Published online 26 April 2022

Nucleic Acids Research, 2022, Vol. 50, No. 8 ***4289–4301***
https://doi.org/10.1093/nar/gkac274

Gene-SCOUT: identifying genes with similar continuous trait fingerprints from phenome-wide association analyses

Lawrence Middleton[1], Andrew R. Harper[1], Abhishek Nag[1], Quanli Wang[2], Anna Reznichenko[3], Dimitrios Vitsios[1,*] and Slavé Petrovski[1,4,*]

[1]Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK, [2]Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Waltham, MA, USA, [3]Translational Science & Experimental Medicine, Research and Early Development, Cardiovascular, Renal and Metabolism, BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden and [4]Department of Medicine, University of Melbourne, Austin Health, Melbourne, Victoria, Australia

Received October 19, 2021; Revised March 31, 2022; Editorial Decision April 07, 2022; Accepted April 09, 2022

ABSTRACT

Large-scale phenome-wide association studies performed using densely-phenotyped cohorts such as the UK Biobank (UKB), reveal many statistically robust gene-phenotype relationships for both clinical and continuous traits. Here, we present Gene-SCOUT, a tool used to identify genes with similar continuous trait fingerprints to a gene of interest. A fingerprint reflects the continuous traits identified to be statistically associated with a gene of interest based on multiple underlying rare variant genetic architectures. Similarities between genes are evaluated by the cosine similarity measure, to capture concordant effect directionality, elucidating clusters of genes in a high dimensional space. The underlying gene-biomarker population-scale association statistics were obtained from a gene-level rare variant collapsing analysis performed on over 1500 continuous traits using 394 692 UKB participant exomes, with additional metabolomic trait associations provided through Nightingale Health's recent study of 121 394 of these participants. We demonstrate that gene similarity estimates from Gene-SCOUT provide stronger enrichments for clinical traits compared to existing methods. Furthermore, we provide a fully interactive web-resource (http://genescout.public.cgr.astrazeneca.com) to explore the pre-calculated exome-wide similarities. This resource enables a user to examine the biological relevance of the most similar genes for Gene Ontology (GO) enrichment and UKB clinical trait enrichment statistics, as well as a detailed breakdown of the traits underpinning a given fingerprint.

INTRODUCTION

Large-scale cohorts, such as the UK Biobank (1) that combine exome sequence data with longitudinal medical records and biomarkers, offer a unique opportunity to identify novel gene-phenotype associations. Phenome-wide association studies (PheWAS) performed on such densely-phenotyped cohorts reveal many statistically robust gene-phenotype relationships (2). Exploiting the associations derived from these analyses, we here develop a tool capable of elucidating causal gene networks through constructing similarity scores between genes. Such a tool not only provides insight into gene variants that share similar phenotypic properties but can also be used to suggest alternative drug targets to those that are otherwise intractable. As such, we here introduce Gene-SCOUT (Gene-Similarities from COntinUous Traits) that takes a user-defined gene and identifies other human genes with a similar continuous trait fingerprint. A fingerprint represents levels of statistical associations between genes and continuous traits across multiple gene-based collapsing models. A range of genetic architectures (termed 'qualifying models') (3) is evaluated in a population of 394 692 exome sequences with 1419 continuous traits available for each participant and 168 metabolomic traits profiled by Nightingale Health on a subset of ∼120K of the participants (4).

Much of the existing methodology around understanding gene similarity have relied on 'semantic similarity', which aims to quantify the similarity or distance between pairs of terms organized in an ontology. Generally, these approaches have utilised observations from low-level annotations derived from the Gene Ontology database (5,6), for ex-

*To whom correspondence should be addressed. Tel: +44 7384 520 047; Email: dimitrios.vitsios@astrazeneca.com
Correspondence may also be addressed to Slavé Petrovski. Email: slav.petrovski@astrazeneca.com

ample, GOSim (7) while others like HPOSim (8) are based on data derived from Human Phenotype Ontology (HPO) (9). Alternative methods to these ontology based similarity scores leverage, for example, protein-protein interaction networks such as STRING (10).

Gene-SCOUT differs from previous approaches by relying solely on the association statistics generated from a large-scale population-based exome sequencing data linked with continuous trait measurements. We provide empirical comparisons against GOSim and HPOSim as well as an additional method for calculating gene similarities based on the STRING protein-protein interaction database. To compare the performance of the methods, we obtain 2662 gene sets spanning a number of resources, including OMIM, KEGG biological pathways, PanelApp and the aforementioned UKB PheWAS. We see Gene-SCOUT's performance vary depending on the resource, though observe substantial outperformance against all other methods on OMIM and UKB PheWAS gene sets, as measured by the overlap of similar genes to a given gene set.

We accompany our method with a user-friendly interactive web resource to help explore the resulting similarities in greater detail. Finally, gene clusters are identified based on the derived similarity measures. Using this approach, relying solely on cohort association statistics, we are able to recapitulate known biologically similar genes thereby providing an opportunity to provide additional insight beyond the positive controls.

MATERIALS AND METHODS

Datasets

The UK Biobank is a prospective, longitudinal cohort study of ∼500 000 participants, aged between 40 and 69 years when recruited between 2006 and 2010 (1). Participant data, based on questionnaires and assessment visits, includes health records capturing various phenotypic endpoints and are periodically updated by the UK Biobank (11). Records range from blood biomarkers to imaging and accelerometer readings, though the following restricts attention to continuous traits listed in Supplementary Table S1. Whole exome sequencing (WES) data for UK Biobank participants were generated at the Regeneron Genetics Center (RGC) as part of the UKB-ESC pre-competitive data generation collaboration between AbbVie, Alnylam Pharmaceuticals, AstraZeneca, Biogen, Bristol Myers Squibb, Pfizer, Regeneron and Takeda with the UK Biobank (12). Genomic DNA underwent paired-end 75 bp whole exome sequencing (WES) at Regeneron Pharmaceuticals using the IDT xGen v1 capture kit on NovaSeq 6000 machines. Details regarding the AstraZeneca Centre for Genomics Research Bioinformatics Pipeline and quality control procedures undertaken are described elsewhere (2). In brief, FASTQ files were aligned to GRCh38 and small variant SNVs and indels called (Illumina DRAGEN Bio-IT Platform Germline Pipeline v3.0.7), and subsequently annotated (SnpEff v4.3 against Ensembl Build 38.92). Distantly related participants (kinship coefficient < 0.0884) of European ancestry were selected for gene-based collapsing analyses, reducing the cohort from approximately 455 000 to 394 692 samples. Collapsing analyses, that aggregate specific 'qualifying variants' (QVs) (3) within a gene, test whether there is a difference in the proportion of individuals carrying one or more QVs between cases and controls. Eleven QV models were evaluated to assess the impact of rare genetic variants on human disease; of these, 10 assess non-synonymous variants (nine dominant and one recessive model), plus an additional synonymous variant model adopted as an empirical negative control. In addition to the aforementioned continuous traits we also include metabolomic data captured across a subset of over 120 000 participants of the UK Biobank cohort (Supplementary Methods, section 10).

Constructing a gene fingerprint

A unique continuous trait fingerprint was generated for each of the 18 862 genes considered. The number of genes was motivated through a previous collapsing analysis of ∼300 000 UKB exomes (2). Each fingerprint was compiled from the regression coefficients derived from univariate linear regressions between pairs of genes and continuous traits (1587 in total). Specifically, regressions were estimated through regressing normally-transformed continuous traits against the binary indicators (presence or absence) of a gene carrying at least one QV. Each regression adjusted for age and sex, and all of these regressions were repeated for each collapsing model. As gene-trait associations are available for multiple QV models, signatures were constructed through the concatenation of regression coefficients across all continuous traits and all QV models (excluding the synonymous model) to form an extended feature set (Supplementary Methods, section 3). The following methodology aims to exploit both the associations and their accompanying p-values to best capture similarities between genes.

Exploring similarity methodologies

There are different ways to measure distance or similarity between pairs of vectors (13–15). Having constructed gene fingerprints from the association scores and corresponding *P*-values for all protein-coding genes there remained a number of possible ways in which to quantify the similarity between two fingerprints. *A priori*, it is unclear what aspect of the signature is most informative—for example, genes could be deemed 'similar' based on a few informative traits specific to each gene (perhaps those that are most significant or have the highest degree of association), or it may be beneficial to aggregate the differences in all association scores, in which case some of this finer detail may be diluted amid sample randomness. Also, given that in general it is possible to recover a notion of similarity from a distance metric (the closer the distance, the more similar the vectors) selecting an appropriate distance metric may also be critical to determining which aspect of the signatures contribute most to their similarities. Given these variables, a preliminary investigation was conducted with the objective of refining the methodology to produce more biologically meaningful similarities (Supplementary Methods, section 3). Improvements to the methodology were guided by a set of 22 expertly-curated genes (Supplementary Methods, Appendix) with well understood biological functions that

were to be used as seed genes. It was then possible to assess the performance of different methods both quantitatively, through Gene Ontology and PheWAS enrichment analyses of reportedly similar genes, and qualitatively, through manual inspection of the reported genes and the relevance of the top enrichment terms, (Supplementary Figures S3 and S4).

The investigation revealed two crucial insights, firstly, that the inclusion of a feature selection strategy could considerably enhance the performance of the similarity calculation (Supplementary Figure S24) and, secondly, that there was merit in opting for a nonstandard distance function (Supplementary Figures S24 and S25). A number of distinct methods were trialed, including normalizing the association statistics by the *P*-values to exaggerate associations that were highly significant, sub-setting features that were significant in either the seed gene or an alternative gene and sub-setting genes based on only those that shared significant associations. At each stage, refinements were made based on the inspection of the most similar genes to the expertly curated ones, as well as their enrichment for Gene Ontology terms and PheWAS binary traits (Supplementary Figures S3 and S4).

Gene similarity calculation

Gene-SCOUT's similarity calculation was chosen based on an exploration of different strategies looking to optimize the biological relevance of the closest genes retrieved to a set of expertly-curated seed genes. We have tried a range of distance metrics, which can be largely represented by either Euclidean or cosine distance (Supplementary Figure S2). Eventually, we selected cosine similarity measure due to its stronger performance in terms of enrichment of similar genes on a set of expertly curated genes (Supplementary Figure S25) and also due to providing a normalized measure of similarity between vectors, which can aid the interpretability of the similarities. One possible reason for its superior performance is that the cosine distance better captures the directionality of two vectors, which can provide an additional degree of robustness when studying the effect of genic variation to continuous traits. This is in addition to other favorable robustness properties of the cosine distance in high dimensions (Supplementary Appendix 1).

The method rests critically on a tuning parameter, α, used to threshold the significance of a gene–trait association, ensuring only those associations that satisfy a degree of population-scale statistical robustness are included in the signatures. We identify an optimal value of this parameter based on a sensitivity analysis, described subsequently. Having chosen its value and proposed a seed gene to identify other similar genes from, the method proceeds as follows (Figure 1):

1. Construct the seed gene's signature:
   a. Identify the continuous trait associations that are significant at level α for this gene
   b. Collect the gene-trait association scores into a vector z
2. Populate a set of other gene signatures:
   a. Each gene must be significantly associated to at least one trait identified in 1a
   b. Collect the gene-trait associations scores into vectors $x_1, x_2, \ldots$ (one for each gene) using the set of traits in 1a (matching the order of traits in z)
   c. Gene-trait associations that are not significant in this set are set to 0
3. Calculate the cosine similarity between pairs (z, x_1), (z, x_2) etc. For two D-dimensional vectors x and y, the cosine distance is defined as

$$d_{cos}(x, y) = 1 - \frac{\sum_{i=1}^{D} x_i y_i}{\sqrt{\left(\sum_{i=1}^{D} x_i^2\right)}\sqrt{\left(\sum_{i=1}^{D} y_i^2\right)}} \quad (1)$$

where the distance varies in the interval [0,2], with $d_{cos}(x, y) = 0$ if and only if $x = y$, $d_{cos}(x, y) = 1$ when x and y are perpendicular to each other and $d_{cos}(x, y) = 2$ when x and y are pointing in exactly opposite directions.

A crucial element of the method is the 'filtering out' of any traits that are not significantly associated to the seed gene (occurring in 1a). Other gene signatures (x_1, x_2, etc.) may not share exactly the same set of significantly associated traits, but due to step 2a there is a guarantee that at least one trait is significantly associated in both z and x_1 for example. Furthermore, there is no requirement for α to be set to, for example, genome-wide significance, indeed a sensitivity analysis revealed that a more relaxed threshold was in fact preferable.

An artefact of the above formulation is that after accounting for feature selection $Sim(A, B) \neq Sim(B, A)$ where A and B are genes and $Sim(\cdot, \cdot)$ is the Gene-SCOUT similarity between them, i.e. the similarity is not symmetric in its arguments. As such, a symmetric similarity measure can be provided by

$$Sim^*(A, B) = \varphi(Sim(A, B), Sim(B, A))$$

where φ is any function satisfying $\phi(a, b) = \phi(b, a)$. For example φ could be either $\min(a, b)$ or $(a + b)/2$. We exploit this symmetrisation further in estimating clusters from Gene-SCOUT distances.

Sensitivity analysis for thresholding significant gene–trait associations

Importantly, in order to utilise the vectors of *P*-values for associations the user must specify a significance level, α, which for these purposes we set to $\alpha = 1 \times 10^{-5}$. We perform a sensitivity analysis to optimize the *P*-value threshold used to calculate similarities between genes. The *P*-value threshold is used twice in calculating similarities. Firstly, a candidate list of similar genes is constructed based on those genes that share at least one significant trait with the seed gene. It is only these genes that Gene-SCOUT provides similarities for. Secondly, based on this set of genes, similarities are calculated using the cosine similarity, based on features that are significant in the seed gene. This requires a second use of the threshold.

To summarise the performance of Gene-SCOUT in evaluating similarities that are biologically relevant, we consider three gene sets, each of which are enriched for separate phenotypes. The three gene sets considered were those relating

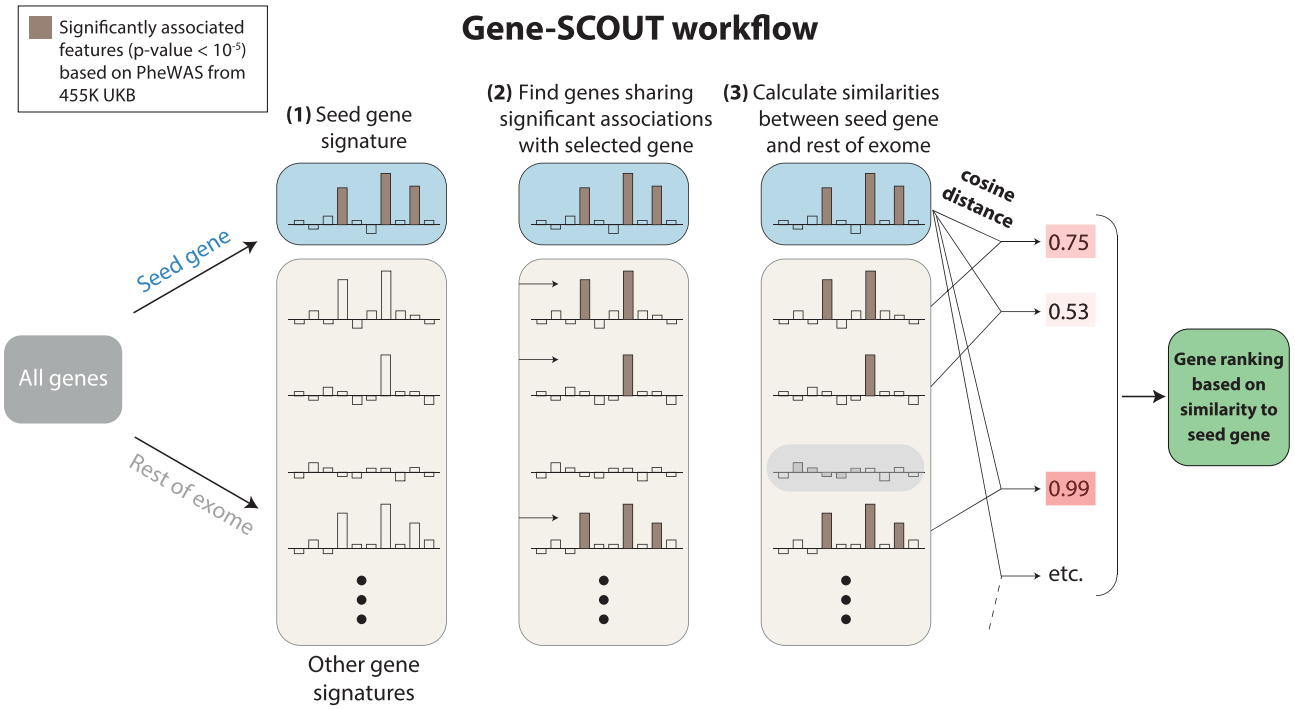

Figure 1. Schematic of Gene-SCOUT's similarity calculation. Each barplot describes a signature specific to a different gene, with bars representing an association score between a gene and a continuous trait (beta coefficient from linear regression between genotype and continuous trait). Significant associations ($P < 10^{-5}$) are denoted with a brown fill color. Gene-SCOUT finds a subset of genes that share at least one significant association with the seed gene and calculates the respective similarities using the cosine distance metric. Genes are eventually ranked based on their Gene-SCOUT similarity score to each seed gene.

to (i) lipoprotein disorders, (ii) diabetes (type I and type II) and (iii) chronic kidney disease. The choice of these gene sets was guided by the requirements that, firstly, there must be a large number of genes known to be associated to a clinically relevant phenotype. This was measured through ranking the PheWAS clinical traits according to the number of significantly associated genes at 1×10^{-8}, for which these gene sets rank among the top 2% of 17 361 phenotypes captured among the previously published PheWAS (Supplementary Methods, section 3). Secondly, to guarantee that a gene's signature could be calculated for all genes in the gene set we required that the diseases could be inferred from a set of relevant continuous traits measured in UKB. A list of all gene sets and relevant continuous traits as well as a breakdown of any genes common to multiple sets is provided (Supplementary Methods, Appendix).

We focus the analysis on a single biologically meaningful PheWAS clinical trait for each gene set and examine how the average PHRED score changes as a function of the p-value threshold (Supplementary Figure S5). We see according to the median PHRED score that 1×10^{-5} achieves strong enrichment across each of the three disorders (Supplementary Figure S6), with gene-set enrichment PHRED scores plateauing or declining at more stringent thresholds. Increasingly stringent thresholds also reduce the gene-space from which similarity scores can be constructed. We also see substantial enrichment for closely related disease phenotypes at this optimal threshold (Supplementary Figure S7).

Similarities between exome genes in the UKB dataset

Having optimised the similarity methodology, we calculate similarities between all exome genes based on the UKB dataset. Importantly, we note that similarities can only be calculated between genes that share at least one significant association. As such, given that not all genes in the exome have at least one significant association at a given significance threshold, we restrict similarity estimation to those genes—2880 (15% of 18 862 genes) – that do (Supplementary Figure S10). We see that 95% of these have under 25 significant traits (Supplementary Figure S1). As a strict subset of this, approximately 11% of all protein-coding genes have at least one other gene that is similar to each of them. The sparsity of similarities is comparable to the sparsity of significant traits in the exome. We see also that within features that have at least one significant association there is a correlation structure (Supplementary Figure S20, Table S2). A correlation analysis revealed both that Gene-SCOUT similarities are robust to the inclusion of correlated features through simulated data and that for real data the enrichment analyses are robust to including or excluding highly correlated features (Supplementary Methods section 4, Supplementary Figures S8 and S9). Furthermore, it is beneficial to retain features that a user could use to establish the biological relevance of a resulting similarity, rather than arbitrarily remove one or more from a set that are highly correlated with each other.

Constructing a similarity network using *k*-nearest neighbor graphs

To aid exploration of similar genes we construct a network whereby an edge between two genes is indicative that the two genes are similar. In particular, the network is constructed such that an edge between two genes, A and B, signifies that B is within the k most similar to A according to the derived similarities. Such a network is referred to as a 'k-nearest neighbor graph' (16) and has been used in other settings to lower the computational cost of the closely related 'k-nearest neighbor' clustering algorithm (17,18). Due to the sub-setting technique in the similarity calculation, where genes must share at least one significant trait with a seed gene in order for the similarity to be evaluated, some genes will not have any other genes on which to evaluate similarities. This is reflected in the network figure, where only genes that have at least one other gene for evaluating similarities are included. More explicitly, if for $GENE_0$, it has genes $GENE_1$, $GENE_2$, …, $GENE_{10}$ that are all in its close list, then the network figure reflects this through an edge from $GENE_0$ to $GENE_1$, an edge from $GENE_0$ to $GENE_2$ etc. It is noted that sometimes there will be edges in the opposite direction, i.e. for $GENE_2$, $GENE_0$ ranked in its top k list, however a repeated edge though them is not a requirement. An edge in either direction is sufficient to aid exploration.

Clustering and visualization based on similarities

Based on the above formulation, it is possible to provide pairwise similarities within a set of genes. Having converted these similarities to distances, it is then possible to use unsupervised machine learning algorithms to estimate clusters within the gene set, where genes in the same cluster indicate a large Gene-SCOUT similarity. To allow for novel distance metrics (since some clustering algorithms are restricted to the Euclidean distance only), we focus on the OPTICS clustering algorithm and demonstrate its superior performance when testing cluster enrichments over an alternative method—Louvain clustering (19) (Supplementary Figure S23).

As part of this process, we note that it is also possible to visualize similarities between genes—irrespective of the dimensionality of the original feature space. Similarities are closely related to distances and for some manifold learning techniques, e.g. t-distributed stochastic neighbor embedding (t-SNE), they require only a precomputed distance matrix—i.e. a matrix where element $[i, j]$ represents the distance between points i and j. In general, the distance matrix may or may not represent Euclidean distances, this depends on the precise application, but for these purposes we ensure distances reflect similarities between genes as defined above (Supplementary Figure S22). Importantly, as the input features may vary, the similarity of X_i to X_j may not be the same as the similarity of X_j to X_i. As such a symmetrisation procedure is required to convert the similarities to distances that are symmetric in X_i and X_j. We describe this symmetrisation along with additional subtleties relating to converting cosine similarities to distances in full in the Supplemental Methods.

As part of the validation of the derived clusters, we explored whether human paralog genes preferentially cluster together compared to random sets of genes. For this analysis, we first extracted all sets of human paralog genes from PANTHER DB (32) and retained those that were assigned to a cluster based on the OPTICS algorithm (leaving eventually a set of 569 seed genes for examination). For each seed gene, we checked whether it has at least one paralog belonging to the same Gene-SCOUT derived cluster and we found 39 such clusters. We repeated the same process using 569 random seed genes and respective random sets of genes (of matched size with the real sets of paralog genes) and cluster co-occurrence was achieved in only five clusters (median number from 10 random iterations; standard deviation: 3.4). That corresponds to an ∼8-fold enrichment of cluster co-occurrence for paralog genes compared to random ones (OR = 8.3; P-value = 8.4×10^{-8} via Fisher's exact test).

Description of benchmarked similarity measures

Three alternative methods for calculating similarities between protein-coding genes were considered, namely GOSim, HPOSim and scores derived from protein-protein interaction networks in StringDB. Both GOSim and HPOSim require seed genes to be expressed in terms of Entrez IDs. We therefore use the R package 'org.Hs.eg.db' (20) to map HGNC gene names to the appropriate nomenclature. While GOSim implements an array of similarity measures between genes, we opt for the feature space embedding approach due to its computational tractability. For HPOSim, we calculate similarities between genes using the function 'getGeneSim', with the default settings of 'method = Resnik' and 'combinemethod = funSimMax', both of which relate to how similarities are calculated using the underlying ontology of gene annotations.

In general, each method provides coverage of different parts of the exome. After translating between gene nomenclatures and calculating all gene similarities where possible we see that of the original 18 930 genes registered within the collapsing analysis, Gene-SCOUT provides similarities between 2097 of these (11%). Similarly, the coverage of GOSim, HPOSim and StringDB is 14 028 (74%), 2,785 (15%) and 18 008 (95%) respectively. We note that the Pearson's correlation coefficient between Gene-SCOUT similarities and other methods is under 0.05 when taking the intersection of available gene-gene pairs (Supplementary Figure S26), noting that the intersection is allowed to vary between two compared methods. This suggests that our proposed method is capturing similarities that are generally orthogonal to the other methods.

Quantifying performance of similarity methods

The performance of Gene-SCOUT (and any other method) on a particular gene set was calculated as follows. Take an example gene set from a resource (e.g. all genes associated with 'Deafness' in OMIM), denote this as G. Then for each gene in G recover the (at most) n most similar genes, ensuring they do not include the seed gene. This results in $m = size(G)$ sets of similar genes each with length $\leq n$.

We then take the union of all of these m gene sets, denoted as H_n for a given value of n, and quantify the overlap with G using Fisher's exact test. A positive overlap can arise for example if a seed gene in G appears in one of the remaining seed genes' lists.

The contingency table for Fisher's exact test is constructed through considering a reference set of either (i) 2097 genes with Gene-SCOUT similarities or (ii) 18 862 protein-coding genes. For a given reference set, the contingency table counts the number of genes that are positive or negative for inclusion in G as well as positive or negative for inclusion in H_n.

External gene sets used for benchmarking

The benchmarking procedure makes extensive use of external gene sets to establish how successfully Gene-SCOUT retrieves similar genes that are known to be enriched for the same phenotype. In particular, four resources are included, namely OMIM (21), KEGG (22), PanelApp (23) and a PheWAS collapsing analysis of UKB exomes (2). OMIM gene sets were constructed through parsing an OMIM dump (accessed 19 April 2021). KEGG gene sets comprised 186 gene sets taken from MSigDB's canonical pathways. PanelApp gene sets were extracted through downloading 326 panels from the PanelApp website and then restricting gene sets to only those with a 'GEL_status' of 2 or 3 (i.e. the association is either amber or green), accessed on 30 November 2021.

Further gene sets were also constructed using the PheWAS collapsing analysis of UKB exome data – as used to generate the original gene signatures. In this case, however, binary traits were considered rather than the continuous ones used to generate the signatures. Each binary phenotype association was quantified using Fisher's exact test p-values from 11 different collapsing models, including a synonymous model which was subsequently excluded for these purposes. As such, a single measure of association between a given gene and phenotype was provided through taking the minimum P-value over all of these collapsing models. These summary p-values were then used to construct three collections of gene sets, where each set corresponded to a different significance threshold (10^{-4}, 10^{-5} and 10^{-8}). For example, given a significance threshold of 10^{-4}, a gene set 'Non-insulin-dependent diabetes mellitus' comprises all those genes with summary P-value $\leq 10^{-4}$ for their association significance with that particular binary trait.

Web resource with pre-calculated results

We provide for each input gene a network of closest genes, similarities, enrichments and gene signatures. Taking *LDLR* as an example we describe the outputs of these analyses in the web resource (Figure 1). Firstly, we are able to explore the portion of the network that immediately surrounds *LDLR*. The red node represents the seed gene (*LDLR* in this case) and the blue genes represent its 'one-hop' neighbors. The remaining grey nodes represent 'two-hop' neighbors of *LDLR*. Finally, for *LDLR*'s 10 most similar genes we color the corresponding edge pink and weight the thickness according to the strength of the similarity. Furthermore, for the list of nearest genes reported, we are able to test whether they are statistically enriched for other biologically relevant properties. We do this first of all by performing a Gene Ontology enrichment analysis for biological processes and secondly for clinical traits studied using the same exome sequences in the UK Biobank. More descriptions of the web resource and details of the enrichment tests are provided in Supplemental Methods.

Trait finder

Trait finder provides a means to match genes that satisfy a certain quantitative trait fingerprint as measured by the direction of significant associations. The user provides two sets of traits, looking for either positive or negative (significant) associations with underlying genotypes, to construct the desired fingerprint. Specifically, each trait is used to query the entire exome and identify those genes that positively or negatively associate with it in the event of a variation in its sequence (with variations being captured by the qualifying variant models used in this study). The significance level for Trait finder (unlike the similarity calculations) can be set by the user.

Genes are scored based on this desired fingerprint. Each trait in each of the two sets is iterated over and if, for example, a gene is positively associated at the given threshold with the trait (and the trait is in the desired positively associated set), then it scores an additional point. Having scored each gene based on the traits in each of the sets they are ranked with highest points representing the most closely matching fingerprint.

RESULTS

We here leverage the vast scale of the UKB cohort to identify similarities for each gene based on their phenotypic signatures. The dataset itself expands from the 281 104 samples studied previously (2), now containing samples from 394 692 individuals and also incorporates metabolomic features from a subset of 121 394 individuals. This rich dataset allows us to quantify the associations between many underlying genetic architectures and continuous traits with the objective of better capturing the 'macro' effect of genetic variation, measuring the association with continuous traits as driven by rare-variant mutations. For each gene we have a multi-dimensional vector quantifying these associations to over 1587 different traits. In order to capture biologically meaningful similarities between genes we explored different methodologies and strategies to best exploit both the association scores, and their accompanying P-values (see Materials and Methods). The latter was instrumental in helping to improve the statistical robustness of the resulting similarities. Other variables involved in the exploration included comparing different distance metrics, different ways to filter out non-significant associations and varying significance levels. The final methodology was tuned through a sensitivity analysis using a set of 22 manually curated gene sets as a gold standard (Supplementary Methods, Appendix).

We apply this optimised methodology across the whole exome, based on association scores derived from the UKB dataset. As the calculation relies on genes possessing at least one significant association across the 1587 traits measured and 10 collapsing models, we restrict the calculation

to 2097 of the original 18 862 genes (11%). In the following, we first illustrate the power of the approach through construction of a gene network and identify clusters driven by these similarities – genes are assigned to the same cluster if they share a high degree of similarity. Next, we benchmark the similarities against competing methods, namely GOSim (7), HPOSim (8) and a similarity derived from protein-protein interaction scores using StringDB (10). We conclude with an exposition of a publicly available web-utility, built to both explore and validate Gene-SCOUT similarities through a user-friendly interface.

Validation of clusters using known gene sets

Similarities were used to cluster genes and map their fingerprints through a 2D projection to visualize the clusters, applicable only to those genes that have at least one other similar gene. To verify the results of the projection we exploit three expert-curated gene sets around dyslipidemia, diabetes and chronic kidney associated genes. We plot the low-dimensional projection of Gene-SCOUT similarity for genes that have at least one other similar gene (Figure 2A), showing in addition the location of the genes that are enriched for each of these disorders (some genes may be shared across more than one validation set). We observe a consistent grouping of the gene members of each gene set, providing a degree of confidence that the similarities are based on biologically meaningful associations. Additionally, we observed that human paralog genes, as defined in Panther DB (32), preferentially group into the same Gene-SCOUT derived clusters compared to random pairs of genes (Fisher's exact odds ratio = 8.3, $P = 8.4\times10^{-8}$; see Materials and Methods).

Biological function of most enriched clusters

Gene clustering is first performed (using the OPTICS algorithm) and then each gene (point) is projected into the plane (using t-SNE). We see that natural clusters arise (Figure 2A, Supplementary Figure S12); however, to test the meaning of these clusters we perform enrichment analyses on each cluster, investigating the enrichment of a cluster for both GO terms and PheWAS clinical traits. We evaluate the top performing clusters based on the maximum PHRED score achieved across all traits. Reassuringly, the top three clusters contain well-established gene-phenotype relationships (Figure 2B). For example, cluster one demonstrates enrichment for blood disorders, specifically hereditary spherocytosis, through the clustering of known disease genes such as *ANK1, SLC4A1, SPTA1, SPTB* and *EPB42*. Cluster two shows enrichment for lipid related disorders, with *ANGPTL3, APOB, HMGCR, NPC1L1* and *PCSK9* all well-established disease genes underlying lipoprotein metabolism. Similarly, cluster three shows strong enrichment for chronic myeloproliferative disease, with *CALR, JAK2* and *MPL* as established disease genes. An extended list of clusters and their enrichments with PheWAS clinical traits shows other meaningful clusters (Supplementary Figure S13) such as *PKD1, PKD2* and *CLDN10* sharing a cluster enriched for kidney disorders as an exemplar. To gauge the uniqueness of the clusters, we extract the top 1,000 most strongly enriched traits across all clusters and count their co-occurrence across multiple clusters, varying the significance threshold between 10^{-2} to 10^{-6} (Supplementary Figure S28). We see that at the weakest threshold, 81.7% of the traits are enriched in only one cluster (increasing to 100% at 10^{-5} threshold), suggesting that the majority of the clusters capture unique biology between them. Furthermore, we also provide a negative control through permuting the similarities between genes and then repeating both the clustering and enrichment analyses based on the permuted similarities (Supplementary Methods, section 6). We see that restricting to the top 3 enrichment terms in each cluster, corresponding to approximately the highest 5% of terms, the mean PHRED score was significantly larger for Gene-SCOUT in both PheWAS clinical traits and GO terms ($P \leq 10^{-5}$; Supplementary Figure S14).

Based on the derived similarity scores from Gene-SCOUT we are able to recover biologically meaningful clusters, thereby further validating the methodology. Genes sharing the same cluster are also linked through sharing similar significant associations with continuous traits. For example, in cluster two, *ANGPTL3, APOB, HMGCR, NPC1L1* and *PCSK9* are all known to be related to lipid metabolism and, as such, it is possible that other genes highlighted within this cluster also relate to lipid-related disorders, but may not be as well appreciated yet.

Benchmarking against other similarity scores on over 2500 gene sets

We compare the performance of Gene-SCOUT's similarity scores to those provided by GOSim (7), HPOSim (8) as well as StringDB's protein-protein interaction scores (10). All benchmarked methods provide a similarity score for a given seed gene to other genes in the exome, though they do not necessarily provide similarities for all protein-coding genes and may cover different parts of the exome (see Materials and Methods). We therefore focus predominantly on the portion of the exome for which Gene-SCOUT provides similarities for.

After an initial benchmarking procedure provided on the three gene sets described in the sensitivity analysis (Supplementary Figure S29) we extend the scope of the procedure to accommodate diverse collections of gene sets from multiple public resources. Each phenotype represents a ground truth from which it is possible to establish the performance of a method through checking the relevance of similar genes to the original gene set. These gene sets were collected from four different resources – KEGG, OMIM, PanelApp and UKB PheWAS – leading to over 15 000 initial gene sets. Phenotypes in KEGG tend to reflect biological pathways, those in OMIM are taken from the biomedical literature, PanelApp are ascertained through consensus among experts and UKB PheWAS are determined through a collapsing analysis of ∼400 000 exomes. The vast majority of these gene sets, however, were redundant as we specifically focus on those with two or more genes between which Gene-SCOUT can provide similarities for, reducing this figure to 2,662 different gene sets over all of the resources. Performance was measured through calculating the overlap of the 'n' most similar genes to all other genes in a gene set and

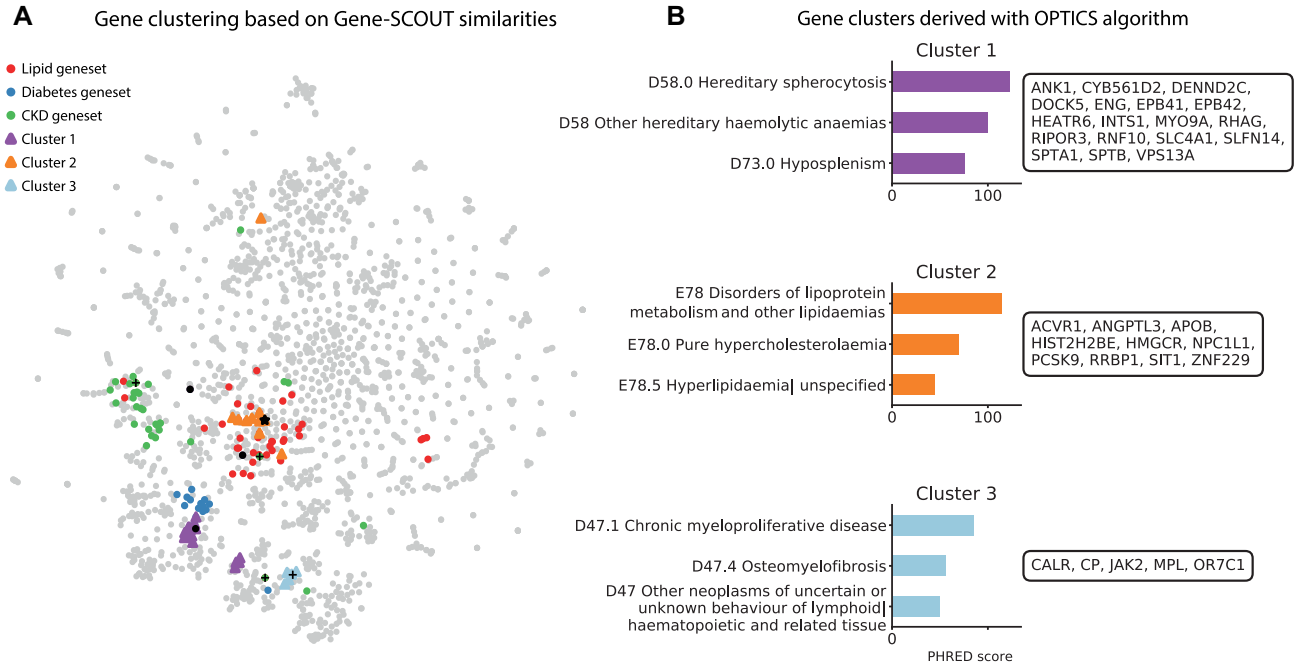

Figure 2. Gene-SCOUT scores validation using known manually curated gene sets as a gold standard. (**A**) t-SNE projection of similarities, with known gene sets (dyslipidemia, diabetes and chronic kidney disease associated genes) as well as top 3 OPTICS-derived clusters overlayed. Genes common to multiple sets are in black with (*, •, +) denoting membership in (all, dyslipidemia & CKD or dyslipidemia & diabetes) genes sets, respectively. (**B**) PheWAS enrichments for top performing clusters extracted with the OPTICS algorithm (clusters are ranked by the highest performing PHRED score).

quantified using Fisher's exact test (see Materials and Methods). To evaluate the relative performance of Gene-SCOUT compared to an alternative method, we compare methods only on gene sets that satisfy Fisher's exact P-value ≤ 0.05 for both Gene-SCOUT and an alternative method. This ensures only gene sets that reach some minimal level of overlap in both methodologies are included, helping to enable a fairer comparison. We see that the resulting number of gene sets for a comparison varies depending on the collection (OMIM, PanelApp etc.), the alternative method (GOSim, HPOSim etc.) and the value of 'n', with the median gene set size being 51 genes (Supplementary Figure S27).

Of these comparisons, we are then able to observe when the overlap was stronger in Gene-SCOUT compared to an alternative method, which we summarise as a fraction of the number of gene sets. A fraction of $>50\%$ indicates that on average, Gene-SCOUT is more likely to recover meaningful similarities compared to an alternative method. We see that in general (Figure 3A), Gene-SCOUT achieves this for gene sets arising from UKB PheWAS and OMIM phenotypes and to some extent for gene sets derived from KEGG pathways or PanelApp gene panels. To summarize, any comparison with Gene-SCOUT is performed on four different collections of gene sets and with four values of 'n', leading to 16 different settings. We saw that Gene-SCOUT outperforms GOSim, HPOSim and StringDB on 13, 11 and 7 of these settings, respectively.

We acknowledge a potential limitation of the previous benchmarking procedure, with regards to potential biases induced between similarity methods and gene sets serving as ground truths. In particular HPO (the data-source for HPOSim) employs data from OMIM, similarly, StringDB scores are derived using data from KEGG, GO and OMIM. In terms of UKB PheWAS, the same set of exome samples were used to estimate Gene-SCOUT similarities and also to estimate the phenotypic gene sets used to validate them. This may introduce an undesirable circularity into the validation, with any sample noise potentially affecting both the continuous trait associations in the signatures and phenotype associations used to construct the validation gene sets—moderated to some extent by the large sample size. As such, we eliminate any circularities and biases through splitting the samples into two buckets of ∼200 000 exomes, stratified by age and sex, to provide an unbiased validation. The first bucket was used to estimate Gene-SCOUT similarities while the second was used to identify gene sets to test against (see Materials and Methods). We do not alter other settings relating to how similarities are calculated in the process—e.g. the significance level used to determine relevant features in Gene-SCOUT. We see that Gene-SCOUT performs almost uniformly stronger compared to all other methods on this independent data set (Figure 3B). Here, we perform comparisons using 12 different settings: three PheWAS significance thresholds (10^{-4}, 10^{-5} and 10^{-8}) and four values of 'n'. Gene-SCOUT outperforms the alternative methods in 11, 12 and 12 of these settings (GOSim, HPOSim & StringDB, respectively). Also, it is noted that the seemingly stronger performance of Gene-SCOUT when similarities are estimated using fewer UKB exomes (Figure 3A and B) may in part be attributed to the fact that Gene-SCOUT's feature selection is potentially more stringent when using the same significance threshold on fewer

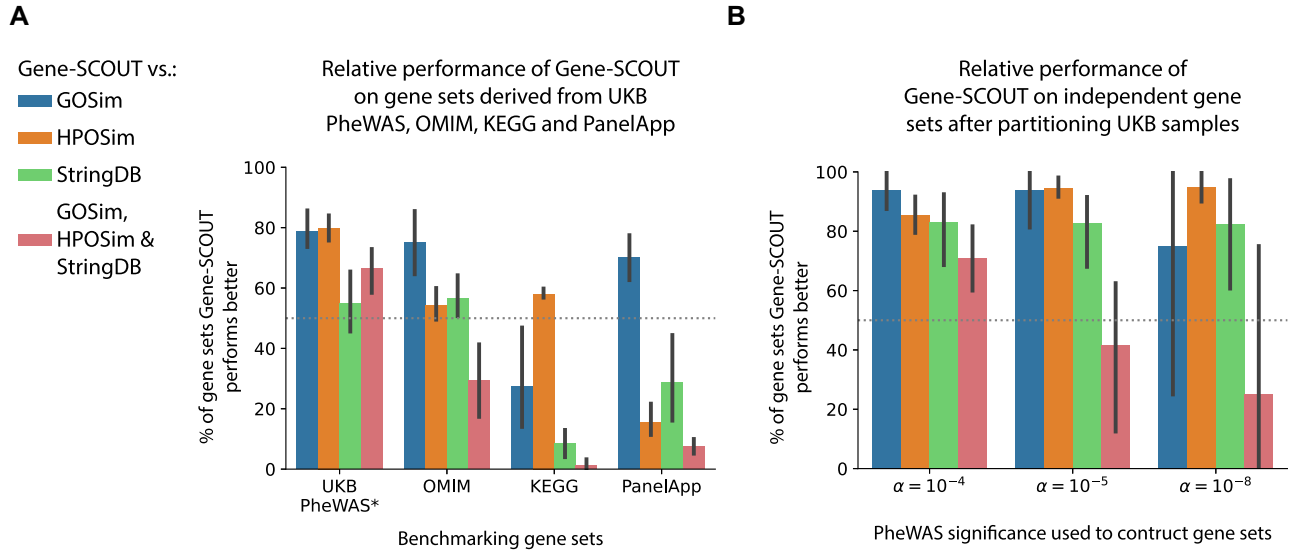

Figure 3. Comparison of Gene-SCOUT with alternative methods across different collections of gene sets. Bars represent the fraction of the gene sets Gene-SCOUT outperformed an alternative method (or collection of methods). Vertical ranges capture the performance while varying the 'n closest' genes in {5, 10, 20, 50}. (**A**) Gene sets provided by four different resources (*the same UKB cohort was used to obtain gene sets and estimate Gene-SCOUT similarities). (**B**) Gene sets derived from only PheWAS collapsing analysis of UKB data. Samples were randomly split into two subsets, one to estimate the Gene-SCOUT similarities and one to establish gene sets used in validation, thereby eliminating any risk of circularity.

samples, thereby increasing the genetic signal. This occurs at the cost of supplying similarities between a smaller number of genes (1695 compared to 2097 previously), as a more stringent threshold used in feature selection causes fewer genes to have any significant quantitative traits, and fewer still sharing the same ones. Finally, we are able to manually inspect the phenotypes that Gene-SCOUT attains the best performance over other methods (Supplementary Table S3), where we see strong performance on many lipid- and blood-related disorders.

Web-utility

We accompany our method with a rich web-resource capable of exploring the gene similarities. Similarities between genes have been pre-computed and all associated results are accessible through the web app. As a result, the resource provides a user-friendly interface for querying similarities across the whole exome, explore the enrichment of groups of similar genes for PheWAS and/or Gene Ontology terms and extract the continuous trait fingerprint underpinning each similarity association (Figure 4, Supplementary Figures S15 and S16).

The user provides a gene name based on HGNC nomenclature (24), after which the results page displays information related to genes that are similar to the input gene. Based on the ranked list of (at most) 10 genes plus the seed gene we are able to perform a number of additional analyses. Firstly, we are able to construct a network of neighboring genes. The network provides a visual representation of genes that are close to each other. We utilize a k-nearest neighbor graph (see Methods) to construct the network with $k = 10$, chosen based on a robustness analysis (Supplementary Methods, Supplementary Figure S18 and S19) and inspection of the distribution of number of similar genes to each seed gene (Supplementary Figure S11). Secondly, based on this list of nearest genes, we are able to quantitatively verify whether that collection of genes is enriched for established biological processes and/or clinical traits studied in the UK Biobank. In the former case, we utilise the automated Gene Ontology enrichment provided by the GOATools (25) package to test whether the list of closest genes is enriched for various biological processes. For the latter, we perform Fisher's exact test between genes that are significant for a given clinical trait and the list of nearest genes (Supplementary Methods). We repeat this for all previously studied clinical traits in the UK Biobank (2), thereby constructing a P-value score measuring the overlap between close genes and those that are enriched for a given clinical trait. Finally, we display the raw vector of significant associations and the charts that show the associations that are significant in both the seed gene and the gene that is close to it. As such, the user can manually inspect which continuous trait associations gave rise to the measure of similarity between any two genes, providing an additional degree of interpretability.

One of the main utilities of the web resource is in offering tractable alternatives to potentially intractable targets. We explore this through leveraging druggability annotation data from the PHAROS druggability resource (26) and the druggable genome work (27). We identify a set of 413 of Gene-SCOUT genes for which there is no supporting evidence of tractability (denoted as 'Tdark' by PHAROS) and use these as seed genes in Gene-SCOUT. We then compile a list of 416 likely tractable genes from the union of PHAROS ('Tclin' and 'Tchem') and the druggable genome ('Tier 1'). We see that over 50% of Gene-SCOUT genes annotated as difficult-to-drug retained at least one other druggable target gene within the closest 10 most similar genes (Supplementary Figure S21).

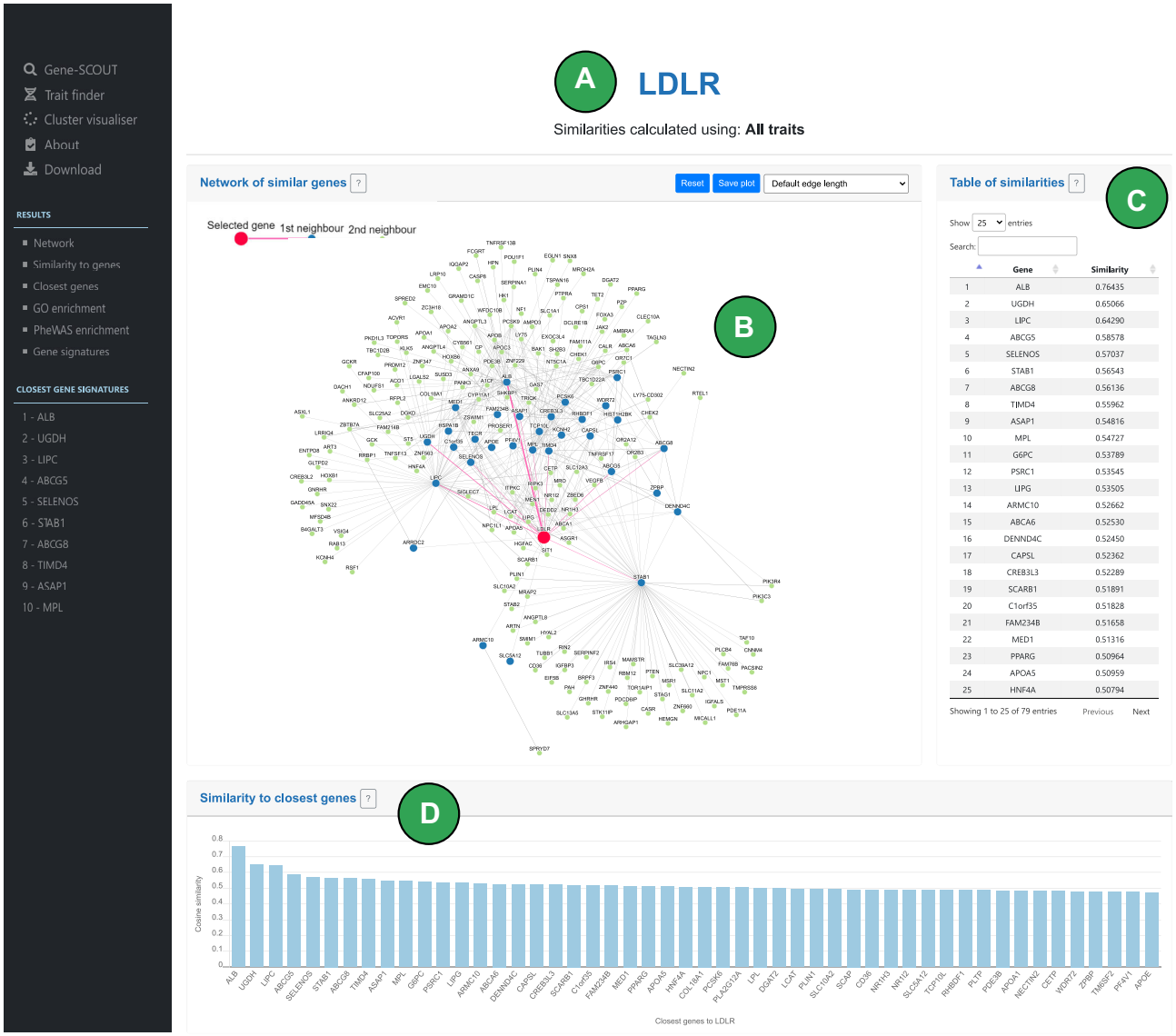

Figure 4. Screenshot of web-resource for Gene-SCOUT pre-calculated results. (**A**) Seed gene. (**B**) Gene network – this is populated through considering the (at most) 10 closest genes to each other gene (an edge in the network implies one of the genes was in the 10 closest for the other). The network only shows the 'two-hop' neighbors to the seed gene. (**C**) Table of genes closest to seed gene, ranked by Gene-SCOUT similarities. (**D**) Plot of most similar genes to seed gene.

Trait finder

In addition to this gene-specific information, we provide a tool to retrieve a list of genes that closely match a desired continuous trait fingerprint, where we consider only the directionalities of significant associations to construct the fingerprint (see Materials and Methods). In this way, it is possible to rank genes that most closely satisfy a desired continuous trait profile (Supplementary Figure S17). For example, clinicians and other researchers may wish to find genes that when mutated are negatively associated with a set of particular continuous traits while being positively associated with an alternative set. We refer to this tool as 'Trait finder' and comprises a separate section of the web resource.

We illustrate the technique, with the following example, illustrating Trait Finder (Figure 5A) in two example configurations. The first (Figure 5B) considers positive and negative traits as Glucose and LDL direct, respectively. It can be see that *GCK, GIGYF1* and *HNF1A* can all be matched to this particular signature. Such a result is consistent with recent findings, noting that *GIGYF1* mutations are associated with an increased risk of Type II Diabetes and produce protective effects in terms of hypocholesterolemia (28). Another setting (Figure 5C), comprising of Vitamin D as a positive trait and two negative traits (total cholines and LDL direct) reveals three matching genes again—*ANGPTL3, APOB* and *PCSK9*, all of which are known to be associated with cholesterol levels.

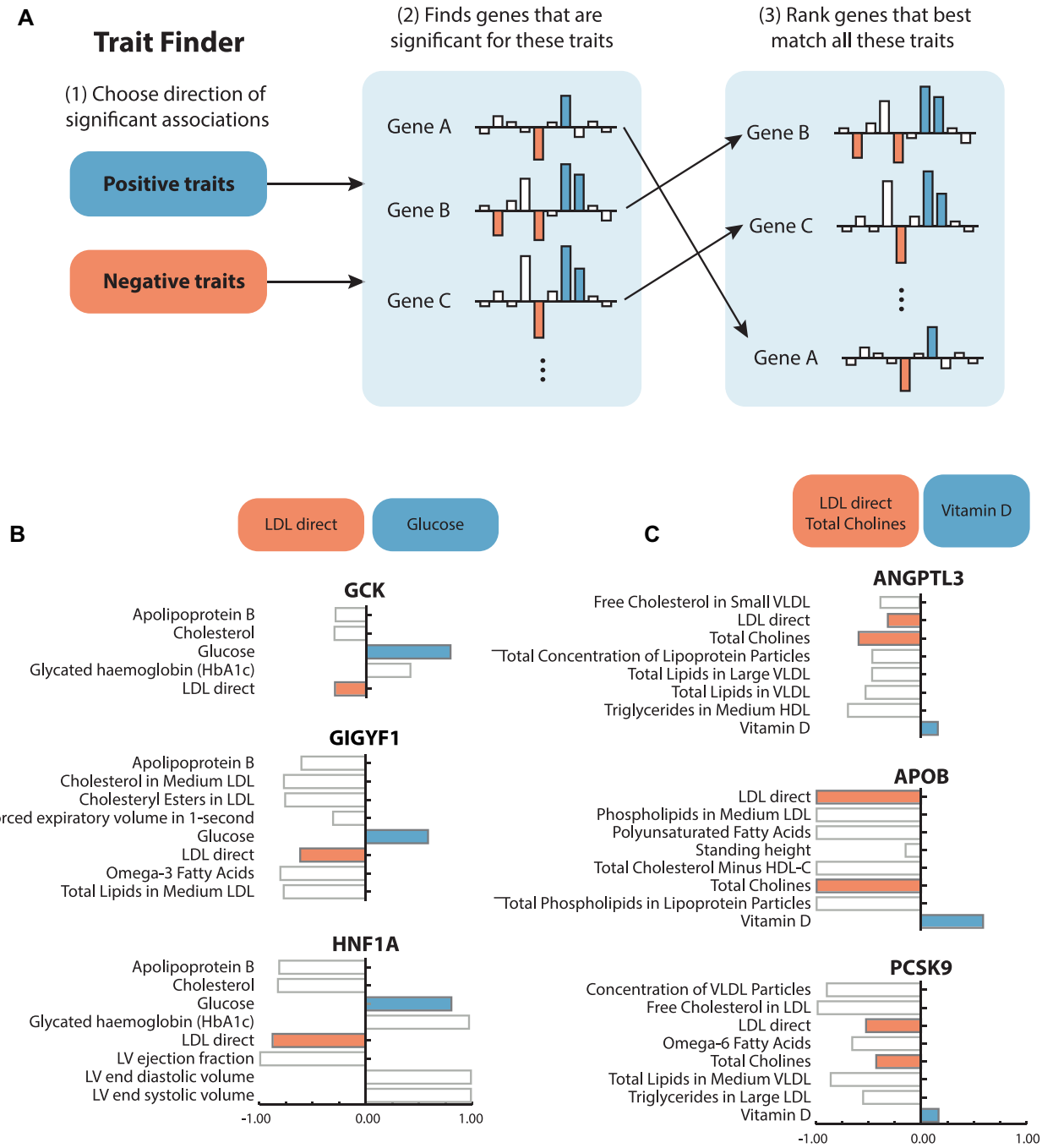

Figure 5. Schematic workflow of Gene-SCOUT's Trait finder utility. (**A**) Traits are first selected by the user based on their direction of association. A subset of genes is then identified that are significant in one or more of these traits. Finally, these genes are ranked so that genes with the most number of positive association with 'Positive traits' (and/or negative association for 'Negative traits') appear towards the top. (**B**) Example gene signatures returned by Trait Finder, using Glucose and LDL direct as the positive and negative traits. Three genes satisfied all these conditions (*GCK, GIGYF1, HNF1A*) at a significance level of 10^{-4} with a subset of each of their signatures shown (the matching traits plus a selection of other traits that were also significant at the same value). (**C**) The same process though with different positive (Vitamin D) and negative traits (Total Cholines, LDL direct) returns three matching genes: *ANGPTL3*, *APOB* and *PCSK9*. Signature coefficients are capped between –1 and 1 to aid with visualisation.

DISCUSSION

Estimating similarities between genes based on their fingerprint for biomarkers and other continuous traits can not only facilitate identifying alternative targets for less tractable index genes, but also provide a deeper understanding of the shared biomarker fingerprints for proof of mechanism and alike experimental medicine studies (29). Gene-SCOUT uses ∼400K exomes from the UK Biobank to develop a unique continuous trait fingerprint for each gene in the exome. Using this fingerprint, it computes similarities with all other genes in the exome based on statistical relationships with continuous traits. We demonstrated that gene similarity estimations from Gene-SCOUT provide stronger enrichments for clinical traits in comparison to other existing methods. We provide a web-resource capable of exploring the similarities calculated from this dataset, where the user is able to examine the biological relevance of the reportedly close genes in terms of GO enrichment and clinical trait enrichment analyses. The user is also able to explore the similarities through a fully interactive gene network, known as a *k*-nearest neighbor graph. We provide the user with a detailed breakdown of the traits used to construct each fingerprint. In parallel to this, we also offer functionality that lets the user visualise the similarities through a nonlinear projection and perform automated clustering of the genes based on the derived similarities, with ontology and trait-specific enrichments also provided for each cluster. Finally, we have introduced additional functionality ('Trait finder') that allows the user to search for genes that satisfy a certain user-defined fingerprint, as specified by the associations and their direction of effects for a set of continuous traits.

Further methodological extensions to the current approach could involve incorporating gene ontology and interactome-based similarity measures while also formulating it within a supervised machine learning framework to learn an optimal distance function between pairs of genes (30,31) based on an expert curated truth dataset.

DATA AVAILABILITY

Gene-SCOUT results are provided in a fully interactive web resource with pre-calculated gene similarities and networks, based on continuous trait fingerprints, which is freely available at: http://genescout.public.cgr.astrazeneca.com.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the participants and investigators in the UKB study who made this work possible (Resource Application Number 26041); the UKB Exome Sequencing Consortium (UKB-ESC) members AbbVie, Alnylam Pharmaceuticals, AstraZeneca, Biogen, Bristol-Myers Squibb, Pfizer, Regeneron and Takeda for funding the generation of the data and Regeneron Genetics Center for completing the sequencing and initial quality control of the exome sequencing data; the AstraZeneca Centre for Genomics Research Analytics and Informatics team for processing and analysis of sequencing data.

FUNDING

L.M.'s work was funded by the AstraZeneca post-doctorate program. Funding for open access charge: AstraZeneca.
Conflict of interest statement. L.M., A.R.H., A.N., Q.W., A.R., D.V. and S.P. are employees of AstraZeneca. A.R.H., A.N., Q.W., A.R., D.V. and S.P. are shareholders of AstraZeneca.

REFERENCES

1. Sudlow,C., Gallacher,J., Allen,N., Beral,V., Burton,P., Danesh,J., Downey,P., Elliott,P., Green,J., Landray,M. *et al.* (2015) UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.*, **12**, e1001779.
2. Wang,Q., Dhindsa,R.S., Carss,K., Harper,A.R., Nag,A., Tachmazidou,I., Vitsios,D., Deevi,S.V.V., Mackay,A., Muthas,D. *et al.* (2021) Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature*, **597**, 527–532.
3. Petrovski,S., Todd,J.L., Durheim,M.T., Wang,Q., Chien,J.W., Kelly,F.L., Frankel,C., Mebane,C.M., Ren,Z., Bridgers,J. *et al.* (2017) An exome sequencing study to assess the role of rare genetic variation in pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.*, **196**, 82–93.
4. Nightingale Health Plc (2021) Nightingale health metabolic biomarkers: phase 1 release. *UK Biobank Tech. Rep*.
5. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The Gene Oncology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
6. Mazandu,G.K., Chimusa,E.R. and Mulder,N.J. (2017) Gene Ontology semantic similarity tools: survey on features and challenges for biological knowledge discovery. *Brief. Bioinform.*, **18**, 886–901.
7. Fröhlich,H., Speer,N., Poustka,A. and Beißbarth,T. (2007) GOSim - an R-package for computation of information theoretic GO similarities between terms and gene products. *BMC Bioinformatics*, **8**, 166.
8. Deng,Y., Gao,L., Wang,B. and Guo,X. (2015) HPOSim: an R package for phenotypic similarity measure and enrichment analysis based on the human phenotype ontology. *PLoS One*, **10**, e0115692.
9. Köhler,S., Doelken,S.C., Mungall,C.J., Bauer,S., Firth,H. V., Bailleul-Forestier,I., Black,G.C.M., Brown,D.L., Brudno,M., Campbell,J. *et al.* (2014) The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res.*, **42**, D966–D974.
10. Szklarczyk,D., Franceschini,A., Wyder,S., Forslund,K., Heller,D., Huerta-Cepas,J., Simonovic,M., Roth,A., Santos,A., Tsafou,K.P. *et al.* (2015) STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, **43**, D447–D452.
11. Bycroft,C., Freeman,C., Petkova,D., Band,G., Elliott,L.T., Sharp,K., Motyer,A., Vukcevic,D., Delaneau,O., O'Connell,J. *et al.* (2018) The UK Biobank resource with deep phenotyping and genomic data. *Nature*, **562**, 203–209.
12. Szustakowski,J.D., Balasubramanian,S., Kvikstad,E., Khalid,S., Bronson,P.G., Sasson,A., Wong,E., Liu,D., Wade Davis,J., Haefliger,C. *et al.* (2021) Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. *Nat. Genet.*, **53**, 942–948.
13. Hu,L.Y., Huang,M.W., Ke,S.W. and Tsai,C.F. (2016) The distance function effect on k-nearest neighbor classification for medical datasets. *Springerplus*, **5**, 1304.
14. Pandit,S. and Gupta,S. (2011) A comparative study on distance measuring approaches for clustering. *Int. J. Res. Comput. Sci.*, **2**, 29–31.
15. Kumar,V., Chhabra,J.K. and Kumar,D. (2014) Performance evaluation of distance metrics in the clustering algorithms. *Infocomp*, **13**, 38–52.

16. Connor,M. and Kumar,P. (2010) Fast construction of κ-nearest neighbor graphs for point clouds. *IEEE Trans. Vis. Comput. Graph.*, **16**, 599–608.
17. Hajebi,K., Abbasi-Yadkori,Y., Shahbazi,H. and Zhang,H. (2011) Fast approximate nearest-neighbor search with k-nearest neighbor graph. In: *IJCAI International Joint Conference on Artificial Intelligence*. 1312–1317.
18. Vajda,S. and Santosh,K.C. (2017) A fast k-nearest neighbor classifier using unsupervised clustering. *Commun. Comput. Inf. Sci.*, **709**, 185–193.
19. Blondel,V.D., Guillaume,J.L., Lambiotte,R. and Lefebvre,E. (2008) Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.*, **2008**, P10008.
20. Carlson,M. (2019) org.Hs.eg.db: genome wide annotation for Human. *R Package version 3.8.2*.
21. Hamosh,A., Scott,A.F., Amberger,J., Valle,D. and McKusick,V.A. (2000) Online Mendelian Inheritance in Man (OMIM). *Hum. Mutat.*, **15**, 57–61.
22. Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.
23. Martin,A.R., Williams,E., Foulger,R.E., Leigh,S., Daugherty,L.C., Niblock,O., Leong,I.U.S., Smith,K.R., Gerasimenko,O., Haraldsdottir,E. *et al.* (2019) PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat. Genet.*, **51**, 1560–1565.
24. Tweedie,S., Braschi,B., Gray,K., Jones,T.E.M., Seal,R.L., Yates,B. and Bruford,E.A. (2021) Genenames.org: the HGNC and VGNC resources in 2021. *Nucleic Acids Res.*, **49**, D939–D946.
25. Klopfenstein,D. V., Zhang,L., Pedersen,B.S., Ramírez,F., Vesztrocy,A.W., Naldi,A., Mungall,C.J., Yunes,J.M., Botvinnik,O., Weigel,M. *et al.* (2018) GOATOOLS: a Python library for Gene Ontology analyses. *Sci. Rep.*, **8**, 10872.
26. Nguyen,D.T., Mathias,S., Bologa,C., Brunak,S., Fernandez,N., Gaulton,A., Hersey,A., Holmes,J., Jensen,L.J., Karlsson,A. *et al.* (2017) Pharos: collating protein information to shed light on the druggable genome. *Nucleic Acids Res.*, **45**, D995–D1002.
27. Hopkins,A.L. and Groom,C.R. (2002) The druggable genome. *Nat. Rev. Drug Discov.*, **1**, 727–730.
28. Deaton,A.M., Parker,M.M., Ward,L.D., Flynn-Carroll,A.O., BonDurant,L., Hinkle,G., Akbari,P., Lotta,L.A., Baras,A. and Nioi,P. (2021) Gene-level analysis of rare variants in 379,066 whole exome sequences identifies an association of GIGYF1 loss of function with type 2 diabetes. *Sci. Rep.*, **11**, 21565.
29. Perlman,L., Gottlieb,A., Atias,N., Ruppin,E. and Sharan,R. (2011) Combining drug and gene similarity measures for drug-target elucidation. *J. Comput. Biol.*, **18**, 133–145.
30. Yang,L. and Jin,R. (2006) In: *Distance Metric Learning: A Comprehensive Survey*. Michigan State Universiy, https://doi.org/10.1073/pnas.0809777106.
31. Xing,E.P., Ng,A.Y., Jordan,M.I. and Russell,S. (2003) Distance metric learning, with application to clustering with side-information. In: *Advances in Neural Information Processing Systems*. pp. 521–528.
32. Mi,H., Ebert,D., Muruganujan,A., Mills,C., Albou,L.P., Mushayamaha,T. and Thomas,P.D. (2021) PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res.*, **49**, D394–D403.