The Gene Expression Landscape of Disease Genes

Judit García-González[1], Alanna C. Cote[1], Saul Garcia-Gonzalez[1,2], Lathan Liou[1], Paul F. O'Reilly[1]

[1]Department of Genetics and Genomic Sciences, Icahn School of Medicine, Mount Sinai, New York City, NY 10029, USA
[2]Center for Excellence in Youth Education, Icahn School of Medicine, Mount Sinai, New York City, NY 10029, USA

Abstract

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, which, in previous studies, has mostly been investigated by testing for enrichment of GWAS signal among genes with cell/tissue specific expression. By integrating GWAS summary statistics, gene prioritisation results, and RNA-seq data from 46 tissues and 204 cell types, we directly analyse the gene expression of putative disease genes across the body in relation to 11 major diseases and cancers. In tissues and cell types with established disease relevance, disease genes show higher and more specific gene expression compared to control genes. However, we also detect elevated expression in tissues and cell types without previous 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 < 10^{-3}), the cause for others is unclear and warrants further investigation. To support functional follow-up studies of disease genes, we identify technical and biological factors influencing their expression, and highlight tissues in which higher expression is associated with increased odds of inclusion in drug development programs. We provide our systematic testing framework as an open-source, publicly available tool that can be utilised to offer novel insights into the genes, tissues and cell types involved in any disease, with the potential for informing drug development and delivery strategies.

Introduction

Genome-wide association studies (GWAS) have identified thousands of risk loci for complex diseases over the past two decades[1]. An important aspect of translating these findings into biological and clinical insights is pin-pointing the causal variants and genes. To this end, fine-mapping and gene prioritisation strategies have been developed [2] and – only in recent years – lists of high-confidence causal genes have been compiled across multiple diseases [3–11]. These curated lists of putative disease genes provide the opportunity to systematically and directly examine where in the body disease genes are active, offering insights into disease initiation, aetiological mechanisms and therapeutic opportunities. For example, successful drug targets are enriched for GWAS prioritised genes[12], as well as tissue specific genes[13–15]. This suggests that mapping the tissues and cell types in which disease genes are active could further enhance the identification of new drug targets [16,17] with reduced off-target effects[18,19].

Previous studies integrating GWAS and gene expression data to infer the tissues/cells involved in disease have done this more indirectly. Typically, such studies identify genes with the greatest relative (i.e. specific) expression in each cell type or tissue, irrespective of disease, and then assess their enrichment in GWAS signals for the disease of interest [20–23]. While this approach provides insight into the tissues and cell types where disease may initiate or progress, it does not directly capture the question of most interest: *how and where disease genes are expressed*. Existing methods do not leverage lists of high-confidence causal disease genes, and, thus, they include many genes with no involvement in the disease under study in their testing and cannot explore how different definitions of disease genes influence the identification of disease-relevant tissues and cell types. Moreover, most current approaches focus on relative gene expression (i.e. specificity) only. Although investigating relative expression may highlight disease genes less likely to have pleiotropic effects, absolute gene expression, capturing overall abundance, may

be more important for diseases that involve multiple cell/tissue types. Thus, while a gene expressed at high levels across multiple tissues could clearly play a pivotal role in a disease, it would be missed by methods restricted to tissue-specific signal.

In this study, we use three alternative approaches that leverage GWAS and RNA-seq data to infer where disease genes are expressed: the first – that we call "*GWAS to Gene Expression"* – is a novel approach that tests whether putative disease genes, according to recent large-scale GWAS, exhibit distinct expression levels across the body compared to control genes (**Fig 1a**). We ensure the robustness of this approach by employing three distinct definitions for disease genes and utilising three different sets of control genes. The second approach – that we call "*Gene Expression to GWAS"* – examines whether high-expression genes are enriched for GWAS signal, as calculated by MAGMA[22] (**Fig 1b**). In the third approach, we perform a systematic PubMed search to assess the evidence for tissue-disease associations reported in the literature (**Fig 1c**). Contrasting results systematically across the three approaches allows for a degree of 'triangulation' of results. We apply our testing framework in 46 tissues and 204 cell types obtained from the GTEx[24], ARCHS4[25] and Tabula Sapiens[26] resources (**Fig 1d**). We also characterise, for each individual gene, to what extent different predictors (batch ID, subject ID, age, sex etc) contribute to gene expression (**Fig 1e**) and examine how both absolute and tissue-specific (relative) expression of disease genes influence their odds for being included in drug development programs (**Fig 1f**).

Analyses are performed across eight major diseases and three cancers, for which powerful GWAS and curated lists of high-confidence disease genes are available: Schizophrenia (SCZ), Inflammatory Bowel Disease (IBD), Alzheimer's Disease (AD), Coronary Artery Disease (CAD),

Bipolar Disorder (BD), Type 2 Diabetes (T2D), Attention-Deficit/Hyperactivity Disorder (ADHD), Serum 25 Hydroxyvitamin D (Vitamin D), and Breast, Prostate and Colorectal cancer.

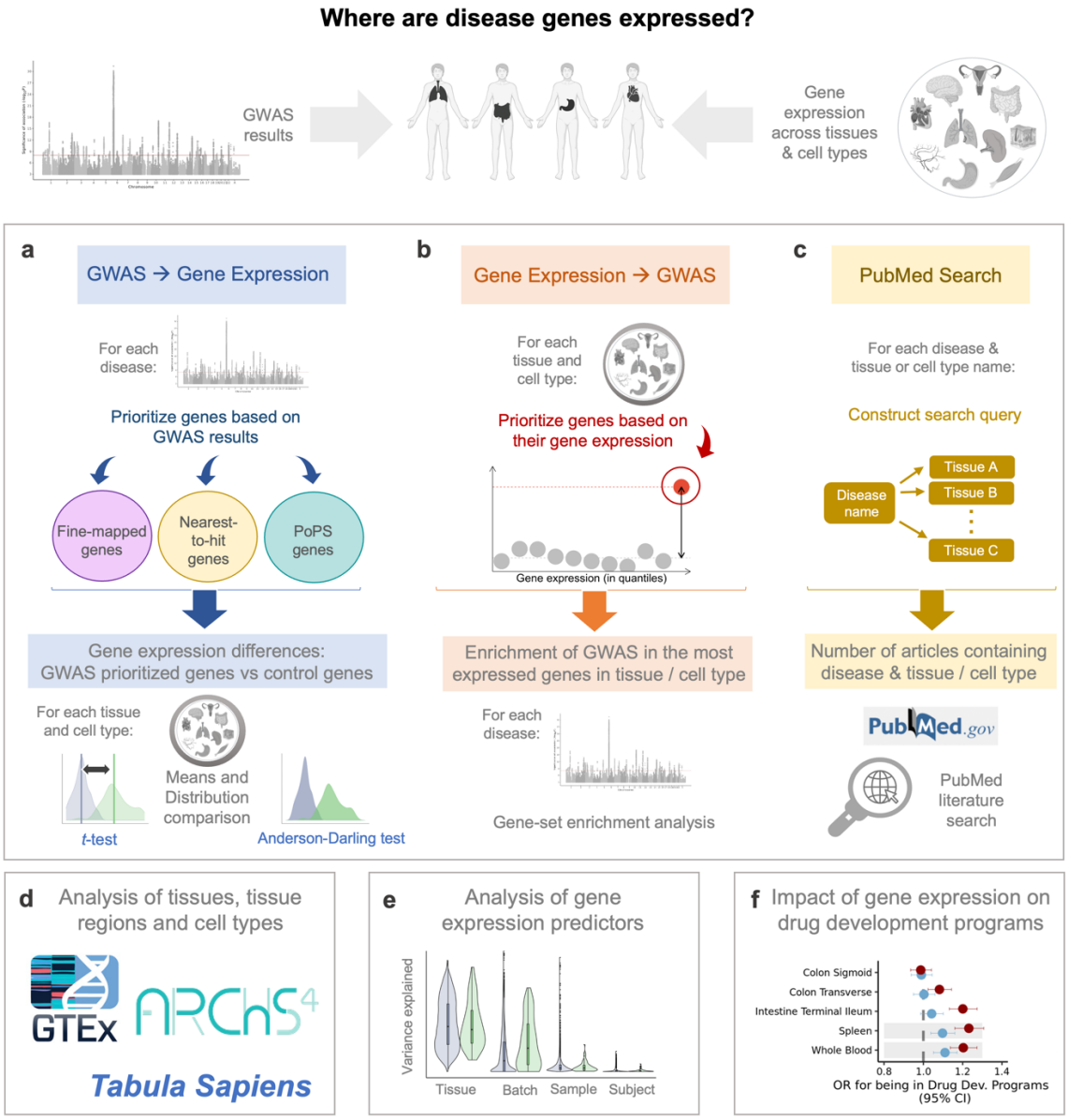


Figure 1. Study overview. **a**, Approach using lists of prioritised, high-confidence disease genes, derived from GWAS results. **b**, Approach using genes with the highest absolute or relative expression in each cell and tissue. **c**, PubMed-based literature search to assess the tissues most often cited in relation to each disease. **d,** Primary analyses apply approaches a-c to the tissue gene expression data of GTEx, while follow-up analyses are applied to the cell types of the ARCHS4 and Tabula Sapiens datasets. **e,** Factors affecting gene expression variance are analysed for each disease gene. **f**, For tissues where disease genes are significantly more expressed than control genes (*t*-test, Bonferroni-corrected), we assess how absolute and relative expression of disease genes influence their likelihood of inclusion in drug development programs. Figure partially created using BioRender.com.

Results

Exploring the expression profiles of disease genes

To infer where in the body disease genes operate, here we leverage GWAS and gene expression data utilising three alternative strategies of inference: (i) *'GWAS to Gene Expression'*; (ii) *'Gene Expression to GWAS'*; and (iii) *'PubMed Search'*. The first part of the results presents our primary analyses, which examine gene expression in tissues across the three strategies. We then extend these analyses to the cell-type level to refine our findings within relevant tissues.

GWAS to Gene Expression

A range of statistical and computational approaches have been developed[27–31] to link GWAS-associated variants to their target genes and to prioritise genes based on GWAS results. Given the absence of a single gold standard approach for gene prioritisation, we employ and describe below three alternative approaches to define 'disease genes' inferred from GWAS results: (i) *nearest-to-hit genes*, (ii) *fine-mapped genes*, and (iii) *PoPS genes*. All remaining protein-coding genes are used as control genes (**Methods**).

(i) Nearest-to-hit genes: While sophisticated methods to prioritise causal genes exist (see below), previous research suggests that the nearest gene to the lead variant is far more likely to be the causal gene than others at the locus [27,31,32]. We performed clumping on the GWAS hits and identified an average of 162 'nearest-to-hit' genes for each disease/trait, with an average distance of 28 kb between the SNP and the closest protein coding gene (See **Methods, Table 1** and **Supplemental Table 1**).

85 *(ii) Fine-mapped genes*: The latest landmark GWAS of the major diseases[3–11] have incorporated
86 sub-studies performing a range of gene prioritisation analyses to produce lists of around 50-300
87 high-confidence causal genes for each disease. While different studies use a different selection
88 of methods to prioritise putatively causal genes, most of them integrate GWAS results with the
89 latest functional genomics data (**Methods**). Our literature search on fine-mapped genes resulted
90 in gene lists ranging from 49 genes prioritised for BD to 281 genes prioritised for IBD, with an
91 average of 126 genes across the 8 diseases investigated (**Supplemental Table 2**).
92
93 *(iii) PoPS genes:* The method 'Polygenic Priority Scores' (PoPS) method[31] leverages gene-level
94 *Z*-scores from GWAS (calculated using the software MAGMA[22]), as well as gene features from
95 single-cell gene expression data, biological pathways and predicted protein-protein interaction
96 networks to prioritise putatively causal disease genes. For each trait, we extracted the genes
97 with the top 1% PoPS scores, corresponding to 184 protein-coding genes with the highest PoPS
98 scores for each trait. The top 1% threshold was chosen because it provides a similar number of
99 genes as the other two prioritisation approaches, limiting biases due to differences in the
100 number of genes included for each group of putative disease genes (**Supplemental Table 3**).
101
102 For each disease, we performed *t*-tests comparing the absolute and relative expression of high-
103 confidence disease genes (fine-mapped, nearest-to-hit and PoPS genes) with those of the
104 control genes. Disease genes show higher expression in tissues with known relevance to the
105 corresponding disease (**Fig 2,** *t*-test *P*-values shown in blue columns). Overall, the disease
106 genes' relative gene expression *t*-test results are similar to those for absolute gene expression
107 but generally have smaller *P*-values. Unless otherwise noted, the results and *P*-values
108 discussed below refer to absolute gene expression.
109

110 SCZ genes are more expressed in the brain, although *P*-values vary substantially across brain
111 tissues (from *P*-value = 3.1×10^{-2} for the nearest gene list in substantia nigra to *P*-value = $2.0 \times 10^{-}$
112 ^{29} for PoPS genes in the brain cortex). CAD genes are more expressed in the aorta, coronary
113 and tibial arteries (*P*-values < 10^{-7}). IBD genes are most expressed in the small intestine and
114 colon transverse (*P*-values < 10^{-4} except for nearest-to-hit genes). Vitamin D genes are most
115 expressed in the skin and in the liver, with the latter tissue being where the biologically inactive
116 vitamin D_3 is activated to produce 25-hydroxyvitamin D_3 [33].
117
118 Other significant results point to tissues not typically linked to the disease: AD genes show
119 highest expression in the blood, spleen (PoPS genes *P*-value < 10^{-13}, nearest-to-hit genes *P*-
120 value < 10^{-5}) and adipose tissues (PoPS genes *P*-values < 10^{-13}) but not in the brain. SCZ
121 genes are highly expressed in the pituitary (PoPS genes *P*-values = 3.7×10^{-5}). CAD genes
122 exhibit higher expression across multiple tissues including reproductive, adipose, and digestive
123 systems, as well as in the lung (*P*-values < 10^{-6}). IBD genes have highest expression in lung (*P*-
124 values < 10^{-4}), blood and spleen (*P*-values < 10^{-6}). Differences between T2D genes vs control
125 genes appear most significant in breast mammary tissues (*P*-values < 10^{-7}), but these results
126 are not driven by differences between males and females, or the presence of some genes with
127 high expression in only one sex (**Additional File 1, Supplemental Material** and **Supplemental**
128 **Figure 1**).
129
130 We also applied the Anderson-Darling test[34] to evaluate whether the expression *distribution* of
131 high-confidence disease genes differs from that of control genes. In most cases, results were
132 consistent with the *t*-tests but yielded smaller *P*-values (**Supplemental Figure 2**). Full results for
133 both tests, across all diseases, tissues and gene lists are provided in **Supplemental Table 4**.
134

135 We performed several sensitivity analyses to ensure the robustness of our findings. First, we
136 tested the impact of varying the PoPS top percentage threshold (**Methods** and **Supplementary**
137 **Figure 3**). Second, we examined alternative definitions of control groups, repeating our
138 analyses using genes prioritised by the Open Targets initiative[41] (**Methods**, **Supplemental**
139 **Figures 4-11** and **Supplemental Table 5**). Third, we investigated whether the observed results
140 could be explained by a small number of genes with extremely high expression. We repeated
141 the analyses of Fig.2 excluding genes in the top 10% of absolute and relative expression for
142 tissues with significant results. (**Additional File 1, Supplemental Material, Supplemental**
143 **Table 6**). Results show that, while the *P*-values increase for all the tests, results are qualitatively
144 consistent (**Supplemental Figures 12-13** and **Supplemental Table 7**). For example, the cortex
145 and cingulate cortex are significantly associated with SCZ (Absolute expression *P*-values < 10^{-}
146 ^{20}), and the tissues most associated with AD remain the small intestine (*P*-value = 3.8×10^{-9}),
147 spleen (*P*-value = 4.9×10^{-8}) and lung (*P*-value = 5.0×10^{-8}).
148

149 **Gene Expression to GWAS**

150 In this section, we test whether genes with high absolute and relative expression in a tissue are
151 enriched for GWAS signal. In an early example of this approach, Skene et al.[35] assessed the
152 enrichment of GWAS signal using MAGMA[23] and LDSC [21], among genes in the top decile of
153 absolute and relative (also called specific) gene expression (see **Methods**). We expand the
154 approach here to investigate *absolute gene expression*, in addition to *relative gene expression*.
155 Our findings are overall consistent with the results of the GWAS to Gene Expression approach.
156 (**Figure 2,** red columns). For SCZ, BP and ADHD, MAGMA results were nominally significant in
157 numerous brain tissues (*P*-value < 0.05), such as cortex, anterior cingula, hippocampus,
158 amygdala, or cerebellum or nucleus accumbens. The strongest results were observed for the
159 cortex and anterior cingula tissues and SCZ (*P*-value < 10^{-12}). For CAD, numerous tissues show
160 significant *P*-values for both absolute and relative expression, with arteries (*P*-value < 10^{-6}),

161 colon sigmoid (*P*-value = 4.7×10^{-6}) and esophagus (*P*-value = 8.4×10^{-5}) having the strongest
162 enrichment of GWAS signal among their most specific genes. In IBD, the intestine, blood, testis,
163 liver, lung, and spleen are the tissues with strongest enrichment (*P*-values < 10^{-3}), while AD
164 shows the strongest result in spleen and blood (*P*-values < 10^{-6}). For Vitamin D, liver is the most
165 relevant tissue (*P*-value = 2.5×10^{-3}). For T2D, none of the tissues showed significant results.
166 Overall, MAGMA enrichments for relative gene expression are more pronounced than for
167 absolute gene expression. MAGMA results for each disease, each tissue, and each gene list are
168 included in **Supplemental Table 8**.
169

170 **PubMed Search**

171 Next, we investigated disease-tissue associations by performing a systematic literature search
172 via PubMed, using two approaches to construct PubMed search queries. First, we use Medical
173 Subject Headings (MeSH) terms, a standardized vocabulary from the National Library of
174 Medicine. Second, we identify tissue/disease pair names in the title and abstract of the PubMed
175 articles. These two approaches provide consistent results (**Figure 2,** yellow columns). While the
176 PubMed search results largely support our findings, there are tissues identified in our main
177 analyses that may be understudied, as the support in the literature is low (e.g. spleen in AD and
178 IBD, digestive tissues in CAD). There are also disease-tissue pairs that have been well-studied
179 (e.g. Skeletal muscle and Vitamin D, or Liver and Inflammatory Bowel Disease), but that have
180 limited support in our analyses. Results with the number of papers found for each query are
181 included in **Supplemental Table 9 & 10**.
182

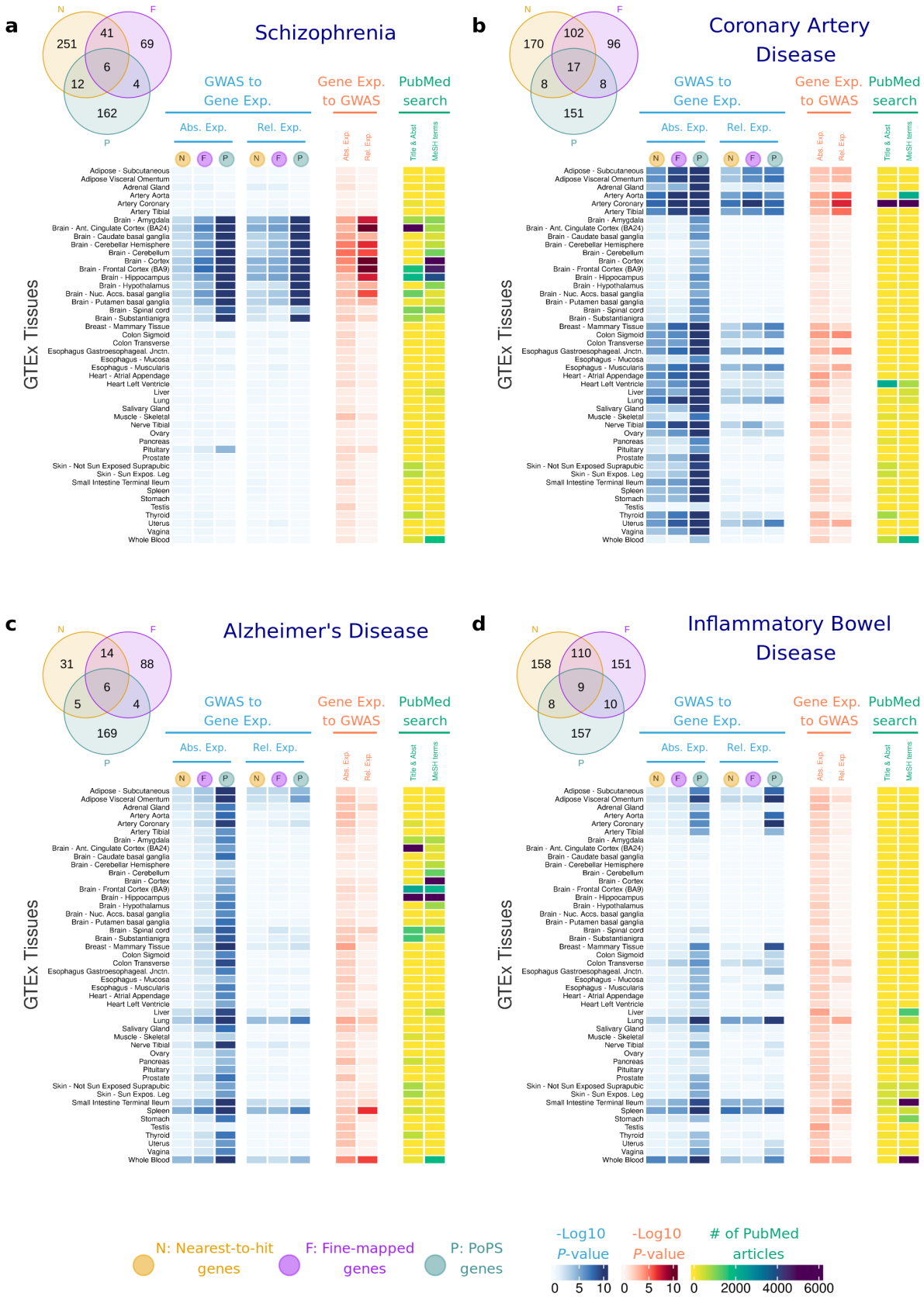
183 **Correlation in results across the three alternative strategies**

184 Across the different sets of tests, we observe strong correlations between the *t*-tests and
185 Anderson-Darling tests (mean r = 0.73) and between *t*-tests and MAGMA (mean r = 0.55), with

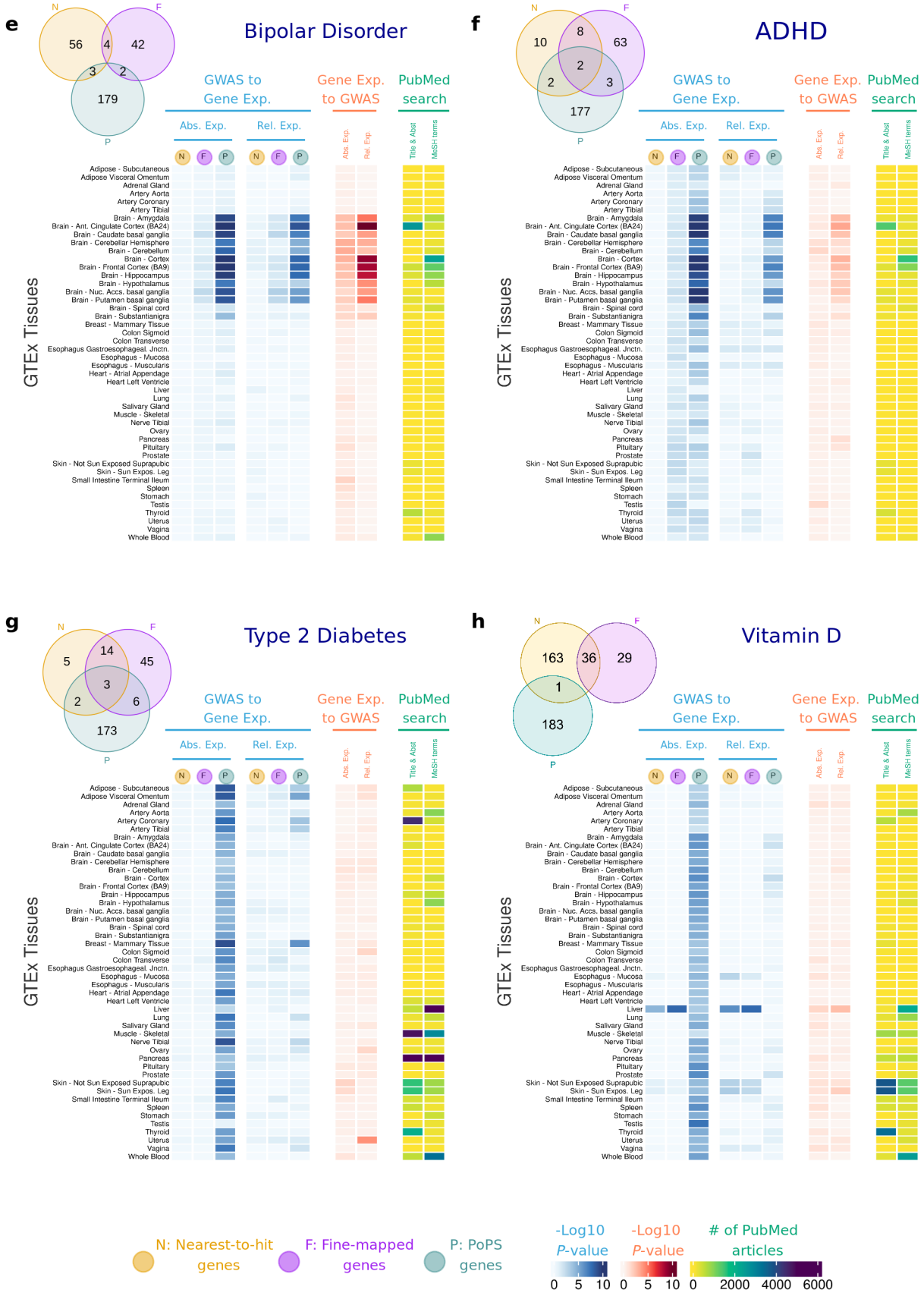
186 most P-values < $1x10^{-3}$. Correlations across the three approaches varied by disease, with the

187 highest correlations in SCZ (r = 0.83), CAD (r = 0.71), and IBD (r = 0.69), while T2D exhibited

188 the lowest correlation (r = 0.27). Detailed results are provided in **Supplemental Table 11**.


a
Schizophrenia
N
F
P
251
41
69
6
12
4
162
GWAS to Gene Exp.
Gene Exp. to GWAS
PubMed search
Abs. Exp.
Rel. Exp.
Abs. Exp.
Rel. Exp.
Title & Abst
MeSH terms
GTEx Tissues
b
Coronary Artery Disease
170
102
96
17
8
8
151
c
Alzheimer's Disease
31
14
88
6
5
4
169
d
Inflammatory Bowel Disease
158
110
151
9
8
10
157
N: Nearest-to-hit genes
F: Fine-mapped genes
P: PoPS genes
-Log10 P-value
0 5 10
-Log10 P-value
0 5 10
# of PubMed articles
0 2000 4000 6000

189


e
Bipolar Disorder
N
F
P
56
4
42
3
2
179
f
ADHD
10
8
63
2
2
3
177
g
Type 2 Diabetes
5
14
45
3
2
6
173
h
Vitamin D
163
36
29
1
183
GWAS to Gene Exp.
Gene Exp. to GWAS
PubMed search
Abs. Exp.
Rel. Exp.
Abs. Exp.
Rel. Exp.
Title & Abst
MeSH terms
GTEx Tissues
Adipose - Subcutaneous
Adipose Visceral Omentum
Adrenal Gland
Artery Aorta
Artery Coronary
Artery Tibial
Brain - Amygdala
Brain - Ant. Cingulate Cortex (BA24)
Brain - Caudate basal ganglia
Brain - Cerebellar Hemisphere
Brain - Cerebellum
Brain - Cortex
Brain - Frontal Cortex (BA9)
Brain - Hippocampus
Brain - Hypothalamus
Brain - Nuc. Accs. basal ganglia
Brain - Putamen basal ganglia
Brain - Spinal cord
Brain - Substantianigra
Breast - Mammary Tissue
Colon Sigmoid
Colon Transverse
Esophagus Gastroesophageal. Jnctn.
Esophagus - Mucosa
Esophagus - Muscularis
Heart - Atrial Appendage
Heart Left Ventricle
Liver
Lung
Salivary Gland
Muscle - Skeletal
Nerve Tibial
Ovary
Pancreas
Pituitary
Prostate
Skin - Not Sun Exposed Suprapubic
Skin - Sun Expos. Leg
Small Intestine Terminal Ileum
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
Whole Blood
N: Nearest-to-hit genes
F: Fine-mapped genes
P: PoPS genes
-Log10 P-value
0 5 10
-Log10 P-value
0 5 10
# of PubMed articles
0 2000 4000 6000

190

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

The gene expression landscape at the cell-type level

To further investigate where in the body disease genes operate, we repeated our testing framework in a set of cell types and tissue regions extracted from the (i) Tabula Sapiens[26], a resource combining data on nearly 500,000 cells from 24 different tissues and organs, many from the same donor, and (ii) ARCHS4[25], a resource that aggregates RNA-seq data from the Gene Expression Omnibus and the Sequence Read Archive. Gene expression in cell types may have affected the results of our tissue-level analyses because the same cell types are found across different tissues and tissue-level expression is more likely to reflect abundant, rather than rare, cell types.

Figure 3 presents results of the analyses conducted across tissues from Tabula Sapiens for Vitamin D, IBD and CAD, and **Figure 4** presents results of the analyses for AD using tissue regions and cell-type data from ARCHS4 (**Figure 4a)** and Tabula Sapiens (**Figure 4b)**. Results for the other diseases using ARCHS4 can be found in **Supplemental Figure 13** and **Supplemental Tables 12-14**. We focus on Vitamin D, AD and IBD here because significant results were observed in unexpected tissues, such as the spleen and lung, and on CAD because it shows the largest number of associated tissues.

Overall, the results for ARCHS4, Tabula Sapiens and GTEx are consistent, although the *P*-values for all cell-type analyses tend to be larger. In the Tabula Sapiens, PoPS genes typically show higher expression than control genes, but these results are often not replicated in fine-mapped genes or nearest-gene lists (**Supplemental Tables 15-17**). Therefore, only results that replicate in at least two gene lists are discussed below.

GWAS to Gene Expression results for IBD in cell types derived from the small intestine, large intestine, and lung show that IBD genes present higher relative expression in T-cells (*P*-values for relative expression < 0.02), B cells (*P*-values < 0.038), neutrophils (*P*-values < 0.034) and dendritic cells (*P*-values < 0.01). Results from the *gene expression to GWAS* analyses were largely consistent with these findings.

For CAD, the most significant differences between disease and control genes were observed for relative expression of T-cells in the vasculature system (*P*-value = 0.0028), smooth muscle cells (*P*-value < 0.027) and cardiac fibroblasts (*P*-value < 0.045) in the heart, and fat tissues (*P*-value < 0.044 for fine-mapped genes, *P*-value < 0.014 for PoPS and nearest genes). Various immune cell types, such as macrophages, mast cells, and NK cells in the vasculature system were also significant (*P*-value < 0.015) but only for PoPS genes.

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

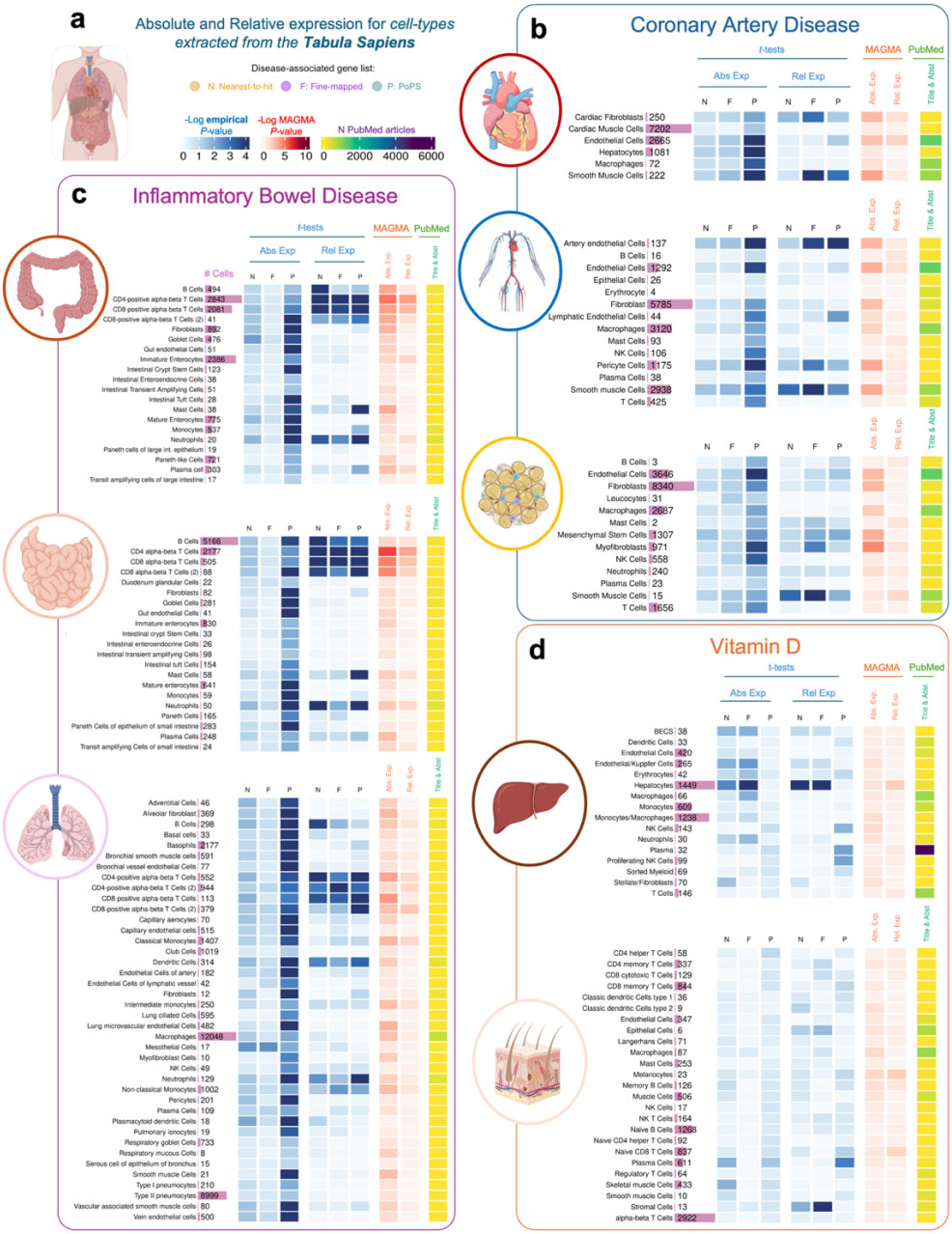


Figure 3. Heatmap showing results of the association between gene expression in each cell type obtained from the Tabula Sapiens dataset **a**, Tabula Sapiens datasets were downloaded for each tissue, and absolute and relative expression was calculated for each cell type in each tissue. **b**, Results for Coronary Artery Disease. **c**, Results for Inflammatory Bowel Disease. **d**. Results for Vitamin D. In blue, results showing the Log_{10} empirical *P*-value after running 10,000 permutations, testing the null hypothesis that disease-associated genes are not more expressed than other protein-coding genes expressed in that tissue. In red, results showing the Log_{10} *P*-value for enrichment of GWAS signal across the set of genes with highest expression for each tissue. In yellow, results for the Literature Search using PubMed. Abs. Exp, Absolute expression; Rel. Exp, relative expression; F, Fine-mapped genes; N, Nearest-to-hit genes, P, Polygenic Priority Scores genes; Title & Abst, Title and Abstract are used in the PubMed Search.

Figure 4a shows results for AD analyses using ARCHS4, showing high absolute and relative expression of AD genes in immune cell types, including dendritic cells, macrophagues and neutrophils (e.g. *P*-values < 0.05). **Figure 4b** shows similar findings using Tabula Sapiens, with significant expression in macrophages (in the spleen, blood, and fat) and neutrophils (in fat). Due to the lack of brain tissue data in Tabula Sapiens, brain-related cell types were examined only in the ARCHS4 dataset. Microglia – known to play a critical role in AD[36] – is the only brain tissue significantly associated for both absolute (*P*-value < 0.005) and relative expression (*P*-value < 0.05).

For SCZ, BP, and ADHD, disease genes show significant absolute and relative expression in motor neurons in ARCHS4 (*P*-value = 0.00021 for PoPS genes; *P*-value = 0.036 for fine-mapped genes), but not in the broader category of neurons, despite their larger sample size (**Supplemental Figure 13**).

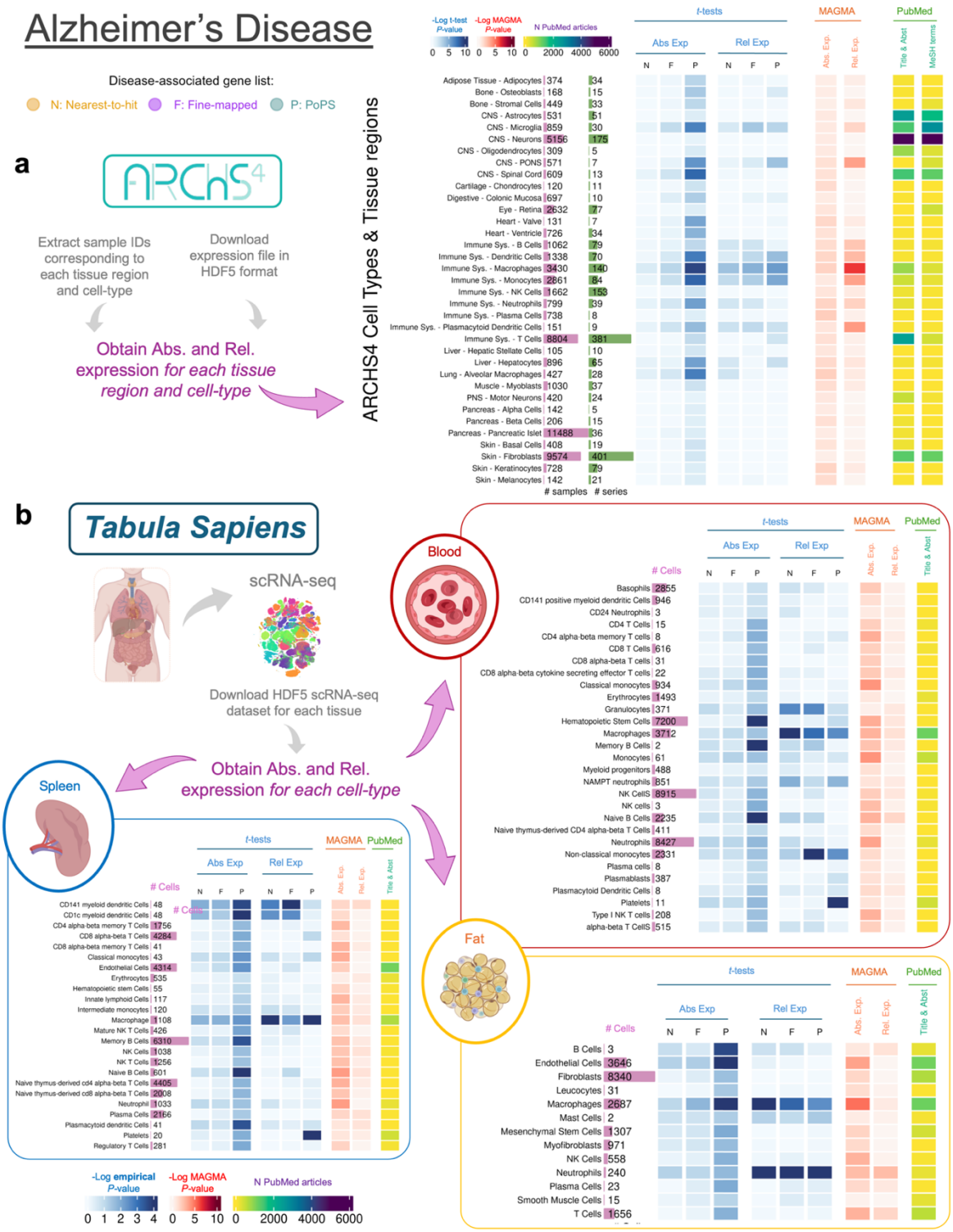


Figure 4. Heatmap showing results of the association between gene expression in each cell type obtained from the Tabula Sapiens dataset. **a**, Results using RNA-seq data from cell types and tissue regions extracted from the ARCHS4 resources. **b**, Results using RNA-seq data from the Tabula Sapiens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

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

267 We extended our analysis framework to test the gene expression landscape of eight common
268 cancer types (**Additional File 1, Supplemental Materials**). For three of the cancers, breast,
269 prostate, and colorectal cancer, we applied our full testing pipeline across GTEx, ARCHS4 and
270 Tabula Sapiens datasets since robust GWAS, prioritised gene lists, and high PoPS score genes
271 were available. For five additional cancers (bladder, kidney, lung, ovarian, and pancreatic
272 cancer), no gene prioritisation analyses had been performed[37], so we tested the "*GWAS to gene*
273 *expression*" approach using the 'nearest-to-hit' method on GTEx. Since cancers have well-
274 established primary tissues of origin (e.g., the colon for colorectal cancer), these analyses serve
275 as a positive control for benchmarking the different approaches and tests.

276

277 For the three cancer types with full prioritised gene lists, GTEx results show the most significant
278 and consistent associations with their respective cancer-relevant tissues (P-values < 10^{-8}):
279 mammary tissue for breast cancer (**Supplemental Figure 14)**, prostate for prostate cancer
280 (**Supplemental Figure 15**), and colon transverse for colorectal cancer (**Supplemental Figure**
281 **16**). Additional significant associations are observed in other tissues, with cell-type level
282 analyses from ARCHS4 and Tabula Sapiens providing further resolution. For example, in
283 prostate cancer, epithelial cells from prostate tissue (luminal, CPE-high luminal, and CGB3A1-
284 high club) had the strongest associations (*P*-values < 10^{-4}). For breast cancer, results vary by
285 gene prioritisation method, but B cells and mature luminal cells in the Tabula Sapiens mammary
286 tissue were consistently implicated in both the "*GWAS to Gene Expression*" and the "*Gene*
287 *Expression to GWAS*" approaches. In colorectal cancer, fibroblasts, monocytes and neutrophils
288 show the most significant associations, supporting the role of immune cells in the disease
289 mechanism. Results from analyses stratified by sex are similar (**Supplemental Figure 17**).

290

291 No significant associations are observed for the five cancers in which we employed the 'nearest-
292 to-hit' prioritisation method, probably due to the low number of genes available for each
293 analysis: 12 genes were prioritised for bladder, 13 for kidney, 16 for lung, 22 for pancreas and
294 28 for ovary (**Supplemental Figure 18**).
295

296 **Predictors of gene expression of disease genes**

297 To characterise the biological factors (such as tissue, sample ID, subject ID) and technical
298 factors (such as batch ID) that contribute to variability in the gene expression of each high-
299 confidence disease gene, we used the '*variancePartition*' R package (v.4.3)[38] (**Methods**).
300

301 Disease genes show considerable variability in the main factors that contribute to expression
302 variability (**Figure 5a** and **Supplemental Table 18**): for some, tissue type explains more than
303 70% of the expression variance, whereas for others, factors such as batch or collection method
304 are the primary predictors, and tissue type accounts for less than 20% of the variance. **Figure**
305 **5b, c** shows schizophrenia as an example, and **Supplemental Figures 19-21** and the R Shiny
306 website associated with this manuscript https://juditgg.shinyapps.io/diseasegenes/ include
307 gene-level results quantifying the contribution of each variable to the variation in expression of
308 each gene and disease.
309

310 Given the gene-level differences observed, we tested whether *variancePartition* results differ
311 between high-confidence disease genes and control genes (**Figure 5d**). Results were
312 consistent across individual gene lists and diseases, except for Vitamin D, where tissue type
313 explained more variance in disease genes than in control genes (**Supplemental Figs 22-24**).

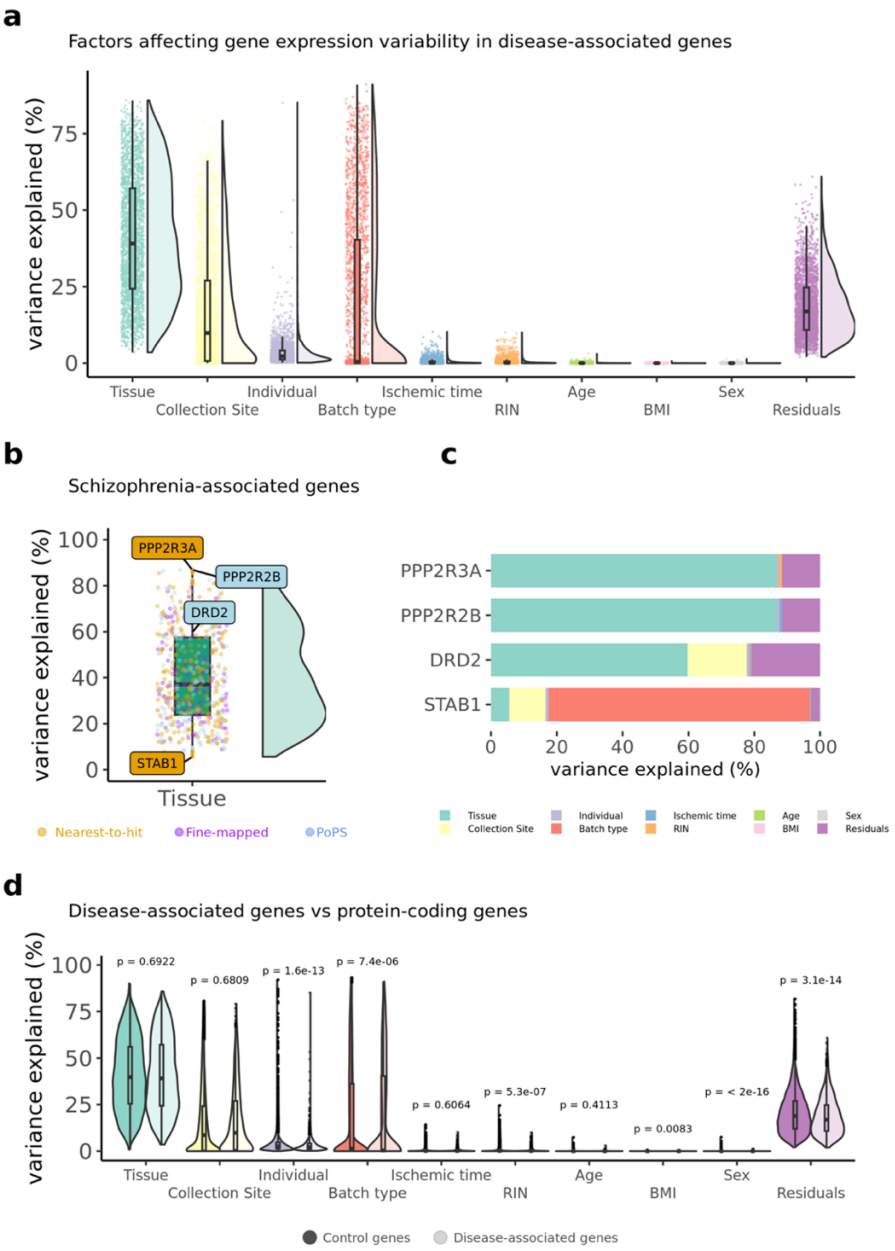


Figure 5. **Variance partition is used to investigate the factors influencing gene expression variability in disease-associated genes. a,** Violin plots representing the distribution of variance partition across all disease-associated genes for the eight diseases investigated. **b,** variance partition results for genes associated with schizophrenia. Genes labelled represent: two genes encoding protein phosphatases (PPP2R3A & PPP2R2B) where tissue-type explain a large fraction in gene expression variance, and a gene (STAB1) where tissue explains less than 10% in gene expression variance. The dopamine receptor 2 (DRD2) is also included because it is the main receptor for most antipsychotic drugs[39,40]. **c,** Bar plots of individual genes showing the variance partition estimates at the individual gene level for genes highlighted in panel b, **d,** Violin plots showing the differences in variance partition results between disease-associated genes and control genes.

Impact of gene expression on the inclusion of disease genes in drug development programs

Previous research has demonstrated that genes with support from GWAS are more likely to become successful drug targets [12,16], important given the high failure rates and associated substantial financial losses of the drug discovery process[41]. Similarly, genes with higher cell and tissue-type expression have been found to be more likely to enter drug development programs[42] and succeed as drug targets[19]. However, no studies to date have examined the association between gene expression and drug development success among disease genes prioritised by GWAS.

For each tissue, we conducted logistic regression analyses using gene expression levels as independent variables and the presence (yes/no) of each gene in drug development programs as the dependent variable. These regressions were performed for tissues in which disease genes exhibited significantly higher expression than control genes after Bonferroni correction. (referred as '*GTEx Tissues relevant to the disease*' in **Figure 6**). Gene–drug target pairs and their corresponding presence in drug development programs were extracted from a dataset with all target–indication pairs since 2000 (**Methods**).

Figure 6 presents results for CAD, SCZ, and IBD. CAD genes with higher absolute and relative expression in subcutaneous adipose, visceral adipose (omentum), breast mammary tissue, liver, and lung are more likely to be included in drug development programs – often for indications unrelated to CAD (**Figure 6a**). We further investigated this association and find that, out of the 552 CAD genes, 180 were investigated as potential drug targets for 518 different indications, including cancers (neoplasms), nervous system disorders, cardiovascular diseases, and other pathological conditions (**Figure 6b**). **Figure 6c** highlights in red CAD genes included

349 in drug development programs relative to their absolute and relative expression in adipose
350 tissues, liver and lung.
351
352 While SCZ genes with high relative gene expression are not more likely to be included in drug
353 development programs, those with higher *absolute expression* in key brain tissues are (**Figure**
354 **6d**). Of the 545 SCZ genes, 108 have entered drug development programs, primarily for
355 indications related to nervous system disorders, cancer, immune system disorders, pathological
356 conditions, and mental health disorders (**Figure 6e**).
357
358 IBD genes with higher relative expression in the small intestine (terminal ileum), blood and
359 spleen are more likely to be included in drug development programs (**Figure 6f** and
360 **Supplemental Table 19**). These results are driven by 198 out of 603 IBD genes that are already
361 included in drug development programs. Drugs targeting these genes were investigated for 529
362 indications related to cancer, the immune system, the nervous system, pathological conditions,
363 signs and symptoms, hemic-lymphatic diseases, and infections (**Figure 6f**).

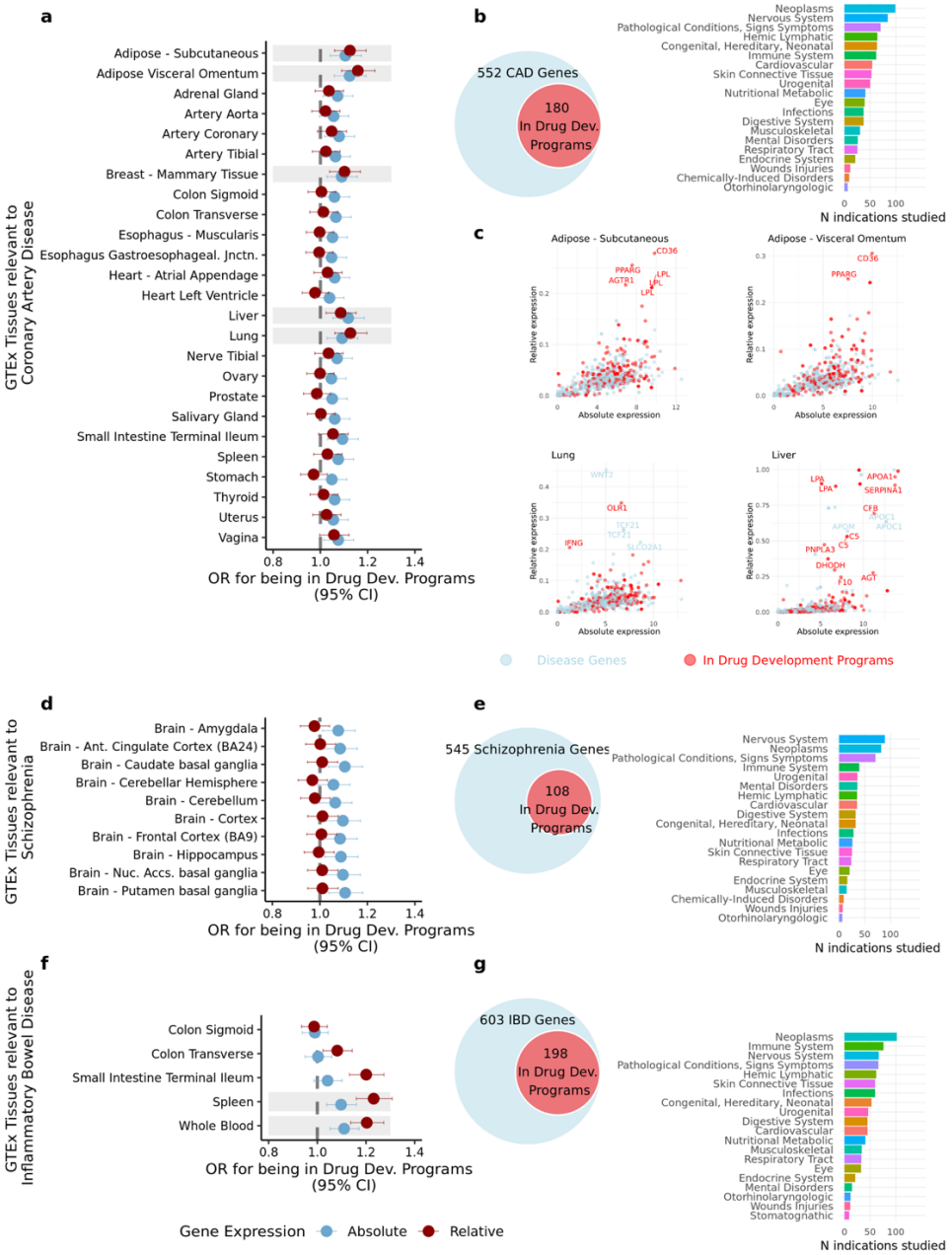


Figure 6. Impact of gene expression features on the inclusion of disease genes in drug target development.

a, d, f, Forest plots displaying the odds ratios (OR) for absolute and relative gene expression affecting the likelihood of disease-associated genes being incorporated into drug development programs for coronary artery disease (CAD), schizophrenia, and inflammatory bowel disease (IBD). Only relevant GTEx tissues – defined as those in which disease genes are significantly more expressed than control genes (based on the GWAS to Gene Expression approach) – are included. Shaded areas indicate tissues where both absolute and relative expression have a significant OR. **b, e, g,** Bar plots showing the number of disease-associated genes already included in drug development programs and the MeSH categories of the indications for which drugs were studied. **c,** Scatter plots depicting the relationship between absolute and relative gene expression in CAD-associated genes. Shown are four tissues where both measures significantly increased the likelihood of CAD genes being incorporated into drug development programs. Genes highlighted in red correspond to those that have been included in drug development programs.

375 **Discussion**

376 In this study, we systematically investigated where disease genes, prioritised using GWAS
377 results, are most expressed in the body, enabling insights into the initiation of disease and the
378 pleiotropic effects of disease genes. Using different definitions for disease and control genes,
379 three gene expression datasets at the tissue and cell type level, two measurements of gene
380 expression, and alternative analytical strategies, we consistently find that disease genes are
381 highly expressed in a broad range of tissues and cell types, including some without known links
382 to disease. Higher expression of disease genes in several of those tissues is also associated
383 with a greater OR for being included in drug development programs – often for indications
384 unrelated to the original disease in which the genes were prioritised. Altogether, these findings
385 could inform strategies for predicting side effects[43] and repurposing drugs targeting disease
386 genes.

387

388 While we use different analytical strategies to interrogate where disease genes act, a key novel
389 aspect of our study is the *GWAS to Gene Expression* approach. Previous work has largely
390 focused on analyses performed in the reverse direction – identifying tissues or cell types
391 enriched for GWAS signals (i.e., the *Gene Expression to GWAS* approach) [20,21,44] – which we
392 also use here to triangulate our results. However, the *GWAS to Gene Expression* approach that
393 we introduce here is based on high-confidence causal disease genes, identified through state-
394 of-the-art prioritisation methods applied to large-scale GWAS data. By focusing on genes most
395 likely to be causal, this approach offers a more targeted framework for drug discovery than
396 broader enrichment analyses. Our work also differs from other approaches that combine GWAS
397 and functional genomic data to make inference about disease aetiology. For example,
398 Transcriptome-Wide Association Studies (TWAS) integrate GWAS findings with expression
399 quantitative loci (eQTLs) to investigate how genetic variants influence gene expression[30]. TWAS
400 relies on the availability of eQTL data and on genes with highly heritable gene expression[45]. In

401 contrast, our approach provides a gene expression profile of disease-associated genes
402 regardless of eQTL data availability and gene heritability, broadening the scope for combining
403 GWAS signal and gene expression.
404
405 We first focus our study on tissue-level analysis because extensive prior knowledge of disease-
406 tissue associations provides a “ground truth” and, thus, informs the benchmarking of our
407 approach. Next, we apply our approach to Tabula Sapiens and ARCHS4 datasets, which offer
408 insights not easily discerned at the tissue level. For example, the results that we observe
409 between AD and IBD in spleen, blood and lung are likely driven by the high fraction of
410 macrophages and other innate immune system cells present within those tissues. However,
411 cell-type level datasets also have limitations: in the case of ARCHS4, the representation of cell
412 types and tissue regions is less systematic than in GTEx. Furthermore, ARCHS4's
413 heterogeneity may have reduced power to detect associations in other diseases despite
414 correction of batch effects. While *Tabula Sapiens*[25] may provide a more systematic multiorgan
415 dataset at the cellular level, the scale of this dataset is smaller (they measured single-cell RNA-
416 sequence data for a total of 15 individuals, respectively, in contrast to e.g. GTEx, which
417 assessed more than 700 individuals). Moreover, some tissues of interest were not available
418 here, such as brain tissues to interrogate cell types related to AD, SCZ, BP or ADHD.
419
420 There are two key considerations when interpreting our results and implementing our pipeline.
421 First, no gene prioritisation method is perfect, and therefore some genes categorised as
422 ‘disease genes’ may not, in fact, affect disease. Moreover, prioritisation methods that used
423 functional genomics datasets may be more powered, but they may provide circular results,
424 especially when using functional data from tissues hypothesised to be implicated in the disease.
425 This could be particularly relevant for the analyses of cancer traits, since prioritisation
426 approaches often rely exclusively on the specific cancer tissue RNA-seq dataset. Second, this

427 study mainly used expression datasets that included tissues and individuals considered to be
428 healthy. The gene expression profiles represent a healthy state and cannot be used to explore
429 the dynamics of gene expression during disease states. However, before studying the dynamics
430 of gene expression between cases and controls, we argue that it is important to understand the
431 tissues and cell types in which disease genes show high expression and specificity.
432
433 In conclusion, we present a comprehensive expression profiling of genes identified as high-
434 confidence disease genes. We identify known and novel tissues and cell types that could have
435 causal links to disease, factors that contribute to variability in the gene expression of disease
436 genes, and explore the potential relevance of gene expression to the inclusion of disease genes
437 in drug development programs. All together, we hope that this work opens new avenues for
438 guiding experimental follow-up and informs future drug development efforts targeting genes with
439 support from GWAS data, ultimately advancing our understanding of disease mechanisms and
440 response to treatment.
441

Methods

RNA-seq datasets

GTEx dataset

Gene expression measurements were obtained for 50 tissues from the GTEx project[24] version 8. Median gene TPMs for each tissue were downloaded from https://gtexportal.org/home/datasets. Standard RNA-seq processing steps were applied to the dataset as follows: (1) we filtered out all non-protein-coding genes and genes not expressed in any tissue; (2) we removed the tissues with less than 100 samples, cancer or cell related tissue types (i.e. EBV-transformed lymphocytes and Leukemia cell lines); (3) we scaled the expression of each tissue such that the total is 10^6 TPM. 45 tissues remained after quality control.

Sex-stratified analyses in the GTEx dataset

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

467 **ARCHS4 dataset**

468 Gene expression across cell types and tissue regions were extracted from the ARCHS4[25]
469 resource, which provides access to uniformly processed gene counts from human RNA-seq
470 experiments stored in the Gene Expression Omnibus (GEO) and Sequence Read Archive
471 (SRA). Using the ARCHS4 web browser (https://maayanlab.cloud/archs4/), we systematically
472 identified all the cell types available in the metadata search menu. For some disease-relevant
473 cell types like microglia in AD, we entered the cell type name directly into the metadata search
474 bar. ARCHS4 generates R scripts listing the samples related to each cell type, to facilitate their
475 extraction from the main repository –a HDF5 file named "human_gene_v2.2.h5" available for
476 download at https://maayanlab.cloud/archs4/download.html (Downloaded version date: 5-30-
477 2023).

478

479 The quality control of ARCHS4 datasets and *per-gene* TPM calculation was as follows: Only cell
480 types with more than 100 samples across all experiments available in ARCHS4 were included in
481 our analyses. Upon obtaining the counts expression matrix for each cell type, we performed
482 quantile normalisation of samples using the function '*normalise.quantiles'* available on the R
483 package ("preprocessCore"). Quantile normalisation was performed on raw counts. Given that
484 samples from a specified cell type may originate from multiple experiments with slightly different
485 conditions, we (1) excluded experiments containing less than 10 samples, and (2) adjusted for
486 batch effects using the package ComBat_seq[46], which is an improved version of the popular
487 ComBat[47]. Unlike its predecessor ComBat (designed for microarray data), ComBat_seq is
488 tailored for RNA-Seq studies and it does not assume a normal distribution of gene expression
489 data.

490

491 After batch correction, median TPM values for each gene were calculated in each cell type
492 using the formula:

493 $$\text{TPM} = \left(\frac{Number\ of\ mapped\ reads\ for\ gene}{Gene\ length\ in\ kilobases\ (kb)}\right) x\ 10^6 \div (Total\ number\ of\ mapped\ reads)$$

494

495 where 'Gene lengths in kilobases' were calculated using the genomic coordinates indicated in a

496 GTF file (built GRCh38, version 106), downloaded from ENSEMBL

497 (https://ftp.ensembl.org/pub/release-106/gtf/homo_sapiens/Homo_sapiens.GRCh38.106.gtf.gz).

498

499 **Tabula Sapiens datasets**

500 scRNA-seq datasets were obtained from the *Tabula Sapiens* figshare

501 (https://figshare.com/articles/dataset/Tabula_Sapiens_release_1_0/14267219). These datasets,

502 initially in '.h5ad' format, contained gene counts for each cell and metadata, and were converted

503 into Seurat objects and then into '*SingleCellExperiment*' objects to ensure the input data files

504 were in a similar format as the ones used for ARCHS4 and GTEx.

505

506 Data quality processing was performed: cells with zero counts across all genes were removed,

507 and outlier cells with an extreme number of detected genes were excluded. Pseudobulk data

508 was then generated using the *aggregateToPseudoBulk* function from the *dreamlet* R package to

509 aggregate expression counts across cell types, according to the free cell type annotation

510 included in the original datasets.

511

512 **Calculating 'Absolute gene expression 'and 'Relative gene expression' values**

513 Since the expression patterns of protein-coding genes tend to follow a negative binomial

514 distribution, we calculated *absolute levels of gene expression* by taking the Log_2 of the median

515 TPM+1 values. To calculate *relative gene expression*, we divided the *absolute levels of gene*

516 *expression* of each gene by its total expression across tissues. The resulting relative gene

517 expression ranged from 0 (gene is not expressed) to 1 (gene is exclusively expressed in this

518 tissue). The Log_2 of the absolute and relative expression measures were used in subsequent
519 analyses.
520
521 To calculate absolute and relative gene expression values for each gene in the *Tabula Sapiens*
522 dataset, the total expression counts for each cell type was obtained, and normalisation was
523 performed using the *calcNormFactors* function from the *edgeR* R package. Absolute gene
524 expression values were calculated as the total number of counts per million for each gene in
525 each cell type (after normalisation). Data was organised in data tables where each row
526 represented a gene, and each column represented a cell type. For relative expression of a gene
527 in a cell type, the *cellTypeSpecificity* function – which calculates the number of counts of a gene
528 in a cell type divided by the total number of counts across cell types in that tissue – was utilised
529 after applying the same normalisation procedure as for absolute expression.
530

531 **GWAS to Gene Expression**

532 **Definition of disease-associated genes inferred from GWAS results.**
533 Three different types of gene lists were inferred for each disease using GWAS results: nearest-
534 to-hit genes, fine-mapped genes, and PoPS genes. For all the gene definitions, we obtained
535 each gene ENSEMBL IDs using a GTF file (built GRCh37, version 75), downloaded from
536 ENSEMBL https://ftp.ensembl.org/pub/release-
537 75/gtf/homo_sapiens/Homo_sapiens.GRCh37.75.gtf.gz). The list of prioritised genes for each
538 approach and disease can be found in **Supplemental Tables 1-3.**
539
540 - *Definition of 'nearest-to-hit' genes*: To find the genes closes to the GWAS hit, we obtained
541 publicly available GWAS summary statistics for the diseases investigated. We performed
542 clumping using PLINK 1.9[48] and individual level genotype data from the 1000 Genomes

543 Project[49] as a reference linkage-disequilibrium (LD) panel. During clumping, variants with *P*-
544 values ≤ $5x10^{-8}$ were retained, and variants within a 250 Kbp window correlated ≥ 0.5 with the
545 index variant or variants with *P*-value ≥ 0.01 were removed. For each clump, the nearest
546 protein-coding gene to the index variant was identified. We used the GTF file (built GRCh37,
547 version 75) to extract the gene start and gene end coordinates of each protein-coding gene.
548 Information about the GWAS used, number of clumped variants, genes identified and distance
549 between variant and nearest gene is included in **Table 1**.
550
551 **Table 1**. Gene prioritisation based on the nearest gene to GWAS hit. Table shows references for each GWAS summary statistics,
552 the number of clumped SNPs, genes and median distance between SNP and selected gene. CAD, Coronary Artery Disease; SCZ,
553 Schizophrenia; IBD, Inflammatory Bowel Disease; AD, Alzheimer's Disease; BD, Bipolar Disorder; ADHD, Attention-
554 deficit/hyperactivity disorder; T2D, Type 2 Diabetes.

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[50]	487	285	11,724
AD	Bellenguez et al, 2022[11]	239	99	8,363
BD	Mullins et al, 2021[51]	73	63	20,660
ADHD	Demontis et al, 2023[8]	32	22	104,008
T2D	Suzuki et al, 2023[52]	50	31	26,633
Vitamin D	Revez et al, 2020[6]	517	200	22,111

555
556
557 - *Definition of 'fine-mapped genes'*: We acquired gene lists that previously fine-mapped GWAS
558 for CAD[3], SCZ[4], IBD[5], AD[53], BD[7], ADHD[8], T2D[9] and Vitamin D[54]. Most gene lists were
559 constructed using a combination of statistical fine-mapping, transcriptome association studies,
560 and mendelian randomization.

For CAD, the integration of eight gene prioritisation predictors enabled the identification of 220 likely causal genes[3]. For SCZ, statistical fine-mapping was integrated with summary Mendelian randomization and Hi-C interaction data to obtain a list of 120 prioritised genes[4]. For IBD, we extracted the list of genes linked to variants fine-mapped, available in the Supplemental material of the study conducted by Huang and colleagues[5]. For AD, we used a review of that reported a list of genes prioritised via fine mapping of GWAS in two previous studies[10,11]. The list is available in https://github.com/sjfandrews/ADGenetics/blob/main/results/adgwas_loci.csv. For BP, fine-mapping of the GWA signals was performed and seven complementary approaches were used to prioritise 47 credible genes that were mapped to loci by at least three of the seven approaches[7]. For ADHD, fine-mapping of the most recent ADHD GWAS[8] identified sets of credible variants for each risk locus. Credible sets were subsequently linked to genes based on genomic position, information about eQTLs, and chromatin interaction mapping in human brain tissue as implemented in FUMA. For T2D, we used a gene list containing the nearest gene of the results of a fine-mapping approach used in 380 independent association signals[9]. For Vitamin D, we extracted a list of genes published by *Manousaki* and colleagues[54], who prioritised genes using the DEPICT method[28] on a GWAS of serum 25 hydroxyvitamin D.

- *Definition of 'PoPS genes'*: The PoPS method[31] prioritises disease-associated genes by integrating gene-level z-scores from MAGMA[22], single-cell gene expression data, biological pathways, and predicted protein-protein interaction networks. The original PoPS publication suggests combining PoPS scores with location information would provide a list of high-confidence genes. However, the combined PoPS+location approach leads to a low recall (it detects very few genes for each disease). Therefore, we used the top 1% PoPS, because it results in a list of 184 prioritised genes per disease. This number of genes is similar to the

586 number of fine-mapped and nearest-to-hit genes, reducing differences in power due to the
587 number of genes assessed. The choice of the parameters top 1% of PoPS scores is further
588 discussed in the **Supplemental Material**; our results are robust across thresholds (**Additional**
589 **File 1, Supplemental Figure 3**). Full PoPS scores are available at:
590 https://www.finucanelab.org/data.
591
592 **Definition of control genes**
593 In this study, three control groups are defined. The control group used for our primary analyses
594 was defined by all protein-coding genes. Additionally, two additional control groups were tested,
595 using genes associated with other diseases, identified by the Open Targets initiative.
596
597 - *Definition of 'control genes' in our primary analysis:* We defined the control group as all protein-
598 coding genes expressed in the tissue or cell type being analysed. These results form the basis
599 of our main findings (e.g., Figures 2, 3, and 4). To identify these genes, we used the publicly
600 available GTF file (built GRCh38, version 106), downloaded from ENSEMBL
601 (https://ftp.ensembl.org/pub/release-106/gtf/homo_sapiens/Homo_sapiens.GRCh38.106.gtf.gz).
602 Using this file, we selected genes located on autosomes with the biotype designation "protein-
603 coding." Any protein-coding genes not included in the list of disease-associated genes were
604 classified as part of the control group.
605
606 - *Definition of 'control genes' groups, utilising genes associated with other diseases:*
607 To generate a list of control genes associated with multiple traits and diseases, we extracted two
608 lists of genes from the Open Targets resource[55]. The first group uses the *Open Targets 'Gold*
609 *Standards'*. The second group uses a list of genes prioritised via *Open Targets Genetics*
610 *evidence,* using a machine learning method[56] that calculates a disease-specific score to
611 prioritise genes.

612

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

620

621 Open Targets Genetics evidence: This list of genes was extracted from the results of a machine-
622 learning method used to identify the most likely causal genes[56]. This method integrates the
623 results of 1) fine-mapping credible set analysis, 2) functional genomics data such as
624 pathogenicity prediction, colocalization with molecular QTLs, genomic distance and chromatin
625 interaction data to generate predictive features. The machine-learning model is supervised
626 using the gold-standard positive GWAS loci, and a score is computed for each gene (named
627 Locus to gene (L2G) score). The L2G score is calibrated so that a gene's score indicates the
628 fraction of genes at or above the score that would be expected to be true positives. Thus, we
629 selected genes with a score >= 0.8, which assumes that 80% of the genes associated with a
630 trait or disease in our list are causal.

631

632 Data was downloaded from the publicly available website
633 https://platform.opentargets.org/downloads/data, section "Target - Disease evidence / Integrated
634 list of target - disease evidence from all data sources" (version 23/09), which provides several
635 directories with different evidence sources for the target-disease associations. However, only
636 the ones indicating 'genetics evidence' were used in our analyses. From those, 3,862 protein-
637 coding genes from 1,582 traits had L2G scores >= 0.8. The traits with the largest number of

genes were height (580 genes), blood protein measurement (359 genes), and heel bone mineral density (286 genes).

Statistical analyses to compare disease-associated genes vs control genes

The *t*-test is an inferential statistic used to evaluate whether the means of two independent samples are significantly different. Here, we run one-side *t*-tests in R, testing the null hypothesis that the expression of disease-associated genes is higher than those of the control group. *t*-tests assume that the sample means are normally distributed. Since gene expression follows a negative binomial distribution, we normalised the gene expression values by taking the Log_2 of the median TPM+1 before applying the *t*-tests.

The Anderson-Darling[34] is a non-parametric test to evaluate whether the gene expression of disease-associated genes originates from the same distribution than the control group of genes. It tests the null hypothesis that both groups were drawn from populations with identical distributions. The Anderson-Darling test is similar to other tests assessing differences between empirical distributions (such as the two-sample Kolmogorov-Smirnov test[57]), but it is more sensitive to differences in the tails of the distribution.

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

Gene Expression to GWAS

Definition of genes in the top decile of expression

For each tissue, we defined to gene-sets that are in the top decile of gene expression: one gene-set is composed by the 10% of genes with the highest *absolute* gene expression. The second gene-set is composed by the 10% of genes with the highest *relative* gene expression. To obtain these gene-sets, we first classified all protein-coding genes into 11 quantiles. In this classification, the 1st quantile is composed by genes without expression in a specific tissue, whereas the 11th quantile encompasses the genes with the highest expression values. We then grouped genes from the top quantile and tested their GWAS enrichment using MAGMA and the 1000Genomes as a reference panel. We expanded the gene coordinates by adding a 35 kb window upstream and a 10kb window downstream of the gene. The Major Histocompatibility Complex (MHC) region was excluded from the analyses due to their long-range LD.

Gene set enrichment analyses with MAGMA

MAGMA[22] is a software designed for gene-set enrichment analysis using GWAS data. It provides enrichment results at the gene-level and at the gene-set level. In gene-level analyses, MAGMA employs GWAS *P*-values to compute gene test statistics, accounting for LD structure via a reference dataset. For gene-set analyses, gene-level association stats are transformed into Z-scores, reflecting the strength of gene-phenotype associations. MAGMA uses a competitive gene-set test formula: $Z = \beta_0 + I\beta_p + C\beta_k + \epsilon$ where I is an indicator (1 if a gene is in gene-set p, 0 if not), and C is a covariate matrix. The resulting *P*-value originates from a test on coefficient β_p, evaluating if the phenotype shows a stronger association with genes included in the gene-set of interest versus other genes.

688 In our analyses, SNPs from GWAS data were mapped to genes based on genomic coordinates
689 defined in a GTF file (Genome build GRCh37). Gene boundaries were extended 35kb and 10Kb
690 around the 5' and 3' gene boundaries, respectively. The reference dataset used to account for
691 LD was the genotype data from the 1000 Genomes Project.
692

693 **PubMed search**

694 To determine which tissues are associated with specific diseases based on previous knowledge,
695 we interrogated how often a combination of tissue and disease terms appeared together in
696 published articles found on PubMed. To count the PubMed occurrences of a tissue being
697 mentioned in relation to a disease, we used the Python library *Beautiful Soup*[58], taking the
698 queries defined above as input. The script performs the following tasks: it generates
699 combinations of tissue-disease pairs, constructs search queries, sends requests to the PubMed
700 website based on these queries, and subsequently extracts the number of search results from
701 the webpage. The resulting count shows how frequently the tissue-disease pair appears in the
702 body of literature. Two types of PubMed search analyses we conducted based on the type of
703 query constructed: one using Medical Subject Headings (MeSH) Terminology, and another one
704 using keywords.
705
706 - *Search queries utilising MeSH Terminology:* The Medical Subject Headings (MeSH) thesaurus
707 is a curated collection of terms established by the National Library of Medicine. MeSH terms are
708 valuable in recognizing content that uses different words but refers to the same concept,
709 enhancing the accuracy and consistency of the literature search process. Leveraging MeSH
710 terminology, we prioritised the list of 45 tissues from the GTEx dataset based on their frequency
711 of occurrence within MeSH terms connected to scientific articles. For each pairing of tissue and
712 disease, a search query in the format *'<tissue name> [Mesh] AND <disease name> [Mesh]'* was
713 used.

714

715 - *Search queries utilising Keyword-based Literature Search:* The list of tissues and cell types
716 were ranked based on their citation frequency within the titles or abstracts of relevant scientific
717 articles. The construction of search queries followed the format '*<tissue name> [Title/Abstract]*
718 *AND <disease name> [Title/Abstract]'* for each unique tissue-disease pair.

719

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

721 To uncover the key contributors to the variability in gene expression among disease-associated
722 genes, we performed variance partition analyses using the R package '*variancePartition*'[38]. This
723 package assesses drivers of variation for each gene by fitting a linear fixed model to quantify
724 the contribution of tissues, individuals, technical variables etc. in gene expression.

725

726 We calculated the variance partition for each disease-associated and control gene. We used as
727 predictors uncorrelated variables ($r^2 < 0.75$) that explained the largest proportion of variance in
728 gene expression, as calculated by the Canonical Correlation Analysis in the *variancePartition*
729 package and reported in the original *variancePartition* publication[38] (which also used the GTEx
730 dataset). The variance partition analysis results in a data table where each row is a gene, and
731 each column is the predictor variable included in the model. The results show, for each gene,
732 the percentage of variance explained for each predictor.

733

734 **Assessing the impact of gene expression on the inclusion of disease genes in**
735 **drug development programs**

736 To evaluate whether absolute and relative gene expression influence the likelihood of disease
737 genes being considered as drug targets, we assessed whether higher absolute and relative

738 expression values across GTEx tissues increase the odds of disease genes being included in
739 drug development programs.
740
741 We extracted drug development data for all disease genes from the Citeline Pharmaprojects
742 database, as curated by Minikel et al.[12]. This dataset includes monotherapy programs initiated
743 since 2000, annotated with the human gene target for each drug under study and for each
744 disease indications the program was developed. The table was downloaded from the study's
745 Github repository (https://github.com/ericminikel/genetic_support/blob/main/data/pp.tsv).
746
747 For each disease investigated, we combined the three lists of disease-associated genes used
748 for the "GWAS to gene expression" approach: (i) nearest-to-GWAS hit genes, (ii) fine-mapped
749 genes, and (iii) PoPS-prioritised genes (top 1% PoPS threshold). Disease genes absolute and
750 relative gene expression for each GTEx relevant tissue were extracted. Relevant tissues were
751 defined using the same criteria as in the sensitivity analysis that excluded genes highly
752 expressed in disease-relevant tissues (**Additional File 1, Supplemental Material**). Specifically,
753 tissues were considered relevant if they showed significantly higher expression in all statistical
754 tests performed. A Bonferroni-corrected significance threshold was set at 0.05 / (45 x 3 x 2) to
755 account for 45 tissues, 3 gene prioritisation approaches, and 2 statistical tests (t-test and
756 Anderson-Darling test). Tissues where both tests yielded *P*-values < threshold were selected.
757 Expression values were classified into quantiles (10 and 100 quantiles were tested, leading
758 consistent results) to mitigate the effect of genes with extreme expression values inflating the
759 OR.
760
761 To assess the impact of gene expression on the inclusion of disease genes in drug development
762 programs, we performed logistic regression separately for each gene expression feature
763 (absolute or relative), for each disease and for each relevant tissue. The dependent variable

764 was a binary indicator, where each disease gene was assigned 1 if included in a drug
765 development program and 0 otherwise. The independent variable was either absolute or relative
766 gene expression quantiles of that gene in the relevant tissue. OR and 95% Confidence intervals
767 were calculated from the output of the logistic regression.
768
769 To categorize the indications for which drug targets had been investigated into broader groups
770 (as shown in **Figure 6**, panels b, e, and g), we extracted the MeSH IDs associated with all
771 indications from the Minikel dataset and mapped them to their ancestor categories in the MeSH
772 ontology tree. To retrieve the ontology hierarchy for each MeSH ID, we queried the MeSH
773 Resource Description Framework (RDF) dataset using SPARQL.
774
775

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 Aβ, 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 underutilised 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. Calderon, D. *et al.* Inferring Relevant Cell Types for Complex Traits by Using Single-Cell Gene Expression. *Am. J. Hum. Genet.* **101**, 686–699 (2017).
21. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
22. 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).
23. Shang, L., Smith, J. A. & Zhou, X. Leveraging gene co-expression patterns to infer trait-relevant tissues in genome-wide association studies. *PLoS Genet.* **16**, e1008734 (2020).
24. THE GTEX CONSORTIUM. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
25. Lachmann, A. *et al.* Massive mining of publicly available RNA-seq data from human and mouse. *Nat. Commun.* **9**, 1366 (2018).
26. THE TABULA SAPIENS CONSORTIUM. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).
27. Gazal, S. *et al.* Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. *Nat. Genet.* **54**, 827–836 (2022).
28. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6**, 5890 (2015).
29. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).
30. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
31. 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).
32. Backman, J. D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).
33. 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).
34. Stephens, M. A. EDF Statistics for Goodness of Fit and Some Comparisons. *J. Am. Stat. Assoc.* **69**, 730–737 (1974).

35. Skene, N. G. *et al.* Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet.* **50**, 825–833 (2018).
36. Hansen, D. V., Hanson, J. E. & Sheng, M. Microglia in Alzheimer's disease. *J. Cell Biol.* **217**, 459–472 (2018).
37. 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).
38. Hoffman, G. E. & Schadt, E. E. variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics* **17**, 483 (2016).
39. Creese, I., Burt, D. R. & Snyder, S. H. Dopamine receptor binding predicts clinical and pharmacological potencies of antischizophrenic drugs. *Science* **192**, 481–483 (1976).
40. 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).
41. Sun, D., Gao, W., Hu, H. & Zhou, S. Why 90% of clinical drug development fails and how to improve it? *Acta Pharm. Sin. B* **12**, 3049 (2022).
42. Dann, E. *et al.* Estimating the impact of single-cell RNA sequencing of human tissues on drug target validation. 2024.04.04.24305313 Preprint at https://doi.org/10.1101/2024.04.04.24305313 (2024).
43. 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).
44. Wang, R., Lin, D.-Y. & Jiang, Y. EPIC: Inferring relevant cell types for complex traits by integrating genome-wide association studies and single-cell RNA sequencing. *PLoS Genet.* **18**, e1010251 (2022).
45. Wainberg, M. *et al.* Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* **51**, 592–599 (2019).
46. Zhang, Y., Parmigiani, G. & Johnson, W. E. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genomics Bioinforma.* **2**, lqaa078 (2020).
47. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
48. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, s13742-015-0047–8 (2015).
49. The 1000 Genomes Project Consortium. A map of human genome variation from population scale sequencing. *Nature* **467**, 1061 (2010).
50. 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).
51. 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).

52. 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.
53. Andrews, S. J. *et al.* The complex genetic architecture of Alzheimer's disease: novel insights and future directions. *eBioMedicine* **90**, (2023).
54. Manousaki, D. *et al.* Genome-wide Association Study for Vitamin D Levels Reveals 69 Independent Loci. *Am. J. Hum. Genet.* **106**, 327–337 (2020).
55. Koscielny, G. *et al.* Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res.* **45**, D985–D994 (2017).
56. 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).
57. 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).
58. Richardson, L. beautifulsoup4: Screen-scraping library.

Code availability

We have released an open-source implementation of our approach, along with all derived gene lists from publicly available data (https://gitlab.com/JuditGG/GeneExpressionLandscape). This pipeline provides a flexible framework for analysing prioritised gene sets across diseases and can be adapted for future studies.

Acknowledgements

This work was supported by a grant from the National Institutes of Health (R01MH122866) to PFO, by a 2022 NARSAD Young Investigator Grant (Number 30749) by the Brain & Behavior Research Foundation to JGG, and a grant from the National Human Genome Research Institute (1K99HG013547-01) 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 ARCHS4, Tabula Sapiens and the GTEx projects and the scientists that shared their results 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. AC: Investigation, Software, Validation. 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.