Organisation of gene programs revealed by unsupervised analysis of diverse gene-trait associations

Dalia Mizikovsky[1], Marina Naval Sanchez[1], Christian M. Nefzger[1], Gabriel Cuellar Partida[2,*,#], Nathan J. Palpant[1,#]

[1] Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia
[2] University of Queensland, Diamantina Institute
[#] Co-corresponding authors

*Current address: 23andMe Inc.

Corresponding authors

Nathan Palpant
Institute for Molecular Bioscience
University of Queensland
Brisbane, QLD, Australia
E: n.palpant@uq.edu.au
T: 61 04 39 241 069

Gabriel Cuellar Partida
Diamantina Institute
University of Queensland
Brisbane, QLD, Australia
E: g.cuellarpartida@uq.edu.au

ABSTRACT

Genome wide association studies provide statistical measures of gene-trait associations that reveal how genetic variation influences phenotypes. This study develops an unsupervised dimensionality reduction method called UnTANGLeD (Unsupervised Trait Analysis of Networks from Gene Level Data) which organises 16,849 genes into discrete gene programs by measuring the statistical association between genetic variants and 1,393 diverse complex traits. UnTANGLeD reveals 173 gene clusters enriched for protein-protein interactions and highly distinct biological processes governing development, signalling, disease, and homeostasis. We identify diverse gene networks with robust interactions but not associated with known biological processes. Analysis of independent disease traits shows that UnTANGLeD gene clusters are conserved across all complex traits, providing a simple and powerful framework to predict novel gene candidates and programs influencing orthogonal disease phenotypes. Collectively, this study demonstrates that gene programs co-ordinately orchestrating cell functions can be identified without reliance on prior knowledge, providing a method for use in functional annotation, hypothesis generation, machine learning and prediction algorithms, and the interpretation of diverse genomic data.

INTRODUCTION

Generation of consortium-scale data such as ENCODE (1), the Human Cell Atlas (2) and the UKBiobank (3) coupled with the development of advanced computational methods is enabling the creation of transformative models that harness the natural diversity of biological systems. These models draw on the relationships and patterns derived from biological data to establish quantitative frameworks that can make highly accurate predictions, with implications for nearly every field of biology. For example, in the field of structural biology, patterns in the sequences and structures of proteins' evolutionary homologs reveal how amino acids interact, enabling prediction of protein structure with atomic accuracy (4). Similarly, patterns of repressive histone methylation (H3K27me3) across hundreds of human cell types enable identification of genes governing cell decisions and functions for any cell type and organism (5).

Genome wide association studies (GWAS) characterise the genomic variation underlying complex traits and diseases, providing insights into how genes affect biological processes (6). Despite the wealth of variant-trait association information, GWAS studies predominantly focus on elucidating the genetic basis of a single trait or a group of highly related traits (6, 7). Here, we utilize patterns of genomic variation across hundreds of diverse phenotypes as the basis for an unsupervised method to parse the organisation of gene programs in cells.

We hypothesised that complex traits are underpinned by conserved gene programs that can be identified by studying associations between genetic variation and phenotypic variation. To test this, we developed UnTANGLeD (Unsupervised Trait Analysis of Networks from Gene Level Data), which identifies patterns of association between genes and hundreds of diverse phenotypes. UnTANGLeD creates a phenotypic signature to cluster genes with similar associations across many traits in an unsupervised manner into gene programs controlling cell biological processes **(Figure 1)**.

We used a gene-trait association matrix derived from GWAS data for 1,393 complex traits to infer co-ordinately acting gene programs that represent both known and novel biological processes. While the scale of associations available from public GWAS data is underpowered to saturate the accuracy of our model, we demonstrate that UnTANGLeD can be applied to any orthogonal GWAS data to predict the genetic basis of disease including in underpowered and transethnic GWAS data. UnTANGLeD provides a powerful analytical framework for studies in population genetics, cell biology, and genomics, that will improve as more data emerges.

Collectively, this study provides a statistical framework for defining genes orchestrating biological processes by evaluating genetic signatures across diverse complex traits.

MATERIALS AND METHODS

Data Collection

S-MultiXcan results for 1,393 phenotypes with statistically significant SNP-based heritability ($p < 0.05$) were downloaded from CTG-VL (http://vl.genoma.io). Phenotypes are listed in **Table S2**. SNP-based heritability was estimated using linkage disequilibrium score regression (LDSR). The significance values reflecting the strength of the association between each gene and trait across all tissues were compiled into a gene-trait association matrix.

Dimensionality Reduction Analysis Pipeline

All genes with fewer than 2 significant associations across all phenotypes ($p < 10^{-4}$) were removed, leaving 16849 genes. Following this, all values in the gene-trait association matrix were chi-squared transformed. Infinite values produced when transforming very small p-value (<1e-300) due to floating point precision were replaced with 1,415, which was 5 greater than the largest non-infinite value. The data was then normalised by the sum of chi-squared values per phenotype and scaled by a factor of 10,000. 10 principal components were retained from the principal component analysis (PCA). Clustering of genes was performed using the native *Seurat* shared-nearest neighbour algorithm. Clustering iterations were performed at increasing resolutions from 0.2 to 20 in increments of 0.2. The resolution is a parameter from Seurat where increased values lead to a greater number of clusters. Cluster assignments were compiled into a consensus distance matrix, where each gene pair had a value representing how often they were grouped together out of 100 potential matches. 100 was then subtracted from the values and they were made absolute to transform the matrix into a dissimilarity matrix. Agglomerative clustering using Ward's minimum variance method, as implemented in the *stats* package, was applied to the consensus matrix directly. The average silhouette score (a metric used to calculate how well a data point relates to its cluster) across all genes was calculated using the *cluster* package from 2 to 300 clusters. The *inflection* package was used to calculate the plateau point, which was determined to be the optimal number of clusters. Pearson's correlation was used to determine the correlation of a gene with the other genes in the same cluster based on chi-squared association values.

126 ***Enrichment Analyses***

127 GO, DO, KEGG enrichment, colocalization and tissue specificity enrichment were performed
128 using *ClusterProfiler* (8). An FDR corrected significance value of $p < 0.01$ was used.
129 Colocalization was determined using *ClusterProfiler* enrichment for the Molecular Signatures
130 Database collection 3: positional gene sets (9). The largest proportion of genes within a cluster
131 belonging to a single genomic region was divided by the total number of genes within the
132 cluster to identify the maximum degree of colocalization. STRING enrichment analysis was
133 performed using the *STRINGdb* package, with a significance threshold of $p < 0.001$ and a
134 confidence threshold of 0.400. STRING enrichment analysis without the text-mining
135 component was performed using the online STRING interface (https://string-db.org/) for
136 clusters found to have PPI enrichment in the prior analysis with a confidence threshold of 0.150
137 to preserve predicted interactions reinforced by other components. For the calculation of the
138 correlation between the loss of enrichment and the degree of colocalization, clusters 111 and
139 173 were removed due to having well established biological functions despite being highly
140 colocalised. Broad enrichment analysis for more specialised gene sets was performed using
141 *EnrichR* (https://maayanlab.cloud/Enrichr/) across all 192 libraries. Redundant libraries,
142 including GO, KEGG, chromosomal location and NIH-grant associated libraries were
143 excluded. The top significant term from each library for each cluster are reported in **Table S9**.
144 A significance value threshold of 0.01, after correction for multiple testing, was used. For
145 identification of genes possessing the same protein domains or belonging to the same family,
146 the EnrichR library 'Pfam_Domains_2019' was used. A distinct protein family or domain was
147 defined by collating the family or domain terms together that shared genes until there was no
148 overlap between them. Protein terms did not need to be significantly enriched, but two or more
149 members of a protein family had to be present in a single cluster.

150

151 ***Permutations***

152 Five permutations were generated by re-ordering the values within the gene-trait association
153 matrix. These permutations were analysed as described above. A one-way ANOVA with FDR
154 corrected pairwise comparisons was performed to identify significant differences in the number
155 of enriched clusters, total enriched GO terms and the most significant GO enrichment of any
156 cluster.

157

158 ***Phenotype Associations***

159 The gene-trait association matrix containing p-values was -log10 transformed. All infinite
160 values generated due to floating point precision were windsorized with 315, which was 5
161 greater than the maximum finite value. The phenotypic associations for the genes within a
162 cluster were extracted, averaged, normalised for their average associations across the dataset
163 and ranked.

164

165 ***Clustering quality in dimensionality reduction methods***

166 We extracted the UMAP coordinates for all genes as calculated by *Seurat*. Following this, we
167 identified the 10 closest neighbours for each gene and calculated the average correlation of chi-
168 squared association values between the gene and its neighbours. The UMAP was re-plotted
169 representing the average correlation with each point colour. We repeated the process, instead
170 colouring by the number of significant associations for each gene.

171

172 ***Prediction of novel genes using an underpowered GWAS of the same trait***

173 *Data collection and S-MultiXcan Analysis*

174 We selected 13 phenotypes for which GWAS studies had been performed at differing cohort
175 sizes or ethnicities for the same, or comparable traits. The specific studies and their respective
176 details can be found in **Table S1.** Summary statistics were downloaded from various sources
177 and harmonised using MetaXcan's in-built harmonization
178 (https://github.com/hakyimlab/MetaXcan) to be compatible with the MASHR models. We then
179 performed S-MultiXcan analysis of each trait using the MASHR models built off the V8 GTEx
180 release. Associated genes were defined as those found to have a significance of $p < 10^{-4}$
181 by S-MultiXcan.

182

183 *Global Clustering Coefficient Calculation*

184 The genes identified for an independent GWAS were projected onto the 173 identified clusters.
185 Following this, we generated an unweighted adjacency matrix in which genes in the same
186 cluster were represented by a 1, and genes in different clusters by a 0. A comparison between
187 the same gene was represented by 0. Finally, the global clustering coefficient (GCC) for the
188 genes was calculated. To derive a statistical significance, we randomly sampled the same
189 number of genes as there were significant genes for the phenotype and calculated the GCC one-
190 hundred times. A Z score was calculated from the curve generated by the sampled values.

191

192 *Gene Prediction*

193 We took a simple approach of predicting which clusters were associated with the trait using the
194 S-MultiXcan associations from the smaller GWAS and then checking whether novel gene
195 associations identified by the larger GWAS were in those clusters. A chi-squared enrichment
196 test was used where the minimum expected frequency was greater than 5, and a fisher's test if
197 not. Several approaches to predict clusters associated with the trait were trialled. The first was
198 to identify any of the 173 clusters with a significant gene in it. The second was to integrate the
199 additional phenotype into the trait-gene association matrix. Next, clusters were identified which
200 had an overall significance signature > 1.5 times the average or were significantly ($p < 0.05$)
201 higher than the average signature. Different values were tested for these thresholds, with these
202 providing the best performance. The third approach was to predict associated clusters from the
203 previously established 173 clusters using the thresholds taken in approach two. A one-way
204 ANOVA was performed with pairwise comparisons to determine the best approach. Approach
205 three was the most effective, albeit not significantly, while maintaining a low computational
206 burden. In instances where transethnic GWAS were compared, the East-Asian GWAS was
207 used to predict the trait relevant clusters, and the European GWAS was used as the test set.

208

209 ***Gene Prioritization Analysis***

210 The GWAS with the largest sample size for each of the 13 traits listed in **Table S1** was used to
211 determine the potential of our pipeline for prioritizing genes within a locus. Clumping was
212 performed on each summary statistic using PLINK (https://www.cog-genomics.org/plink/) and
213 1000 genomes phase 1 genotype data with an LD threshold of 0.5. This was followed by
214 clumping for long distance LD, at the same threshold. Next, we identified individual 500kb
215 regions around the lead SNPs and the genes within that region.

216

217 We took a leave one chromosome out (LOCO) approach, where we removed all potential genes
218 on one chromosome. With the remaining genes, we identified which clusters were enriched for
219 genes associated with the trait. To calculate enrichment, we treated all genes associated with
220 one locus as one positive, so that enrichment was for different loci and not genes at the same
221 locus. A fisher's enrichment test was used to determine significance. Finally, we assessed at
222 what proportion of loci the UnTANGLeD clusters identified a gene when that chromosome
223 was left out of the analysis.

224

225 ***Normalisation***

226 We trialled relative count, centralised-log ratio and logarithmic normalisation on the chi-
227 squared transformed values of the gene-trait matrix across phenotypes. We evaluated their
228 effects on the following metrics: correlation score, silhouette score, GO and STRING
229 enrichment, global clustering coefficient, prediction of GWAS. A Kruskal Wallis one-way
230 analysis of variance was used to evaluate differences. Relative count was used for the final
231 pipeline.

232

233 ***Phenotype Filtering Based on Euclidean Distance***

234 A distance matrix between phenotypes using chi-squared transformed, RC-normalised data was
235 generated using the Euclidean distance formula from the package *wordspace*. Phenotypes with
236 a Euclidean distance below a set threshold, which indicated a high degree of relatedness, were
237 removed from the data, leaving the phenotype with the highest number of significant
238 associations. This was performed for thresholds 0 to 62, at which too few phenotypes remained
239 to cluster the genes using the dimensionality reduction methods. GO enrichment was used to
240 evaluate the clustering efficacy at each threshold.

241

242 ***Phenotype Subsampling and Sensitivity Analyses***

243 Phenotype subsampling was performed on two datasets; MultiXcan results for 1393 phenotypes
244 across 16,849 genes generated in this paper, and another dataset containing MultiXcan results
245 for 4091 phenotypes across 15,734 genes (phenomexcan.org). For the data containing 1393
246 phenotypes, subsampling was performed randomly without replacement from 50 to 1393
247 phenotypes in 20 equal increments across 5 replicates for each number of traits. The full
248 UnTANGLeD clustering pipeline was applied to each subsampled matrix. Adjusted rand index
249 (ARI) was calculated for each of the subsampled clustering configurations compared to the full
250 dataset. This analysis was repeated for the data containing 4091 phenotypes; however,
251 subsampling was performed from 50 phenotypes to 4091 phenotypes in 50 equal increments.

252

253 ***Cluster Conservation***

254 To explore the cause for the marked increase in ARI between 1322 phenotypes and 1393
255 phenotypes, cluster conservation was calculated between them. For each cluster from 1393
256 phenotypes, the proportion of genes that remained grouped together in each of the clusters from

257 1322 phenotypes was calculated. That proportion was used to assign a conservation score to
258 each gene, depending on how large the proportion of cluster the specific gene remained with
259 was. The same approach was applied between 4091 phenotypes and 4009 phenotypes.
260
261 **RESULTS**
262
263 **Unsupervised identification of gene groups with shared complex trait associations**
264 We used MultiXcan results from CTG-VL (10) derived from publicly available GWAS
265 (primarily from UK Biobank, on ~400,000 individuals) to create a gene-trait association matrix
266 for 16,849 genes and 1,393 traits **(Figure 1, Figure S1, Table S2)**. For each gene trait pair,
267 MultiXcan predicts whether trait-associated variants alter the gene's expression. The chi-
268 squared transformed significance value for each gene-trait association pair was compiled into
269 the gene-trait association matrix **(Figure 2A)**. These values were normalised using relative
270 count normalisation to account for the difference in power between phenotypes. Performance
271 was not significantly different using other normalisation methods including centralised log ratio
272 or log normalised data **(Figure S2)**. The data was then clustered using *Seurat,* a dimensionality
273 reduction method commonly used to analyse single cell RNA sequencing data to cluster cells
274 into related groups (11). Here, we use *Seurat* to test whether the calculated gene-trait
275 associations could be simplified into biologically enriched gene clusters. Clustering was
276 performed across 100 stepwise increases in resolution, a parameter which increases the number
277 of gene clusters. Repeat iterations provided an opportunity to survey both the broad scope of
278 biological processes that could be identified, as well as the specificity that could be achieved
279 with each biological process **(Figure 2B)**.
280
281 To test the biological validity of the derived clusters, we used positive gene sets as defined by
282 gene ontology (GO) (12) and STRING (13) to show that gene clusters have significant
283 enrichment for GO biological processes and STRING protein-protein interactions **(Figure 2C**
284 **and 2D)**. To demonstrate that the observed enrichment is driven by distinct gene-trait
285 association signatures rather than chance, we performed permutation analyses in which the
286 values in the data matrix were randomly re-ordered. Permutations had significantly fewer
287 enriched clusters, GO terms and a lower strongest significance compared to the real data ($p <$
288 $4x10^{-27}$) **(Figure 2C, Figure S3A-B)**. Furthermore, we validated that GO enriched clusters
289 were more likely to also have enrichment for protein-protein interactions, suggesting the
290 enrichment is robust **(Figure 2D)**. This analysis revealed that genes possessing similar

291 associations to complex trait phenotypes cluster meaningfully into biologically enriched groups
292 and the enrichment is not stochastic.
293
294 We next developed an ensemble learning method we call "consensus clustering" that
295 incorporates a measure of clustering robustness and quality. Across each of 100 stepwise
296 increases in clustering resolution we evaluated the robustness of clustering by assessing how
297 often every possible gene combination was clustered together ranging from 100 (always) to 0
298 (never) and compiled these values into a consensus matrix **(Figure 2E)**. Following this, we
299 performed agglomerative hierarchical clustering, evaluating the average silhouette score at
300 each possible number of clusters. The silhouette score quantifies how consistent genes within
301 the same cluster are across *Seurat* resolutions. To derive the optimal number of gene clusters,
302 we calculated the plateau point of the average silhouette score, which informs the number of
303 clusters at which point further splitting no longer improves the stability of clustering
304 assignments **(Figure S3C)**. Applying this methodology to gene-trait associations for 16,849
305 genes, we identified 173 clusters with an average of 97 genes **(Figure 2F, G).** Across each
306 cluster, we measured the silhouette score, a metric of cluster robustness and the correlation
307 score, a metric of relation across phenotypes, thereby providing two metrics to quantify the
308 quality of clustering **(Figure 2H)**. Collectively, we call this approach UnTANGLeD:
309 Unsupervised Trait Analysis of Networks from Gene Level Data.
310
311 **Consensus clustering identifies robust gene groups enriched for known gene sets**
312 We analysed each cluster by reference to curated annotations of gene programs (GO, disease
313 ontology (14)), signalling pathways (KEGG (15)), protein-protein interactions (STRING), and
314 tissue specificity (16) to evaluate the ability of UnTANGLeD to identify distinct, biologically
315 established gene programs in an unsupervised manner **(Figure 3A, Figure S3D, Tables S3-8)**.
316 This analysis revealed significant enrichment of cell biological pathways and networks across
317 gene clusters, with stronger enrichment among clusters with higher silhouette and correlation
318 scores **(Figure 3A)**. We further performed enrichment analysis of the UnTANGLeD clusters
319 using the EnrichR database (17) **(Figure S4A-C, Table S9),** finding considerable enrichment
320 for disease-associated genes, gene-expression perturbations associated with disease states or
321 drugs and protein domains and families. We note that although many clusters contain multiple
322 members of a protein family (18) the proportion of any one protein family in the cluster is
323 minor **(Figure S4D, Table S10)**.
324

325 We next investigated the relationship between individual gene clusters and the traits most
326 strongly influencing the genes within the clusters, using enriched GO processes as a proxy for
327 the functional profile of a cluster **(Figure 3B).** Each cluster is defined by a distinct gene-trait
328 association 'signature' indicated by the variation and strength of association across 1,393
329 diverse complex traits. In some instances, the enriched biological processes for certain gene
330 clusters are clearly related to the cluster's most significantly associated complex trait
331 phenotypes (e.g., cluster 119: GO enrichment: Cholesterol Homeostasis; Dominant complex
332 trait phenotypes: Low-density lipoprotein, Alipoprotein B quantile).
333
334 Since UnTANGLeD draws on associations across diverse phenotypes to inform gene-gene
335 relationships, the method can identify gene groups with enriched functions that are apparently
336 biologically independent of the phenotypes most significantly associated with the genes in the
337 cluster. For example, cluster 80, enriched for embryonic morphogenesis (GO:004859), is most
338 significantly associated to the phenotype Bone Mineral Density and cluster 111, enriched for
339 nucleosome organisation (GO:0034728), is most significantly associated to the phenotype
340 Mean Corpuscular Haemoglobin. These results support the central hypothesis that genes with
341 shared effects across diverse phenotypes can be clustered into gene groups controlling shared
342 biological functions and processes in an unsupervised manner **(Figure 3B)**.
343
344 Importantly, we show that the GO enriched gene clusters show no overlap in their strongest
345 enriched biological functions, and almost no overlap in their top 5 enriched terms,
346 demonstrating the use of gene-trait association data to parse novel biological gene programs
347 encoded within the genome **(Figure 3C)**.
348
349 Stratifying clusters by their silhouette and correlation scores reveals a higher level of GO,
350 STRING, KEGG, DO and tissue specificity enrichment with higher clustering quality,
351 indicating that the metrics provide an accurate representation of cluster quality **(Figure 3A, D)**.
352 Furthermore, both the robustness of clustering and the presence and strength of GO and
353 STRING enrichment are correlated with the number of significant associations to phenotypes
354 per gene (Pearson's correlation, $r > 0.65$), as well as the stability of clustering (Pearson's
355 correlation, $r > 0.69$) (**Figure 3E-G, Figure S5A-F)**.
356
357 Lastly, we note that there is considerable colocalization of genes within clusters, with a stronger
358 relationship between the correlation score and the degree of colocalization for the genes in a

359 cluster (Pearson's correlation, r = 0.77), than the cluster robustness (Pearson's correlation, r =
360 0.34) **(Figure 3H, Table S11).** STRING enrichment may also be inflated due to the text-
361 mining component, as findings from GWAS may be incorporated into the database, with genes
362 in proximity often being reported together. Indeed, we find that the loss of enrichment due to
363 removal of the text-mining component is correlated with the colocalization of the cluster (r =
364 0.60) **(Figure S6A-B)**. However, clusters with a high degree of colocalization are not
365 necessarily artefacts of false-positive associations identified by MultiXcan. For example,
366 clusters 173 and 111 are strongly enriched for immune processes and chromatin organisation
367 respectively, despite being highly colocalised **(Figure S6C-D).**
368
369 **Subsampling reveals need for more data to improve accuracy of UnTANGLeD**
370 We next sought to determine how the number and diversity of phenotypes influences the
371 accuracy and utility of UnTANGLeD clusters. We show that the number of GO enriched
372 clusters is highly correlated with the number of phenotypes utilised in the analysis (Pearson's
373 correlation, r = 0.85), even when phenotypic diversity is preserved **(Figure S7A)**. To further
374 test this, we performed phenotype subsampling and evaluated clustering accuracy using an
375 adjusted rand index (ARI) analysis. We found that clustering accuracy compared to the full
376 data improved with the addition of more phenotypes, however a marked increase in ARI
377 between 1322 and the full data set suggests that inaccuracy in clustering that isn't determined
378 by phenotypic diversity can be attributed to genes which have weak signatures and few
379 significant associations **(Figure S7B)**. We repeated subsampling in a larger dataset containing
380 MultiXcan analysis of 4091 phenotypes retrieved from Pividori *et al*. (2020) which resulted in
381 the same outcome **(Figure S7C)**. Comparison of the two data sets revealed that genes already
382 having many significant associations simply had more associations in the larger dataset with
383 both datasets possessing an equal proportion of genes with few to no significant associations
384 **(Figure S7D)**. Further, that genes with higher numbers of significant associations have higher
385 degrees of conservation **(Figure S7E-F)**. It's likely that the effective number of traits is similar
386 between the two datasets, as both mostly draw on the UK Biobank and have many highly
387 correlated phenotypes
388
389 Cumulatively, these findings indicate that the quality of gene clustering is dependent on the
390 scale and quality of data needed to derive high silhouette and correlation scores as a basis for
391 efficient enrichment of functional gene clusters. Accordingly, as more data becomes available,
392 the quality and accuracy of UnTANGLeD will improve. However, simply increasing the

393 number of phenotypes leads to an increase in redundant associations, and therefore strategies
394 to increase the number of significant gene-trait associations across the genome should be
395 employed, such as diversifying phenotypes and increasing cohort size.
396
397 **UnTANGLeD clusters are conserved across traits and can predict novel trait associated**
398 **genes**
399 GWAS require collections of large cohorts comprising thousands of individual-level genotype
400 data to characterise the genetic architecture of a trait. Furthermore, collecting enough samples
401 can prove challenging for many diseases, and as such they are often underrepresented in
402 biobanks.
403
404 We hypothesised that UnTANGLeD gene clusters would be conserved across complex traits.
405 To test this, we investigated an independent GWAS of ulcerative colitis (UC) (19) **(Figure 4A).**
406 We show that the 278 genes associated with UC ($p < 10^{-4}$) **(Figure 4B)** were significantly more
407 clustered within the UnTANGLeD clusters than expected by chance ($p = 2x10^{-9}$) **(Figure 4C)**.
408 The result shows that despite not being used to construct the clusters, UC associated genes
409 nevertheless group within the UnTANGLeD clusters, demonstrating that the defined gene
410 programs are conserved. We replicate our findings in 6 additional independent GWAS
411 phenotypes, highlighting that the UnTANGLeD clusters are conserved across a broad
412 phenotypic space (3, 20–28) (**Figure 4G)**.
413
414 We next tested whether the gene clusters can be used to predict novel genes and cellular
415 processes underpinning independent complex trait data. To test this hypothesis, we examined
416 two GWAS for UC. The first was performed in 2013 with 6,687 cases and 19,718 controls (29),
417 and the latter in 2017 with 12,366 cases and 33,609 controls (19) **(Figure 4A)**. MultiXcan
418 analysis of the summary statistics identified 153 and 278 genes respectively, with an overlap
419 of 53 genes **(Figure 4B)**. We projected the MultiXcan associations for the 2013 GWAS onto
420 the 173 clusters, identifying clusters were statistically associated with UC **(Figure S8)**. Finally,
421 we tested whether the clusters predicted from the 2013 GWAS contained novel genes identified
422 by the 2017 GWAS. Of the 225 novel genes identified by the 2017 GWAS, our approach was
423 able to use the 2013 GWAS to predict 120 with a significant enrichment for predicting UC
424 associated genes compared to other genes ($p < 3x10^{-121}$, chi-squared test) **(Figure 4D)**.
425

426 GWAS of the same complex trait conducted in populations of differing ancestries may
427 implicate both shared and distinct loci. We tested whether UnTANGLeD clusters are conserved
428 for genes specific to non-European ancestries, given that the UnTANGLeD gene clusters are
429 built from a European cohort. To test this, we examined a GWAS for triglyceride levels in an
430 East Asian population, which identified 34 genes (30) **(Figure S9A-B)**. Mirroring our findings
431 in a GWAS conducted on a European population, we found that the genes associated with
432 triglyceride levels in an East Asian population are significantly more clustered than expected
433 ($p = 1x10^{-9}$) and replicate this finding in 4 other GWAS conducted in populations of non-
434 European ancestry (30–32). We further tested whether the GWAS conducted in the East Asian
435 cohort could be used to predict novel genes identified in a European cohort. We found that
436 clusters implicated in triglyceride levels using the East Asian GWAS were highly enriched for
437 genes identified by the European GWAS ($p = 6x10^{-109}$) **(Figure 4F)**.
438
439 All together, we show significant enrichment for prediction of novel genes across GWAS
440 performed for 7 traits in differing cohort sizes in a European population, and 4 traits for which
441 GWAS were performed in different ancestries (3, 20-28, 30–33) **(Figure 4G, Figure S9C)**.
442
443 We further tested whether the UnTANGLeD clusters could be used to prioritize causal genes
444 at any given locus. It is difficult to accurately identify the causal genes from GWAS identified
445 variants due to linkage disequilibrium and complex regulatory effects of intergenic variants.
446 For each independent trait, we identified potential gene candidates within 500kb of each
447 independent significant SNP then took a leave one chromosome out approach (LOCO) to
448 investigate whether genes on the removed chromosome would be implicated in the clusters
449 associated with the remaining genes. **(Figure 4H).** We are able to identify a major proportion
450 of loci independently across all traits and reduce the potential candidates at each locus
451 considerably, further highlighting the utility of UnTANGLeD **(Figure 4I)**.
452

453 **DISCUSSION**

454 This study demonstrates that gene programs governing biological processes can be identified
455 without reliance on prior knowledge, by analysing the association between genetic variation
456 and a large range of diverse complex traits. Several prior studies have constructed small gene
457 networks using a limited number of disease phenotypes and their associated genes from curated
458 GWAS databases and restricted sources of rare genetic variants. Other studies, like PheWAS
459 (34, 35) and PhenomeXcan (36) have collated genomic associations across numerous
460 phenotypes to create resources of variant-trait and gene-trait associations.

461

462 Here, we construct a gene-trait association matrix for 16,849 genes across 1,393 complex traits
463 similarly to PhenomeXcan, and further the concept by using UnTANGLeD to identify gene
464 programs. We apply dimensionality reduction methods, which can harness the high
465 dimensional, complex gene-trait association data, allowing us to greatly expand on the scale of
466 studies previously attempting to build gene networks. By increasing the scale of data, we not
467 only identify gene programs enriched for biological processes specific to associated phenotypes
468 but also reveal gene programs enriched for central processes governing diverse mechanisms of
469 cellular development and homeostasis.

470

471 The UnTANGLeD framework is a powerful approach to identify gene programs orchestrating
472 key biological processes. We implicate novel genes in clusters enriched for known processes
473 and identify numerous novel gene programs with enrichment for protein-protein interactions
474 and no known function. We further highlight the utility of UnTANGLeD for hypothesis
475 generation and functional annotation of genes, which may be particularly valuable for non-
476 coding genes, as they are notoriously difficult to annotate *in silico* (37). Finally, the
477 UnTANGLeD framework reveals relationships between complex traits, linking phenotypes by
478 the gene programs that underpin them.

479

480 We demonstrate the utility of UnTANGLeD for predicting genes associated with complex traits
481 and diseases using a low-powered GWAS of the same trait. Currently, standard methods use
482 gene-set analysis to improve power to identify genes and pathways involved in a phenotype,
483 such as MAGMA, or GIGSEA (38–41). Our method eliminates the need to define gene-sets
484 and instead uses gene-trait association data to learn gene sets governing complex traits (39),
485 enabling us to implicate novel trait associated genes and loci from a much smaller cohort size.

486

487 We further highlight the use of the UnTANGLeD clusters for gene prioritization, showing that
488 they effectively select gene candidates at different loci related to the same phenotype. Current
489 gene prioritisation approaches use either distance-based metrics or mapping to eQTLs to predict
490 changes in gene expression (42). However, these also suffer from a considerable false positive
491 rate and may not always distinguish between two genes in proximity, as noted in our data (42).
492 Some recent methods have integrated biological data, such as gene sets, RNA sequencing and
493 protein-protein interaction databases to further prioritise genes at a locus (43). Our framework
494 can be used independently or integrated with any of these approaches to advance understanding
495 of complex trait biology.
496
497 Outside of its utility in GWAS analyses, UnTANGLeD may provide key mechanisms for data
498 analysis in medical and industry pipelines including genetic testing and drug discovery. For
499 example, polygenic risk scores (PRS) are an emerging method that evaluate an individual's
500 disease risk from genetic variants (44). Methods such as UnTANGLeD may help reveal genes
501 and hence genetic variants governing cell programs underlying disease risk and hence improve
502 prediction accuracy. In the context of pharmacogenomics, studies have shown that drug targets
503 with genetic support from either rare or common diseases are more than twice as likely to pass
504 through clinical trials (45, 46). Since UnTANGLeD captures gene programs associated with all
505 complex traits and diseases, its predictive power may help de-risk candidates and thereby
506 decrease cost associated with the drug discovery pipeline. Overall, UnTANGLeD represents a
507 powerful and versatile framework for studying cellular gene programs to interpret diverse
508 sources of orthogonal genetic data.
509
510 We note several limitations in our method. Primarily, that the current GWAS data does not
511 represent the whole phenome. Furthermore, many traits are highly correlated, and disease traits
512 are underrepresented in the UK Biobank, the main source of data in this study. Secondly,
513 UnTANGLeD relies on S-MultiXcan to construct the gene-trait association matrix. While S-
514 MultiXcan is powered to detect associations across all tissues, it suffers from a high false
515 discovery rate and may perform poorly in tissues with small sample sizes. Moreover, S-
516 MultiXcan can identify genes colocalised with a causal gene as significant, which can obscure
517 biological signatures. Other approaches such as SMR MR-JTI may remedy this issue (47).
518 Additionally, UnTANGLeD does not account for the predicted directionality of effect or tissue-
519 specific effects, which may help to further increase the quality and biological specificity of the
520 clusters. Biological validation of the method using established gene sets may be inflated due to

GWAS data being included in the definition of the gene sets. Finally, we note that although UnTANGLeD is a powerful tool for identifying clusters in an unsupervised manner, the overall function of the cluster may be difficult to determine. The development of improved gene-based tests and emergence of larger GWAS data spanning the whole phenome will improve the accuracy and utility of UnTANGLeD.

This study provides a powerful framework for the identification of gene programs governing biological processes conserved across all complex traits and diseases, with important applications for functional annotation, hypothesis generation, machine learning and prediction algorithms and interpretation of GWAS and diverse other genomic data types. Our approach can be applied to any collection of gene-trait information, harnessing the power of biological patterns in a diverse landscape of phenotypic variation.

DATA AND MATERIALS AVAILABILITY

All source code available on GitHub (https://github.com/palpant-comp/UnTANGLeD) and all data available on Zenodo (https://doi.org/10.5281/zenodo.6572617).

SUPPLEMENTARY MATERIALS

Supplementary Data are available at NAR online.

FUNDING

We thank support from the National Health and Medical Research Council of Australia (Grant 1143163, N.P.) and the Australian Research Council (Grant SR1101002, N.P.) and the National Heart Foundation of Australia (Grant 101889, N.P.).

ACKNOWLEDGEMENTS

We thank Naomi Wray and Jian Yang for assistance in conceptual development of the study.

AUTHOR CONTRIBUTIONS

DM: Developed the study, performed all analyses, and wrote the manuscript
MS and **CN:** Helped supervise bioinformatics analysis
GCP: Helped supervise and design GWAS data selection and analysis, interpreted data, and wrote the manuscript
NP: Conceived and supervised the project, raised funding, and wrote the manuscript

CONFLICT OF INTEREST STATEMENT

GCP is currently an employee of 23andMe Inc. and holds stock options for the company.

567 **REFERENCE**

568 1. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the
569 human genome. *Nature*, **489**, 57–74.
570 2. Rozenblatt-Rosen,O., Stubbington,M.J.T., Regev,A. and Teichmann,S.A. (2017) The
571 Human Cell Atlas: from vision to reality. *Nat. 2017 5507677*, **550**, 451–453.
572 3. Sudlow,C., Gallacher,J., Allen,N., Beral,V., Burton,P., Danesh,J., Downey,P., Elliott,P.,
573 Green,J., Landray,M., *et al.* (2015) UK Biobank: An Open Access Resource for
574 Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age.
575 *PLoS Med.*, **12**, 1–10.
576 4. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O.,
577 Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A., *et al.* (2021) Highly accurate
578 protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
579 5. Shim,W.J., Sinniah,E., Xu,J., Vitrinel,B., Alexanian,M., Andreoletti,G., Shen,S., Sun,Y.,
580 Balderson,B., Boix,C., *et al.* (2020) Conserved Epigenetic Regulatory Logic Infers Genes
581 Governing Cell Identity. *Cell Syst*, **11**, 625-639.e13.
582 6. Visscher,P.M., Wray,N.R., Zhang,Q., Sklar,P., McCarthy,M.I., Brown,M.A. and Yang,J.
583 (2017) 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum.*
584 *Genet.*, **101**, 5–22.
585 7. Bellomo,T.R., Bone,W.P., Chen,B.Y., Gawronski,K.A.B., Zhang,D., Park,J., Levin,M.,
586 Tsao,N., Klarin,D., Lynch,J., *et al.* (2021) Multi-trait GWAS of atherosclerosis detects
587 novel pleiotropic loci. *medRxiv*, 10.1101/2021.05.21.21257493.
588 8. Yu,G., Wang,L.-G., Han,Y., He,Q.-Y. (2012) clusterProfiler: an R Package for Comparing
589 Biological Themes Among Gene Clusters. *Omi. A J. Integr. Biol.*, **16**, 284–287.
590 9. Liberzon,A., Subramanian,A., Pinchback,R., Thorvaldsdóttir,H., Tamayo,P. and
591 Mesirov,J.P. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**,
592 1739.
593 10. Cuellar-Partida,G., Lundberg,M., Kho,P.F., D'Urso,S., Gutierrez-Mondragon,L.F. and
594 Hwang,L.-D. (2019) Complex-Traits Genetics Virtual Lab: A community-driven web
595 platform for post-GWAS analyses. *bioRxiv*, 10.1101/518027.
596 11. Butler,A., Hoffman,P., Smibert,P., Papalexi,E. and Satija,R. (2018) Integrating single-cell
597 transcriptomic data across different conditions, technologies, and species. *Nat.*
598 *Biotechnol.*, **36**, 411–420.

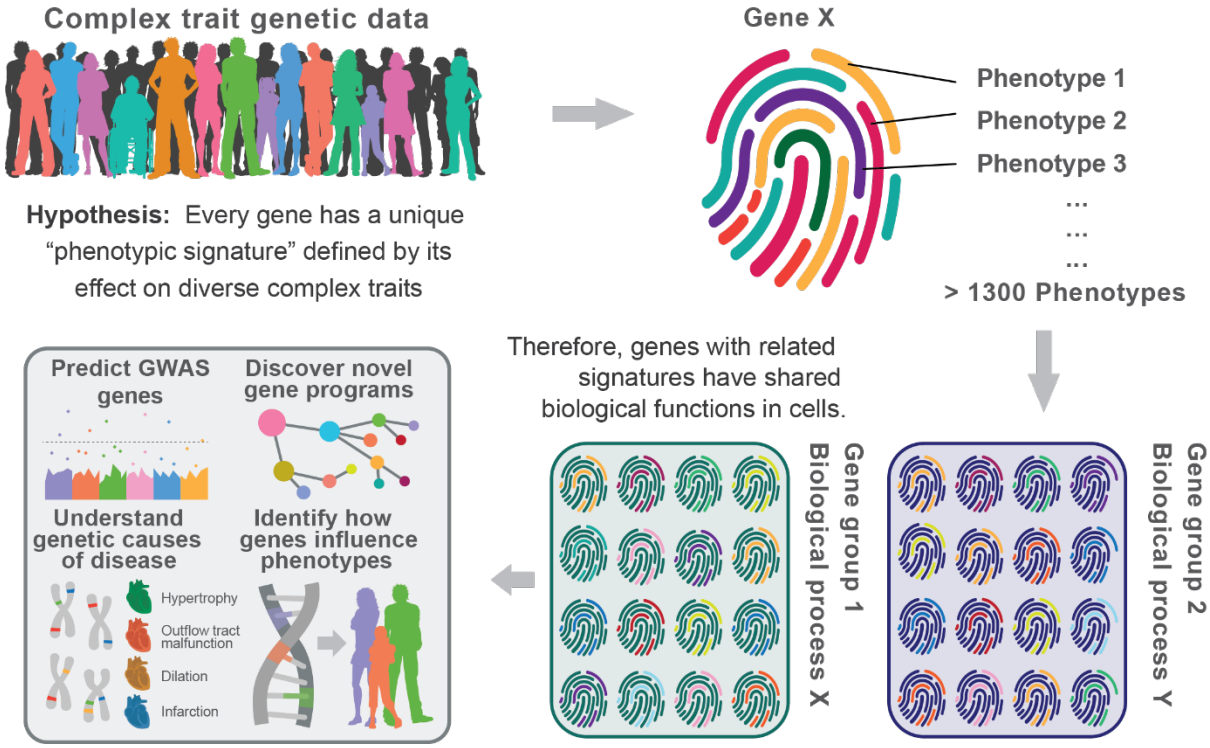
599 12. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P.,
600 Dolinski,K., Dwight,S.S., Eppig,J.T., *et al.* (2000) Gene ontology: tool for the unification
601 of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
602 13. Szklarczyk,D., Gable,A.L., Lyon,D., Junge,A., Wyder,S., Huerta-Cepas,J., Simonovic,M.,
603 Doncheva,N.T., Morris,J.H., Bork,P., *et al.* (2019) STRING v11: protein–protein
604 association networks with increased coverage, supporting functional discovery in genome-
605 wide experimental datasets. *Nucleic Acids Res.*, **47**, D607–D613.
606 14. Schrimi,L.M., Arze,C., Nadendla,S., Wayne Chang,Y.-W., Mazaitis,M., Felix,V., Feng,G.
607 and Kibbe,W.A. (2012) Disease Ontology: a backbone for disease semantic integration.
608 *Nucleic Acids Res.*, **40**.
609 15. Kanehisa,M. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids*
610 *Res.*, **28**, 27–30.
611 16. Jain,A. and Tuteja,G. (2019) TissueEnrich: Tissue-specific gene enrichment analysis.
612 *Bioinformatics*, **35**, 1966–1967.
613 17. Kuleshov,M. V, Jones,M.R., Rouillard,A.D., Fernandez,N.F., Duan,Q., Wang,Z.,
614 Koplev,S., Jenkins,S.L., Jagodnik,K.M., Lachmann,A., *et al.* (2016) Enrichr: a
615 comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.,*
616 **44**, W90–W97.
617 18. El-Gebali,S., Mistry,J., Bateman,A., Eddy,S.R., Luciani,A., Potter,S.C., Qureshi,M.,
618 Richardson,L.J., Salazar,G.A., Smart,A., *et al.* (2019) The Pfam protein families database
619 in 2019. *Nucleic Acids Res.*, **47**, D427–D432.
620 19. de Lange,K.M., Moutsianas,L., Lee,J.C., Lamb,C.A., Luo,Y., Kennedy,N.A., Jostins,L.,
621 Rice,D.L., Gutierrez-Achury,J., Ji,S.G., *et al.* (2017) Genome-wide association study
622 implicates immune activation of multiple integrin genes in inflammatory bowel disease.
623 *Nat Genet*, **49**, 256–261.
624 20. Köttgen,A., Albrecht,E., Teumer,A., Vitart,V., Krumsiek,J., Hundertmark,C., Pistis,G.,
625 Ruggiero,D., O'Seaghdha,C.M., Haller,T., *et al.* (2013) Genome-wide association analyses
626 identify 18 new loci associated with serum urate concentrations. *Nat Genet*, **45**, 145–154.
627 21. Tin,A., Marten,J., Halperin Kuhns,V.L., Li,Y., Wuttke,M., Kirsten,H., Sieber,K.B., Qiu,C.,
628 Gorski,M., Yu,Z., *et al.* (2019) Target genes, variants, tissues and transcriptional pathways
629 influencing human serum urate levels. *Nat. Genet.*, **51**, 1459–1474.
630 22. Shah,S., Henry,A., Roselli,C., Lin,H., Sveinbjörnsson,G., Fatemifar,G., Hedman,Å.K.,
631 Wilk,J.B., Morley,M.P., Chaffin,M.D., *et al.* (2020) Genome-wide association and
632 Mendelian randomisation analysis provide insights into the pathogenesis of heart failure.

633 *Nat. Commun.*, **11**, 163.
634 23. Kathiresan,S., Melander,O., Guiducci,C., Surti,A., Burtt,N.P., Rieder,M.J., Cooper,G.M.,
635 Roos,C., Voight,B.F., Havulinna,A.S., *et al.* (2008) Six new loci associated with blood low-
636 density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in
637 humans. *Nat Genet*, **40**, 189–197.
638 24. Willer,C.J., Schmidt,E.M., Sengupta,S., Peloso,G.M., Gustafsson,S., Kanoni,S., Ganna,A.,
639 Chen,J., Buchkovich,M.L., Mora,S., *et al.* (2013) Discovery and refinement of loci
640 associated with lipid levels. *Nat. Genet.*, **45**, 1274–1283.
641 25. Stahl,E.A., Raychaudhuri,S., Remmers,E.F., Xie,G., Eyre,S., Thomson,B.P., Li,Y.,
642 Kurreeman,F.A., Zhernakova,A., Hinks,A., *et al.* (2010) Genome-wide association study
643 meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat Genet*, **42**, 508– 514.
644 26. Okada,Y., Wu,D., Trynka,G., Raj,T., Terao,C., Ikari,K., Kochi,Y., Ohmura,K., Suzuki,A.,
645 Yoshida,S., *et al.* (2014) Genetics of rheumatoid arthritis contributes to biology and drug
646 discovery. *Nature*, **506**, 376–381.
647 27. Schizophrenia Psychiatric Genome-Wide Association Study (GWAS)Consortium (2011)
648 Genome-wide association study identifies five new schizophrenia loci. *Nat Genet*, **43**, 969–
649 976.
650 28. Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014) Biological
651 insights from 108 schizophrenia-associated genetic loci. *Nature*, **511**, 421– 427.
652 29. Anderson,C.A., Boucher,G., Lees,C.W., Franke,A., D'Amato,M., Taylor,K.D., Lee,J.C.,
653 Goyette,P., Imielinski,M., Latiano,A., *et al.* (2011) Meta-analysis identifies 29 additional
654 ulcerative colitis risk loci, increasing the number of confirmed associations to
655 47. *Nat Genet*, **43**, 246–252.
656 30. Spracklen,C.N., Chen,P., Kim,Y.J., Wang,X., Cai,H., Li,S., Long,J., Wu,Y., Wang,Y.X.,
657 Takeuchi,F., *et al.* (2017) Association analyses of East Asian individuals and transancestry
658 analyses with European individuals reveal new loci associated with cholesterol and
659 triglyceride levels. *Hum Mol Genet*, **26**, 1770–1784.
660 31. Lam,M., Chen,C.-Y., Li,Z., Martin,A.R., Bryois,J., Ma,X., Gaspar,H., Ikeda,M.,
661 Benyamin,B., Brown,B.C., *et al.* (2019) Comparative genetic architectures of
662 schizophrenia in East Asian and European populations. *Nat. Genet.*, **51**, 1670–1678.
663 32. Wang,Y.-F., Zhang,Y., Lin,Z., Zhang,H., Wang,T.-Y., Cao,Y., Morris,D.L., Sheng,Y.,
664 Yin,X., Zhong,S.-L., *et al.* (2021) Identification of 38 novel loci for systemic lupus
665 erythematosus and genetic heterogeneity between ancestral groups. *Nat. Commun.*, **12**, 772.

666 33. Bentham,J., Morris,D.L., Graham,D.S.C., Pinder,C.L., Tombleson,P., Behrens,T.W.,
667 Martín,J., Fairfax,B.P., Knight,J.C., Chen,L., *et al.* (2015) Genetic association analyses
668 implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of
669 systemic lupus erythematosus. *Nat Genet*, **47**, 1457–1464.
670 34. Diogo,D., Tian,C., Franklin,C.S., Alanne-Kinnunen,M., March,M., Spencer,C.C.A.,
671 Vangjeli,C., Weale,M.E., Mattsson,H., Kilpeläinen,E., *et al.* (2018) Phenome-wide
672 association studies across large population cohorts support drug target validation. *Nat.*
673 *Commun. 2018 91*, **9**, 1–13.
674 35. Pendergrass,S.A., Buyske,S., Jeff,J.M., Frase,A., Dudek,S., Bradford,Y., Ambite,J.-L.,
675 Avery,C.L., Buzkova,P., Deelman,E., *et al.* (2019) A phenome-wide association study
676 (PheWAS) in the Population Architecture using Genomics and Epidemiology (PAGE)
677 study reveals potential pleiotropy in African Americans. *PLoS One*, **14**, e0226771.
678 36. Pividori,M., Rajagopal,P.S., Barbeira,A., Liang,Y., Melia,O., Bastarache,L., Park,Y.,
679 Consortium,G., Wen,X. and Im,H.K. (2020) PhenomeXcan: Mapping the genome to the
680 phenome through the transcriptome. *Sci. Adv.*, **6**, eaba2083.
681 37. Perron,U., Provero,P. and Molineris,I. (2017) In silico prediction of lncRNA function using
682 tissue specific and evolutionary conserved expression. *BMC Bioinformatics*, **18**, 29–39.
683 38. Zhu,S., Qian,T., Hoshida,Y., Shen,Y., Yu,J. and Hao,K. (2019) GIGSEA: genotype
684 imputed gene set enrichment analysis using GWAS summary level data. *Bioinformatics*,
685 **35**, 160–163.
686 39. Zhu,X. and Stephens,M. (2018) Large-scale genome-wide enrichment analyses identify
687 new trait-associated genes and pathways across 31 human phenotypes. *Nat. Commun.*
688 (2018) *91*, **9**, 1–14.
689 40. de Leeuw,C.A., Mooij,J.M., Heskes,T. and Posthuma,D. (2015) MAGMA: Generalized
690 Gene-Set Analysis of GWAS Data. *PLOS Comput. Biol.*, **11**, e1004219.
691 41. Sun,R., Hui,S., Bader,G.D., Lin,X. and Kraft,P. (2019) Powerful gene set analysis in
692 GWAS with the Generalized Berk-Jones statistic. *PLoS Genet.*, **15**.
693 42. Broekema,R. V., Bakker,O.B. and Jonkers,I.H. (2020) A practical view of fine-mapping
694 and gene prioritization in the post-genome-wide association era. *Open Biol.*, **10**.
695 43. Schaid,D.J., Chen,W. and Larson,N.B. (2018) From genome-wide associations to
696 candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.*, **19**, 491.
697 44. Lewis,C.M. and Vassos,E. (2020) Polygenic risk scores: From research tools to clinical
698 instruments. *Genome Med.*, **12**, 1–11.

45. Nelson,M.R., Tipney,H., Painter,J.L., Shen,J., Nicoletti,P., Shen,Y., Floratos,A., Sham,P.C., Li,M.J., Wang,J., *et al.* (2015) The support of human genetic evidence for approved drug indications. *Nat. Genet. 2015 478*, **47**, 856–860.
46. King,E.A., Wade Davis,J. and Degner,J.F. (2019) 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.
47. Zhou,D., Jiang,Y., Zhong,X., Cox,N.J., Liu,C. and Gamazon,E.R. (2020) A unified framework for joint-tissue transcriptome-wide association and Mendelian randomization analysis. *Nat. Genet.*, **52**, 1239–1246.

709 **FIGURES**



710

711 **Figure 1. Schematic of central model design.** Complex trait genetic data provide a unique
712 association signature for each gene which can be used to parse the genome into functionally
713 related gene sets.

714

715

716

717

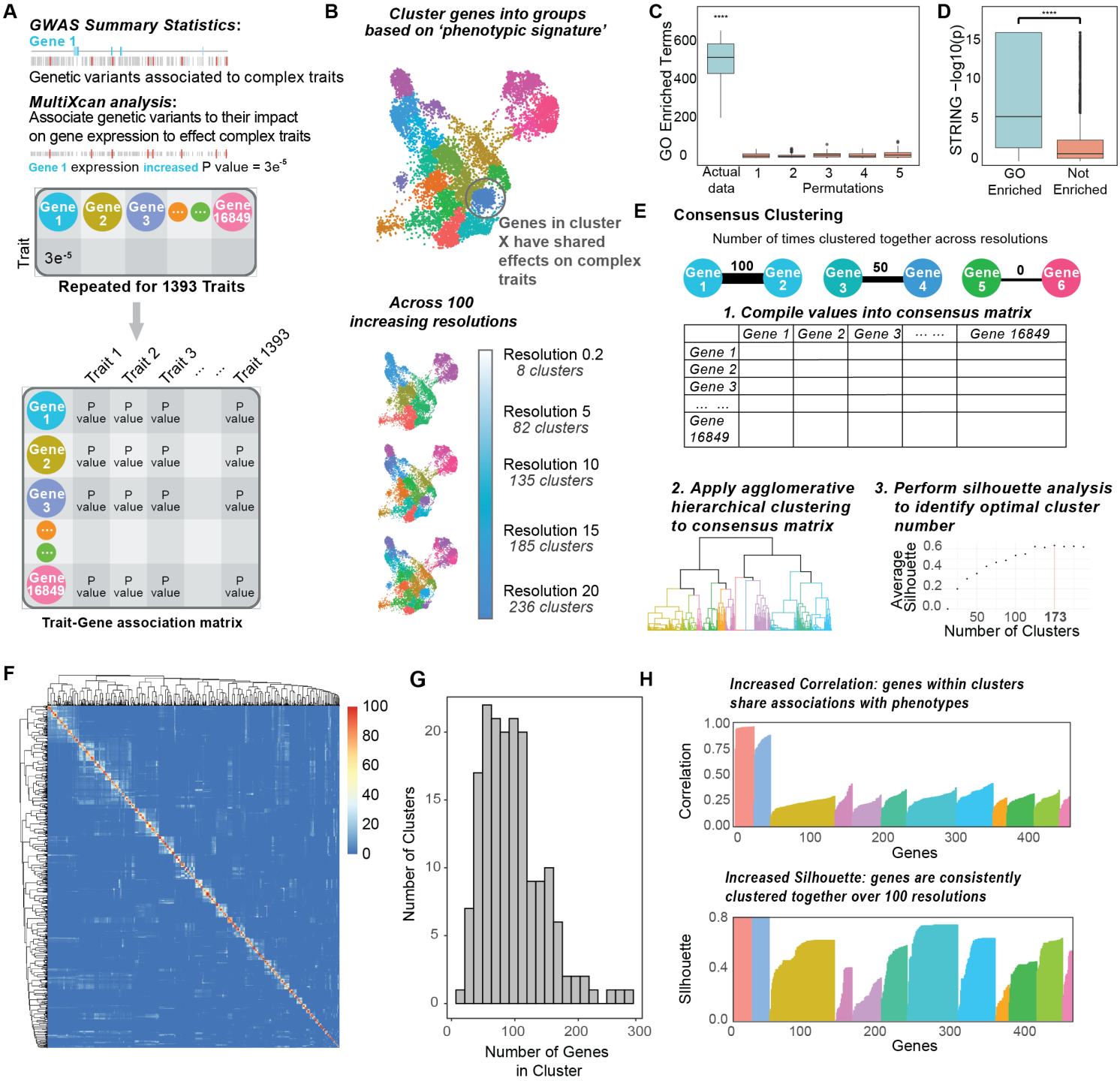
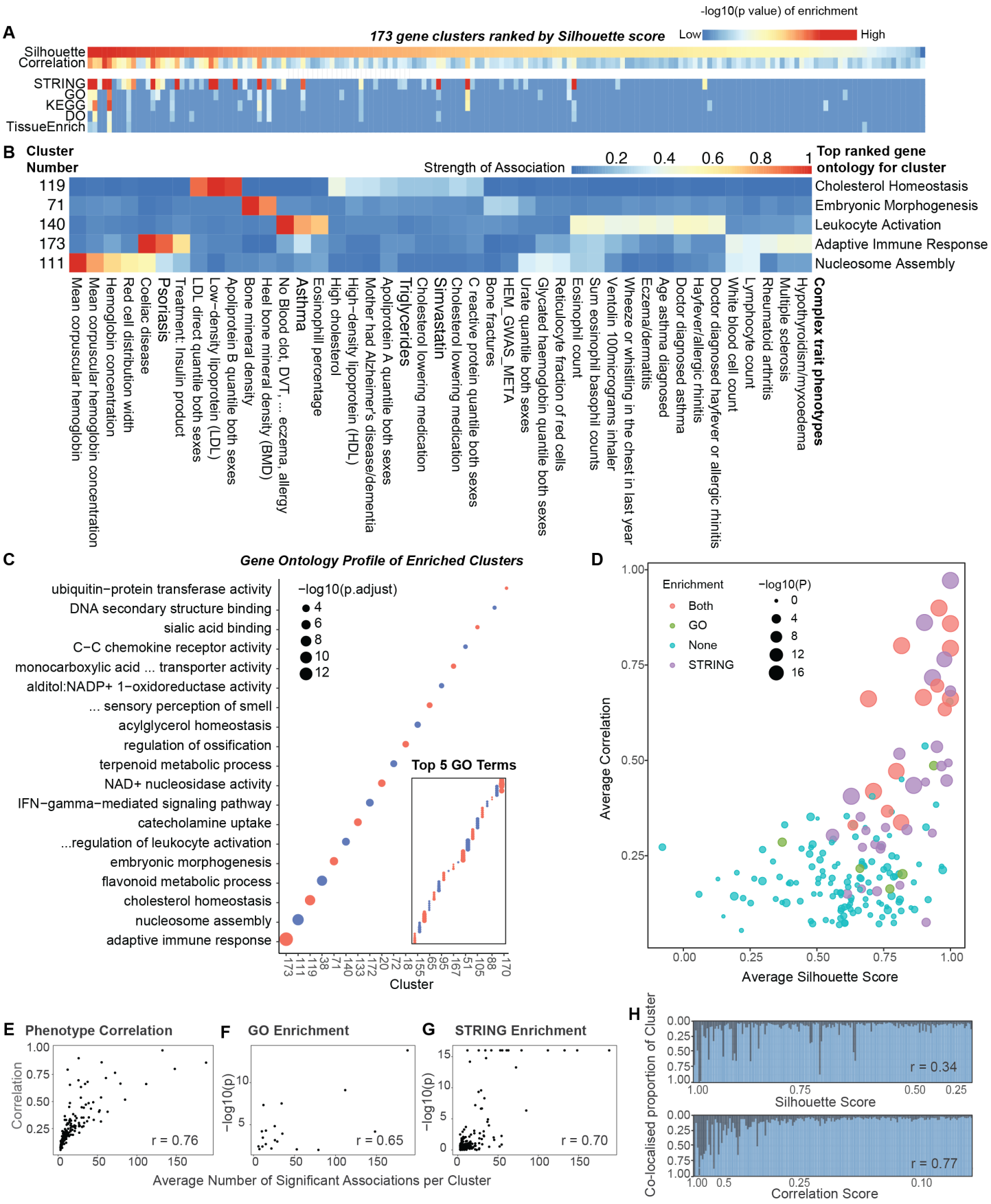


Figure 2. Consensus clustering method identifies biologically enriched gene clusters.

(A) MultiXcan analysis links genetic variants to genes by predicting changes in gene expression using eQTLs. The chi-squared values of the associations between each of 1393 traits and 25851 genes were compiled into a gene-trait matrix.

(B) Dimensionality reduction clustering of genes based on their phenotypic associations was performed using *Seurat* across resolutions 0.2 to 20 in 0.2 increments.

725 **(C)** Five permutations of the dataset were compared to the real data by the number of enriched
726 gene ontology terms per resolution. Enrichment was performed using *ClusterProfiler,* FDR
727 corrected p-value < 0.01. Pairwise comparisons between permutations were performed
728 with Wilcoxon signed rank test.
729 **(D)** Validation of gene ontology enriched clusters with STRING protein-protein interaction
730 enrichment. Wilcoxon signed rank test was used to compare STRING enrichment in gene
731 ontology enriched and non-enriched clusters.
732 **(E)** Each gene pair is given a similarity score based on how often they were clustered together
733 across 100 resolutions and these values are compiled into a consensus matrix.
734 Agglomerative hierarchical clustering is applied to the matrix, with the plateau in the
735 average silhouette score defining the optimal number of clusters.
736 **(F)** Heatmap of consensus matrix as clustered using agglomerative hierarchical clustering for
737 173 clusters.
738 **(G)** Histogram of the number of genes in each of the 173 clusters.
739 **(H)** Silhouette scores and correlation scores calculated for each gene to evaluate the clustering
740 robustness and quality respectively. Data generated for 450 genes selected from 12 random
741 clusters.


A
173 gene clusters ranked by Silhouette score
-log10(p value) of enrichment
Low
High
Silhouette
Correlation
STRING
GO
KEGG
DO
TissueEnrich
B
Cluster Number
Strength of Association
0.2
0.4
0.6
0.8
1
Top ranked gene ontology for cluster
119
71
140
173
111
Cholesterol Homeostasis
Embryonic Morphogenesis
Leukocyte Activation
Adaptive Immune Response
Nucleosome Assembly
Complex trait phenotypes
Mean corpuscular hemoglobin
Mean corpuscular hemoglobin concentration
Hemoglobin concentration
Red cell distribution width
Coeliac disease
Psoriasis
Treatment: Insulin product
LDL direct quantile both sexes
Low-density lipoprotein (LDL)
Apolipoprotein B quantile both sexes
Bone mineral density
Heel bone mineral density (BMD)
No Blood clot, DVT, ... eczema, allergy
Asthma
Eosinophill percentage
High cholesterol
High-density lipoprotein (HDL)
Mother had Alzheimer's disease/dementia
Apolipoprotein A quantile both sexes
Triglycerides
Cholesterol lowering medication
Simvastatin
Cholesterol lowering medication
C reactive protein quantile both sexes
Bone fractures
HEM_GWAS_META
Urate quantile both sexes
Glycated haemoglobin quantile both sexes
Reticulocyte fraction of red cells
Eosinophil count
Sum eosinophil basophil counts
Ventolin 100micrograms inhaler
Wheeze or whistling in the chest in last year
Eczema/dermatitis
Age asthma diagnosed
Doctor diagnosed asthma
Hayfever/allergic rhinitis
Doctor diagnosed hayfever or allergic rhinitis
White blood cell count
Lymphocyte count
Rheumatoid arthritis
Multiple sclerosis
Hypothyroidism/myxoedema
C
Gene Ontology Profile of Enriched Clusters
ubiquitin-protein transferase activity
DNA secondary structure binding
sialic acid binding
C-C chemokine receptor activity
monocarboxylic acid ... transporter activity
alditol:NADP+ 1-oxidoreductase activity
... sensory perception of smell
acylglycerol homeostasis
regulation of ossification
terpenoid metabolic process
NAD+ nucleosidase activity
IFN-gamma-mediated signaling pathway
catecholamine uptake
...regulation of leukocyte activation
embryonic morphogenesis
flavonoid metabolic process
cholesterol homeostasis
nucleosome assembly
adaptive immune response
-log10(p.adjust)
4
6
8
10
12
Top 5 GO Terms
Cluster
D
Enrichment
Both
GO
None
STRING
-log10(P)
0
4
8
12
16
Average Correlation
Average Silhouette Score
E
Phenotype Correlation
Correlation
r = 0.76
F
GO Enrichment
-log10(p)
r = 0.65
G
STRING Enrichment
-log10(p)
r = 0.70
Average Number of Significant Associations per Cluster
H
Co-localised proportion of Cluster
r = 0.34
Silhouette Score
r = 0.77
Correlation Score

742

743 **Figure 3. Enrichment of identified clusters for known gene sets is dependent on data**
744 **quality.**

745 **(A)** Broad enrichment profile of 173 clusters stratified by average silhouette and correlation
746 scores.
747 **(B)** Heatmap showing the relationship between the biological profile of five clusters, as
748 proxied by their top gene ontology term, and the unique phenotypic signature. The top 10
749 phenotypes per cluster were selected. Association strength was calculated using negative log
750 transformed significance values, which were then normalised across phenotypes and then per
751 cluster. **(C)** Top enriched gene ontology biological processes for each cluster have no overlap.
752 Clusters ranked by strength of top enriched term. Specificity of top 5 terms per cluster is also
753 provided. **(D)** 173 clusters stratified by their average correlation and silhouette scores with
754 their gene ontology and STRING enrichment indicated. The P-value represented is specific to
755 the enriched category, and in the case of both represents the more significant of the two.
756 **(E-G)** Correlation between the average number of significant gene-trait associations per cluster
757 and **E)** the average correlation score of each cluster **(F)**, strength of GO enrichment and strength
758 of STRING enrichment per cluster (Pearson's correlation) **(G)**.
759 **(H)** Presence and degree of colocalization within clusters as stratified by their silhouette and
760 correlation scores (Pearson's correlation).
761
762

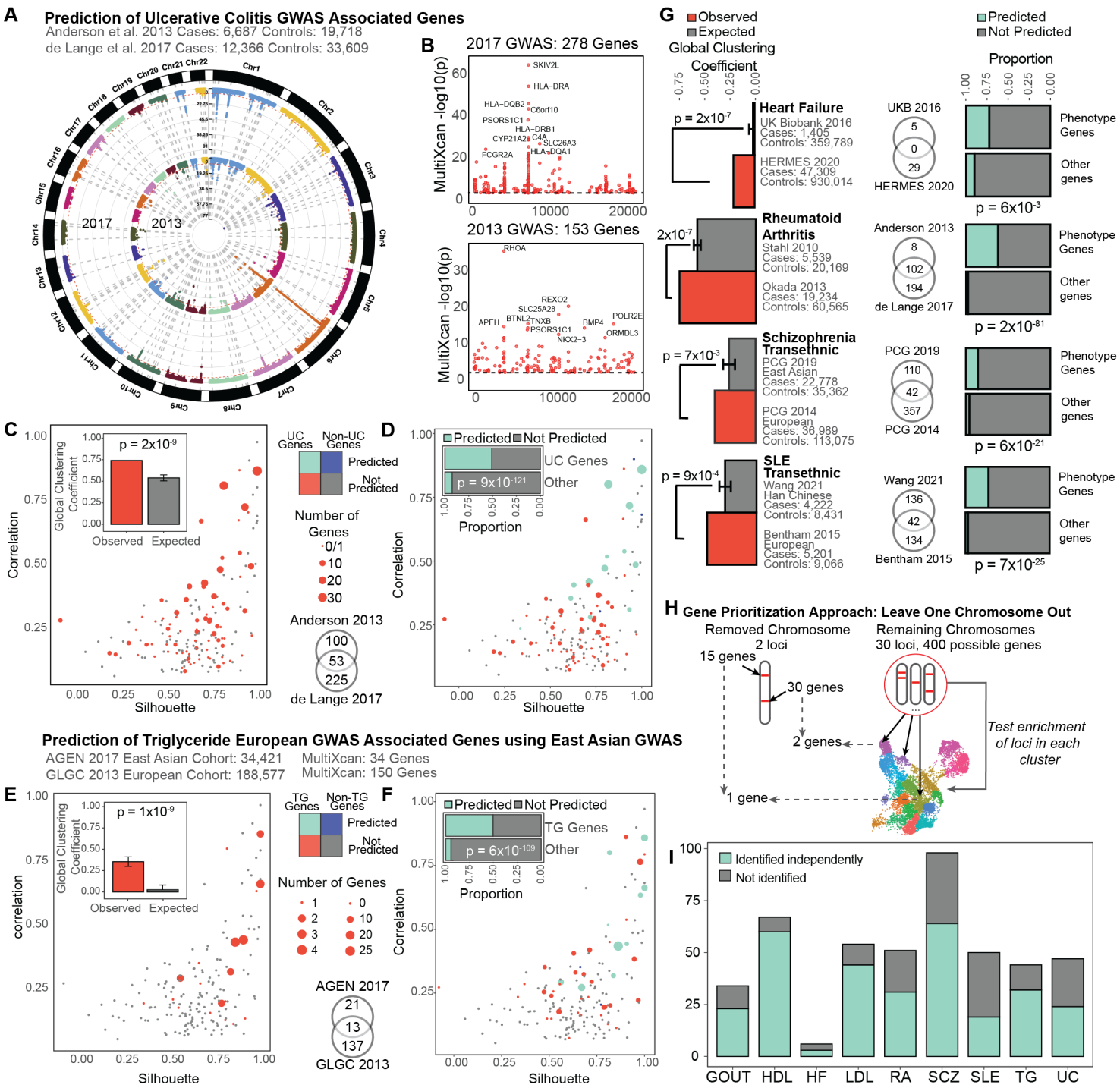


Figure 4. Identified clusters are conserved across all phenotypes and can be used for prediction of genes involved in complex trait biology and prioritization of GWAS genes at implicated loci.

(A) Manhattan plot of loci identified by a 2017 and 2013 GWAS of ulcerative colitis (UC). **(B)** Manhattan plot of S-MultiXcan genes for the two GWAS respectively, genes are ordered according to their genomic positions.

770 **(C)** Distribution of 278 significant genes from the 2017 UC GWAS across 173 clusters.
771 Global clustering coefficient was calculated for the 278 genes. Significance was calculated
772 using 100 bootstrap replicates to establish a distribution from which a Z score was calculated.
773 **(D)** Prediction of 2017 UC GWAS genes using 2013 UC GWAS. Chi-squared enrichment
774 test was used to determine enrichment for prediction of novel genes compared to non-trait
775 associated genes.
776 **(E)** Distribution and global clustering coefficient of 34 significant genes from East Asian
777 GWAS of Triglyceride levels. Significance was calculated using bootstrapping.
778 **(F)** Prediction of 137 novel genes from 2013 European GWAS of triglycerides using 2017
779 East Asian GWAS. Enrichment was calculated using chi-squared test.
780 **(G)** Increase in observed global clustering coefficient compared to expected and prediction
781 enrichment across four additional traits.
782 **(H)** Schematic of gene prioritization strategy using the leave one chromosome out approach.
783 Potential genes at a significant locus were refined using clusters enriched for the trait.
784 **(I)** The proportion of loci at which a gene was successfully identified independently of all
785 genes on the same chromosome.
786