830 ciation study (GWAS) data with existing genomic resources.” In: *Eur J Hum Genet* 22 (Jan.  
831 2014), pp. 144–7.