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 242 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

93 and transethnic GWAS data. UnTANGLeD provides a powerful analytical framework for
94 studies in population genetics, cell biology, and genomics, that will improve as more data
95 emerges.
96 Collectively, this study provides a statistical framework for defining genes orchestrating
97 biological processes by evaluating genetic signatures across diverse complex traits.
98
99 **MATERIALS AND METHODS**
100 ***Data Collection***
101 S-MultiXcan results for 1,393 phenotypes with statistically significant SNP-based heritability
102 ($p < 0.05$) were downloaded from CTG-VL (http://vl.genoma.io). Phenotypes are listed in
103 **Table S2**. SNP-based heritability was estimated using linkage disequilibrium score regression
104 (LDSR). The significance values reflecting the strength of the association between each gene
105 and trait across all tissues were compiled into a gene-trait association matrix.
106
107 ***Dimensionality Reduction Analysis Pipeline***
108 All genes with fewer than 2 significant associations across all phenotypes ($p < 10^{-4}$) were
109 removed, leaving 16849 genes. Following this, all values in the gene-trait association matrix
110 were chi-squared transformed. Infinite values produced when transforming very small p-value
111 (<1e-300) due to floating point precision were replaced with 1,415, which was 5 greater than
112 the largest non-infinite value. The data was then normalised by the sum of chi-squared values
113 per phenotype and scaled by a factor of 10,000. 10 principal components were retained from
114 the principal component analysis (PCA). Clustering of genes was performed using the native
115 *Seurat* shared-nearest neighbour algorithm. Clustering iterations were performed at increasing
116 resolutions from 0.2 to 20 in increments of 0.2. The resolution is a parameter from Seurat
117 where increased values lead to a greater number of clusters. Cluster assignments were
118 compiled into a consensus distance matrix, where each gene pair had a value representing
119 how often they were grouped together out of 100 potential matches. 100 was then subtracted
120 from the values and they were made absolute to transform the matrix into a dissimilarity
121 matrix. Agglomerative clustering using Ward's minimum variance method, as implemented
122 in the *stats* package, was applied to the consensus matrix directly. The average silhouette
123 score (a metric used to calculate how well a data point relates to its cluster) across all genes
124 was calculated using the *cluster* package from 2 to 300 clusters. The *inflection* package was
125 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.

Enrichment Analyses

GO, DO, KEGG enrichment, colocalization and tissue specificity enrichment were performed using *ClusterProfiler* (8). An FDR corrected significance value of $p < 0.01$ was used. Colocalization was determined using *ClusterProfiler* enrichment for the Molecular Signatures Database collection 3: positional gene sets (9). The largest proportion of genes within a cluster belonging to a single genomic region was divided by the total number of genes within the cluster to identify the maximum degree of colocalization. STRING enrichment analysis was performed using the *STRINGdb* package, with a significance threshold of $p < 0.001$ and a confidence threshold of 0.400. STRING enrichment analysis without the text-mining component was performed using the online STRING interface (https://string-db.org/) for clusters found to have PPI enrichment in the prior analysis with a confidence threshold of 0.150 to preserve predicted interactions reinforced by other components. Broad enrichment analysis for more specialised gene sets was performed using *EnrichR* (https://maayanlab.cloud/Enrichr/) across all 192 libraries. Redundant libraries, including GO, KEGG, chromosomal location, WikiPathways and NIH-grant associated libraries were excluded. The top 5 significant terms from each library for each cluster are reported in **Table S9**. A significance value threshold of 0.01, after correction for multiple testing, was used. For identification of genes possessing the same protein domains or belonging to the same family, the EnrichR library 'Pfam_Domains_2019' was used. A distinct protein family or domain was defined by collating the family or domain terms together that shared genes until there was no overlap between them. Protein terms did not need to be significantly enriched, but two or more members of a protein family had to be present in a single cluster.

Permutations

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

159 ***Phenotype Associations***

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

165

166 ***Clustering quality in dimensionality reduction methods***

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

172

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

174 *Data collection and S-MultiXcan Analysis*

175 We selected 13 phenotypes for which GWAS studies had been performed at differing cohort
176 sizes or ethnicities for the same, or comparable traits. The specific studies and their respective
177 details can be found in **Table S1.** Summary statistics were downloaded from various sources
178 and harmonised using MetaXcan's in-built harmonization
179 (https://github.com/hakyimlab/MetaXcan) to be compatible with the MASHR models. We
180 then performed S-MultiXcan analysis of each trait using the MASHR models built off the V8
181 GTEx release. Associated genes were defined as those found to have a significance of
182 $p < 10^{-4}$ by S-MultiXcan.

183

184 *Global Clustering Coefficient Calculation*

185 The genes identified for an independent GWAS were projected onto the 242 identified
186 clusters. Following this, we generated an unweighted adjacency matrix in which genes in the
187 same cluster were represented by a 1, and genes in different clusters by a 0. A comparison
188 between the same gene was represented by 0. Finally, the global clustering coefficient (GCC)
189 for the genes was calculated. To derive a statistical significance, we randomly sampled the
190 same number of genes as there were significant genes for the phenotype and calculated the
191 GCC one-hundred times. A Z score was calculated from the curve generated by the sampled
192 values.

Gene Prediction

We took a simple approach of predicting which clusters were associated with the trait using the S-MultiXcan associations from the smaller GWAS and then checking whether novel gene associations identified by the larger GWAS were in those clusters. A chi-squared enrichment test was used where the minimum expected frequency was greater than 5, and a fisher's test if not. Several approaches to predict clusters associated with the trait were trialled. The first was to identify any of the 242 clusters with a significant gene in it. The second was to integrate the additional phenotype into the trait-gene association matrix. Next, clusters were identified which had an overall significance signature > 1.5 times the average or were significantly ($p < 0.05$) higher than the average signature. Different values were tested for these thresholds, with these providing the best performance. The third approach was to predict associated clusters from the previously established 242 clusters using the thresholds taken in approach two. A one-way ANOVA was performed with pairwise comparisons to determine the best approach. Approach three was the most effective, albeit not significantly, while maintaining a low computational burden. In instances where transethnic GWAS were compared, the East-Asian GWAS was used to predict the trait relevant clusters, and the European GWAS was used as the test set.

Gene Prioritization Analysis

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

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

227

228 ***Normalisation***

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

235

236 ***Phenotype Filtering Based on Euclidean Distance***

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

244

245 ***Phenotype Subsampling and Sensitivity Analyses***

246 Phenotype subsampling was performed on two datasets; MultiXcan results for 1393
247 phenotypes across 16,849 genes generated in this paper, and another dataset containing
248 MultiXcan results for 4091 phenotypes across 15,734 genes (phenomexcan.org). For the data
249 containing 1393 phenotypes, subsampling was performed randomly without replacement from
250 50 to 1393 phenotypes in 20 equal increments across 5 replicates for each number of traits.
251 The full UnTANGLeD clustering pipeline was applied to each subsampled matrix, using a
252 cut-off of 242 clusters universally. Adjusted rand index (ARI) was calculated for each of the
253 subsampled clustering configurations compared to the full dataset. This analysis was repeated
254 for the data containing 4091 phenotypes; however, subsampling was performed from 50
255 phenotypes to 4091 phenotypes in 50 equal increments.

256

257 ***Cluster Conservation***

258 To explore the cause for the marked increase in ARI between 1322 phenotypes and 1393
259 phenotypes, cluster conservation was calculated between them. For each cluster from 1393
260 phenotypes, the proportion of genes that remained grouped together in each of the clusters
261 from 1322 phenotypes was calculated. That proportion was used to assign a conservation
262 score to each gene, depending on how large the proportion of cluster the specific gene
263 remained with was. The same approach was applied between 4091 phenotypes and 4009
264 phenotypes.
265
266 **RESULTS**
267
268 **Unsupervised identification of gene groups with shared complex trait associations**
269 We used MultiXcan results from CTG-VL (10) derived from publicly available GWAS
270 (primarily from UK Biobank, on ~400,000 individuals) to create a gene-trait association
271 matrix for 16,849 genes and 1,393 traits **(Figure 1, Figure S1, Table S2)**. For each gene trait
272 pair, MultiXcan predicts whether trait-associated variants alter the gene's expression. The
273 chi-squared transformed significance value for each gene-trait association pair was compiled
274 into the gene-trait association matrix **(Figure 2A)**. These values were normalised using
275 relative count normalisation to account for the difference in power between phenotypes.
276 Performance was not significantly different using other normalisation methods including
277 centralised log ratio or log normalised data **(Figure S2)**. The data was then clustered using
278 *Seurat,* a dimensionality reduction method commonly used to analyse single cell RNA
279 sequencing data to cluster cells into related groups (11). Here, we use *Seurat* to test whether
280 the calculated gene-trait associations could be simplified into biologically enriched gene
281 clusters. Clustering was performed across 100 stepwise increases in resolution, a parameter
282 which increases the number of gene clusters. Repeat iterations provided an opportunity to
283 survey both the broad scope of biological processes that could be identified, as well as the
284 specificity that could be achieved with each biological process **(Figure 2B)**.
285
286 To test the biological validity of the derived clusters, we used positive gene sets as defined by
287 gene ontology (GO) (12) and STRING (13) to show that gene clusters have significant
288 enrichment for GO biological processes and STRING protein-protein interactions **(Figure 2C**
289 **and 2D)**. To demonstrate that the observed enrichment is driven by distinct gene-trait
290 association signatures rather than chance, we performed permutation analyses in which the
291 values in the data matrix were randomly re-ordered. Permutations had significantly fewer

292 enriched clusters, GO terms and a lower strongest significance compared to the real data ($p <$
293 $4x10^{-27}$) **(Figure 2C, Figure S3A-B)**. Furthermore, we validated that GO enriched clusters
294 were more likely to also have enrichment for protein-protein interactions, suggesting the
295 enrichment is robust **(Figure 2D)**. This analysis revealed that genes possessing similar
296 associations to complex trait phenotypes cluster meaningfully into biologically enriched
297 groups and the enrichment is not stochastic.
298
299 We next developed an ensemble learning method we call "consensus clustering" that
300 incorporates a measure of clustering robustness and quality. Across each of 100 stepwise
301 increases in clustering resolution we evaluated the robustness of clustering by assessing how
302 often every possible gene combination was clustered together ranging from 100 (always) to 0
303 (never) and compiled these values into a consensus matrix **(Figure 2E)**. Following this, we
304 performed agglomerative hierarchical clustering, evaluating the average silhouette score at
305 each possible number of clusters. The silhouette score quantifies how consistent genes within
306 the same cluster are across *Seurat* resolutions. To derive the optimal number of gene clusters,
307 we calculated the plateau point of the average silhouette score, which informs the number of
308 clusters at which point further splitting no longer improves the stability of clustering
309 assignments **(Figure S3C)**. Applying this methodology to gene-trait associations for 16,849
310 genes, we identified 242 clusters with an average of 70 genes **(Figure 2F, G).** Across each
311 cluster, we measured the silhouette score, a metric of cluster robustness and the correlation
312 score, a metric of relation across phenotypes, thereby providing two metrics to quantify the
313 quality of clustering **(Figure 2H)**. Collectively, we call this approach UnTANGLeD:
314 Unsupervised Trait Analysis of Networks from Gene Level Data.
315
316 **Consensus clustering identifies robust gene groups enriched for known gene sets**
317 We analysed each cluster by reference to curated annotations of gene programs (GO, disease
318 ontology (14)), signalling pathways (KEGG (15)), protein-protein interactions (STRING),
319 and tissue specificity (16) to evaluate the ability of UnTANGLeD to identify distinct,
320 biologically established gene programs in an unsupervised manner **(Figure 3A, Figure S3D,**
321 **Tables S3-8)**. This analysis revealed significant enrichment of cell biological pathways and
322 networks across gene clusters, with stronger enrichment among clusters with higher silhouette
323 and correlation scores **(Figure 3A)**. We further performed enrichment analysis of the
324 UnTANGLeD clusters using the EnrichR database (17) **(Figure S4A-C, Table S9),** finding
325 considerable enrichment for disease-associated genes, gene-expression perturbations

326 associated with disease states or drugs and protein domains and families. We note that
327 although many clusters contain multiple members of a protein family (18) the proportion of
328 any one protein family in the cluster is minor **(Figure S4D, Table S10)**.
329
330 We next investigated the relationship between individual gene clusters and the traits most
331 strongly influencing the genes within the clusters, using enriched GO processes as a proxy for
332 the functional profile of a cluster **(Figure 3B).** Each cluster is defined by a distinct gene-trait
333 association 'signature' indicated by the variation and strength of association across 1,393
334 diverse complex traits. In some instances, the enriched biological processes for certain gene
335 clusters are clearly related to the cluster's most significantly associated complex trait
336 phenotypes (e.g., cluster 135: GO enrichment: Lipid Homeostasis; Dominant complex trait
337 phenotypes: Apolipoprotein B quantile, Lipoprotein A quantile).
338
339 Since UnTANGLeD draws on associations across diverse phenotypes to inform gene-gene
340 relationships, the method can identify gene groups with enriched functions that are apparently
341 biologically independent of the phenotypes most significantly associated with the genes in the
342 cluster. For example, cluster 80, enriched for ear morphogenesis (GO:0042471), is most
343 significantly associated to the phenotype Bone Mineral Density and cluster 122, enriched for
344 nucleosome organisation (GO:0034728), is most significantly associated to the phenotype
345 Malabsorption/Celiac Disease. These results support the central hypothesis that genes with
346 shared effects across diverse phenotypes can be clustered into gene groups controlling shared
347 biological functions and processes in an unsupervised manner **(Figure 3B)**.
348
349 Importantly, we show that the GO enriched gene clusters show no overlap in their strongest
350 enriched biological functions, and almost no overlap in their top 5 enriched terms,
351 demonstrating the use of gene-trait association data to parse novel biological gene programs
352 encoded within the genome **(Figure 3C)**.
353
354 Stratifying clusters by their silhouette and correlation scores reveals a higher level of GO,
355 STRING, KEGG, DO and tissue specificity enrichment with higher clustering quality,
356 indicating that the metrics provide an accurate representation of cluster quality **(Figure 3A,**
357 **D)**. Furthermore, both the robustness of clustering and the presence and strength of GO and
358 STRING enrichment are correlated with the number of significant associations to phenotypes

359 per gene (Pearson's correlation, r > 0.71), as well as the stability of clustering (Pearson's
360 correlation, r > 0.69) (**Figure 3E-G, Figure S5A-F**).
361
362 Lastly, we note that there is considerable colocalization of genes within clusters, with a
363 stronger relationship between the correlation score and the degree of colocalization for the
364 genes in a cluster (Pearson's correlation, r = 0.77), than the cluster robustness (Pearson's
365 correlation, r = 0.43) **(Figure 3H).** STRING enrichment may also be inflated due to the text-
366 mining component, as findings from GWAS may be incorporated into the database, with
367 genes in proximity often being reported together. Indeed, we find that the loss of enrichment
368 due to removal of the text-mining component is correlated with the colocalization of the
369 cluster (r = 0.57) **(Figure S6A-B)**. However, clusters with a high degree of colocalization are
370 not necessarily artefacts of false-positive associations identified by MultiXcan. For example,
371 clusters 242 and 122 are strongly enriched for immune processes and chromatin organisation
372 respectively, despite being highly colocalised **(Figure S6C-D).**
373
374 **Subsampling reveals need for more data to improve accuracy of UnTANGLeD**
375 We next sought to determine how the number and diversity of phenotypes influences the
376 accuracy and utility of UnTANGLeD clusters. We show that the number of GO enriched
377 clusters is highly correlated with the number of phenotypes utilised in the analysis (Pearson's
378 correlation, r = 0.85), even when phenotypic diversity is preserved **(Figure S7A)**. To further
379 test this, we performed phenotype subsampling and evaluated clustering accuracy using an
380 adjusted rand index (ARI) analysis. We found that clustering accuracy compared to the full
381 data improved with the addition of more phenotypes, however a marked increase in ARI
382 between 1322 and the full data set suggests that inaccuracy in clustering that isn't determined
383 by phenotypic diversity can be attributed to genes which have weak signatures and few
384 significant associations **(Figure S7B)**. We repeated subsampling in a larger dataset containing
385 MultiXcan analysis of 4091 phenotypes retrieved from Pividori *et al*. (2020) which resulted
386 in the same outcome **(Figure S7C)**. Comparison of the two data sets revealed that genes
387 already having many significant associations simply had more associations in the larger
388 dataset with both datasets possessing an equal proportion of genes with few to no significant
389 associations **(Figure S7D)**. Further, that genes with higher numbers of significant
390 associations have higher degrees of conservation **(Figure S7E-F)**. It's likely that the effective
391 number of traits is similar between the two datasets, as both mostly draw on the UK Biobank

392 and have many highly correlated phenotypes. Further, that genes with higher numbers of
393 significant associations have higher degrees of conservation **(Figure S7E-F)**.
394
395 Cumulatively, these findings indicate that the quality of gene clustering is dependent on the
396 scale and quality of data needed to derive high silhouette and correlation scores as a basis for
397 efficient enrichment of functional gene clusters. Accordingly, as more data becomes
398 available, the quality and accuracy of UnTANGLeD will improve. However, simply
399 increasing the number of phenotypes leads to an increase in redundant associations, and
400 therefore strategies to increase the number of significant gene-trait associations across the
401 genome should be employed, such as diversifying phenotypes and increasing cohort size.
402
403 **UnTANGLeD clusters are conserved across traits and can predict novel trait associated**
404 **genes**
405 GWAS require collections of large cohorts comprising thousands of individual-level genotype
406 data to characterise the genetic architecture of a trait. Furthermore, collecting enough samples
407 can prove challenging for many diseases, and as such they are often underrepresented in
408 biobanks.
409
410 We hypothesised that UnTANGLeD gene clusters would be conserved across complex traits.
411 To test this, we investigated an independent GWAS of ulcerative colitis (UC) (19) **(Figure**
412 **4A).** We show that the 278 genes associated with UC ($p < 10^{-4}$) **(Figure 4B)** were
413 significantly more clustered within the UnTANGLeD clusters than expected by chance ($p =$
414 $8x10^{-19}$) **(Figure 4C)**. The result shows that despite not being used to construct the clusters,
415 UC associated genes nevertheless group within the UnTANGLeD clusters, demonstrating that
416 the defined gene programs are conserved. We replicate our findings in 7 additional
417 independent GWAS phenotypes, highlighting that the UnTANGLeD clusters are conserved
418 across a broad phenotypic space (3, 20–28) (**Figure 4G)**.
419
420 We next tested whether the gene clusters can be used to predict novel genes and cellular
421 processes underpinning independent complex trait data. To test this hypothesis, we examined
422 two GWAS for UC. The first was performed in 2013 with 6,687 cases and 19,718 controls
423 (29), and the latter in 2017 with 12,366 cases and 33,609 controls (19) **(Figure 4A)**.
424 MultiXcan analysis of the summary statistics identified 153 and 278 genes respectively, with
425 an overlap of 53 genes **(Figure 4B)**. We projected the MultiXcan associations for the 2013

426 GWAS onto the 242 clusters, identifying clusters were statistically associated with UC
427 **(Figure S8)**. Finally, we tested whether the clusters predicted from the 2013 GWAS
428 contained novel genes identified by the 2017 GWAS. Of the 225 novel genes identified by
429 the 2017 GWAS, our approach was able to use the 2013 GWAS to predict 118 with a
430 significant enrichment for predicting UC associated genes compared to other genes ($p < 3x10^{-}$
431 ^{95}, chi-squared test) **(Figure 4D)**.
432
433 GWAS of the same complex trait conducted in populations of differing ancestries may
434 implicate both shared and distinct loci. We tested whether UnTANGLeD clusters are
435 conserved for genes specific to non-European ancestries, given that the UnTANGLeD gene
436 clusters are built from a European cohort. To test this, we examined a GWAS for triglyceride
437 levels in an East Asian population, which identified 34 genes (30) **(Figure S9A-B)**. Mirroring
438 our findings in a GWAS conducted on a European population, we found that the genes
439 associated with triglyceride levels in an East Asian population are significantly more
440 clustered than expected ($p = 3x10^{-20}$) and replicate this finding in 4 other GWAS conducted in
441 populations of non-European ancestry (30–32). We further tested whether the GWAS
442 conducted in the East Asian cohort could be used to predict novel genes identified in a
443 European cohort. We found that clusters implicated in triglyceride levels using the East Asian
444 GWAS were highly enriched for genes identified by the European GWAS ($p = 1x10^{-123}$)
445 **(Figure 4F)**.
446
447 All together, we show significant enrichment for prediction of novel genes across GWAS
448 performed for 7 traits in differing cohort sizes in a European population, and 4 traits for
449 which GWAS were performed in different ancestries (3, 20-28, 30–33) **(Figure 4G, Figure**
450 **S9C)**.
451
452 We further tested whether the UnTANGLeD clusters could be used to prioritize causal genes
453 at any given locus. It is difficult to accurately identify the causal genes from GWAS
454 identified variants due to linkage disequilibrium and complex regulatory effects of intergenic
455 variants. For each independent trait, we identified potential gene candidates within 500kb of
456 each independent significant SNP then took a leave one chromosome out approach (LOCO)
457 to investigate whether genes on the removed chromosome would be implicated in the clusters
458 associated with the remaining genes. **(Figure 4H).** We are able to identify a major proportion

459 of loci independently across all traits and reduce the potential candidates at each locus
460 considerably, further highlighting the utility of UnTANGLeD **(Figure 4I)**.
461

462 **DISCUSSION**

463 This study demonstrates that gene programs governing biological processes can be identified
464 without reliance on prior knowledge, by analysing the association between genetic variation
465 and a large range of diverse complex traits. Several prior studies have constructed small gene
466 networks using a limited number of disease phenotypes and their associated genes from
467 curated GWAS databases and restricted sources of rare genetic variants. Other studies, like
468 PheWAS (34, 35) and PhenomeXcan (36) have collated genomic associations across
469 numerous phenotypes to create resources of variant-trait and gene-trait associations.
470
471 Here, we construct a gene-trait association matrix for 16,849 genes across 1,393 complex
472 traits similarly to PhenomeXcan, and further the concept by using UnTANGLeD to identify
473 gene programs. We apply dimensionality reduction methods, which can harness the high
474 dimensional, complex gene-trait association data, allowing us to greatly expand on the scale
475 of studies previously attempting to build gene networks. By increasing the scale of data, we
476 not only identify gene programs enriched for biological processes specific to associated
477 phenotypes but also reveal gene programs enriched for central processes governing diverse
478 mechanisms of cellular development and homeostasis.
479
480 The UnTANGLeD framework is a powerful approach to identify gene programs orchestrating
481 key biological processes. We implicate novel genes in clusters enriched for known processes
482 and identify numerous novel gene programs with enrichment for protein-protein interactions
483 and no known function. We further highlight the utility of UnTANGLeD for hypothesis
484 generation and functional annotation of genes, which may be particularly valuable for non-
485 coding genes, as they are notoriously difficult to annotate *in silico* (37). Finally, the
486 UnTANGLeD framework reveals relationships between complex traits, linking phenotypes
487 by the gene programs that underpin them.
488
489 We demonstrate the utility of UnTANGLeD for predicting genes associated with complex
490 traits and diseases using a low-powered GWAS of the same trait. Currently, standard methods
491 use gene-set analysis to improve power to identify genes and pathways involved in a
492 phenotype, such as MAGMA, or GIGSEA (38–41). Our method eliminates the need to define
493 gene-sets and instead uses gene-trait association data to learn gene sets governing complex
494 traits (39), enabling us to implicate novel trait associated genes and loci from a much smaller
495 cohort size.

We further highlight the use of the UnTANGLeD clusters for gene prioritization, showing that they effectively select gene candidates at different loci related to the same phenotype. Current gene prioritisation approaches use either distance-based metrics or mapping to eQTLs to predict changes in gene expression (42). However, these also suffer from a considerable false positive rate and may not always distinguish between two genes in proximity, as noted in our data (42). Some recent methods have integrated biological data, such as gene sets, RNA sequencing and protein-protein interaction databases to further prioritise genes at a locus (43). Our framework can be used independently or integrated with any of these approaches to advance understanding of complex trait biology.

Outside of its utility in GWAS analyses, UnTANGLeD may provide key mechanisms for data analysis in medical and industry pipelines including genetic testing and drug discovery. For example, polygenic risk scores (PRS) are an emerging method that evaluate an individual's disease risk from genetic variants (44). Methods such as UnTANGLeD may help reveal genes and hence genetic variants governing cell programs underlying disease risk and hence improve prediction accuracy. In the context of pharmacogenomics, studies have shown that drug targets with genetic support from either rare or common diseases are more than twice as likely to pass through clinical trials (45, 46). Since UnTANGLeD captures gene programs associated with all complex traits and diseases, its predictive power may help de-risk candidates and thereby decrease cost associated with the drug discovery pipeline. Overall, UnTANGLeD represents a powerful and versatile framework for studying cellular gene programs to interpret diverse sources of orthogonal genetic data.

We note several limitations in our method. Primarily, that the current GWAS data does not represent the whole phenome. Furthermore, many traits are highly correlated, and disease traits are underrepresented in the UK Biobank, the main source of data in this study. Secondly, UnTANGLeD relies on S-MultiXcan to construct the gene-trait association matrix. While S-MultiXcan is powered to detect associations across all tissues, it suffers from a high false discovery rate and may perform poorly in tissues with small sample sizes. Moreover, S-MultiXcan can identify genes colocalised with a causal gene as significant, which can obscure biological signatures. Other approaches such as SMR MR-JTI may remedy this issue (47). Additionally, UnTANGLeD does not account for the predicted directionality of effect or tissue-specific effects, which may help to further increase the quality and biological

specificity of the 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://zenodo.org/record/5879956#.Yei6X1hBy3I).

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.

578
579 **REFERENCE**

580 1. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the
581 human genome. *Nature*, **489**, 57–74.
582 2. Rozenblatt-Rosen,O., Stubbington,M.J.T., Regev,A. and Teichmann,S.A. (2017) The
583 Human Cell Atlas: from vision to reality. *Nat. 2017 5507677*, **550**, 451–453.
584 3. Sudlow,C., Gallacher,J., Allen,N., Beral,V., Burton,P., Danesh,J., Downey,P., Elliott,P.,
585 Green,J., Landray,M., *et al.* (2015) UK Biobank: An Open Access Resource for
586 Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age.
587 *PLoS Med.*, **12**, 1–10.
588 4. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O.,
589 Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A., *et al.* (2021) Highly accurate
590 protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
591 5. Shim,W.J., Sinniah,E., Xu,J., Vitrinel,B., Alexanian,M., Andreoletti,G., Shen,S., Sun,Y.,
592 Balderson,B., Boix,C., *et al.* (2020) Conserved Epigenetic Regulatory Logic Infers Genes
593 Governing Cell Identity. *Cell Syst*, **11**, 625-639.e13.
594 6. Visscher,P.M., Wray,N.R., Zhang,Q., Sklar,P., McCarthy,M.I., Brown,M.A. and Yang,J.
595 (2017) 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum.*
596 *Genet.*, **101**, 5–22.
597 7. Bellomo,T.R., Bone,W.P., Chen,B.Y., Gawronski,K.A.B., Zhang,D., Park,J., Levin,M.,
598 Tsao,N., Klarin,D., Lynch,J., *et al.* (2021) Multi-trait GWAS of atherosclerosis detects
599 novel pleiotropic loci. *medRxiv*, 10.1101/2021.05.21.21257493.
600 8. Yu,G., Wang,L.-G., Han,Y., He,Q.-Y., G,Y., LG,W., Y,H. and QY,H. (2012)
601 clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters.
602 *Omi. A J. Integr. Biol.*, **16**, 284–287.
603 9. Liberzon,A., Subramanian,A., Pinchback,R., Thorvaldsdóttir,H., Tamayo,P. and
604 Mesirov,J.P. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**,
605 1739.
606 10. Cuellar-Partida,G., Lundberg,M., Kho,P.F., D'Urso,S., Gutierrez-Mondragon,L.F. and
607 Hwang,L.-D. (2019) Complex-Traits Genetics Virtual Lab: A community-driven web
608 platform for post-GWAS analyses. *bioRxiv*, 10.1101/518027.
609 11. Butler,A., Hoffman,P., Smibert,P., Papalexi,E. and Satija,R. (2018) Integrating single-cell
610 transcriptomic data across different conditions, technologies, and species. *Nat.*
611 *Biotechnol.*, **36**, 411–420.

612 12. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P.,
613 Dolinski,K., Dwight,S.S., Eppig,J.T., *et al.* (2000) Gene ontology: tool for the unification
614 of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
615 13. Szklarczyk,D., Gable,A.L., Lyon,D., Junge,A., Wyder,S., Huerta-Cepas,J.,
616 Simonovic,M., Doncheva,N.T., Morris,J.H., Bork,P., *et al.* (2019) STRING v11: protein–
617 protein association networks with increased coverage, supporting functional discovery in
618 genome-wide experimental datasets. *Nucleic Acids Res.*, **47**, D607–D613.
619 14. LM,S., C,A., S,N., YW,C., M,M., V,F., G,F. and WA,K. (2012) Disease Ontology: a
620 backbone for disease semantic integration. *Nucleic Acids Res.*, **40**.
621 15. Kanehisa,M. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids*
622 *Res.*, **28**, 27–30.
623 16. Jain,A. and Tuteja,G. (2019) TissueEnrich: Tissue-specific gene enrichment analysis.
624 *Bioinformatics*, **35**, 1966–1967.
625 17. Kuleshov,M. V, Jones,M.R., Rouillard,A.D., Fernandez,N.F., Duan,Q., Wang,Z.,
626 Koplev,S., Jenkins,S.L., Jagodnik,K.M., Lachmann,A., *et al.* (2016) Enrichr: a
627 comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.,*
628 **44**, W90–W97.
629 18. El-Gebali,S., Mistry,J., Bateman,A., Eddy,S.R., Luciani,A., Potter,S.C., Qureshi,M.,
630 Richardson,L.J., Salazar,G.A., Smart,A., *et al.* (2019) The Pfam protein families database
631 in 2019. *Nucleic Acids Res.*, **47**, D427–D432.
632 19. de Lange,K.M., Moutsianas,L., Lee,J.C., Lamb,C.A., Luo,Y., Kennedy,N.A., Jostins,L.,
633 Rice,D.L., Gutierrez-Achury,J., Ji,S.G., *et al.* (2017) Genome-wide association study
634 implicates immune activation of multiple integrin genes in inflammatory bowel disease.
635 *Nat Genet*, **49**, 256–261.
636 20. Köttgen,A., Albrecht,E., Teumer,A., Vitart,V., Krumsiek,J., Hundertmark,C., Pistis,G.,
637 Ruggiero,D., O'Seaghdha,C.M., Haller,T., *et al.* (2013) Genome-wide association
638 analyses identify 18 new loci associated with serum urate concentrations. *Nat Genet*, **45**,
639 145–154.
640 21. Tin,A., Marten,J., Halperin Kuhns,V.L., Li,Y., Wuttke,M., Kirsten,H., Sieber,K.B.,
641 Qiu,C., Gorski,M., Yu,Z., *et al.* (2019) Target genes, variants, tissues and transcriptional
642 pathways influencing human serum urate levels. *Nat. Genet.*, **51**, 1459–1474.
643 22. Shah,S., Henry,A., Roselli,C., Lin,H., Sveinbjörnsson,G., Fatemifar,G., Hedman,Å.K.,
644 Wilk,J.B., Morley,M.P., Chaffin,M.D., *et al.* (2020) Genome-wide association and
645 Mendelian randomisation analysis provide insights into the pathogenesis of heart failure.

646 *Nat. Commun.*, **11**, 163.
647 23. Kathiresan,S., Melander,O., Guiducci,C., Surti,A., Burtt,N.P., Rieder,M.J., Cooper,G.M.,
648 Roos,C., Voight,B.F., Havulinna,A.S., *et al.* (2008) Six new loci associated with blood
649 low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides
650 in humans. *Nat Genet*, **40**, 189–197.
651 24. Willer,C.J., Schmidt,E.M., Sengupta,S., Peloso,G.M., Gustafsson,S., Kanoni,S.,
652 Ganna,A., Chen,J., Buchkovich,M.L., Mora,S., *et al.* (2013) Discovery and refinement of
653 loci associated with lipid levels. *Nat. Genet.*, **45**, 1274–1283.
654 25. Stahl,E.A., Raychaudhuri,S., Remmers,E.F., Xie,G., Eyre,S., Thomson,B.P., Li,Y.,
655 Kurreeman,F.A., Zhernakova,A., Hinks,A., *et al.* (2010) Genome-wide association study
656 meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat Genet*, **42**, 508–
657 514.
658 26. Okada,Y., Wu,D., Trynka,G., Raj,T., Terao,C., Ikari,K., Kochi,Y., Ohmura,K.,
659 Suzuki,A., Yoshida,S., *et al.* (2014) Genetics of rheumatoid arthritis contributes to
660 biology and drug discovery. *Nature*, **506**, 376–381.
661 27. Schizophrenia Psychiatric Genome-Wide Association Study (GWAS)Consortium (2011)
662 Genome-wide association study identifies five new schizophrenia loci. *Nat Genet*, **43**,
663 969–976.
664 28. Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014)
665 Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, **511**, 421–
666 427.
667 29. Anderson,C.A., Boucher,G., Lees,C.W., Franke,A., D'Amato,M., Taylor,K.D., Lee,J.C.,
668 Goyette,P., Imielinski,M., Latiano,A., *et al.* (2011) Meta-analysis identifies 29 additional
669 ulcerative colitis risk loci, increasing the number of confirmed associations to
670 47. *Nat Genet*, **43**, 246–252.
671 30. Spracklen,C.N., Chen,P., Kim,Y.J., Wang,X., Cai,H., Li,S., Long,J., Wu,Y., Wang,Y.X.,
672 Takeuchi,F., *et al.* (2017) Association analyses of East Asian individuals and
673 transancestry analyses with European individuals reveal new loci associated with
674 cholesterol and triglyceride levels. *Hum Mol Genet*, **26**, 1770–1784.
675 31. Lam,M., Chen,C.-Y., Li,Z., Martin,A.R., Bryois,J., Ma,X., Gaspar,H., Ikeda,M.,
676 Benyamin,B., Brown,B.C., *et al.* (2019) Comparative genetic architectures of
677 schizophrenia in East Asian and European populations. *Nat. Genet.*, **51**, 1670–1678.

678 32. Wang,Y.-F., Zhang,Y., Lin,Z., Zhang,H., Wang,T.-Y., Cao,Y., Morris,D.L., Sheng,Y.,
679 Yin,X., Zhong,S.-L., *et al.* (2021) Identification of 38 novel loci for systemic lupus
680 erythematosus and genetic heterogeneity between ancestral groups. *Nat. Commun.*, **12**,
681 772.
682 33. Bentham,J., Morris,D.L., Graham,D.S.C., Pinder,C.L., Tombleson,P., Behrens,T.W.,
683 Martín,J., Fairfax,B.P., Knight,J.C., Chen,L., *et al.* (2015) Genetic association analyses
684 implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis
685 of systemic lupus erythematosus. *Nat Genet*, **47**, 1457–1464.
686 34. Diogo,D., Tian,C., Franklin,C.S., Alanne-Kinnunen,M., March,M., Spencer,C.C.A.,
687 Vangjeli,C., Weale,M.E., Mattsson,H., Kilpeläinen,E., *et al.* (2018) Phenome-wide
688 association studies across large population cohorts support drug target validation. *Nat.*
689 *Commun. 2018 91*, **9**, 1–13.
690 35. Pendergrass,S.A., Buyske,S., Jeff,J.M., Frase,A., Dudek,S., Bradford,Y., Ambite,J.-L.,
691 Avery,C.L., Buzkova,P., Deelman,E., *et al.* (2019) A phenome-wide association study
692 (PheWAS) in the Population Architecture using Genomics and Epidemiology (PAGE)
693 study reveals potential pleiotropy in African Americans. *PLoS One*, **14**, e0226771.
694 36. Pividori,M., Rajagopal,P.S., Barbeira,A., Liang,Y., Melia,O., Bastarache,L., Park,Y.,
695 Consortium,G., Wen,X. and Im,H.K. (2020) PhenomeXcan: Mapping the genome to the
696 phenome through the transcriptome. *Sci. Adv.*, **6**, eaba2083.
697 37. Perron,U., Provero,P. and Molineris,I. (2017) In silico prediction of lncRNA function
698 using tissue specific and evolutionary conserved expression. *BMC Bioinformatics*, **18**,
699 29–39.
700 38. Zhu,S., Qian,T., Hoshida,Y., Shen,Y., Yu,J. and Hao,K. (2019) GIGSEA: genotype
701 imputed gene set enrichment analysis using GWAS summary level data. *Bioinformatics*,
702 **35**, 160–163.
703 39. Zhu,X. and Stephens,M. (2018) Large-scale genome-wide enrichment analyses identify
704 new trait-associated genes and pathways across 31 human phenotypes. *Nat. Commun.*
705 (2018) *91*, **9**, 1–14.
706 40. de Leeuw,C.A., Mooij,J.M., Heskes,T. and Posthuma,D. (2015) MAGMA: Generalized
707 Gene-Set Analysis of GWAS Data. *PLOS Comput. Biol.*, **11**, e1004219.
708 41. Sun,R., Hui,S., Bader,G.D., Lin,X. and Kraft,P. (2019) Powerful gene set analysis in
709 GWAS with the Generalized Berk-Jones statistic. *PLoS Genet.*, **15**.
710 42. Broekema,R. V., Bakker,O.B. and Jonkers,I.H. (2020) A practical view of fine-mapping
711 and gene prioritization in the post-genome-wide association era. *Open Biol.*, **10**.

43. Schaid,D.J., Chen,W. and Larson,N.B. (2018) From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.*, **19**, 491.
44. Lewis,C.M. and Vassos,E. (2020) Polygenic risk scores: From research tools to clinical 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.

FIGURES

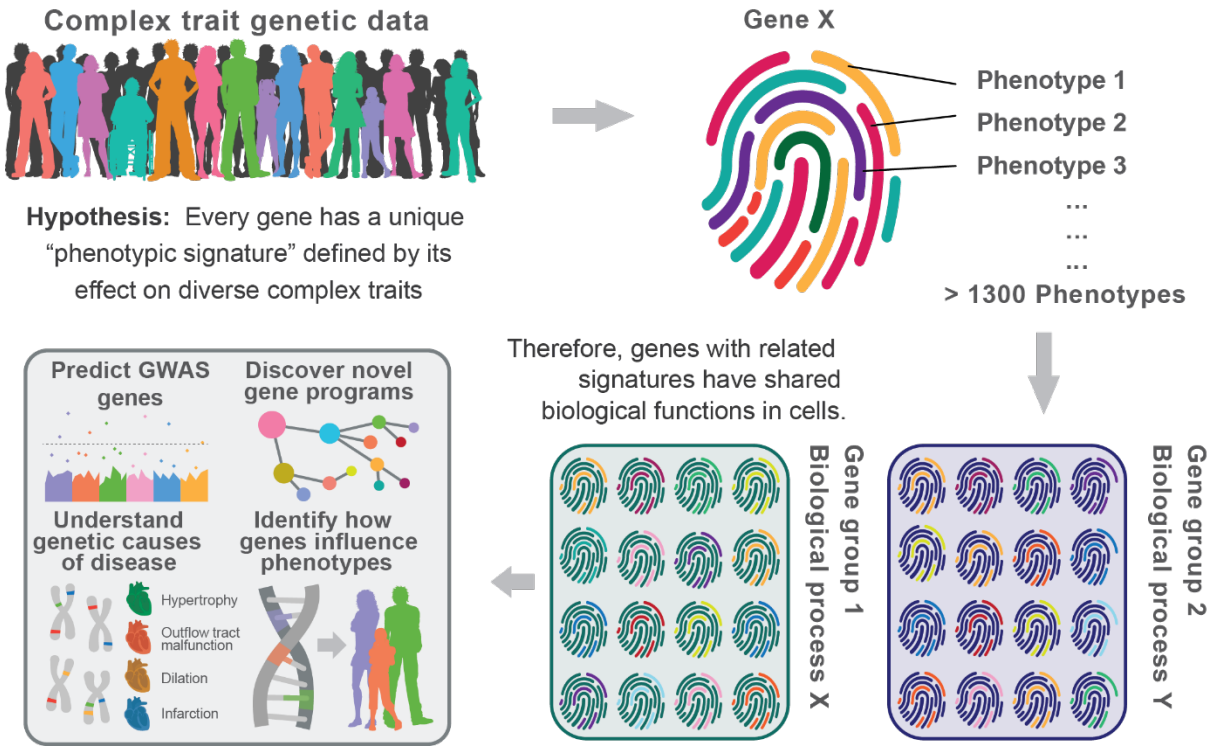


Figure 1. Schematic of central model design. Complex trait genetic data provide a unique association signature for each gene which can be used to parse the genome into functionally related gene sets.

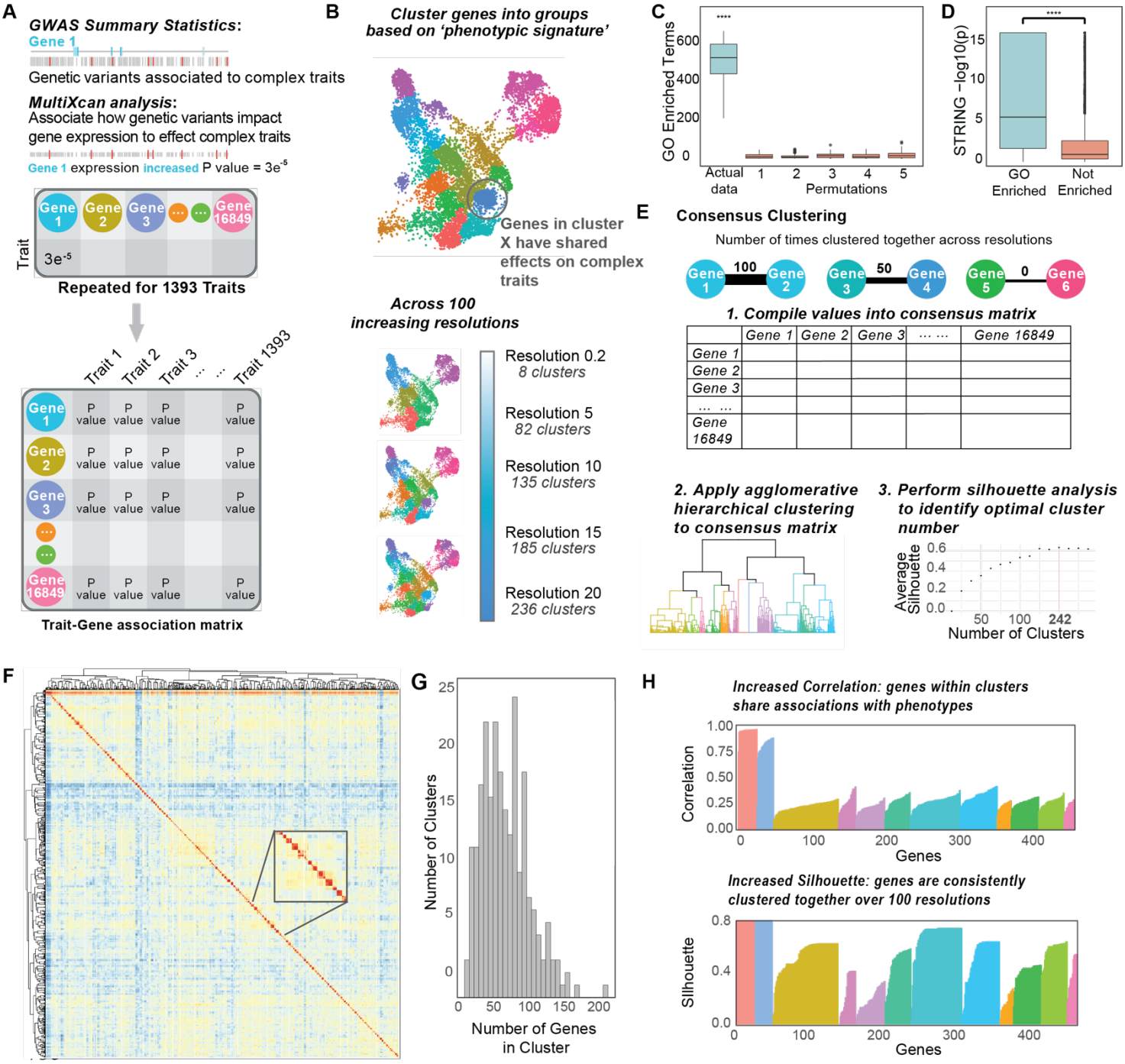
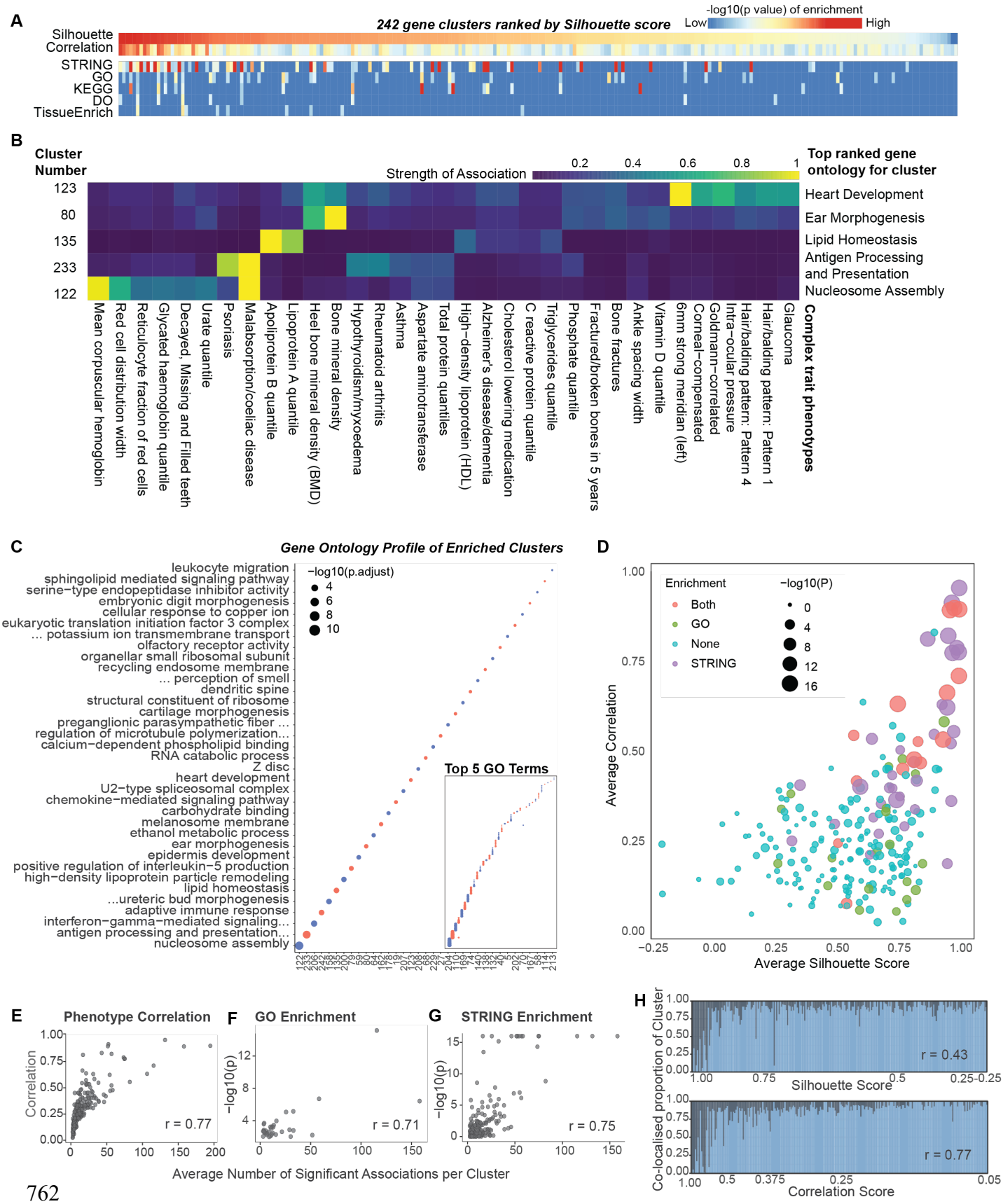


Figure 2. Consensus clustering method identifies biologically enriched gene clusters. (A) MultiXcan analysis links genetic variants to genes by predicting changes in gene expression using eQTLs. The chi-squared values of the associations between each of 1393 traits and 25851 genes were compiled into 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.

743 **(C)** Five permutations of the dataset were compared to the real data by the number of
744 enriched gene ontology terms per resolution. Enrichment was performed using
745 *ClusterProfiler,* FDR corrected p-value < 0.01. Pairwise comparisons between
746 permutations were performed with Wilcoxon signed rank test.
747 **(D)** Validation of gene ontology enriched clusters with STRING protein-protein interaction
748 enrichment. Wilcoxon signed rank test was used to compare STRING enrichment in gene
749 ontology enriched and non-enriched clusters.
750 **(E)** Each gene pair is given a similarity score based on how often they were clustered
751 together across 100 resolutions and these values are compiled into a consensus matrix.
752 Agglomerative hierarchical clustering is applied to the matrix, with the plateau in the
753 average silhouette score defining the optimal number of clusters.
754 **(F)** Heatmap of consensus matrix as clustered using agglomerative hierarchical clustering for
755 242 clusters.
756 **(G)** Histogram of the number of genes in each of the 242 clusters.
757 **(H)** Silhouette scores and correlation scores calculated for each gene to evaluate the
758 clustering robustness and quality respectively. Data generated for 450 genes selected
759 from 12 random clusters.
760

761


A
242 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
123
80
135
233
122
Heart Development
Ear Morphogenesis
Lipid Homeostasis
Antigen Processing and Presentation
Nucleosome Assembly
Complex trait phenotypes
Mean corpuscular hemoglobin
Red cell distribution width
Reticulocyte fraction of red cells
Glycated haemoglobin quantile
Decayed, Missing and Filled teeth
Urate quantile
Psoriasis
Malabsorption/coeliac disease
Apolipoprotein B quantile
Lipoprotein A quantile
Heel bone mineral density (BMD)
Bone mineral density
Hypothyroidism/myxoedema
Rheumatoid arthritis
Asthma
Aspartate aminotransferase
Total protein quantiles
High-density lipoprotein (HDL)
Alzheimer's disease/dementia
Cholesterol lowering medication
C reactive protein quantile
Triglycerides quantile
Phosphate quantile
Fractured/broken bones in 5 years
Bone fractures
Ankle spacing width
Vitamin D quantile
6mm strong meridian (left)
Corneal-compensated
Goldmann-correlated
Intra-ocular pressure
Hair/balding pattern: Pattern 4
Hair/balding pattern: Pattern 1
Glaucoma
C
Gene Ontology Profile of Enriched Clusters
-log10(p.adjust)
4
6
8
10
leukocyte migration
sphingolipid mediated signaling pathway
serine-type endopeptidase inhibitor activity
embryonic digit morphogenesis
cellular response to copper ion
eukaryotic translation initiation factor 3 complex
... potassium ion transmembrane transport
olfactory receptor activity
organellar small ribosomal subunit
recycling endosome membrane
... perception of smell
dendritic spine
structural constituent of ribosome
cartilage morphogenesis
preganglionic parasympathetic fiber ...
regulation of microtubule polymerization...
calcium-dependent phospholipid binding
RNA catabolic process
Z disc
heart development
U2-type spliceosomal complex
chemokine-mediated signaling pathway
carbohydrate binding
melanosome membrane
ethanol metabolic process
ear morphogenesis
epidermis development
positive regulation of interleukin-5 production
high-density lipoprotein particle remodeling
lipid homeostasis
...ureteric bud morphogenesis
adaptive immune response
interferon-gamma-mediated signaling...
antigen processing and presentation...
nucleosome assembly
Top 5 GO Terms
D
Enrichment
Both
GO
None
STRING
-log10(P)
0
4
8
12
16
Average Correlation
Average Silhouette Score
E
Phenotype Correlation
Correlation
r = 0.77
F
GO Enrichment
-log10(p)
r = 0.71
G
STRING Enrichment
-log10(p)
r = 0.75
Average Number of Significant Associations per Cluster
H
Co-localised proportion of Cluster
r = 0.43
Silhouette Score
r = 0.77
Correlation Score

762

763

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

766 **(A)** Broad enrichment profile of 242 clusters stratified by average silhouette and
767 correlation scores.
768 **(B)** Heatmap showing the relationship between the biological profile of five clusters, as
769 proxied by their top gene ontology term, and the unique phenotypic signature. The top 10
770 phenotypes per cluster were selected. Association strength was calculated using negative log
771 transformed significance values, which were then normalised across phenotypes and then per
772 cluster. **(C)** Top enriched gene ontology biological processes for each cluster have no
773 overlap. Clusters ranked by strength of top enriched term. Specificity of top 5 terms per
774 cluster is also provided. **(D)** 242 clusters stratified by their average correlation and silhouette
775 scores with their gene ontology and STRING enrichment indicated. The P-value represented
776 is specific to the enriched category, and in the case of both represents the more significant of
777 the two.
778 **(E-G)** Correlation between the average number of significant gene-trait associations per
779 cluster and **E)** the average correlation score of each cluster **(F)**, strength of GO enrichment
780 and strength of STRING enrichment per cluster (Pearson's correlation) **(G)**.
781 **(H)** Presence and degree of colocalization within clusters as stratified by their silhouette and
782 correlation scores (Pearson's correlation).
783
784

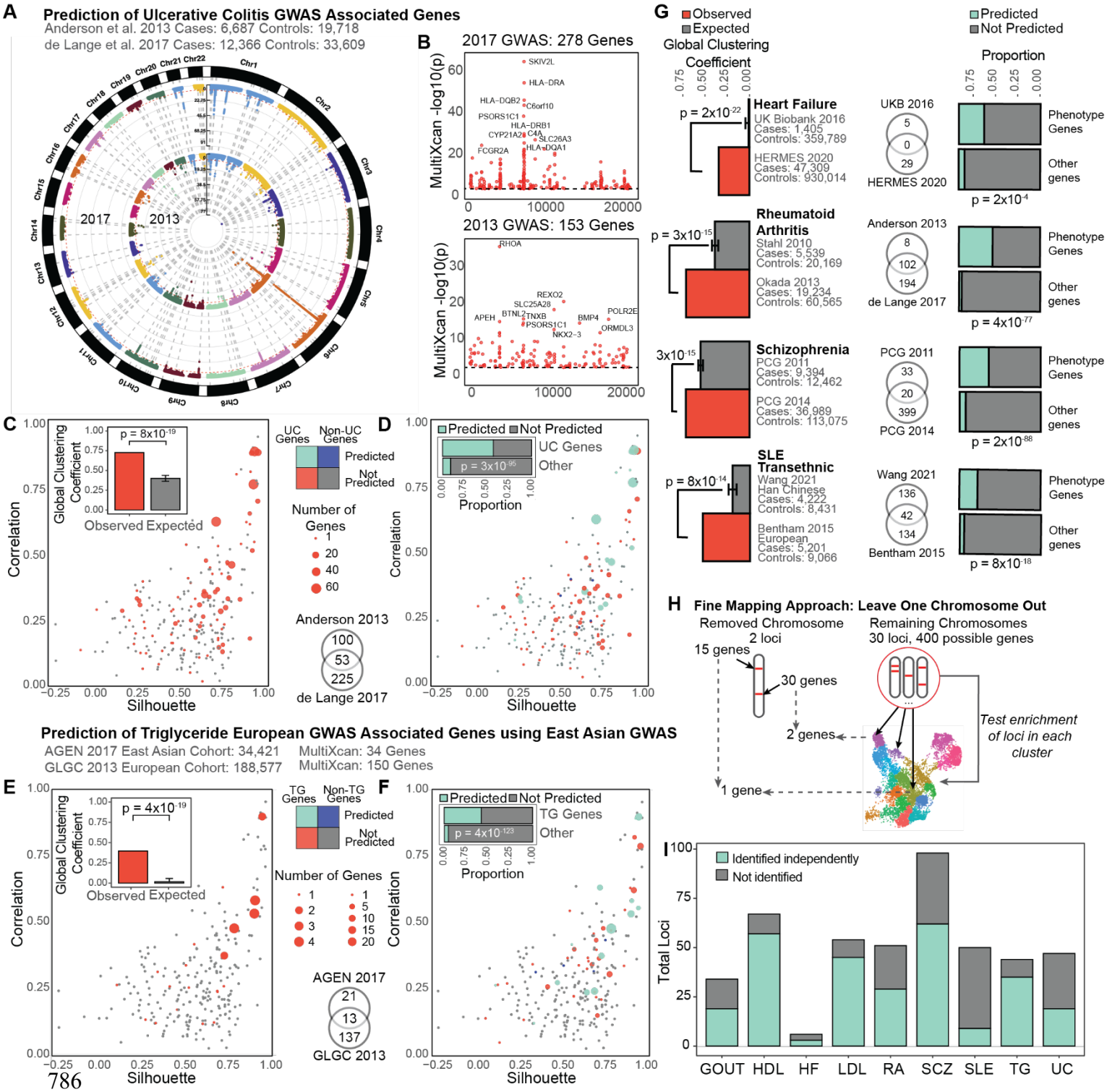


Figure 4. Identified clusters are conserved across all phenotypes and can be used for prediction of genes involved in complex trait biology and prioritization of GWAS genes at implicated loci.

791 **(A)** Manhattan plot of loci identified by a 2017 and 2013 GWAS of ulcerative colitis (UC).
792 **(B)** Manhattan plot of S-MultiXcan genes for the two GWAS respectively, genes are ordered
793 according to their genomic positions.
794 **(C)** Distribution of 278 significant genes from the 2017 UC GWAS across 242 clusters.
795 Global clustering coefficient was calculated for the 278 genes. Significance was calculated
796 using 100 bootstrap replicates to establish a distribution from which a Z score was calculated.
797 **(D)** Prediction of 2017 UC GWAS genes using 2013 UC GWAS. Chi-squared enrichment
798 test was used to determine enrichment for prediction of novel genes compared to non-trait
799 associated genes.
800 **(E)** Distribution and global clustering coefficient of 34 significant genes from East Asian
801 GWAS of Triglyceride levels. Significance was calculated using bootstrapping.
802 **(F)** Prediction of 137 novel genes from 2013 European GWAS of triglycerides using 2017
803 East Asian GWAS. Enrichment was calculated using chi-squared test.
804 **(G)** Increase in observed global clustering coefficient compared to expected and
805 prediction enrichment across four additional traits.
806 **(H)** Schematic of gene prioritization strategy using the leave one chromosome out
807 approach. Potential genes at a significant locus were refined using clusters enriched for the
808 trait.
809 **(I)** The proportion of loci at which a gene was successfully identified independently of all
810 genes on the same chromosome.
811