PrismExp: Predicting Human Gene Function by Partitioning Massive RNA-seq Co-expression Data

Alexander Lachmann¹✉, Kaeli Rizzo¹, Alon Bartal¹, Minji Jeon¹, Daniel J. B. Clarke¹, and Avi Ma'ayan¹✉

¹ Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1603, New York, NY 10029, USA

Gene co-expression correlations from mRNA-sequencing (RNA-seq) can be used to predict gene function based on the covariance structure that exists within such data. In the past, we showed that RNA-seq co-expression data is highly predictive of gene function and protein-protein interactions. We demonstrated that the performance of such predictions is dependent on the source of the gene expression data. Furthermore, since genes function in different cellular contexts, predictions derived from tissue-specific gene co-expression data outperform predictions derived from cross-tissue gene co-expression data. However, the identification of the optimal tissue type to maximize gene function predictions for all mammalian genes is not trivial. Here we introduce and validate an approach we term Partitioning RNA-seq data Into Segments for Massive co-EXpression-based gene function Predictions (PrismExp), for improved gene function prediction based on RNA-seq co-expression data. With co-expression data from ARCHS4, we apply PrismExp to predict a wide variety of gene functions, including pathway membership, phenotypic associations, and protein-protein interactions. PrismExp outperforms the cross-tissue co-expression correlation matrix approach on all tested domains. Hence, PrismExp can enhance machine learning methods that utilize RNA-seq co-expression correlations to impute knowledge about understudied genes and proteins.

Correspondence: avi.maayan@mssm.edu; alexander.lachmann@mssm.edu

Introduction

The automated annotation of gene function can be derived by combining prior knowledge manual annotations with mRNA co-expression patterns based on the concept of guilt-by-association. In general, it is well established that genes that co-express display functional relatedness (1–3). Proteins with high mRNA expression correlations, or high mutual information (4), are often found in the same macro-molecular complexes, cell signaling and metabolic pathways (5, 6). Since proteins predominantly function through the formation of protein complexes and regulatory networks, their activity depends on their expression in different cellular and tissue contexts. Tight control of protein complex stoichiometry is needed for their correct formation (7). While all cells and tissues carry out general housekeeping processes, cells and tissues also have to fulfill specialized roles and subsequently differ in their pathway and biological function composition. Due to their cell and tissue specificity, considering gene regulatory networks in the right cellular and tissue context is im-

portant (8). Reverse engineering of regulatory mechanisms with tissue agnostic approaches lacks fidelity to capture nuanced gene-gene and protein-protein interactions accurately. With the availability of RNA-seq data from resources such as ARCHS4 (9), Genotype-Tissue Expression (GTEx) (10), or The Cancer Genome Atlas (TCGA) (11) featuring cell type and tissue specific gene expression information, it is feasible to leverage the context specific gene expression patterns for improved gene function predictions.

Over decades of biochemistry, pharmacology, molecular and cell biology research, many mammalian genes have been thoroughly characterized. However, there are still many genes that remain obscured due to a lack of research interest. Extensive gene function annotations can be accessed from resources such as the Gene Ontology (12), Human Phenotype Ontology (13), UniProt (14), Harmonizome (15) and Enrichr (16). The Gene Matrix Transposed (GMT) file format, used by Gene Set Enrichment Analysis (GSEA) tools (17) provides a convenient way to organize knowledge about gene-function associations. Each row in a GMT file corresponds to a biological function, followed by a set of associated genes. The assumption is that these gene-sets are generally incomplete, reflecting our partial knowledge about the involvement of genes and their protein products in biological processes.

Computational methods that can utilize gene function annotations to impute the function of understudied genes are an efficient way to form novel hypotheses that can lead to new discoveries. High-content genome-wide data, such as RNA-seq, includes information about under-studied genes and is not biased towards well-studied genes. Combining prior knowledge gene annotations with high-content genome-wide data can translate existing knowledge about well-studied genes to the under-studied genes. Significant efforts are made to produce data and develop tools to shed light on these understudied genes (18–20). One such effort is the NIH Common Fund Program Illuminating the Druggable Genome (IDG) (21). IDG focuses on understudied genes from three prominent gene families: protein kinases, G-protein coupled receptors (GPCRs), and ion channels. These families contain the most promising druggable protein targets with many understudied members (22).

In the past, we have developed methods to predict gene function for all human genes, including members of these three families, using co-expression data that is tissue agnostic. It

is reasonable to expect that by considering the tissue context of co-expression data the quality of such predictions will improve. However, the utilization of tissue specific gene expression patterns poses practical limitations:

- I. When making predictions about a specific biological function, the correct context of the particular function has to be considered. With potentially hundreds of relevant tissues to choose from, tissue specific predictions require significant domain knowledge.
- II. While some gene expression resources are well organized into individual tissues, for example, GTEx (10) or TCGA (23), these resources only cover a fraction of all human tissues and cell types. On the other hand, more diverse datasets such as ARCHS4 (9) lack accurate tissue classification of individual samples.
- III. Gene function prediction algorithms that utilize co-expression data are applied to predict thousands of diverse biological functions. Thus, manual input for individual cases is not feasible.

To address the above limitations, while still leveraging the advantage of tissue specific gene function prediction using co-expression patterns, we developed a novel machine learning approach called Partitioning RNA-seq data Into Segments for Massive co-EXpression-based gene function Predictions (PrismExp). PrismExp generates a high dimensional feature space from large unlabeled RNA-seq mRNA gene expression data. The generated feature space automatically encodes tissue specific information via vertical partitioning of the data matrix. We demonstrate that this approach significantly improves gene function predictions when compared to methods that do not partition the data matrix.

Materials and Methods

A. Data resources. To compute pairwise gene-gene mRNA co-expression correlations, we used the mouse ARCHS4 RNA-seq gene expression dataset. We chose mouse over human because benchmarking (9) demonstrated that predictions made with the mouse co-expression data slightly outperformed predictions derived from human co-expression data. The ARCHS4 mouse dataset is comprised of 284,907 uniformly aligned RNA-seq gene expression samples publicly available from GEO (24). For benchmarking, we utilized all 163 available gene-set libraries from Enrichr (16). For more detailed analysis, we used a subset of the gene-set libraries from Enrichr (Table 1). In total, the collection of the gene-sets in these five libraries contain 19,906 gene functions.

B. Generating the gene-gene mRNA co-expression matrices. We first filter out genes with low gene expression, resulting in G , a set of genes with expression values across samples. The threshold to include genes in G requires a gene to have at least 20 reads in 1% of all samples. This filtering step is intended to remove pseudo-genes and genes with

very low levels of transcript abundance. PrismExp automatically identifies a set of N clusters defined as C using a k-means clustering algorithm. For each cluster defined by the k-means clustering C , we apply quantile normalization on the log2 transformed gene counts to correct for library size variability. For each resulting normalized gene expression matrix, we calculate the pairwise gene-gene correlations for genes in G , C_i for cluster i using Pearson correlation. Additionally, we calculate the Pearson correlation coefficients between genes across a random selection of 10,000 samples, C_{global} . The set of correlation coefficient matrices is: $C = \{C_0, C_1, \dots, C_k, C_{global}\}$.

C. Computing gene $\times$ gene-set features. Gene $\times$ gene-set features are scores for the strength of association between a gene and a gene-set pair. For a gene $g_x \in \{g_1, \dots, g_L\}$ and a gene-set $p_y \in \{p_1, \dots, p_M\}$ the algorithm generates a set of features $\tilde{S}_k$ used in the final prediction step. A gene-set library $S \in L \times M$ is defined as a binary matrix with genes as rows and gene-sets as columns.

$$S_{(g_x, p_y)} = \begin{cases} 1, & \text{if } g_x \in p_y \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Given N gene co-expression matrices $C_k \in \{C_1, \dots, C_N\}$ and a gene-set library S , the corresponding gene-function association features $\tilde{S}_{(C_k, g_x, p_y)}$ are defined as,

$$\tilde{S}_{(C_k, g_x, p_y)} = \frac{\sum_{i=1, i \neq x}^N S_{(g_x, g_i)} \times C_{k(g_i, p_y)}}{\sum_{i=1, i \neq x}^N C_{k(g_i, p_y)}} \quad (2)$$

Alternatively, a gene-set library can also be defined as a non-binary adjacency matrix. Given N co-expression matrices, each pair of gene and gene-set/function has a set of features,

$$F_{(g_x, p_y)} = \{\tilde{S}_{(C_1, g_x, p_y)}, \dots, \tilde{S}_{(C_N, g_x, p_y)}\}. \quad (3)$$

D. Training a Random Forest regression model. PrismExp takes F_{g_x, p_y} as input features to classify whether g_x should be annotated with p_y as a functional association. Support Vector Machine (SVM), artificial neural networks (ANN), and other machine learning algorithms are suitable to apply for constructing predictive models in this context. PrismExp utilizes a Random Forest regression model (30) to accomplish this task. We also tested several other machine learning models (Support Vector Machine, Multi-layer Perceptron regressor, logistic regression, XGBoost, Gaussian Naive Bayes, k-Nearest Neighbor, and Linear Discriminant Analysis) (31–37) and all performed comparable. The model is trained with the following data,

$$T = \{(x_1 = F_{(g_{x_1}, p_{y_1})}, y_1 = S_{(g_{x_1}, p_{y_1})}), \dots, (x_n = F_{(g_{x_n}, p_{y_n})}, y_n = S_{(g_{x_n}, p_{y_n})})\} \quad (4)$$

E. Python package and web interface. PrismExp is available as a Python package. The package supports programmatic prediction of gene function from gene expression

GMT		Description	Literature	Sets	Genes
ChEA 2016	(25)	Transcription factor targets from ChIP-seq experiments	✗	645	17022
KEA 2015	(26)	Kinase-substrate interactions	✗	428	2904
MGI phenotype	(27)	Gene-phenotype associations from phenotype screens	✗	5261	10390
hu.MAP	(28)	Protein-protein interactions	✓	995	2117
GO: Biological Process	(12)	Manually curated ontology of gene function	✓	5103	12096
GWAS Catalog 2019	(29)	Manually curated SNP-trait associations from literature	✓	1737	10744

Table 1. Gene-set libraries from Enrichr with the number of gene-sets and unique genes. Literature based gene-set libraries are manually curated and non-literature based libraries are derived from directly from experimental outcomes.

Fig. 1. PrismExp workflow for gene function prediction.

data utilizing the data from ARCHS4 (9) stored in HDF5 format (38). Predictions can be made for any gene-set library. The PrismExp package also includes gene-set libraries to use directly from the collection of libraries available from Enrichr (16). The PrismExp Python package offers parameter customization, for example, the user can adjust the number of gene expression clusters C . In addition to the PrismExp Python package, PrismExp is also available as an Appyter (<https://appymaayanlab.cloud/PrismEXP/>). Appyters convert Jupyter Notebooks to web-based bioinformatics applications. The PrismExp Appyter enables users to execute the PrismExp computational pipeline using their own data in the cloud without coding. The PrismExp Appyter has access to 51 pre-computed correlation matrices stored in Amazon Web Services (AWS) S3. The Appyter implementation of PrismExp directly extracts the required correlation data from the pre-computed matrices. A user interface enables the selection of a gene of interest, as well as the selection of any gene-set library for target predictions. Alternatively, the user can upload their own custom gene-set library in GMT format.

F. Validation. To assess the quality of the predictions made by PrismExp, we asked how well the prediction method can retrieve the known biological functions of a gene. The PrismExp prediction output is a matrix with genes as rows, and biological functions as columns. The higher a value in the matrix, the more likely the corresponding gene is associated with the respective biological function. Thus, there are two ways to assess the quality of the predictions made by PrismExp:

- Rank the values of each row in descending order

- Rank the values of each column in descending order

Ranking the values for each row is equivalent of ranking gene functions for each gene from likeliest to least likely. Since there is prior knowledge about the associations of genes with biological functions, these incidences are marked as positives. These positives are expected to be highly ranked. Biological functions that were not previously associated with a gene are considered false, and are expected to have lower ranks. PrismExp calculates an area under the receiver operating characteristic (AUROC) curve for each gene. These AUROC values are used to compare the quality of the predictions by different methods, model parameters, and underlying data sets.

Alternatively, we can rank the genes for each biological function by the most likely to the least likely. Similarly, we can calculate the AUROCs by labeling the previously known associations. Both sets of AUROC scores, ranking gene for biological function and ranking biological functions for genes, are critical to consider when evaluating and validating various sources of data and model parameter settings.

Results

G. Gene function prediction workflow. PrismExp produces predictions about gene-function associations at the genome-wide scale. The predictions are based on prior knowledge about gene-function associations and gene-gene co-expression data. The PrismExp workflow is made of two main branches: initialization and prediction 1. During initialization, PrismExp generates the required correlation matrices and pre-trains a Random Forest regression model. The prediction step takes as input a gene-set library stored in

Gene Matrix Transposed (GMT) format, and uses the pre-computed correlation matrices and regression model to make predictions. To make a prediction, PrismExp calculated the average correlation of a gene with a gene-set for each correlation matrix. The resulting feature vector is used as the input for the Random Forest model. As an optional step, PrismExp can validate the prediction performance for a given biological function, or for a specific gene using AUROCs.

H. Performance and hardware requirements. PrismExp is optimized to run on standard desktop computers utilizing multiple cores for the clustering and correlation calculations. Since PrismExp is using large matrices, memory consumption should be considered. For a realistic use-case, using the ARCHS4 mouse gene co-expression dataset, the memory requirements are 8GB. PrismExp was benchmarked using an Intel i7-8750H with 6 cores and 32GB of RAM using Python 3.8.5. PrismExp was also tested on an AWS r5ad.2xlarge instance with 8 cores and 64GB of RAM running Ubuntu. Both tests produce comparable performance (Figure S1). The memory consumption and compute time from the five different stages of PrismExp execution, training it with 50 gene expression clusters, shows that gene filtering and sample clustering is completed in 50 minutes. Calculating the correlation matrices takes the longest time, 122 minutes or 2.4 minutes per cluster. Reducing or increasing the number of clusters affects run-time linearly based on the number of clusters. The prediction step depends on the number of genes, the number of gene-sets, and the average gene-set size of the gene-set library. This is in addition to the number of gene expression clusters. This is determined based on prediction and training applied using the Gene Ontology Biological Processes 2018 library from the Enrichr gene-set library catalog. This gene-set library contains 5,103 gene-sets, 14,433 genes, and has an average of 36 genes per set. Training the Random Forest model with 40,000 positive samples and 200,000 negative samples take about 50 minutes to complete. The steps shown above have to be performed only once to initialize the prerequisite data for making multiple predictions. Custom gene-set libraries can be used to train the Random Forest regression model. The only step that needs to be performed each time for each gene-set library is the prediction step, which calculates the average correlation of a gene to all gene-sets in the library for each correlation cluster.

The number of gene expression clusters, and the resulting co-expression correlation matrices, impact the final prediction quality. We compared the prediction quality for 5, 10, 25, 50, 100, and 300 clusters (Figure 2). These results are compared to predictions made with the global correlation matrix that is not partitioned into clusters. Overall, we observe a significant improvement in the average AUROCs for using the partitioning method over the global co-expression correlation matrix. While the performance of the global correlation matrix outperforms PrismExp for low numbers of clusters, PrismExp significantly beats the prediction quality made by the global correlation matrix when more clusters are considered. We applied a paired sample Student t-test on the average AUROCs from 163 gene-set libraries to compare the perfor-

mance of a single correlation matrix to applying PrismExp with 301 clusters. The results confirm significant improvement for using the partitioning approach, $p = 1.401e^{-37}$ and $p = 2.397e^{-31}$ for sets and genes respectively. The increase in average AUROCs achieved by a higher cluster count can be represented roughly by a logarithmic function of the shape $f(x) = a \times \log_2(bx) + c$ 2. This fitting suggests that the number of clusters is logarithmically proportional to the average AUROCs. Increasing the number of clusters beyond 300 only marginally improves the predictions while increasing the computation time considerably.

By projecting the high dimensional feature-space of PrismExp using t-SNE (39) we observe distinct clusters (Figure 3). In these plots, each point represents a gene $\times$ gene-set pair. Gene $\times$ gene-set pairs with prior known associations are organized into a gradient of average correlation from low on the left to high on the right (Figure 3 b). A similar gradient from left to right is observed for the PrismExp predictions, following the true positives sample distribution (Figure 3 c). The features derived from the 301 co-expression correlation matrices, in combination with the Gene Ontology gene-set library as the training set, show that individual gene co-expression matrices contribute to the performance of the PrismExp model (Figure 4 a and b). The feature importance of each co-expression correlation matrix is related to the prediction performance when using the average correlation of the correlation matrix alone. Co-expression correlation matrices that contribute a high level of feature importance also perform better individually than matrices with low feature importance contribution. The global co-expression correlation matrix computed from 10,000 randomly selected samples extracted from the ARCHS4 resource has the highest single matrix performance and the highest feature importance contribution.

It is expected that a model trained with a specific gene-set library will produce high prediction performance when predicting gene functions from that same library. However, we next asked how well a model trained with one gene library can predict terms from another library (Figure 4 c). A heatmap that compares average AUROCs illustrates that the six PrismExp models trained on one gene-set library listed in 1 can be used to predict terms in other libraries. The prediction performance of a model trained on the same gene-set library is always the best, as expected. For example, a model trained on the hu.MAP gene-set library achieves an average gene AUROC of 0.98 when applied to make predictions for the same gene-set library. However, this model does not achieve good predictions for any of the other tested gene-set library. The results suggest that the hu.MAP gene-set library contains different co-expression patterns than the other libraries, and thus do not generalize gene function predictions across a diverse set of libraries. However, the model trained with GO Biological Processes (BP) achieves good prediction performance when used to predict other libraries' functions. The GO-BP model also outperforms the single matrix predictions using all tested gene-set libraries, suggesting that the GO-BP model is generalizable across multiple gene-set

Fig. 2. a) Average AUROCs of gene-sets for 163 gene-set libraries. **b)** Average AUROC of gene-sets and predicted AUROC performance by the number of gene expression clusters. **c)** Average AUROCs of genes for the 163 gene-set libraries. **d)** Average AUROC of gene-sets and predicted AUROC performance by co-expression clusters.

Fig. 3. a) t-SNE visualization of 301 dimensional feature-space of 17,000 gene $\times$ gene-set pairs. **b)** Distribution of gene $\times$ gene-set pairs with known association (true positives). **c)** Average correlation scores of gene $\times$ gene-set pairs relative to the high dimensional feature space. **d)** PrismExp scores for gene $\times$ gene-set pairs relative to the high dimensional feature space.

libraries S2.

I. Predicting under-studied druggable targets with PrismExp.

The Illuminating the Druggable Genome (IDG) project is focused on furthering our knowledge about understudied potential protein targets from the ion channel, kinase, and GPCR gene families (40). Here we applied PrismEXP to predict the gene functions for one understudied member from each of these three drug-target-rich families. The most significant alpha-kinase 3 (ALPK3) associations are illustrated in a network diagram that connects ALPK3 with the top predicted GO terms and other relevant genes (Figure 5a). To date, there are only 20 publications that mention ALPK3. One of these, reported that biallelic loss of function of this gene is implicated in early-onset cardiomyopathy (41). The top gene functions predicted for ALPK3 by PrismExp are skeletal muscle contraction ($p=1.244e-15$), myosin filament assembly ($p=2.8701e-14$), cardiac muscle hypertrophy ($p=3.526494e-13$), cardiac muscle fiber development ($p=8.0835e-13$), and striated muscle contraction ($p=1.0353e-12$). In addition, ALPK3 is highly related to well-studied genes that form the sarcomere, namely TTN, MYOM1, MYOM2, and MYOM3. Hence, the understudied kinase ALPK3 has an important role in cardiac function that is highly under-appreciated. For GPR183, Prism

Exp predicts involvement in immune system-related processes. The top 5 terms are B-Cell activation ($p=9.5052e-15$), lymphocyte chemotaxis ($p=3.8298e-09$), positive regulation of B-Cell proliferation ($p=4.683e-09$), T-Cell differentiation ($p=5.7275e-09$), and T-Cell activation ($p=1.1794e-08$). GPR183 highly associated genes are the immune-response cytokines IL2, IL4, and CCL19 (Figure 5b). The predicted associated kinases for GPR183 are the non-receptor tyrosine kinases LCK, TXK, BLK, and IT; suggesting a direct or an indirect regulatory role for tyrosine phosphorylation to regulate the function of this GPCR. Finally, for the ion channel CLIC5, PrismExp predicts roles in inner ear receptor stereocilium organization ($p=3.285e-20$), sensory perception of mechanical stimulus ($p=1.7546e-07$), regulation of systemic arterial blood pressure ($p=1.3696e-06$), vasoconstriction ($p=1.1852e-05$), and positive regulation of vasculogenesis ($p=2.4557e-05$) (Figure 5c). Loss of CLIC5 has been associated with deafness (42). Hence, PrismExp is highly predictive for illuminating knowledge about under-studied genes and proteins.

Summary

Here we demonstrated that we could improve the inference of gene function by automatically partitioning the ARCHS4 gene-gene co-expression correlation matrix. Such automated

Fig. 4. a) Feature importance of individual correlation matrices relative to their prediction performance of ranking genes based on predicted known associations with biological functions. **b)** Feature importance of individual correlation matrices relative to their prediction performance in ranking gene-sets based on their associations with genes. **c)** PrismExp Random Forest model prediction performance evaluated as the average gene AUROC when trained on six gene-set libraries. The plot also includes prediction performance when using the global single correlation matrix.

partitioning is likely positively correlated with tissue and cell type partitioning but without the need for metadata labeling. The method treats each gene expression cluster as a source for features to construct machine learning models for gene function prediction. This approach deviates from prior work that required correct labeling of the gene expression profiles, for example, by text mining the metadata (43). Another advantage of PrismExp is that the RNA-seq data is available for all genes. Thus, it is suitable for predicting biological functions for understudied genes, including non-coding genes. The predictions performed by PrismEXP can be used for hypothesis generation and guidance in the design of wet-bench experiments. Specifically, PrismExp generates ranks for genes associated with biological functions, or ranks of biological functions for genes, facilitating many use cases. If a gene of interest is already associated with a biological function, a phenotype, or a disease, PrismExp can be used to highlight other genes with similar profile.

As with all machine learning applications, there is a risk of over-fitting the PrismExp model to patterns embedded within one specific gene-set library used for training. Hence, we aimed to evaluate how generalizable the models created using one library are to make predictions of known gene functions from other libraries. We confirmed that the models are, in general, generalizable. In addition, we observed a significant improvement in the predictions for hu.MAP and KEA due to the partitioning the samples for these libraries. Further research should be able to discern why such significant improvement is observed for these datasets. One postulation is that protein-protein interactions and kinase-substrate phosphorylations are context specific.

In this study, all predictions were made based on features that are derived from RNA-seq mRNA co-expression data. However, other omics data modalities could be incorporated into the same framework. By adding other sources of gene-gene similarity, additional features can be added to the model. We have explored the use of other gene similarity matrices for gene function prediction for other applications. For example, for developing Geneshot (18), we predicted gene functions

with features from gene-gene co-mentions in the literature, and co-occurrence in user-uploaded gene-sets to Enrichr (16). However, we never attempted to partition the gene-gene similarity matrices from these resources.

Finally, to make the PrismExp accessible to the community, we developed a Python package that enables computational biologists access the PrismExp method. In addition, we also provide access to the predictions that are made by PrismExp via an Appyter. The Appyter provides access to gene function predictions for users without programming skills. In the future, we plan to add the PrismExp predictions to the ARCHS4 (9) gene pages.

ACKNOWLEDGEMENTS

This work was partially supported by the National Institutes of Health (NIH) grants U54-HL127624 (LINCS-DCIC) and U24-CA224260 (IDG-KMC) to AM.

Bibliography

1. Jung Kyoong Choi, Ungsik Yu, Oon Joon Yoo, and Sangsoo Kim. Differential coexpression analysis using microarray data and its application to human cancer. *Bioinformatics*, 21(24): 4348–4355, 2005.
2. Kevin Horan, Charles Jang, Julia Bailey-Serres, Ron Mittler, Christian Shelton, Jeff F Harper, Jian-Kang Zhu, John C Cushman, Martin Gollery, and Thomas Girke. Annotating genes of known and unknown function by large-scale coexpression analysis. *Plant Physiology*, 147(1):41–57, 2008.
3. Ugo Ala, Rosario Michael Piro, Elena Grassi, Christian Damasco, Lorenzo Silengo, Martin Oti, Paolo Provero, and Ferdinando Di Cunto. Prediction of human disease genes by human-mouse conserved coexpression analysis. *PLoS computational biology*, 4(3): e1000043, 2008.
4. Alexander Lachmann, Federico M Giorgi, Gonzalo Lopez, and Andrea Califano. Aracneap: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics*, 32(14):2233–2235, 2016.
5. Koh Aoki, Yoshiyuki Ogata, and Daisuke Shibata. Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant and Cell Physiology*, 48(3):381–390, 2007.
6. Dan Zhang, Yan Guo, and Ni Xie. Prognostic value and co-expression patterns of metabolic pathways in cancers. *BMC genomics*, 21(11):1–9, 2020.
7. JM Kirkpatrick, M Mazzetto, M Baumgart, A Bartolome, S Di Sanzo, C Caterino, M Sanguanini, N Papaevgeniou, M Lefaki, D Childs, et al. Reduced proteasome activity in the aging brain results in ribosome stoichiometry loss and aggregation. *Molecular Systems Biology*, 16(6):e9596–e9596, 2020.
8. Abhijeet Rajendra Sonawane, John Platig, Maud Fagny, Cho-Yi Chen, Joseph Nathaniel Paulson, Camila Miranda Lopes-Ramos, Dawn Lisa DeMeo, John Quackenbush, Kimberly Glass, and Marieke Lydia Kuijjer. Understanding tissue-specific gene regulation. *Cell reports*, 21(4):1077–1088, 2017.
9. Alexander Lachmann, Denis Torre, Alexandra B Keenan, Kathleen M Jagodnik, Hoyjin J Lee, Lily Wang, Moshe C Silverstein, and Avi Ma'ayan. Massive mining of publicly available rna-seq data from human and mouse. *Nature communications*, 9(1):1366, 2018.
10. GTEx Consortium et al. The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

Fig. 5. Top 5 predicted Gene Ontology Biological Processes (orange) and most associated other genes (purple) for understudied targets (yellow) **a)** The protein kinase ALPK3, **b)** The GPCR GPR183, and **c)** the ion channel CLIC5. Green nodes represent predicted kinases for GPR183.

- John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.
- Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25, 2000.
- Sebastian Köhler, Nicole A Vasilevsky, Mark Engelstad, Erin Foster, Julie McMurtry, Ségolène Aymé, Gareth Baynam, Susan M Bello, Cornelius F Boerkoel, Kym M Boycott, et al. The human phenotype ontology in 2017. *Nucleic acids research*, 45(D1):D865–D876, 2017.
- UniProt Consortium. Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, 47(D1):D506–D515, 2019.
- Andrew D Rouillard, Gregory W Gunderson, Nicolas F Fernandez, Zichen Wang, Caroline D Monteiro, Michael G McDermott, and Avi Ma'ayan. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database*, 2016, 2016.
- Maxim V Kuleshov, Matthew R Jones, Andrew D Rouillard, Nicolas F Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L Jenkins, Kathleen M Jagodnik, Alexander Lachmann, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, 44(W1):W90–W97, 2016.
- Aravind Subramanian, Heidi Kuehn, Joshua Gould, Pablo Tamayo, and Jill P Mesirov. Gseap: a desktop application for gene set enrichment analysis. *Bioinformatics*, 23(23):3251–3253, 2007.
- Alexander Lachmann, Brian M Schilder, Megan L Wojciechowicz, Denis Torre, Maxim V Kuleshov, Alexandra B Keenan, and Avi Ma'ayan. Geneshot: search engine for ranking genes from arbitrary text queries. *Nucleic acids research*, 47(W1):W571–W577, 2019.
- Christopher Y Park, Aaron K Wong, Casey S Greene, Jessica Rowland, Yuanfang Guan, Lars A Bongo, Rebecca D Burdine, and Olga G Troyanskaya. Functional knowledge transfer for high-accuracy prediction of under-studied biological processes. *PLoS Comput Biol*, 9(3):e1002957, 2013.
- Naoki Nariai, Eric D Kolaczyk, and Simon Kasif. Probabilistic protein function prediction from heterogeneous genome-wide data. *Plos one*, 2(3):e337, 2007.
- Tudor I Oprea. Exploring the dark genome: implications for precision medicine. *Mammalian Genome*, 30(7-8):192–200, 2019.
- Dac-Trung Nguyen, Stephen Mathias, Cristian Bologa, Soren Brunak, Nicolas Fernandez, Anna Gaulton, Anne Hersey, Jayme Holmes, Lars Juhl Jensen, Anneli Karlsson, et al. Pharos: Collating protein information to shed light on the druggable genome. *Nucleic acids research*, 45(D1):D995–D1002, 2017.
- Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wizniewicz. The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary oncology*, 19(1A):A68, 2015.
- Emily Clough and Tanya Barrett. The gene expression omnibus database. In *Statistical genomics*, pages 93–110. Springer, 2016.
- Alexander Lachmann, Huilei Xu, Jayanth Krishnan, Seth I Berger, Amin R Mazloom, and Avi Ma'ayan. Chex: transcription factor regulation inferred from integrating genome-wide chip-x experiments. *Bioinformatics*, 26(19):2438–2444, 2010.
- Alexander Lachmann and Avi Ma'ayan. Kea: kinase enrichment analysis. *Bioinformatics*, 25(5):684–686, 2009.
- Cynthia L Smith, Carroll-Ann W Goldsmith, and Janan T Eppig. The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome biology*, 6(1):R7, 2004.
- Kevin Drew, Chanjae Lee, Ryan L Huizar, Fan Tu, Blake Borgeson, Claire D McWhite, Yun Ma, John B Wallingford, and Edward M Marcotte. Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Molecular systems biology*, 13(6):932, 2017.
- Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Solis, et al. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1):D1005–D1012, 2019.
- Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.
- Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.
- Fionn Murtagh. Multilayer perceptrons for classification and regression. *Neurocomputing*, 2(5-6):183–197, 1991.
- David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Mitchell Klein. *Logistic regression*. Springer, 2002.
- Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.
- Irina Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.
- Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613. ACM, 1998.
- Suresh Balakrishnama and Aravind Ganapathiraju. Linear discriminant analysis-a brief tutorial. In *Institute for Signal and information Processing*, volume 18, pages 1–8, 1998.
- The HDF Group. Hierarchical data format version 5, 2000-2010.
- Espen Mikal Robertsen, Tim Kahlke, Inge Alexander Raknes, Edvard Pedersen, Erik Kjærner Semb, Martin Ernsten, Lars Ailo Bongo, and Nils Peder Willassen. Meta-pipe-pipeline annotation, analysis and visualization of marine metagenomic sequence data. *arXiv preprint arXiv:1604.04103*, 2016.
- Tudor I Oprea, Cristian G Bologa, Søren Brunak, Allen Campbell, Gregory N Gan, Anna Gaulton, Shawn M Gomez, Rajarshi Guha, Anne Hersey, Jayme Holmes, et al. Unexplored therapeutic opportunities in the human genome. *Nature reviews Drug discovery*, 17(5):317, 2018.

41. Khalfan Al Senaidi, Niranjani Joshi, Maryam Al-Nabhani, Ghalia Al-Kasbi, Abdullah Al Farqani, Khalid Al-Thihli, and Almundher Al-Maawali. Phenotypic spectrum of alpk3-related cardiomyopathy. *American Journal of Medical Genetics Part A*, 179(7):1235–1240, 2019.
42. Celia Zazo Seco, Anne MM Oonk, María Domínguez-Ruiz, Jos MT Draaisma, Marta Gandía, Jaap Oostrik, Kornelia Neveling, Henricus PM Kunst, Lies H Hoefsloot, Ignacio Del Castillo, et al. Progressive hearing loss and vestibular dysfunction caused by a homozygous nonsense mutation in clic5. *European Journal of Human Genetics*, 23(2):189–194, 2015.
43. Casey S Greene, Arjun Krishnan, Aaron K Wong, Emanuela Ricciotti, Rene A Zelaya, Daniel S Himmelstein, Ran Zhang, Boris M Hartmann, Elena Zaslavsky, Stuart C Sealfon, et al. Understanding multicellular function and disease with human tissue-specific networks. *Nature genetics*, 47(6):569–576, 2015.

Fig. S1. PrismExp runtime and memory consumption for 50 gene expression clusters.

Supplementary Note 1: Supplementary Material

Fig. S2. Random Forest model prediction performance (average set AUROC), when PrismExp is trained on six gene-set libraries as well as prediction performance of single correlation matrix (global).