

The Gene Expression Landscape of Disease Genes

Judit García-González¹, Saul Garcia-Gonzalez^{1,2}, Lathan Liou¹, Paul F. O'Reilly¹,

¹Department of Genetics and Genomic Sciences, Icahn School of Medicine, Mount Sinai, New York City, NY 10029, USA

²Center for Excellence in Youth Education, Icahn School of Medicine, Mount Sinai, New York City, NY 10029, USA

Fine-mapping and gene-prioritisation techniques applied to the latest Genome-Wide Association Study (GWAS) results have prioritised hundreds of genes as causally associated with disease. Here we leverage these recently compiled lists of high-confidence causal genes to interrogate where in the body disease genes operate. Specifically, we combine GWAS summary statistics, gene prioritisation results and gene expression RNA-seq data from 46 tissues and 204 cell types in relation to 16 major diseases (including 8 cancers). In tissues and cell types with well-established relevance to the disease, the prioritised genes typically have higher absolute and relative (i.e. tissue/cell specific) expression compared to non-prioritised 'control' genes. Examples include brain tissues in psychiatric disorders (P -value $< 1 \times 10^{-7}$), microglia cells in Alzheimer's Disease (P -value = 9.8×10^{-3}) and colon mucosa in colorectal cancer (P -value $< 1 \times 10^{-3}$). We also observe significantly higher expression for disease genes in multiple tissues and cell types with no established links to the corresponding disease. While some of these results may be explained by cell types that span multiple tissues, such as macrophages in brain, blood, lung and spleen in relation to Alzheimer's disease (P -values $< 1 \times 10^{-3}$), the cause for others is unclear and motivates further investigation that may provide novel insights into disease etiology. For example, mammary tissue in Type 2 Diabetes (P -value $< 1 \times 10^{-7}$); reproductive tissues such as breast, uterus, vagina, and prostate in Coronary Artery Disease (P -value $< 1 \times 10^{-4}$); and motor neurons in psychiatric disorders (P -value $< 3 \times 10^{-4}$). In the GTEx dataset, tissue type is the major predictor of gene expression but the contribution of each predictor (tissue, sample, subject, batch) varies widely among disease-associated genes. Finally, we highlight genes with the highest levels of gene expression in relevant tissues to guide functional follow-up studies. Our results could offer novel insights into the tissues and cells involved in disease initiation, inform drug target and delivery strategies, highlighting potential off-target effects, and exemplify the relative performance of different statistical tests for linking disease genes with tissue and cell type gene expression.

1 Introduction

2 Genome-wide association studies (GWAS) for complex diseases have identified thousands of
3 risk loci in the last two decades¹. An important first step in translating GWAS findings into
4 biological and clinical insights is to take broadly identified risk loci, incorporating associations
5 across usually many genes due to linkage disequilibrium (LD), and interrogate them to pin-point
6 the causal variants and genes. To identify causal disease genes, fine-mapping and gene
7 prioritisation strategies have been developed² and – only in recent years – lists of high-
8 confidence causal genes for multiple diseases have been compiled^{3–11}. Since (i) the probability
9 of success in drug development increases with support for the relevant gene in GWAS¹², and (ii)
10 tissue-specific genes are more likely to become drug targets than broadly expressed genes^{13–15},
11 profiling the gene expression of these disease-associated genes across multiple tissues and cell
12 types could aid the development of new drugs^{16,17} and limit off-target effects^{18,19}. However, a
13 systematic characterisation of gene expression for GWAS prioritised genes has yet to be
14 performed.

15
16 While understanding the specific cell types involved in disease and their spatial distribution has
17 been of intense interest in recent years^{20–25}, owing to the technological and computational
18 advances of single-cell genomics (reviewed here²⁶), we first focus here on identifying the
19 relevant tissues in which disease-associated genes are expressed. Tissues known to be
20 implicated in diseases serve as positive controls, essential to benchmark and optimise
21 approaches that assess the relevance of cell and tissue types. Such positive controls are scarce
22 for cell types^{27,28}. Moreover, growing evidence highlights the intricate interconnections among
23 the body's various systems (nervous, immune, metabolic, hematopoietic, endocrine),
24 suggesting that multiple tissues could typically be involved in disease^{29–31}. After investigating
25 tissues, we apply the same systematic approach at the higher resolution of the cell type.

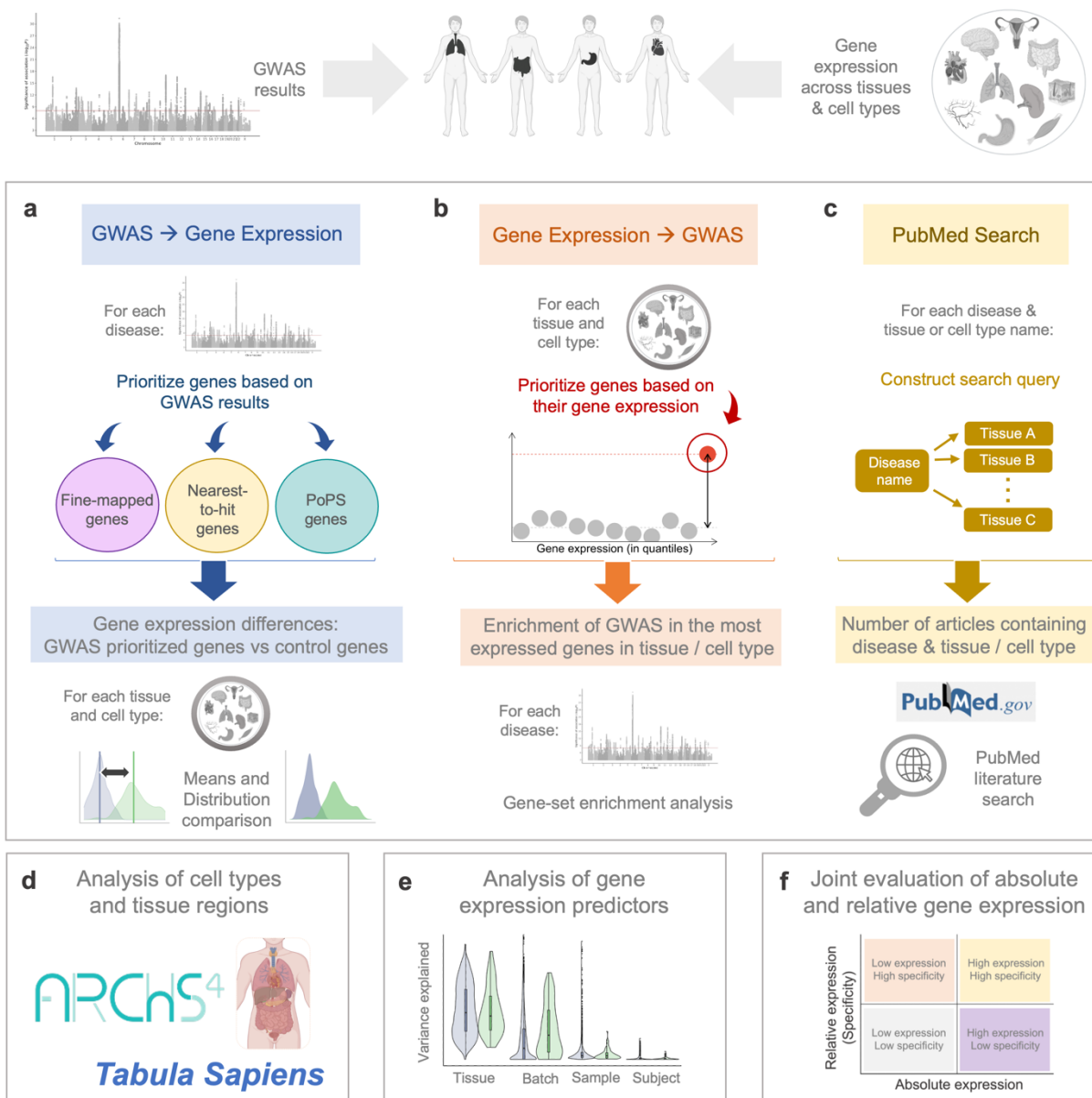
26
27 In this study, we characterise the tissues and cell types in which disease-associated genes are
28 expressed. Our primary analysis uses three alternative approaches that leverage RNA-seq and
29 GWAS data: the first – that we call “*from GWAS to Gene Expression*” – interrogates whether
30 genes prioritised as causal in the latest landmark GWAS of major diseases have distinct gene
31 expression features from those of other protein-coding genes (**Fig 1a**). The second approach –
32 that we call “*from Gene Expression to GWAS*” – examines whether high-expression genes are
33 enriched for GWAS signal, as calculated by MAGMA³² (**Fig 1b**). In the third approach, we

34 perform a systematic PubMed scan to assess the evidence for tissue-disease associations
35 reported in the literature (**Fig 1c**). Contrasting results systematically across the three
36 approaches allows for triangulation of results. Furthermore, we ensure the robustness of our
37 study by employing three distinct definitions for disease-associated genes and utilising three
38 different sets of control genes. We apply our testing framework to more than 200 cell types and
39 tissue regions obtained from the ARCHS4³³ and Tabula Sapiens³⁴ resources (**Fig 1d**). We also
40 characterise, for each individual gene, to what extent different predictors (batch ID, subject ID,
41 age, sex etc) contribute to gene expression (**Fig 1e**) and identify genes with the highest
42 absolute and relative gene expression in relevant tissues (**Fig 1f**).

43
44 Analyses are performed across eight cancers – for which there is a strongly implicated tissue for
45 each – as well as Schizophrenia (SCZ), Inflammatory Bowel Disease (IBD), Alzheimer’s
46 Disease (AD), Coronary Artery Disease (CAD), Bipolar Disorder (BD), Type 2 Diabetes (T2D),
47 Attention-Deficit/Hyperactivity Disorder (ADHD) and Serum 25 Hydroxyvitamin D (Vitamin D).
48 These outcomes were selected to optimise the power of the relevant GWAS and the availability
49 of curated lists of high-confidence disease genes.

50

Where are disease genes expressed?



51
 52 **Figure 1.** Overview of the study to characterise the gene expression features of genes associated to diseases. **a**, Approach using
 53 lists putatively causal genes, prioritized based on GWAS results. **b**, Approach using genes with the highest absolute expression or
 54 highest relative expression in each tissue and cell type. **c**, PubMed-based literature search to assess the tissues that are most often
 55 cited for each disease. **d**, The approaches in a-c were repeated for the cell types and tissue regions available in the ARCHS4 and
 56 Tabula Sapiens resources. **e**, The predictors of gene expression were analysed for each gene using the variance partition R
 57 package. **f**, For each gene, absolute and relative gene expression was evaluated. Figure partially created with BioRender.com.
 58
 59

60 **Results**

61 **Defining absolute and relative expression across tissues and cell types**

62 We obtained bulk-tissue, RNA-seq gene expression data from the GTEx consortium³⁵.
63 Throughout this study, we assess two gene expression measurements: (1) *absolute gene*
64 *expression*, representing the median number of transcripts per million (TPM) of each gene in
65 each tissue, and (2) *relative gene expression*, calculated by dividing the *absolute gene*
66 *expression* (median TPM) of each gene in a tissue by the total expression of that gene across
67 all the other tissues (**Methods**). The relative gene expression measure, often referred to as
68 gene expression specificity, has been widely used to map genes to their specific tissue and cell
69 type expressions^{14,36,37}.

70 71 **Exploring the expression profiles of disease-associated genes**

72 To identify where in the body disease-associated genes operate, here we leverage GWAS and
73 gene expression data utilising three alternative strategies: (i) '*GWAS to Gene Expression*'; (ii)
74 '*Gene Expression to GWAS*'; and (iii) '*Systematic Literature Search*'.

75 76 **GWAS to gene expression**

77 Heritability across the genome is influenced by polygenicity and the genome's correlation
78 structure (LD), causing signals from single causal variants in key disease genes to spread
79 across wide regions and many genes. Although the 'omnigenic model' suggests there may be
80 few key ('*core*') disease genes despite widespread genetic associations³⁸, identifying those
81 causal variants and genes remains challenging. To link regulatory SNPs to their target genes
82 and prioritize genes based on GWAS results, comprehensive annotations of genome
83 function^{35,39–43} and a range of statistical and computational approaches have been developed^{44–}
84 ⁴⁸.

85
86 Given the absence of a single gold standard approach, we use three alternative methods to
87 collect, for each disease, lists of putatively causal genes inferred from GWAS results: (i)
88 *nearest-to-hit genes*: genes are prioritized by physical proximity to each GWAS hit, (ii) *fine-*
89 *mapped genes*: we extracted from published studies – often produced by large GWAS consortia
90 – lists of genes prioritized on the basis of multiple statistical and functional genomic prioritization
91 strategies, and (iii) *PoPS genes*: derived using the recently published method Polygenic Priority
92 Scores (PoPS)⁴⁸, which leverages polygenic enrichment and functional gene features to
93 prioritise genes.

94
95 - *Nearest-to-hit genes*: While functional data can be incorporated to improve precision in
96 assigning SNPs to genes⁴⁴, previous research suggests that the nearest gene to the lead SNP
97 is the most likely causal gene^{44,48}. We performed clumping on the GWAS hits and identified an
98 average of 162 ‘nearest-to-hit’ genes for each trait, with an average distance of 28Kbp between
99 the SNP and the protein coding gene (See **Methods, Table 1** and **Supplemental Table 1**).

100
101 - *Fine-mapped genes*: The latest landmark GWAS of the major diseases³⁻¹¹ have incorporated
102 sub-studies performing gene prioritization analyses to produce lists of approximately 50-300
103 high-confidence causal genes for each disease. While different studies use a different selection
104 of methods to prioritize putatively causal genes, most of them integrate GWAS results with the
105 latest functional genomics data (**Methods**). Our literature search on fine-mapped genes resulted
106 in gene lists ranging from 49 genes prioritized for Bipolar Disorder (BD) to 281 genes prioritized
107 for Inflammatory Bowel Disorder (IBD), with an average of 126 across the 8 diseases
108 investigated (**Supplemental Table 2**).

109
110 - *PoPS genes*: The method PoPS⁴⁸ leverages gene-level Z-scores from GWAS (calculated
111 using the software MAGMA³²), as well as gene features from single-cell gene expression data,
112 biological pathways and predicted protein-protein interaction networks to prioritize putatively
113 causal genes. For each trait, we extracted the genes with the top 1% PoPS scores,
114 corresponding to 184 protein-coding genes with highest PoPS scores for each trait. The top 1%
115 of genes threshold was selected because it provides a similar number of genes as the other two
116 prioritization approaches, preventing biases due to differences in the number of genes included
117 for each group of disease-associated genes (**Supplemental Table 3**).

118
119 The Venn diagrams in **Fig 2** shows the overlap of genes for each of the diseases examined.
120 Across the eight non-cancer diseases, fine-mapped genes and the nearest-to-hit genes showed
121 the highest overlap, likely due to the use of genomic distance to fine-map variants and prioritize
122 genes in the previously published studies from which we extracted the lists of fine-mapped
123 genes. Differences in the criteria used for prioritizing genes may have partially led to differences
124 in the number of overlapping genes. For example, PoPS scores were one of the eight strategies
125 used for creating the list of fine-mapped genes for CAD, but PoPS scores were not used for any
126 of the other outcomes.

127

128 For each disease, we performed *t*-tests to compare the *absolute gene expression* and *relative*
129 *gene expression* between disease-associated genes (fine-mapped, nearest-to-hit and PoPS
130 genes) vs control genes. In **Fig 2**, we report the *t*-test *P*-values, whereas the effect size,
131 measured using Cohen's *D*, is included in **Supplemental Figure 1**. The *absolute gene*
132 *expression* of disease-genes vs other genes is higher in tissues with established links to each
133 disease (**Fig 2**, blue columns). For instance, SCZ-associated genes are more expressed in the
134 brain, although the *P*-values vary significantly across brain tissues (from *P*-value = 3.09×10^{-2}
135 for the nearest gene list in substantia nigra to *P*-value = 2.01×10^{-29} for PoPS genes in the brain
136 cortex). CAD-associated genes are more expressed in the aorta, coronary and tibial arteries (*P*-
137 values < 10^{-7}). IBD-associated genes are most expressed in the small intestine and colon
138 transverse (*P*-values < 10^{-4} except for nearest-to-hit genes). Vitamin D genes are most
139 expressed in the skin and in the liver, with the latter tissue being where the biologically inactive
140 vitamin D₃ is activated to produce 25-hydroxyvitamin D₃⁴⁹.

141
142 Intriguingly, other significant results point to tissues not typically linked to the disease: AD-
143 associated genes do not show higher expression across brain tissue. Instead, the most
144 significant differences in expression appear in the blood and spleen (PoPS genes *P*-value < 10^{-13} ,
145 nearest-to-hit genes *P*-value < 10^{-5}) and in adipose tissues (PoPS genes *P*-values < 10^{-13}).
146 SCZ-associated genes are more expressed in the pituitary (PoPS genes *P*-values = 3.66×10^{-5}).
147 CAD-associated genes present higher expression across multiple tissues including
148 reproductive, adipose, and digestive systems, as well as in the lung (*P*-values < 10^{-6}). IBD-
149 associated genes present highest expression in lung (*P*-values < 10^{-4}), blood and spleen (*P*-
150 values < 10^{-6}). Differences between T2D associated genes vs control genes appear most
151 significant in breast mammary tissues (*P*-values < 10^{-7}). Disease genes' *relative gene*
152 *expression t*-test results are similar to those for *absolute gene expression*, but show smaller *P*-
153 values.

154
155 We also applied the Anderson-Darling test⁵⁰ to assess whether disease-associated genes have
156 a distribution of expression that differs from that of other protein-coding genes (i.e. not only in
157 terms of mean expression). This test is more sensitive to detect differences in the tails of the
158 distribution, in comparison to tests focusing only on mean differences (e.g. *t*-tests) or at the
159 shape of the cumulative distribution (e.g. two-sample Kolmogorov-Smirnov)⁵¹. In most cases,
160 the Anderson Darling tests reported lower *P*-values than the *t*-tests (**Supplemental Figure 2**).

161 Full results for the *t*-tests and Anderson-Darling tests for each disease, each tissue, and each
162 gene list are included in **Supplemental Table 4**.

163
164 Overall, genes prioritized by PoPS show smaller *P*-values across a wider range of tissues,
165 especially for CAD, AD, T2D and Vitamin D, suggesting that PoPS prioritizes genes with higher
166 levels of expression. In the original PoPS publication⁴⁸, PoPS scores are combined with the
167 genomic location to provide a list of high-confidence causal genes. However, this list has a low
168 recall (it detects few genes for each disease). Therefore, we used the top 1% PoPS, that results
169 in 184 genes per disease. This number of genes is similar to the GWAS fine-mapping and
170 closest gene to locus approaches and would not lead to biases in statistical power for our tests.

171
172 One potential explanation for the differences between the disease-associated genes and all
173 other protein coding genes is that the genes associated with any disease have distinct gene
174 expression profiles only because they are related to human traits and diseases. Therefore, we
175 hypothesized that if we compare our lists of disease genes with other genes that are more
176 similar in their connection to human traits (i.e. we run *t*-tests comparing disease genes vs other
177 disease genes, instead of disease genes vs any protein coding gene) the differences in gene
178 expression levels would be attenuated. To further investigate this, we repeated our analyses
179 using two more stringent control groups. These control groups consisted of genes associated
180 with diseases identified through GWAS, and were extracted from the Open Targets resource⁵²
181 (**Methods**). Results using Open Target control genes are consistent with analyses using all
182 protein-coding genes. In fact, the differences in absolute and relative expression between
183 disease-associated genes and Open Targets control genes are more significant (**Supplemental**
184 **Figs 3-10** and **Supplemental Table 5**). For example, AD-associated genes are more expressed
185 than Open Target control genes across all tissues. For CAD, only the brain does not show a
186 significant difference between CAD-associated genes and control genes. For SCZ, disease-
187 associated genes show higher expression in testis and pituitary, in addition to the brain tissues.
188 Taken together, these results demonstrate that disease genes present higher expression in
189 particular set of tissues, even if a more restrictive criteria for the control genes group is used.

190

191 **Gene Expression to GWAS**

192 In this section we test whether genes with high absolute and relative expression in a tissue are
193 enriched in GWAS signal. Originally proposed by Skene et al.⁵³, this approach utilizes the
194 software MAGMA³² to assess the enrichment of GWAS among genes in the top decile of

195 absolute and relative (also called specific) gene expression (see **Methods**). We expand the
196 original approach here since we also investigate *absolute gene expression* – in addition to
197 *relative gene expression* – to assess whether absolute expression may also provide valuable
198 insights for highlighting relevant tissues.

199
200 Our findings are overall consistent with the results obtained by the *t*-tests and the Anderson-
201 Darling tests (**Fig 2**, red columns). For SCZ, BP and ADHD, MAGMA results were nominally
202 significant in numerous brain tissues (P -value < 0.05), such as cortex, anterior cingula,
203 hippocampus, amygdala, or cerebellum or nucleus accumbens. The strongest results were
204 between cortex and anterior cingula tissues and SCZ (P -value $< 10^{-12}$). For CAD, numerous
205 tissues show significant P -values for both absolute and relative expression, with arteries (P -
206 value $< 10^{-6}$), colon sigmoid (P -value = 4.67×10^{-6}) and esophagus (P -value = 8.44×10^{-5}) having
207 the strongest enrichment of GWAS signal among their most specific genes. In IBD, the intestine,
208 blood, testis, liver, lung, and spleen are the tissues with strongest enrichment (P -values $< 10^{-3}$),
209 while AD shows the strongest result in spleen and blood (P -values $< 10^{-6}$). For Vitamin D, liver is
210 the most relevant tissue (P -value = 2.52×10^{-3}). For T2D, none of the tissues showed significant
211 results. Overall, MAGMA enrichments for relative gene expression are more pronounced than
212 for absolute gene expression. MAGMA results for each disease, each tissue, and each gene list
213 are included in **Supplemental Table 6**.

214

215 **Systematic Literature Search**

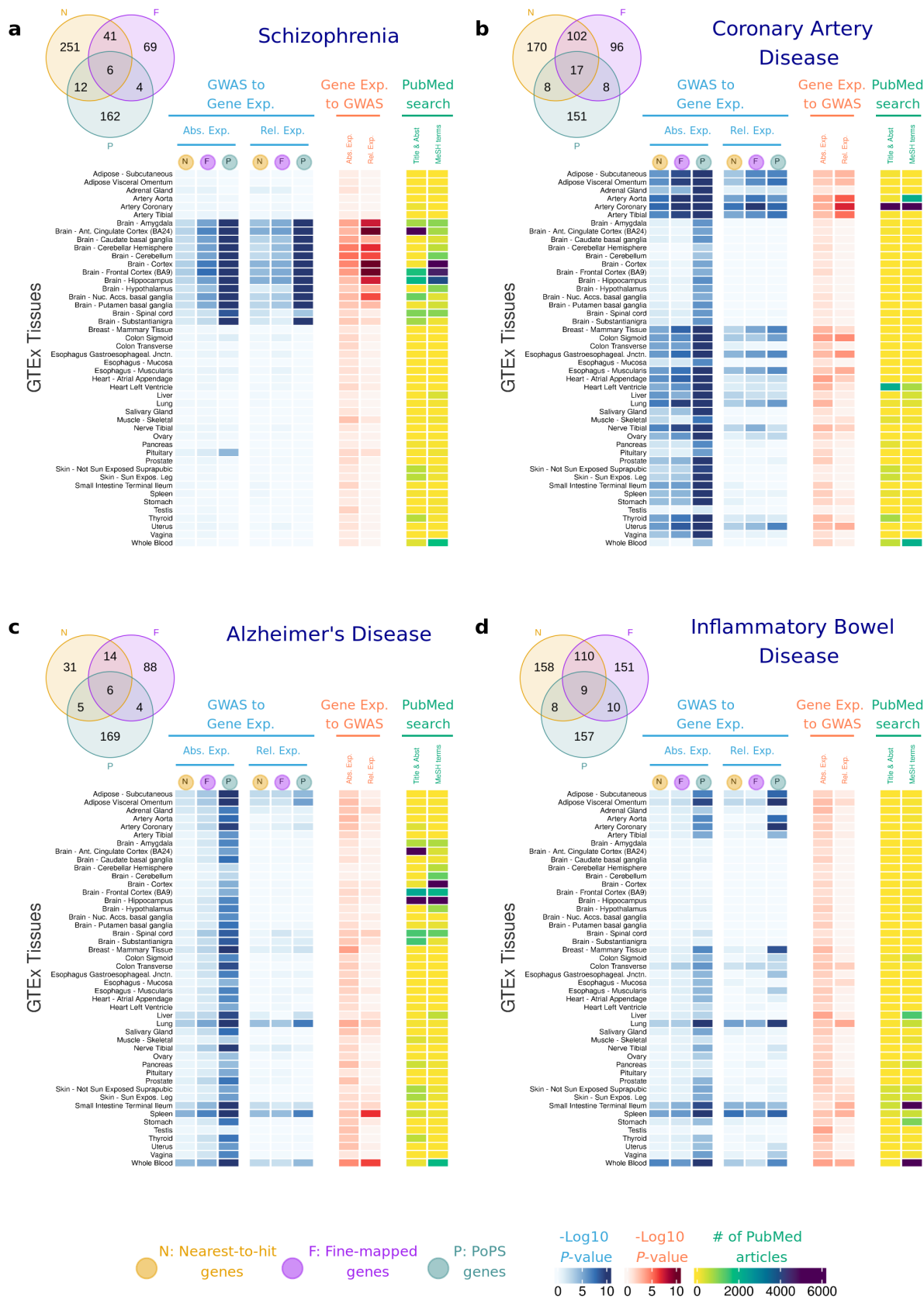
216 We investigated disease-tissue associations by cross-referencing our findings with PubMed
217 data using two methods to construct the PubMed search queries. For the first method, we use
218 Medical Subject Headings (MeSH) terms, a standardized vocabulary from the National Library
219 of Medicine. For the second method, we identify tissue/disease pair names in the title and
220 abstract of the PubMed articles. Both methods provide consistent results (**Fig 2**, yellow
221 columns). While the PubMed search results largely support our findings, there are additional
222 tissues identified that may be understudied, as the number of occurrences in the literature is
223 low. For instance, the spleen in relation to AD and IBD, as well as tissues associated with the
224 digestive system in the context of CAD, offer promising avenues for further exploration. Results
225 with the number of papers found for each query are included in **Supplemental Table 7 & 8**.

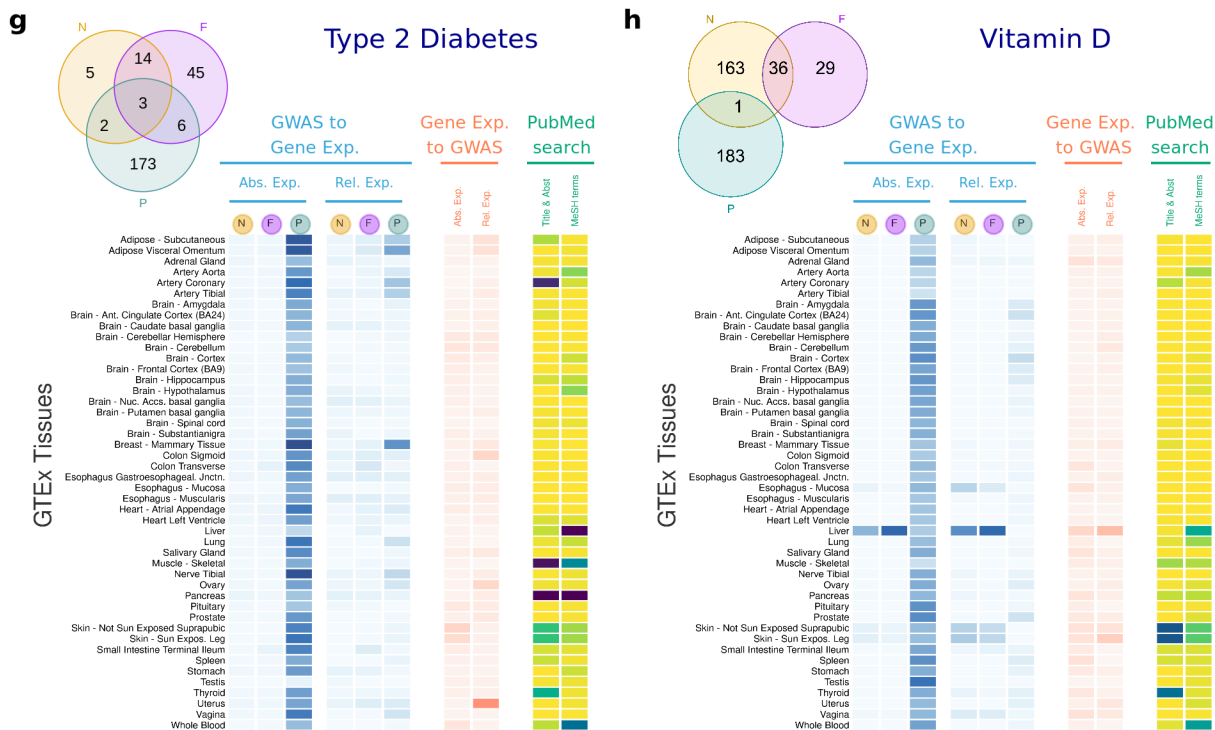
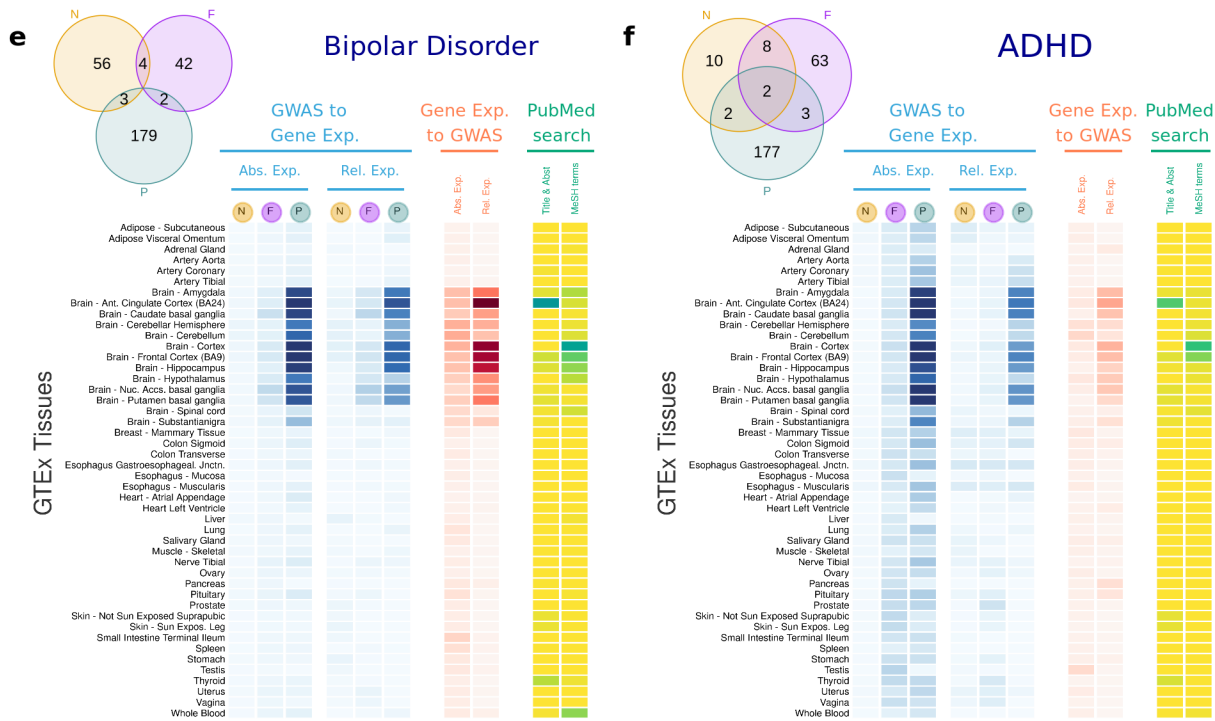
226

227 **Correlation in results across the three alternative strategies**

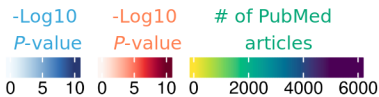
228 Across the different sets of tests, we observe the strongest correlation of results between the *t*-
229 test and Anderson-Darling tests (mean correlation across diseases and gene lists $r = 0.729$ and
230 42/48 correlation tests P-values $< 1 \times 10^{-3}$), and between *t*-tests and MAGMA (mean $r = 0.547$
231 and 35/48 correlation test P-values $< 1 \times 10^{-3}$). Correlations among the three approaches (GWAS
232 to gene expression, gene expression to GWAS, and PubMed search) varied widely across
233 diseases: the disease with highest correlation of results was SCZ, followed by CAD and IBD
234 (mean correlation across tests $r = 0.825$, $r = 0.707$ and $r = 0.691$, respectively with most P-
235 values $< 1 \times 10^{-5}$). Finally, the trait with the lowest correlation was T2D, ($r = 0.269$). Detailed
236 correlation results are included in **Supplemental Table 9**.

237





N: Nearest-to-hit genes F: Fine-mapped genes P: PoPS genes



240 **Figure 2.** Heatmap showing results of the association between gene expression in each GTEx tissue and **a**, Schizophrenia; **b**,
241 Coronary Artery Disease; **c**, Alzheimer's Disease; **d**, Inflammatory Bowel Disease; **e**, Bipolar Disorder; **f**, ADHD; **g**, Type 2 Diabetes;
242 **h**, Vitamin D. In blue, results showing the Log_{10} P -value for a one-side t -tests, testing the null hypothesis that disease-associated
243 genes are not more expressed than other protein-coding genes expressed in that tissue. In red, results showing the Log_{10} P -value
244 for enrichment of GWAS signal across the set of genes with highest absolute and relative expression for each tissue. In yellow,
245 results for the Literature Search using PubMed. Abs. Exp, Absolute expression; Rel. Relative expression; F, Fine-mapped genes; N,
246 Nearest-to-hit genes, P, Polygenic Priority Scores genes; Title & Abst, Title and Abstract; MeSH, MeSH terms.

247

248 **The gene expression landscape of cancer genes**

249 Given that for a cancer we have a strong hypothesis of what is the primary tissue involved (e.g.
250 colon for colorectal cancer), we applied the 'nearest-to-hit' prioritization method to cancer traits.
251 Genetic variants associated with eight common cancers were prioritized via GWAS⁵⁴, and the
252 gene closest to each GWAS hit was selected (**Supplemental Table 10**). The number of genes
253 found was low for most cancer traits: 12 for bladder, 13 for kidney, 16 for lung, and 28 for ovary.
254 Colorectal (80 genes), prostate (120 genes) and breast (228 genes) were the cancers with
255 largest number of genes associated, and significant results were observed in these three
256 diseases with better powered GWAS: prioritized genes exhibit higher expression in prostate for
257 prostate cancer (P -value = 2.94×10^{-6}), in breast for breast cancer (P -value = 1.27×10^{-4}), and gut
258 tissues for colorectal cancer GWAS (colon P -value = 4.54×10^{-5} , small intestine P -value =
259 7.07×10^{-5}). The other cancer traits with low number of GWAS associations – and therefore
260 lower number of genes – showed non-significant tissue-trait association results (**Supplemental**
261 **Figure 11** and **Supplemental Table 11**).

262

263 We also performed sex stratified analyses for cancers where the cancerous tissue is only
264 available in one of the sexes (i.e. ovary, prostate, breast), and assessed the expression of
265 disease genes across all GTEx tissues in men and women separately. Results were as
266 expected (**Supplemental Figure 12** and **Supplemental Table 12**): Despite the low number of
267 genes identified in ovary cancer, the Anderson-Darling test shows a significant association for
268 vagina in women (P -value = 7.62×10^{-6}), and genes related to prostate cancer are highly
269 expressed in men. In addition, the association of breast and breast cancer is less significant in
270 men (Rel. Expression P -value=0.0018) than in women (Rel. Expression P -value=0.00012).

271

272 **Sex-stratified analyses for T2D**

273 Since results show T2D genes in breast present higher expression than control genes, we
274 performed sex-stratified analyses to see whether results were driven by one of the sexes. We

275 repeated our analysis and compared the gene expression of disease vs control genes in men
276 and women separately (**Methods**).

277

278 For T2D, disease associated genes are more expressed in breast for both sexes
279 (**Supplemental Figure 13, panels a and b**), although P -values were slightly lower for men (P -
280 values for Rel. and Abs. Expression $< 10^{-6}$) than for women (P -values for Rel. and Abs.
281 Expression $< 10^{-4}$). Significant results were also observed for adipose tissues (P -values range:
282 10^{-3} to 10^{-13}), pituitary in the case of women (PoPS P -value=0.001), and testis in the case of
283 men (PoPS P -value= 5.68×10^{-12}). Significant results were observed only for the PoPS gene list,
284 and had smaller P -values in the Anderson-Darling test than in the t -tests (**Supplemental Table**
285 **13**).

286

287 To investigate whether specific genes are driving the T2D results in a sex-specific manner, we
288 examined the expression of individual genes for each gene list (nearest-to-hit, Fine-mapped,
289 PoPS). Across all tissues with significant P -values, the gene Thymosin Beta 10 (TMSB10) is
290 highly expressed (**Supplemental Fig 13, panel c**). TMSB10 plays an important role in the
291 organization of the cytoskeleton by binding to actin monomers, and therefore inhibiting actin
292 polymerization. Multiple studies have reported TMSB10 upregulation in cancer⁵⁵⁻⁵⁸, including
293 pancreatic cancer^{59,60}. Moreover, repositories like *MalaCards* and *Gene Cards* report the
294 association between pancreatic cancer and TMSB10 as highly relevant. The relationship
295 between TMSB10, T2D and pancreatic cancer is particularly interesting, given that T2D has
296 been consistently associated with pancreatic cancer in previous epidemiological studies, with a
297 two-fold higher risk of developing pancreatic cancer among diabetes patients^{61,62}.

298

299 **Impact of highly expressed genes**

300 To investigate whether the observed signal is primarily influenced by a small subset of genes
301 that are highly expressed, we excluded genes within the top 10% of absolute and relative
302 expression in *relevant tissues*, defined as the tissues where disease-associated genes were
303 significantly more expressed than control genes (**Methods**). The number of tissues removed,
304 and number of disease-associated genes are listed in **Supplemental Table 14**. We repeated
305 the t -test and Anderson-Darling tests with these new lists of disease-associated genes. Results
306 show that, while the P -values increase for all the tests, results remain consistent after removing
307 the top 10% expressed genes (**Supplemental Fig 14** and **Supplemental Table 15**). For
308 example, the cortex and cingulate cortex are significantly associated with schizophrenia

309 (Absolute expression P -values $< 10^{-20}$), and the tissues most associated with AD remain being
310 small intestine (P -value = 3.82×10^{-9}), spleen (P -value = 4.91×10^{-8}) and lung (P -value = 5.01×10^{-8}).
311

312

313 **The gene expression landscape at the cell-type level**

314 Tissues are composed by different cell types, each of them expressing gene expression
315 programs to perform specific functions. The cell types in each tissue and their relative
316 proportions may affect the results observed in the previous sections, since abundant cell types
317 will be better powered than rare cell types for detecting differences between disease-associated
318 genes and control genes. To further investigate where in the body disease genes operate, we
319 repeated our testing framework in a set of cell types and tissue regions extracted from the (i)
320 Tabula Sapiens³⁴, a dataset which accrues nearly 500,000 cells from 24 different tissues and
321 organs, many from the same donor, and (ii) ARCHS4³³, a resource that aggregates RNA-seq
322 data from the Gene Expression Omnibus and the Sequence Read Archive.

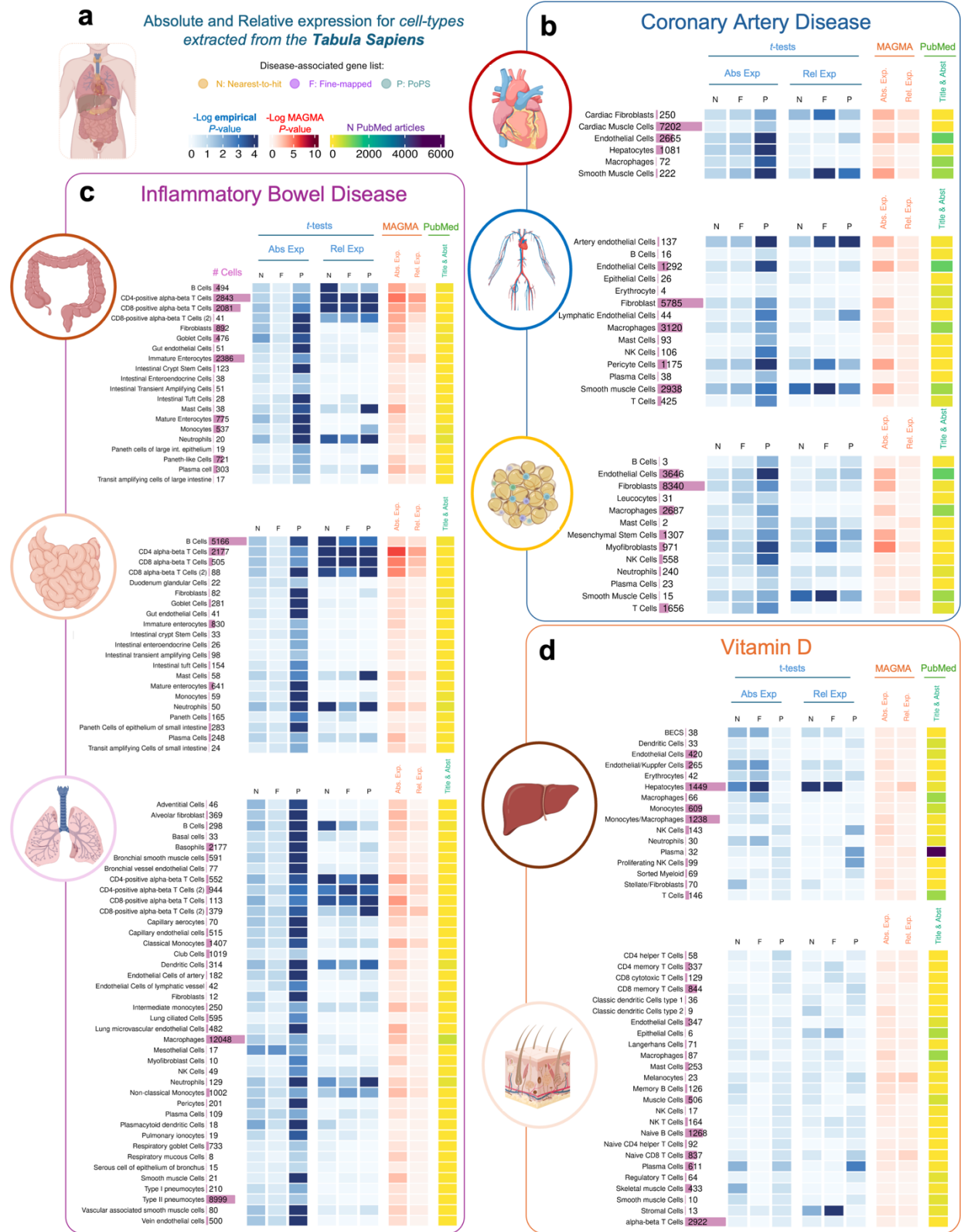
323

324 **Fig 3** presents the results of the analyses conducted with a set of tissues from Tabula Sapiens
325 for Vitamin D, IBD and CAD, and **Fig 4** presents the results of the analyses for AD using tissue
326 regions and cell type-level datasets from both ARCHS4 (**Fig 4a**) and Tabula Sapiens (**Fig 4b**).
327 Results for the other diseases using ARCHS4 can be found in **Supplemental Fig 15** and
328 **Supplemental Tables 16-18**. We focus on Vitamin D, AD and IBD because significant results
329 were observed in “non-typical” tissues such as the spleen (primarily composed of immune cells)
330 and lung. We also focus on CAD because this disease shows the largest number of associated
331 tissues. Overall, the results for ARCHS4, Tabula Sapiens and GTEx are consistent, although the
332 P -values for all cell-type analyses tend to be higher. In the Tabula Sapiens, PoPS genes show
333 higher expression than control genes, but these results are often not replicated in fine-mapped
334 genes or nearest gene lists (**Supplemental Tables 19-21**). Therefore, only results that replicate
335 in at least two gene lists are reported in the following paragraphs.

336

337 Results for IBD in cell types derived from the small intestine, large intestine, and lung show that
338 IBD-associated genes have higher expression in T-cells (P -values for Relative expression $<$
339 0.02). Other immune cell types such as B cells (P -values < 0.038), neutrophils (P -values $<$
340 0.034) and dendritic cells (P -values < 0.01) also showed significant differences, albeit with
341 larger P -values. These findings were consistent with MAGMA results.

342



343
 344
 345
 346

Figure 3. Heatmap showing results of the association between gene expression in each cell-type obtained from the Tabula Sapiens dataset **a**, Tabula Sapiens datasets were downloaded for each tissue, and absolute and relative expression was calculated for each cell type in each tissue. **b**, Results for Coronary Artery Disease. **c**, Results for Inflammatory Bowel Disease. **d**, Results for Vitamin

347 D. In blue, results showing the Log_{10} empirical P -value after running 10,000 permutations, testing the null hypothesis that disease-
348 associated genes are not more expressed than other protein-coding genes expressed in that tissue. In red, results showing the
349 Log_{10} P -value for enrichment of GWAS signal across the set of genes with highest expression for each tissue. In yellow, results for
350 the Literature Search using PubMed. Abs. Exp, Absolute expression; Rel. Exp, relative expression; F, Fine-mapped genes; N,
351 Nearest-to-hit genes, P, Polygenic Priority Scores genes; Title & Abst, Title and Abstract are used in the PubMed Search.

352
353 For CAD, the most significant differences in expression between disease-associated and control
354 genes were observed in the relative expression of T cells in the vasculature system (P -value =
355 0.0028). Additionally, we found significant differences in relative expression in smooth muscle
356 cells in the heart (P -value < 0.0271) and fat tissues (P -value < 0.0438 for fine-mapped genes,
357 P -value < 0.0138 for PoPS and nearest genes) and cardiac fibroblasts in the heart (P -value <
358 0.045). Various immune cell types such as macrophages, mast cells, and NK cells in the
359 vasculature system were significant (P -value < 0.0146) but only for PoPS genes.

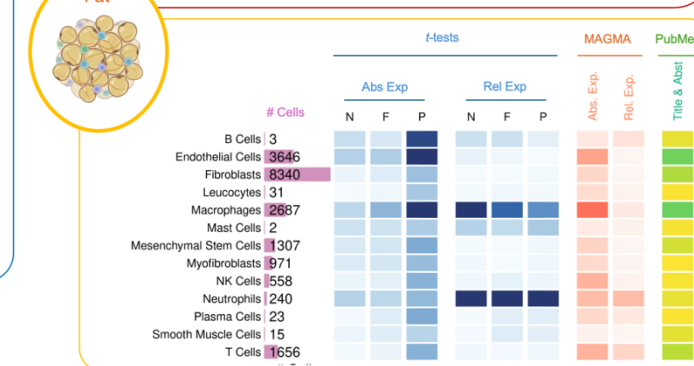
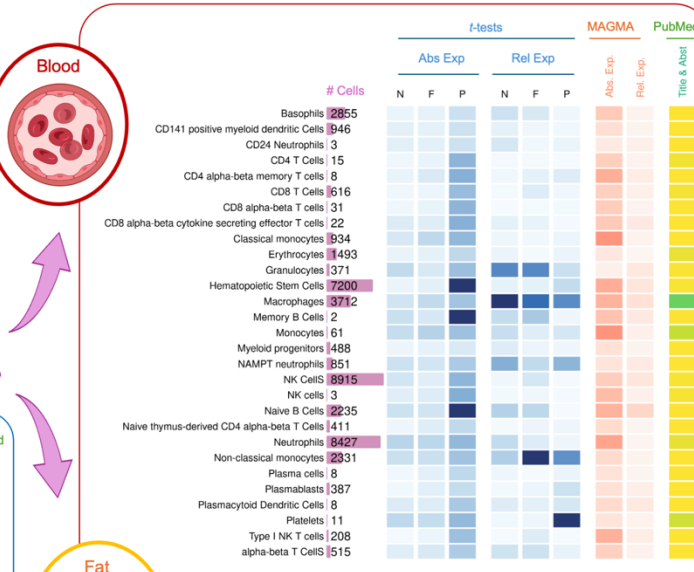
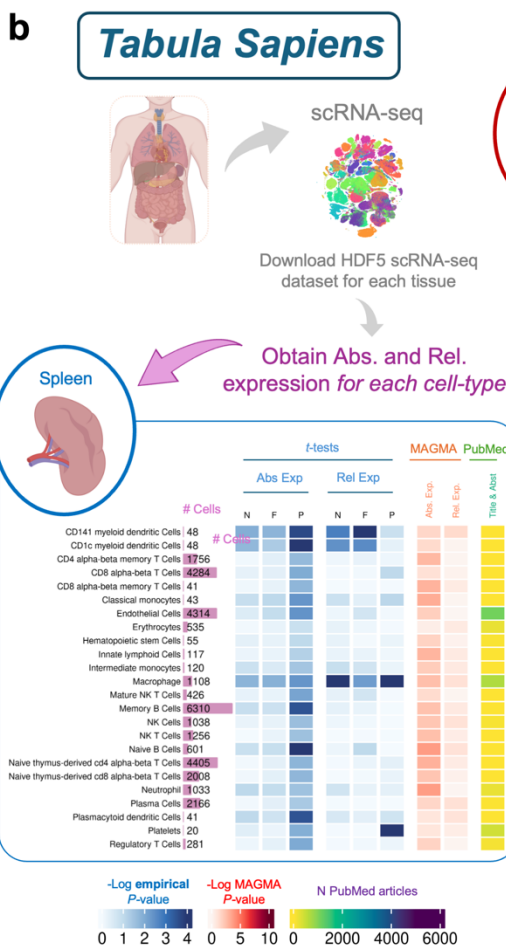
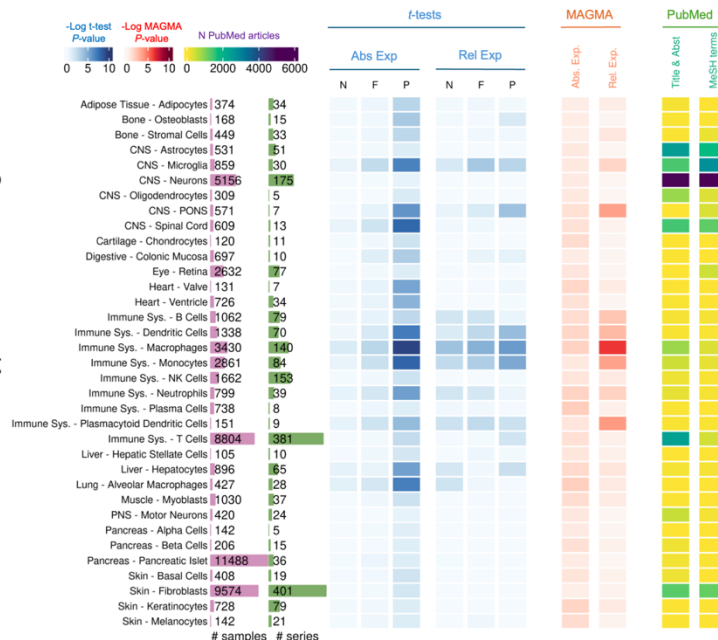
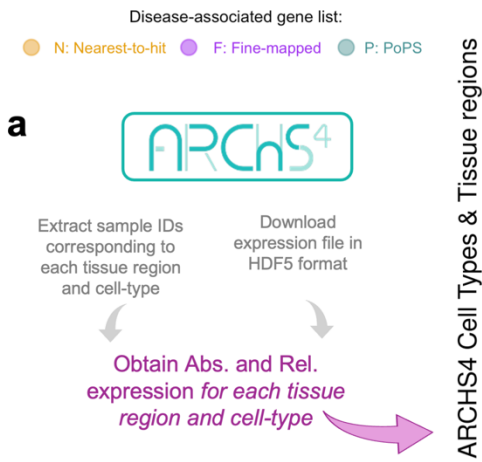
360
361 For Vitamin D, the results aligned with the GTEx and ARCHS4 analyses. We observed the most
362 significant difference in expression in the liver, specifically in hepatocytes, between disease-
363 associated and control genes (P -values < 0.0012).

364
365 In the case of AD, **Fig 4a** shows results using ARCHS4, where disease-associated genes
366 present high absolute and relative expression in immune-related cell types (e.g. Absolute and
367 relative expression P -values < 0.05 for dendritic cells, macrophages and neutrophils). **Fig 4b**
368 shows results using Tabula Sapiens, where similar patterns emerge as in ARCHS4. For
369 instance, macrophages (in the spleen, blood, and fat) and neutrophils (in fat) show significant
370 differences between disease-associated and control genes (P -values < 0.03). Since brain
371 tissues were not available in Tabula Sapiens, we could only look at brain-related cell types in the
372 ARCHS4 dataset (**Supplemental Fig 15**). Microglia – known to play a critical role in AD⁶³ – is the
373 sole brain tissue significantly associated (P -value = 9.72×10^{-3} for relative expression). Genes
374 linked to SCZ, BP, and ADHD show significant absolute and relative expression in motor
375 neurons in ARCHS4 (P -value = 0.00021 for PoPS genes; P -value = 0.036 for fine-mapped
376 genes; P -value > 0.05 for nearest-to-hit genes), but not in the broader category of neurons
377 (even though the sample size and number of studies is larger for this cell type).

378
379 Tabula Sapiens and ARCHS4 results offer insights not easily discerned at the tissue level. For
380 example, the results that we observe between AD and IBD in spleen, blood and lung are
381 probably driven by the high fraction of macrophages and other innate immune system cells

382 present within those tissues. However, both datasets have their own limitations: In the case of
383 ARCHS4, the representation of cell types and tissue regions is less systematic than in GTEx:
384 after quality control (**Methods**), 8 immune and 6 CNS-related cell types are included, with only
385 one related to the digestive system. Furthermore, ARCHS4's heterogeneity may have reduced
386 power to detect associations in other diseases despite correction of batch effects. While *Tabula*
387 *Sapiens*³⁴ may provide a more systematic multiorgan dataset at the cellular level, the scale of
388 this dataset is smaller (they measured single-cell RNA-seq data for a total of 15 individuals,
389 respectively, in contrast to e.g. GTEx, which assessed more than 700 individuals). Moreover,
390 some tissues of interest were not available here, such as brain tissues to interrogate cell types
391 related to AD, SCZ, BP or ADHD.

Alzheimer's Disease



392
393
394
395

Figure 4. Heatmap showing results of the association between gene expression in each cell-type obtained from the Tabula Sapiens dataset. **a**, Results using RNA-seq data from cell types and tissue regions extracted from the ARCHS4 resources. **b**, Results using RNA-seq data from the Tabula Sapiens for blood (red section), spleen (blue section), and fat (yellow section).

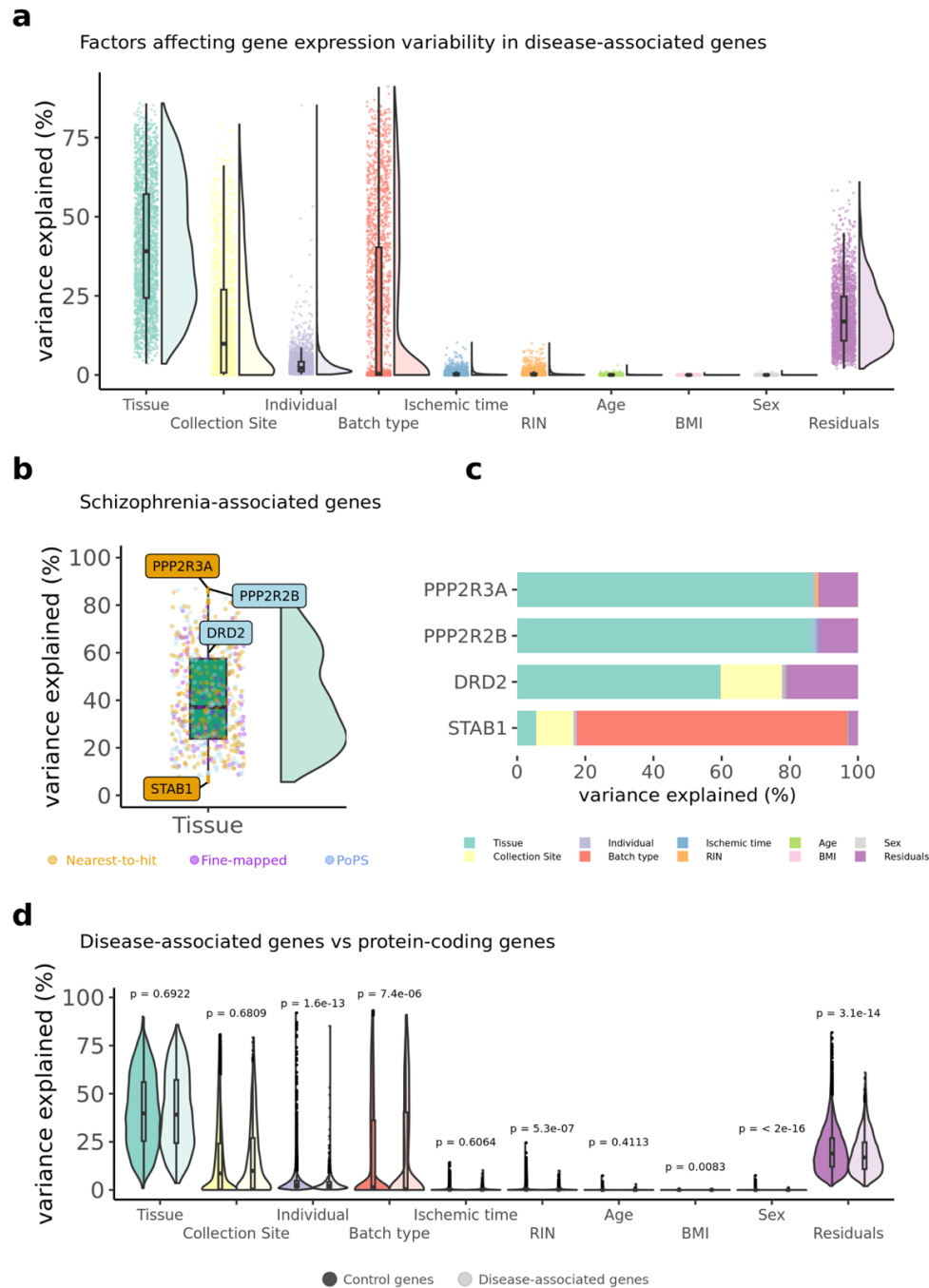
396 **Predictors of gene expression of disease genes**

397 In the previous sections, we demonstrated that disease-associated genes exhibit both high
398 absolute and high relative expression in certain tissues and cell types. Although previous
399 studies have shown that tissue type is an important predictor in gene expression⁶⁴, in this
400 section we expand this work by evaluating multiple factors that may contribute to gene
401 expression variability in disease-associated genes specifically. Additionally, we assess whether
402 the relative contributions of gene expression predictors differ significantly between disease-
403 associated and control genes.

404
405 To characterize the biological factors (such as tissue, sample ID, subject ID) and technical
406 factors (such as batch ID) that contribute to variability in the gene expression of each disease-
407 associated gene, we used the '*variancePartition*' R package (v.4.3)⁶⁴. *variancePartition* uses a
408 linear mixed model framework in which the expression values of *each gene* are the dependent
409 variable, and distinct sources of variation – such as those driven by tissue type, individual
410 differences, and technical effects – are the independent variables. Overall, tissue type explains
411 the highest proportion of the variance in gene expression (**Fig 5a** and **Supplemental Table 22**).
412 However, there is a lot of variability across individual genes; while for some genes tissue type
413 explains more than 70% of the variability in their expression, for other genes factors such as the
414 type of batch, collection site, explain <20% of the variance. **Fig 5b** and **c** shows SCZ fine-
415 mapped genes as an example of such variation in variance explained. **Supplemental Figs 16-**
416 **18** and the R Shiny website associated with this manuscript
417 <https://juditgg.shinyapps.io/diseasegenes/> include gene-level results quantifying the contribution
418 of each variable to the variation in expression of each gene and disease.

419
420 Given the differences in results across genes, we tested whether the *variancePartition* results
421 for disease-associated genes are different from all other protein-coding genes. The variance
422 explained by biological factors (e.g. tissue, individual) and technical factors (collection site,
423 batch type) in disease-associated genes was compared vs the variance in other protein coding
424 genes. Individuals, batch type, RNA Integrity Number (RIN), and sex exhibited small yet
425 significant differences in contributing to gene expression variability between disease-associated
426 and control genes. (**Fig 5d**). When comparisons are assessed for each gene list and disease
427 individually, results remain consistent. Only for Vitamin D, tissue-type explains a greater
428 proportion of the variance for disease-associated genes than for control genes (**Supplemental**
429 **Figs 19-21**).

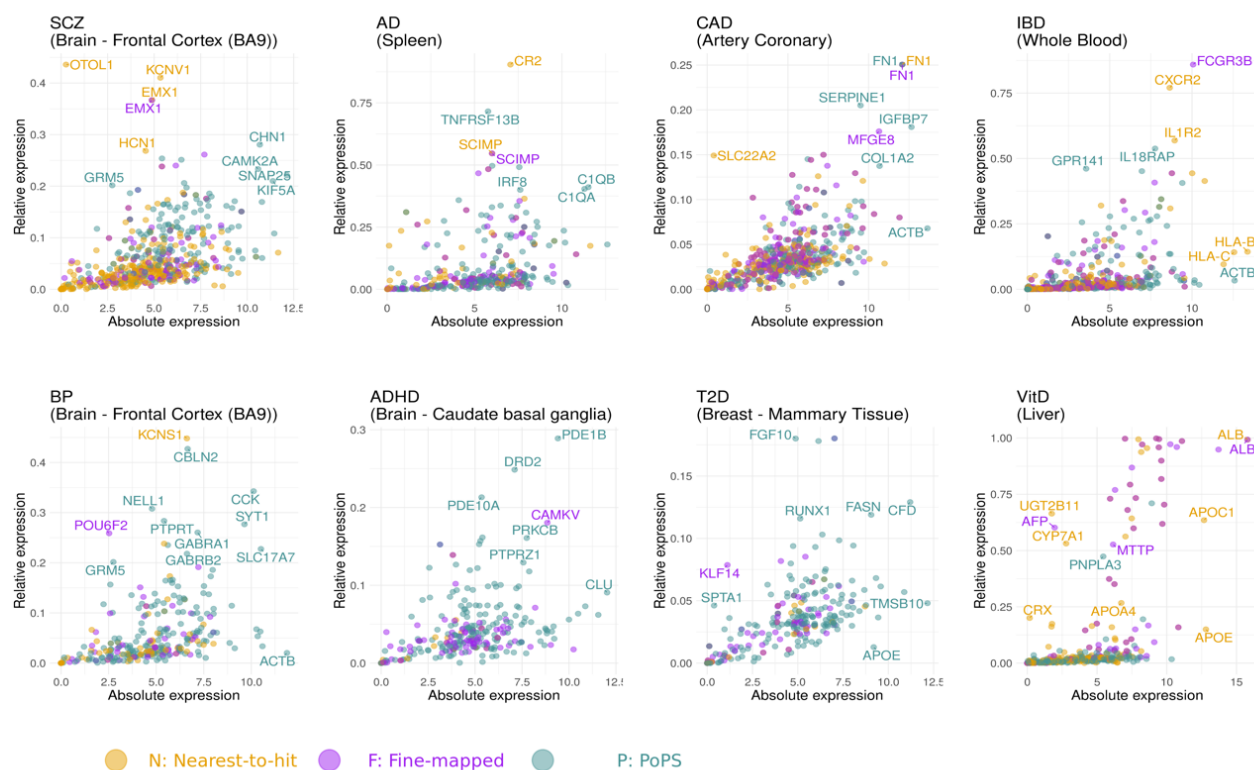
430



431

432 **Figure 5. Variance partition is used to investigate the factors influencing gene expression of disease-associated genes. a,**
 433 **Violin plots representing the distribution of variance partition across all disease-associated genes for the eight diseases investigated.**
 434 **b,** variance partition results for genes associated with schizophrenia. Genes labelled represent: two genes encoding protein
 435 phosphatases (PPP2R3A & PPP2R2B) where tissue-type explain a large fraction in gene expression variance, and a gene (STAB1)
 436 where tissue explains less than 10% in gene expression variance. The dopamine receptor 2 (DRD2) is also included because it is
 437 the main receptor for most antipsychotic drugs^{65,66}. **c,** Bar plots of individual genes showing the variance partition estimates at the
 438 individual gene level for genes highlighted in panel b, **d,** Violin plots showing the differences in variance partition results between
 439 disease-associated genes and control genes.

440
 441 **Joint evaluation of absolute and relative gene expression**
 442 New drugs underpinned by genetic evidence have a significantly higher success rate in clinical
 443 trials¹⁶. Consequently, genes prioritized via GWAS are often examined in experimental studies
 444 to validate their causal role in disease and understand their biological function. Given that (i) the
 445 *VariancePartition* results show wide variability among genes in the contribution that tissue type
 446 infers on gene expression, and (ii) the relative gene expression (a.k.a. tissue- and cell-type-
 447 specificity) of candidate target genes can inform drug efficacy^{18,67} and side effect prediction¹⁹,
 448 here we identify the genes with both high absolute and high relative expression across tissues.
 449 In **Figure 6**, we present the genes with both high absolute and high relative expression for the
 450 tissue with the most significant differences between disease-associated and control genes.
 451 Results for the rest of tissues can be found at the R Shiny website associated with this
 452 manuscript <https://juditgg.shinyapps.io/diseasegenes/>.
 453



454
 455 **Figure 6. Exploring disease-associated genes showing both high absolute and high relative expression.** Scatter plots
 456 showing the relationship between absolute gene expression and relative gene expression in disease-associated genes. Only the
 457 tissue with the most significant differences between disease-associated and control genes is shown.
 458

459 When assessing both high absolute and relative gene expression, we find that absolute
460 expression provides useful information beyond specificity. For example, ALB is a gene
461 associated with vitamin D that presents high absolute and relative expression in liver, FN1 is a
462 gene associated with CAD with high absolute and relative expression in the coronary artery. In
463 contrast, APOE presents high expression in tissues such as breast and liver but is not
464 specifically expressed in any of them (low relative expression), and OTOL1 is a gene associated
465 with schizophrenia, with high relative expression in the Frontal Cortex, but low absolute
466 expression.

467

468 Nevertheless, absolute expression alone is typically insufficient to identify the most suitable
469 tissues for validating individual disease-associated genes because these genes often maintain
470 elevated expression across various tissues. **Supplemental Figs 22 to 29** show the co-
471 occurrence of the 10 most expressed fine-mapped genes across diseases, showing a high
472 overlap of highly expressed genes across multiple tissues. These results – showing that many
473 disease genes are often expressed across many tissues – are in line with previous studies
474 showing that 46% of protein-coding genes are expressed in all tissues⁶⁸.

475

476

477 Discussion

478 In this study, we have systematically characterized the gene expression features of GWAS
479 prioritized genes. Genes associated with diseases exhibit higher absolute and relative gene
480 expression not only in the anticipated tissues and cell types (e.g. brain in SCZ, BP and ADHD),
481 but also in tissues and cell types not typically associated with the diseases (e.g. lung and spleen
482 in AD and IBD, motor neurons in psychiatric disorders, cells in the PONS associated with IBD).
483 Additional analyses removing genes with the highest expression and using more stringent
484 criteria for the control group showed similar results. Next, we explored which biological and
485 technical factors are significant predictors of gene expression in disease-associated genes.
486 Although tissue-type is a consistent key contributor to gene expression variability of disease-
487 associated genes in GTEx, results varied widely, with some disease gene showing batch and
488 subject ID as important predictors of gene expression. Finally, and given that (i) highly
489 expressed genes tend to maintain their elevated expression level across multiple tissues, and
490 (ii) tissue-specific genes are reported to be twice as likely as broadly expressed genes to be
491 drug targets^{16,17}, we highlight disease genes with both absolute and relative gene expression –
492 as these properties will be important for further experimental validation and drug target
493 development.

494
495 We first focus our study on tissue-level analysis, despite the intense focus in the field on cell-
496 types. We focus on tissues because: (1) extensive prior knowledge of disease-tissue
497 associations provides a “ground truth” and thus informs the benchmarking of our approach, (2)
498 multiple-organ, single-cell transcriptomic atlases – that systematically characterize the cell type
499 composition of tissues – have been performed on a limited number of individuals, since they
500 require high-coverage sequencing to obtain highly accurate single-cell expression profiles, (3)
501 cell types present different phenotypic properties at multiple levels, which make them difficult to
502 define and categorize²⁸. Despite these challenges, we also leveraged RNA-seq data from the
503 ARCHS4 and Tabula Sapiens resources to explore in what cell types and tissue regions GWAS
504 signal and gene expression converges. ARCHS4 and Tabula Sapiens results highlight the gain
505 in specificity that can be obtained when the relevant cell types are rare (e.g. microglia in the
506 brain, where much of the GWAS signal for Alzheimer’s disease resides, but composes only ~7%
507 of non-neuronal cells in the brain⁶⁹). They also explain some tissue-level associations driven by
508 the presence of relevant cell types that can be found in multiple tissues (e.g. the presence of
509 monocytes, relevant for AD and IBD, in spleen, lung, blood etc.). However, the ARCHS4 and
510 Tabula Sapiens analyses also highlight the challenges related to using single cell and cell-type

511 datasets stated above. Examples of these challenges include capturing measures of gene
512 expression for cell types with a very low number of cells that may be obtained from the same
513 individual, and accurately defining what is a ‘cell-type’ given a gene expression program (e.g.
514 the cell type ‘neuron’ is abundant in ARCHS4 but likely heterogenous, resulting in no significant
515 results between neurons and psychiatric disorders).

516
517 Our approach differs from some other approaches that combine GWAS and functional genomic
518 data to make inference disease etiology. For example, Transcriptome-Wide Association Studies
519 (TWAS) integrate GWAS findings with expression quantitative loci (eQTLs) to investigate how
520 genetic variations influence gene expression⁴⁷. TWAS relies on the availability of the eQTL data
521 and on genes with highly heritable gene expression⁷⁰. In contrast, our strategy provides a gene
522 expression profile of disease-associated genes regardless of eQTL data availability and gene
523 heritability, broadening the scope for combining GWAS signal and gene expression.

524
525 Identifying disease-associated genes that are active in a wide range of tissues, including
526 unexpected ones, is crucial because drugs often cause side effects in the tissues where their
527 target genes are active⁷¹. By providing comprehensive expression profiles of disease-
528 associated genes, we aim to support future research in validating candidate genes and
529 developing drug targets more effectively. However, our study has several limitations. First, no
530 gene prioritization methods are perfect, and therefore it is possible that some genes categorized
531 as ‘disease-associated’ may not significantly contribute to disease. Additionally, the fine-mapped
532 and PoPS prioritization approaches used functional genomics data such as tissue and cell-type
533 specific RNA-seq, which may lead to some circularity in the analyses. Unlike previous studies⁷²,
534 the focus of this work is not to benchmark different prioritization methods, but to follow up on
535 previously prioritized genes to assess their expression in the body; Second, our analyses were
536 performed at the gene level, and therefore alternative mRNA transcripts were not explored here;
537 Third, this study mainly used the GTEx dataset, considered a population control that is “normal”
538 relative to the age of the individual and where the tissues are considered healthy. The gene
539 expression profiles represent a healthy state and does not explore the dynamics of gene
540 expression during disease states. However, we propose that – before studying the dynamics of
541 gene expression between cases and controls – it is important to understand the tissues and cell
542 types where these genes show high expression and specificity; Fourth, we observe that factors
543 such as age, sex, and batch have minimal but varying effects on different genes. However, we
544 did not regress out these covariates prior t-tests or AD tests: Instead, for the GTEx dataset and

545 ARCHS4, we obtained the median TPM for each gene across all samples. For the Tabula
546 Sapiens analyses, we calculated *P*-values based on a permutation procedure, which generates
547 a null hypothesis drawn from the data itself but does not explicitly adjust for covariates.

548
549 In conclusion, our study on ‘the gene expression landscape of disease genes’ not only confirms
550 established links between diseases and tissues, but also identifies unexplained tissue and cell-
551 type-disease associations that warrant further investigation. This systematic characterization of
552 the gene expression features of high-confidence disease genes opens new avenues for guiding
553 experimental follow-up and drug design, ultimately advancing our understanding of disease
554 mechanisms and response to treatment.

555

556 **Methods**

557

558 **RNA-seq datasets**

559 **GTEX dataset**

560 Gene expression measurements were obtained for 50 tissues from the GTEx project³⁵ version
561 8. Median gene TPMs for each tissue were downloaded from
562 <https://gtexportal.org/home/datasets>. Standard RNA-seq processing steps were applied to the
563 dataset as follows: (1) we filtered out all non-protein-coding genes and genes not expressed in
564 any tissue; (2) we removed the tissues with less than 100 samples, cancer or cell related tissue
565 types (i.e. EBV-transformed lymphocytes and Leukemia cell lines); (3) we scaled the expression
566 of each tissue such that the total is 10^6 TPM. 45 tissues remained after quality control.

567

568 **Sex-stratified analyses in the GTEX dataset**

569 In tissues for which we wanted to test whether the disease-associated genes present higher
570 absolute or relative expression in men and women specifically (i.e. T2D), we extracted the
571 GTEx expression dataset, and for each tissue, the median gene expression of all genes was
572 calculated for women and men separately. For each sex, we then created absolute and relative
573 expression datasets (where columns represent tissues, and rows represent genes). To prepare
574 sex-specific inputs for the MAGMA analyses, we generated GMT files for men and women
575 separately, which contain the gene sets composed of the top decile of absolute and relative
576 gene expression for each sex. T-tests, Anderson-darling tests, and MAGMA analyses were also
577 performed separately for men and women.

578

579 **ARCHS4 datasets**

580 Gene expression across cell types and tissue regions were extracted from the ARCHS4³³
581 resource, which provides access to uniformly processed gene counts from human RNA-seq
582 experiments stored in the Gene Expression Omnibus (GEO) and Sequence Read Archive
583 (SRA). Using the ARCHS4 web browser (<https://maayanlab.cloud/archs4/>), we systematically
584 identified all the cell types available in the metadata search menu. For some disease-relevant
585 cell types like microglia in AD, we entered the cell type name directly into the metadata search
586 bar. ARCHS4 generates R scripts listing the samples related to each cell type, to facilitate their
587 extraction from the main repository –a HDF5 file named "human_gene_v2.2.h5" available for
588 download at <https://maayanlab.cloud/archs4/download.html> (Downloaded version date: 5-30-
589 2023).

590
591 The quality control of ARCHS4 datasets and *per-gene* TPM calculation was as follows: Only cell
592 types with more than 100 samples across all experiments available in ARCHS4 were included in
593 our analyses. Upon obtaining the counts expression matrix for each cell type, we performed
594 quantile normalization of samples using the function '*normalize.quantiles*' available on the R
595 package ("preprocessCore"). Quantile normalization was performed on raw counts. Given that
596 samples from a specified cell type may originate from multiple experiments with slightly different
597 conditions, we (1) excluded experiments containing less than 10 samples, and (2) adjusted for
598 batch effects using the package *ComBat_seq*⁷³, which is an improved version of the popular
599 *ComBat*⁷⁴. Unlike its predecessor *ComBat* (designed for microarray data), *ComBat_seq* is
600 tailored for RNA-Seq studies and it does not assume a normal distribution of gene expression
601 data.

602
603 After batch correction, median TPM values for each gene were calculated in each cell type
604 using the formula:

$$605 \quad \text{TPM} = \left(\frac{\text{Number of mapped reads for gene}}{\text{Gene length in kilobases (kb)}} \right) \times 10^6 \div (\text{Total number of mapped reads})$$

606 where 'Gene lengths in kilobases' were calculated using the genomic coordinates indicated in a
607 GTF file (built GRCh38), downloaded from ENSEMBL.

608
609 **Tabula Sapiens datasets**
610 scRNA-seq datasets were obtained from the *Tabula Sapiens* figshare
611 (https://figshare.com/articles/dataset/Tabula_Sapiens_release_1_0/14267219). These datasets,

612 initially in '.h5ad' format, contained gene counts for each cell and metadata, and were converted
613 into Seurat objects and then into '*SingleCellExperiment*' objects to ensure compatibility with
614 downstream analysis tools.

615
616 Data quality processing was performed: cells with zero counts across all genes were removed,
617 and outlier cells with an extreme number of detected genes were excluded. Pseudobulk data
618 was then generated using the *aggregateToPseudoBulk* function from the *dreamlet* R package to
619 aggregate expression counts across cell types, according to the free cell type annotation
620 included in the original datasets.

621
622 **Calculating 'Absolute gene expression' and 'Relative gene expression' values**

623 Since the expression patterns of protein-coding genes tend to follow a negative binomial
624 distribution, we calculated *absolute levels of gene expression* by taking the Log_2 of the median
625 TPM+1 values. To calculate *relative gene expression*, we divided the *absolute levels of gene*
626 *expression* of each gene by its total expression across tissues. The resulting relative gene
627 expression ranged from 0 (gene is not expressed) to 1 (gene is exclusively expressed in this
628 tissue). The Log_2 of the absolute and relative expression measures were used in subsequent
629 analyses.

630
631 To calculate absolute and relative gene expression values for each gene in the *Tabula Sapiens*
632 dataset, the total expression counts for each cell type was obtained, and normalization was
633 performed using the *calcNormFactors* function from the *edgeR* R package. Absolute gene
634 expression values were calculated as the total number of counts per million for each gene in
635 each cell type (after normalization). Data was organized in data tables where each row
636 represented a gene, and each column represented a cell type. For relative expression of a gene
637 in a cell type, the *cellTypeSpecificity* function – which calculates the number of counts of a gene
638 in a cell type divided by the total number of counts across cell types in that tissue – was utilized
639 after applying the same normalization procedure as for absolute expression.

640
641 **GWAS to Gene Expression**

642 **Definition of disease-associated genes inferred from GWAS results.**

643 Three different types of gene lists were inferred for each disease using GWAS results: nearest-
644 to-hit genes, fine-mapped genes, and PoPS genes. For all the gene definitions, we obtained

645 each gene ENSEMBL IDs using a GTF file obtained from ENSEMBL (built GRCh37.75). The list
 646 of prioritized genes for each approach and disease can be found in **Supplemental Tables 1-3**.

647

648 - *Definition of ‘nearest-to-hit’ genes*: To find the genes closest to the GWAS hit, we obtained
 649 publicly available GWAS summary statistics for the diseases investigated. We performed
 650 clumping using PLINK 1.9⁷⁵ and individual level genotype data from the UK Biobank as a
 651 reference linkage-disequilibrium (LD) panel (UK Biobank Resource under application number
 652 18177). During clumping, variants with P -values $\leq 5 \times 10^{-8}$ were retained, and variants within a
 653 250 Kbp window correlated ≥ 0.5 with the index variant or variants with P -value ≥ 0.01 were
 654 removed. For each clump, the nearest protein-coding gene to the index variant was identified.
 655 We used a GTF file obtained from ENSEMBL (built GRCh37.75) to extract the gene start and
 656 gene end coordinates of each protein-coding gene. Information about the GWAS used, number
 657 of clumped variants, genes identified and distance between variant and nearest gene is
 658 included in **Table 1**.

659

660 **Table 1**. Gene prioritization based on the nearest gene to GWAS hit. Table shows references for
 661 each GWAS summary statistics, the number of clumped SNPs, genes and median distance
 662 between SNP and selected gene.

Trait	GWAS summary statistics used	N clumped SNPs	N unique ensemble IDs	Median distance between SNP and nearest protein coding gene (bp)
CAD	Aragam et al, 2022 ³	525	297	14,734
SCZ	Trubetskoy et al. 2022 ⁴	452	310	18,125
IBD	Liu et al, 2015 ⁷⁶	487	285	11,724
AD	Bellenguez et al, 2022 ¹¹	239	99	8,363
BD	Mullins et al, 2021 ⁷⁷	73	63	20,660
ADHD	Demontis et al, 2023 ⁸	32	22	104,008
T2D	Suzuki et al, 2023 ⁷⁸	50	31	26,633
Vitamin D	Revez et al, 2020 ⁶	517	200	22,111

663

664 - *Definition of ‘fine-mapped genes’*: We acquired gene lists that previously fine-mapped GWAS
 665 for CAD³, SCZ⁴, IBD⁵, AD⁷⁹, BD⁷, ADHD⁸, T2D⁹ and Vitamin D⁸⁰. Most gene lists were
 666 constructed using a combination of statistical fine-mapping, transcriptome association studies,
 667 and mendelian randomization.

668

669 For CAD, the integration of eight gene prioritization predictors enabled the identification of 220
 670 likely causal genes³. For SCZ, statistical fine-mapping was integrated with summary Mendelian

671 randomization and Hi-C interaction data to obtain a list of 120 prioritized genes⁴. For IBD, we
672 extracted the list of genes linked to variants fine-mapped, available in the Supplemental material
673 of the study conducted by Huang and colleagues⁵. For AD, we used a review of that reported a
674 list of genes prioritized via fine mapping of GWAS in two previous studies^{10,11}. The list is
675 available in https://github.com/sjfandrews/ADGenetics/blob/main/results/adgwas_loci.csv. For
676 BP, fine-mapping of the GWA signals was performed and seven complementary approaches
677 were used to prioritize 47 credible genes that were mapped to loci by at least three of the seven
678 approaches⁷. For ADHD, fine-mapping of the most recent ADHD GWAS⁸ identified sets of
679 credible variants for each risk locus. Credible sets were subsequently linked to genes based on
680 genomic position, information about eQTLs, and chromatin interaction mapping in human brain
681 tissue as implemented in FUMA. For T2D, we used a gene list containing the nearest gene of
682 the results of a fine-mapping approach used in 380 independent association signals⁹. For
683 Vitamin D, we extracted a list of genes published by *Manousaki* and colleagues⁸⁰, who
684 prioritized genes using the DEPICT method⁴⁵ on a GWAS of serum 25 hydroxyvitamin D.

685
686 - *Definition of 'PoPS genes'*: The PoPS method⁴⁸ prioritizes disease-associated genes by
687 integrating gene-level z-scores from MAGMA³², single-cell gene expression data, biological
688 pathways, and predicted protein-protein interaction networks. The original PoPS publication
689 suggests combining PoPS scores with location information would provide a list of high-
690 confidence genes. However, the combined PoPS+location approach leads to a low recall (it
691 detects very few genes for each disease). Therefore, we used the top 1% PoPS, because it
692 results in a list of 184 prioritized genes per disease. This number of genes is similar to the
693 number of fine-mapped and nearest-to-hit genes, reducing differences in power due to the
694 number of genes assessed. Full PoPS results can be accessed at:

695 <https://www.finucanelab.org/data>.

696

697 **Statistical analyses to compare disease-associated genes vs other genes**

698 The *t*-test is an inferential statistic used to evaluate whether the means of two independent
699 samples are significantly different. Here, we run one-side *t*-tests in R, testing the null hypothesis
700 that the expression of disease-associated genes is higher than those of the control group. *t*-
701 tests assume that the sample means are normally distributed. Since gene expression follows a
702 negative binomial distribution, we normalized the gene expression values by taking the Log₂ of
703 the median TPM+1 before applying the *t*-tests.

704

705 The Anderson-Darling⁵⁰ is a non-parametric test to evaluate whether the gene expression of
706 disease-associated genes originates from the same distribution than the control group of genes.
707 It tests the null hypothesis that both groups were drawn from populations with identical
708 distributions. The Anderson-Darling test is similar to other tests assessing differences between
709 empirical distributions (such as the two-sample Kolmogorov-Smirnov test⁵¹), but it is more
710 sensitive to differences in the tails of the distribution.

711
712 For the Tabula Sapiens dataset, *P*-values to compare disease-associated genes vs control
713 genes were calculated using a permutation approach (10,000 permutations). We calculate
714 empirical *P*-values here to account for the small sample size of the dataset (up to 15 individuals,
715 although typically only 2 individuals were used to extract scRNA-seq measurements for each
716 tissue). This method avoids bias in cases where cell types are obtained from cells derived from
717 the same individual, ensuring that results are not affected by violated assumptions of
718 independence –which would invalidate a t-test.

719

720 **Gene Expression to GWAS**

721 **Definition of genes in the top decile of expression**

722 For each tissue, we defined to gene-sets that are in the top decile of gene expression: one
723 gene-set is composed by the 10% of genes with the highest *absolute* gene expression. The
724 second gene-set is composed by the 10% of genes with the highest *relative* gene expression.
725 To obtain these gene-sets, we first classified all protein-coding genes into 11 quantiles. In this
726 classification, the 1st quantile is composed by genes without expression in a specific tissue,
727 whereas the 11th quantile encompasses the genes with the highest expression values. We then
728 grouped genes from the top quantile and tested their GWAS enrichment using MAGMA and the
729 UK Biobank as a reference panel. We expanded the gene coordinates by adding a 35 kb
730 window upstream and a 10kb window downstream of the gene. The Major Histocompatibility
731 Complex (MHC) region was excluded from the analyses due to their long-range LD.

732

733 **Gene set enrichment analyses with MAGMA**

734 MAGMA³² is a software designed for gene-set enrichment analysis using GWAS data. It
735 provides enrichment results at the gene-level and at the gene-set level. In gene-level analysis,
736 MAGMA employs GWAS *P*-values to compute gene test statistics, accounting for LD structure
737 via a reference dataset. For gene-set analysis, gene-level association stats are transformed into
738 Z-scores, reflecting the strength of gene-phenotype associations. MAGMA uses a competitive

739 pathway test formula: $Z = \beta_0 + I\beta_p + C\beta_k + \epsilon$ where I is an indicator (1 if a gene is in pathway p , 0
740 if not), and C is a covariate matrix. The resulting P -value originates from a test on coefficient β_p ,
741 evaluating if the phenotype shows a stronger association with genes included in the gene-set of
742 interest versus other genes.

743

744 **Systematic Literature search**

745 **Search queries utilizing Medical Subject Headings (MeSH) Terminology**

746 The Medical Subject Headings (MeSH) thesaurus is a curated collection of terms established by
747 the National Library of Medicine. MeSH terms are valuable in recognizing content that uses
748 different words but refers to the same concept, enhancing the accuracy and consistency of the
749 literature search process. Leveraging MeSH terminology, we prioritized the list of 45 tissues
750 from the GTEx dataset based on their frequency of occurrence within MeSH terms connected to
751 scientific articles. For each pairing of tissue and disease, a search query in the format '*<tissue
752 name> [Mesh] AND <disease name> [Mesh]*' was used.

753

754 **Search queries utilizing Keyword-based Literature Search**

755 The list of tissues and cell types were ranked based on their citation frequency within the titles
756 or abstracts of relevant scientific articles. The construction of search queries followed the format
757 '*<tissue name> [Title/Abstract] AND <disease name> [Title/Abstract]*' for each unique tissue-
758 disease pair.

759

760 **PubMed literature Search**

761 To determine which tissues are associated with specific diseases based on previous knowledge,
762 we interrogated how often a combination of tissue and disease terms appeared together in
763 published articles found on PubMed. To count the PubMed occurrences of a tissue being
764 mentioned in relation to a disease, we used the Python library *Beautiful Soup*⁸¹, taking the
765 queries defined above as input. The script performs the following tasks: it generates
766 combinations of tissue-disease pairs, constructs search queries, sends requests to the PubMed
767 website based on these queries, and subsequently extracts the number of search results from
768 the webpage. The resulting count shows how frequently the tissue-disease pair appears in the
769 body of literature. Two types of PubMed scrapping analyses we conducted based on the type of
770 query constructed.

771

772 **Utilizing genes associated with other diseases as control group**

773 To generate a list of control genes associated with multiple traits and diseases, we extracted two
774 lists of genes from the Open Targets resource⁵². The first group uses the *Open Targets 'Gold*
775 *Standards'*. The second group uses a list of genes prioritized via *Open Targets Genetics*
776 *evidence*, using a machine learning method⁸² that calculates a disease-specific score to
777 prioritize genes.

778

779 - *Open Targets Gold standards*: This list of genes represents a repository of >400 published
780 GWAS loci for which there is high confidence in the gene functionally implicated. The list of gold
781 standard genes was downloaded from [https://github.com/opentargets/genetics-gold-](https://github.com/opentargets/genetics-gold-standards/blob/master/gold_standards/processed/gwas_gold_standards.191108.tsv)
782 [standards/blob/master/gold_standards/processed/gwas_gold_standards.191108.tsv](https://github.com/opentargets/genetics-gold-standards/blob/master/gold_standards/processed/gwas_gold_standards.191108.tsv). The final
783 set of genes was composed by 519 protein-coding genes from 284 traits were used as gold
784 standard control gene list. The traits with the largest number of genes were 2 diabetes (44
785 genes), breast carcinoma (43 genes) and prostate carcinoma (23 genes).

786

787 - *Open Targets Genetics evidence*: This list of genes was extracted from the results of a
788 machine-learning method used to identify the most likely causal genes⁸². This method integrates
789 the results of 1) fine-mapping credible set analysis, 2) functional genomics data such as
790 pathogenicity prediction, colocalization with molecular QTLs, genomic distance and chromatin
791 interaction data to generate predictive features. The machine-learning model is supervised
792 using the gold-standard positive GWAS loci, and a score is computed for each gene (named
793 Locus to gene (L2G) score). The L2G score is calibrated so that a gene's score indicates the
794 fraction of genes at or above the score that would be expected to be true positives. Thus, we
795 selected genes with a score ≥ 0.8 , which assumes that 80% of the genes associated with a
796 trait or disease in our list are causal.

797

798 Data was downloaded from the publicly available website
799 <https://platform.opentargets.org/downloads/data>, section "Target - Disease evidence / Integrated
800 list of target - disease evidence from all data sources" (version 23/09), which provides several
801 directories with different evidence sources for the target-disease associations. However, only
802 the ones indicating 'genetics evidence' were used in our analyses. From those, 3,862 protein-
803 coding genes from 1,582 traits had L2G scores ≥ 0.8 . The traits with the largest number of
804 genes were height (580 genes), blood protein measurement (359 genes), and heel bone
805 mineral density (286 genes).

806

807 **Profiling the gene expression landscape of cancer-associated genes**

808 We applied the '*nearest-to-hit*' prioritization method to cancer traits by extracting genetic
809 variants associated with eight common cancers through GWAS. These datasets are publicly
810 available and included lists of independent, GWAS significant SNPs used to construct polygenic
811 risk scores⁵⁴. The SNP lists range from 22 SNPs for pancreatic cancer to 288 SNPs for breast
812 cancer. Unlike the other diseases analysed in this study, for which we had the full summary
813 statistics instead of only the top SNPs, we didn't perform clumping on these SNP lists. The
814 procedure for assigning the closest gene to each GWAS hit was the same as for '*nearest-to-hit*'
815 genes.

816

817 **Definition of disease-relevant tissues for removing genes in the top decile of expression**

818 To define the tissues that showed significant higher expression across all the tests, we defined a
819 *P*-value threshold for association (threshold = 0.05/45x3x2, corresponding to 45 tissues, 3 lists
820 of gene prioritization approaches, and 2 test statistics (*t*-test and Anderson-Darling test). Then,
821 we identified the tissues for which the Anderson-Darling and the *t*-test showed *P*-values <
822 threshold.

823

824 To test whether our association results were driven by only a few genes, we removed the genes
825 that are in the top decile of absolute or relative expression in the relevant tissues, and repeated
826 the '*GWAS to gene expression*' analyses.

827

828 **Calculating predictors of disease-associated gene expression**

829 To uncover the key contributors to the variability in gene expression among disease-associated
830 genes, we performed variance partition analyses using the R package '*variancePartition*'⁶⁴. This
831 package assesses drivers of variation for each gene by fitting a linear fixed model to quantify
832 the contribution of tissues, individuals, technical variables etc. in gene expression.

833

834 We calculated the variance partition for each disease-associated and control gene. We used as
835 predictors uncorrelated variables ($r^2 < 0.75$) that explained the largest proportion of variance in
836 gene expression, as calculated by the Canonical Correlation Analysis in the *variancePartition*
837 package and reported in the original *variancePartition* publication⁶⁴ (which also used the GTEx
838 dataset). The variance partition analysis results in a data table where each row is a gene, and

839 each column is the predictor variable included in the model. The results show, for each gene,
840 the percentage of variance explained for each predictor.

841

842

References

1. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
2. Qi, T., Song, L., Guo, Y., Chen, C. & Yang, J. From genetic associations to genes: methods, applications, and challenges. *Trends Genet.* S0168952524000957 (2024)
doi:10.1016/j.tig.2024.04.008.
3. Aragam, K. G. *et al.* Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. *Nat. Genet.* **54**, 1803–1815 (2022).
4. Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022).
5. Huang, H. *et al.* Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547**, 173–178 (2017).
6. Revez, J. A. *et al.* Genome-wide association study identifies 143 loci associated with 25 hydroxyvitamin D concentration. *Nat. Commun.* **11**, 1647 (2020).
7. Bipolar Disorder Working Group of the Psychiatric Genomics Consortium, 23andMe Research Team *et al.* Genetic diversity enhances gene discovery for bipolar disorder. 2023.10.07.23296687 Preprint at <https://doi.org/10.1101/2023.10.07.23296687> (2023).
8. Demontis, D. *et al.* Genome-wide analyses of ADHD identify 27 risk loci, refine the genetic architecture and implicate several cognitive domains. *Nat. Genet.* **55**, 198–208 (2023).
9. Mahajan, A. *et al.* Fine-mapping of an expanded set of type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).
10. Kunkle, B. W. *et al.* Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates A β , tau, immunity and lipid processing. *Nat. Genet.* **51**, 414–430 (2019).

11. Bellenguez, C. *et al.* New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat. Genet.* **54**, 412–436 (2022).
12. Minikel, E. V., Painter, J. L., Dong, C. C. & Nelson, M. R. Refining the impact of genetic evidence on clinical success. *Nature* **629**, 624–629 (2024).
13. Gashaw, I., Ellinghaus, P., Sommer, A. & Asadullah, K. What makes a good drug target? *Drug Discov. Today* **16**, 1037–1043 (2011).
14. Dezso, Z. *et al.* A comprehensive functional analysis of tissue specificity of human gene expression. *BMC Biol.* **6**, 49 (2008).
15. Yang, L. *et al.* Comparative analysis of housekeeping and tissue-selective genes in human based on network topologies and biological properties. *Mol. Genet. Genomics MGG* **291**, 1227–1241 (2016).
16. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860 (2015).
17. King, E. A., Davis, J. W. & Degner, J. F. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLOS Genet.* **15**, e1008489 (2019).
18. Ryaboshapkina, M. & Hammar, M. Tissue-specific genes as an underutilized resource in drug discovery. *Sci. Rep.* **9**, 7233 (2019).
19. Duffy, Á. *et al.* Tissue-specific genetic features inform prediction of drug side effects in clinical trials. *Sci. Adv.* **6**, eabb6242 (2020).
20. Eraslan, G. *et al.* Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. *Science* **376**, eabl4290 (2022).
21. Travaglini, K. J. *et al.* A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587**, 619–625 (2020).
22. Kadur Lakshminarasimha Murthy, P. *et al.* Human distal lung maps and lineage hierarchies reveal a bipotent progenitor. *Nature* **604**, 111–119 (2022).

23. Winkler, E. A. *et al.* A single-cell atlas of the normal and malformed human brain vasculature. *Science* **375**, eabi7377 (2022).
24. Perez, R. K. *et al.* Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. *Science* **376**, eabf1970 (2022).
25. de Paiva Lopes, K. *et al.* Genetic analysis of the human microglia transcriptome across brain regions, aging and disease pathologies. *Nat. Genet.* **54**, 4–17 (2022).
26. Cuomo, A. S. E., Nathan, A., Raychaudhuri, S., MacArthur, D. G. & Powell, J. E. Single-cell genomics meets human genetics. *Nat. Rev. Genet.* **24**, 535–549 (2023).
27. Cano-Gamez, E. & Trynka, G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front. Genet.* **11**, 424 (2020).
28. Zeng, H. What is a cell type and how to define it? *Cell* **185**, 2739 (2022).
29. Zhang, T., Perkins, M. H., Chang, H., Han, W. & de Araujo, I. E. An inter-organ neural circuit for appetite suppression. *Cell* **185**, 2478-2494.e28 (2022).
30. Stanley, S. A. *et al.* Bidirectional electromagnetic control of the hypothalamus regulates feeding and metabolism. *Nature* **531**, 647–650 (2016).
31. Jimenez-Gonzalez, M. *et al.* Mapping and targeted viral activation of pancreatic nerves in mice reveal their roles in the regulation of glucose metabolism. *Nat. Biomed. Eng.* **6**, 1298–1316 (2022).
32. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
33. Lachmann, A. *et al.* Massive mining of publicly available RNA-seq data from human and mouse. *Nat. Commun.* **9**, 1366 (2018).
34. THE TABULA SAPIENS CONSORTIUM. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).
35. THE GTEx CONSORTIUM. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

36. Watanabe, K., Umićević Mirkov, M., de Leeuw, C. A., van den Heuvel, M. P. & Posthuma, D. Genetic mapping of cell type specificity for complex traits. *Nat. Commun.* **10**, 3222 (2019).
37. Kryuchkova-Mostacci, N. & Robinson-Rechavi, M. A benchmark of gene expression tissue-specificity metrics. *Brief. Bioinform.* **18**, 205–214 (2017).
38. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
39. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
40. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).
41. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
42. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**, 1369-1384.e19 (2016).
43. Boix, C. A., James, B. T., Park, Y. P., Meuleman, W. & Kellis, M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* **590**, 300–307 (2021).
44. Gazal, S. *et al.* Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. *Nat. Genet.* **54**, 827–836 (2022).
45. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6**, 5890 (2015).
46. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).
47. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
48. Weeks, E. M. *et al.* Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. *Nat. Genet.* **55**, 1267–1276 (2023).

49. Norman, A. W. From vitamin D to hormone D: fundamentals of the vitamin D endocrine system essential for good health¹. *Am. J. Clin. Nutr.* **88**, 491S-499S (2008).
50. Stephens, M. A. EDF Statistics for Goodness of Fit and Some Comparisons. *J. Am. Stat. Assoc.* **69**, 730–737 (1974).
51. Chakravarti, I. M., Laha, R. G. & Roy, J. *Handbook of Methods of Applied Statistics. 1, Techniques of Computation, Descriptive Methods, and Statistical Inference.* (John Wiley & Sons, New York [etc.], 1967).
52. Koscielny, G. *et al.* Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res.* **45**, D985–D994 (2017).
53. Skene, N. G. *et al.* Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet.* **50**, 825–833 (2018).
54. Jia, G. *et al.* Evaluating the Utility of Polygenic Risk Scores in Identifying High-Risk Individuals for Eight Common Cancers. *JNCI Cancer Spectr.* **4**, pkaa021 (2020).
55. Sawanyawisuth, K. *et al.* Serial analysis of gene expression reveals promising therapeutic targets for liver fluke-associated cholangiocarcinoma. *Asian Pac. J. Cancer Prev. APJCP* **13 Suppl**, 89–93 (2012).
56. Lee, S. M. *et al.* Hypomethylation of the thymosin $\beta(10)$ gene is not associated with its overexpression in non-small cell lung cancer. *Mol. Cells* **32**, 343–348 (2011).
57. Huang, L. *et al.* Identification of a gene-expression signature for predicting lymph node metastasis in patients with early stage cervical carcinoma. *Cancer* **117**, 3363–3373 (2011).
58. Zhang, X.-J. *et al.* Thymosin beta 10 correlates with lymph node metastases of papillary thyroid carcinoma. *J. Surg. Res.* **192**, 487–493 (2014).
59. Li, M. *et al.* Thymosin beta 10 is Aberrantly Expressed in Pancreatic Cancer and Induces JNK Activation. *Cancer Invest.* **27**, 251 (2009).

60. Alldinger, I. *et al.* Gene expression analysis of pancreatic cell lines reveals genes overexpressed in pancreatic cancer. *Pancreatol. Off. J. Int. Assoc. Pancreatol. IAPAI* **5**, 370–379 (2005).
61. Ben, Q. *et al.* Diabetes mellitus and risk of pancreatic cancer: A meta-analysis of cohort studies. *Eur. J. Cancer Oxf. Engl. 1990* **47**, 1928–1937 (2011).
62. Tan, J. *et al.* Association of elevated risk of pancreatic cancer in diabetic patients: A systematic review and meta-analysis. *Oncol. Lett.* **13**, 1247–1255 (2017).
63. Hansen, D. V., Hanson, J. E. & Sheng, M. Microglia in Alzheimer's disease. *J. Cell Biol.* **217**, 459–472 (2018).
64. Hoffman, G. E. & Schadt, E. E. variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics* **17**, 483 (2016).
65. Creese, I., Burt, D. R. & Snyder, S. H. Dopamine receptor binding predicts clinical and pharmacological potencies of antischizophrenic drugs. *Science* **192**, 481–483 (1976).
66. Meltzer, H. Y., Matsubara, S. & Lee, J. C. Classification of typical and atypical antipsychotic drugs on the basis of dopamine D-1, D-2 and serotonin₂ pKi values. *J. Pharmacol. Exp. Ther.* **251**, 238–246 (1989).
67. Dezső, Z. *et al.* A comprehensive functional analysis of tissue specificity of human gene expression. *BMC Biol.* **6**, 49 (2008).
68. Fagerberg, L. *et al.* Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics *. *Mol. Cell. Proteomics* **13**, 397–406 (2014).
69. Santos, S. E. D. *et al.* Similar Microglial Cell Densities across Brain Structures and Mammalian Species: Implications for Brain Tissue Function. *J. Neurosci.* **40**, 4622–4643 (2020).
70. Wainberg, M. *et al.* Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* **51**, 592–599 (2019).

71. Nguyen, P. A., Born, D. A., Deaton, A. M., Nioi, P. & Ward, L. D. Phenotypes associated with genes encoding drug targets are predictive of clinical trial side effects. *Nat. Commun.* **10**, 1579 (2019).
72. Hemerich, D. *et al.* An integrative framework to prioritize genes in more than 500 loci associated with body mass index. *Am. J. Hum. Genet.* **0**, (2024).
73. Zhang, Y., Parmigiani, G. & Johnson, W. E. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genomics Bioinforma.* **2**, lqaa078 (2020).
74. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
75. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, s13742-015-0047–8 (2015).
76. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).
77. Mullins, N. *et al.* Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. *Nat. Genet.* **53**, 817–829 (2021).
78. Suzuki, K. *et al.* Multi-ancestry genome-wide study in >2.5 million individuals reveals heterogeneity in mechanistic pathways of type 2 diabetes and complications. *medRxiv* 2023.03.31.23287839 (2023) doi:10.1101/2023.03.31.23287839.
79. Andrews, S. J. *et al.* The complex genetic architecture of Alzheimer’s disease: novel insights and future directions. *eBioMedicine* **90**, (2023).
80. Manousaki, D. *et al.* Genome-wide Association Study for Vitamin D Levels Reveals 69 Independent Loci. *Am. J. Hum. Genet.* **106**, 327–337 (2020).
81. Richardson, L. beautifulsoup4: Screen-scraping library.
82. Mountjoy, E. *et al.* An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat. Genet.* **53**, 1527–1533 (2021).

Code availability

The scripts used in the current study are available at https://gitlab.com/JuditGG/gene_expr_landscape

Acknowledgements

This work was supported by a grant from the National Institutes of Health (R01MH122866) to PFO, and by a 2022 NARSAD Young Investigator Grant (Number 30749) by the Brain & Behavior Research Foundation to JGG.

Additionally, this work was supported in part through the computational and data resources and staff expertise provided by Scientific Computing and Data at the Icahn School of Medicine at Mount Sinai and supported by the Clinical and Translational Science Awards (CTSA) grant UL1TR004419 from the National Center for Advancing Translational Sciences. Research reported in this publication was also supported by the Office of Research Infrastructure of the National Institutes of Health under award number S10OD026880 and S10OD030463.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

We thank the participants of the UK Biobank, ARCHS4, Tabula Sapiens and the GTEx projects and the scientists involved in the construction of these resources.

Author contributions

JGG: Conceptualization, Funding Acquisition, Data Curation, Formal Analysis, Investigation, Software, Validation, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing. SGG: Investigation. LL: Writing – Review & Editing, Data Curation. PFO: Conceptualization, Funding Acquisition, Formal Analysis, Supervision, Writing – Review & Editing.

Competing interests

The authors declare that they have no competing interests.