DATA-DRIVEN COMPUTATIONAL APPROACHES TO UNRAVEL AND INTERPRET THE POLYGENIC ARCHITECTURE OF HUMAN COMPLEX TRAITS AND DISEASES

By

Alexander Patrick McKim

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Computational Mathematics, Science, and Engineering—Doctor of Philosophy
Genetics and Genome Sciences—Dual Major

2024

ABSTRACT

The genetics and mechanisms underlying most human traits and diseases can be very complex, where the number of true gene associations can number in the many hundreds. Thus, when trying to better understand a trait/disease, researchers are faced with two main challenges; missing knowledge of which genes are truly related to the trait/disease, and understanding how those genes work together through molecular pathways. These knowledge gaps make it hard to translate large scale genetic information into actionable hypotheses. The overarching goal of the research presented in this dissertation is to develop methods that address these challenges in order to gain a better understanding of the etiology of complex traits and diseases. We worked towards this goal by developing general-purpose computational frameworks that leverage vast publicly-available datasets — genome-scale gene networks, gene functional annotations, thousands of gene expression signatures, and experimentally-derived gene-phenotype associations in humans and model organisms — to resolve large gene lists associated with highly polygenic disease into relevant genes, pathways, and critical interactions. Together, these findings reveal nuanced understanding of disease mechanisms. Overall, this research helps get away from treating each disease as a single well-defined condition, and instead find mechanism-based disease subtypes and use these insights to find novel diagnostic and treatment avenues.

This dissertation is dedicated to my father, who encouraged me from the beginning until the end.

ACKNOWLEDGEMENTS

I would like to acknowledge the people that most helped me get this dissertation written and completed. Dr. Christopher Mancuso, thank you for helping me not just with the projects here, but keeping me sane and really pushing me forward to complete this. To Keenan Manpearl, thank you for coming through for me while I was finishing this. This thesis would not have been completed without you two. I would also like to thank Dr. Kayla Johnson for talking to me about thesis worries and encouraging me that I am not in a hopeless position. I also want to thank her for helping to teach me biology when I was starting out and super interested in the subject, but knew nothing. I thank Dr. Stephanie Hickey for being my co-first author on the published work in chapter 2, and for also teaching so many cool things about biology. I thank Hao Yuan for being amazing to talk to and for providing great feedback over the years on my work. I also want to thank my committee members Dr. Wen Huang, Dr. Shin-han Shiu, and Dr. Jianrong Wang for their feedback on the ideas I have presented and in discussing with me development opportunities for my post-PhD life. I would also like to thank Dr. Janani Ravi for introducing me to topics involving the intersection of human genetics and infectious disease, and for all the conversations about how my work can contribute. Last but not least, I want to thank my advisor, Dr. Arjun Krishnan, for taking a chance on a student who knew nothing about biology to start with and has taught me so much.

TABLE OF CONTENTS

CHAPTER 1: BACKGROUND, RESEARCH QUESTIONS, AND JUSTIFICATION

Background

Over the past 15 years, large-scale studies, including gene expression, genome-wide association (GWAS), and long-term biobank research have shown that complex diseases (e.g., type II diabetes and coronary artery disease) are linked to hundreds of genes. Further, individuals with the same disease can exhibit a range of phenotypes. However, we still lack a thorough understanding of how disease genes lead to deregulated molecular pathways and cellular functions and which mechanisms underlie specific disease phenotypes. Closing this knowledge gap is crucial because we need a mechanism-based framework to understand how, despite having the same disease, patients can have unique genetic mutations and how these mutations lead to different functional and phenotypic disruptions that then lead to the disease. Consequently, addressing this gap also has an impact on moving diagnoses and treatments from the current "one size fits all" paradigm towards designing measures and interventions that work for individual patients that may be different from the status quo based on population-level understanding of disease. Therefore, there is a critical need for developing general-purpose computational approaches that can analyze large, heterogeneous data collections to connect genes, pathways, cell functions, phenotypes, diseases, and drug candidates.

The Post-GWAS era

The last fifteen years have seen major advancements in the knowledge of the genetic architecture underneath complex diseases, including genetic mutations, genes, mechanisms, and clinical phenotypes. Examples include the growing catalog of millions of genetic variants identified in GWAS experiments for thousands of complex diseases[1,2] and the growing number of gene function annotations in tissue-specific networks[3]. A significant challenge in the post-GWAS era is learning the mechanistic relationships that underlie complex diseases and trait variation across individuals. This problem is due to the extreme polygenicity[4,5] behind complex traits that GWAS has revealed and the lack

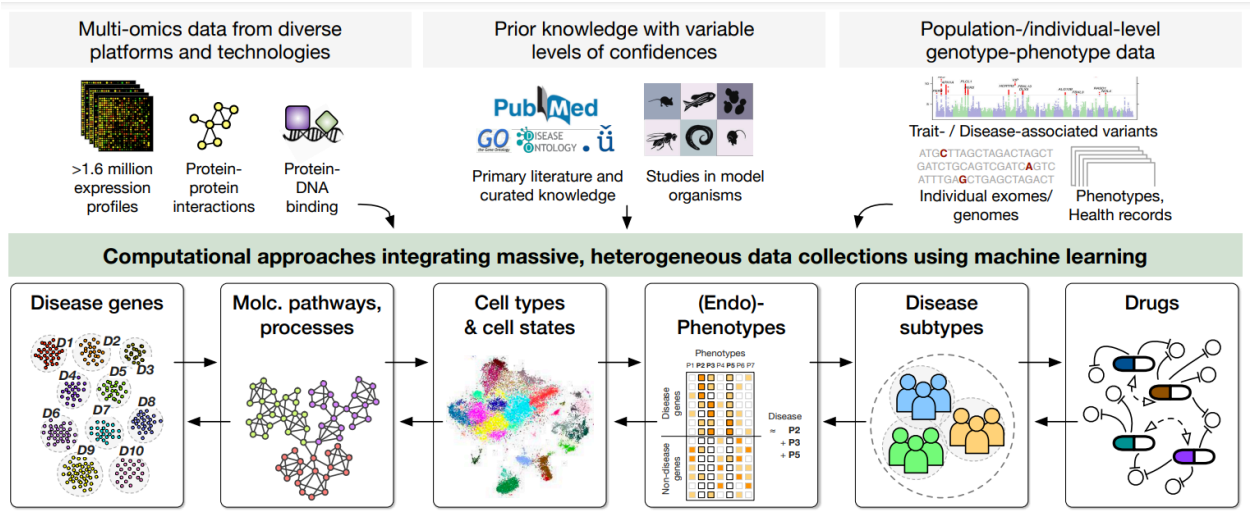

Figure 1.1: An overview of the biological data and concepts used and discussed in this dissertation. The explosion of various types of diverse -omics data, the new knowledge gained and that addition to new genesets, and the ability to cheaply sequence individuals and populations has created the foundation for creating computational methods to unravel the biology of human complex traits and disease.

of functional understanding of the thousands of important single nucleotide polymorphisms (SNPs) found by GWAS without additional biological annotation[5]. When GWAS is conducted, techniques like fine mapping[5,6] are often employed to predict causal SNPs for further study. However, the omnigenic model, motivated by the extreme polygenicity behind complex traits revealed through GWAS, for example challenges many of the assumptions behind the usefulness of fine mapping variants alone[7]. Isolating causal variants if there is truth behind all loci in the genome contributing to trait manifestation at some level[8] may be at odds with underlying biology. Even with simplifying biological reality, there has been genuine progress computationally utilizing new genomic data consortiums to elucidate our functional knowledge of many complex diseases. An example of this is annotating significant SNPs from a GWAS to the closest (or a very close) gene for downstream analysis[9,10]. While the integration of a variety of different data sources, such as eQTL data in conducting TWAS or chromatin capture data[7,9,11,12] are vital tools for learning genes that are associated with a phenotype, it is also true that SNPs very often have their causative gene or relatively close to its physical location[13,14] . Biswas and colleagues further argue that genes associated with

relevant disease pathways also are often near significant SNPs[15]. Thus, predicting genes near the significant SNPs is often reasonable. Integrating more biological information is not because gene prioritization metrics based on only proximity performs poorly, but because using additional biological data is complementary to simpler gene proximity methods. Integrating other biological data finds relationships would not have the statistical power to find consistently[16]. The challenge is figuring out the best biological information to help interpret results into something actually meaningful and usable.

Post-GWAS methods

Important methods have been developed using GWAS data to relate common variant data to other biological data types, aiming to unravel human genetics for disease and design treatments[17]. Linking implicated loci from GWAS to important biological pathways[9,10,18] provides a more interpretable explanation behind disease phenotypes. One approach to interpret GWAS results has been to utilize networks. Numerous methods have been implemented to map implicated GWAS loci with gene-gene networks to prioritize disease genes and pathways[19–22]. Discovered GWAS variants can be linked to expression data from both bulk RNA and single-cell RNA studies[23–27] to refine predictions about important genes and discover what each data set alone cannot. Polygenic risk scores —calculated using variant association data from GWAS— are able to perform reasonably well for at least some diseases because of the ability to predict disease risk of individual patients based on what variants they specifically have[28,28,29]. Polygenic risk scores have powerful applications for precision medicine when they can begin to work for patients of all backgrounds and risk factors[30,31]. In summary, GWAS results are powerful not only to interpret disease at the variant level but for translating genome scale results to the gene and pathway level to study mechanisms behind human disease and traits. While these diverse findings and methods using many different models have opened the door for many avenues of study, there are still key questions about polygenic diseases that must be investigated.

Disease Heterogeneity

A complex disease is not a single well defined condition, and the more complex it is, the less defined it will be. This is because each disease manifests uniquely in each

individual, leading to extreme heterogeneity of the disease across the population. Further, translating quantitative models for disease risk across individuals becomes even harder when the discovery and target individuals have different genetic ancestries, i.e. low cross-ethnic portability[29]. An example of this heterogeneity can be seen with autism spectrum disorder (ASD), which has extreme patient stratification and subtypes. Many of these differences are explained because of a different genetic makeup across individuals[32]. ASD has hundreds of genes associated with it[33], but individuals do not have mutations that target each of those hundreds of associated genes. Rather, a patient has mutations in a subset of genes that lead to a general diagnosis of autism, but in actuality manifests as a specific subtype of autism within the spectrum and within that patient. The genetic variation across patients in understanding disease mechanism and treating individuals. Pharmaceutical companies recognize the heterogeneity that can be seen in a complex disease, and are undertaking designing drug treatments for sub-groups of patients[34]. This can be seen in defining subtypes of diseases as endotypes. While what makes a good definition for an *endotype* is still controversial[35]**,** the motivation to classify patients with the same disease into different subtypes is in order to facilitate better treatment in groups of people with different types of gene and mechanistic disruptions.

Cross-Disease relationships

It is clear that complex diseases do not operate in isolation from one another. Cross trait work is motivated by finding similar genetic architectures between traits or diseases, which potentially leads to similar treatments[36]. There has been great success in demonstrating this shared architecture across many complex traits and diseases[36–38]. However, relationships between complex diseases are often not obvious. Seemingly distinct or unrelated complex diseases often manifest together more often than expected by chance, such as inflammatory bowel syndrome and cardiovascular disease[39]. Another type of comorbidity involves phenotypes that are either symptoms of the disease or are common traits that are involved in cellular pathways relating to a complex disease, known as endophenotypes[40]. Endophenotypes are typically complex, having multiple relevant genes or cellular mechanisms, but are less polygenic compared to the relevant complex disease. They often can be better annotated in terms of

underlying cellular mechanisms[41]. Endophenotypes can be shared across distinct complex diseases. For example chronic inflammation, a commonly observed disease phenotype, is involved with autoimmune disorders, type II diabetes, coronary artery disease, and various cancers[42]. Additionally, while an endophenotype should share genetic variance with the disease of interest[41], it can be difficult to define if an endophenotype is a symptom or a cause of a disease. Despite this, investigating underlying phenotypes can lead to developing treatments that work commonly across diseases.

Context-Specificity

It is crucial to find and annotate disease subtypes and underlying mechanisms to relevant tissue and cell types. This is because proteins and processes are important in specific contexts[43,44]. Understanding relevant contexts of complex diseases gives insight to the disruptions that lead to disease manifestation, which then helps identify treatment opportunities for a larger percentage of patients. Zhu et. al. point out that genetic variants influence the phenotypes that underlie complex diseases in a context-specific manner[45]. For an individual disease, associated variants do not necessarily influence the same context – implying that complex diseases can have multiple relevant tissue and cell types. This means that even if you know *a priori* tissues and cells that variants work in for a disease, newly discovered variants do not necessarily influence the same contexts. Thus, improving treatment opportunities requires researchers to understand not only what contexts a disease operates in, but context-specific functions.

Genotypic vs phenotypic first approaches

A common goal is to gain better functional understanding of complex traits through deconvolution of heterogeneity. However, methodologies for understanding the functional makeup of complex diseases and the genotype-phenotype relationships can differ[46]. With the advent of high-throughput sequencing, the genotype-first approach has allowed researchers to use less functionally interpretable biological information like variants to both define diseases and to learn about the functional makeup of a disease. GWAS is an example of a phenotype-first approach, where the study is conducted with patients chosen on observed disease status or some measured quantitative phenotype[5]. After the case/control groups are determined, associated variants are

identified and annotated through various methods such as mendelian randomization[47] and colocalization[48] to more biologically interpretable units like pathways. Conversely, a genetics-first approach is PHeWAS[49,50], which selects SNPs of interest and then finds associated phenotypes. Both approaches have advanced the understanding of complex diseases, including genetic annotations and pathways that underlie them, but additional analysis has to be done to predict effective treatments on an individual level.

Network biology

The vast complexity of biological entities and systems has led to the development of biological networks and graphs. These networks are used to capture relationships between entities – such as genes or proteins. Various networks, including protein-protein interaction networks[51,52], coexpression networks[53,54], and gene regulatory networks[55], serve this purpose. Network biology aims to understand genetics by utilizing biological interactions in a multi-dimensional framework. Interactions/connections (edges) between genes or proteins, or other entities represent functional association or relationship[56,57]. Networks are vital for interpreting complex disease biology because highly connected genes or proteins likely participate in similar biological mechanisms due to these functional relationships[58]. Networks are also diverse, such as being able to represent specific biological contexts. For example, protein-protein interaction networks differ across biological contexts because proteins interact differently in different cells. One specific use case is a study in which networks were created representing specific cellular and tissue contexts, and this has shed light on implicated disease-associated processes and mechanisms[59,60] that a context-naive network could not. Another useful property is that biological networks are modular in nature[61], where genes within a module are more densely connected with one another than other genes in the network. Complex trait/disease genes often map to network neighborhoods because disease-gene associations are not randomly scattered[62]. This occurs because disease genes are known to work together in disease-relevant processes and pathways. The discovered modules/clusters of interacting genes can implicate important phenotypes involved with the disease. For complex diseases involving hundreds of genes, multiple distinct modules are likely to be found, each enriched for distinct biological pathways and phenotypes that contribute to the disease[63–65]. Ultimately, the utility of networks is to

provide context to gene lists retrieved from experiments or some dataset, showing how the important genes discovered interact with one another in high-dimensional space.

Definitions of Biological Network Modules

As previously cited, biological networks tend to be modular in nature[56,63]. Researchers can discover how their genes of interest fall into modules within a network, and multiple distinct but related types of modules can be uncovered. Topological modules are locally dense neighborhoods in the network, found using only the network structure. This network is created through biological data, but additional outside information is not being used to influence module discovery. Functional modules consist of nodes that work within some specific biological concept, typically defined by user-provided external data, that are in a network neighborhood. Disease modules correspond to genes relevant to disease phenotype manifestation - i.e all disease associated genes. These concepts are interconnected, as functional gene modules will have relationship to topological modules in a network, and these will be relevant in disease modules the genes are annotated to. Given the high-dimensional nature of complex diseases and biology, there is no gold standard of “true” modules[63]. Disease modules are particularly challenging to discover due to network and data limitations. An important observation is that within a truly complete disease module including every relevant disease gene, this module would contain multiple submodules that relate to distinct functional processes, where submodule genes are more interconnected in the network relative to other disease genes. Therefore, many module detection methods discover topological modules within a network and use user-defined genes to identify functionally related topological modules[63,66,67]. An additional strategy is to define an initial disease module as the user seed list, and expand utilizing network properties[62]. In short, disease modules are those made of all disease genes - and this disease module will contain submodules that are related to distinct functions, phenotypes, cell types, tissues, and other biological concepts. It is clear that network modules and gene relationships are crucial for understanding complex disease biology. Surprisingly however, they are even relevant for mendelian and rare disease. While mendelian diseases are typically defined as being caused by single points of mutation, it has been observed that some specific mendelian diseases such as sickle cell disease are heterogeneous in the population

and associated with multiple phenotypes[56,68]. This suggests that even mendelian diseases potentially have a “true disease module” of genes related to functional processes disrupted by network effects of gene perturbations. This elucidates the challenge of developing effective treatment for many diseases and the importance of networks in biologically interpreting the unique mutations that lead to phenotype heterogeneity.

Network-based gene classification

Gene classification is the task of computationally predicting the association of genes to biological genesets, such as traits, processes, pathways, or diseases. These computational predictions are possible because of the emergence of large databases of publicly available data relating to gene function[69,70], and the integration of this functional information into genome wide networks[43,52,71,72]. This task is vital for two primary reasons: 1) Complex diseases can be highly polygenic on the order of hundreds of loci, variants, and genes[8,73]. Our known gene associations for very complex diseases are incomplete, with most genetic heritability of complex disease still being unexplained[73–75]. 2) Many genes are understudied in experimental settings[76–80]. Additionally, even if all genes were known, gene classification has utility in finding the genes most functionally related to experimental results that may be missing due to lack of power, noise, or immeasurability. Biological networks address these issues by providing functional context to genes of interest. Using a guilt-by-association approach[20,58], if a gene has ample edge connections to known positive genes, that gene is likely associated with the biological geneset due to these network relationships. This principle is based on the observation that genes in biological networks highly connected to one another have functional relationships and participate in the same or similar higher level biological concepts such as processes and phenotypes[81]. Using functional networks for gene classification assumes that discovering genes that work together in similar contexts or have similar function are likely to also be disease associated. Multiple studies and experiments have helped validate this assumption and the validity of this approach[3,33,82–86].

GenePlexus

GenePlexus is a supervised learning approach for network-based gene

classification[81,82,87,88]. It has been robustly demonstrated that supervised learning outperforms label propagation for gene classification[81]. Given a list of genes, GenePlexus predicts the association of all genes in a genome-wide network to that list based on the list's genes connectivity to other genes in the network. The model is a L2-regularized logistic regression classifier that distinguishes between positive and negative labels. The user-supplied genelist are used as positive labels. Negative labels are chosen through a hypergeometric test – where in taking the positive labels, a hypergeometric test is conducted with all gene sets in gene set collections (GSCs) such as DisGeNet[89,90] and GO[70]. If the positive genelist has significant overlap ($p < .05$) with a geneset, the genes in that set which are not part of the initial user geneset are defined as neutral. All other genes are given negative labels. Multiple gene-level feature vectors can be used. One feature type being the rows of the adjacency matrix, where each row corresponds to a gene. The rows of the node embedding matrix determined from node2vec[91,92] is another example. GenePlexus offers two major benefits: prioritizing genes that are good candidates for experimental study by investigating the top-ranked genes, and giving each gene in the genome a prediction of how related that gene is to the user geneset.

Omnigenic model

Fully explaining disease mechanisms is challenging for complex diseases due to their extremely polygenic nature. The omnigenic model[8,93,94] is motivated by the particular challenges revealed in GWAS studies, where most statistically significant loci have a small effect size, and these small effect size SNPs explain most of the genetic heritability. In addition, GWAS results are highly dependent on sample size. As experiments get bigger, the number of small effect size loci discovered increases, while large effect size loci actually have been found to decrease with larger sample sizes[14]. The question raised here is what these observations imply about the genetic architecture that underlies complex traits, and this motivates the omnigenic model – viewing diseases as networks. Specifically, the model proposes there are a relatively small number of "core" disease genes with direct, mechanistically interpretable genes that are influenced by perturbations in a much larger number of "peripheral" disease genes. From a GWAS perspective, a multitude of low effect size loci would be targeting

peripheral genes and the relatively high effect size loci would be enriched with core gene relationships[14]. A challenge associated with this model is delivering refined, objective definitions of what are truly core and peripheral genes. There is no gold standard list where the genes of the disease are categorized and validated in this way. Some distinctions that have some agreement are that core genes will be those which are more relevant for designing treatments – and will be enriched for drug targets[95]. Additionally they will be thought of as typically being more conserved[96]. Multiple noticeable and novel attempts to define candidate core genes have been created[95,97,98]. However, methods to either "discover" core genes or investigate biological statistics such as conservation typically utilize small numbers of validated important genes or pathways – and neglect the modular nature of diseases within a network – where multiple distinct modules enriched for distinct phenotypes and pathways are discovered when mapping disease genes. In other words, they assume important known genes are core and evaluate them in the context of other data[99]. Because the definition of core genes includes the set being "small", it is tempting to think of only a singular pathway being of note. Given the number of disease genes and the modular nature of both networks and disease genes when mapped to networks, it is unlikely a singular set of core genes would have direct effects on all of these diverse phenotypes. This suggests the classical definition and way of thinking of core genes actually underestimates the genetic complexity and heterogeneity of disease.

Research Questions

The explosion of -omics data for complex traits and diseases has created enormous opportunities for fueling new computational models. In ideal circumstances, computational models can both interpret lab experiment results and guide researchers in planning new ones. A primary challenge of working with human genetic data is that for some complex diseases and traits, the polygenic genetic architecture is enormously complex on a large multi-dimensional scale of hundreds or thousands of loci. Developing methods that manage and leverage that complexity is a vital challenge that must be addressed. Nearly every patient with a disease has unique mutations – where the loci do not interact independently but influence and interact with one another. Unraveling that complexity is what will allow precision medicine to become a truly viable

way to treat patients, where the patient with completely unique, unseen mutations can be stratified to a specific disease subtype[100]. Two outstanding, essential questions are thus: *How do we implement computational methods for seemingly impossible-to-unravel genetic complexity?* and *How do we interpret diseases at the gene and pathway level?* Our proposal is that complex traits are best interpreted as smaller, meaningful subnetworks of genes. This is true across the large number of complex traits that exist out there, and can be leveraged to build general-purpose methods that work across many datasets. Working with these subnetworks is what will provide viable and better answers to a wide range of computational problems from gene classification and prioritization, to discovering relevant pathways, and discovering which disease genes are likely to have important and notable perturbation effects.

Dissertation Contributions and Significance

Firstly, we show a use case of the sheer complexity that human diseases can have. Inflammation is a common pathway implicated in many diseases. While some diseases are defined by inflammation pathways, such as autoimmune diseases, many complex diseases have inflammation components in a way that is less obvious genetically but has been phenotypically observed. Some common examples of this include alzheimer's disease, coronary heart disease, and endometriosis. We investigate whether networks and modules can unravel the inflammation components of these non-autoimmune complex diseases, and define the specific inflammation pathways involved. Additionally, we predict approved drugs in autoimmune diseases that could be repurposed to target discovered inflammation components in complex disease. This project demonstrates the utility of networks in discovering meaningful subsets of genes to unravel specific pathways and phenotypes of note within highly complex human disease. We show that networks are a useful tool for making novel discoveries at a refined level – enabling researchers to investigate specific processes in disease subnetworks rather than the disease as a whole.

Secondly, we address the challenges of interpreting biology, namely the discovery of novel genes, from large gene sets. Real world experiments – such as differential expression – of complex disease data not only have hundreds of relevant gene results meeting significance thresholds, but also have noise and false positive results. To

address these issues, we implemented ModGenePlexus, which is an extension of the GenePlexus method, to discover modules of a disease for gene discovery experimental data. We leverage discovering functional modules of diseases to not only find meaningful biological subsets of genes, but to additionally perform a form of label propagation - a robust, well validated semi-supervised form of gene classification. This removes genes that aren't well connected in a network – false positives – and finds genes that have highly robust connections – false negatives. These semi-supervised results are then run with GenePlexus where a supervised learning model is created for each discovered module. We demonstrate that performing gene classification in this way is superior to running GenePlexus on the initial experimental results as a whole, an uncovers unique biology of diseases that is missed when considering the entire gene list as a singular unit. Our method removes poorly connected genes, and discovers genes that were not found in the experiment, performing geneset refinement for downstream analysis.

Thirdly, we use ModGenePlexus to predict candidate core and peripheral genes by finding and defining a proposed disease module using GWAS data. Assuming this module has every gene that could ever possibly be associated with the disease in question, from an omnigenic perspective the likely 'core' genes are those which have a large number of network connections to other disease genes within the module. This is because disease core genes are influenced by a much larger number of disease peripheral genes – whose main reason for disease association is through network connection to the core genes. This definition allows us to categorize both core and peripheral genes within the proposed disease module, and we demonstrate that we predict meaningful core genes for atrial fibrillation. Similarly, our method allows us to make predictions for important genes of other species utilizing network connections in a multi-species network framework. Starting with a human disease gene list, we are able to find the genes of other organisms that are highly connected to the human disease gene list. These model organism genes and their network connections are used to shed light on how the human core gene orthologs transfer into the model organism space, which has major implications for designing experiments and predicting genes with functional relevance across species.

Across all these projects, we have either released or plan to release code to both reproduce our methods and expand on them. Code for chapter 2 is already publicly available on github repository (https://github.com/krishnanlab/chronic-inflammation) and in Zenodo record (https://zenodo.org/record/6858073). Chapter 2 has additionally been published[100]. ModGenePlexus will be integrated into the publicly released software package PyGenePlexus[88] python package and into the GenePlexus webserver[87], enabling researchers to obtain top gene hits on a gene module basis - benefiting common and existing computational pipelines.

Dissertation Structure and Research Summary

The rest of this dissertation is organized as follows: Chapter 2 demonstrates utilizing a computational approach that integrates networks, large complex disease associations, and drug-target information to isolate aspects of diseases that correspond to chronic inflammation phenotypes and genes. We integrate a drug prioritization method with our module predictions to discover inflammation related gene targets of these diseases. Chapter 3 describes an innovation to the GenePlexus method, allowing GenePlexus to be used with large scale experimental -omics data. We demonstrate that our new method systematically outperforms original GenePlexus performance for real world experiments by using semi-supervised classification to discover gene modules for supervised learning and that unique, additional biology is uncovered when using this new method. Chapter 4 dives into interpreting complex diseases utilizing the omnigenic model by predicting and categorizing core genes within the discovered disease module using GWAS experimental data. We demonstrate the functional relevance of the predicted core genes and elucidate how they relate to other biological data. Additionally we demonstrate the utility of discovering cross species phenotypes through demonstrating how they relate to mechanistically relevant human genes. Chapter 5 discusses the broader impact, some noticeable limitations that we designed our methods to work around, and proposes future directions possible based on the results of this dissertation. Lastly, Chapter 6 gives an overview of some gene prioritization methods that can be used with GWAS summary statistics data, and goes over a project that motivated the biological concepts and some methods discussed in this dissertation.

REFERENCES

1. Caliskan, M., Brown, C. D. & Maranville, J. C. A catalog of GWAS fine-mapping efforts in autoimmune disease. *Am. J. Hum. Genet.* **108**, 549–563 (2021).

2. Sudlow, C. *et al.* 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**, (2015).

3. Greene, C. S. *et al.* Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* **47**, 569–576 (2015).

4. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).

5. Uffelmann, E. *et al.* Genome-wide association studies. *Nat. Rev. Methods Primer* **1**, 1–21 (2021).

6. Genetic Investigation of ANthropometric Traits (GIANT) Consortium *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012).

7. Broekema, R. V., Bakker, O. B. & Jonkers, I. H. A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. *Open Biol.* **10**, 190221 (2020).

8. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).

9. Leeuw, C. A. de, Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Comput. Biol.* **11**, e1004219 (2015).

10. Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z. & Bergmann, S. Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. *PLOS Comput. Biol.* **12**, e1004714 (2016).

11. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).

12. GTEx Consortium *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).

13. Brodie, A., Azaria, J. R. & Ofran, Y. How far from the SNP may the causative genes be? *Nucleic Acids Res.* **44**, 6046–6054 (2016).

14. Sinnott-Armstrong, N., Naqvi, S., Rivas, M. & Pritchard, J. K. GWAS of three molecular traits highlights core genes and pathways alongside a highly polygenic background. *eLife* **10**, e58615 (2021).

15. Biswas, S., Pal, S., Majumder, P. P. & Bhattacharjee, S. A framework for pathway knowledge driven prioritization in genome-wide association studies. *Genet.*

Epidemiol. **44**, 841–853 (2020).

16. Gerring, Z. F., Mina-Vargas, A., Gamazon, E. R. & Derks, E. M. E-MAGMA: an eQTL-informed method to identify risk genes using genome-wide association study summary statistics. *Bioinformatics* **37**, 2245–2249 (2021).

17. Adam, Y., Samtal, C., Brandenburg, J.-T., Falola, O. & Adebiyi, E. Performing post-genome-wide association study analysis: overview, challenges and recommendations. *F1000Research* **10**, 1002 (2021).

18. Salvi, R., Gawde, U., Idicula-Thomas, S. & Biswas, B. Pathway Analysis of Genome Wide Association Studies (GWAS) Data Associated with Male Infertility. *Reprod. Med.* **3**, 235–245 (2022).

19. Kim, S. S. *et al.* Genes with High Network Connectivity Are Enriched for Disease Heritability. *Am. J. Hum. Genet.* **104**, 896–913 (2019).

20. Lee, I., Blom, U. M., Wang, P. I., Shim, J. E. & Marcotte, E. M. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.* **21**, 1109–1121 (2011).

21. Vanunu, O., Magger, O., Ruppin, E., Shlomi, T. & Sharan, R. Associating Genes and Protein Complexes with Disease via Network Propagation. *PLoS Comput. Biol.* **6**, e1000641 (2010).

22. Taşan, M. *et al.* Selecting causal genes from genome-wide association studies via functionally coherent subnetworks. *Nat. Methods* **12**, 154–159 (2015).

23. De Vries, D. H. *et al.* Integrating GWAS with bulk and single-cell RNA-sequencing reveals a role for LY86 in the anti-Candida host response. *PLOS Pathog.* **16**, e1008408 (2020).

24. Jagadeesh, K. A. *et al.* Identifying disease-critical cell types and cellular processes by integrating single-cell RNA-sequencing and human genetics. *Nat. Genet.* **54**, 1479–1492 (2022).

25. Boltz, T. *et al.* Cell type deconvolution of bulk blood RNA-Seq to reveal biological insights of neuropsychiatric disorders. *BioRxiv Prepr. Serv. Biol.* 2023.05.24.542156 (2023) doi:10.1101/2023.05.24.542156.

26. Mai, J., Lu, M., Gao, Q., Zeng, J. & Xiao, J. Transcriptome-wide association studies: recent advances in methods, applications and available databases. *Commun. Biol.* **6**, 899 (2023).

27. Zhou, Y. *et al.* Integrating RNA-Seq With GWAS Reveals a Novel SNP in Immune-Related HLA-DQB1 Gene Associated With Occupational Pulmonary Fibrosis Risk: A Multi-Stage Study. *Front. Immunol.* **12**, 796932 (2022).

28. Collister, J. A., Liu, X. & Clifton, L. Calculating Polygenic Risk Scores (PRS) in UK Biobank: A Practical Guide for Epidemiologists. *Front. Genet.* **13**, 818574 (2022).

29. Hu, X. *et al.* Polygenic transcriptome risk scores for COPD and lung function improve cross-ethnic portability of prediction in the NHLBI TOPMed program. *Am. J. Hum. Genet.* S0002929722001033 (2022) doi:10.1016/j.ajhg.2022.03.007.

30. Richardson, T. G., O'Nunain, K., Relton, C. L. & Davey Smith, G. Harnessing Whole Genome Polygenic Risk Scores to Stratify Individuals Based on Cardiometabolic Risk Factors and Biomarkers at Age 10 in the Lifecourse—Brief Report. *Arterioscler. Thromb. Vasc. Biol.* **42**, 362–365 (2022).

31. Patel, A. P. *et al.* A multi-ancestry polygenic risk score improves risk prediction for coronary artery disease. *Nat. Med.* **29**, 1793–1803 (2023).

32. Sandin, S. *et al.* The Heritability of Autism Spectrum Disorder. *JAMA* **318**, 1182 (2017).

33. Krishnan, A. *et al.* Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat. Neurosci.* **19**, 1454–1462 (2016).

34. Hartl, D. *et al.* Translational precision medicine: an industry perspective. *J. Transl. Med.* **19**, 245 (2021).

35. Genkel, V. V. & Shaposhnik, I. I. Conceptualization of Heterogeneity of Chronic Diseases and Atherosclerosis as a Pathway to Precision Medicine: Endophenotype, Endotype, and Residual Cardiovascular Risk. *Int. J. Chronic Dis.* **2020**, 5950813 (2020).

36. Caberlotto, L. *et al.* Cross-disease analysis of Alzheimer's disease and type-2 Diabetes highlights the role of autophagy in the pathophysiology of two highly comorbid diseases. *Sci. Rep.* **9**, 3965 (2019).

37. Pividori, M. *et al.* PhenomeXcan: Mapping the genome to the phenome through the transcriptome. *Sci. Adv.* **6**, eaba2083 (2020).

38. Guo, M. *et al.* Analysis of disease comorbidity patterns in a large-scale China population. *BMC Med. Genomics* **12**, 177 (2019).

39. Román, A. L. S. Comorbidity in inflammatory bowel disease. *World J. Gastroenterol.* **17**, 2723 (2011).

40. Kendler, K. S. & Neale, M. C. Endophenotype: a conceptual analysis. *Mol. Psychiatry* **15**, 789–797 (2010).

41. Iacono, W. G., Malone, S. M. & Vrieze, S. I. Endophenotype best practices. *Int. J. Psychophysiol.* **111**, 115–144 (2017).

42. Furman, D. *et al.* Chronic inflammation in the etiology of disease across the life span. *Nat. Med.* **25**, 1822–1832 (2019).

43. Wong, A. K., Krishnan, A. & Troyanskaya, O. G. GIANT 2.0: genome-scale integrated analysis of gene networks in tissues. *Nucleic Acids Res.* **46**, W65–W70

(2018).

44. Guan, Y. *et al.* Tissue-Specific Functional Networks for Prioritizing Phenotype and Disease Genes. *PLoS Comput. Biol.* **8**, e1002694 (2012).

45. Zhu, H., Shang, L. & Zhou, X. A Review of Statistical Methods for Identifying Trait-Relevant Tissues and Cell Types. *Front. Genet.* **11**, (2021).

46. Kohane, I. S. Finding a new balance between a genetics-first or phenotype-first approach to the study of disease. *Neuron* **109**, 2216–2219 (2021).

47. Benn, M. & Nordestgaard, B. G. From genome-wide association studies to Mendelian randomization: novel opportunities for understanding cardiovascular disease causality, pathogenesis, prevention, and treatment. *Cardiovasc. Res.* (2018) doi:10.1093/cvr/cvy045.

48. Kanduri, C., Sandve, G. K., Hovig, E., De, S. & Layer, R. M. Editorial: Genomic Colocalization and Enrichment Analyses. *Front. Genet.* **11**, 617876 (2021).

49. Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26**, 1205–1210 (2010).

50. Legault, M.-A., Perreault, L.-P. L. & Dubé, M.-P. *ExPheWas: A Browser for Gene-Based pheWAS Associations*. http://medrxiv.org/lookup/doi/10.1101/2021.03.17.21253824 (2021) doi:10.1101/2021.03.17.21253824.

51. Safari-Alighiarloo, N., Taghizadeh, M., Rezaei-Tavirani, M., Goliaei, B. & Peyvandi, A. A. Protein-protein interaction networks (PPI) and complex diseases. *Gastroenterol. Hepatol. Bed Bench* **7**, 17–31 (2014).

52. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).

53. Ovens, K., Eames, B. F. & McQuillan, I. Comparative Analyses of Gene Co-expression Networks: Implementations and Applications in the Study of Evolution. *Front. Genet.* **12**, 695399 (2021).

54. Russell, M., Aqi, A., Saitou, M., Gokcumen, O. & Masuda, N. Gene communities in co-expression networks across different tissues. *ArXiv* arXiv:2305.12963v2 (2023).

55. Emmert-Streib, F., Dehmer, M. & Haibe-Kains, B. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. *Front. Cell Dev. Biol.* **2**, (2014).

56. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **12**, 56–68 (2011).

57. Vidal, M., Cusick, M. E. & Barabási, A.-L. Interactome Networks and Human

Disease. *Cell* **144**, 986–998 (2011).

58. Wang, X., Gulbahce, N. & Yu, H. Network-based methods for human disease gene prediction. *Brief. Funct. Genomics* **10**, 280–293 (2011).

59. Cvijovic, M. & Polster, A. Network medicine: facilitating a new view on complex diseases. *Front. Bioinforma.* **3**, 1163445 (2023).

60. Krishnan, A., Taroni, J. N. & Greene, C. S. Integrative Networks Illuminate Biological Factors Underlying Gene–Disease Associations. *Curr. Genet. Med. Rep.* **4**, 155–162 (2016).

61. Alcalá-Corona, S. A., Sandoval-Motta, S., Espinal-Enríquez, J. & Hernández-Lemus, E. Modularity in Biological Networks. *Front. Genet.* **12**, 701331 (2021).

62. Ghiassian, S. D., Menche, J. & Barabási, A.-L. A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome. *PLOS Comput. Biol.* **11**, e1004120 (2015).

63. The DREAM Module Identification Challenge Consortium *et al.* Assessment of network module identification across complex diseases. *Nat. Methods* **16**, 843–852 (2019).

64. Chen, Y. *et al.* Variations in DNA elucidate molecular networks that cause disease. *Nature* **452**, 429–435 (2008).

65. Pe'er, D., Regev, A., Elidan, G. & Friedman, N. Inferring subnetworks from perturbed expression profiles. *Bioinformatics* **17**, S215–S224 (2001).

66. Levi, H., Elkon, R. & Shamir, R. DOMINO: a network‐based active module identification algorithm with reduced rate of false calls. *Mol. Syst. Biol.* **17**, e9593 (2021).

67. Levi, H., Elkon, R. & Shamir, R. *DOMINO: A Novel Algorithm for Network-Based Identification of Active Modules with Reduced Rate of False Calls*. http://biorxiv.org/lookup/doi/10.1101/2020.03.10.984963 (2020) doi:10.1101/2020.03.10.984963.

68. Kato, G. J. *et al.* Sickle cell disease. *Nat. Rev. Dis. Primer* **4**, 18010 (2018).

69. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).

70. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).

71. Oughtred, R. *et al.* The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci. Publ. Protein Soc.*

30, 187–200 (2021).

72. Szklarczyk, D. *et al.* The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **49**, D605–D612 (2021).

73. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five Years of GWAS Discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).

74. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).

75. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).

76. Dolgin, E. The most popular genes in the human genome. *Nature* **551**, 427–431 (2017).

77. Dunham, I. Human genes: Time to follow the roads less traveled? *PLOS Biol.* **16**, e3000034 (2018).

78. Edwards, A. M. *et al.* Too many roads not taken. *Nature* **470**, 163–165 (2011).

79. Hoffmann, R. & Valencia, A. Life cycles of successful genes. *Trends Genet.* **19**, 79–81 (2003).

80. Haynes, W. A., Tomczak, A. & Khatri, P. Gene annotation bias impedes biomedical research. *Sci. Rep.* **8**, 1362 (2018).

81. Liu, R., Mancuso, C. A., Yannakopoulos, A., Johnson, K. A. & Krishnan, A. Supervised learning is an accurate method for network-based gene classification. *Bioinformatics* **36**, 3457–3465 (2020).

82. Mancuso, C. A., Johnson, K. A., Liu, R. & Krishnan, A. Joint representation of molecular networks from multiple species improves gene classification. *PLOS Comput. Biol.* **20**, e1011773 (2024).

83. Leiserson, M. D. M. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47**, 106–114 (2015).

84. Guan, Y., Ackert-Bicknell, C. L., Kell, B., Troyanskaya, O. G. & Hibbs, M. A. Functional Genomics Complements Quantitative Genetics in Identifying Disease-Gene Associations. *PLoS Comput. Biol.* **6**, e1000991 (2010).

85. Köhler, S., Bauer, S., Horn, D. & Robinson, P. N. Walking the Interactome for Prioritization of Candidate Disease Genes. *Am. J. Hum. Genet.* **82**, 949–958 (2008).

86. Warde-Farley, D. *et al.* The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* **38**, W214-220 (2010).

87. Mancuso, C. A. *et al.* GenePlexus: a web-server for gene discovery using network-based machine learning. *Nucleic Acids Res.* **50**, W358–W366 (2022).

88. Mancuso, C. A., Liu, R. & Krishnan, A. PyGenePlexus: a Python package for gene discovery using network-based machine learning. *Bioinformatics* **39**, btad064 (2023).

89. Piñero, J. *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* **45**, D833–D839 (2017).

90. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* **48**, D845–D855 (2020).

91. Grover, A. & Leskovec, J. node2vec: Scalable Feature Learning for Networks. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* 855–864 (ACM Press, San Francisco, California, USA, 2016). doi:10.1145/2939672.2939754.

92. Liu, R. & Krishnan, A. PecanPy: a fast, efficient and parallelized Python implementation of node2vec. *Bioinformatics* **37**, 3377–3379 (2021).

93. The Omnigenic Model: Response from the Authors. *J. Psychiatry Brain Sci.* (2017) doi:10.20900/jpbs.20170014S8.

94. Wray, N. R., Wijmenga, C., Sullivan, P. F., Yang, J. & Visscher, P. M. Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model. *Cell* **173**, 1573–1580 (2018).

95. Ratnakumar, A., Weinhold, N., Mar, J. C. & Riaz, N. Protein-Protein interactions uncover candidate 'core genes' within omnigenic disease networks. *PLOS Genet.* **16**, e1008903 (2020).

96. Mostafavi, H., Spence, J. P., Naqvi, S. & Pritchard, J. K. Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nat. Genet.* **55**, 1866–1875 (2023).

97. Ratajczak, F. *et al.* Speos: an ensemble graph representation learning framework to predict core gene candidates for complex diseases. *Nat. Commun.* **14**, 7206 (2023).

98. The Schizophrenia Working Group of the Psychiatric Genomics Consortium 2, *et al.* The role of polygenic risk score gene-set analysis in the context of the omnigenic model of schizophrenia. *Neuropsychopharmacology* **44**, 1562–1569 (2019).

99. Zhang, W., Reeves, G. R. & Tautz, D. Testing Implications of the Omnigenic Model for the Genetic Analysis of Loci Identified through Genome-wide Association. *Curr. Biol.* **31**, 1092-1098.e6 (2021).

100. Hickey, S. L., McKim, A., Mancuso, C. A. & Krishnan, A. A network-based approach for isolating the chronic inflammation gene signatures underlying complex diseases

towards finding new treatment opportunities. *Front. Pharmacol.* **13**, 995459 (2022).

CHAPTER 2: A NETWORK-BASED APPROACH FOR ISOLATING THE CHRONIC INFLAMMATION GENE SIGNATURES UNDERLYING COMPLEX DISEASES TOWARDS FINDING NEW TREATMENT OPPORTUNITIES

Abstract

Complex diseases are associated with a wide range of cellular, physiological, and clinical phenotypes. To advance our understanding of disease mechanisms and our ability to treat these diseases, it is critical to delineate the molecular basis and therapeutic avenues of specific disease phenotypes, especially those that are associated with multiple diseases. Inflammatory processes constitute one such prominent phenotype, being involved in a wide range of health problems including ischemic heart disease, stroke, cancer, diabetes mellitus, chronic kidney disease, non-alcoholic fatty liver disease, and autoimmune and neurodegenerative conditions. While hundreds of genes might play a role in the etiology of each of these diseases, isolating the genes involved in the specific phenotype (e.g. inflammation “component”) could help us understand the genes and pathways underlying this phenotype across diseases and predict potential drugs to target the phenotype. Here, we present a computational approach that integrates gene interaction networks, disease-/trait-gene associations, and drug-target information to accomplish this goal. We apply this approach to isolate gene signatures of complex diseases that correspond to chronic inflammation and use SAveRUNNER to prioritize drugs to reveal new therapeutic opportunities.

Introduction

Acute inflammation is an organism's healthy response to invasion by pathogens or to cellular damage caused by injury[1]. Systemic chronic inflammation (CI) occurs when these inflammatory responses do not resolve, resulting in persistent, low-grade immune activation that causes collateral damage to the affected tissue over time[2]. While the direct connection of CI to auto-immune diseases has been well known for some time, only recently has the medical community uncovered the prevalence of CI in a multitude of complex diseases and disorders[2,3]. Therefore, it is imperative to better understand the different molecular mechanisms of CI manifestation across diseases.

Network-based methods are a powerful collection of tools in both elucidating specific pathways and processes that may underlie a complex phenotype[4–6] as well as for drug repurposing[7–9]. For instance, HotNet2 is a pan-cancer network analysis in which active network modules in a genome-wide molecular network are determined by guiding the module detection algorithm with thousands of genes scored with how prevalent they are across 12 cancers in TCGA[4]. HotNet2 is then able to determine if any module is enriched for a given cancer type, pathway, or process. In a similar vein, another approach, DIAMOnD, starts with a genome-wide network, and then creates a disease specific network using an expanded set of known disease-gene annotations[5]. This disease specific network is then analyzed and compared to other disease specific networks generated using the same technique. Both approaches find regions of a genome-wide network that are enriched for disease-related genes.

Inflammation is an example of an endophenotype, or intermediate phenotype, of a complex disease. Ghiassian *et al.* studied endophenotype network models by starting with a genome-wide network and constructing modules for sets of seed genes related to three endophenotypes: inflammation, thrombosis, and fibrosis[6]. The authors showed that the network modules derived from the three endophenotypes have strong overlap in the network and that these modules are enriched for genes differentially expressed in various complex diseases. While the above methods provide invaluable insight in disease mechanisms using a disease-focussed and a phenotype-focussed approach, respectively, they raise the critical question of identifying phenotypic signatures specific to individual diseases. For instance, can we identify the CI-signature that is specific to a given disease and use that to find avenues for therapeutic intervention?

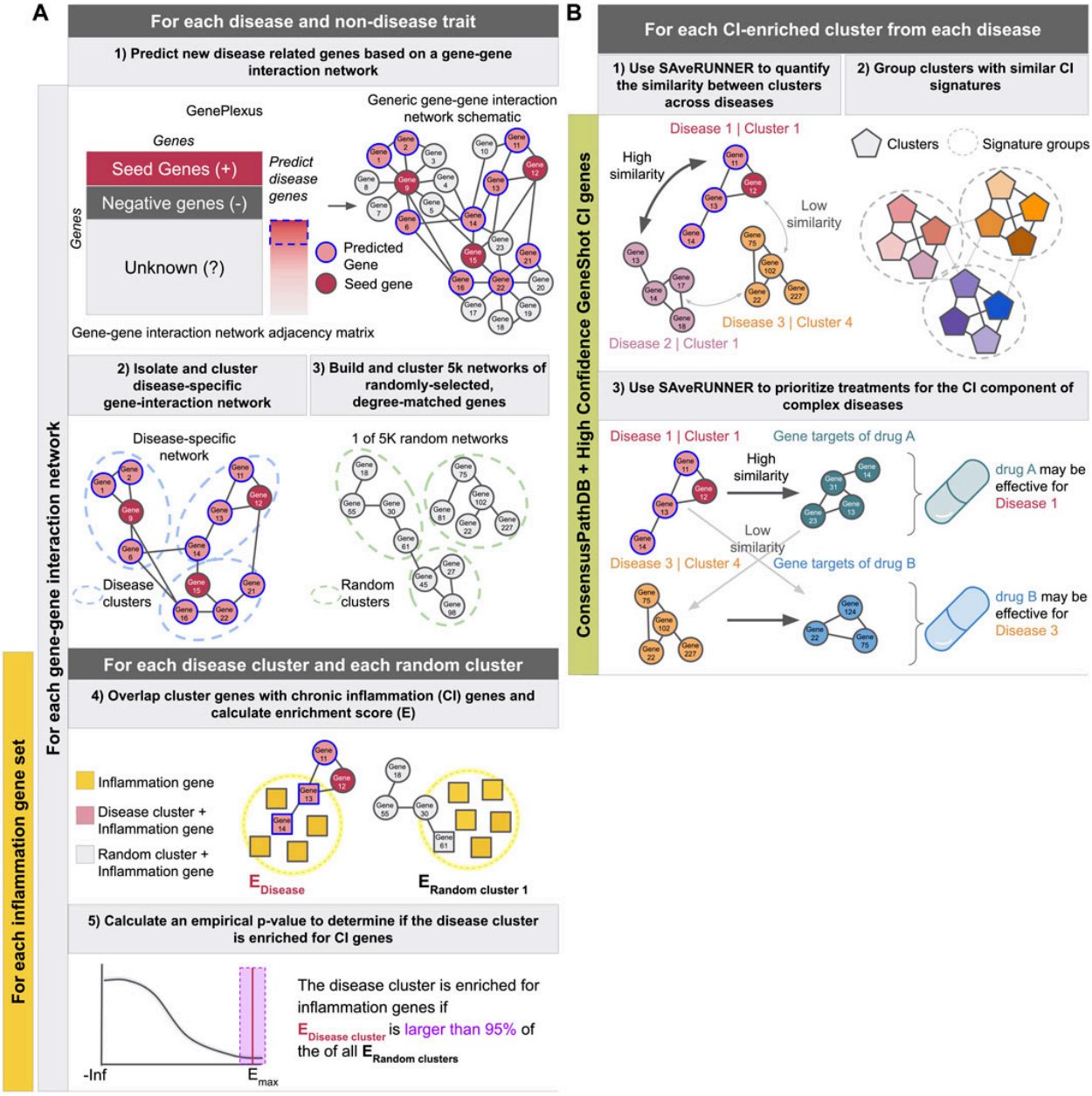

Figure 2.1: Schematics describing the experimental pipeline. **(A)** Describes predicting new disease related genes (step 1), clustering the disease specific interaction network (step 2) as well as 5,000 networks made from randomly-selected degree-matched genes for each disease, identifying CI-enriched clusters (steps 4 and 5), and calculating the proportion of diseases with at least one CI-enriched cluster. These steps were performed for each gene-gene interaction network in combination with each inflammation gene set described in the methods. **(B)** Describes using SAveRUNNER to find groups of CI-enriched clusters from all diseases with similar CI-signatures (steps 1

Figure 2.1 (cont'd)

and 2), and prioritize treatments for the CI component of complex diseases (step 3). Using ConsensusPathDB with the high-confidence GeneShot derived CI gene set resulted in the highest proportion of autoimmune diseases and the lowest proportion of non-disease traits with at least one CI-enriched cluster. Therefore, steps 1-3 were performed with clusters from that network-CI gene st combination only.

In this work, we address this question using a network-based approach. We first generate a network consisting of only genes associated with a single disease (**Figure 2.1A, steps 1-2**) Here, like in DIAMOnD[5], we expand our original disease-gene annotations to build more robust networks and glean insight into unstudied genes. We use a network-based supervised machine learning model to expand our gene sets[10], which has been shown to systematically outperform label propagation methods like DIAMOnD. We then cluster the disease specific network, and find clusters that are significantly enriched for known CI genes, and compare these CI signatures across diseases (**Figure 2.1A, 2.1B steps 1-2)**. We then use the SAveRUNNER[7] method on these enriched clusters to predict drugs that might help treat the CI-component specific to a given disease (**Figure 2.1B, step 3**).

Methods

Selection of complex and autoimmune diseases and associated seed genes

We searched the literature[2,11–15] and curated examples of 17 complex diseases associated with chronic inflammation (CI) and 9 common autoimmune diseases. Some of these diseases are quite broad (i.e "Malignant neoplasm of lung"), and to add more narrowly defined diseases to our list, we used the Human Disease Ontology[16] to identify child terms of these diseases. The chosen diseases were not meant to be comprehensive, but examples of autoimmune diseases and complex diseases thought to have immune components. We then identified genes annotated to each disease by the DisGeNet database, which is a database that stores a collection of disease-gene annotations from expert curated repositories, GWAS catalogs, animal models and the scientific literature (Piñero et al. 2020). To ensure that our disease gene sets were largely non-overlapping, we created a network such that nodes were diseases and an edge was created between two diseases if the two gene sets had ≥ 0.6 overlap (

$|A \cap B|/min(|A|, |B|)$). We then chose the most representative disease from each connected component. This resulted in 10 autoimmune diseases and 37 complex diseases (**Table S1**).

Selection of non-disease traits

Two lab members manually curated 113 non-disease-traits that are unlikely to be related to SNPs associated with CI (i.e. handedness, coffee intake, and average household income) from the list of traits with GWAS results from the UK Biobank[17] to be used as negative controls. Based on GWAS summary statistics from the Neale group[18], we used Pascal[19] (upstream and downstream windows of 50 KB with the sum-of-chi-squared statistics method; only autosomal variants) to associate genes with the non-disease traits. Genes with $p < 0.001$ were included as seed genes for that trait.

GenePlexus

To predict new genes associated with a set of input seed genes, we used GenePlexus, a tool that builds an L2-regularized logistic regression model using features from a gene interaction network[10]. As input features, we used the adjacency matrices from STRING, STRING with only experimentally derived edges (STRING-EXP)[20], BioGRID[21], and ConsensusPathDB[22]. For predicting disease genes, positive examples were disease/trait seed genes and negative example genes were generated by: (i) finding the union of all genes annotated to all diseases in DisGeNET[23], (ii) removing genes annotated to the given seed genes, and (iii) removing genes annotated to any disease in the collection that significantly overlapped with the given seed genes ($p < 0.05$ based on the one-sided Fisher's exact test)[10]. We tested the performance of the above features for predicting new genes associated with our diseases and traits of interest using three-fold cross validation and only included diseases in subsequent analyses if the diseases/traits had ≥ 15 associated genes and median $log2(auPRC/prior) \geq 1$. (i.e. the area under the precision-recall curve '$auPRC$' is at least twice as much as expected by random chance '$prior$'[10]). See **Figure 2.1A, step 1**.

Identifying clusters of interacting genes within a disease-specific network

One list of disease-associated genes was formed for each of the four biological networks used as features in GenePlexus. Specifically, we added genes with a GenePlexus prediction probability of ≥ 0.80 on the network of interest to the original

disease or trait seed gene list to create our final set of associated genes for each disease or trait for that network. We formed disease/trait-specific networks by subsetting a given network to include only the disease/trait associated genes and any edges connecting those genes based on direct interactions (**Figure 2.1A, step 2**). We tested five prediction-network–cluster-network combinations: Genes predicted on each of the four networks were clustered on the same network. Genes predicted on STRING were also clustered on both STRING and STRING-EXP to test if using the full network for novel gene prediction but only experimentally derived gene-gene associations for clustering would improve performance. We then used the Leiden algorithm[24] to partition the disease/trait-specific networks into clusters (**Figure 1A, step 2**). Specifically, we used the *leiden_find_partition* function from the leidenbase R package (v 0.1.3) (https://github.com/cole-trapnell-lab/leidenbase) with 100 iterations and ModularityVertex Partition as the partition type. We retained clusters containing ≥ 5 genes.

Cluster GOBP enrichment analysis

We used the R package topGO[25] (v 2.44.0) to find enrichment of genes annotated to GO biological processes (min size = 5, max size = 100) among disease gene clusters. The annotations were taken from the Genome wide annotation for Human bioconductor annotation package, org.Hs.eg.db[26] (v 3.13.0). The background gene set included all human genes present in the network of interest.

Defining chronic-inflammation-associated genes

We tested several different sets of chronic inflammation associated genes for this study including the GO[27] biological process (GOBP) terms GO:0002544 (”chronic inflammatory response”) and GO:0006954 (“inflammatory response”). These were collected from the Genome wide annotation for Human bioconductor annotation package, org.Hs.eg.db[26] (v 3.13.0) with and without propagation of gene-term relationships from the descendent terms (org.Hs.egGO2ALLEGS and org.Hs.egGO2EG, respectively). GO:0006954 was also filtered to retain gene-term relationships inferred from experiments (evidence codes EXP, IDA, IPI, IMP, IGI, IEP, HTP, HDA, HMP, HGI, and HEP). As GO:0002544 without propagation contained < 15 genes, this list was ultimately not included in the study. We also identified genes

associated with chronic inflammation using Geneshot which, given the search term "chronic inflammation", searches Pubmed using manually collected GeneRif gene-term associations to return a ranked list containing genes that have been previously published in association with the search term[28]. We tested both the entire Geneshot generated list, and the subset of genes with > 10 associated publications ("High confidence GeneShot"). As with the disease genes, we predicted additional chronic-inflammation-associated genes using GenePlexus with features from each network. Negative examples for GenePlexus were derived from non-overlapping GOBP terms. We added genes with a prediction probability of ≥ 0.80 to the seed gene list to create our final sets of CI-associated genes.

Creating random traits

After running GenePlexus to predict new genes for each trait, the gene lists for each trait were used to generate 5,000 random gene lists that have matching node degree distributions to the original traits (**Figure 2.1A, step 3**). That is, a random gene list was generated for a given trait by replacing each of its genes in the network of interest with a (randomly chosen) gene that has the same node degree, or a gene that has a close node degree if there are a small number of genes with the exact node degree[4,7]. We clustered the random traits as described in section 2.3. Only clusters with ≥ 5 genes were included. Real traits with no corresponding permuted traits with clusters containing ≥ 5 genes were excluded from the analysis.

Finding CI-gene enriched disease clusters

For each prediction-network—cluster-network pair and each CI gene list expanded on the prediction network of interest, for each disease and random trait cluster containing ≥ 5 genes, we calculated an enrichment score score, $E = log_2(\frac{(CG \cap CI)/CG}{CI/background})$ where CG are the genes in a disease cluster, CI are the CI genes, and $background$ are all of the genes present in the clustering network (**Figure 2.1A, step 4**). For each real disease or trait cluster, we used the matching random trait clusters to calculate a p-value, $p = \frac{n\,random\,cluster\,score \geq real\ \ cluster\,score}{n\,random\ \ clusters\ + 1}$ [4,7]. We corrected for multiple comparisons across clusters within a disease using the Benjamini-Hochberg procedure[29] (**Figure 2.1A, step 5**). Clusters with an $FDR < 0.05$ and $E > 0$ were considered

chronic-inflammation-associated disease clusters and were deemed to represent the 'CI signature' of the disease.

Identifying the optimal prediction-network/cluster-network/CI gene source combination

We chose the network/inflammation gene set combination that resulted in the highest proportion of autoimmune diseases and lowest proportion of non-disease traits with at least one CI-enriched cluster of any network/CI-gene set combination, ConsensusPathDB and the Geneshot generated list, subset with genes with associated publications.

Comparing chronic inflammation signatures across diseases

For CI-enriched clusters identified using ConsensusPathDB and the high-confidence Geneshot CI genes, we used the SAveRUNNER R package to quantify the similarity between each pair of CI-enriched clusters using ConsensusPathDB as the base network[7] (**Figure 2.1B, step 1**). For each pair, SAveRUNNER computes the average shortest path between each gene in $cluster\ A$ and the closest gene in $cluster\ B$ and uses this value to calculate an adjusted similarity score. Then, a p-value is estimated based on a null distribution of adjusted similarity scores between randomly generated clusters with the same node degree distributions as $clusters\ A$ and B. Because the similarity scores and p-values are not symmetric, ie. $A \rightarrow B \neq B \rightarrow A$, we used Stouffer's method to combine p-values for the same pair of clusters and averaged the adjusted similarities. We then used the Leiden algorithm as described in section 2.3 to group related clusters (**Figure 2.1B, step 2**). For each group, we took the union of the genes belonging to the resident CI-enriched clusters. Using genes unique to each group, with all of the ConsensusPathDB genes as background, we used TopGO as in section 2.4 to identify enriched GOBPs.

Identifying expert-curated drug-target associations

The known drug-gene interactions used in this study are the subset of the interactions present in the DrugCentral database[30] that are also among the expert curated interactions in the Drug-Gene Interaction database (DGIdb)[31]. Specifically, we used the DGIdb API to retrieve only drug-gene interactions that were marked "*Expert curated*" (based on the source trust levels endpoint). Intersecting these interactions with those in

DrugCentral (through a list of drug synonyms from DrugCentral) resulted in the final list of expert-curated drug-gene pairs.

Treatment prediction and scoring

We predicted treatment opportunities for the inflammatory component of complex diseases by using the SAveRUNNER R package [7] (**Figure 1B, step 3**). SAveRUNNER builds a bipartite drug-disease network by utilizing the previously determined expert-curated drug targets, the CI-associated cluster disease genes, and the ConsensusPathDB network as a human interactome. Network similarity scores returned by SAveRUNNER represent the proximity between disease and drug modules, where a high similarity score means that the disease and drug modules have high proximity in ConsensusPathDB. SAveRUNNER calculates a p-value where a significant value represents the disease genes and drug targets are nearby in the network more than expected by chance (based on an empirical null distribution built using 200 pairs of randomly selected groups of genes with the same size and degree distribution of the original sets of disease genes and drug targets). Using the list of final predicted associations after normalization of network similarity, the p-values were corrected for multiple comparisons within each disease using the Benjamini-Hochberg procedure. Drugs were associated to diseases based on the disease cluster with the lowest FDR. Predicted treatments are disease-drug pairs with an $FDR < 0.01$.

Evaluating SAveRUNNER prediction performance

We calculated $log2(auPRC/prior)$ by ranking disease-drug pairs by $-log_{10}(SAveRUNNER\ FDR)$ and using either previously indicated drug-disease pairs (both approved and off-label) or drug-disease pairs tested in a clinical trial as positive labels. Approved and off-label drug-disease pairs were collected from DrugCental[30]. Only drugs with expert curated target genes were included (see section *2.6.1*). The Unified Medical Language System (UMLS) Concept Unique Identifiers (CUI) were limited to diseases (T047) and neoplastic processes (T191), and our diseases were matched to diseases in DrugCentral using UMLS CUIs. Drug-disease pairs tested in a clinical trial were collected from the database for Aggregate Analysis of Clinical Trials (AACT) [32]. AACT reports the Medical Subject Headings (MeSH) vocabulary names for diseases. We used disease vocabulary mapping provided by DisGeNET to translate

UMLS CUIs for our diseases to MeSH vocabulary names, further restricted to only those that were present in AACT. We filtered AACT for trials with “Active, not recruiting”, “Enrolling by invitation”, “Recruiting”, or “Completed” status.

Enrichment of predicted drug-disease pairs among previously indicated drug-disease pairs

To check for an enrichment of predicted drug-disease pairs among previously indicated drug-disease pairs for each disease, we tallied the total number of unique drugs previously indicated to any disease, the number of those drugs indicated to the disease of interest, the number of drugs predicted to treat the disease by our method, and the number of drugs predicted to treat the disease by our method that were also previously indicated for that disease. We calculated a p-value using a one tailed Fisher’s exact test, and corrected for multiple comparisons within each disease across drugs using the Benjamini-Hochberg procedure.

Enrichment of anti-inflammatory drugs and immunosuppressants among predicted treatments

We searched the DrugBank database for the ATC codes for anti-inflammatory drugs and immunosuppressants including Immunosuppressants (L04), Corticosteroids for systemic use (H02), Antiinflammatory and antirheumatic products (M01), and Antihistamines (R06)[33]. We used these codes to pull all of the drugs in these categories from our expert curated drug to target gene database. For each disease we ranked predicted drugs by $-log_{10}(SAveRUNNER\ FDR)$ and used the fgsea R package (v 1.20.0) to perform gene set enrichment analysis for drugs belonging to each of the four classes[34,35].

Reference to tables created for this chapter

In this chapter we reference multiple tables that display genelists for the complex diseases and UK Biobank GWAS. These tables are too big to display in the chapter and can be downloaded from the Github repository: (https://github.com/krishnanlab/chronic-inflammation/tree/main/figures/supplemental).

Results

Expanding lists of disease-related genes and identifying disease-specific gene subnetworks

Our first goal was to establish a comprehensive list of genes associated with the complex diseases of interest and resolve the genes linked to each disease into subsets of tightly-connected genes in an underlying molecular network. Towards this goal, we selected 37 complex diseases associated with underlying systemic inflammation (**see methods**). To ensure that we correctly isolate chronic inflammation (CI) signatures, we devised a set of positive and negative controls. We selected 10 autoimmune disorders as positive controls because autoimmune disorders are characterized by CI and should have an easily identifiable CI gene signature. For negative controls, we selected 113 traits from UK Biobank[17] that are unlikely to be associated with CI (i.e. Right handedness, filtered coffee intake, and distance between home and workplace). The full list of traits is included in supplementary material in Zenodo record (https://zenodo.org/record/6858073).

While thousands of genes may play a role in the etiology of a chronic disease, it is unlikely that all of these genes have been cataloged in available databases such as DisGeNET or identified by GWAS. Hence, we expanded the lists of disease-or-trait-associated genes using GenePlexus[10] (**Figure 2.1A, step 1**). Briefly, GenePlexus performs supervised machine learning using network-based features to predict novel genes related to a set of input seed genes. Here, we built one GenePlexus model per disease using disease-associated genes from DisGeNET or trait-associated genes from the UK Biobank GWAS as seed genes (positive examples). To test the robustness of this method for identifying CI enriched clusters, we tested four different biological interaction networks of varying sizes and edge densities — STRING, STRING with only experimentally derived edges (STRING-EXP)[20], BioGRID[21], and ConsensusPathDB[22] (**Figure 2.1A, step 1, see methods** section “GenePlexus”) Genes predicted by the GenePlexus model with a probability ≥ 0.80 were added to the seed gene list to create an expanded list of disease- or trait-associated genes. **Figure 2.2** shows results for ConsensusPathDB. The proportion of genes predicted by GenePlexus

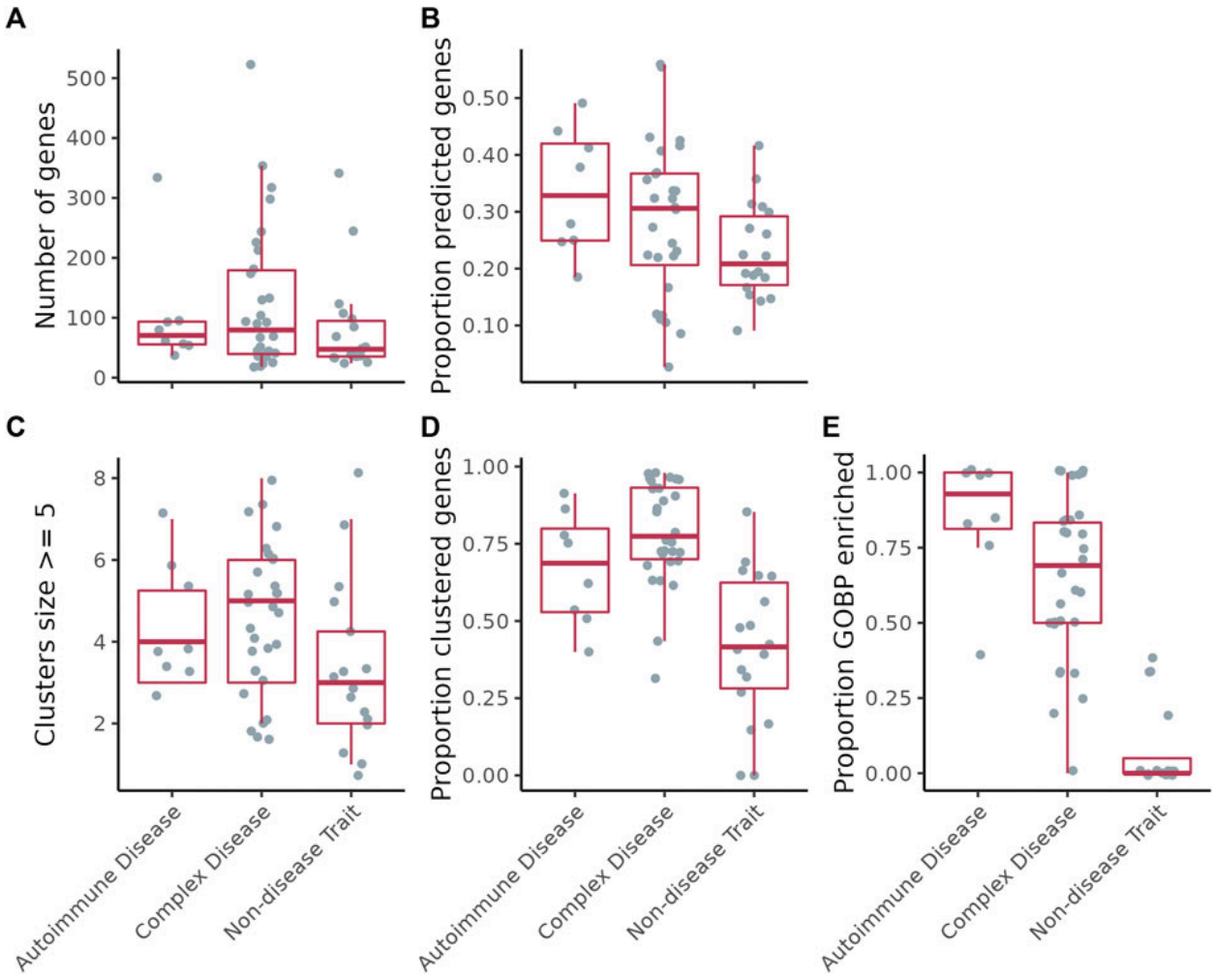

Figure 2.2: (A) Number of genes per disease/trait. **(B)** Proportion of the genes per disease/trait that were predicted by GenePlexus. **(C)** Number of clusters per disease/trait containing at least 5 genes. **(D)** Proportion of total genes assigned to a cluster containing at least 5 genes. **(E)** Proportion of clusters per disease/trait enriched with genes from at least one GO biological process.

for the non-disease traits is lower than those for the autoimmune and complex diseases (**Figure 2.2B**). This observation indicates that genes associated with a specific autoimmune/complex disease tend to have more similar network neighborhoods than genes associated with non-disease traits. All disease-associated genes after GenePlexus prediction are listed in **Table S3**.

Next, for each disease/trait, we clustered the expanded lists of genes based on their interactions in the gene-gene interaction network (**Fig 2.1 A step 2** and **Fig 2.2C**; **Table**

S3). On ConsensusPathDB, the complex diseases had the highest proportion of genes grouped into clusters of ≥ 5 genes, followed by autoimmune diseases and non-disease traits (**Figure 2.2D**). To assess whether clusters are biologically meaningful, we performed an enrichment analysis between every cluster and hundreds of GO Biological Process (GOBP) gene sets. We theorize that significant enrichment of a cluster with a GOBP means the genes in the cluster likely function together to carry out a specific cellular process or pathway. On ConsensusPathDB, for autoimmune and complex diseases, the median proportion of GOBP enriched clusters are > 0.75 and > 0.60, respectively, suggesting most clusters are biologically relevant (**Figure 2.2E**). In contrast, most clusters in non-disease traits are not enriched for a GOBP (**Figure 2.2E**).

Isolating CI-enriched disease clusters

Clusters of related, disease-associated genes on functional gene interaction networks are likely to correspond to the pathways and biological processes disrupted during disease progression. For complex disorders, multiple pathways are likely to be affected. Our next goal was to identify which cluster(s) within a set of disease-associated genes corresponds to the CI component of the disease. For this analysis, similar to the expansion of disease- or trait-associated genes, we used GenePlexus to predict novel inflammation genes for each of the 5 sets of inflammation-related seed genes (**see methods** section “Defining chronic-inflammation-associated genes”, **Table S4**). We then scored the enrichment of CI genes in each disease cluster and performed a permutation test using 5,000 random gene sets for each disease to determine the significance of the enrichment score (**see methods** sections “Creating random traits” and “Finding CI-gene enriched disease clusters”, **Figure 2.1A steps 3-5**, **Table S5**).

With various base networks and CI gene sources, we tested all network–CI-geneset combinations and chose the one that resulted in the highest proportion of autoimmune diseases and lowest proportion of non-disease traits with at least one CI-enriched cluster. Based on this test, we picked ConsensusPathDB as the base network and ‘high-confidence Geneshot’ as the source of CI genes (**Figure A2.2**). We were able to identify clusters enriched for CI genes in all of the autoimmune disorders surveyed (9/9), while finding no CI-enriched clusters among the non-disease traits (**Figure 2.3A**). We identified at least one CI-enriched cluster in 18 of 30 of the complex diseases (**Figure**

2.3A). Twelve out of the 27 diseases with at least one CI-enriched cluster had two or more CI-enriched clusters, and the median proportion of CI-enriched clusters out of the total clusters is higher for autoimmune diseases than complex diseases (**Figure 2.3B**). The number of diseases with at least one CI-enriched cluster varied with different combinations of prediction network, cluster network, and inflammation gene set (**Table S6**). In every case, however, the proportion of autoimmune diseases with at least one CI-enriched cluster was higher than that for non-disease traits suggesting that our method is robust to changes in base-network and inflammation geneset (**Figure A2.2**).

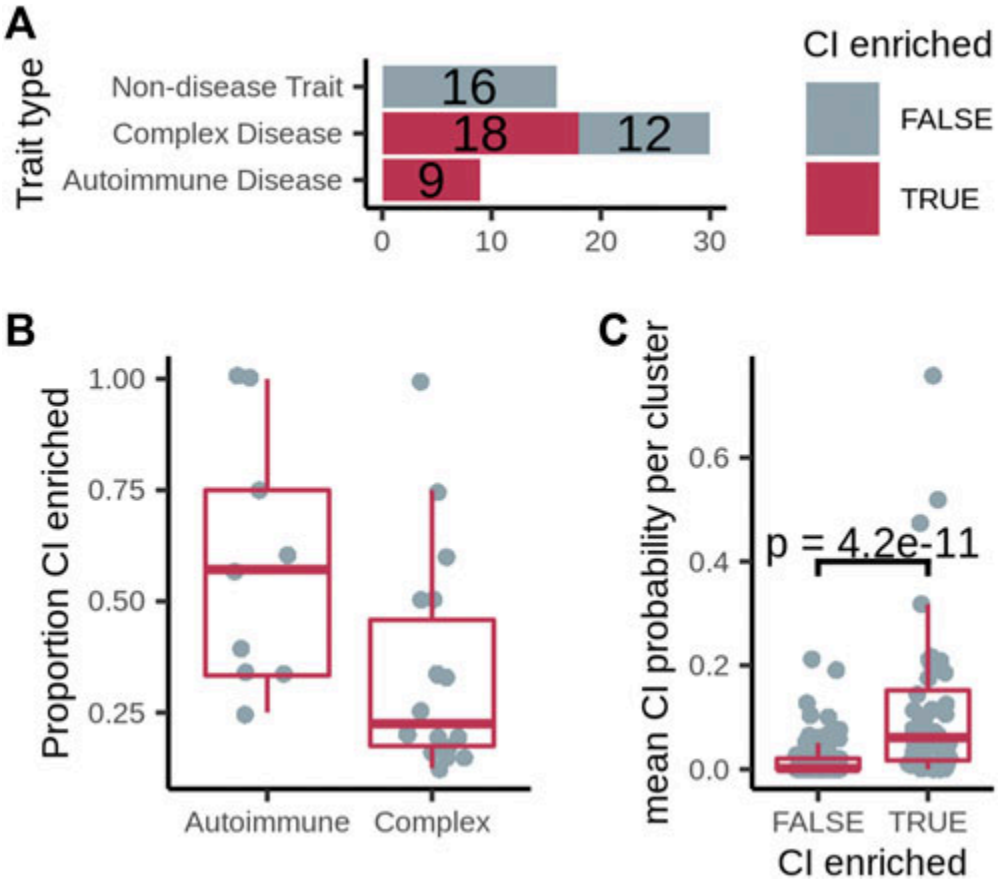

Figure 2.3: **(A)** Number of diseases/traits with at least one cluster overlapping the expanded chronic inflammation (CI) geneset (dark pink), out of the total number of diseases/traits. **(B)** The proportion of CI-enriched disease clusters among all disease clusters per disease. **(C)** Mean probability that genes with no known relationship with chronic inflammation residing in a CI-enriched cluster or non-CI-enriched cluster are associated with CI. P-value calculated using a one-sided Fisher's Exact test.

We hypothesized that, through guilt-by-association, even the genes with no known relationship with chronic inflammation residing in a CI-enriched cluster should have a higher probability of being CI-associated than those in non-CI-enriched clusters. To test this hypothesis, we used GenePlexus with features from each gene-gene interaction network to calculate the probability that every gene is associated with each inflammation gene set. Then, focusing on the genes in disease clusters that were not present in the inflammation geneset, we found that the mean CI probability of these genes in CI-enriched clusters is significantly higher for CI-enriched clusters than non-enriched clusters in 24 out of 25 network/CI-geneset combinations (**Figure A2.3-S7**), including ConsensusPathDB with the high-confidence Geneshot CI geneset (**Figure 2.3C**). This observation suggests that the CI-enriched clusters as a whole, and not just the genes in the high-confidence Geneshot CI geneset residing within them, are CI-associated in the disease of interest. Knocking out putative inflammation associated genes in animal models of the diseases the genes have already been associated with and testing for an increase in known inflammation markers would confirm this result.

Comparing CI gene signatures across diseases

To determine if related diseases have similar chronic inflammation signatures, we used a network-based approach to quantify the similarity between each pair of ConsensusPathDB/high-confidence GeneShot CI-enriched disease clusters across diseases and grouped similar clusters together using the Leiden algorithm[24,36] (**Figure 2.1B, steps 1-2**). Several diseases have more than one CI-enriched cluster and none of these diseases have clusters belonging only to one group (**Figure 2.4A, Table S7**). Moreover, diseases belonging to the same broad category — i.e. autoimmune, cancer, or cardiovascular disease — do not have a larger proportion of clusters belonging to a particular group than expected by chance (one-sided Fisher's exact test, **Figure 2.4A**). This suggests that one disease can harbor more than one type of chronic-inflammation signature, and that the same signatures can be found in very different diseases. For example, rheumatoid arthritis, myocardial ischemia, atherosclerosis and chronic obstructive airway disease all have CI-enriched clusters belonging to each of the three signature groups. To determine the biological significance of these signature groups, we performed enrichment analyses for genes unique to each group among GO biological

processes (**Figure 2.4B, Table S8**). The top 10 significantly enriched terms for each group are largely distinct, with group 1 being enriched for immune relevant signaling pathways, group 2 for regulation of immune cell proliferation, and group 3 for regulation of immune cell chemotaxis (**Figure 2.4B**).

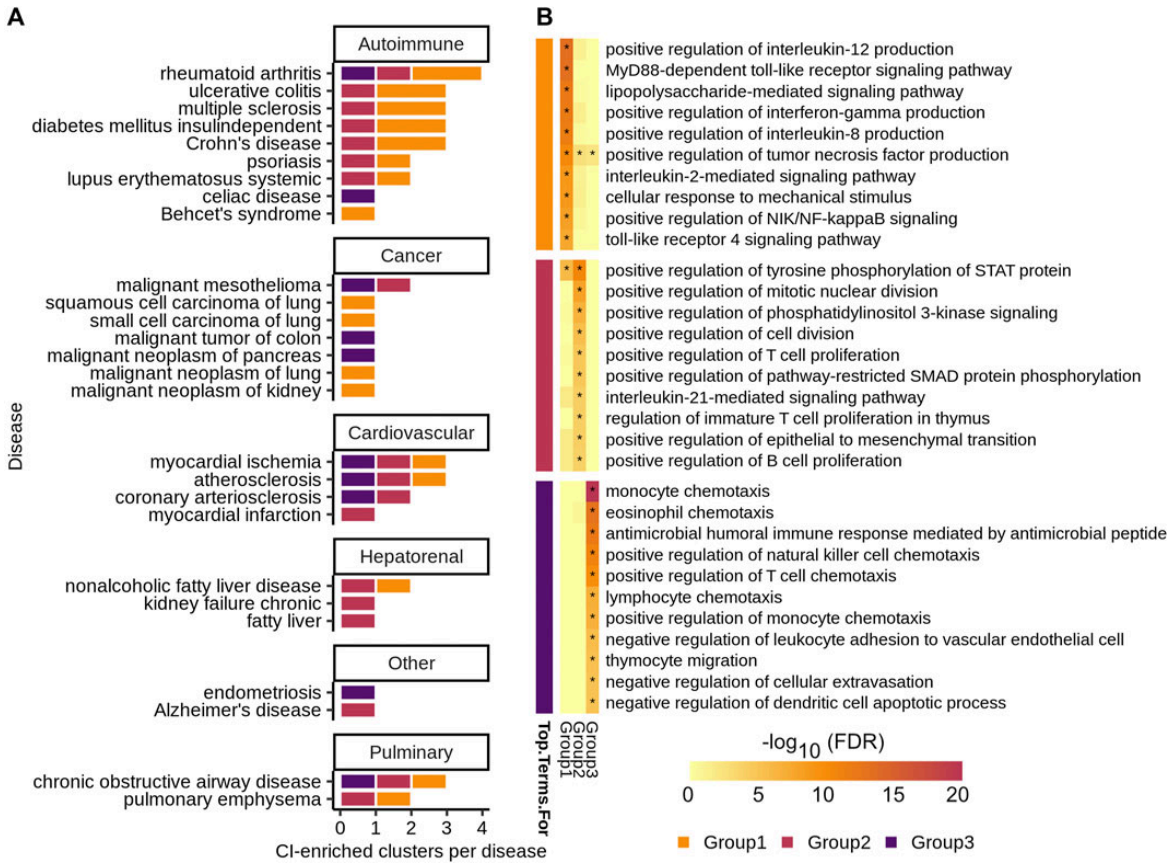

Figure 2.4: (A) Number of CI-enriched clusters per disease colored by CI-signature group. **(B)** Top ten enriched GOBP categories by Benjamini-Hochberg procedure corrected FDR for each CI-signature group — the group is denoted by the colored blocks to the left of the heatmap. The heatmap shows the -log10(FDR) of the enrichment for each CI-signature group — * denotes $FDR < .05$.

Predicting novel treatment opportunities

Our final goal was to leverage the ConsensusPathDB/high confidence GeneShot CI-enriched disease clusters we discovered to find potential avenues for repurposing approved drugs to therapeutically target systemic inflammation underlying complex

diseases (**Fig 2.1B, step 3**). Towards this goal, we used SAveRUNNER to find associations between CI-enriched clusters and FDA approved drugs through each drug's target genes[36]. We found that SAveRUNNER predictions for known treatments were better than random chance — $log_2(auPRC/prior > 0)$ — for diseases with more than five known treatments (**Figure 2.5A**). Moreover, with the exception of myocardial ischemia, SAveRUNNER predicted drugs in Phase IV clinical trials better than random chance (**Figure 2.5A**)[32]. Drugs in Phase IV are those that have already been proved effective for treating a disease (in Phase III) and are being monitored for long-term safety and efficacy.

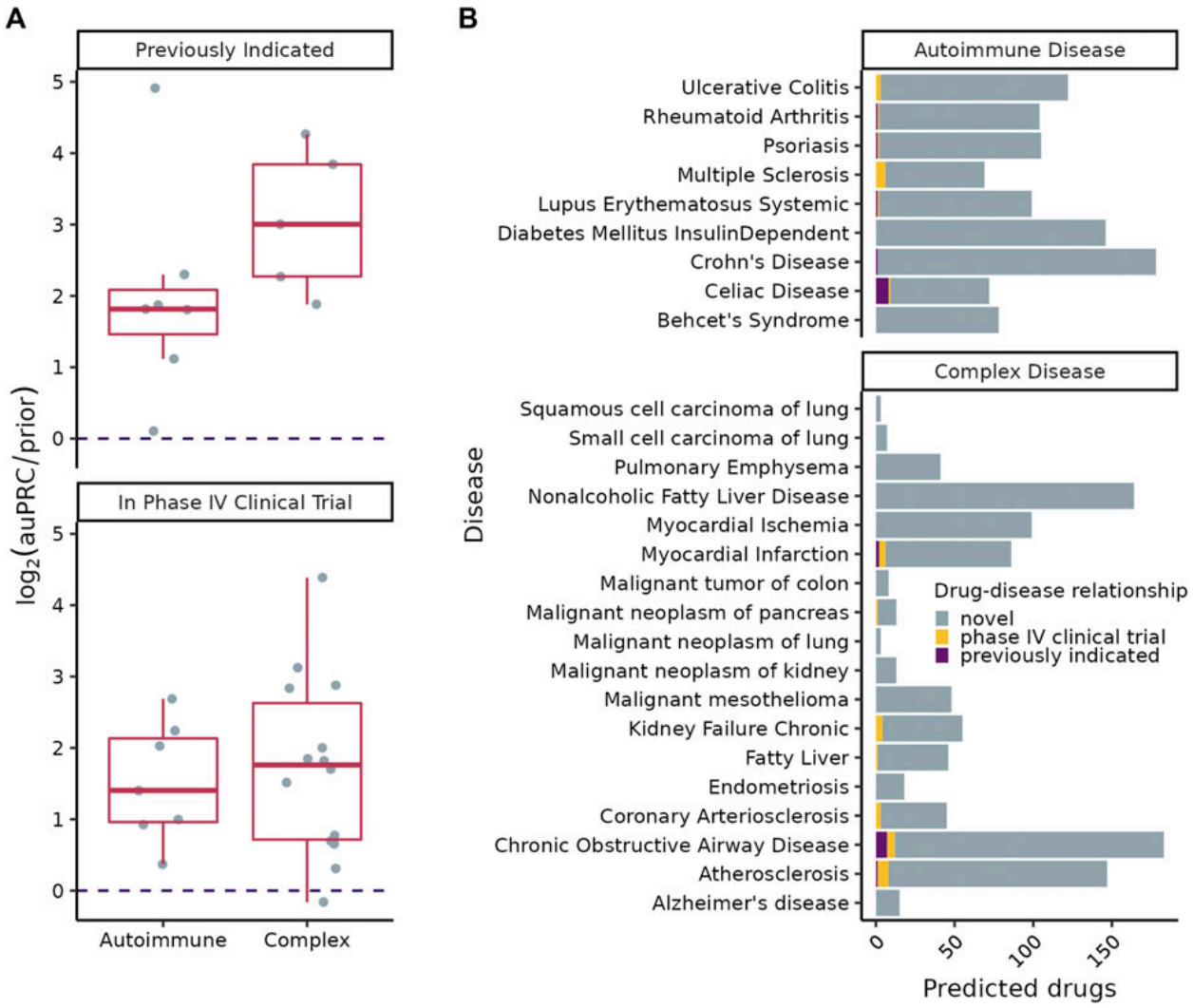

Figure 2.5: (A) $log2(auPRC/prior)$ of SAveRUNNER predictions using drugs previously indicated for the disease (top) or drugs ever in Phase IV clinical trials for a

Figure 2.5 (cont'd)

disease (bottom) as positive examples. The dotted line is at $log2(auPRC/prior) = 0$. $log2(auPRC/prior) > 0$ denotes predictions better than random chance. **(B)** Number of SAveRUNNER predicted genes (Benjamini-Hochberg procedure corrected $FDR < .01$) per disease.

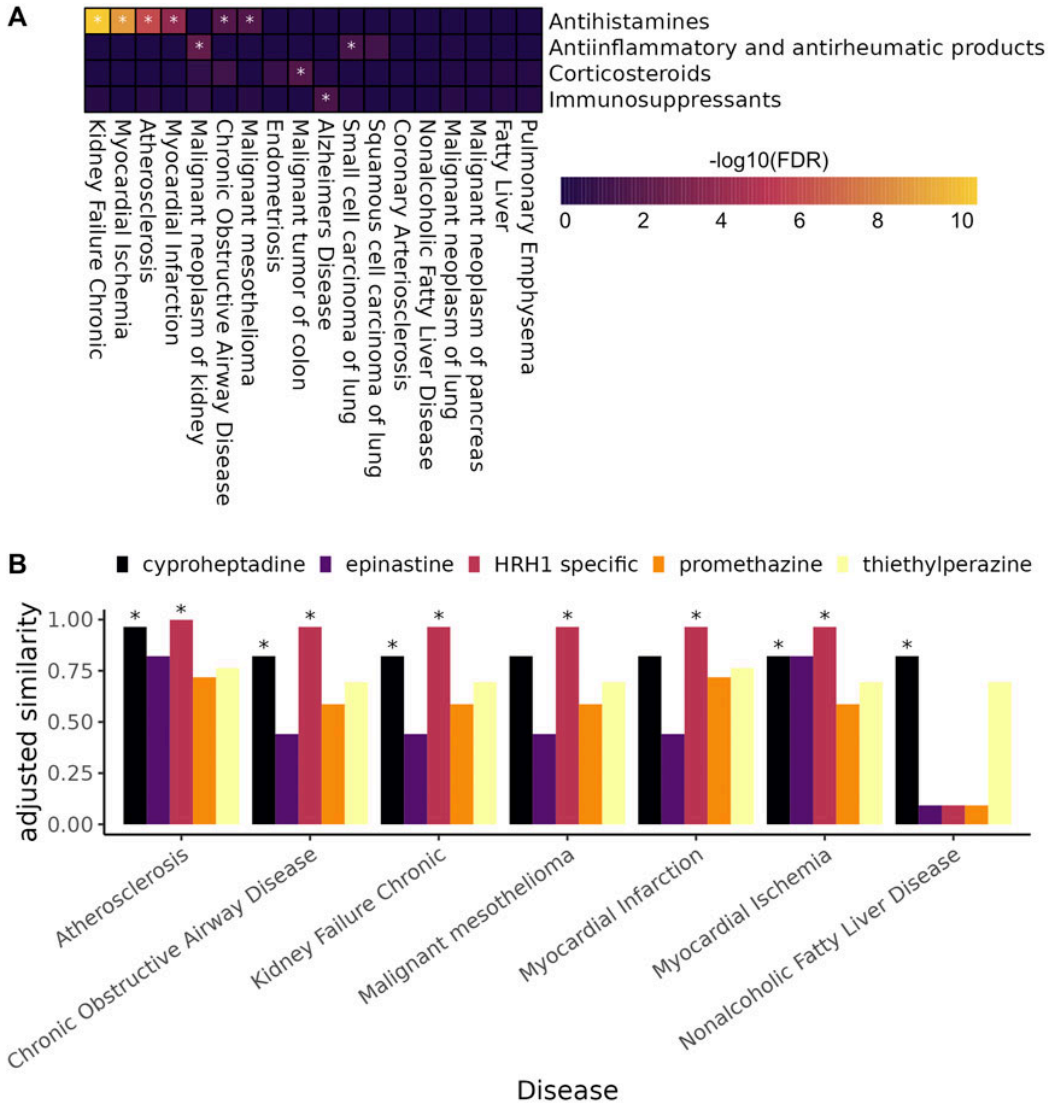

Figure 2.6: (A) Heat map showing the enrichment of anti-inflammatory and immunomodulating drugs among highly ranked SAveRUNNER predicted drugs (gene set enrichment analysis, * denotes adjusted p-value < .05). **(B)** Bar plot showing the adjusted similarity scores of antihistamines for complex diseases with at least one

Figure 2.6 (cont'd)

antihistamine among drugs predicted by SAveRUUNER to treat the disease — * denotes $FDR < .01$. HRH1 specific antihistamines are those listed in our high-confidence drug target database as only targeting *HRH1.*

SAveRUNNER predicted between 3 and 178 high-confidence ($FDR < .01$) treatments for each disease and identified previously indicated drugs for 5 of the 9 autoimmune disorders (**Figure 2.5B, Table S9**), with significant enrichment among drug predictions for celiac disease (one-sided Fisher's exact test, BH corrected $FDR < .001$). SAveRUNNER found previously indicated treatments for only 3 of the 18 complex diseases (**Figure 2.5B, Table S9**). This result is expected given that, unlike for autoimmune disorders, most known treatments for these complex disorders are not likely to target the immune system. Treatments previously tested in a clinical trial were predicted for 6 autoimmune disorders and 7 of the complex disorders (**Figure 2.5B**).

We tested for enrichment of drugs belonging to four immune-related drug classes among treatment predictions highly ranked by SAveRUNNER for each complex disorder (**Figure 2.6A**). SAveRUNNER allows for drug prioritization based both on the p-value and on the adjusted similarity score between drug target genes and CI-enriched cluster genes. Highly scoring drug-cluster pairs have genes that are closely related in the gene interaction network, which increases the likelihood that the drug will be on-target for the paired disease[36]. We found that antihistamines as a whole are enriched for 6 of the 18 complex disorders (**Figure 2.6A**). Antihistamines that specifically target H_1-receptor (*HRH1)* have the highest adjusted similarity score for 6 of the 7 complex disorders with any antihistamine among their high-confidence targets (**Figure 2.6B**). SAveRUNNER predicted that cyproheptadine, which targets both *HRH1* and the serotonin 5-HT(2A) receptor gene, *HTR2A*, instead of *HRH1* alone would be the best antihistamine for treating non-alcoholic fatty liver disease (**Figure 2.6B**). While cyproheptadine is also a high-confidence predicted treatment for atherosclerosis, myocardial ischemia, and chronic obstructive airway disease, it is unlikely to be an effective treatment for myocardial infarction or malignant mesothelioma (**Figure 2.6B**). Interestingly, of the 8 diseases, only myocardial infarction and malignant mesothelioma do not have a CI-enriched cluster belonging to CI-signature group 2 (**Figure 2.4A**). This finding

suggests that, even among drugs in the same class, we are able to predict disease-specific treatments for the chronic inflammation component of the disease etiology.

Discussion

Complex diseases exhibit a staggering amount of heterogeneity, being associated with hundreds of genes and with a range of phenotypes. Therefore, to continue advancing our understanding of disease mechanisms and our ability to treat these diseases, it is critical to deconvolve disease heterogeneity by: a) resolving subset of disease genes (and cellular processes/pathways) that underlie specific disease-associated phenotypes, and b) identifying avenues to diagnostically and/or therapeutically target those specific phenotypes. Here, we present a computational data-driven approach to address this critical need (**Figure 2.1**). We used our approach to study chronic inflammation (CI) — a major phenotype present across many complex diseases. We generated comprehensive lists of (known and predicted) disease-associated genes and identified and classified the CI signal among these genes. We used these signatures to predict novel treatment options to target the inflammatory components of 18 complex diseases.

Validating our method with autoimmune disorders

A key aspect of our approach is ensuring its sensitivity to detect CI disease signatures using autoimmune diseases as positive controls. In autoimmune diseases, the immune system mistakenly attacks healthy tissue causing long-term systemic inflammation. Thus, we expect that the underlying CI disease signatures would be easily identifiable by a valid approach. Indeed, in each of the nine autoimmune diseases analyzed, our approach isolated gene clusters enriched for CI genes (**Figure 2.1A**), and identified drugs already used to treat a number of these disorders (**Figure 2.5B**). This finding is encouraging given that we conservatively matched drugs to diseases only based on expert-curated drug-target data from DGIDb[31] rather than using all drug-target information in DrugCentral[30].

Non-disease GWAS are contrasted to disease GWAS in biological networks

To show that our method was not erroneously uncovering CI signals where there were none, we identified UK Biobank traits not patently associated with CI (along with their

genes) to use as negative controls. Following this analysis, we found that the median fraction of trait-associated genes predicted by GenePlexus and the median fraction of genes assigned to sizable clusters were lower for these traits than for autoimmune and complex diseases (**Figure 2.2C** and **D**). Given that GenePlexus is a method that leverages network connectivity for predicting new genes belonging to a set, these results suggest that the genes associated with non-disease traits may not be as highly connected to one another in ConsensusPathDB as the autoimmune and complex disease genes. Moreover, most of the non-disease trait clusters were not enriched with genes annotated to GO biological processes, suggesting that these clusters are diffuse and that the member genes are unlikely to work together to support a coherent biological task. While non-disease traits like coffee intake and handedness have been associated with inflammation[37,38], this analysis (using GWAS-based trait-associated genes) suggests it is unlikely that SNPs in a coordinated inflammation pathway influence non-disease traits and more likely that any association with inflammation is environmental, not genetic. Taken together, these results suggest that these chosen traits serve as reasonable negative controls and offer a way to meaningfully contrast the results from complex diseases. Ideally, diseases or traits with no underlying inflammatory component but with associated genes that cluster in a network (as well as the autoimmune and complex disease) will serve as better negative controls. Given how common inflammatory processes are in disease, however, such diseases are difficult to definitively identify.

CI pathways are shared across diseases

Complex disorders like cardiovascular diseases, diabetes, cancer, and Alzheimer's disease are among the leading causes of death and disability among adults over 50 years of age, and all are associated with underlying systemic inflammation [2,3]. Patients with systemic inflammation caused by autoimmune disorders are more likely to have another CI disorder like cardiovascular disease, type 2 diabetes mellitus, and certain types of cancer[11–13]. Further, treating one chronic-inflammatory disease can reduce the risk of contracting another, suggesting a common underlying pathway [39]. For example, treating rheumatoid arthritis with tumor necrosis factor (TNF) antagonists lowers the incidence of Alzheimer's disease and type II diabetes[14,40].

To better understand how CI-associated disorders relate to one another, we used a network-based approach to quantify the similarity between their CI-enriched clusters. We hypothesized, for example, that Crohn's disease and “malignant tumor of colon” would have similar CI-signatures, given that patients with inflammatory bowel disease are at increased risk for developing colorectal cancer[41]. However, Crohn’s disease CI-enriched clusters are members of signature groups 1 and 2, while the “malignant tumor of colon” CI-enriched cluster belongs to group 3 (**Figure 2.4A**). Instead of sharing CI-signatures, related CI diseases may, instead, have complementary signatures. Indeed, the group 1 signature, which characterizes two of the three Crohn’s disease CI-enriched clusters, is enriched for genes that positively regulate proinflammatory cytokines like tumor necrosis factor (TNF) and in interferon-gamma (IFNɣ) (**Figure 2.4B**). When these cytokines bind to their respective receptors, reactive oxygen species are generated causing oxidative stress[42]. Oxidative stress, in turn, induces DNA-damage that can induce tumor formation. Colorectal tumors are infiltrated with lymphocytes, which mediate the recruitment of immune cells that suppress tumor growth[43]. Immune cell infiltration likely leads to our ability to detect the group 3 CI-signature among genes associated with “malignant tumor of colon”, given that group 3 is enriched for immune cell migration and chemotaxis (**Figure 2.4A** and **B**). Alternatively, there is a possibility that every CI-associated disease actually exhibits all three CI-signatures, and our method is only sensitive enough to detect these in a handful of diseases.

Predicting treatments for CI components of disease reveals meaningful relationships

Common treatments for systemic inflammation, including non-steroidal anti-inflammatory drugs (NSAIDs), corticosteroids, and biologics like tumor necrosis factor (TNF) antagonists, can cause adverse effects when used long term. For instance, patients treated with corticosteroids or TNF antagonists have increased risk of infection[41,44,45], and corticosteroid use increases both the risk of fracture [46,47] and the risk of developing type II diabetes[48]. NSAIDs present a unique set of side effects, particularly in elderly patients, including gastrointestinal problems ranging from indigestion to gastric bleeding, and kidney damage[49–51]. Therefore, the search for better treatment options for

CI is ongoing. Here, we leverage the CI-signatures to identify novel treatment opportunities for the CI-component of 18 complex diseases (**Figure 2.5B**). Interestingly, antihistamines were among the top drug associations for 6 of 18 complex diseases (**Figure 2.6A**), including atherosclerosis. Atherosclerosis is characterized by the deposition of cholesterol plaques on the inner artery walls. Mast cells, immune cells best known for their response to allergens, are recruited to arteries during plaque progression, where they release histamines. Histamines then activate the histamine H_1-receptor, increasing vascular permeability, which allows cholesterol easier access to arteries promoting plaque buildup[52]. Mepyramine, one of the HRH1-specific antihistamines highly associated with atherosclerosis, has already been shown to decrease the formation of atherogenic plaques in a mouse model of atherosclerosis[52]. Interestingly, it is not predicted as a treatment for myocardial ischemia, which occurs when plaque buildup obstructs blood flow to a coronary artery, suggesting disease-specific antihistamine efficacy even among related diseases. Cetirizine and fexofenadine are also HRH1-specific antihistamines highly associated with atherosclerosis but neither prevented or reduced atherosclerosis progression in a mouse model of atherosclerosis, and both increased atherosclertotic lesions at low doses[53]. In the expert-curated drug-target database used in this study, the histamine H_1-receptor is the only target listed for all three drugs; however, the contradictory results from Rosenberg *et al.* and Raveendran *et al.* suggests that drug-specific off-target effects are mediating atherosclerosis treatment outcomes. A more complete understanding of drug-gene targets would allow for better predictions of novel disease treatments.

For example, unlike the other diseases with antihistamines as predicted treatments, only cyproheptadine, and not the HRH1-specific drugs, is likely to be an effective treatment for non-alcoholic fatty liver disease (NAFLD) (**Figure 2.6B**). Cyproheptadine is an antagonist for both the *HRH1* and the serotonin 5-HT(2A) receptor (*HTR2A)*, suggesting that blocking 5-HT(2A) could be specifically helpful for ameliorating symptoms of NAFLD. Indeed, liver-specific *Htr2a* knockout mice are resistant to HFD-induced hepatic steatosis, increased fat in the liver [54], and increased serum

serotonin levels were correlated with increased disease severity in patients with NAFLD[55].

Conclusion

Overall, we have shown that our method is capable of isolating the chronic inflammation gene signature of a complex disease using a network-based strategy and, by integrating information across multiple complementary sources of data, it can predict and prioritize potential therapies for the systemic inflammation involved in that specific disease. Importantly, our approach provides a blueprint for identifying and prioritizing therapeutic opportunities for any disease endophenotype. This work has been published[56]

This project was foundational for validating the usefulness of using gene modules and networks to isolate particular processes within diseases. The pipeline has multiple steps, including module discovery, gene classification, and analyzing important processes and phenotypes at the module level. We used well validated methods for this task – being leiden clustering for module discovery and GenePlexus for gene classification. However, during this project we started considering how to improve various aspects of this pipeline. For module discovery, we are heavily reliant on the gene annotations from DisGeNet and did not implement a method to remove genes that external sources imply are low evidenced. Similarly, we run GenePlexus on each disease as a whole. However, given that we view diseases as collections of biologically related phenotypes and cellular pathways, this may be an inappropriate way to run GenePlexus. For example, in the analysis performed in **Figures A2.3-7,** we see that non-CI genes in CI clusters are more likely to be associated with CI than genes in non-CI clusters. If it is known that CI is part of a disease, it may make more sense to perform gene classification on subsets of a disease that correspond to specific phenotypes, rather than the disease as a whole. This would have implications for doing enrichment on clusters and discovering/annotating phenotypes at a modular level. Deeply considering the steps of chapter 2's pipeline is what motivates chapters 3 and 4 of this dissertation.

REFERENCES

1. Rock, K. L. & Kono, H. The inflammatory response to cell death. *Annu Rev Pathol* **3**, 99–126 (2008).

2. Furman, D. *et al.* Chronic inflammation in the etiology of disease across the life span. *Nat Med* **25**, 1822–1832 (2019).

3. Vos, T. *et al.* Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet* **396**, 1204–1222 (2020).

4. Leiserson, M. *et al.* Pan-Cancer Network Analysis Identifies Combinations of Rare Somatic Mutations across Pathways and Protein Complexes. *Nature Genetics* **47**, 106–114 (2015).

5. Ghiassian, S. D., Menche, J. & Barabási, A.-L. A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome. *PLOS Computational Biology* **11**, e1004120 (2015).

6. Ghiassian, S. D. *et al.* Endophenotype Network Models: Common Core of Complex Diseases. *Sci Rep* **6**, 27414 (2016).

7. Fiscon, G. & Paci, P. SAveRUNNER: an R-based tool for drug repurposing. *BMC Bioinformatics* **22**, 150 (2021).

8. Chen, H., Zhang, H., Zhang, Z., Cao, Y. & Tang, W. Network-Based Inference Methods for Drug Repositioning. *Computational and Mathematical Methods in Medicine* **2015**, e130620 (2015).

9. Cheng, F. *et al.* Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nat Commun* **9**, 2691 (2018).

10. Liu, R., Mancuso, C. A., Yannakopoulos, A., Johnson, K. A. & Krishnan, A. Supervised-learning is an accurate method for network-based gene classification. *Bioinformatics* (2020) doi:10.1093/bioinformatics/btaa150.

11. Dregan, A., Charlton, J., Chowienczyk, P. & Gulliford, M. C. Chronic Inflammatory Disorders and Risk of Type 2 Diabetes Mellitus, Coronary Heart Disease, and Stroke. *Circulation* **130**, 837–844 (2014).

12. Armstrong, A. W., Harskamp, C. T. & Armstrong, E. J. Psoriasis and the Risk of Diabetes Mellitus: A Systematic Review and Meta-analysis. *JAMA Dermatology* **149**, 84–91 (2013).

13. Yashiro, M. Ulcerative colitis-associated colorectal cancer. *World J Gastroenterol* **20**, 16389–16397 (2014).

14. Chou, R. C., Kane, M., Ghimire, S., Gautam, S. & Gui, J. Treatment for Rheumatoid

Arthritis and Risk of Alzheimer's Disease: A Nested Case-Control Analysis. *CNS Drugs* **30**, 1111–1120 (2016).

15. Autoimmune Diseases: Causes, Symptoms, What Is It & Treatment. *Cleveland Clinic* https://my.clevelandclinic.org/health/diseases/21624-autoimmune-diseases.

16. Schriml, L. M. *et al.* Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res* **47**, D955–D962 (2019).

17. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine* **12**, (2015).

18. Abbot, L. *et al.* UK Biobank GWAS. *nealelab* http://www.nealelab.is/uk-biobank/.

19. Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z. & Bergmann, S. Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. *PLOS Computational Biology* **12**, (2016).

20. Szklarczyk, D. *et al.* The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res* **45**, D362–D368 (2017).

21. Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* **34**, D535–D539 (2006).

22. Kamburov, A., Stelzl, U., Lehrach, H. & Herwig, R. The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Research* **41**, D793–D800 (2013).

23. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research* **48**, D845–D855 (2020).

24. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* **9**, 5233 (2019).

25. Alexa, A. & Rahnenfuhrer, J. topGO: Enrichment Analysis for Gene Ontology. Bioconductor version: Release (3.14) https://doi.org/10.18129/B9.bioc.topGO (2022).

26. Carlson, M. org.Hs.eg.db: Genome wide annotation for Human. (2019).

27. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat Genet* **25**, 25–29 (2000).

28. Lachmann, A. *et al.* Geneshot: search engine for ranking genes from arbitrary text queries. *Nucleic Acids Research* **47**, W571–W577 (2019).

29. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).

30. Avram, S. *et al.* DrugCentral 2021 supports drug discovery and repositioning. *Nucleic Acids Research* **49**, 1160–1169 (2021).

31. Freshour, S. *et al.* Integration of the Drug-Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts. *Nucleic Acids Research* **49**, 1144–1151 (2021).

32. AACT Database | Clinical Trials Transformation Initiative. https://aact.ctti-clinicaltrials.org/.

33. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research* **46**, D1074–D1082 (2018).

34. Korotkevich, G. *et al.* Fast gene set enrichment analysis. 060012 Preprint at https://doi.org/10.1101/060012 (2021).

35. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15545–15550 (2005).

36. Fiscon, G., Conte, F., Farina, L. & Paci, P. SAveRUNNER: A network-based algorithm for drug repurposing and its application to COVID-19. *PLOS Computational Biology* **17**, e1008686 (2021).

37. Paiva, C. *et al.* Consumption of coffee or caffeine and serum concentration of inflammatory markers: A systematic review. *Crit Rev Food Sci Nutr* **59**, 652–663 (2019).

38. Searleman, A. & Fugagli, A. K. Suspected autoimmune disorders and left-handedness: Evidence from individuals with diabetes, Crohn's disease and ulcerative colitis. *Neuropsychologia* **25**, 367–374 (1987).

39. Fullerton, J. N. & Gilroy, D. W. Resolution of inflammation: a new therapeutic frontier. *Nat Rev Drug Discov* **15**, 551–567 (2016).

40. Antohe, J. l. *et al.* Diabetes mellitus risk in rheumatoid arthritis: Reduced incidence with anti–tumor necrosis factor α therapy. *Arthritis Care & Research* **64**, 215–221 (2012).

41. Shah, S. C. & Itzkowitz, S. H. Colorectal Cancer in Inflammatory Bowel Disease: Mechanisms and Management. *Gastroenterology* **162**, 715-730.e3 (2022).

42. Chatterjee, S. Chapter Two - Oxidative Stress, Inflammation, and Disease. in *Oxidative Stress and Biomaterials* (eds. Dziubla, T. & Butterfield, D. A.) 35–58 (Academic Press, 2016). doi:10.1016/B978-0-12-803269-5.00002-4.

43. Idos, G. E. *et al.* The Prognostic Implications of Tumor Infiltrating Lymphocytes in Colorectal Cancer: A Systematic Review and Meta-Analysis. *Sci Rep* **10**, 3360 (2020).

44. Rosenblum, H. & Amital, H. Anti-TNF therapy: Safety aspects of taking the risk.

Autoimmunity Reviews **10**, 563–568 (2011).

45. Murdaca, G. *et al.* Infection risk associated with anti-TNF-α agents: a review. *Expert Opinion on Drug Safety* **14**, 571–582 (2015).

46. Kanis, J. A. *et al.* A Meta-Analysis of Prior Corticosteroid Use and Fracture Risk. *Journal of Bone and Mineral Research* **19**, 893–899 (2004).

47. Mitra, R. Adverse Effects of Corticosteroids on Bone Metabolism: A Review. *PM&R* **3**, 466–471 (2011).

48. Blackburn, D., Hux, J. & Mamdani, M. Quantification of the Risk of Corticosteroid-induced Diabetes Mellitus Among the Elderly. *Journal of General Internal Medicine* **17**, 717–720 (2002).

49. Griffin, M. R. Epidemiology of Nonsteroidal Anti-inflammatory Drug–Associated Gastrointestinal Injury. *The American Journal of Medicine* **104**, 23S-29S (1998).

50. Griffin, M. R., Yared, A. & Ray, W. A. Nonsteroidal antiinflammatory drugs and acute renal failure in elderly persons. *Am J Epidemiol* **151**, 488–496 (2000).

51. Marcum, Z. A. & Hanlon, J. T. Recognizing the Risks of Chronic Nonsteroidal Anti-Inflammatory Drug Use in Older Adults. *Ann Longterm Care* **18**, 24–27 (2010).

52. Rozenberg, I. *et al.* Histamine H1 receptor promotes atherosclerotic lesion formation by increasing vascular permeability for low-density lipoproteins. *Arterioscler Thromb Vasc Biol* **30**, 923–930 (2010).

53. Raveendran, V. V. *et al.* Chronic Ingestion of H1-Antihistamines Increase Progression of Atherosclerosis in Apolipoprotein E-/- Mice. *PLoS One* **9**, e102165 (2014).

54. Choi, W. *et al.* Serotonin signals through a gut-liver axis to regulate hepatic steatosis. *Nat Commun* **9**, 4824 (2018).

55. Wang, L. *et al.* Gut-Derived Serotonin Contributes to the Progression of Non-Alcoholic Steatohepatitis via the Liver HTR2A/PPARγ2 Pathway. *Frontiers in Pharmacology* **11**, (2020).

56. Hickey, S. L., McKim, A., Mancuso, C. A. & Krishnan, A. A network-based approach for isolating the chronic inflammation gene signatures underlying complex diseases towards finding new treatment opportunities. 2022.02.10.479987 Preprint at https://doi.org/10.1101/2022.02.10.479987 (2022).

APPENDIX A2: INFLAMMATION

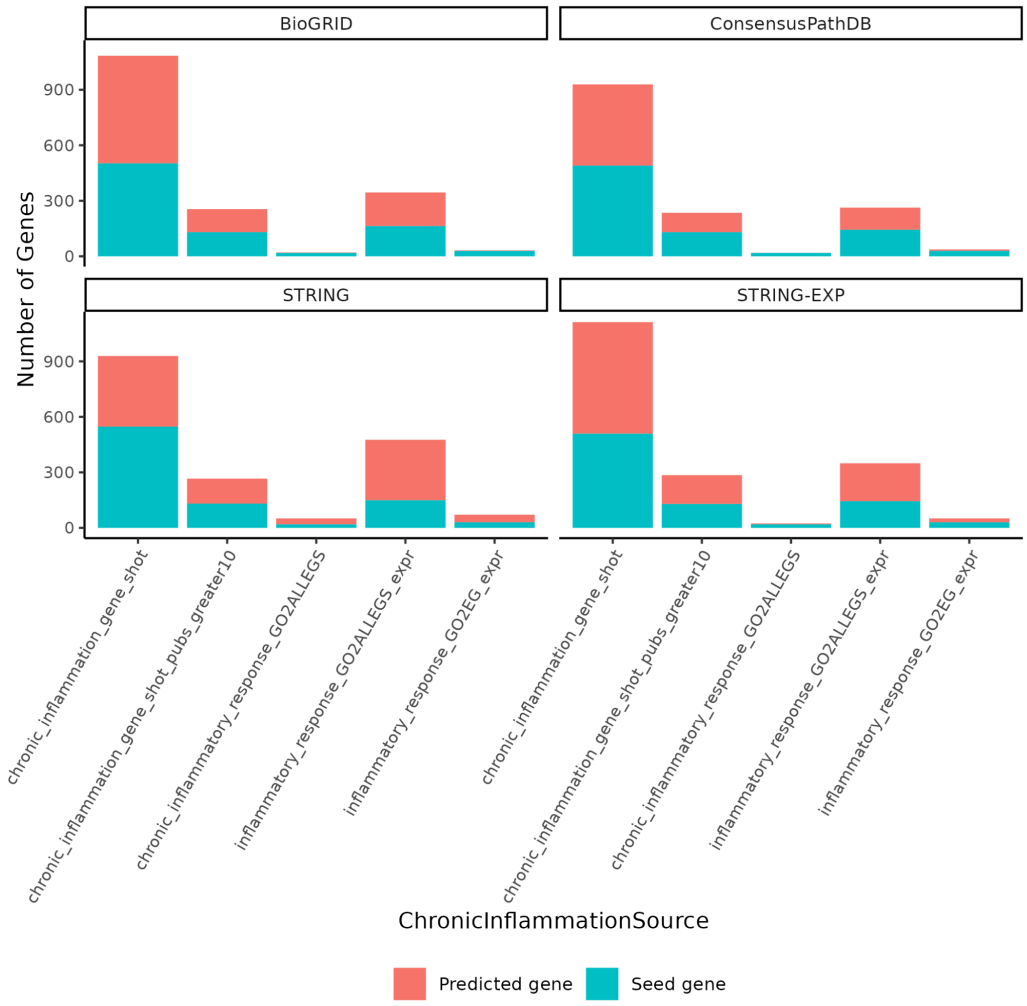

Figure A2.1: Bar charts displaying the number of gene annotations and predicted genes (from GenePlexus) of five different sets for chronic inflammation genes. 1) chronic_inflammation_gene_shot includes all genes from Geneshot for chronic inflammation. 2) chronic_inflammation_gene_shot_pubs_greater10 includes genes from Geneshot that are in at least 10 publications. 3) chronic_inflammatory_response_GO2ALLEGS is the propagated chronic inflammatory response GOBP genes. 4) inflammatory_response_GO2ALLEGS is the propagated inflammatory response GOBP genes. 5) inflammatory_response_GO2EGS is the propagated inflammatory response GOBP genes.

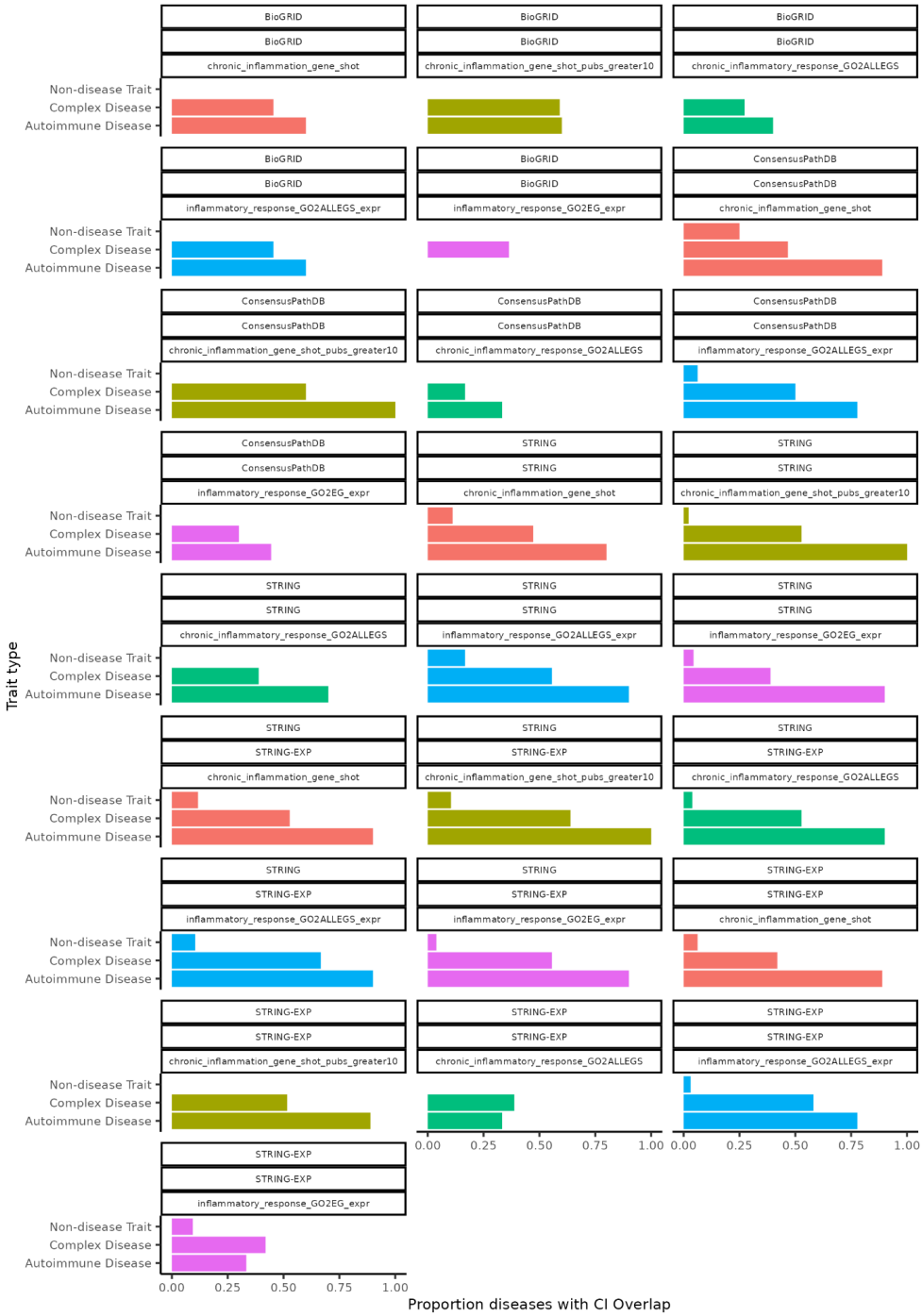

Figure A2.2: Showing proportion of clusters of each trait type with at least one CI-enriched cluster for each network-CI-geneset combination.

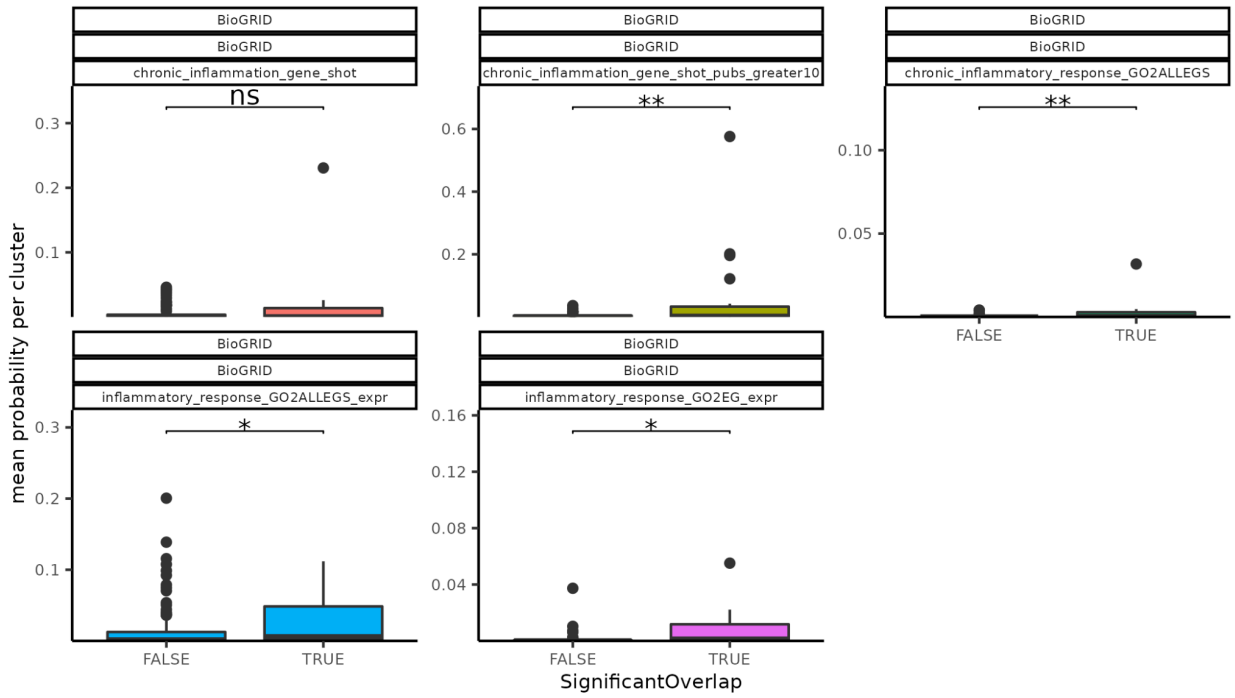

Figure A2.3: For disease gene clusters created from running GenePlexus on the diseases utilizing the BioGRID network, and then clustered using BioGRID, plotting the mean CI probability of the genes of CI-enriched clusters (TRUE) with non-CI enriched clusters (FALSE).

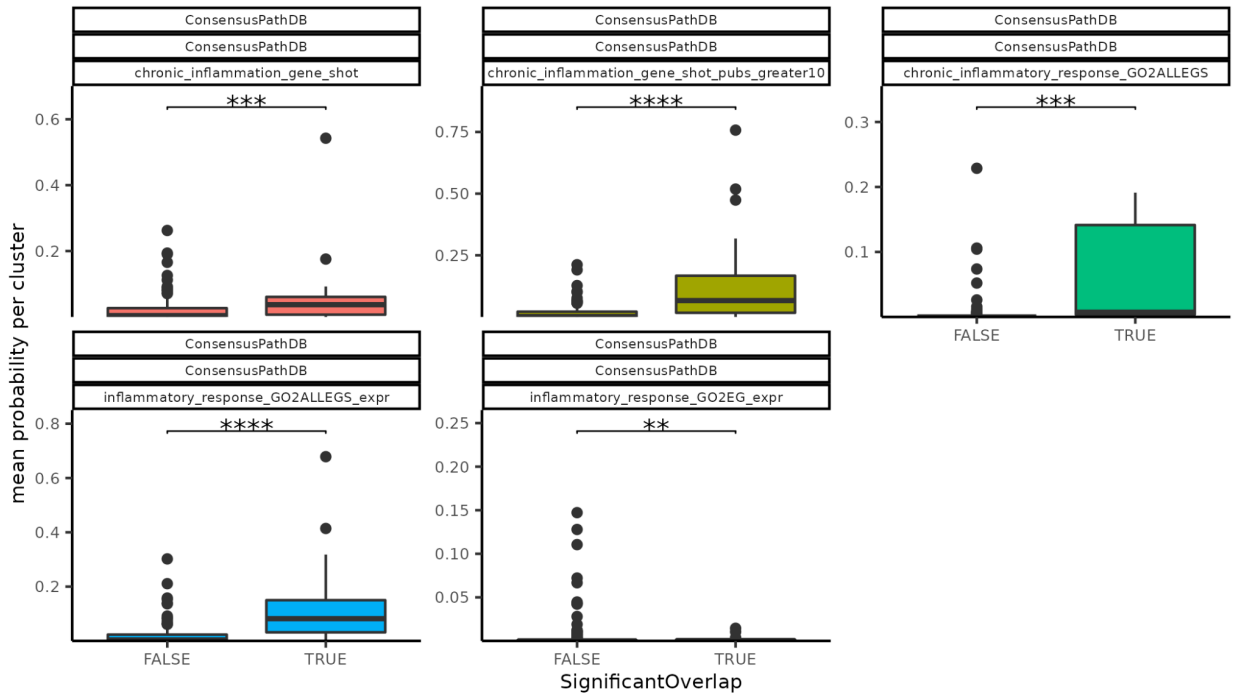

Figure A2.4: For disease gene clusters created from running GenePlexus on the diseases utilizing the ConsensusPathDB network, and then clustered using ConsensusPathDB, plotting the mean CI probability of the genes of CI-enriched clusters (TRUE) with non-CI enriched clusters (FALSE).

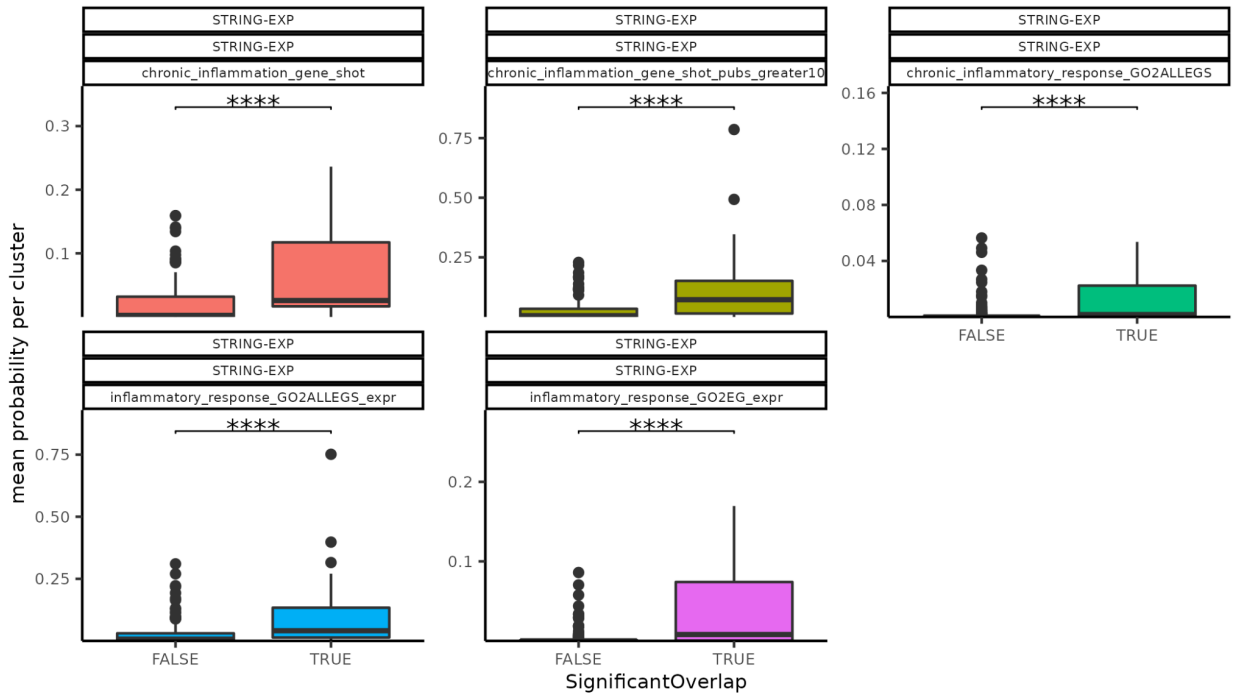

Figure A2.5: For disease gene clusters created from running GenePlexus on the diseases utilizing the STRING-EXP network, and then clustered using STRING-EXP, plotting the mean CI probability of the genes of CI-enriched clusters (TRUE) with non-CI enriched clusters (FALSE).

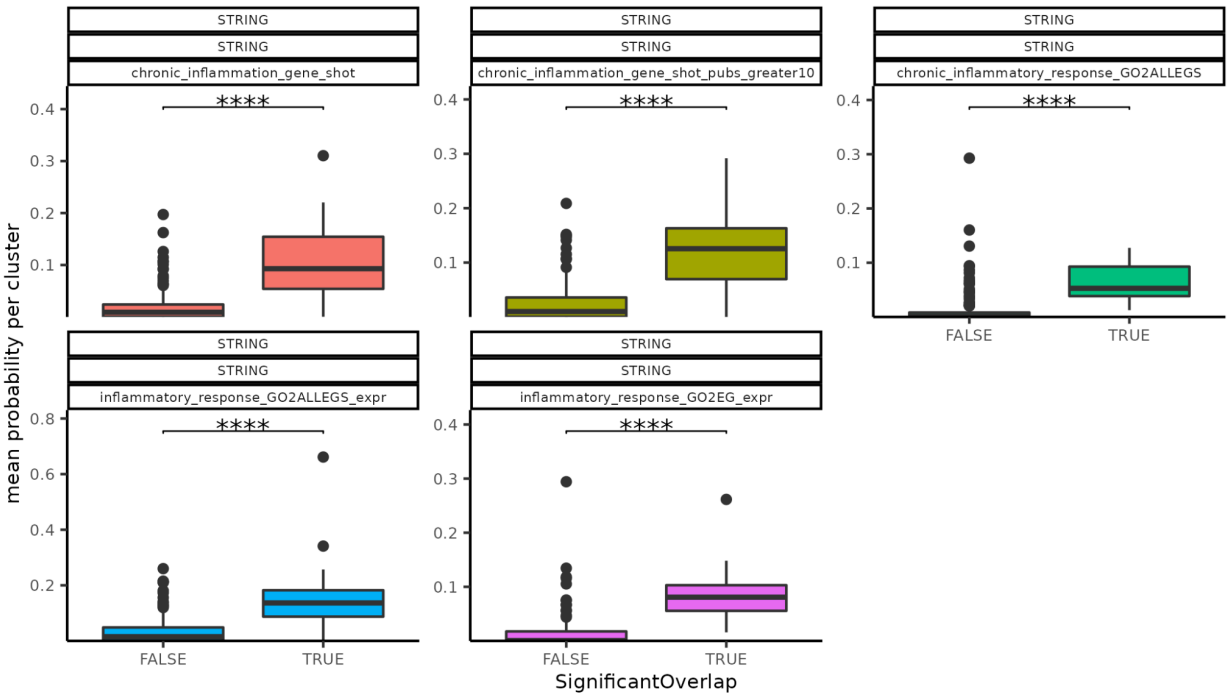

Figure A2.6: For disease gene clusters created from running GenePlexus on the diseases utilizing the STRING network, and then clustered using STRING, plotting the mean CI probability of the genes of CI-enriched clusters (TRUE) with non-CI enriched clusters (FALSE).

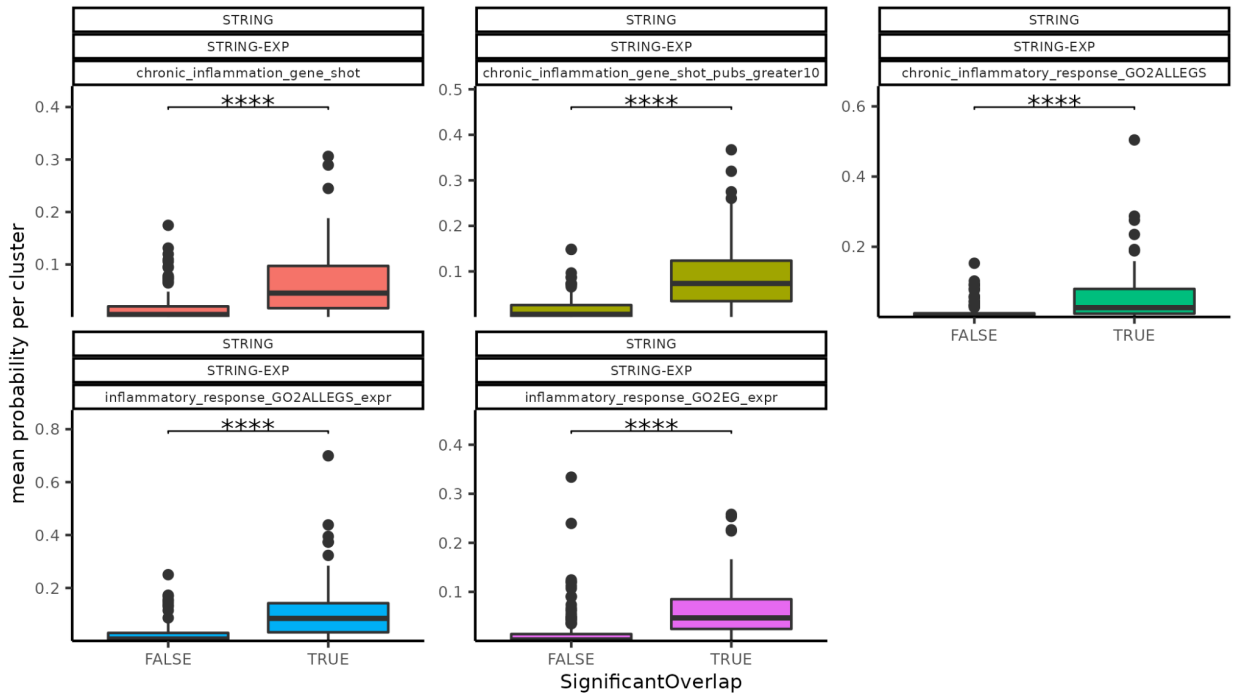

Figure A2.7: For disease gene clusters created from running GenePlexus on the diseases utilizing the STRING network, and then clustered using STRING-EXP, plotting the mean CI probability of the genes of CI-enriched clusters (TRUE) with non-CI enriched clusters (FALSE).

CHAPTER 3: MODGENEPLEXUS: A MODULAR NETWORK-BASED APPROACH FOR GENE CLASSIFICATION IMPROVES POST-OMICS AND POST-GWAS HYPOTHESIS GENERATION FOR DISEASE GENES AND MECHANISMS

Abstract

Network-based gene classification is a method to computationally predict associations of genes to cellular pathways, complex traits, and diseases. Established methods like label propagation and GenePlexus have demonstrated success in using genome-wide networks to classify known genes and uncovering novel associations for various biological datasets. However, directly applying this network-based approach to complex diseases has proven challenging because, unlike genes related to pathways and phenotypes, complex disease genes are not localized in gene networks within a shared neighborhood. This is confirmed by the fact that building a single network-based ML classifier for the disease as a whole leads to poor gene discovery performance. We propose a novel network-based ML approach, ModGenePlexus, to address this challenge. ModGenePlexus works in two stages. First, using semi-supervised learning, it decomposes the large disease gene list into coherent ‘modules’ that each contain a subset of original disease genes and some new candidate genes tightly connected to each other in the underlying network. Second, using supervised learning, ModGenePlexus trains one ML classifier per disease module to learn network patterns unique to that module and predict additional novel genes related to that module. We applied ModGenePlexus to large disease gene lists derived from several transcriptomics studies and GWASs. Using systematic and rigorous evaluations, we demonstrate that ModGenePlexus yields improved performance and facilitates a more nuanced and biologically specific interpretation of the known and novel disease genes. Additionally, we show that, by using networks to provide biological context, ModGenePlexus is capable of using even genes with nominal p-values to improve prediction performance. Lastly, we demonstrate using type-2 diabetes that ModGenePlexus allows for the enrichment of unique and relevant biology that using the experiment result as a whole would miss. These findings underscore the advantages of ModGenePlexus over conventional methods that consider the entire disease gene set as a single unit, as it further refines the geneset and thereby expands the applicability of

network-based ML to post-GWAS or post-omics hypothesis-generation, especially for answering questions about real world data–elucidating functional, cellular, and phenotypic convergence of disease genes.

Introduction

Gene classification is the task of computationally predicting associations of genes to cellular pathways, complex traits, and diseases. In chapter 1 we showed that numerous methods have been created for the purpose of allowing researchers to take genes and predict additional genes likely to be associated with the user list[1–8]. These methods are possible because of large publicly available data collections, including ontologies of biological terms and molecular interaction networks. Two methods of note include label propagation[9] methods, which are a semi-supervised approach that propagates known genes in a network, and supervised learning methods, which characterize positive gene labels and negative non-related genes by seeing how these two sets of labeled genes are related within a network. GenePlexus[1,10] is a recent network-based supervised learning approach to gene classification validated through extensive benchmarking[1] that outperforms label propagation for diverse genesets of cellular pathways, diseases, and traits. It uses a guilt-by-association approach for gene classification – which assumes that genes strongly connected to each other in a network are likely to be involved in similar underlying pathways and functions. GenePlexus' purpose is to prioritize genes for lab study by predicting the association of every gene in the genome/network to an input gene list. GenePlexus returns a prediction for every gene in the genome that represents how well a gene is connected to that original input geneset. However, GenePlexus has not undergone robust evaluation with extremely large genesets, a property characteristic of many real world -omics studies. Additionally, it has been demonstrated that disease and complex trait datasets with lower edge densities of genes within each geneset, such as DisGeNET[11–13] and GWAS, demonstrate inferior performance compared to Gene Ontology[14] Biological Process (GOBP) for gene classification. (**Figure 3.1-3**). **Figure 3.2** additionally shows GenePlexus performance is correlated with the overall size of the genesets – where larger size inputs give worse performance. GenePlexus ran using sets from experimental data such as CREEDS differential expression data shows very poor performance (**Figure 3.1-2)** relative to

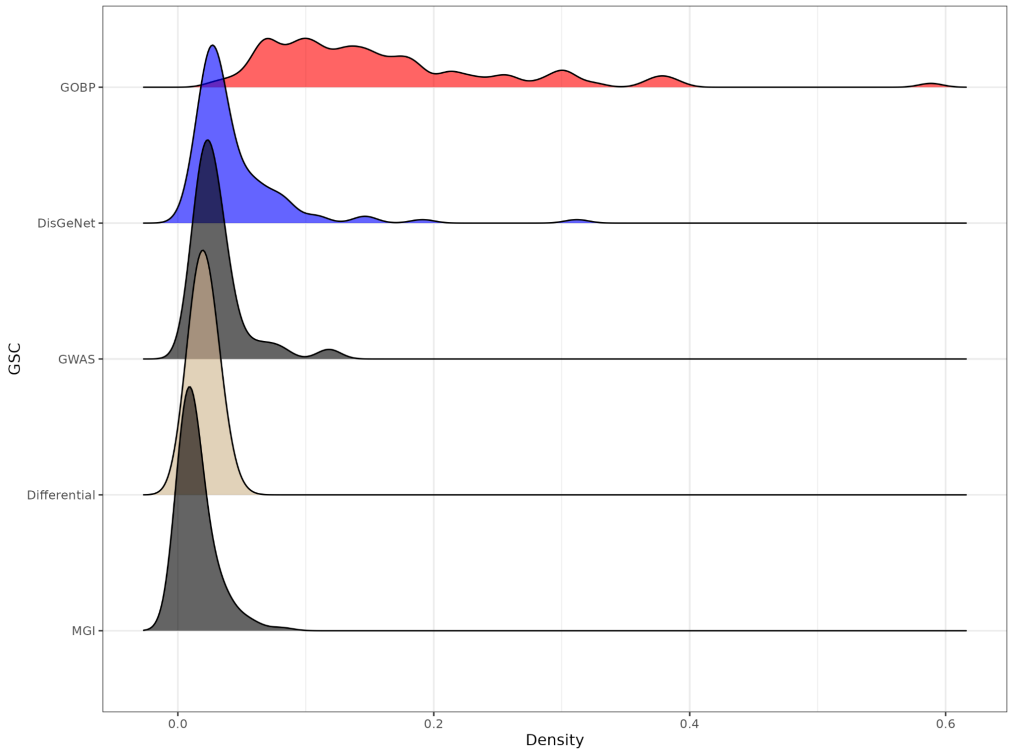

Figure 3.1: Ridge plot demonstrating distribution of network edge density of genesets within each GSC. Function (red) and disease (blue) genesets are more well connected in meaningful neighborhoods in STRING relative to GWAS/MGI (black) and perform better with GenePlexus relative to CREEDS differential expression experimental data (tan).

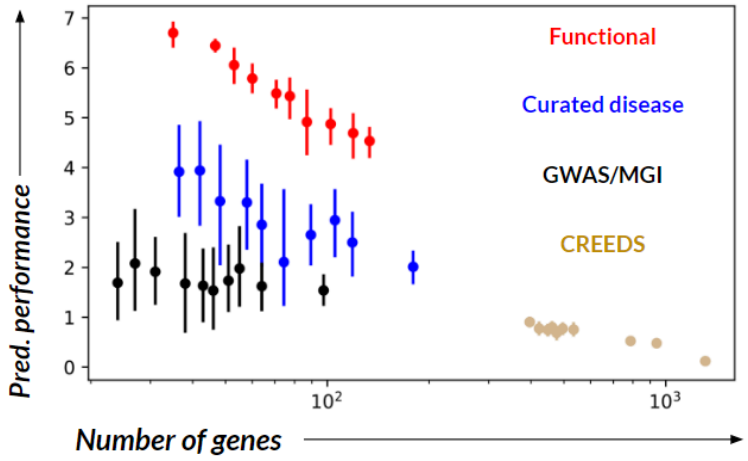

Figure 3.2: Plotting GenePlexus performance with the number of genes in a geneset. The different colored dots represent geneset types. Functional genesets

Figure 3.2 (cont'd)

(red) are GOBP, curated disease genesets (blue) are DisGeNET, GWAS and MGI genesets are (black), and CREEDS (tan) are differential expression results for diseases.

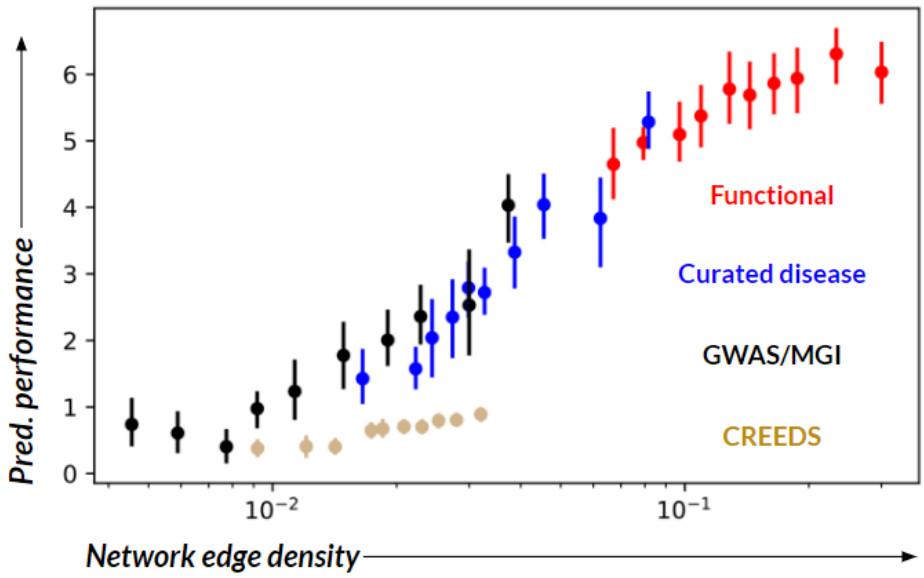

Figure 3.3: Plotting GenePlexus performance with the edge density of the genesets within the STRING network. Performance is negatively correlated with geneset size but it positively correlates with the edge density. Different geneset sources also have different performance. Notably, CREEDS datasets are relatively large and have relatively poor performance even as network edge density increases.

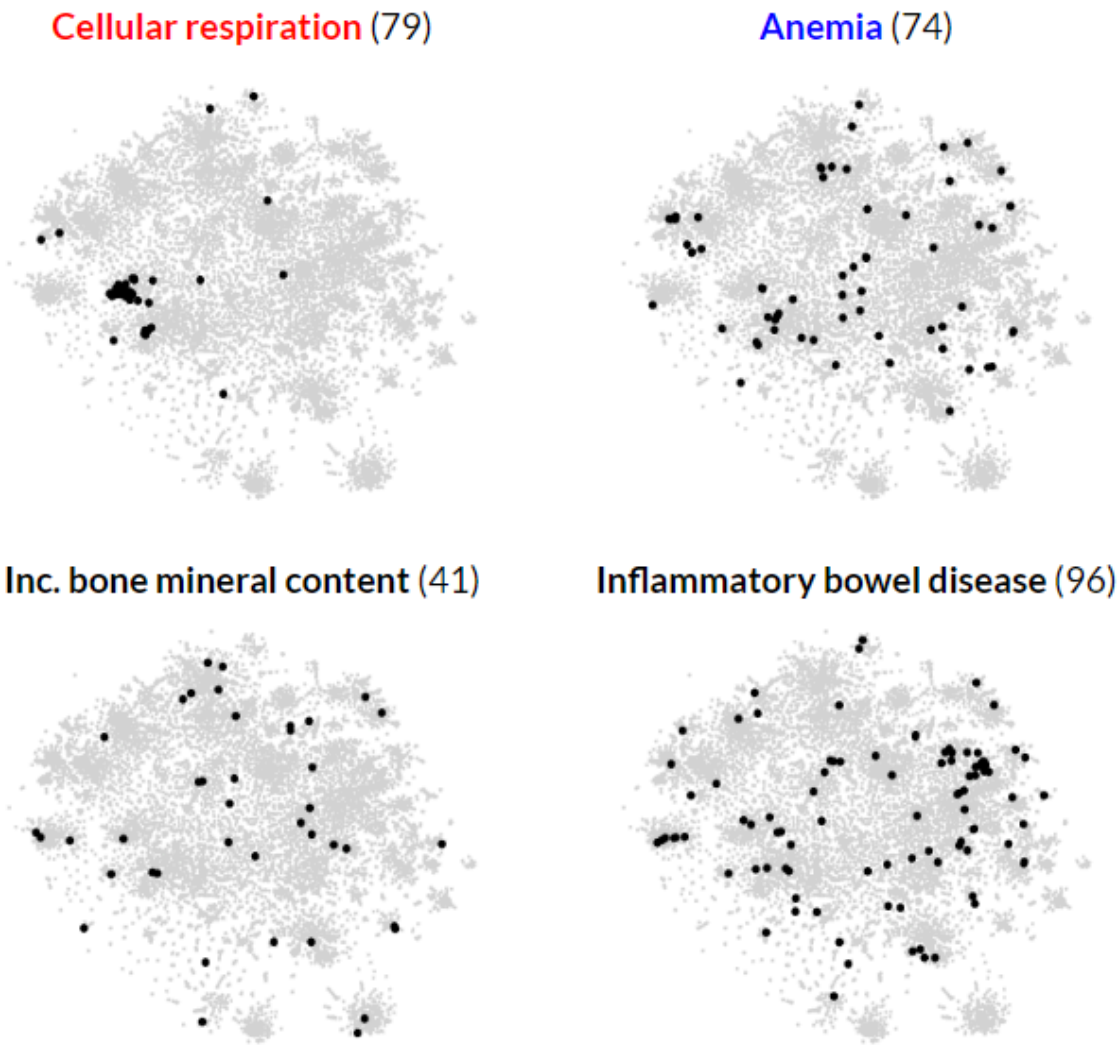

Figure 3.4: t-SNE of STRING network node embeddings, with genes of sample datasets from different sources plotted. Different genesets have different measures of how well genes are densely mapped to neighborhoods in the network. Functional datasets (red) tend to have genes that map to neighborhoods better than disease (blue) or MGI/GWAS (black).

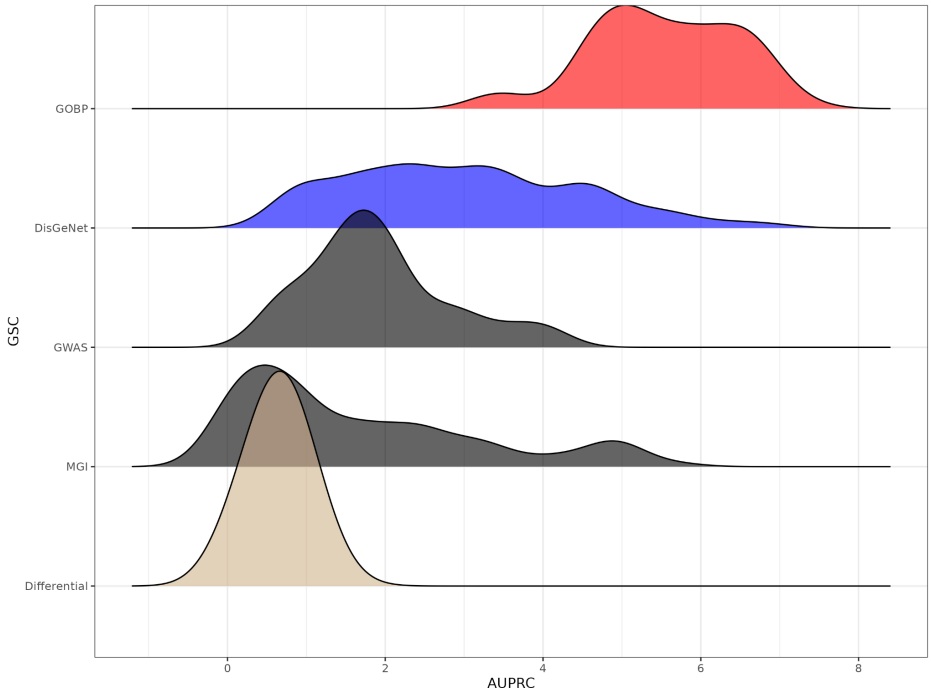

Figure 3.5: Ridge plot demonstrating distribution of performance. X-axis of AUPRC refers to "log2(auPRC/prior)", with each GSC. Function (red) and disease (blue) genesets are more well connected in meaningful neighborhoods in STRING relative to GWAS/MGI (black) and perform better with GenePlexus relative to experimental data (tan).

other GSCs tested in the original study. Performance of GenePlexus models can be poor for two primary reasons. The first being that there is incomplete data at both the network level and the disease/trait-gene annotation levels – which is exacerbated the more complex a trait is. The second primary concern is that complex disease/trait genes are often not localized in the same neighborhood. Rather, a complex trait geneset is made up of multiple gene subsets with varying degrees of localization. **Figure 3.4** shows representations of average sized genesets within GO[14,15], DisGeNET[11,13], MGI[16], and GWAS. GO and DisGeNET have high edge densities (**Figure 3.1, 3.3**) relative to MGI and GWAS, and in these examples the representative GO term is highly localized to one cluster, while the other three genesets have genes scattered across the network embeddings. In other words, GenePlexus tends to perform better on smaller genesets than larger ones, and for genesets that localize in the network – like GO – rather than genesets that are less localized in network neighborhoods like GWAS (**Figure 3.3-5**).

Experimental results add an additional problem that needs to be dealt with, where gene results will not only have unclear network patterns and understudied genes, but also a noticeable level of noise – false positives – compared to defined annotations compiled through multiple sources like GO and DisGeNET. The original GenePlexus application assumes all user provided genes are biologically meaningful and does nothing to filter low quality genes to improve model performance.

GenePlexus and disease modularity

As mentioned in chapter 1, the complexity of human diseases, involving hundreds or thousands of genes, exacerbates incomplete annotations and non-localized genes. Disease gene lists exhibit modularity[17] within biological networks where subsets of genes collaborate with each other, forming modules that are not uniformly or densely connected with every true disease gene. This modularity indicates that complex diseases have multiple and distinct biological processes, phenotypes, and causal mechanisms that are relevant. Essentially, disease gene lists exhibit modularity within biological networks and are very large, which means they are not initially suited for GenePlexus. The additional implication for GenePlexus results is that genes ranked higher in the prediction list are those connected to many input genes in the initial set. Conversely, genes connected to only a small subset are likely to have lower probabilities and may be overlooked by the user. Potentially meaningful processes in a disease are difficult to discover when considering the disease as a whole due to statistical power not because they are unimportant, but because the processes have less true gene annotations relative to other disease processes.

ModGeneplexus: Using GenePlexus with biologically relevant subsets

In this study, we propose ModGenePlexus (**Figure 3.6)**, an extension of GenePlexus that uses subsets of the initial input geneset for model creation and prediction – rather than the entire geneset at once. Recognizing that large genesets encompass multiple distinct biological processes, we aim to enhance results by determining and predicting on these smaller, biologically meaningful subsets. Specifically, ModGenePlexus initially clusters genes into modules (termed gene modules) using a functional gene-gene interaction network and subsequently runs GenePlexus on each discovered gene module. Each model makes a prediction for every gene whether that gene is associated

with the model's respective gene module. The results of each genes' predictions across modules are then aggregated. If the gene has a very high probability in one of the models, the max probability is used. Otherwise the average probability score across modules is used for that gene. When discovering gene modules, ModGenePlexus denoises the initial geneset by excluding genes that lack meaningful network connection to other disease genes, and propagates the initial geneset by finding new genes that were not initially labeled as disease genes – but have dense network connections to disease gene enriched modules. These methods are crucial for datasets that have not gone through any processing and are expected to have false positives or less confident results. We validate ModGenePlexus through showing ModGenePlexus's performance with large, real-world experimental datasets. First, we ran a simulation where we created large genesets that contain modules known to be biologically meaningful and robust to validate utilizing multi-model prediction and aggregation of the results into a final prediction. Second, we improved neutral/negative gene label classification in GenePlexus by determining negative genes utilizing gene modules – rather than with all positive labeled genes at once. Third, we show significant performance improvements in real-world differential expression data from the CREEDS database, encompassing manually extracted gene expression profiles from GEO[18] for gene/drug knockouts and disease patients. We also implemented ModGenePlexus for GWAS experiments, , using MAGMA gene prioritization results and varying p-value thresholds to show performance differences. Next, we apply ModGenePlexus to the same MAGMA results to determine if training models for prioritized genes that meet a less stringent p-value threshold improves performance when prediction for high-threshold, more confidently associated genes. Our analysis revealed that employing models incorporating more genes (less stringent MAGMA thresholds) leads to improved classification of stringent test genesets. We then visualize and investigate how network edge density and geneset size statistics affect performance of ModGenePlexus. Finally, we show using ModGenePlexus with real experimental results for type-2 diabetes highlights unique enriched processes and biology that using the whole experiment would not recover, and we further discuss the implications of using ModGenePlexus and modules in general for interpreting human disease. Our findings demonstrate that module expansion enables

more robust gene classification of large-scale -omics genesets and real experimental results – facilitating the identification of mechanistically important genes in many diverse contexts.

Defining terms and models used in this chapter

This chapter utilizes multiple different models and notation that has particular meaning in the ModGenePlexus pipeline. **Figure 3.6** shows the entire pipeline for ModGenePlexus. However, multiple types of models are created for purposes of comparison and validating ModGenePlexus performance which tests various components of ModGenePlexus, We additionally create models for the original GenePlexus implementation. We provide definitions here:

- **AllAssign**: Creating a single model for all genes originally assigned to the geneset. This is also referred to as “GenePlexus” because this model is running GenePlexus normally.

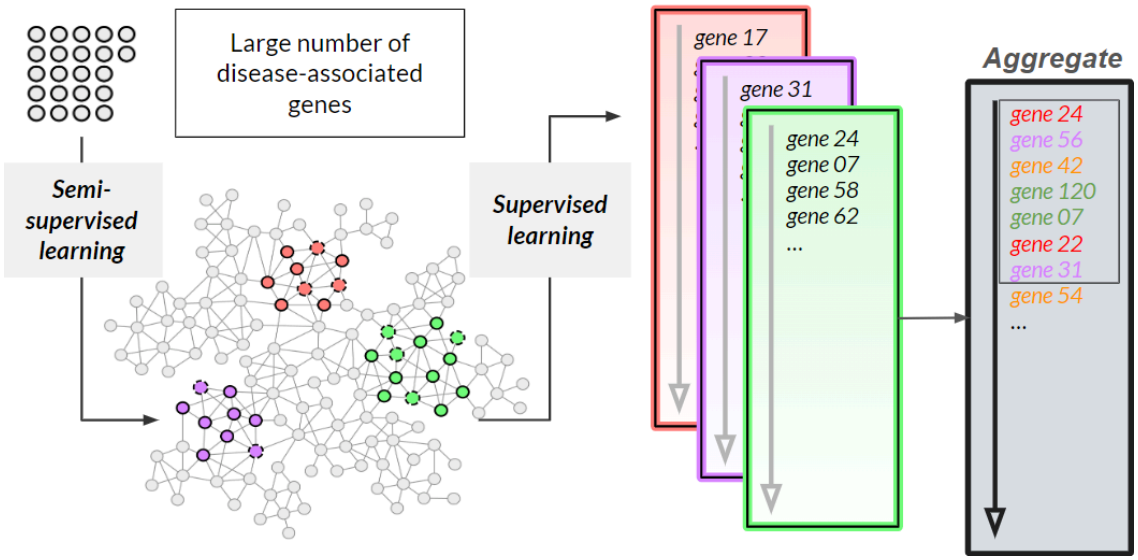

Figure 3.6: A schematic for modGenePlexus. A functional network (STRING) is used to discover gene modules based on a user input list utilizing the DOMINO software. This is a semi-supervised technique which will remove genes from the initial user input list - and discover novel genes in user geneset-enriched clusters. Nodes with a dashed border are those discovered through semi-supervised learning. GenePlexus trains a logistic regression model for each individual cluster containing original seed genes and those found through the label propagation. Supervised learning is conducted on each

Figure 3.6 (cont'd)

individual module for genome-wide gene classification. The gene predictions across clusters are finally aggregated into a final prediction that will typically contain originally positive labels from each individual cluster, and novel predictions.

- **AllClus**: Creating a single model trained for all genes assigned to a cluster.
- **MaxAvgProb**: Multiple models are created, one model for each cluster. Also referred to as ModGenePlexus.
- **noprop**: A model where after cluster assignment, additional propagated genes are not used in training. "noprop" will appear as a prefix to either AllClus or MaxAvgProb – such as noprop_AllClus or noprop_MaxAvgProb
- **domino**: A model where after cluster assignment, additional propagated genes are not used in training. "domino" will appear as a prefix to either AllClus or MaxAvgProb – such as domino_AllClus or domino_MaxAvgProb.
- If either **AllClus** or **MaxAvgProb** models are not labeled with a prefix, it is utilizing propagated genes (assume a prefix of domino_).

Terms and their definitions can be used interchangeably depending on context. For example, running a single model for all genes can be referred to as either GenePlexus or AllAssign.

Methods

Compiling and using the STRING Network

Version 10 of the STRING[19] network was used in this project. This is the version of STRING that also underlies GenePlexus. Specifically, we used a version of STRING where we only kept edges that have an edge weight greater than 0.70. This was done to use only confident edges for cluster assignment of genesets. STRING was chosen as the method for this project over alternatives such as BioGRID[20,21] because (i) the amount of biological annotations integrated in STRING – including protein-protein interactions, conservation data, sequence homology, etc. – is more comprehensive, and (ii) STRING performs the best with GenePlexus compared to alternatives and the goal with ModGenePlexus is to show it outperforms GenePlexus under rigid evaluation.

Compiling CREEDS differential expression datasets

Genesets were compiled from CRowd Extracted Expression of Differential Signatures (CREEDS)[22]. We utilized manual signatures for single gene perturbations, single drug perturbations, and diseases. Manual signatures correspond to those where gene expression profiles from NCBI Gene Expression Omnibus (GEO)[18] were obtained through manual validation rather than through an automated process. Genesets with at least 100 genes were used, and both genes that were up-regulated and down-regulated were included in the same positive label set for GenePlexus models. We utilize CREEDS datasets covering three biological domains: Gene knockout, human disease, and drug expression profiles, where 972, 311, and 234 genesets respectively were run with ModGenePlexus.

Compiling MAGMA gene prioritization results

MAGMA[23] gene prioritization scores for summary GWAS results were compiled from GWAS Atlas[24]. We created three GSCs utilizing these gene predictions based on thresholds of $p < 1 \times 10^{-2}$, $p < 1 \times 10^{-5}$, and $p < 1 \times 10^{-8}$. Sets with at least 100 genes were used, and the number of sets for each threshold were 671, 32, and 15 respectively. Multiple thresholds were used because for most MAGMA gene scores, using very strict thresholds will not give very large genesets – which is the primary use case of ModGenePlexus. While it is common to use very strict thresholds with GWAS data due to multiple correction, it has been shown that genes with more nominal significance values can have biological meaning[25]. Additionally, when analyzing how ModGenePlexus utilizing looser thresholds can predict strict threshold test genes, we added two more thresholds of $p < 5 \times 10^{-2}$ and $p < 1 \times 10^{-1}$. Because there were only 32 genesets that met size criteria with a threshold of $p < 1 \times 10^{-5}$, the same 32 GWAS were used across the looser thresholds for this analysis specifically.

Evaluation metric for measuring model performance

In this chapter, we present results of performance across all models with the auPRC where it is normalized with the prior and the ratio is log-transformed.

The metric log2(auPRC/prior) is defined by:

$$auPRC = \sum_n (Recall_n - Recall_{n-1}) Precision_n$$

$$Precision = TP/(TP + FP)$$

$$Recall = TP/(TP + FN)$$

$$prior = \frac{P}{P+N} = \frac{TP + FN}{(TP+FN)+(FP+TN)}$$

Where *TP* is true positives, *FN* is false negatives, *FP* is false positives, and *TN* is true negatives. *P* and *N* are the ground truth labels. The curve is constructed using multiple thresholds based on the GenePlexus predictions, and at each threshold the true positives, false positives, true negatives, and false negatives are calculated. These values are used to calculate precision and recall at each threshold and construct the precision recall curve. This metric was chosen because gene classification is highly imbalanced, where there are many more negatives than positives. Methods like auROC and accuracy are unsuited for this problem[26]. Additionally, optimizing for this metric controls for false positives[27]. This is vital since the most important goal in gene classification is to make sure the top candidate genes in particular have few errors.

Determining pairs of GOBP far away in a network

For Gene Ontology[14] Biological Processes (GOBPs) A and B, we utilize the influence matrix to determine if genes in GOBP A are far away from genes in GOBP B relative to each geneset's distance to random genes in the network. The values of a diffusion-based representation of the network – an influence matrix – are used as scores. To determine how far GOBP A is from GOBP B, two z-scores are calculated:

$$\frac{A_{source}B_{target} - mean(A_{source}Random)}{sd(A_{source}Random)} \text{ and } \frac{B_{source}A_{target} - mean(B_{source}Random)}{sd(B_{source}Random)}$$

where $A_{source}B_{target}$ refers to influence matrix values for genes in GOBP A where A genes are the start within the influence matrix, and B_{target} are the genes of GOBP B being targeted as the destination. $B_{source}A_{target}$ refers to influence matrix values for

genes in GOBP B where genes in B are the start, and A_{target} are the genes of GOBP A being targeted. $A_{source}Random$ is a list of the mean scores of random sampled genes being targeted from A, the size of the random geneset being the size of GOBP B. $B_{source}Random$ is a list of the mean scores of random sampled genes being targeted by B, the size of the random geneset being the size of GOBP A. This method is utilizing the central limit theorem to determine if the mean connection of one GOBP to another is greater than the connection to random genes. Meaning, we take the average of the mean distances across a large number of randomly sampled genesets, and calculate z-scores based on these averages. The mean of these two z-scores is then taken. To interpret the results, the z-scores that are negative are those GOBP genesets less connected to one another within a network relative to the genesets connections to many randomly sampled genesets. Simulated traits can contain up to 10 GOBPs if all of the GOBPs have a negative z-score to one another and have no gene intersections. GenePlexus was run on every GOBP term and every simulated trait.

Positive gene evaluation in a study-biased holdout split

Evaluations are created using a study-biased holdout. This evaluation was chosen because it is more stringent and reflective of the real world task of novel gene discovery. Understudied genes are less likely to have robust network connections compared to well-studied ones. For a given geneset, genes are put into that trait's training set if they are in the top ⅔ of genes mentioned in pubmed abstracts. Genes that are in the bottom ⅓ of pubmed abstract mentions are classified as test genes. All models implemented in this project use this initial positive test holdout for evaluation.

Negative gene assignment and using negative genes in a study-biased holdout

Negative genes were identified using the algorithm implemented in PyGenePlexus, which intersects the user-input genelist with a geneset collection to identify genes that, while not originally part of the user list, are part of sets in the collection that exhibit significant overlap with the user list. More specifically, after finding the union of all genes annotated to a gene set collection, the seed genes of the relevant geneset are removed and negative genes are those not annotated to any geneset that significantly overlaps ($p < 5 \times 10^{-2}$; one-sided fisher's exact test) with the given seed genes. To ensure

robustness, this geneset collection was narrowed down to only include genes within the positive training universe, excluding genes that are understudied. The neutrals discovered through this process were not utilized as negatives for model creation or testing. For assignment to training and test bins, given that all negative genes are well studied they were assigned to training and test bins through random assignment. This assignment maintains the same ratio of negative train genes to negative test genes as that of the positive train genes to positive test gene ratio for the respective geneset. For evaluation, the negative test set used is a set of genes that are negative in both AllAssign and in methods that utilize modules in some way. Genes that are in this negative test split will never be considered for training in any model.

Determining modules and performing semi-supervised gene classification

We utilized the DOMINO[28,29] method for module detection using the STRING network. Briefly, DOMINO first uses Louvain clustering[30] to split a network into initial slices. A hypergeometric test is run with these slices to determine if there is at least some enrichment with user input genes $FDR \leq 0.3$. Each enriched slice has a single smaller, sub-component extracted through an iterative process with the Prize Collecting Steiner Tree algorithm[31]. The goal of this algorithm is to find a subnetwork of genes that maximizes the sum of prizes that nodes in the current iterative subnetwork give while minimizing the penalties of edges connected to nodes not in the subnetwork. The prizes of active nodes are obtained using influence propagation – meaning genes well connected to active nodes will have a higher prize associated with it. Starting with user-defined genes in the slice as a subnetwork, each iteration adds a new non-user gene to the subnetwork until $\sum v \in T p(v)$-$\sum e \in T c(e)$ – where *v* is a node, $p(v)$ is the prize of a node, *e* is an edge, and $c(e)$ is the cost of edge *e* – is maximized. These obtained sub-slices are further refined using the Newman-Girvan algorithm[32], where edges are iteratively removed using the betweenness centrality metric. Each iteration, a modularity score is computed on the new sub-graph, with the stopping criterion being $\frac{log(\#ofnodesinsub\text{-}slice)}{log(\#ofnodesinnetwork)} \leq M$. Lastly, final modules are determined by each sub-slice passing a stricter hypergeometric test using Bonferroni correction ($q < 0.05$). For ModGenePlexus, the importance of the DOMINO method is that it is a way to discover gene modules where genes are related in the network, and contains genes that while

not part of the original set, were discovered utilizing user gene labels and network topology. Similarly, genes that are not part of a significantly enriched module aren't used. This is because the genes are not well connected to other disease genes in the network. These genes are considered false positives in the context of this project, as there is little reason to think they will contribute to network-based gene classification and their relationship to the disease is unclear since they aren't well related to other disease genes.

Refined neutral selection using cluster information

Individual cluster gene assignments were passed into a previously described algorithm (**see methods**, section "Negative gene assignment and using negative genes in a study-biased holdout"). For a trait, each relevant cluster was intersected with the geneset collection to discover neutral genes for that cluster – rather than the trait as a whole. If a gene is considered neutral in at least one cluster, then it is used as a neutral for all models for each module. Additionally, negative assignments for AllClus models were obtained using this method. The source code of PyGenePlexus was directly modified to use this new method to determine neutrals and negative gene assignments.

Types of created models and aggregation of module predictions

We utilized multiple types of splits for evaluation or for running ModGenePlexus and its variations. The first is creating one model for all genes originally annotated to the disease. This is running the set with GenePlexus normally – called "AllAssign". A second model type is utilizing all genes that were assigned to a cluster in the DOMINO method. This is "AllClus", where only one model is created for all modules at once. In a normal DOMINO run, this will include a subset of the original geneset, and those genes propagated. ModGenePlexus is when models are created for each obtained module and the predictions are aggregated. The aggregation is created by finding two types of genes. If a gene has a prediction where $Prob > 0.80$ in any module, the max score across modules is used in the aggregation. For all other genes, the average module probability score across modules is the final score. This is the complete ModGenePlexus method and is also known as "MaxAvgProb". Lastly, two additional model ways to filter genesets can be run with AllClus and MaxAvgProb. For models that have a prefix "domino_", the propagated genes from DOMINO are used. If it has the

prefix “noprop_”, then the only propagated genes included are those that are test genes of the geneset in question.

Running GenePlexus and ModGenePlexus

GenePlexus and ModGenePlexus was run using a modified version of PyGenePlexus[10,33]. In this, we ran all models with the STRING network and with the adjacency matrix settings. These choices were made due to STRING and the adjacency matrix having superior performance relative to other networks and to allow for each gene to be considered a feature within the model – which embeddings would not allow.

Compiling GWAS for strict threshold gene prediction

The strict threshold was chosen to be $p < 1 \times 10^{-5}$ because $p < 1 \times 10^{-8}$ did not have enough GWAS that had MAGMA predicted geneset sizes of at least 100. The looser thresholds used were $p < 1 \times 10^{-2}$, $p < 5 \times 10^{-2}$, and $p < 1 \times 10^{-1}$. Only traits that had at least 100 genes with a threshold of $p < 1 \times 10^{-5}$ were considered, thus 32 GWAS were used for all thresholds and compared in this analysis.

Calculating edge density

Edge density, D, for a geneset, G, is given by: $D = \sum_{\{(u,v) \in G\}} W_{uv} / (|T| * (|T| - 1)/2)$ where W_{uv} is the edge weight between genes u and v. This measures how connected the genes in a geneset are within itself.

GOBP enrichment of genesets

GOBP enrichment was determined using the ClusterProfiler[34] software package in R. The GSC was subsetted to include processes that had at least 10 genes and a maximum of 500 genes. Enrichment was run for the AllAssign geneset and for each module of type-2 diabetes.

Results

Module expansion replicates performance in combined GOBP simulated traits

To validate our hypothesis that module expansion gives superior performance to regular GenePlexus for large genesets, we implemented a simulation where we create large genesets with known biologically meaningful subsets. An initial difficulty of designing and validating the proposed methodology for ModGenePlexus is that there is no gold standard of modules that we can use for diseases to test the method[35]. As such, we

chose to design a method that creates large genesets that have truly biological meaningful datasets, being modules, underlying them. Using GOBPs, we used a network to discover pairs of GOBPs that are far away from each other in the context of a biological network (**see methods**). Individual GOBPs are treated as clusters of this disease. GOBPs were additionally combined into simulated traits only if there were no genes overlapping between them. This assures that there is no consideration of fuzzy cluster assignment and that the GOBPs should have distinct network signals. We found that our module expansion method ModGenePlexus recreated the performance of running GenePlexus normally on the entire geneset at once using the evaluation metric log2(auPRC/prior) (**Figure 3.7-8)**. We additionally see that models run with the larger, combined genesets have much lower performance compared to running on each term individually. Given the equivalent overall performance of our method with the original geneplexus implementation, we next looked into how different categories of genes, whether they are positive or negative, ranked across the genome-wide predictions. We can see that positive test genes of the relevant term tend to have higher ranks than positive genes of the other combined terms, which themselves have higher rankings than negative test genes (**Figure 3.9**). This is evidence that true negative genes will cluster at the bottom of genome wide rankings, while genes that are possibly relevant can fall anywhere in the genome ranking. Lastly, we hypothesized with ModGenePlexus we will more easily discover genes that are highly ranked at a module level, but are a lower rank in the full disease set. We show that genes are most likely to be ranked highest in the relevant GOBP term the gene originates from – and the genes that GenePlexus gives very high predictions to are nearly always higher ranked in the relevant term – rather than being ranked higher in the irrelevant one (**Figure 3.10**). This experiment shows that ModGenePlexus can recreate the performance of GenePlexus, and that the individual modules are able to rank genes higher relative to a larger set comprising multiple distinct modules. However, this simulation does not do anything to remove genes that are potentially false positive, as with GOBP we are assuming these genesets are quite robust and validated when compared to finding functional modules of real diseases.

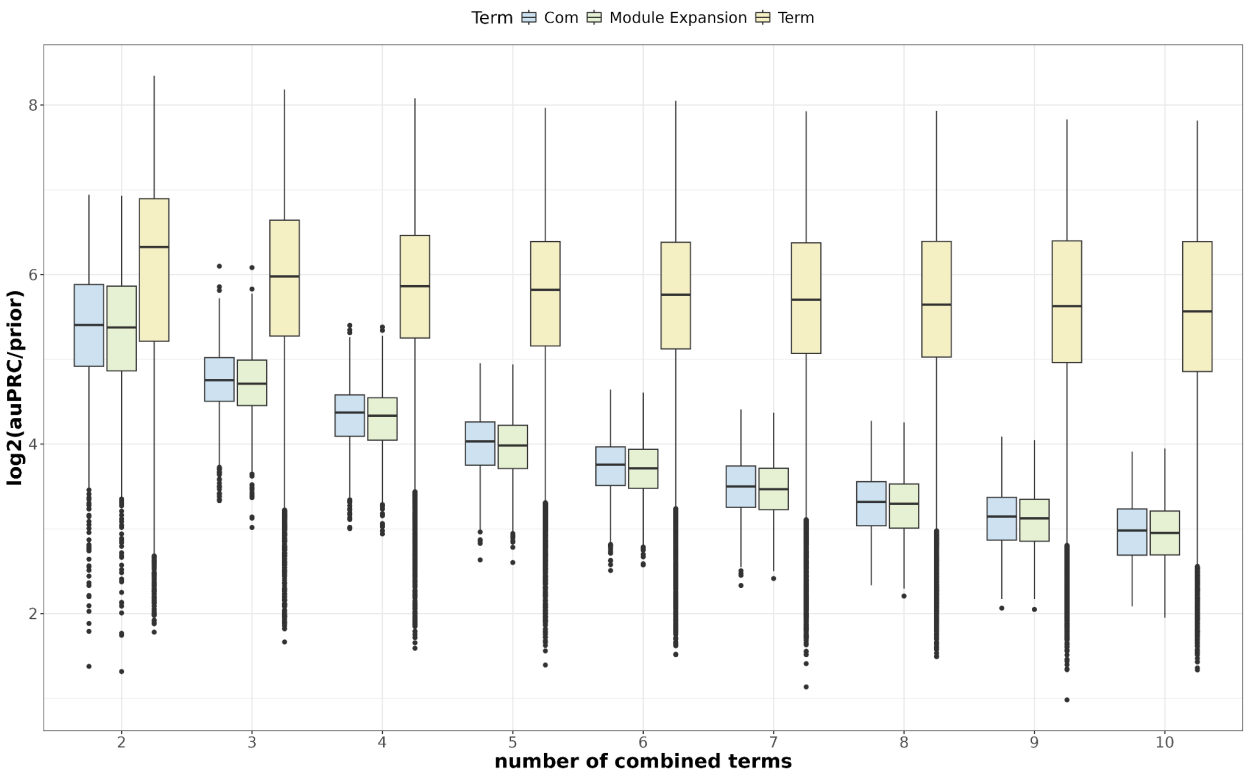

Figure 3.7: Simulation results for models built from individual GOBP terms (Term), a single model built from combining the terms (Com), and module expansion of each term individually (Module Expansion). Module expansion recreates the performance of creating one model for the entire combined GOBP list.

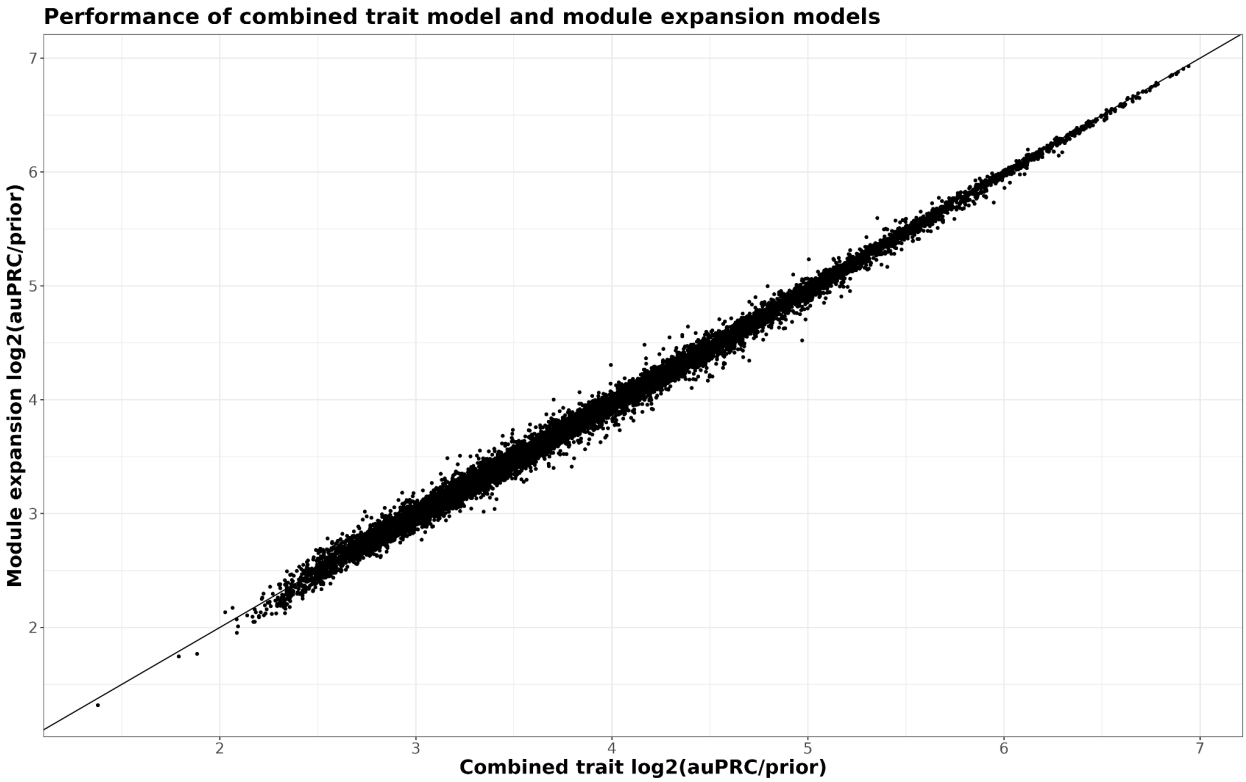

Figure 3.8: log2(auPRC/prior) of simulated trait GenePlexus performance for the Combined simulated trait mode (x-axis) and for ModGenePlexus (y-axis).

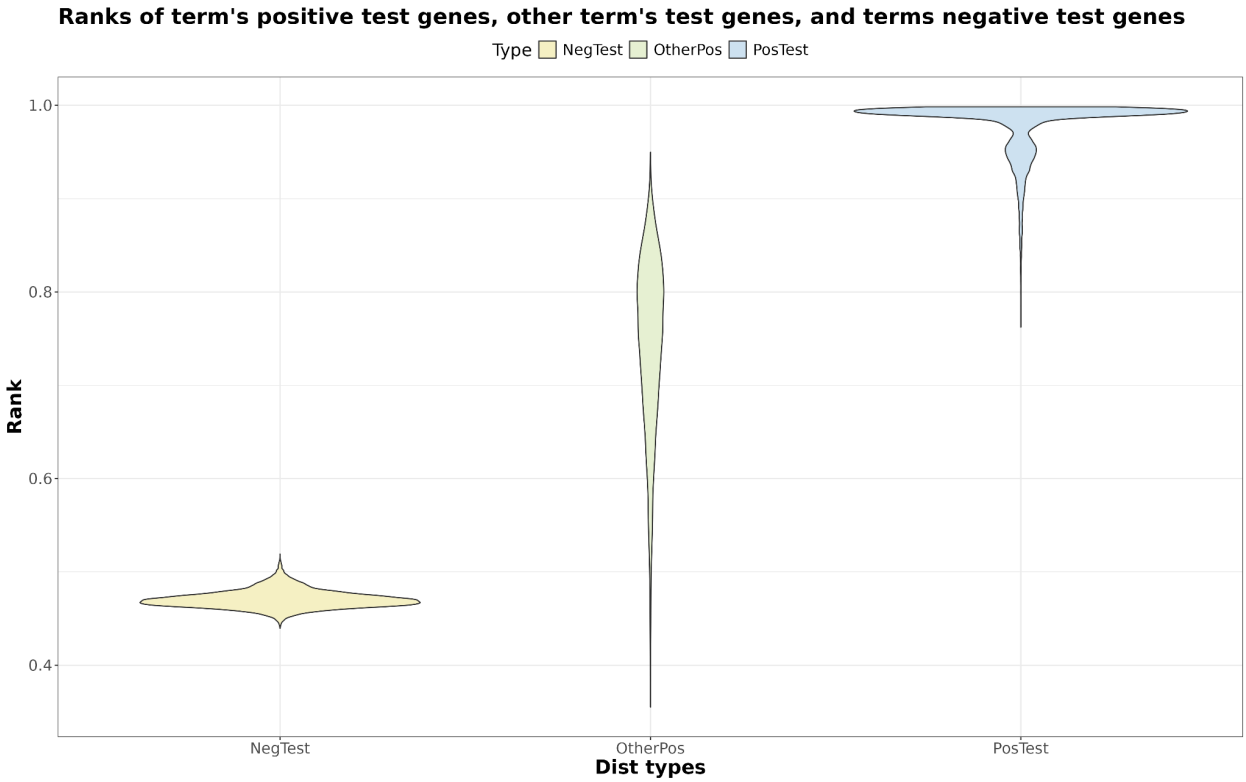

Figure 3.9: The ranks of gene predictions are meaningful across the GenePlexus ranking. The Y-axis is the genome wide rankings scaled from 0-1, where genes closer to one have higher ranks in the GenePlexus prediction output. Genes that are positive test genes (PosTest) – those that have true association – are ranked highly in the genome wide results. Positive test genes of other GOBPs (OtherPos) but are not negative in a GOBP have ranks that encompass the entire GenePlexus prediction output. Those that are negative test genes (NegTest) tend to have low ranks.

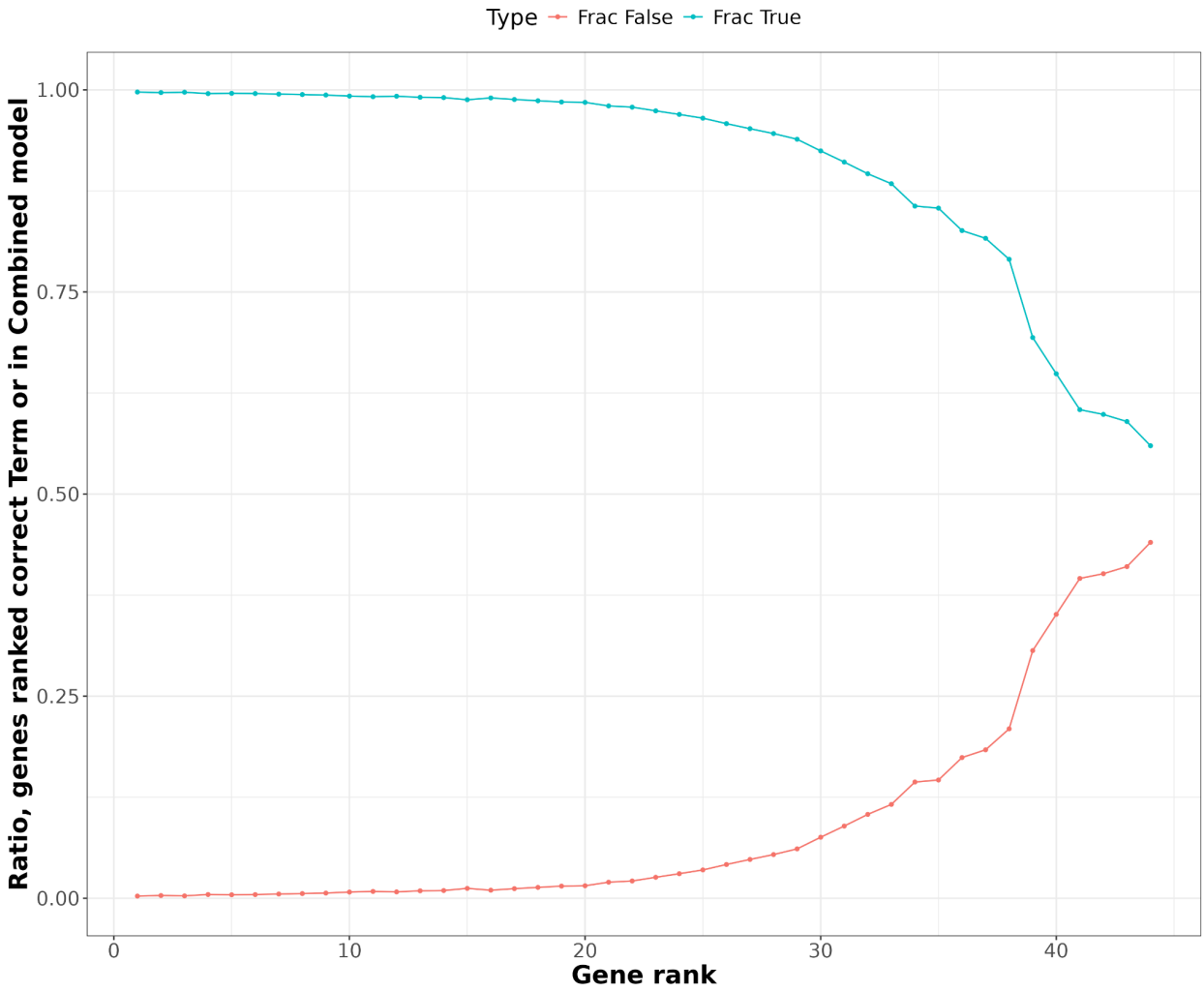

Figure 3.10: Genes that are highly ranked in GenePlexus results of the term the gene is associated with will have a higher rank in those predictions than in unrelated terms the gene is not annotated to. Genes typically have the highest rank in model output corresponding to the term they are positive labels in. If the gene was not highly predicted in the term it corresponds to, then the instances of a gene being predicted higher in an incorrect term increases as the corresponding term rank decreases. X-axis represents the genes being binned in equal frequency bins based on their ranks in GenePlexus predictions.

Improving negative and neutral selection of GenePlexus label assignment using module assignments

Using modules offers an additional benefit by letting us improve how the model determines neutral and negative labels for genes. If diseases are interpreted as a collection of smaller subsets of related genes, then using the entire disease genelist to determine neutrals is problematic. GenePlexus determines neutral genes by taking the positive genes (the user input genelist) and performing a hypergeometric test with genes in a GSC like DisGeNET or GOBP. If there is significant overlap between the positive labels and a geneset in this collection, then the genes in that set that are non-positive become neutral genes, meaning they aren't used as negative labels when training the model (**see methods**). However, these genesets are unlikely to overlap significantly with the entire disease as a whole, so relatively few genes are made neutral for large user genelists **(Figure 3.11)**. Given that modules discovered are filled with biologically related genes, we evaluated whether calculating negatives on a per-module basis where using neutrals discovered in any module, rather than finding neutrals using the geneset as a whole, improves performance. For larger traits of combined GOBP, there is strong correlation between the number of combined terms and the number of discovered neutrals using our new method (**Figure 3.11),** but not when determining neutrals from using the entire simulated trait geneset at once. We modified the PyGenePlexus source code to utilize our new neutral selection method, and found that using neutrals found on a per-module basis gives better results for all different models (**Figure 3.12-13)**. In other words, using other meaningful biological datasets to discover neutrals, rather than only the positive gene list, improves GenePlexus and ModGenePlexus performance. Additionally, there is significant improvement in simulated traits irrespective of how many GOBP terms were combined (**Figure 3.13**). This is especially interesting given that the more modules a trait has, the more neutrals are discovered. This suggests that having a lower number of negatives can improve results when the GSC and biological data is used at the module level. Since complex diseases have multiple modules, this method is integrated in PyGenePlexus when doing neutral selection and when using either ModGenePlexus or any other model that utilizes module assignments.

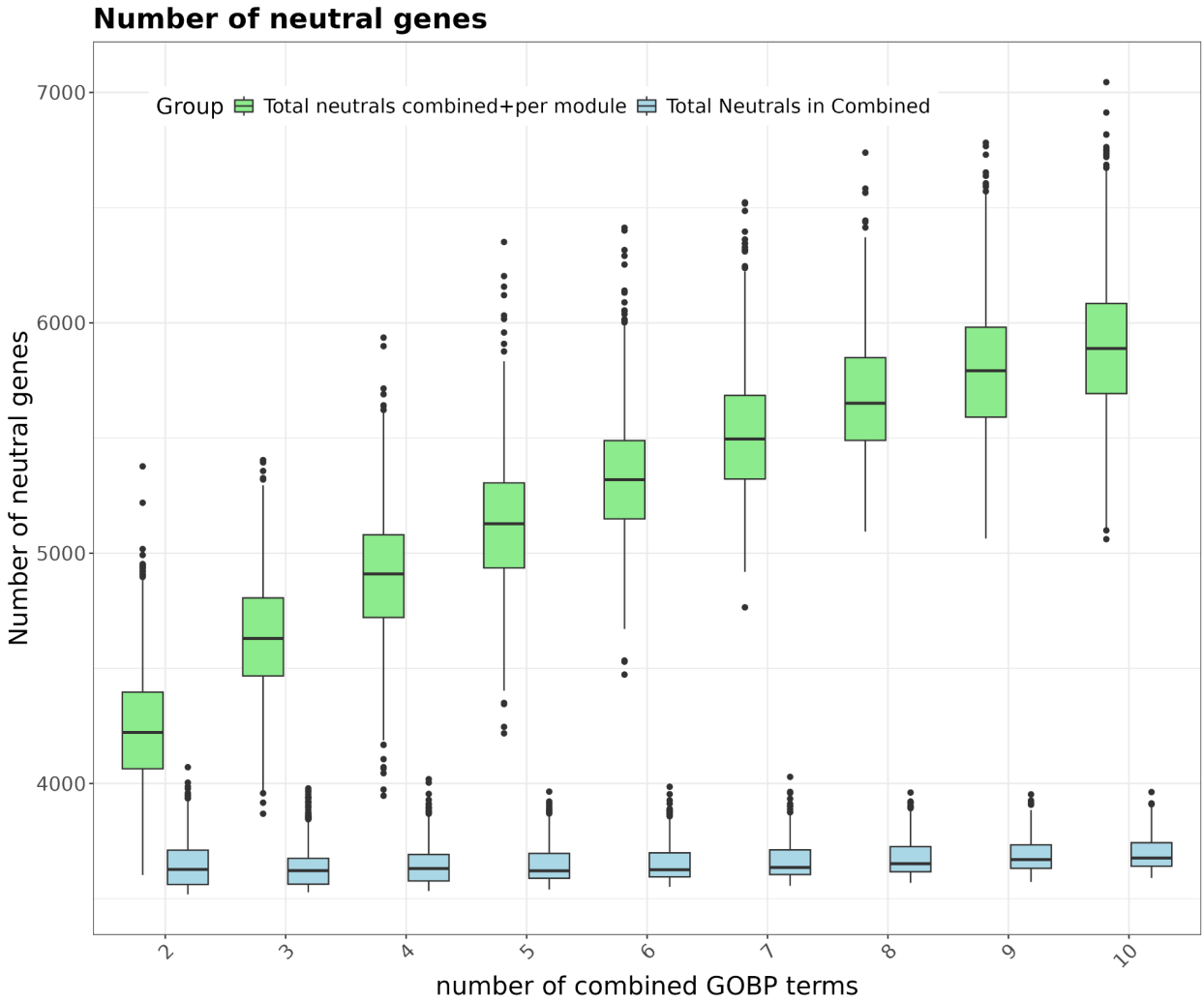

Figure 3.11: The number of neutrals discovered when performing a hypergeometic test with the entire disease list at once or on a per module basis. The more meaningful modules there are within one of the traits, the more neutrals are discovered. Even though simulated traits with 10 terms are significantly larger than those containing 2, the number of neutrals across simulated traits does not change significantly.

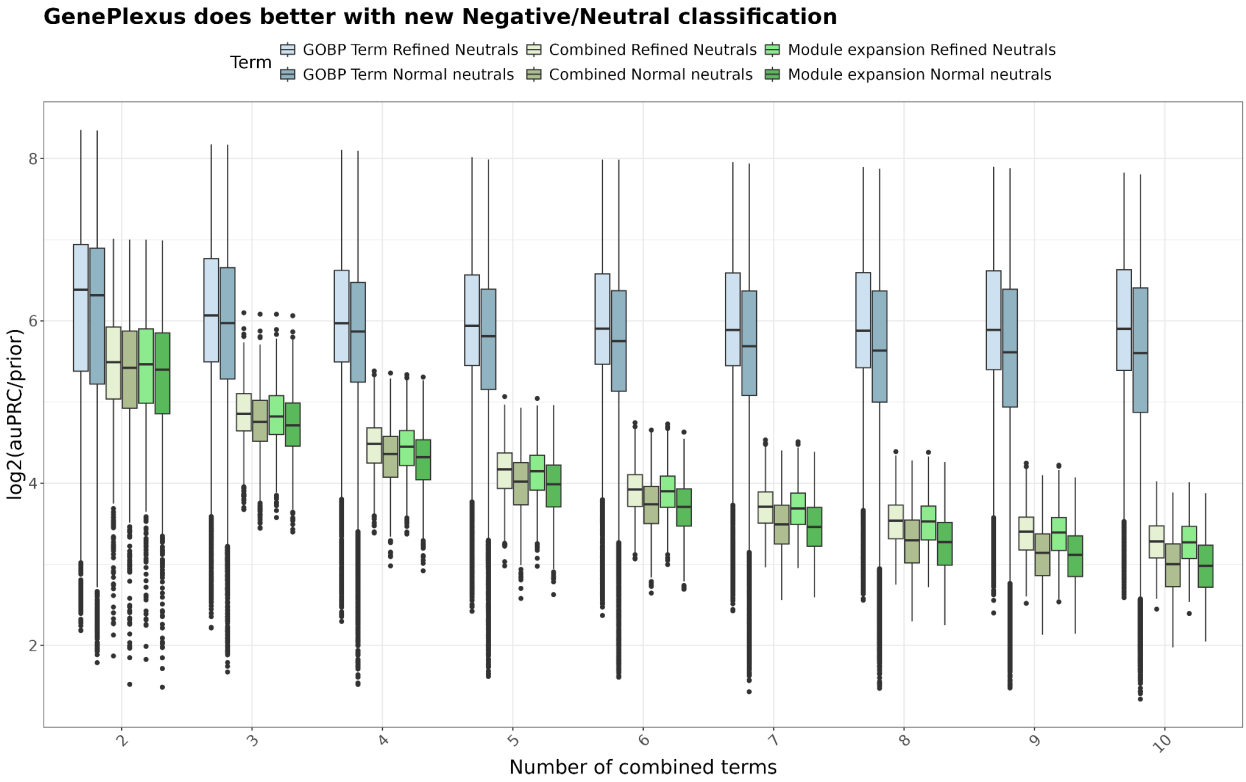

Figure 3.12: Simulation results when using our per-module neutral selection method vs. GenePlexus's. For each model – GOBP term, the Combined terms, and Module expansion of the terms – utilizing neutrals that are obtained using the per-module method has improvement. The more modules there are, the more drastic the improvement.

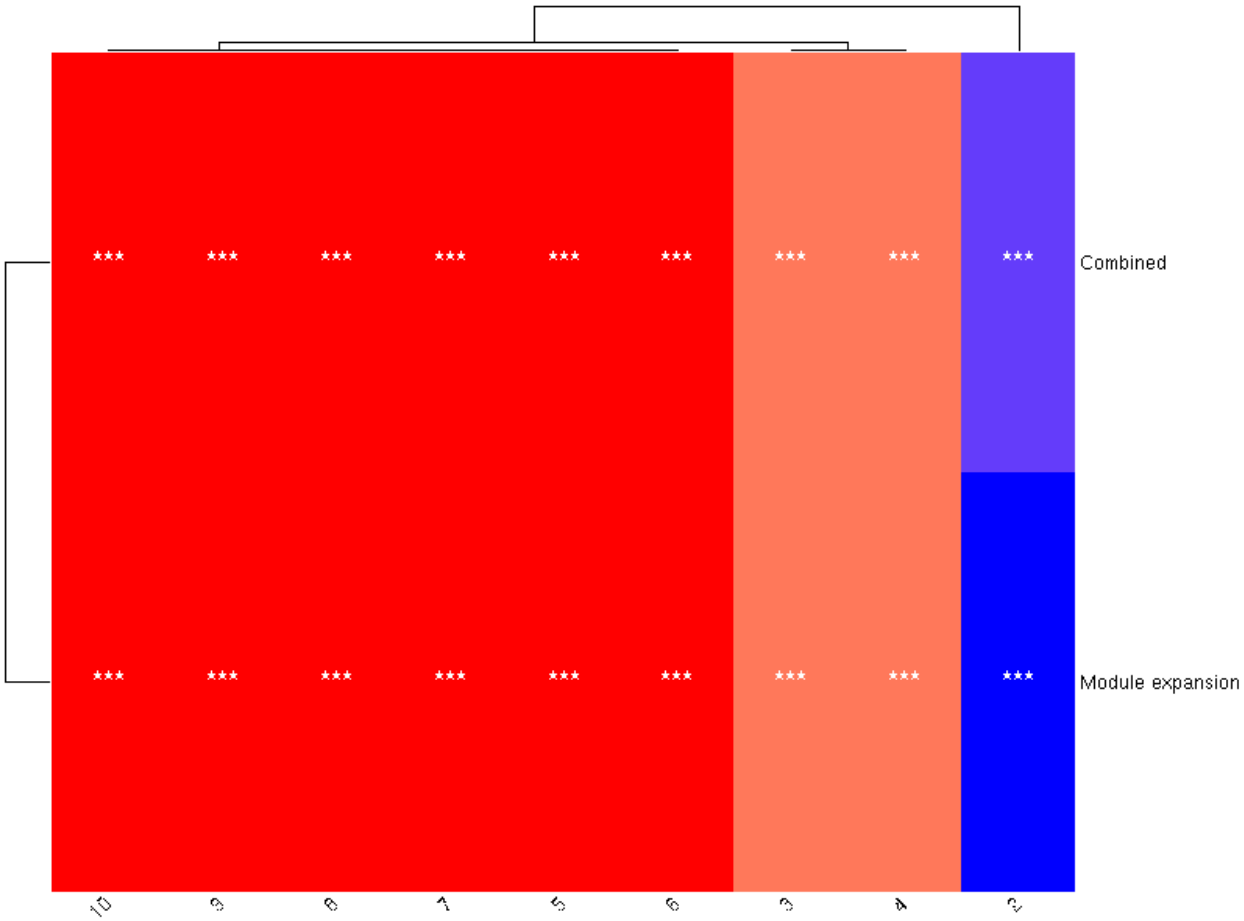

Figure 3.13: Wilcoxon test that shows significance of result improvement of using the per-module neutral selection method for the combined (GenePlexus) and module expansion (ModGenePlexus) outputs. When at least 2 terms are put into a simulated trait, the performance increase with using per-module neutral selection is significantly improved at a threshold of $p < 1 \times 10^{-8}$ (***)**.** This performance increase is true for both the “Combined” trait (GenePlexus on the whole geneset at once) and for ModGenePlexus (creating multiple models for each term in the simulated trait).

Module expansion gives significantly improved results with differential expression data

ModGenePlexus was designed with real-world experimental results in mind. Our first goal was to be able to show that ModGenePlexus can improve gene classification for differential expression datasets under stringent conditions. Differential expression datasets are an ideal use case because they involve hundreds or thousands of gene

associations and are noisy[36,37]. We used CREEDS manually extracted differential expression datasets for diseases, gene knockouts, and drug signatures. After filtering for geneset sizes and experiment number, we ended up with 311, 972, and 234 genesets, respectively, suitable for creating study-biased holdout splits (**see methods**). These traits were chosen based on having a large initial input size of at least 100 genes – as we want to have traits that capture the etiology of complex diseases – and having at least 10 genes that are understudied in literature to use as positive test genes in the evaluation. For CREEDS disease datasets, ModGenePlexus outperformed a single model for all genes – AllAssign – with the log2(auPRC/prior) evaluation statistic (**Figure 3.14).** Similar results were seen for the CREEDS gene knockout and drug profile datasets (**Figure A3.1)**. Next, we created genesets for model AllClus, which creates a single model for all genes that were assigned in a cluster using the DOMINO module discovery algorithm. This geneset includes genes that were part of the initial experiment set that were assigned to a cluster, and propagated genes discovered in the network. ModGenePlexus also does better than this AllClus method for multiple datasets (**Figure 3.15**). AllClus is a useful metric because these genesets have the false positive genes removed, so a comparison to ModGenePlexus shows that creating multiple models for each module gives an additional significant benefit to the gene classification results. We also tested models where we remove all propagated genes that were not part of the disease's test set to determine whether the improved performance came from simply propagating test genes in the semi-supervised DOMINO module identification, or if the novel propagated genes and supervised learning in GenePlexus add anything to performance. Using all the propagated genes – not just discovered disease test genes – improves performances for the CREEDS disease dataset and the CREEDS drug and gene datasets (**Figures 3.16**). Lastly, we visualize results for an AllClus version of the non-propagated cluster assignments displayed in **Figure 3.17**, displaying all discussed model types at once. These results show that ModGenePlexus – where models are created for each module and uses the propagated genes found through semi-supervised learning – performed best. In other words, using all aspects ModGenePlexus is better than only parts of the method. We ran the Wilcoxon signed-rank test on the log2(auPRC/prior) results for each disease to test if

ModGenePlexus results were significantly better than other simpler models. For the CREEDS disease, gene, and drug datasets, and for MAGMA gene predictions of GWASs with prediction threshold $p < 1 \times 10^{-2}$ **,**ModGenePlexus performs better than AllAssign in both its propagated and non-propagated forms (**Figure 3.18).** Additionally, propagated and non-propagated AllClus methods perform better for the three CREEDS datasets but not on any of the MAGMA results compared to AllAssign. The performance of ModGenePlexus is better than the propagated version of AllClus for all three CREEDS datasets and for MAGMA gene predictions of GWASs with prediction threshold $p < 1 \times 10^{-2}$ **(Figure 3.19)**. ModGenePlexus and the propagated version of AllClus is significantly better for the CREEDS datasets compared to the non-propagated modGenePlexus (**Figure 3.20**). **Figure 3.21-22**, shows the correlation of results with the number of understudied genes in the CREEDS disease dataset. For **Figure 3.21**, AllAssign performance is correlated with test geneset size, in line with observations from the original GenePlexus paper, and in **Figure 3.22** we show there is negative correlation with the number of test genes and performance for all model types, where traits with more test genes have worse performance. Notably, the disparity between AllAssign and ModGenePlexus methods performance gets more notable for larger test set sizes, indicating that while ModGenePlexus still performs relatively worse with larger datasets, it does better than original GenePlexus. Overall, we demonstrate that ModGenePlexus is beneficial for most experimental results through each of its important additions. ModGenePlexus and AllClus perform well, indicating that removing genes with minimal network connections and adding new genes from propagation significantly improve results. Adding new genes from the propagation improves the supervised learning predictions. Lastly, creating models for each module rather than a single module at once improves overall performance of genome-wide gene rankings even when having to do an additional step of aggregating the predictions. Notably however, the GWAS results where genes were assigned using p-value thresholds had mixed results, raising questions about if modules can be used to improve GWAS gene prioritization and classification. **Figure A3.1-4** contains results showing model performance for each CREEDS and MAGMA geneset collection.

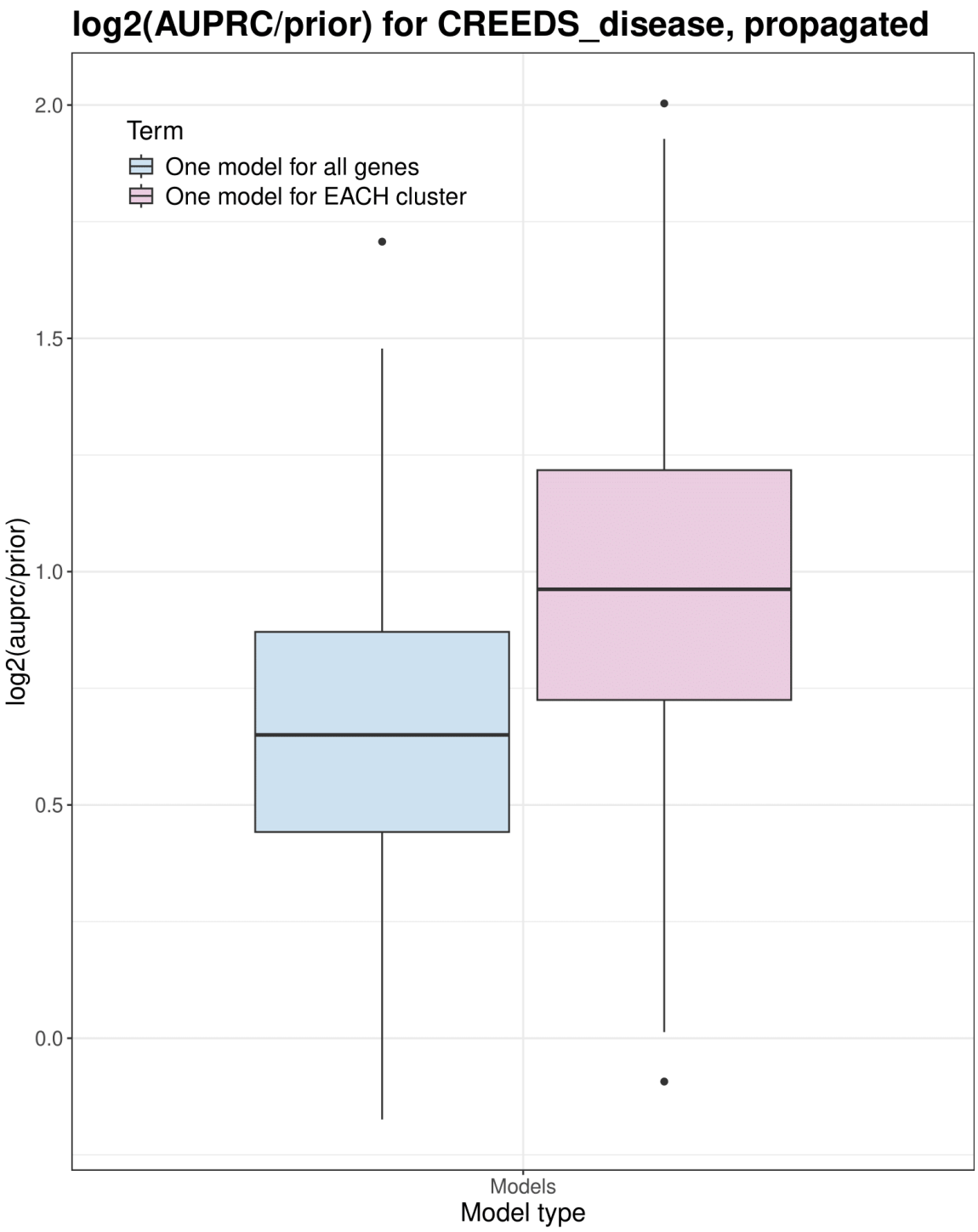

Figure 3.14: Comparing GenePlexus and ModGenePlexus performance. GenePlexus (blue) is creating one model for all genes in the experimental result. ModGenePlexus is creating

Figure 3.14 (cont'd)

multiple models for each module discovered through DOMINO and aggregating gene predictions across each model.

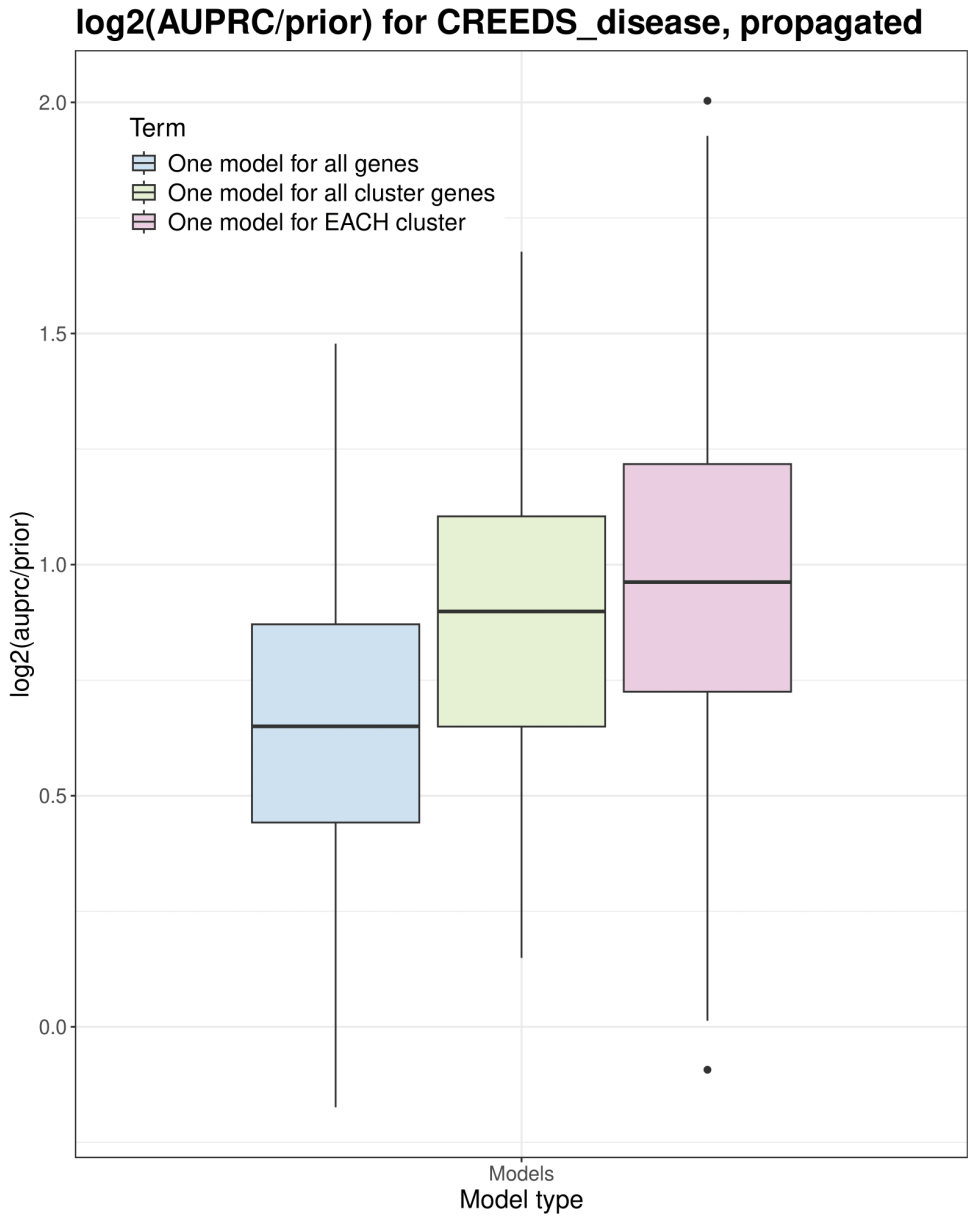

Figure 3.15: Comparing performance between GenePlexus, ModGenePlexus, and AllClus. AllClus (green) creates a single model for all genes that were assigned to a disease-gene enriched cluster.

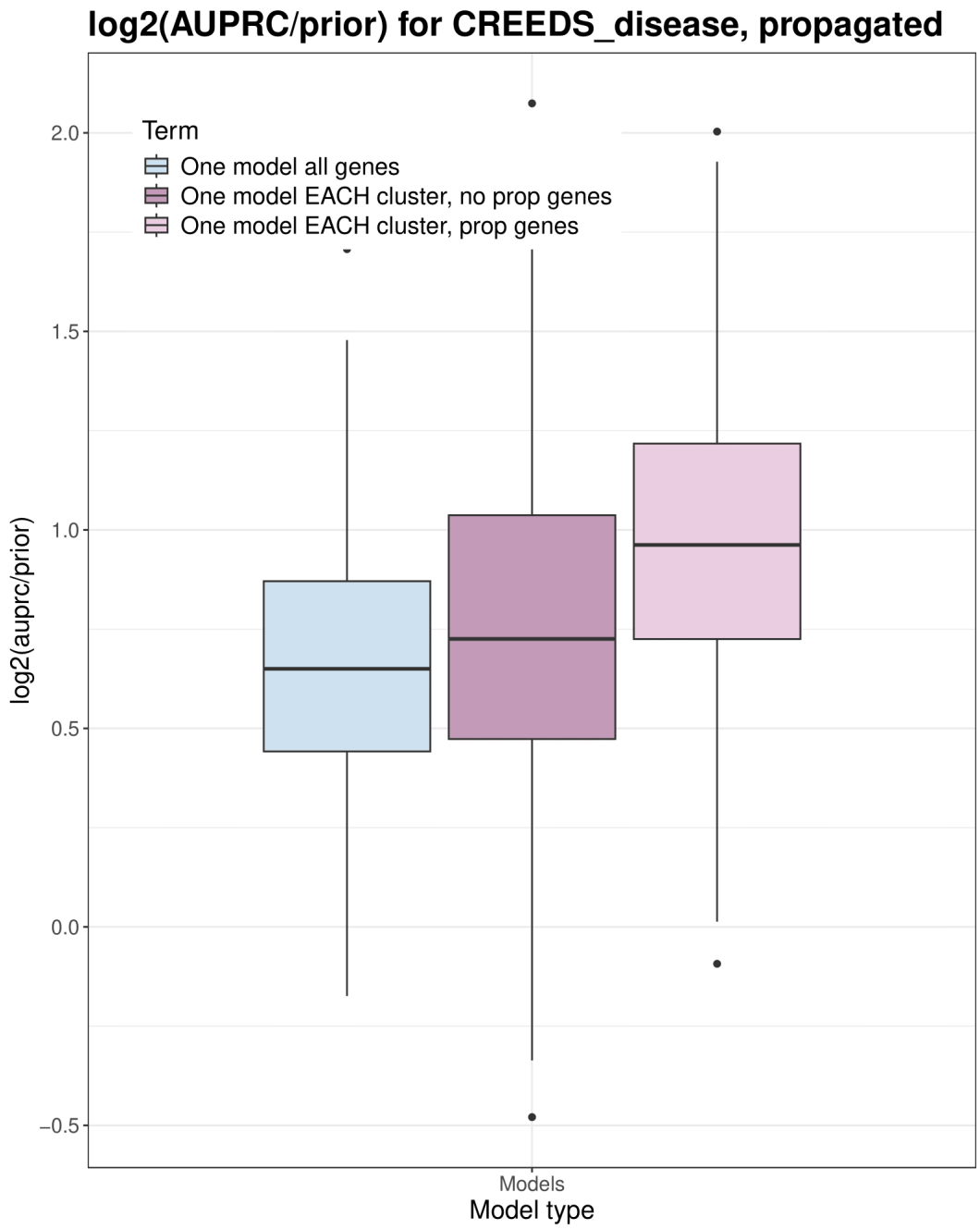

Figure 3.16: Comparing model performance of AllAssign, ModGenePlexus, and a version of ModGenePlexus that does not use additional propagated genes from DOMINO (dark purple).

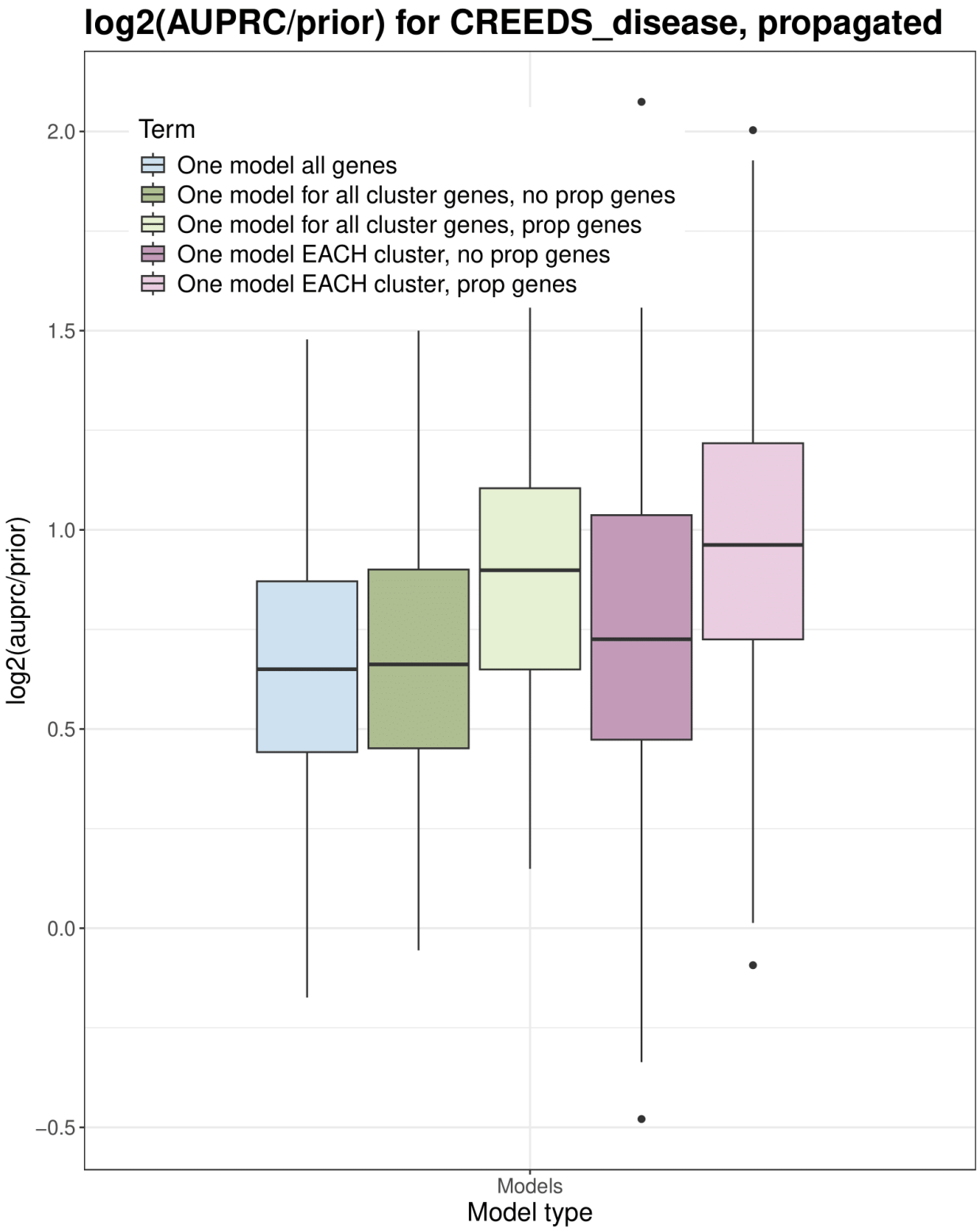

Figure 3.17: Comparing model performance with the inclusion of AllClus for both propagated (light green) and non-propagated (dark green) genelists.

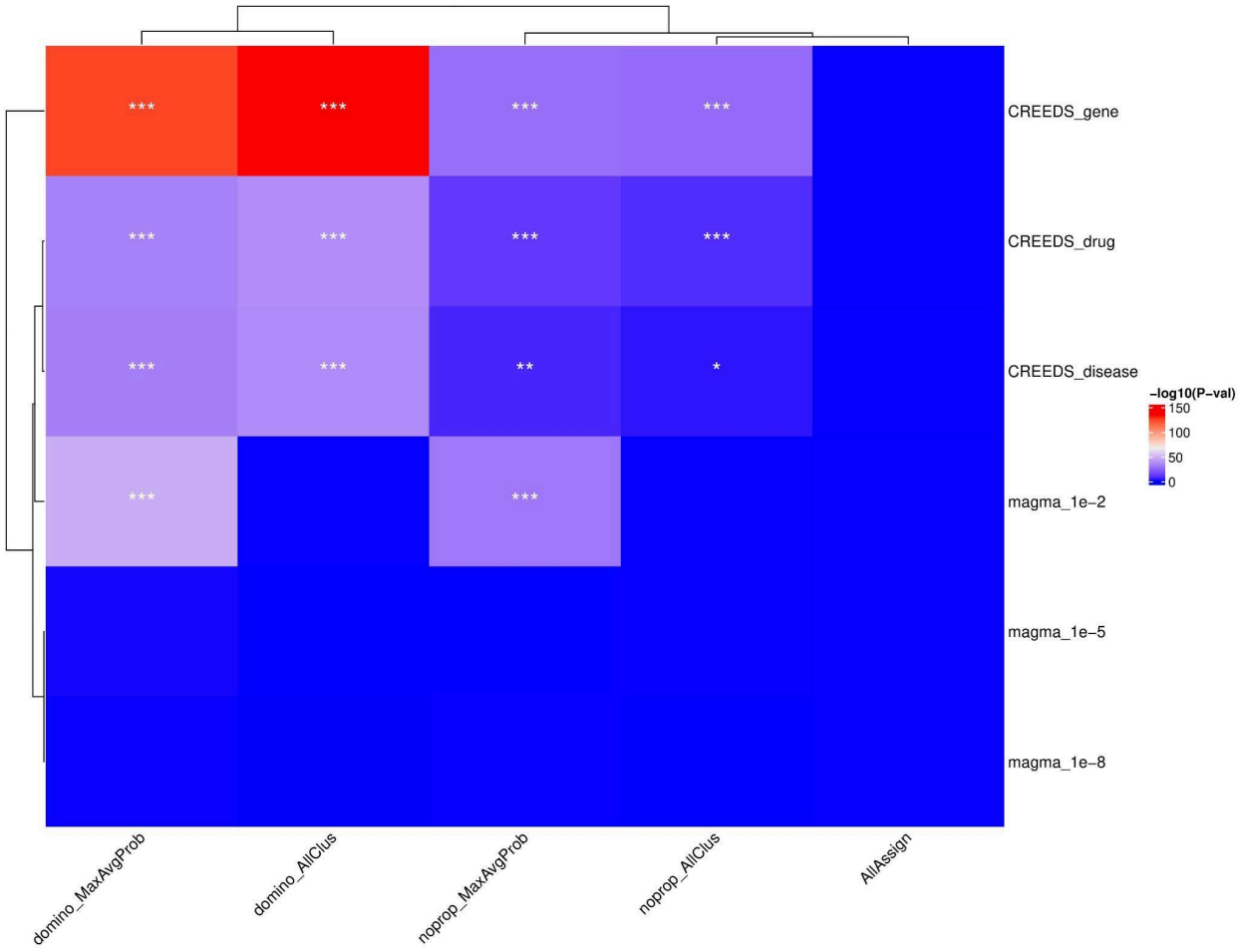

Figure 3.18: Wilcox test demonstrating if results of models are significantly better than GenePlexus (AllAssign) models for all genesets. The cells are annotated if the model is significantly better to AllAssign for the row's geneset. One star corresponds to a significance threshold of $p < 1 \times 10^{-3}$, two stars a threshold of $p < 1 \times 10^{-5}$, and three stars a threshold of $p < 1 \times 10^{-8}$.

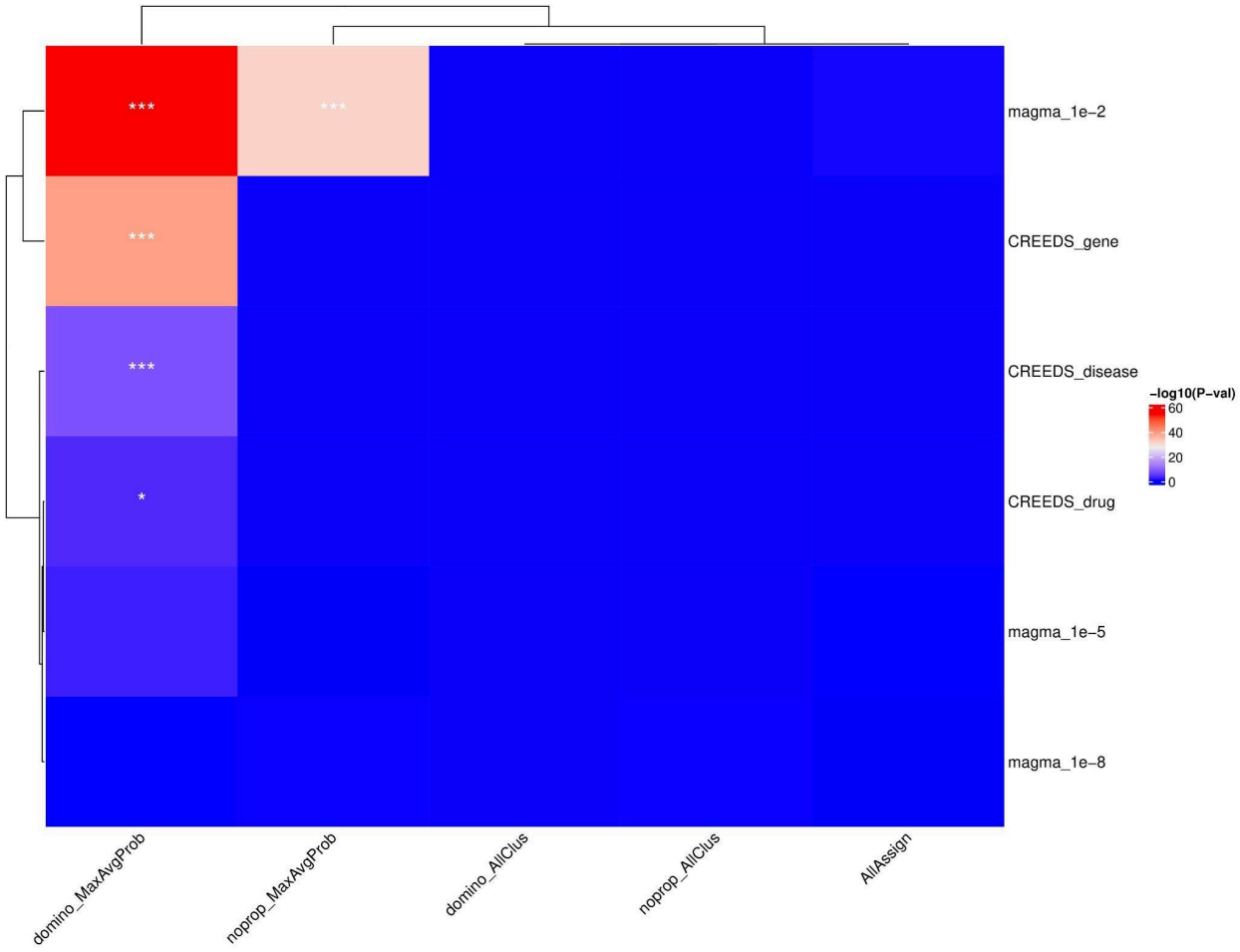

Figure 3.19: Wilcox test demonstrating if results of models are significantly better than the propagated version of AllClus models for all genesets. The cells are annotated if the model is significantly better to domino_AllClus for the row's geneset. One star corresponds to a significance threshold of $p < 1 \times 10^{-3}$, two stars a threshold of $p < 1 \times 10^{-5}$, and three stars a threshold of $p < 1 \times 10^{-8}$.

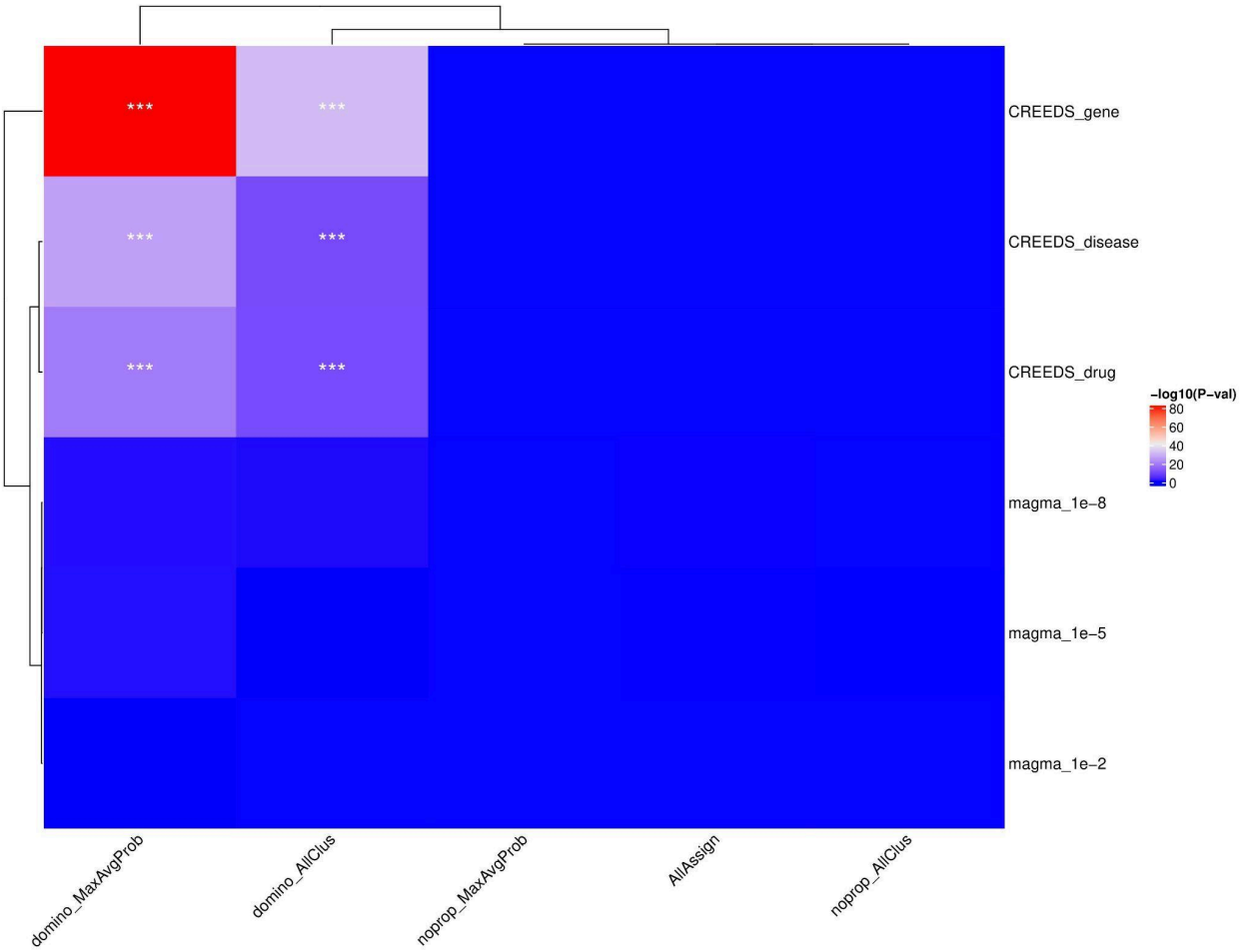

Figure 3.20: Wilcox test demonstrating if results of models are significantly better than the non-propagated version of modGenePlexus models for all genesets. The cells are annotated if the model is significantly better to noprop_MaxAvgProb for the row's geneset. One star corresponds to a significance threshold of $p < 1 \times 10^{-3}$, two stars a threshold of $p < 1 \times 10^{-5}$, and three stars a threshold of $p < 1 \times 10^{-8}$.

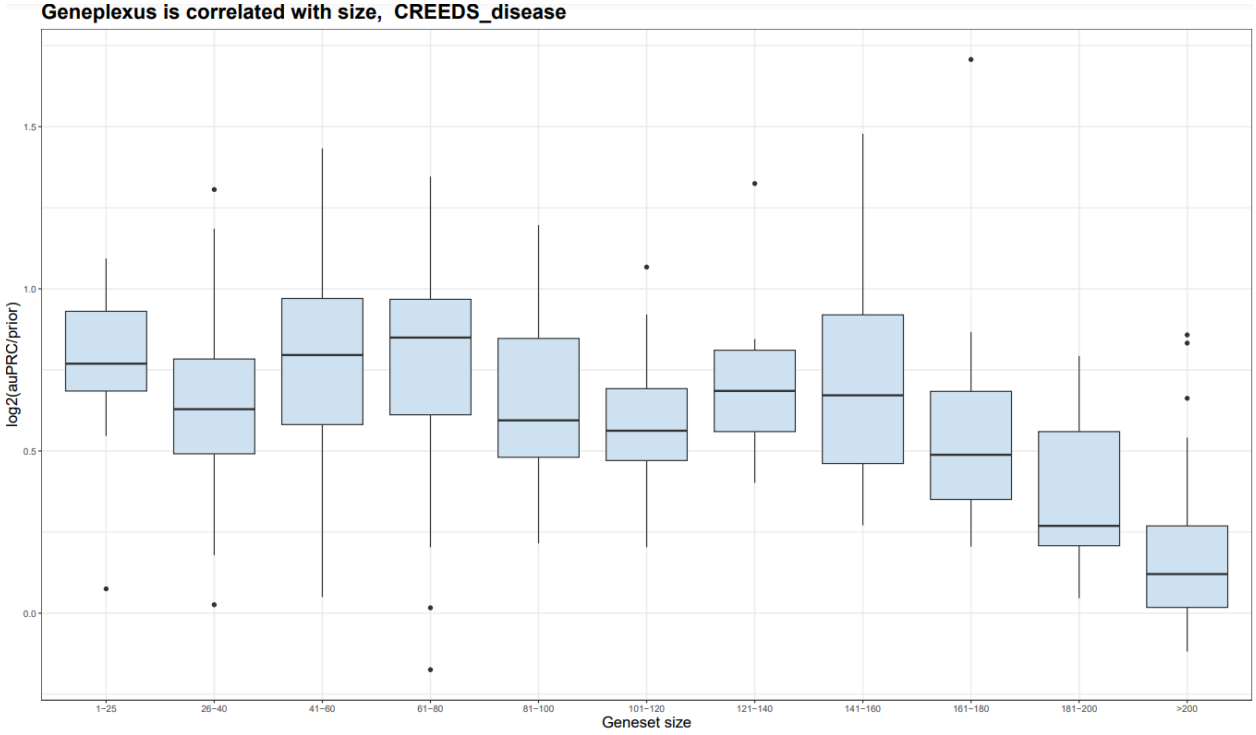

Figure 3.21: GenePlexus performance for CREEDS disease datasets is correlated with size. Performance of GenePlexus for AllAssign is negatively correlated with the number of understudied genes that were part of the experimental set.

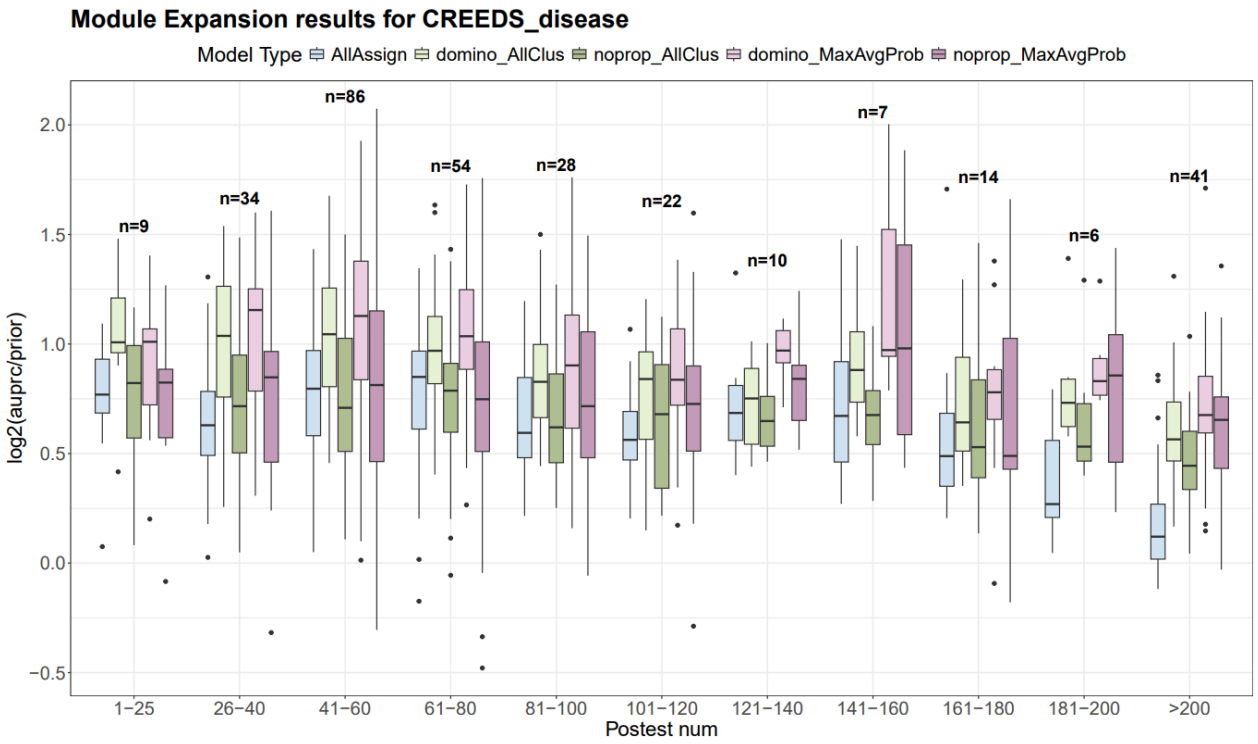

Figure 3.22: Comparing model performance of each model type with the number of understudied genes for CREEDS Disease experiments. Module expansion methods show more improvement when there are a larger number of test genes. The annotation "n=X" refers to the number of CREEDS disease genesets in the bin.AllAssign (blue), AllClus (green) where light green is the propagated version, and ModGenePlexus (purple) where light purple is using propagated genes.

ModGenePlexus allows for better prediction of stringent GWAS data

The previous results show that ModGenePlexus gives superior performance for CREEDS datasets, but the results for GWAS data were more mixed. They were better when choosing genes from MAGMA gene prioritization results using a threshold of $p < 1 \times 10^{-2}$, but not the stricter thresholds. GWAS results are challenging to work with due to needing correction for multiple hypothesis testing at both the summary statistic level and with gene prioritization algorithms like MAGMA. Notably, one of the strict thresholds for MAGMA genesets of $p < 1 \times 10^{-5}$ performed well in GenePlexus – outperforming the looser threshold and the CREEDs datasets (**Figure 3.23)**. The log2(auPRC/prior) of the MAGMA methods are often above 1.0 for thresholds $p < 1 \times 10^{-2}$ and $p < 1 \times 10^{-5}$, indicating fairly good model performance, whereas

CREEDS data models perform worse than random prediction using GenePlexus. Crucially, ModGenePlexus gives improved performance for the looser threshold of $p < 1 \times 10^{-2}$ (**Figure 3.18**) for MAGMA genesets. This geneset collection actually had the most significant improvement over the AllClus models (**Figure 3.19**). We hypothesized that discovering clusters with only the stringent threshold worsened the quality of the input data for downstream supervised learning. Stringent thresholding is done to remove false positive results, but has the disadvantage of increasing the number of false negatives. However, the semi-supervised module discovery can be used instead to remove bad hits and false positives, where biological data is used to determine relevance rather than a naive p-value decision threshold. We tested if utilizing genes that passed looser thresholds in the MAGMA results could be used in the ModGenePlexus pipeline to improve predictions of test genes in a more stringent set. In other words, could utilizing genes that meet a threshold of $p < 1 \times 10^{-2}$ help classify understudied genes with $p < 1 \times 10^{-5}$. A $p < 1 \times 10^{-5}$ was chosen rather than a stricter $p < 1 \times 10^{-8}$ due to limitations in the number of available genesets, as very few MAGMA results had hundreds of genes predicted using the strictest threshold (**see methods**). These models were created and performance was evaluated for all GWAS that met geneset size requirements(**Figure 3.24**), and we visualize those GWAS traits which had notably poor performance in AllAssign models - defined as having log2(auPRC/prior) of below 2.0 **(Figure 3.25)**. **Figure 3.24-25** are visualized separately due to the observation that some of the GWAS did genuinely well with GenePlexus – and we wanted to see if ModGenePlexus is helpful for those datasets that did poorly. There was no significant performance increase for all GWAS, but poorly performing GWAS showed significant improvement (**Figure 3.26**). Interestingly, both ModGenePlexus utilizing the $p < 1 \times 10^{-2}$ MAGMA genes and using an AllClus version of this model gives significantly improved performance. In **Figure 3.27**, we added additional looser thresholds of $p < 5 \times 10^{-2}$ and $p < 1 \times 10^{-1}$ to see their impact on ModGenePlexus performance and if ModGenePlexus continues to improve classification at looser thresholds. We see that while the ModGenePlexus performance increase does indeed level off, the performance still does increase with these very loose

thresholds. In contrast, AllAssign models performed worse with looser thresholds. Plots including the GWAS that had an log2(auPRC/prior) > 2 originally are in **Figure 3.28**. These results demonstrate that using networks, propagation, and module discovery provides biological context to low-powered GWAS results, allowing recovery of previously defined false negatives for novel discovery of genes and can be a tool for post-GWAS analysis. Notably, this analysis provides further evidence that ModGenePlexus is most beneficial for genesets that perform notably poorly in normal GenePlexus.

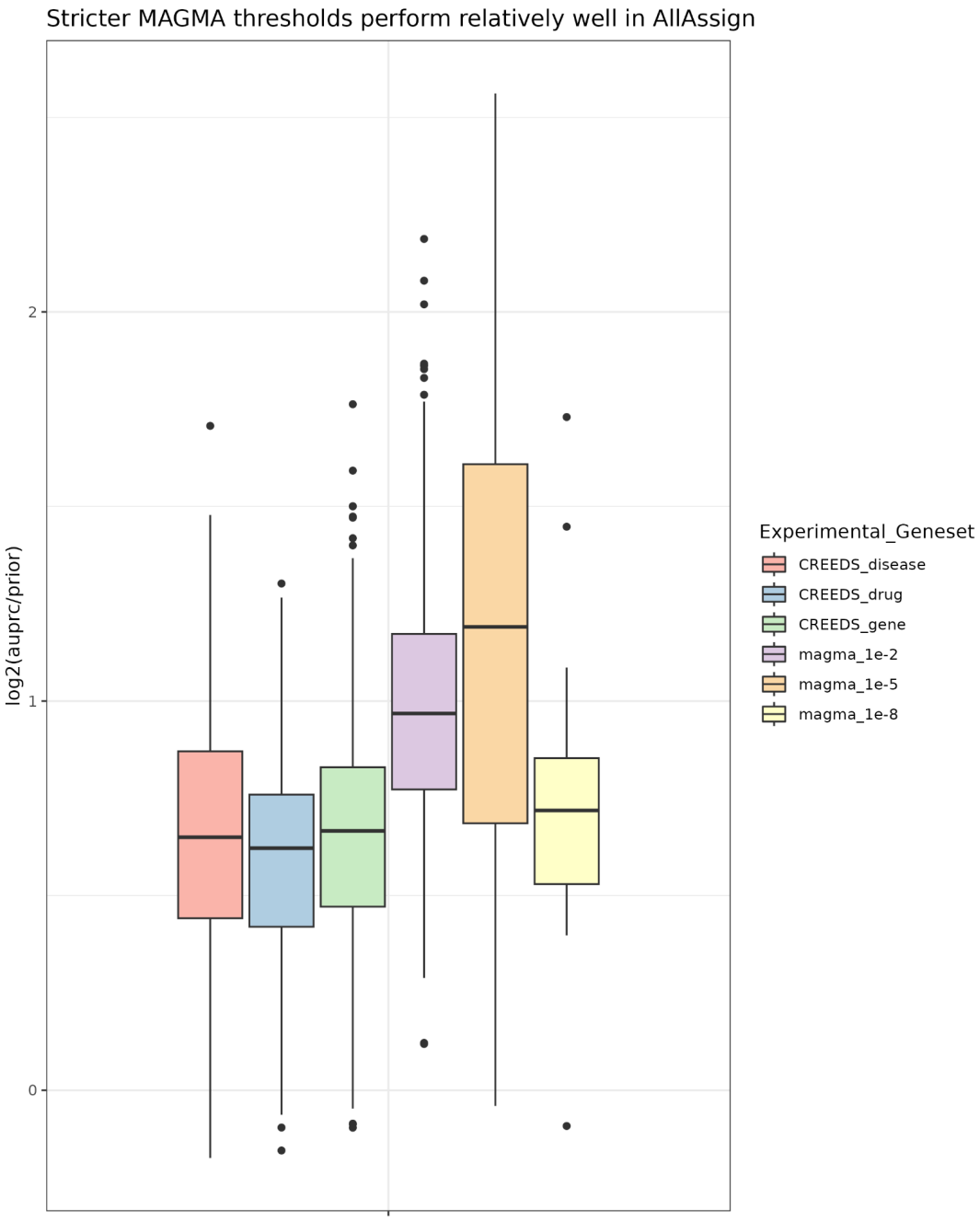

Figure 3.23: Comparing GenePlexus results of each geneset. The MAGMA genesets obtained from using thresholds of $p < 1 \times 10^{-2}$ and $p < 1 \times 10^{-2}$ perform notably better than CREEDS datasets. The strictest threshold also performs notably poorly in GenePlexus.

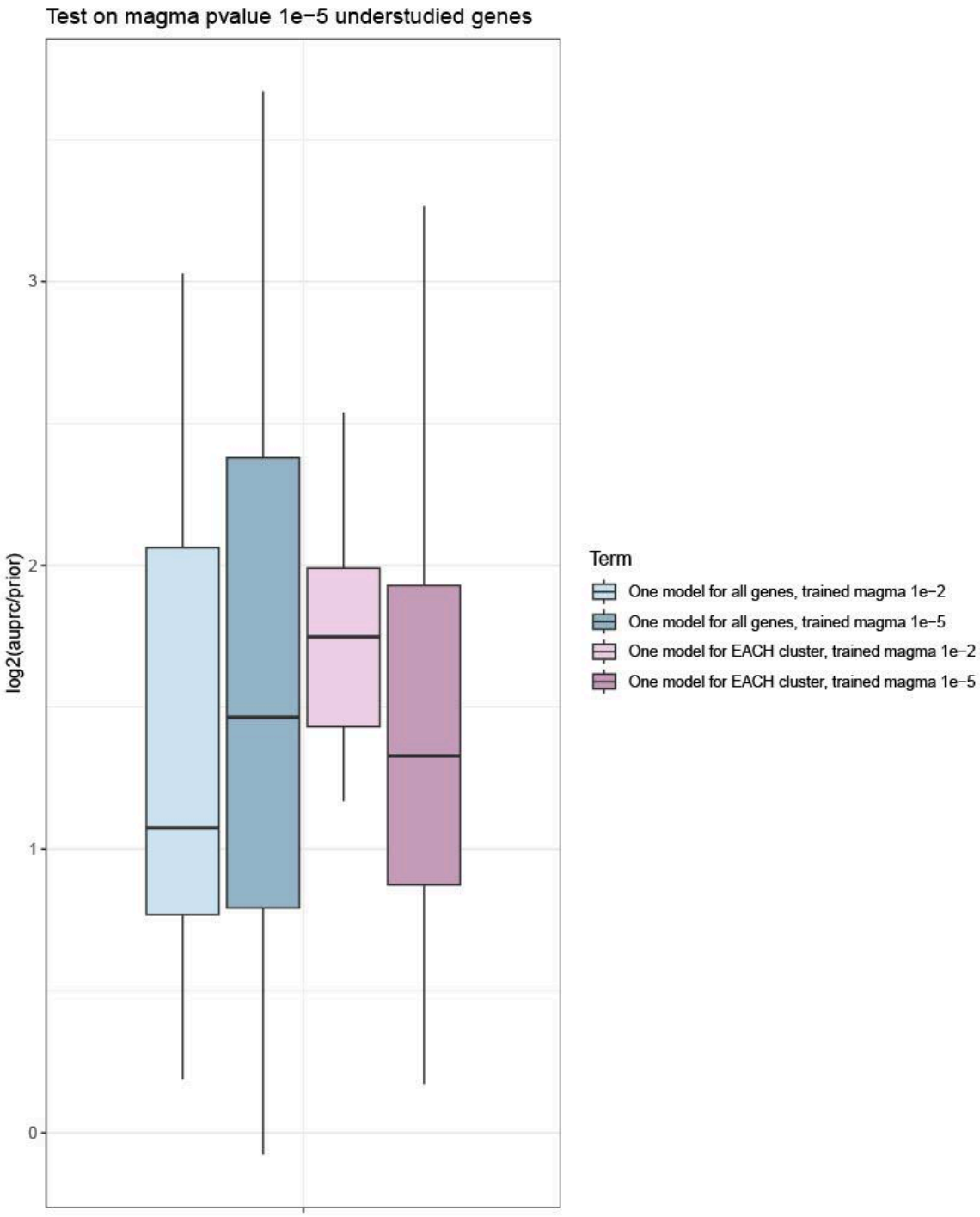

Figure 3.24: Performance of using models trained on nominal threshold genes when evaluated on stringent test sets for all GWAS. The GWAS here include those that perform well in GenePlexus with an log2(auPRC) > 2.0.

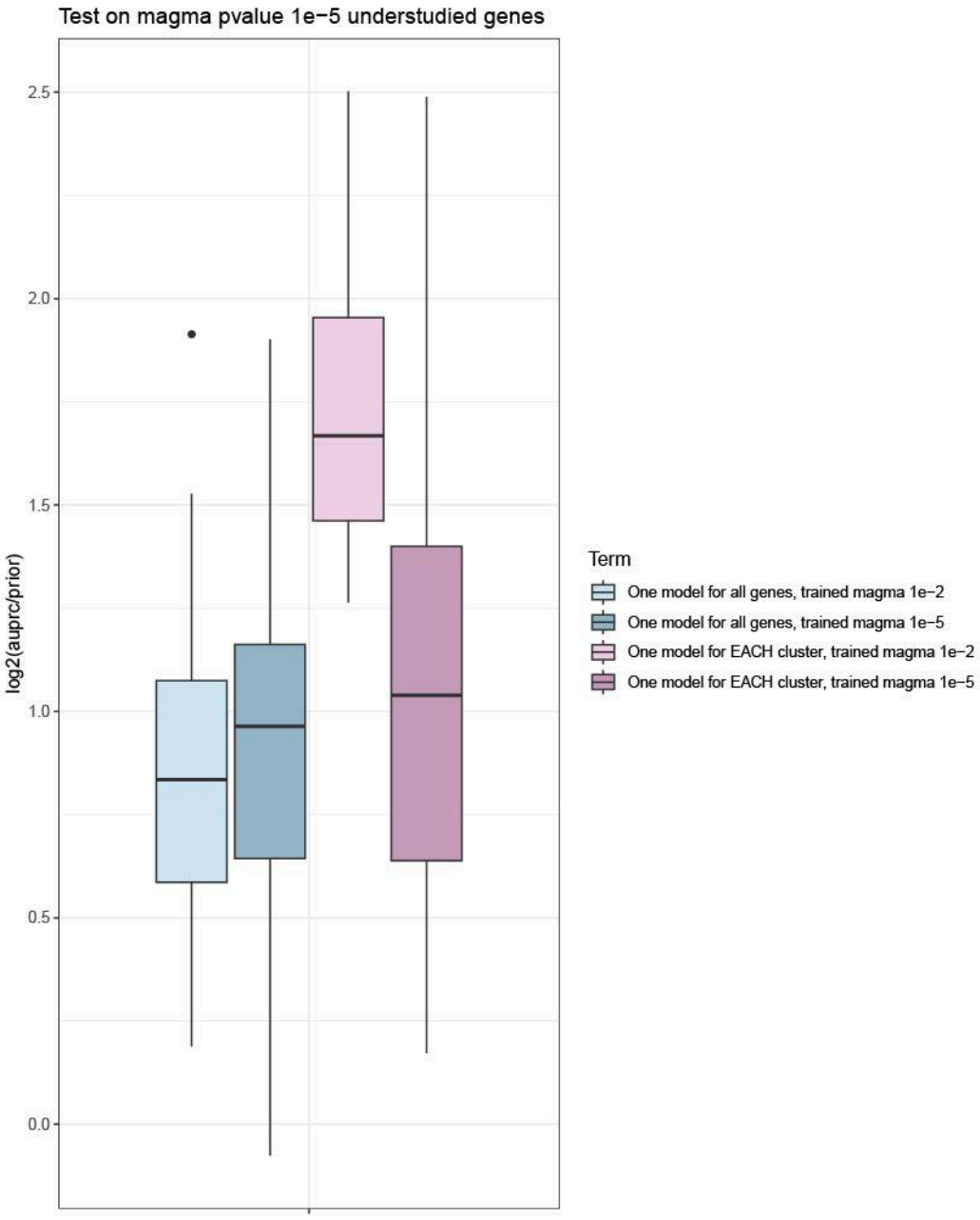

Figure 3.25: Performance of using models trained on nominal threshold genes when evaluated on stringent test sets for all GWAS log2(auPRC/prior) < 2. ModGenePlexus shows significant improvement for genesets that performed poorly in GenePlexus.

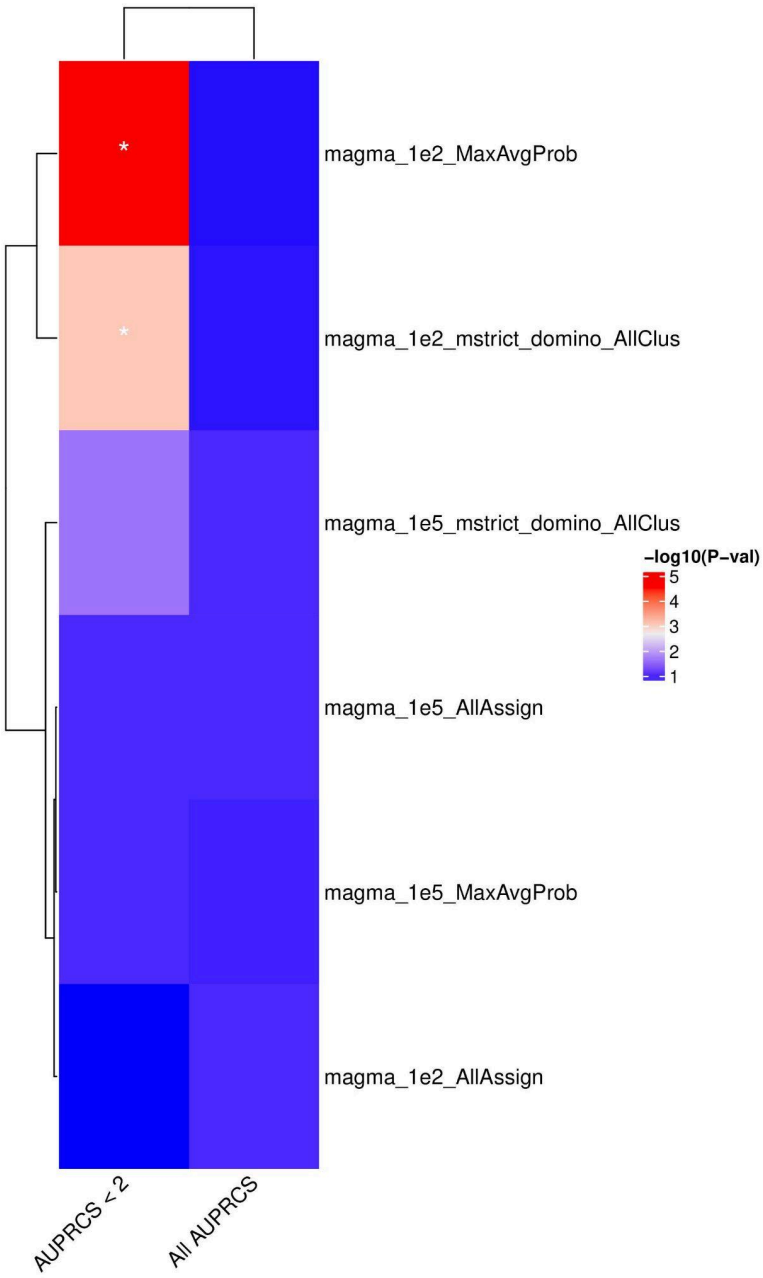

Figure 3.26: Wilcox test demonstrating if results of models are significantly better than the non-propagated version of modGenePlexus models for all genesets. The cells are annotated if the model is significantly better to domino_AllClus for the row's geneset. One star corresponds to a significance threshold of $p < 1 \times 10^{-3}$, two stars a threshold of $p < 1 \times 10^{-5}$, and three stars a threshold of $p < 1 \times 10^{-8}$.

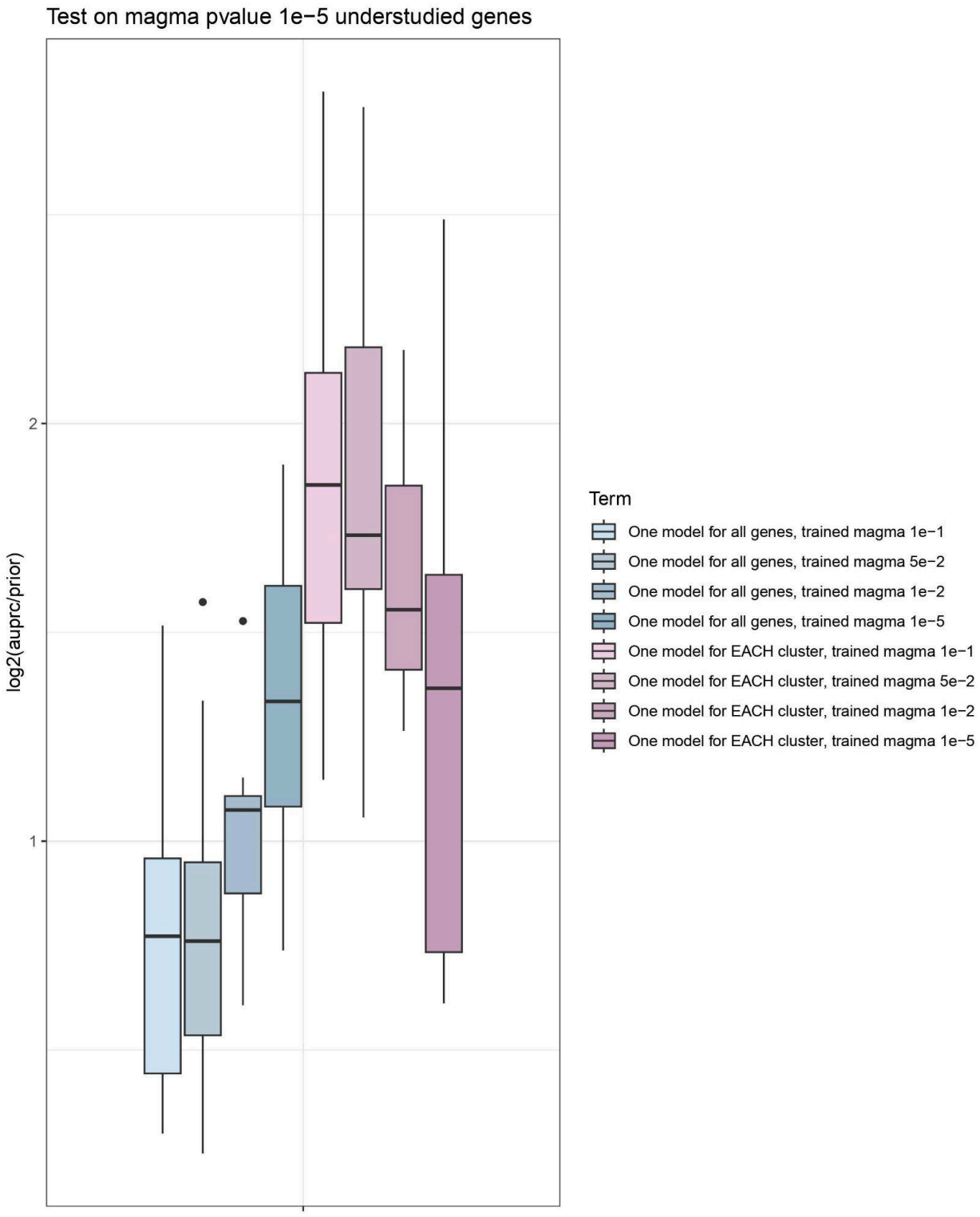

Figure 3.27: Model performance for models trained on various loose thresholds on the stringent test set for GWAS where the log2(auPRC < 2)**.** In addition to

Figure 3.27 (cont'd)

training on $p < 1 \times 10^{-2}$, additional models were trained using thresholds of $p < 5 \times 10^{-2}$ and $p < 1 \times 10^{-1}$. For GenePlexus, using additional genes discovered in looser thresholds makes performance worse, while ModGenePlexus performs better when including these genes.

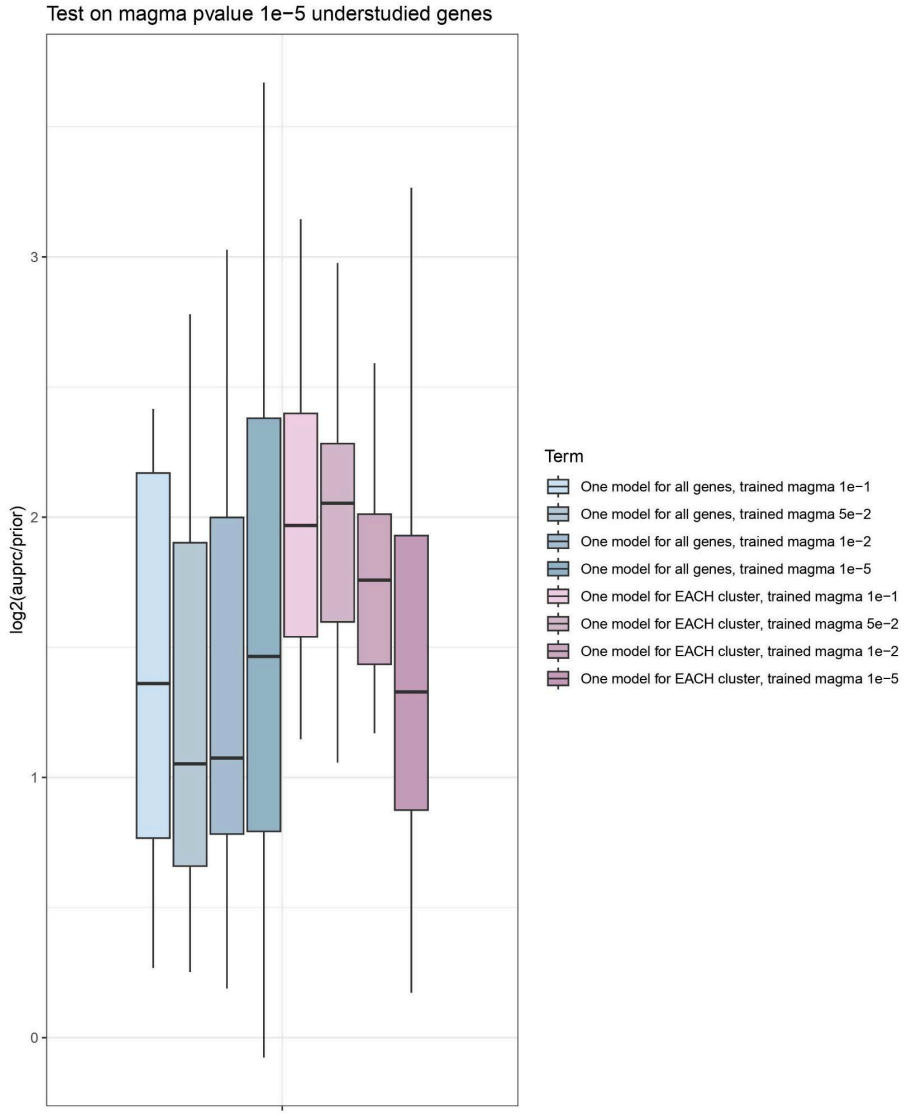

Figure 3.28: Model performance for models trained on various loose thresholds on the stringent test set for all GWAS. In addition to training on $p < 1 \times 10^{-2}$, additional models were trained using thresholds of $p < 5 \times 10^{-2}$ and $p < 1 \times 10^{-1}$.

Figure 3.28 (cont'd)

For GWAS that performed well in GenePlexus, there is less of a relationship between adding more genes or not.

Network edge density of genes shows net benefit of module discovery and propagation before supervised learning

In **Figure 3.3**, we showed that GenePlexus performance is correlated with edge density of the input geneset. A motivation for using modules for gene classification is to leverage smaller subsets of biologically significant genes localized to network neighborhoods, which means that genes in modules are more densely connected relative to the disease as a whole. **Figure 3.29** shows that for complex diseases from the CREEDS disease set, the average module edge density is higher for the larger diseases. This suggests the bigger the geneset, the better module discovery is at finding dense subsets of meaningful genes. For each model type, we calculated the network edge density (**see methods**) of the genesets used as inputs for the respective supervised learning models. For methods that use multiple cluster models, edge density is represented by either the average or max edge density across the trait modules depending on the figure. **Figures 3.30-32** shows that non-propagated clusters have higher average edge densities than propagated ones. Similarly, the non-propagated AllClus has higher edge density than the propagated AllClus **(Figure 3.33)**. AllAssign, as expected, has the lowest edge density. This means that discovering modules does indeed modify the genesets to be more dense for GenePlexus input. **Figure 3.34** shows a possible cause of this ranking of edge densities across the genesets – which is that the noprop genesets are notably smaller than the propagated ones. Because DOMINO propagates genes without the goal of maximizing edge density (**see methods**), the additional genes have meaning in terms of a modularity statistic and for other network properties, but this lowers the overall cluster densities. Despite the lower densities, using propagated genes in clusters give superior results for ModGenePlexus (**Figure 3.35**). We plotted the correlation of average edge density across clusters for propagated and non-propagated cluster genesets with performance (**Figure 3.36)**. There is no correlation between performance and average cluster edge density for propagated (**Figure 3.38**, Pearson Correlation; -0.03*)*, or non-propagated genesets (**Figure 3.39**,

Pearson Correlation; -0.09), even though a positive relationship between edge density and performance exists for the experimental disease differential expression results with the whole CREEDS disease geneset considered (**Figure 3.4)**. Using ModGenePlexus with smaller and denser clusters yields superior results compared to GenePlexus, but traits with more dense clusters tend to be bigger, and (Mod)GenePlexus performance has negative correlation with overall trait size (**Figures 3.29, 3.34-35)**. This could explain the lack of correlation between edge density of clusters with performance, as the bigger traits lead to more dense inputs in ModGenePlexus. This could also explain why ModGenePlexus/GenePlexus performance disparity gets more pronounced for larger inputs.

We further tested if there is a relationship with the max cluster density for each disease. **Figure 3.37** shows the relationship between max cluster density of CREEDS diseases with the total set size for propagated and non-propagated clusters, with similar results to the average cluster edge density.. Correlations for propagated and non-propagated genesets' max density values with performance are in **Figures 3.38-39** (Pearson; -0.13 and -0.10, respectively). We see that like with the average density across clusters, there is little/no correlation with performance and the max cluster density for either geneset type. Overall we demonstrated that geneset size impacts how dense the average discovered module will be, explaining the more significant improvement of ModGenePlexus for large traits (**Figure 3.22**) relative to normal GenePlexus. While edge density is no longer a directly good predictor of overall performance, it helps explain when ModGenePlexus improves relative to GenePlexus for large-scale data that has hundreds or thousands of gene annotations.

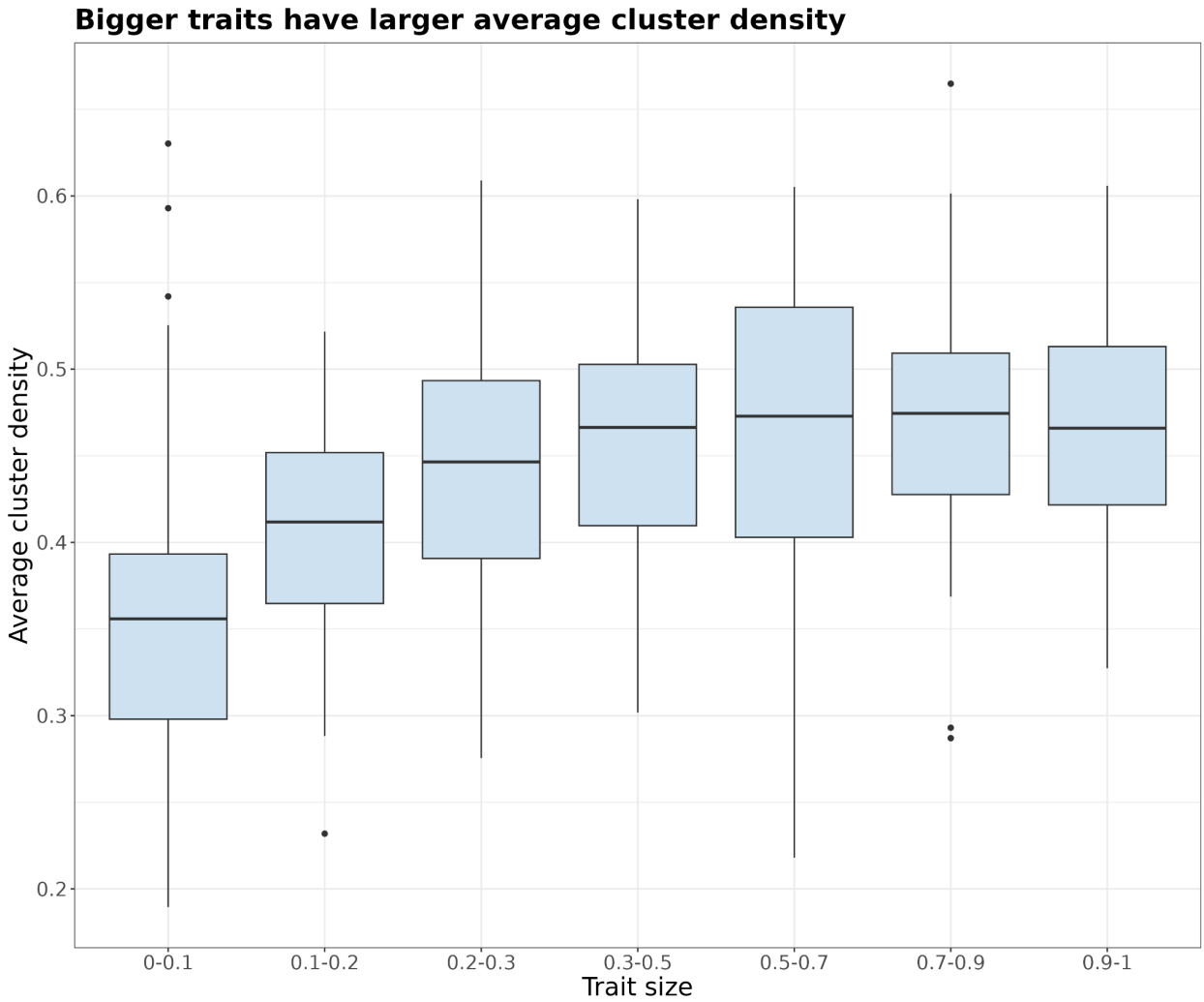

Figure 3.29: Plotting the relationship of trait size and the average edge density of obtained clusters with DOMINO. There is a positive correlation between size and average cluster edge densities.

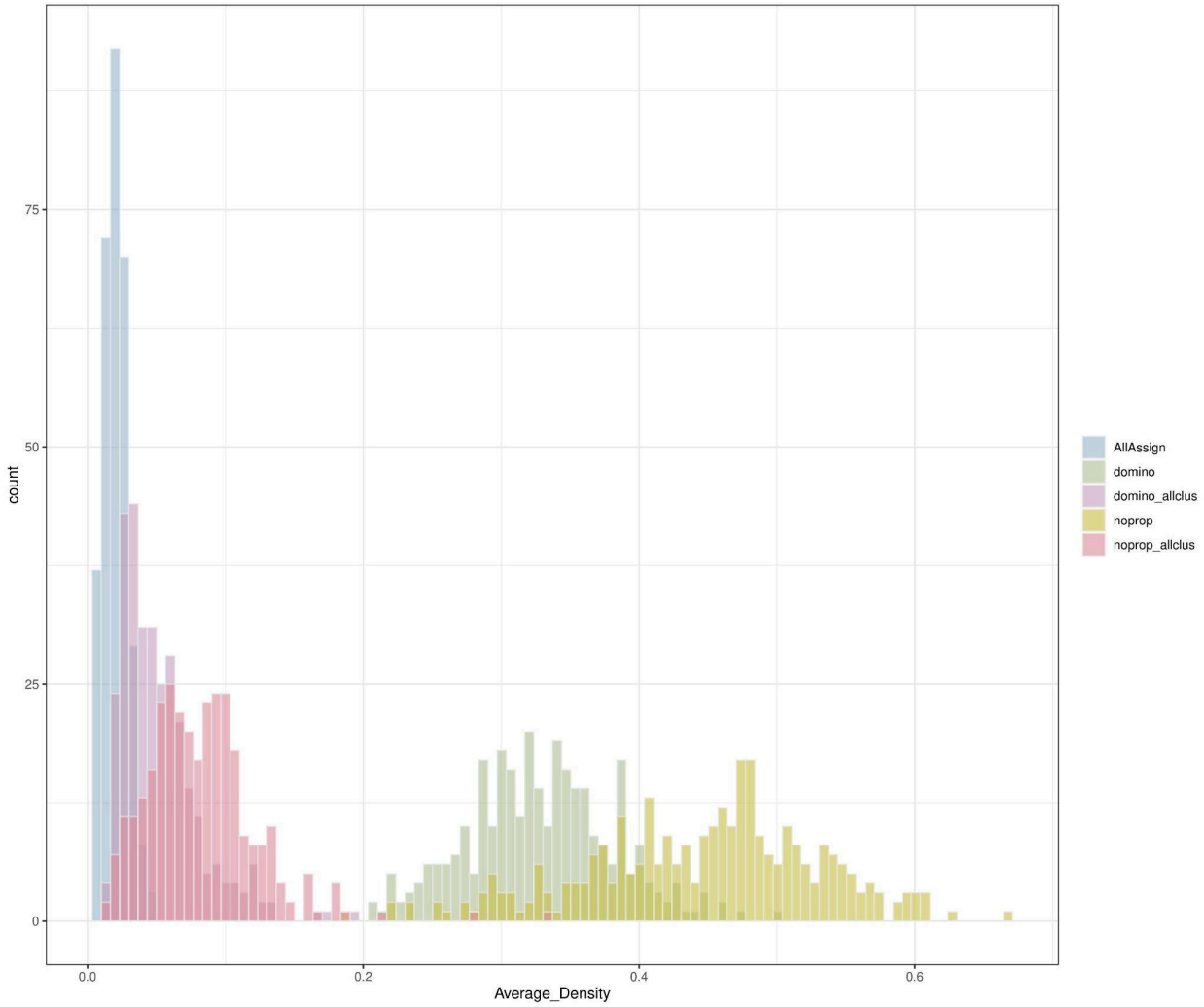

Figure 3.30: Histogram of edge densities for genesets used in models. AllAssign have the smallest edge densities. For the genesets domino_ and noprop_, the average cluster edge density of each geneset is plotted.

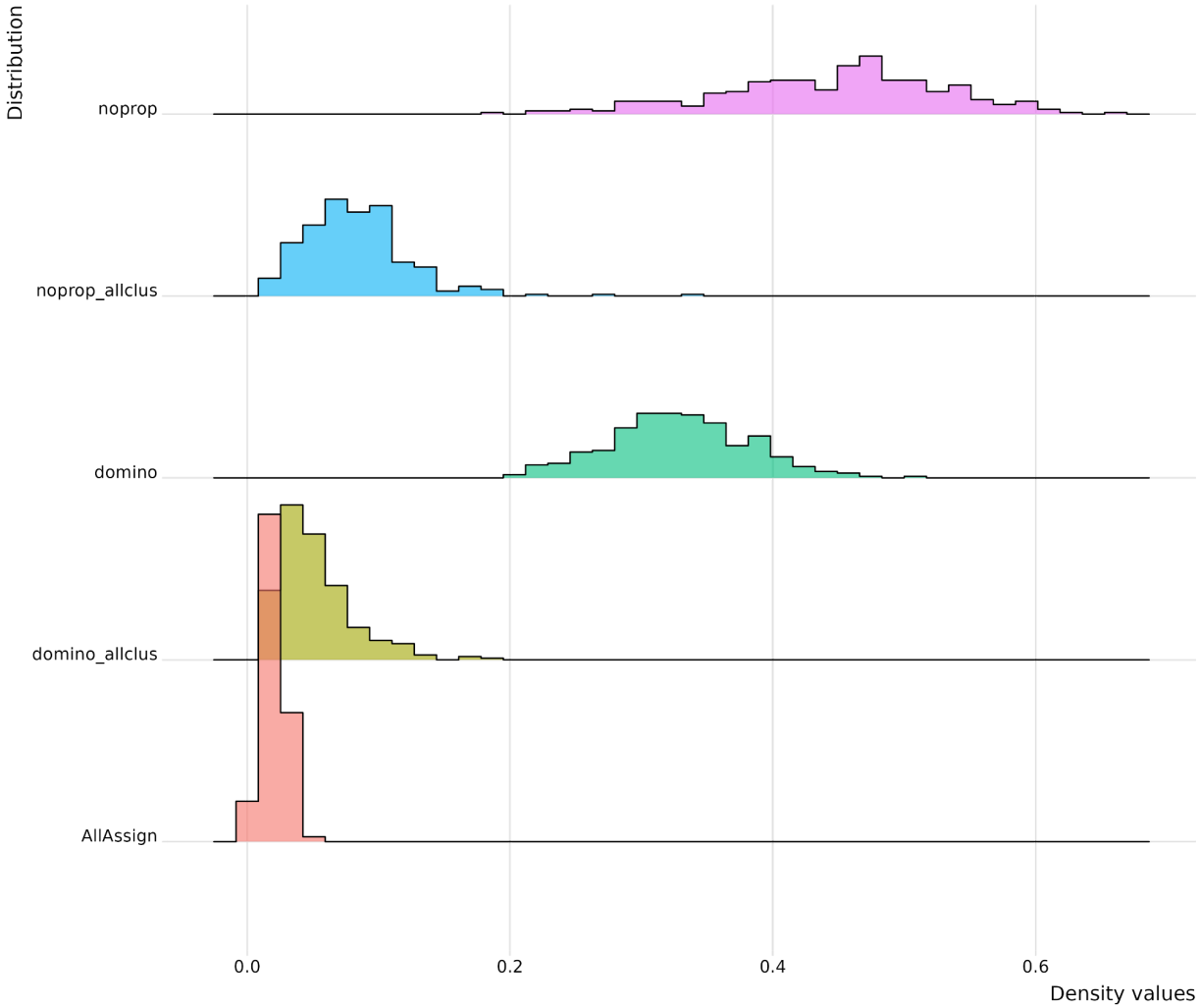

Figure 3.31: Ridge plot of densities across genesets. For the genesets domino_ and noprop_, the average cluster edge density of each geneset is plotted.

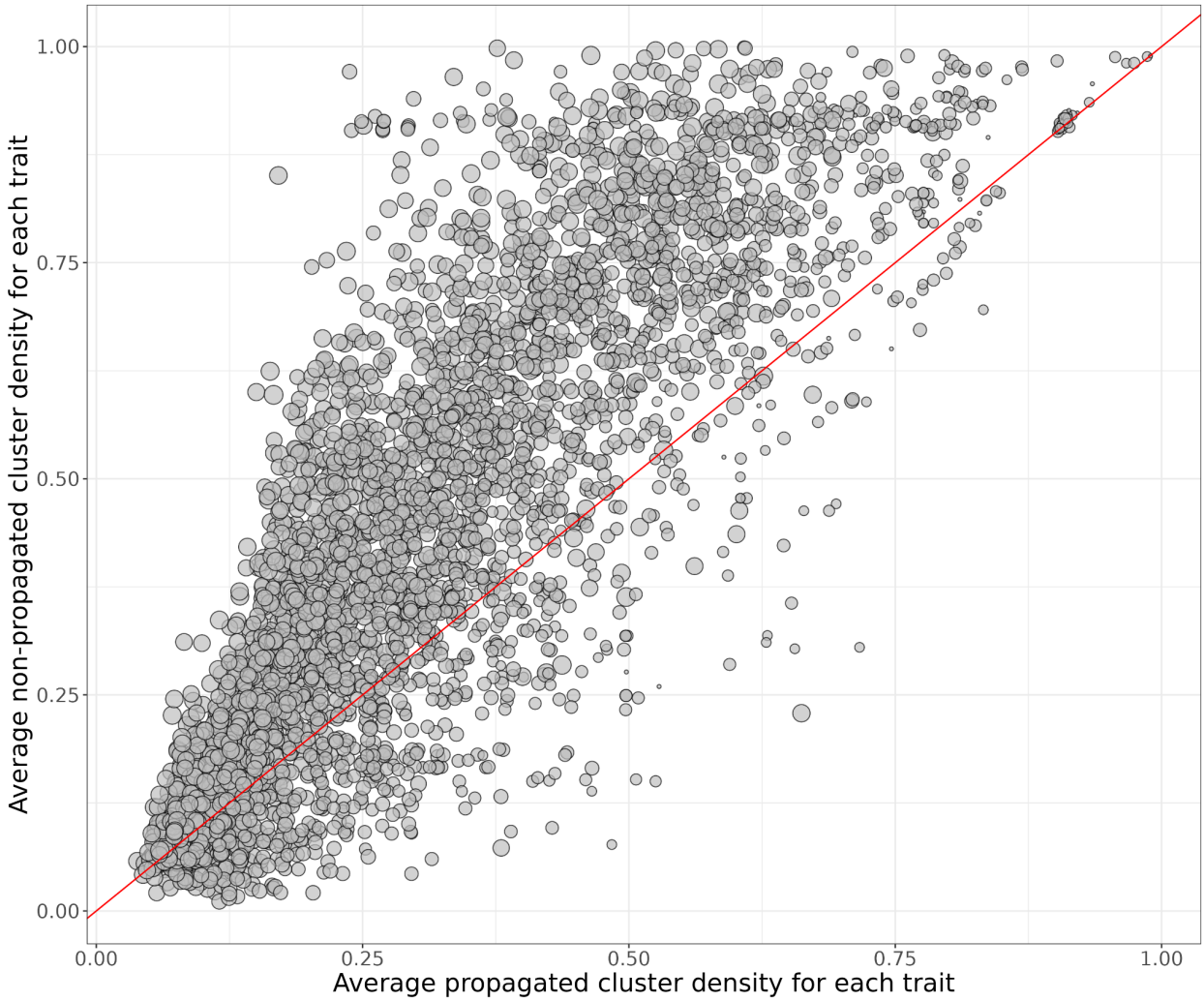

Figure 3.32: Contrasting the average cluster edge density of propagated and non-propagated clusters. For most traits, the average edge density of discovered clusters is higher in the non-propagated versions.

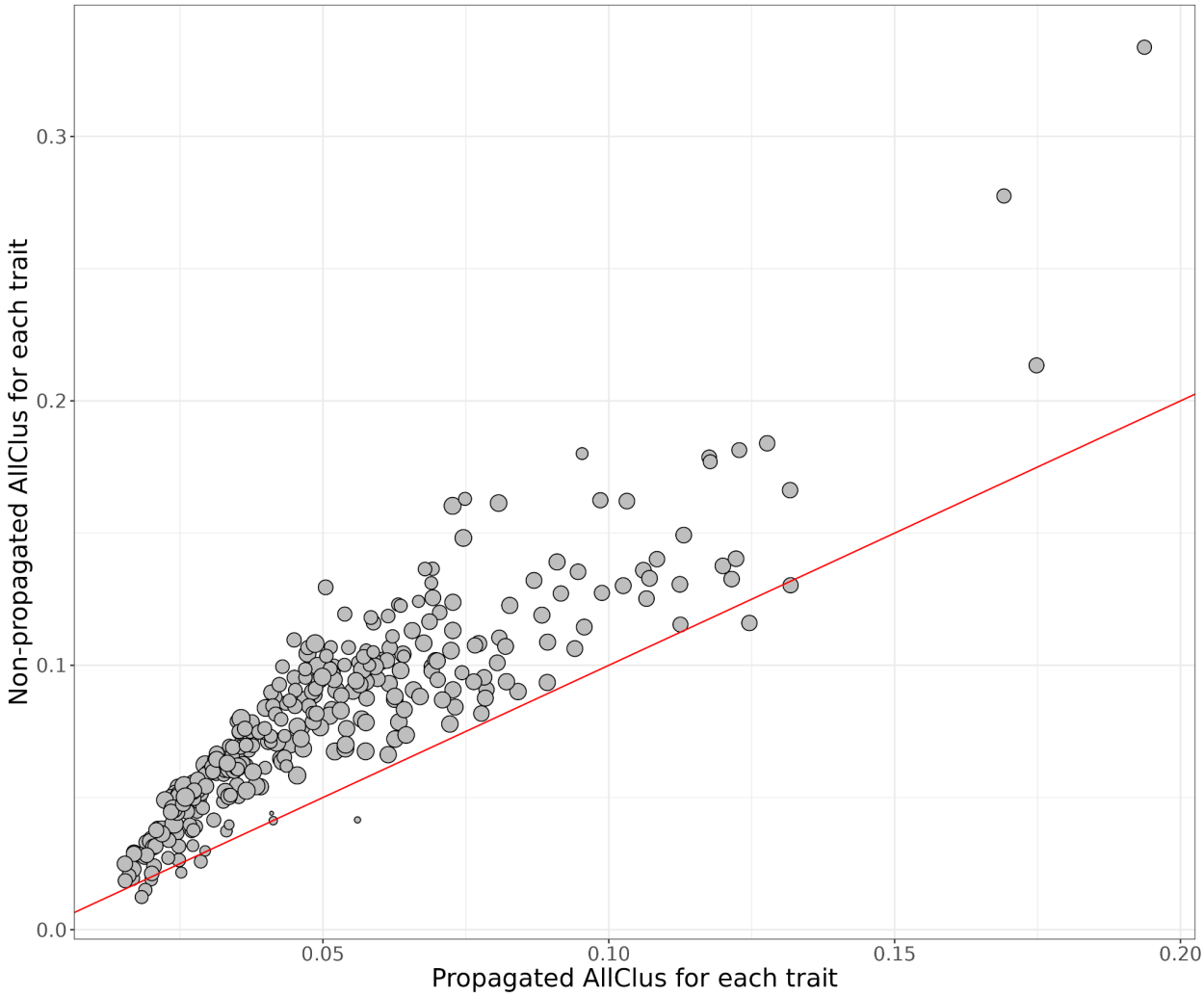

Figure 3.33: Contrasting the propagated and non-propagated versions of AllClus. For most traits, the non-propagated version has a higher edge density.

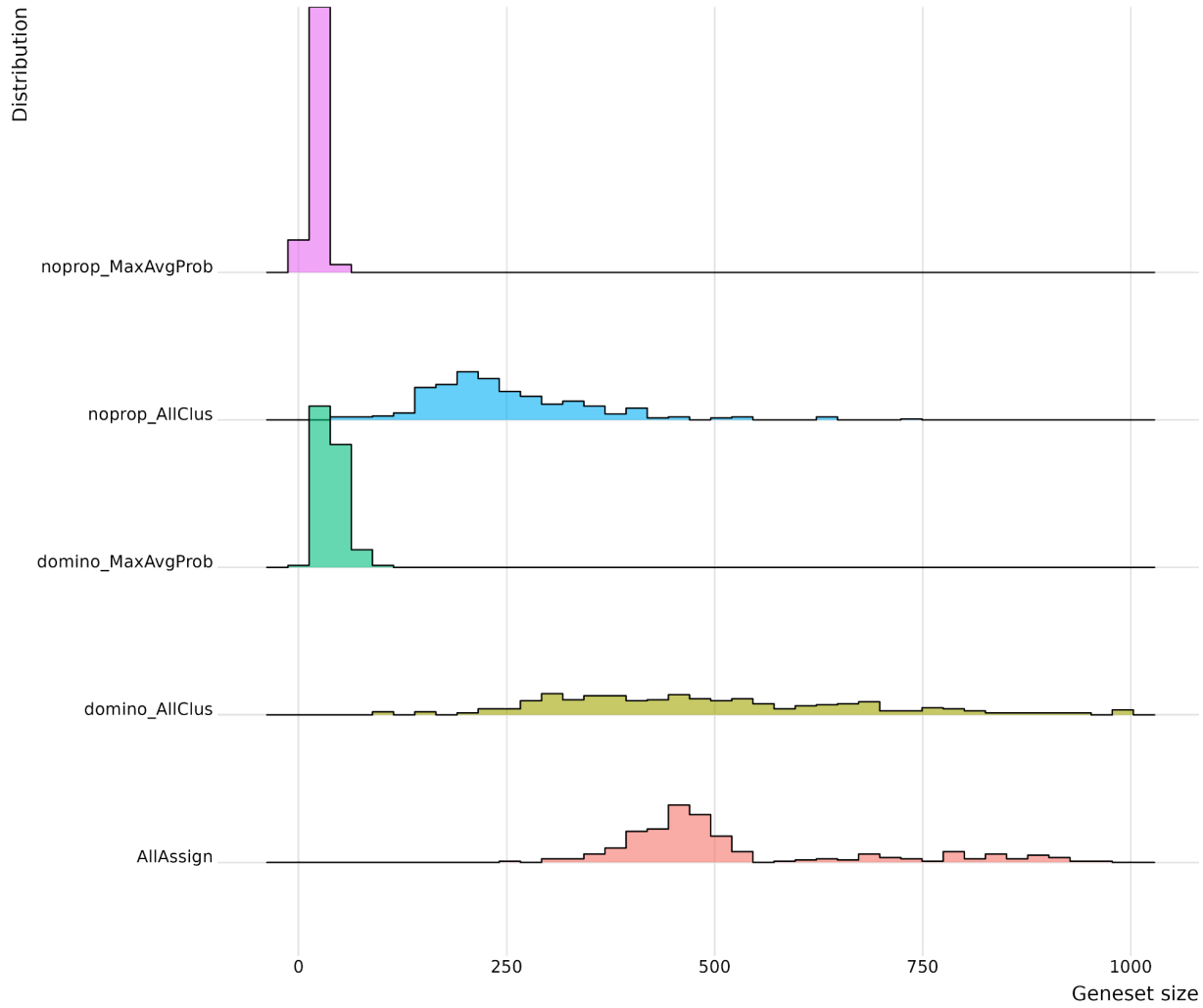

Figure 3.34: Ridge plot of size distributions for each geneset, subsetted to only include in AllAssign < 1,000 genes for readability. The order of the ridge plots is based on edge density distribution.

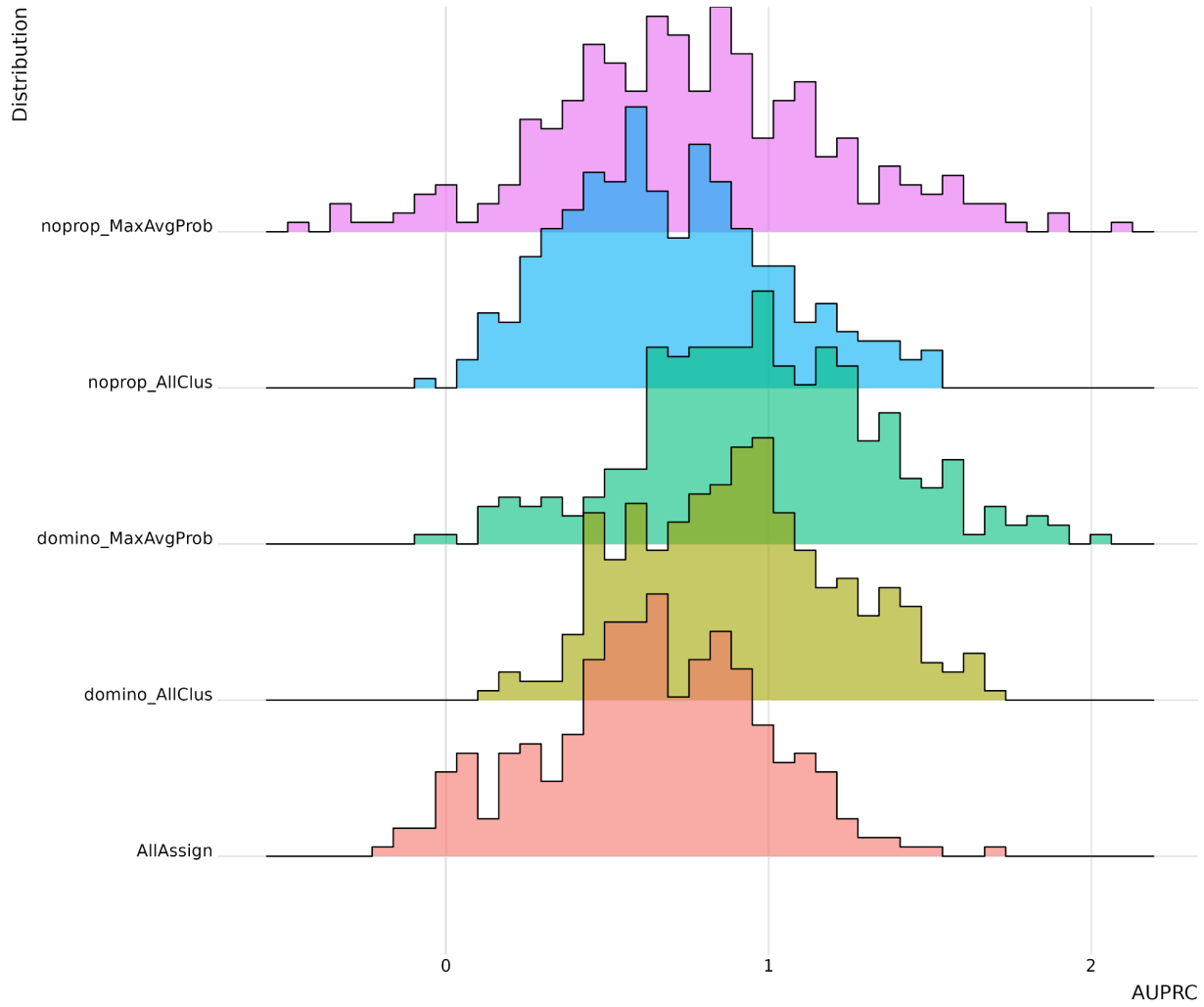

Figure 3.35: Ridge plot of model performance for each geneset, order of the ridge plots is based on edge density distribution.

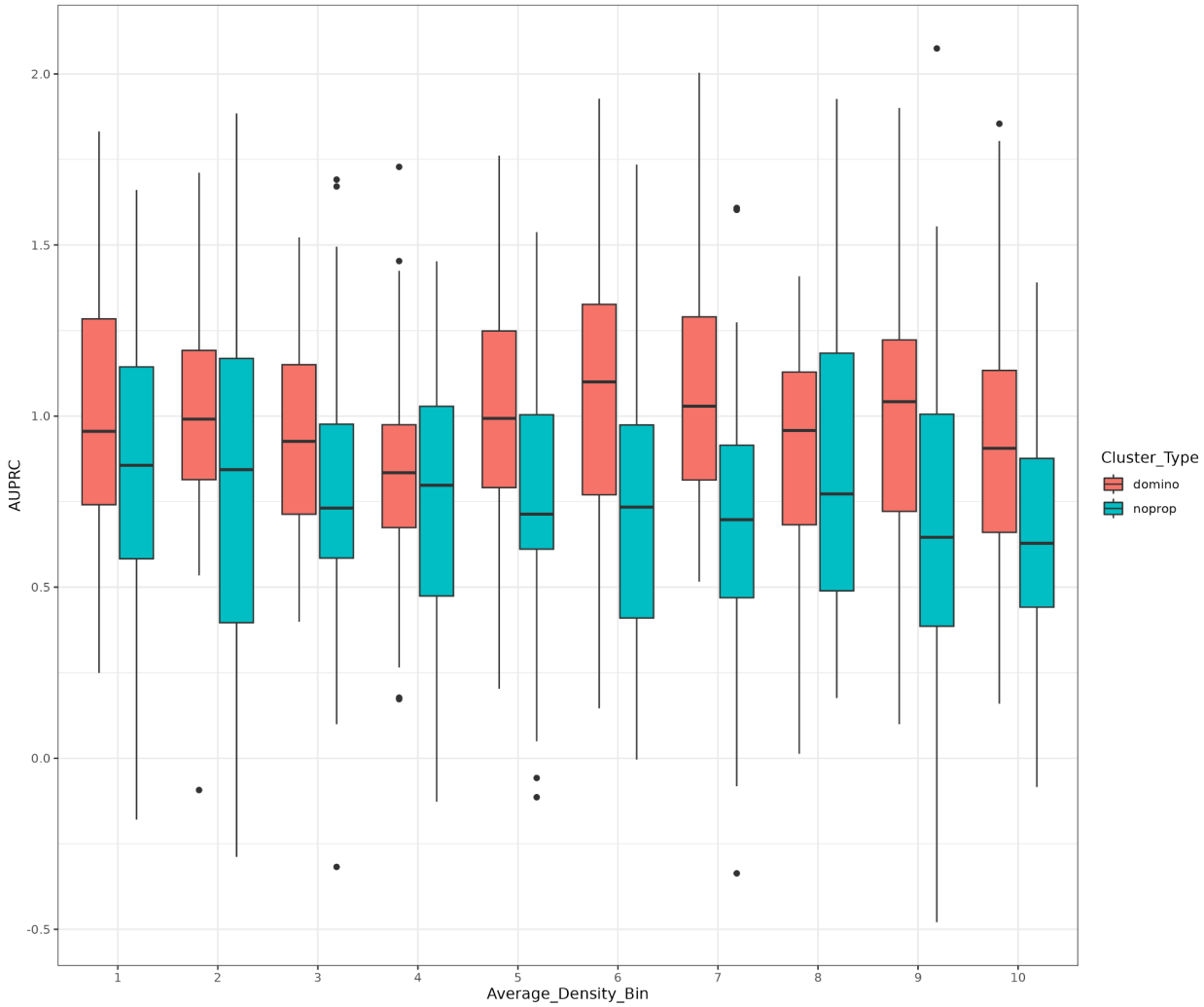

Figure 3.36: Relationship between the average cluster edge density and performance for propagated (domino) and non-propagated (noprop). There is no correlated between the average cluster density and performance for either geneset.

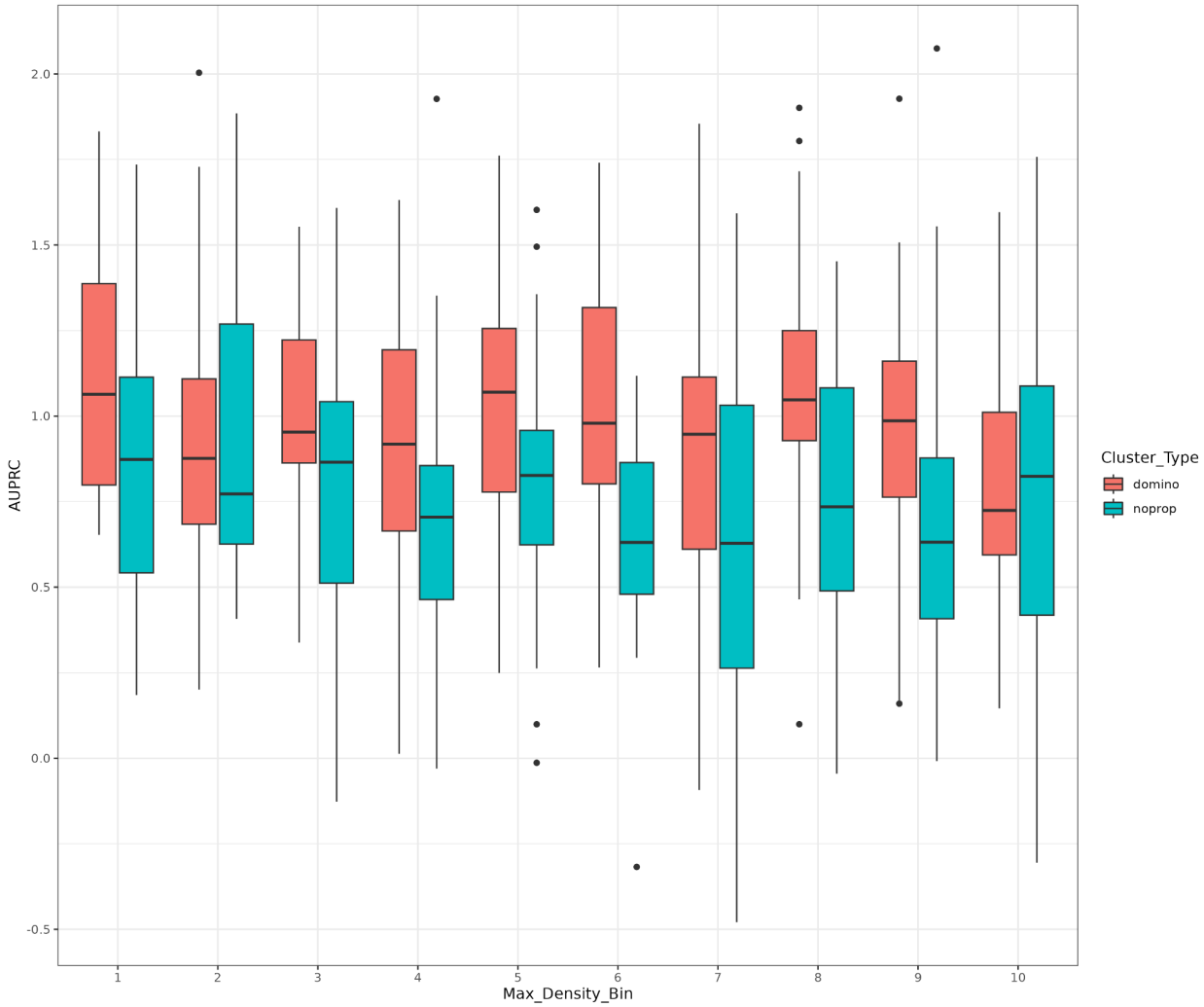

Figure 3.37: Relationship between the max cluster edge density and performance for propagated (domino) and non-propagated (noprop). There is no correlated between the max cluster density and performance for either geneset.

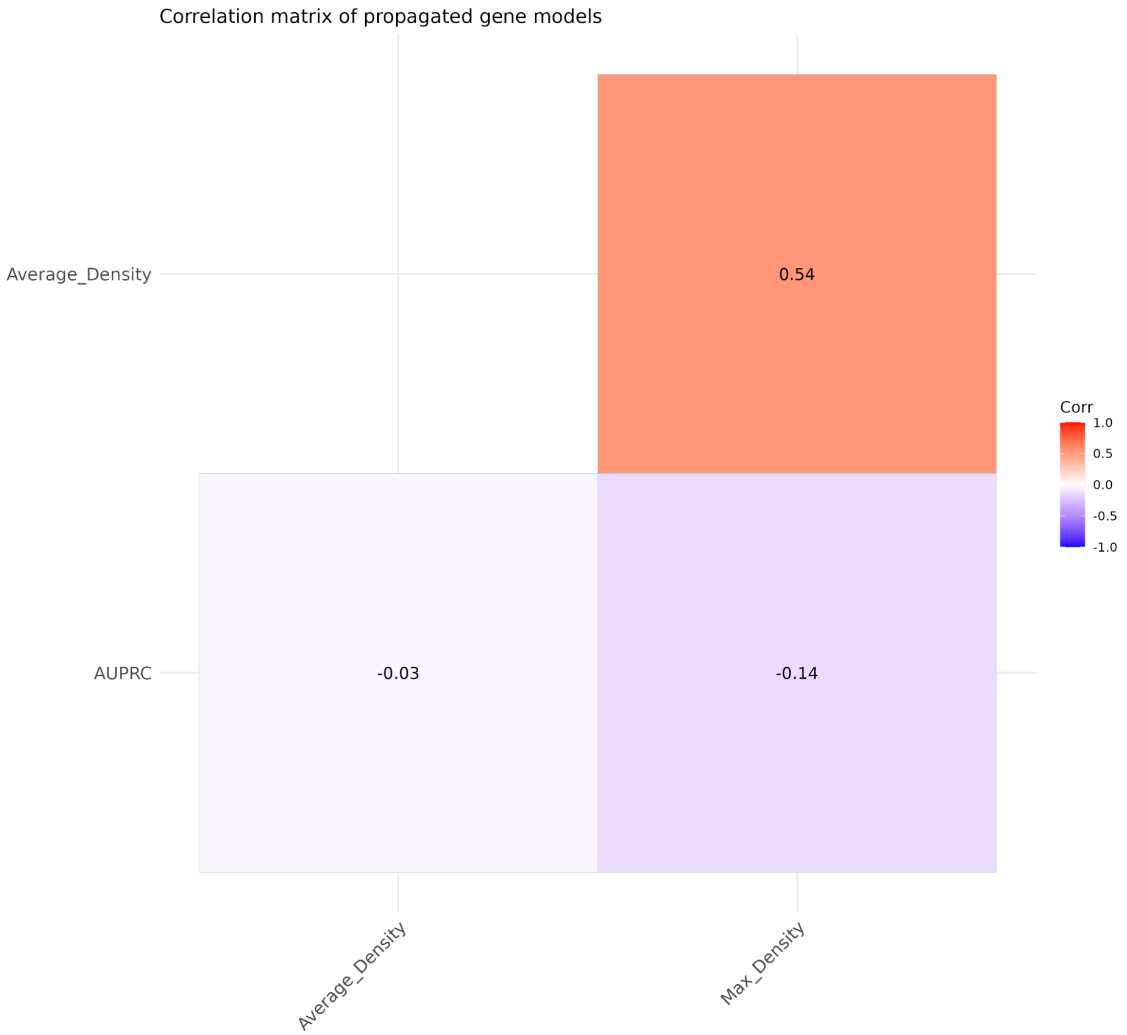

Figure 3.38: Correlation matrix of performance and average/max cluster edge density statistics for propagated genesets.

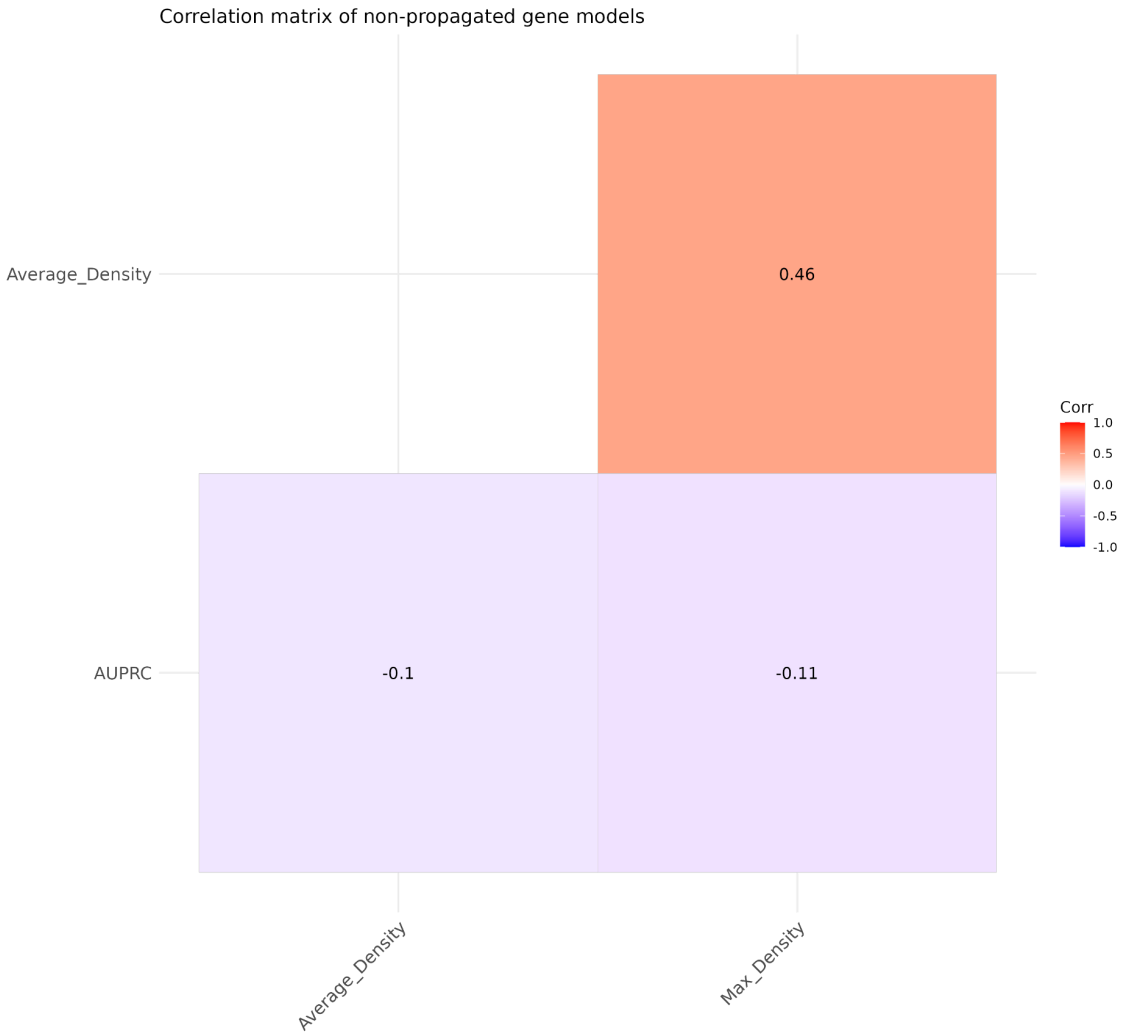

Figure 3.39: Correlation matrix of performance and average/max cluster edge density statistics for non-propagated genesets.

Modules contain unique and specific GOBP information

To understand how ModGenePlexus can allow biologists to better understand their gene list of interest, we performed a case study for type 2 diabetes. This is a list of differentially expressed genes from the CREEDS database and contains 3,837 genes. The gene set was run in the ModGenePlexus pipeline which removed 2,584 genes and resulted in 14 clusters, ranging in size from 54 to 152 genes (**Figure 3.40)** after DOMINO propagation. We additionally ran GenePlexus on the full set in AllAssign. Both AllAssign and the cluster gene lists were passed into the enrichment software in R package ClusterProfiler[34] (**see methods**). We ran enrichment with GOBPs, looking at terms with at least 10 genes and a max size of 500 genes.

A key feature of ModGenePlexus is that in addition to improving gene classification on large, noisy gene sets, the gene level predictions are attached to a specific cluster before the aggregation step. Doing enrichment at the cluster level can give a more refined view of the biology behind the initial experimental gene set. To determine if more specific or more general biology is being discovered, we looked at the number of genes annotated to the GOBPs that are enriched. This assumption is based on the ontology being propagated up the ontology, where more general ontology go terms contain the gene annotations of any term below it in the ontology. Aside from cluster 10, AllAssign enriches for larger GO terms, which are substantially larger when compared to most of the clusters (**Figure 3.41).** Additionally, since both the enriched GOBPs and the cluster sizes are smaller, the statistical power of the enrichment method is increased and gives more significant q-values than AllAssign (**Figure 3.42)**. For AllAssign we showed in **Figure 3.43** that the percentage of the AllAssign gene list inside of enriched GOBPs is quite small in a given GOBP. For this analysis, we see that cluster 10 has a large percentage of its genes in the enriched GOBP terms, which also explains why cluster 10 enriches for even larger genesets tha AllAssign in **Figure 3.41**.

One concern with ModGenePlexus is that it can drop many genes during the DOMINO module discovery and propagation process. We demonstrate in this chapter that this step does improve gene classification results and justify that the genes being lost have poor network connections, but losing this many genes could mean a lot of potentially important biology is being lost. To demonstrate if this concern is valid, we counted how many enriched GOBPs are found in AllAssign and across the clusters (**Figure 3.44**). For AllAssign, the number of GOBPs enriched in the gene list were simply counted as is. For ModGenePlexus, we collected the enriched GOBPs found across all clusters and counted the number of unique terms. Interestingly, the number of terms enriched in ModGenePlexus is greater than for AllAssign, 331 to 222, respectively. In addition there is an overlap between the two sets of 79, which is around 35% of the AllAssign enriched terms. We then tested if the enriched terms in both methods were capturing biology at different scales. This was done by compiling the set sizes of the enriched terms unique to both AllAssign and ModGenePlexus, as well as the terms enriched in both (**Figure 3.45**). We see that the largest GOBPs on average are found by both methods. This is

followed closely by the terms found only by AllAssign, and the terms found by ModGenePlexus were substantially smaller. This suggests that ModGenePlexus finds more specific terms and biology. In sum, this demonstrates that ModGenePlexus can allow researchers to understand their geneset of interest in a more nuanced way, allowing for further studies at the molecular level.

Specific Biology of Type-2 Diabetes is unraveled with ModGenePlexus

The analyses done indicate that for general metrics of enrichment results that ModGenePlexus may implicate more specific biology for diseases. To provide further evidence, we take the top 20 enriched terms for AllAssign of type-2 diabetes (**Figure 3.46**). There were 222 total enriched terms in AllAssign, and in the top 20 there are 2 relatively specific terms in muscle filament sliding and actin-myosin filament sliding. The other terms are quite general, including neutrophil activation, muscle system process, and response to oxygen levels. Looking at two clusters, 5 and 3, we see that they are enriched for smaller GOBP and see the more specific biology that is recovered. In cluster 5, 19 total terms were enriched with only 2 also being enriched in AllAssign, where these 2 are the largest terms in cluster 5 (**Figure 3.47**). It is enriched for cellular processes that relate to copper ions[47–50] and the mitochondrial membrane[51–54], and both have been enriched for type-2 diabetes. In cluster 3, there were a total of 43 enriched terms of which 4 are also enriched in AllAssign and are very general (proton transmembrane transport, ATP metabolic processes, macroautophagy, and mitochondrial transport). Cluster 3 has biology known to be related to type-2 diabetes (**Figure 3.48**). Response to insulin[38–41] and iron/transferring processes[42–45] are well documented to be associated with diabetes. Interestingly, many enriched terms are also related to cytidine processes. A literature search only found one recent study[46] that showed using whole blood metabolomics that cytidine could be a marker for type-2 diabetes, suggesting that the cytidine process in type-2 diabetes might be important yet understudied.

In cluster 10, we saw enrichment for larger genesets, and it recovers a number of processes also seen in AllAssign (**Figure 3.49**). Six out of eleven terms enriched in cluster 10 are enriched in AllAssign. However, the terms are generally more enriched in

cluster 10, an example being translational initiation, which has a q-value of 1.13e-30 in AllAssign, but a much lower value of 1.27e-99 in cluster 10.

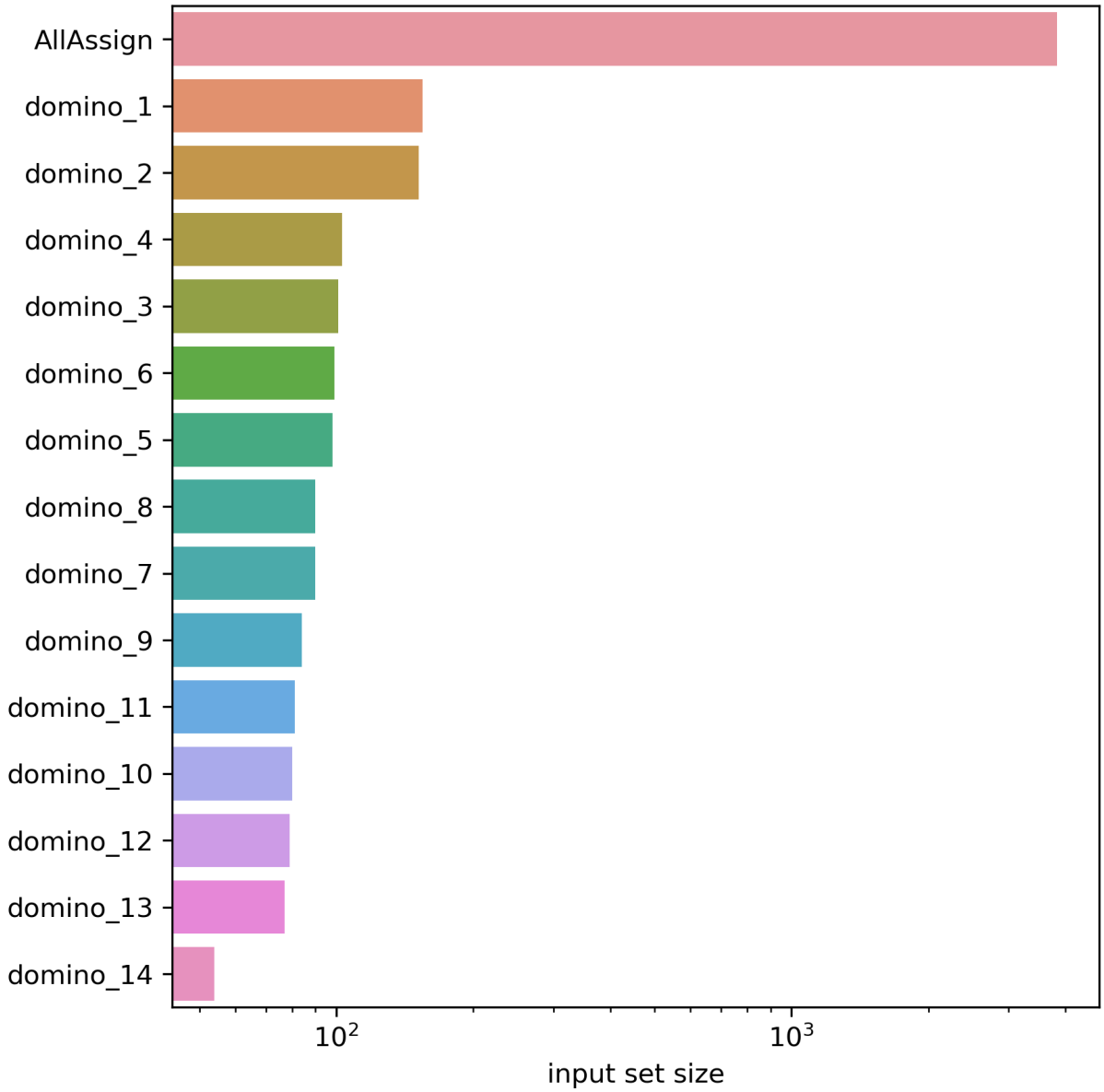

Figure 3.40: The number of genes in each propagated domino cluster and in AllAssign. These propagated assignments were used for enrichment analysis as inputs.

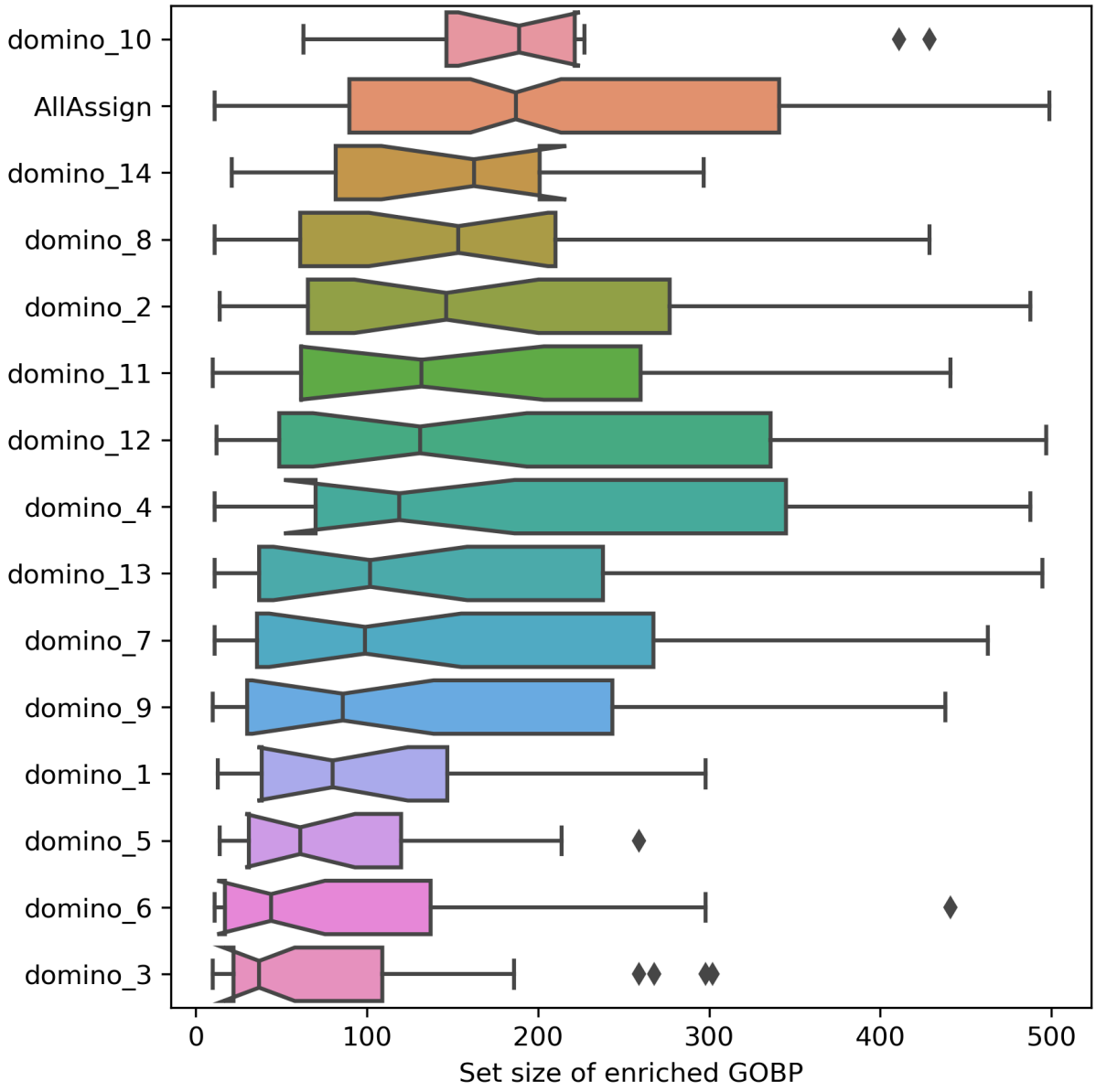

Figure 3.41: The distribution of enriched GOBP term sizes for each cluster and AllAssign.

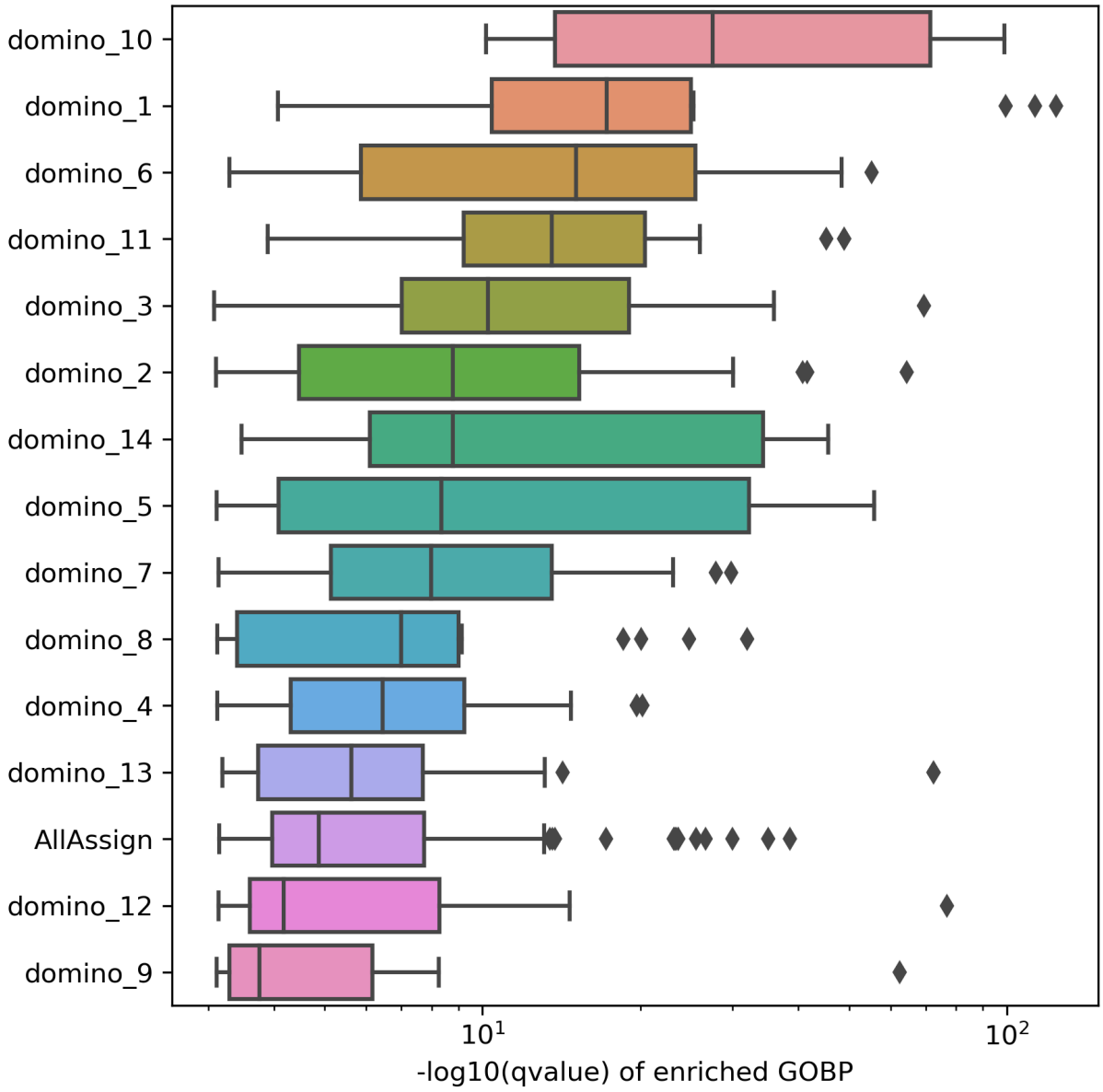

Figure 3.42: The -log10(q-value) for enriched GOBP in each cluster and AllAssign.

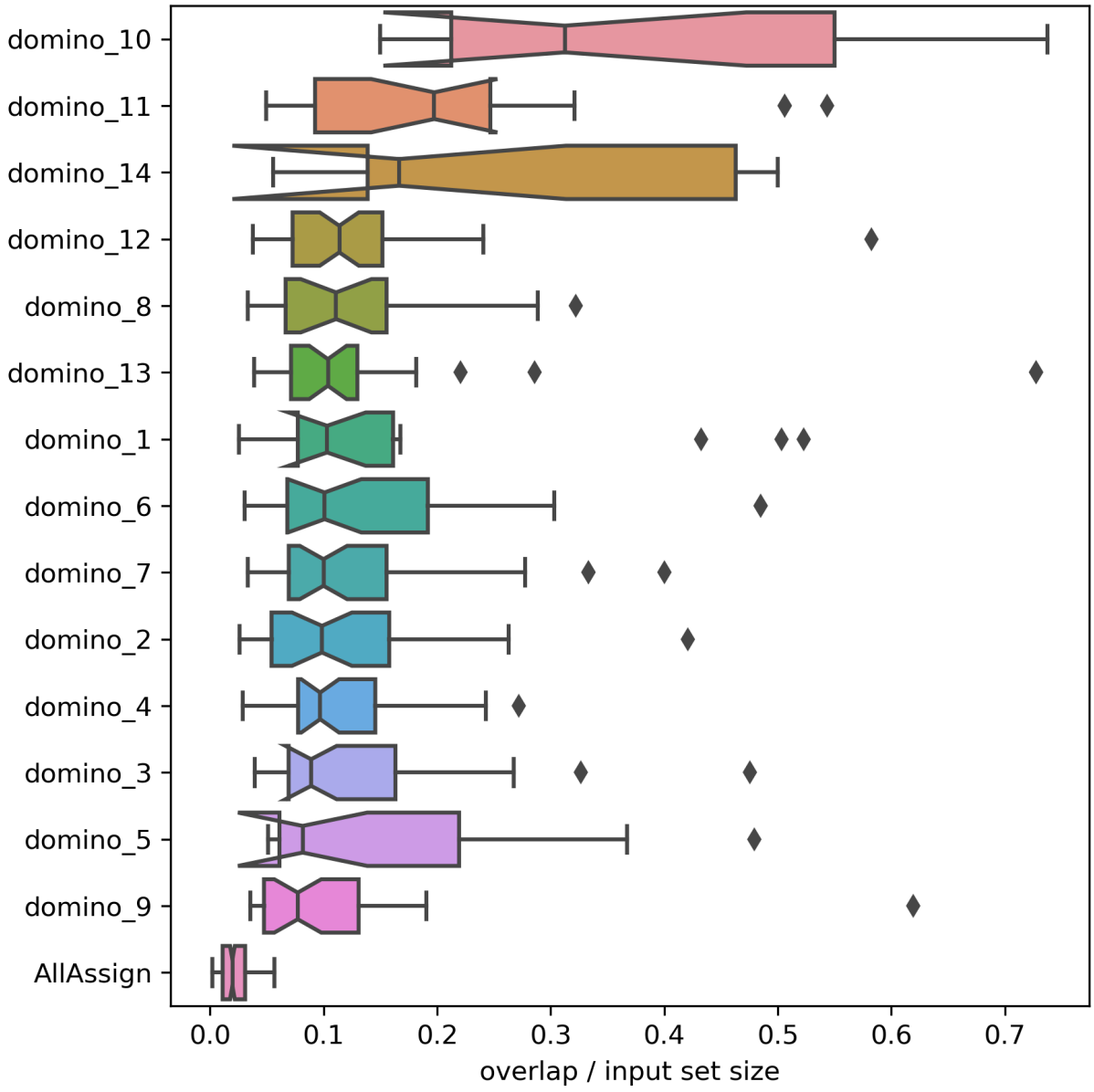

Figure 3.43: Distribution of the percent of the input gene list that is in enriched GOBPs for each cluster and AllAssign.

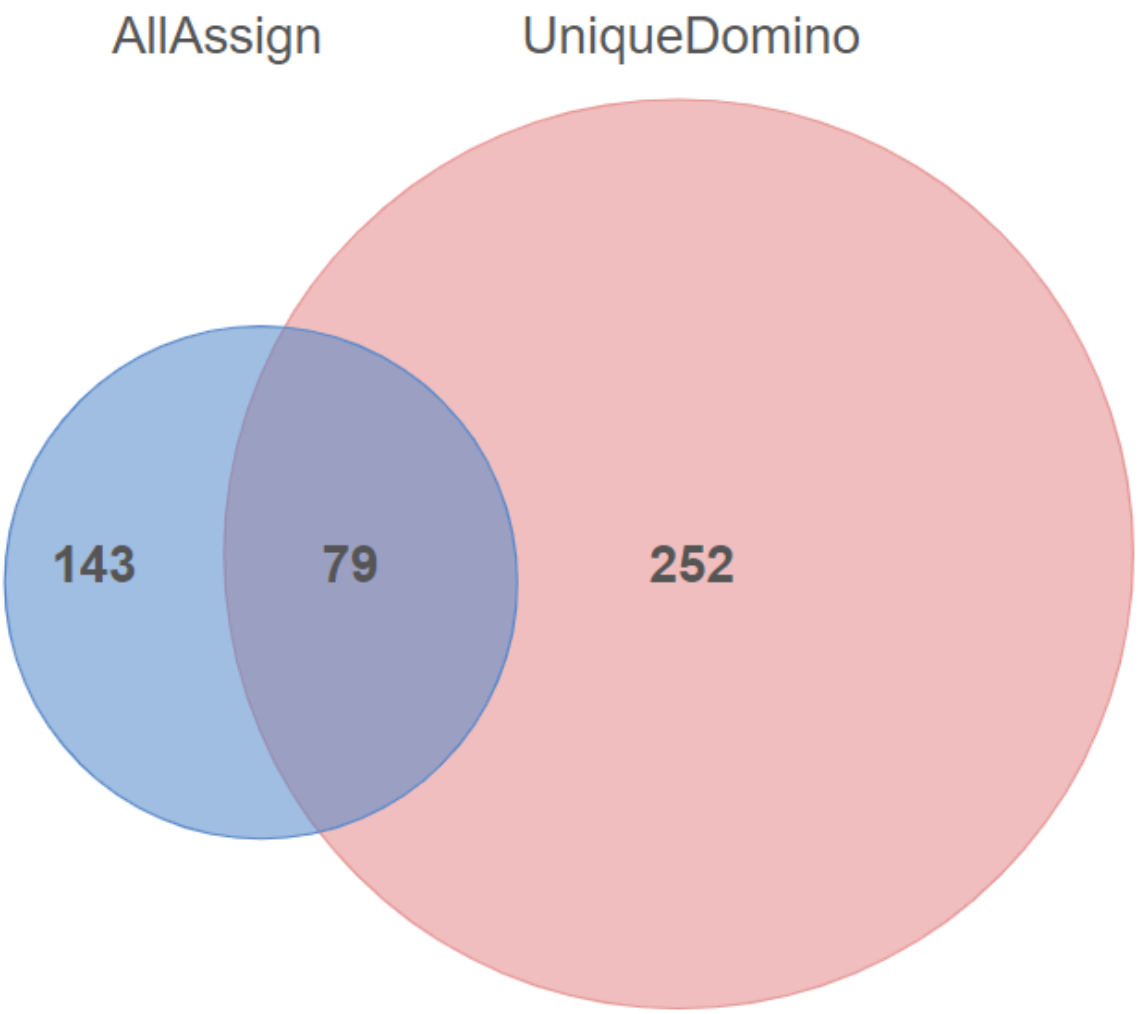

Figure 3.44: Counting the GOBPs that are enriched in the clusters and AllAssign, and the overlap between them. UniqueDomino refers to unique GOBPS enriched across all propagated clusters.

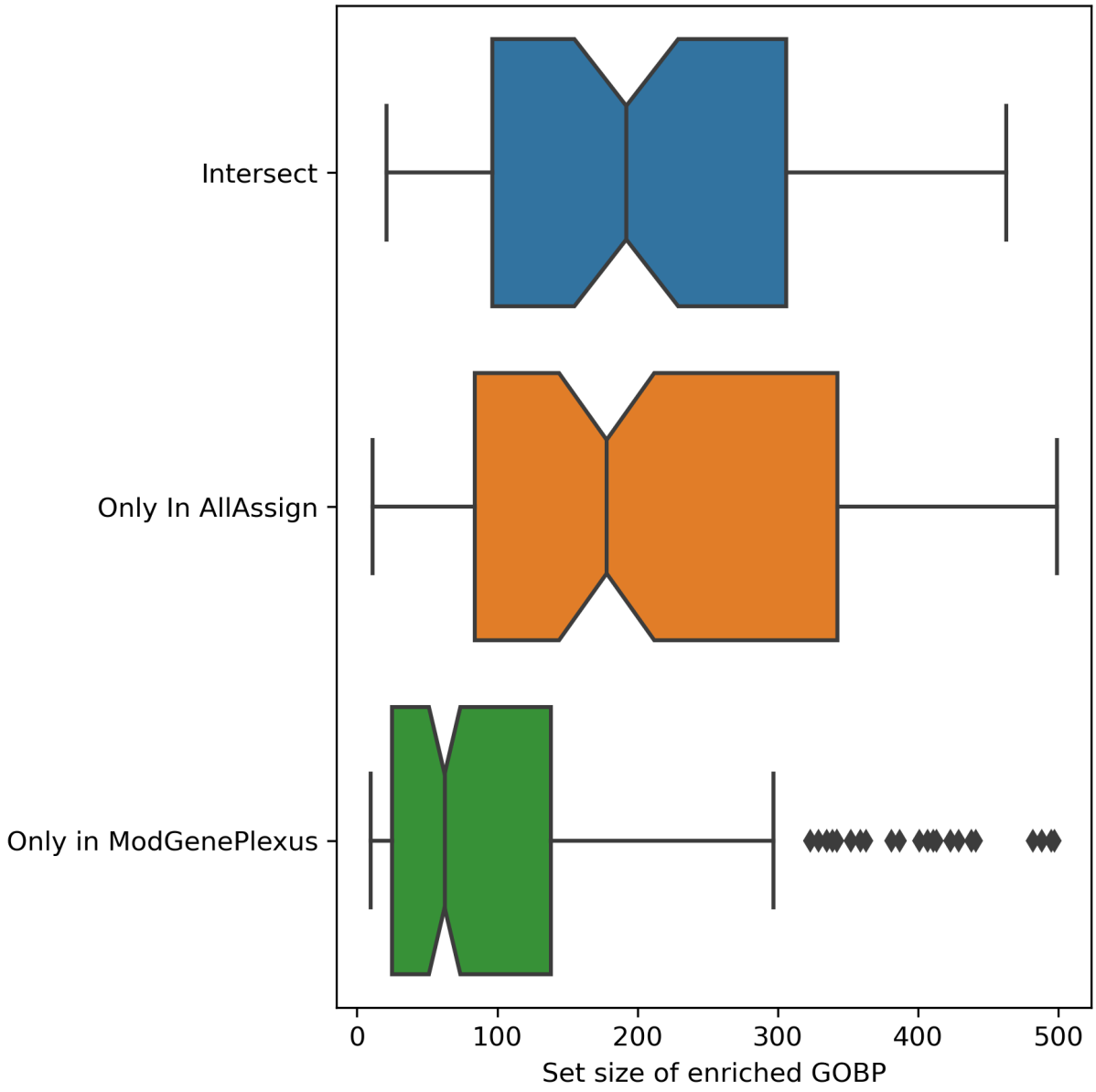

Figure 3.45: The number of genes annotated to each enriched GOBP in ModGenePlexus, AllAssign, and in both methods (intersect).

Top Enriched GOBPs Using AllAssign					
InputSet	Rank	Description	qvalue	Count	TermSize
AllAssign	1	cotranslational protein targeting to membrane	2.93e-39	83	92
AllAssign	2	energy derivation by oxidation of organic compounds	9.26e-36	160	267
AllAssign	3	translational initiation	1.13e-30	117	180
AllAssign	4	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	2.47e-27	83	113
AllAssign	5	muscle system process	2.78e-26	201	424
AllAssign	6	neutrophil degranulation	2.35e-24	214	475
AllAssign	7	ATP metabolic process	2.55e-24	153	298
AllAssign	8	neutrophil activation	3.74e-24	218	489
AllAssign	9	neutrophil activation involved in immune response	5.28e-24	214	478
AllAssign	10	neutrophil mediated immunity	6.75e-24	217	488
AllAssign	11	viral gene expression	6.31e-18	101	187
AllAssign	12	response to oxygen levels	1.76e-14	159	379
AllAssign	13	response to metal ion	2.46e-14	150	352
AllAssign	14	response to oxidative stress	3.55e-14	176	436
AllAssign	15	muscle filament sliding	7.85e-14	33	39
AllAssign	16	actin-myosin filament sliding	7.85e-14	33	39
AllAssign	17	steroid metabolic process	1.62e-13	135	312
AllAssign	18	monosaccharide metabolic process	3.70e-13	126	287
AllAssign	19	monosaccharide biosynthetic process	1.25e-12	57	95
AllAssign	20	response to peptide hormone	2.08e-12	168	427

Figure 3.46: Top 20 enriched GOBP terms in AllAssign. The total number of enriched terms is 222, but we visualize only the top 20 here. Significance is shown in q-value, Count is the intersection of the input gene list and GOBP, and TermSize is the number of GOBP-gene annotations. The total number of genes in type-2 diabetes for AllAssign is 3,837.

Top Enriched GOBPs for Cluster 5						
Rank	Description	qvalue	Count	TermSize	AllAssignRank	AAQvalues
1	mitochondrial transport	1.38e-56	47	259	175	1.47e-04
2	protein targeting to mitochondrion	8.64e-53	35	97		
3	establishment of protein localization to mitochondrion	1.24e-48	36	138		
4	mitochondrial transmembrane transport	1.07e-34	26	98		
5	inner mitochondrial membrane organization	1.10e-34	22	51		
6	protein transmembrane transport	3.79e-31	21	58		
7	cellular copper ion homeostasis	3.47e-11	7	14		
8	copper ion homeostasis	1.82e-10	7	17		
9	copper ion transport	2.78e-10	7	18		
10	protein insertion into membrane	4.46e-09	9	63		
11	mitochondrial respiratory chain complex assembly	1.62e-08	10	102		
12	ATP transport	1.72e-05	5	28		
13	purine ribonucleotide transport	2.62e-05	5	31		
14	adenine nucleotide transport	2.62e-05	5	31		
15	cellular respiration	2.57e-04	8	180		
16	ATP biosynthetic process	2.74e-04	5	51		
17	oxidative phosphorylation	4.74e-04	7	143		
18	purine ribonucleoside triphosphate biosynthetic process	6.26e-04	5	61		
19	protein folding	7.79e-04	8	214	81	2.07e-06

Figure 3.47: Top 20 enriched GOBP terms in Cluster 5. Significance of the term in cluster 5 is shown in q-value, Count is the intersection of the input gene list and GOBP, and TermSize is the number of GOBP-gene annotations. AAQvalues is the q-value in AllAssign, where if the cell is blank then the term was not enriched in AllAssign. 98 genes were used as input for enrichment.

Top Enriched GOBPs for Cluster 3						
Rank	Description	qvalue	Count	TermSize	AllAssignRank	AAQvalues
1	proton transmembrane transport	3.53e-70	48	151	85	2.61e-06
2	phagosome acidification	1.09e-36	20	28		
3	transferrin transport	6.17e-36	21	36		
4	pH reduction	9.31e-35	23	57		
5	oxidative phosphorylation	3.89e-31	27	143		
6	ATP metabolic process	9.48e-31	33	298	7	2.55e-24
7	phagosome maturation	1.34e-30	20	48		
8	iron ion transport	7.31e-28	21	76		
9	aerobic electron transport chain	9.02e-27	14	18		
10	insulin receptor signaling pathway	1.06e-25	23	133		
11	response to insulin	6.43e-20	24	268		
12	cytidine to uridine editing	1.48e-19	10	12		
13	cytidine catabolic process	6.76e-17	9	12		
14	cytidine deamination	6.76e-17	9	12		
15	cytidine metabolic process	6.76e-17	9	12		
16	respiratory chain complex IV assembly	6.76e-17	11	26		
17	DNA cytosine deamination	1.79e-15	8	10		
18	regulation of macroautophagy	1.96e-13	16	174		
19	pyrimidine-containing compound catabolic process	1.91e-11	9	37		
20	mitochondrial ATP synthesis coupled proton transport	2.10e-11	8	24		

Figure 3.48: Top 20 enriched GOBP terms in Cluster 3. Significance of the term in cluster 3 is shown in q-value, Count is the intersection of the input gene list and GOBP, and TermSize is the number of GOBP-gene annotations. AAQvalues is the q-value in AllAssign, where if the cell is blank then the term was not enriched in AllAssign. 101 genes were used as input for enrichment.

Top Enriched GOBPs for Cluster 10						
Rank	Description	qvalue	Count	TermSize	AllAssignRank	AAQvalues
1	translational initiation	1.27e-99	59	180	3	1.13e-30
2	SRP-dependent cotranslational protein targeting to membrane	1.15e-78	43	88		
3	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	7.85e-78	45	113	4	2.47e-27
4	viral gene expression	1.77e-66	45	187	11	6.31e-18
5	ribonucleoprotein complex biogenesis	4.26e-33	35	429		
6	ribonucleoprotein complex assembly	3.19e-28	25	189	150	6.77e-05
7	ribonucleoprotein complex subunit organization	7.79e-28	25	196	155	7.03e-05
8	ribosome assembly	3.62e-15	12	63	185	1.93e-04
9	regulation of cellular amide metabolic process	9.30e-14	20	411		
10	rRNA processing	3.74e-11	14	216		
11	rRNA metabolic process	7.09e-11	14	227		

Figure 3.49: Top 20 enriched GOBP terms in Cluster 10. Significance of the term in cluster 10 is shown in q-value, Count is the intersection of the input gene list and GOBP, and TermSize is the number of GOBP-gene annotations. AAQvalues is the q-value in AllAssign, where if the cell is blank then the term was not enriched in AllAssign. 80 genes were used as input for enrichment.

Discussion

Human diseases can be very complex and could have hundreds of true gene annotations, making it common for experiments like GWAS and differential expression to implicate numerous loci or genes. The goal is to interpret these results biologically, explaining how genes or loci work, relate with one another, or relate to a disease. A common pipeline is to take experimental results and visualize them in a network using software. A typical pipeline involves visualizing experimental results in a network using software, identifying modules within the network, and performing enrichment analysis for biological processes or phenotypes. Researchers then select interesting modules for further study. This standard pipeline raises questions about module definition and quality, which are determined through the relevance and quality of genes implicated in the initial experiment. Our project aimed to improve input in standard computational and bioinformatics methods by improving and expanding GenePlexus in a general-purpose framework that works across multiple experiments for diverse diseases and traits..

GenePlexus is a method aimed to recover false negatives and add genes to genesets that relate to biological entities like diseases, but does nothing on its own to remove false positive genes. ModGenePlexus is a powerful method because it recognizes that not only are genesets incomplete, but that genesets are filled with either false positives, noise, or genes that are so understudied as to not have confident association to many other genes in the genome. We integrated multiple well validated methods – DOMINO, propagation, GenePlexus/supervised learning, and genome-wide networks – for geneset refinement, keeping genes related within external biological data and networks. ModGenePlexus and using gene modules improves gene classification in multiple unique and distinct ways. For example, training on loose threshold MAGMA genes to predict highly significant but understudied genes gave good results with ModGenePlexus, but very poor results with GenePlexus. Rather than using a stringent threshold to decide what GWAS loci or genes to use, using networks to give biological context to determine genes with biological meaning is a powerful tool to predict and classify genes significant and thus more likely to be mechanistically relevant.

Working around gold standard limitations and module definitions

An open question is how to define modules in a robust way. As there is no good gold standard that can be used to confidently define modules for diseases of interest, our goal was not to determine final gene modules underlying a disease, but show that if modules are defined simply as a subset of biologically related genes, they can be used to make better predictions for the disease disease holistically. We demonstrated extensively that gene classification is improved for a variety of datasets in ModGenePlexus under different conditions. Additionally, running GenePlexus for individual modules also uncovers unique, specific processes not found when enriching the whole disease. We argue that the importance of these enrichments is not the module is a final product in itself, but that the modules can be used as tools to gain insight into novel biological insights for disease.

The simulated traits shows ModGenePlexus matches GenePlexus performance

In this project, we began by creating a study where if GOBPs are treated as modules, and if we combine distinct GOBPs into simulated traits, then ModGenePlexus would recreate performance of running GenePlexus on the large geneset (**Figure 3.7**). This

result was initially surprising as we expected that if a trait had multiple distinct modules, that this would substantially hurt GenePlexus performance, and that ModGenePlexus would show noticeable improvement. However, GenePlexus proved to be a robust method. The highly multidimensional nature of the network allows the linear regression model used by GenePlexus to consider boundaries effectively even when there are multiple distinct modules in the network. Therefore, we applied module information to improve performance for particular genesets or in other aspects of the pipeline, and we presented many ways where ModGenePlexus is a general improvement for large-scale experimental genesets and gives meaningful and novel biological associations..

Using module assignments improves negative gene discovery

The robustness of GenePlexus motivated us to consider other ways to use modules to improve model performance by improving the neutral and negative gene label selection process. Interpreting large complex diseases as collections of meaningful gene subsets, rather than one large set where all genes are equal, raises an issue of doing geneset overlap with other meaningful sets. In a network, possible disease genes are not necessarily connected to every single meaningful subset within a disease. Genes that are connected to a single module in a disease, but are not connected in any significant way to other modules, could be given a negative label because from the perspective the "disease as a whole" (all disease genes), it does not have dense connections to most of them and is classified as a negative. We argue that if this gene is well connected to a module of the disease, that negative label is inappropriate, and using modules to determine neutral labels prevents this mislabeling. This method increases runtime, but **Figure 3.12** shows that it dramatically improves results of the simulation for all model types. Notably, the more modules a simulated trait had, the better the new method performed, which is relevant since real-world diseases have many underlying modules. This approach enhances GenePlexus by increasing the statistical power of hypergeometric tests to refine negative label determination to better leverage the biological data in GO and DisGeNET.

ModGenePlexus allows for better gene classification of experimental data

The discussed simulation has an advantage in that the GOBP "modules" underneath the traits do not have many, if any, false positives that need to be removed. This

motivated our focus on experimental datasets, which were not used in GenePlexus due to their large size and known false positives. GenePlexus genesets were originally chosen for their good annotation, in sharp contrast to the realities of initial experimental results. We chose DOMINOfor module discovery not just because it is a good and benchmarked module discovery method, but because it finds more genes using network connections and removes genes with minimal network relationships. Using ModGenePlexus, we show that the ability to remove “bad” genes significantly improves GenePlexus. **Figure 3.19** shows that only the propagated version ModGenePlexus outperforms propagated AllClus – meaning AllClus is also a direct improvement over GenePlexus. **Figure 3.20** shows that using the recovered false negative genes through propagation improves GenePlexus results further. Both removing false positives and recovering some false negatives heavily improves GenePlexus results and both additions are essential to have consistently better results across geneset collections of experimental data. The more high quality gene annotations, the better GenePlexus can do in classifying genes. We implemented and combined multiple computational methods to refine genesets that outperforms GenePlexus in multiple ways.

Training on nominally associated genes better predicts more significant genes

As mentioned previously, the mixed results for ModGenePlexus improving GWAS motivated us to train on looser threshold genes for stringent data. We saw that ModGenePlexus improves the loose threshold genesets, and we show in **Figure 3.29** that bigger traits are likely to have denser clusters discovered. Training on only highly stringent results may mean we get poor clusters, which would explain why ModGenePlexus does poorly. In using the looser threshold genes, we see noticeable improvement in those GWAS that performed poorly in GenePlexus **(Figure 3.24-25**). Ultimately, ModGenePlexus can recreate performance when GenePlexus does well, but improves GenePlexus when it fails. Interestingly, loosening the threshold to an extremely loose degree continues to improve ModGenePlexus results but worsens them for GenePlexus. The improvement of ModGenePlexus has a trade-off – it better recovers truly relevant genes, but given the large number of genes added it is quite likely that the modules become quite big and possibly broad in terms of biological interpretation. This raises the question of the utility of big modules versus small

modules. Small modules are more useful in terms of looking at enrichment results, but we argue that this is evidence for big modules being useful for computational methods. The goal should be to use modules to improve disease interpretation, as such we argue that using large genesets from GWAS analysis can be useful depending on the goal of the post-GWAS computational methods being employed. We provide evidence that for gene classification of mechanistically important genes, large modules are a useful tool when they are coherent in the network.

Edge densities of modules and performance goes against expectations of the original GenePlexus study

When analyzing the edge densities of the clusters and inputs used for ModGenePlexus, it was initially surprising to see that the propagated clusters are less dense than non-propagated ones. There are multiple reasons why this could be happening. One explanation is that DOMINO is not using edge density to build its modules, but is making sure the modularity of the subgraph is higher than the ratio of genes in the subgraph relative to the entire network. Modularity is associated with edge density, but genesets can be modular and non-densely connected in the network. The question arises why utilizing these propagated genes significantly improves results of ModGenePlexus. A likely reason is because the propagated genes have network relationships to the understudied, held out genes of the geneset. Understudied genes are those that are less likely to have robust or ubiquitous network connections, but should in actuality be connected to at least some of the disease genes. They are not connected, however, due to incomplete information about either the gene or within the network itself. This reasoning is likely valid because the very reason the understudied genes are discovered in the first place is because in some way they are disease-associated, and to be disease-associated they must work together with other disease genes. In other words, using DOMINO to “fill in” modules with other non-disease genes with direct network connection may make the module less dense, but utilizing less dense network connections is what better recovers understudied genes that are also not densely connected to the gene modules. The relationship between edge density and geneset size is also interesting, particularly where larger genesets tend to have higher cluster densities on average. Finding these higher density clusters

was a major motivation for this project, and we can see that the biggest genesets are the ones with the most drastic improvements in ModGenePlexus. These results and the original evaluations in GenePlexus show that while edge density is a useful statistic to improve GenePlexus results, other genes in the network are truly relevant and additional methods such as ModGenePlexus are needed to find genes in spite of their less dense connections to most disease genes.

ModGenePlexus finds unique GOBP predictions compared to GenePlexus

Lastly, we show off for a specific disease unique gene predictions and enrichment results. Doing enrichment at a module level allows for increased statistical power for common computational enrichment methods. Large, whole diseases are composed of numerous processes and phenotypes. Those processes that are most likely to be at the top of enrichment results are broad, ubiquitous processes. This is because they often involve large numbers of genes, with many of these genes appearing across disease lists. In addition, specific process gene lists are so small that when a disease has many hundreds of genes, it is extremely difficult to get a significant enrichment result. With modules, because they are a small subset of the disease, it is possible to find significant enrichment between modules and more specific processes. We demonstrated that using ModGenePlexus removes genes in the propagation process, but this actually finds additional biology when doing enrichment at the cluster level. ModGenePlexus is able to find significant biology in the clusters because it utilizes a network created from biological data to determine meaningful subsets, in this case the STRING network. STRING integrates vast amounts of data, ranging from physical interactions, co-expression data, annotation databases, to build the network. This is why the clusters discovered have meaningful biology, and we leverage this network to improve gene classification of a researcher's genelist of interest.

Specific biology of type-2 diabetes revealed interesting pathways in the modules

We demonstrated that the modules created from type-2 diabetes differential expression results led to interesting biological enrichment relative to the gene list as a whole. An interesting discovery was cluster 10, which we saw had different properties compared to other clusters in **Figures 3.41-43**. It was enriched for many processes that were enriched in AllAssign. It seems that this cluster was a fairly general cluster, having

biology that is enriched with many of the more general GOBP terms. Clusters 3 and 5 had specific processes that are relevant to type-2 diabetes that were not discovered by AllAssign. ModGenePlexus discovered both general and specific processes that are enriched for disease genes. How the implicated general and specific processes affect one another in the context of disease manifestation, patient phenotypes and outcome, and molecular pathways are questions that need to be answered in the context of each specific complex disease.

Conclusion: ModGenePlexus gives GenePlexus direct application in hypothesis generation for experiments

Overall, we demonstrate that ModGenePlexus is a superior method to GenePlexus for classifying the genes of experimental results. This means modules are a computational tool to improve interpretation of diseases as a whole, and this information is what will allow for better targeting of mechanistically important genes discussed in chapter 4.

Future Applications

An important goal is to get ModGenePlexus working in the current PyGenePlexus python implementation and Webserver. This runs into multiple practical and technical issues due to the additional memory and computational resources needed. For use by a researcher, it's important to predict whether ModGenePlexus or GenePlexus should be used. ModGenePlexus performs better for most genesets we tested in this project, but not for all – even within the same domain. Having a quantitative way to predict which method should be used is an important task to make the software usable on the server.

REFERENCES

1. Liu, R., Mancuso, C. A., Yannakopoulos, A., Johnson, K. A. & Krishnan, A. Supervised-learning is an accurate method for network-based gene classification. *Bioinformatics* (2020) doi:10.1093/bioinformatics/btaa150.

2. Lee, I., Blom, U. M., Wang, P. I., Shim, J. E. & Marcotte, E. M. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.* **21**, 1109–1121 (2011).

3. Krishnan, A. *et al.* Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat Neurosci* **19**, 1454–1462 (2016).

4. Wang, X., Gulbahce, N. & Yu, H. Network-based methods for human disease gene prediction. *Brief Funct Genomics* **10**, 280–293 (2011).

5. Mancuso, C. A., Johnson, K. A., Liu, R. & Krishnan, A. Joint representation of molecular networks from multiple species improves gene classification. *PLoS Comput Biol* **20**, e1011773 (2024).

6. Leiserson, M. *et al.* Pan-Cancer Network Analysis Identifies Combinations of Rare Somatic Mutations across Pathways and Protein Complexes. *Nature Genetics* **47**, 106–114 (2015).

7. Guan, Y., Ackert-Bicknell, C. L., Kell, B., Troyanskaya, O. G. & Hibbs, M. A. Functional Genomics Complements Quantitative Genetics in Identifying Disease-Gene Associations. *PLOS Computational Biology* **6**, e1000991 (2010).

8. Warde-Farley, D. *et al.* The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res* **38**, W214–W220 (2010).

9. Šubelj, L. Label Propagation for Clustering. in *Advances in Network Clustering and Blockmodeling* (eds. Doreian, P., Batagelj, V. & Ferligoj, A.) 121–150 (Wiley, 2019). doi:10.1002/9781119483298.ch5.

10. Mancuso, C. A., Liu, R. & Krishnan, A. PyGenePlexus: A Python package for gene discovery using network-based machine learning. 2022.07.02.498552 Preprint at https://doi.org/10.1101/2022.07.02.498552 (2022).

11. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research* **48**, D845–D855 (2020).

12. Piñero, J. *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res* **45**, D833–D839 (2017).

13. Piñero, J. *et al.* DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database (Oxford)* **2015**, (2015).

14.The Gene Ontology Consortium *et al.* The Gene Ontology knowledgebase in 2023. *GENETICS* **224**, iyad031 (2023).

15.Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).

16.Law, M. & Shaw, D. R. Mouse Genome Informatics (MGI) Is the International Resource for Information on the Laboratory Mouse. in *Eukaryotic Genomic Databases* (ed. Kollmar, M.) vol. 1757 141–161 (Springer New York, New York, NY, 2018).

17.Gustafsson, M. *et al.* Modules, networks and systems medicine for understanding disease and aiding diagnosis. *Genome Med* **6**, 82 (2014).

18.Barrett, T. *et al.* NCBI GEO: Archive for Functional Genomics Data Sets—Update. *Nucleic Acids Research* **41**, 991–95 (2013).

19.Szklarczyk, D. *et al.* STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* **43**, D447–D452 (2015).

20.Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Research* **34**, D535–D539 (2006).

21.Oughtred, R. *et al.* The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci* **30**, 187–200 (2021).

22.Wang, Z. *et al.* Extraction and Analysis of Signatures from the Gene Expression Omnibus by the Crowd. *Nature Communications* **7**, 12846 (2016).

23.Leeuw, C. A. de, Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Computational Biology* **11**, e1004219 (2015).

24.Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat Genet* **51**, 1339–1348 (2019).

25.Chen, Z., Boehnke, M., Wen, X. & Mukherjee, B. Revisiting the genome-wide significance threshold for common variant GWAS. *G3 Genes|Genomes|Genetics* **11**, jkaa056 (2021).

26.Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* **10**, e0118432 (2015).

27.Davis, J. & Goadrich, M. The relationship between Precision-Recall and ROC curves. in *Proceedings of the 23rd international conference on Machine learning - ICML '06* 233–240 (ACM Press, Pittsburgh, Pennsylvania, 2006). doi:10.1145/1143844.1143874.

28.Levi, H., Elkon, R. & Shamir, R. *DOMINO: A Novel Algorithm for Network-Based*

Identification of Active Modules with Reduced Rate of False Calls. http://biorxiv.org/lookup/doi/10.1101/2020.03.10.984963 (2020) doi:10.1101/2020.03.10.984963.

29. Levi, H., Elkon, R. & Shamir, R. DOMINO: a network‐based active module identification algorithm with reduced rate of false calls. *Molecular Systems Biology* **17**, e9593 (2021).

30. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, P10008 (2008).

31. Johnson, D. S., Minkoff, M. & Phillips, S. The prize collecting Steiner tree problem: theory and practice. *SODA '00: Proceedings of the eleventh annual ACM-SIAM symposium on Discrete algorithms* 760–769 (2000).

32. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 7821–7826 (2002).

33. Mancuso, C. A., Liu, R. & Krishnan, A. PyGenePlexus: a Python package for gene discovery using network-based machine learning. *Bioinformatics* **39**, btad064 (2023).

34. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).

35. The DREAM Module Identification Challenge Consortium *et al.* Assessment of network module identification across complex diseases. *Nat Methods* **16**, 843–852 (2019).

36. Raser, J. M. & O'Shea, E. K. Noise in gene expression: origins, consequences, and control. *Science* **309**, 2010–2013 (2005).

37. Eling, N., Morgan, M. D. & Marioni, J. C. Challenges in measuring and understanding biological noise. *Nat Rev Genet* **20**, 536–548 (2019).

38. Boucher, J., Kleinridders, A. & Kahn, C. R. Insulin receptor signaling in normal and insulin-resistant states. *Cold Spring Harb Perspect Biol* **6**, a009191 (2014).

39. Brännmark, C. *et al.* Insulin signaling in type 2 diabetes: experimental and modeling analyses reveal mechanisms of insulin resistance in human adipocytes. *J Biol Chem* **288**, 9867–9880 (2013).

40. Dall'Agnese, A. *et al.* The dynamic clustering of insulin receptor underlies its signaling and is disrupted in insulin resistance. *Nat Commun* **13**, 7522 (2022).

41. Batista, T. M., Haider, N. & Kahn, C. R. Defining the underlying defect in insulin action in type 2 diabetes. *Diabetologia* **64**, 994–1006 (2021).

42. Feng, J. *et al.* Association of Body Iron Metabolism with Type 2 Diabetes Mellitus in Chinese Women of Childbearing Age: Results from the China Adult Chronic Disease and Nutrition Surveillance (2015). *Nutrients* **15**, 1935 (2023).

43.Miao, R. *et al.* Iron metabolism and ferroptosis in type 2 diabetes mellitus and complications: mechanisms and therapeutic opportunities. *Cell Death Dis* **14**, 186 (2023).

44.Ma, Y., Cai, J., Wang, Y., Liu, J. & Fu, S. Non-Enzymatic Glycation of Transferrin and Diabetes Mellitus. *Diabetes Metab Syndr Obes* **14**, 2539–2548 (2021).

45.Liu, J., Li, Q., Yang, Y. & Ma, L. Iron metabolism and type 2 diabetes mellitus: A meta-analysis and systematic review. *J Diabetes Investig* **11**, 946–955 (2020).

46.Teruya, T., Sunagawa, S., Mori, A., Masuzaki, H. & Yanagida, M. Markers for obese and non-obese Type 2 diabetes identified using whole blood metabolomics. *Sci Rep* **13**, 2460 (2023).

47.Tanaka, A. *et al.* Role of Copper Ion in the Pathogenesis of Type 2 Diabetes. *Endocr J* **56**, 699–706 (2009).

48.Gembillo, G. *et al.* Potential Role of Copper in Diabetes and Diabetic Kidney Disease. *Metabolites* **13**, 17 (2022).

49.Gong, D. *et al.* A copper(II)-selective chelator ameliorates diabetes-evoked renal fibrosis and albuminuria, and suppresses pathogenic TGF-β activation in the kidneys of rats used as a model of diabetes. *Diabetologia* **51**, 1741–1751 (2008).

50.Gong, D. *et al.* Quantitative proteomic profiling identifies new renal targets of copper(II)-selective chelation in the reversal of diabetic nephropathy in rats. *Proteomics* **9**, 4309–4320 (2009).

51.Rovira-Llopis, S. *et al.* Mitochondrial dynamics in type 2 diabetes: Pathophysiological implications. *Redox Biol* **11**, 637–645 (2017).

52.Prasun, P. Role of mitochondria in pathogenesis of type 2 diabetes mellitus. *J Diabetes Metab Disord* **19**, 2017–2022 (2020).

53.Pinti, M. V. *et al.* Mitochondrial dysfunction in type 2 diabetes mellitus: an organ-based analysis. *American Journal of Physiology-Endocrinology and Metabolism* **316**, E268–E285 (2019).

54.Patti, M.-E. & Corvera, S. The Role of Mitochondria in the Pathogenesis of Type 2 Diabetes. *Endocrine Reviews* **31**, 364–395 (2010).

APPENDIX A3: MODGENEPLEXUS

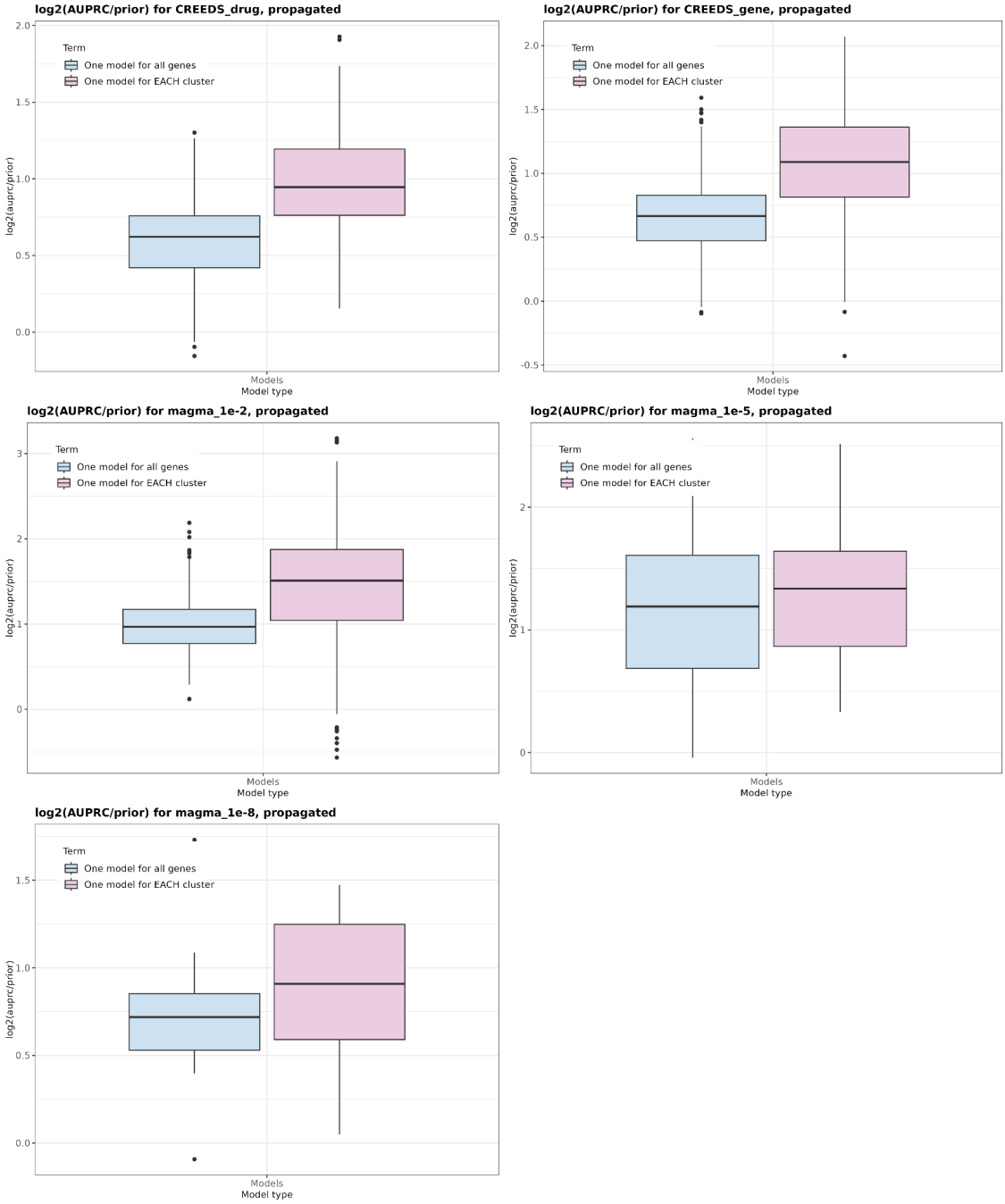

Figure A3.1: Comparing GenePlexus and ModGenePlexus performance for CREEDS drug and gene sets, and for the three magma thresholds.

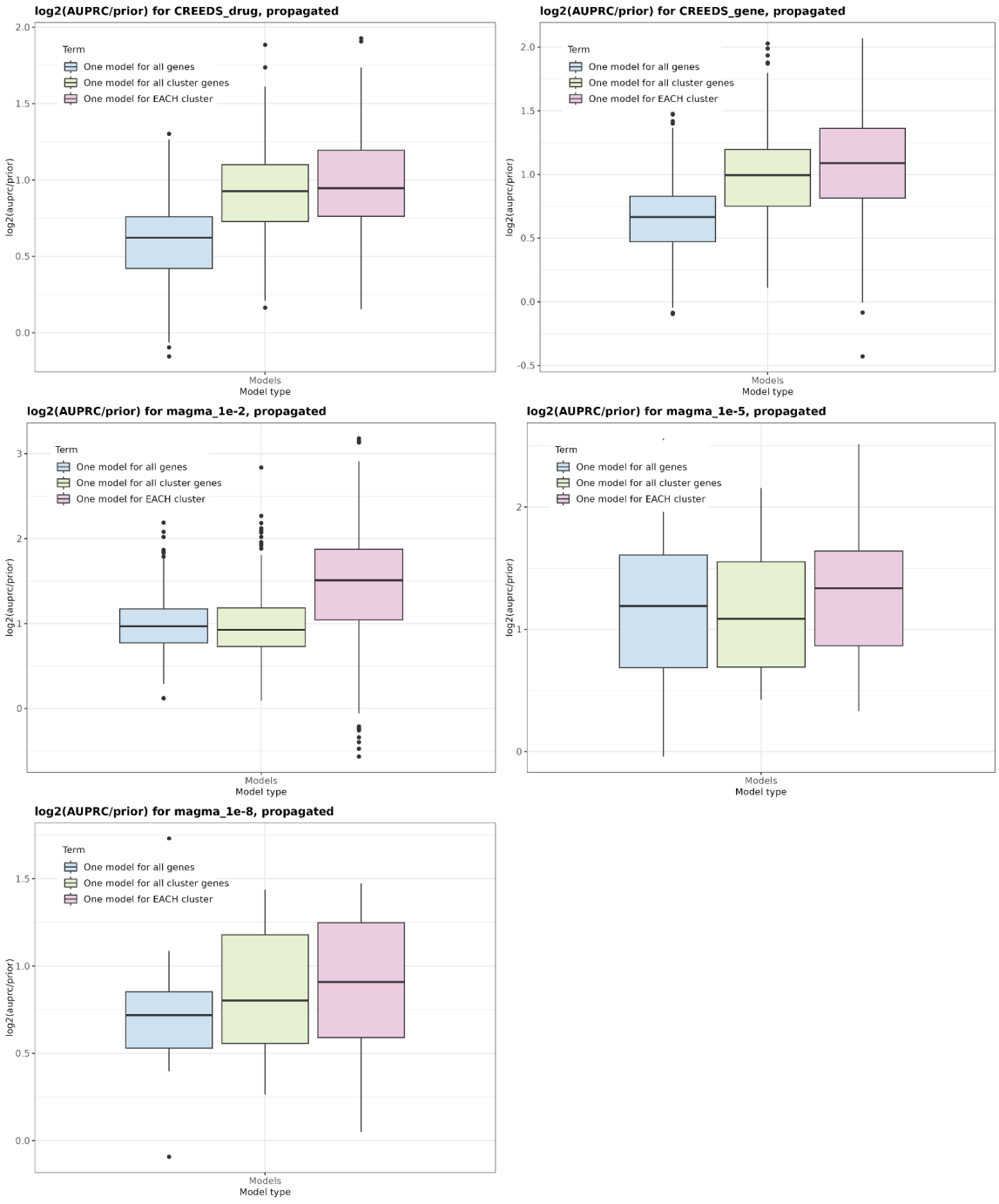

Figure A3.2: Comparing performance between GenePlexus, ModGenePlexus, and AllClus for CREEDS drug and gene sets, and for the three magma thresholds.

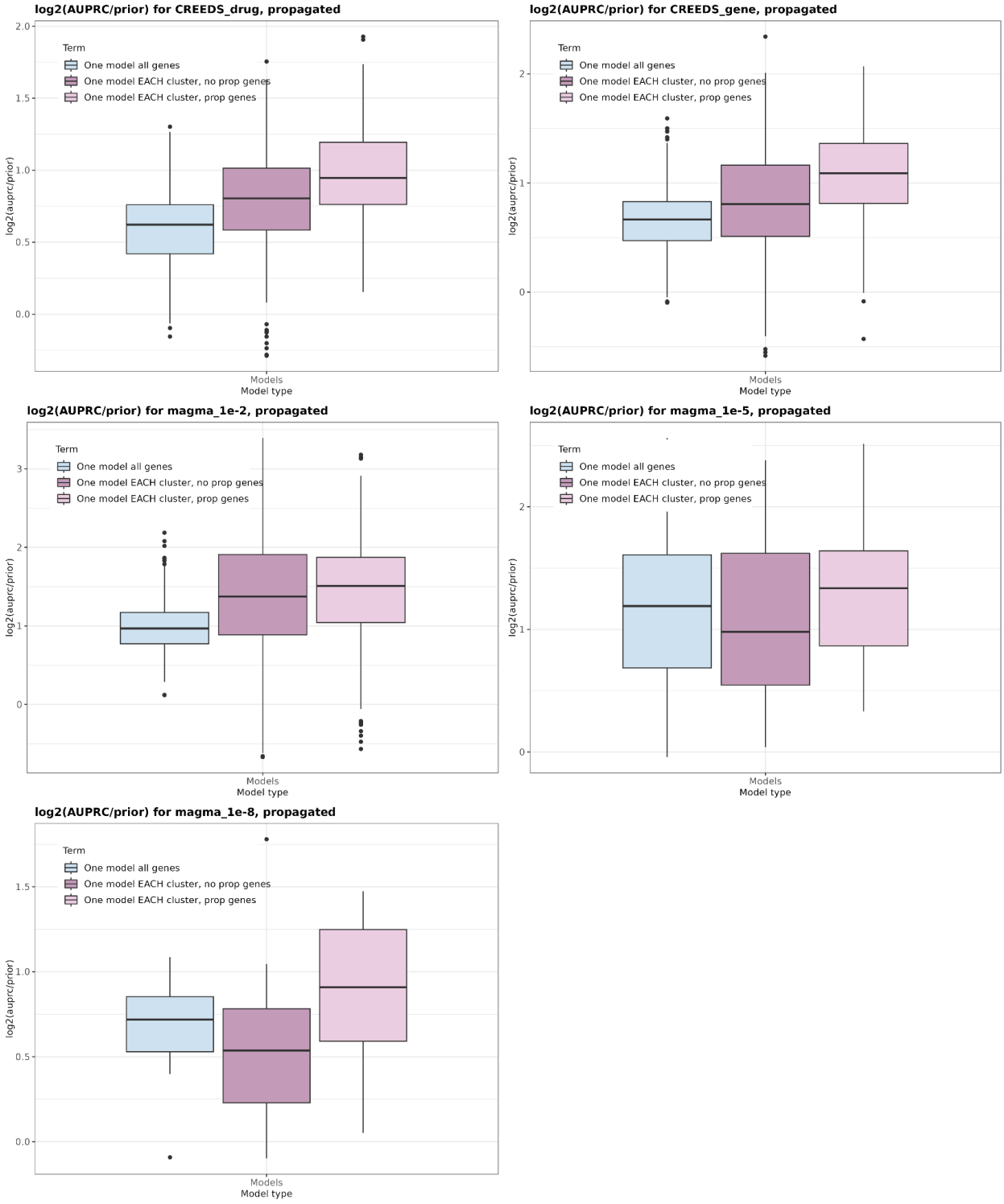

Figure A3.3: Comparing model performance of AllAssign, noprop_ModGenePlexus and ModGenePlexus for CREEDS drug and gene sets, and for the three magma thresholds.

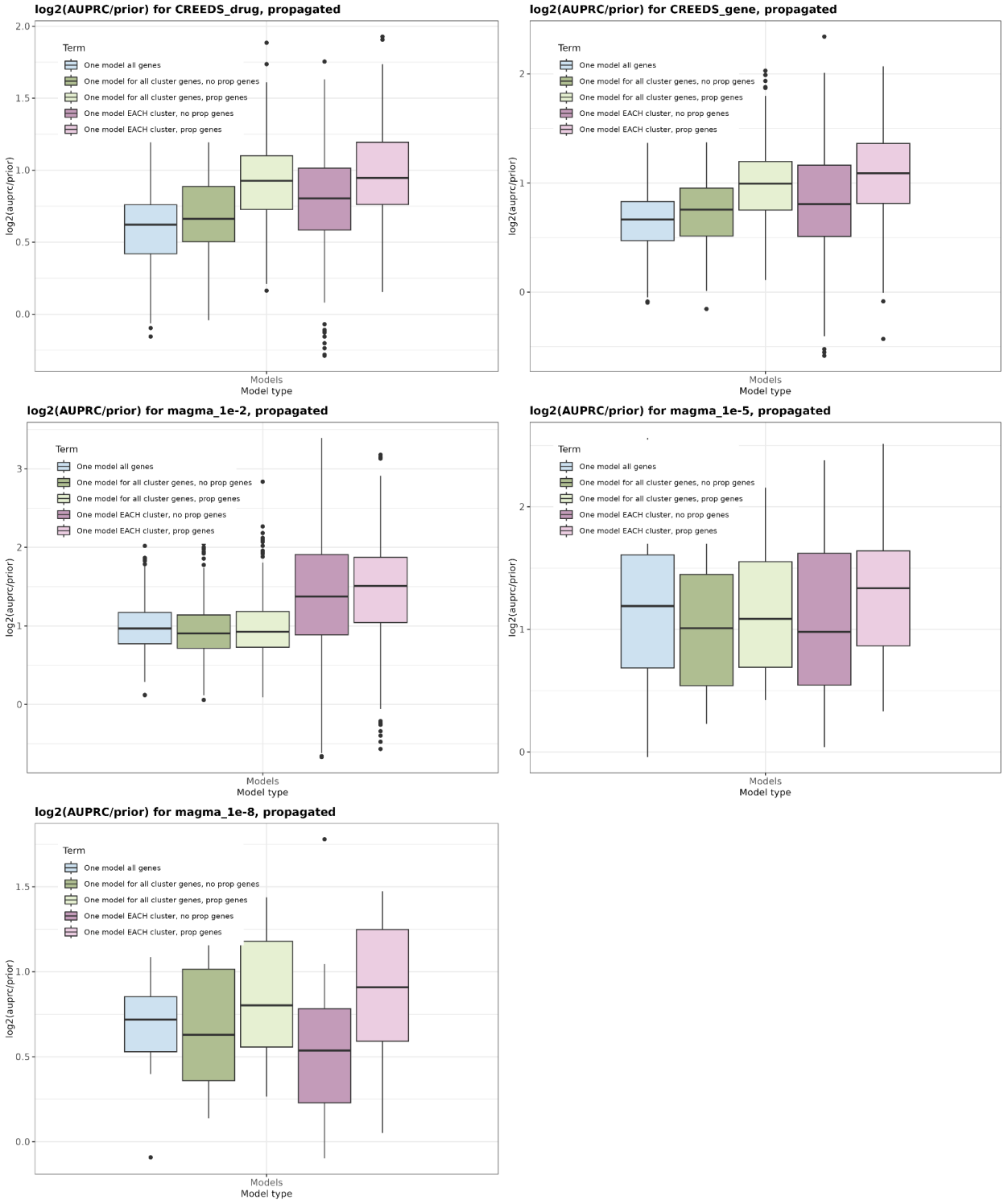

Figure A3.4: Comparing all model types for CREEDS drug and gene sets, and for the three magma thresholds.

CHAPTER 4: DISCOVERING CORE AND PERIPHERAL GENES USING A NETWORK-BASED OMNIGENIC MODEL AND TRANSLATING FINDINGS ACROSS SPECIES

Introduction

Complex diseases manifest through numerous interactions of genetic factors. The genetic interpretation of disease can be conducted at the variant, gene, process, or phenotype levels. Experiments at one of these levels can be integrated with other biological data using computational methods to gain a holistic view of the disease at all levels of biological organization. GWAS have revolutionized our understanding of the etiology of complex diseases and highlighted why discovering robust, consistently successful treatments is challenging. This method is highly useful, not only for its genetic findings but also because sequencing patient genomes has become exceedingly affordable. A flaw of GWAS has emerged, ironically caused by the high scale of complexity that GWAS has been essential in revealing. For very complex diseases, a "naive" GWAS cannot keep up due to a lack of statistical power and the sheer number of implicated loci found in large populations. As GWAS sample size has increased, the number of loci that meet nominal p-values has increased, but the effect size of implicated SNPs has decreased[1,2]. Establishing the biological mechanisms of these many small-effect SNPs is challenging, as their influence on the trait is thought to be indirect, involving downstream functions. Classic post-GWAS methods, such as fine mapping, also struggle when the effect sizes are small. The key question is, given that GWAS results show few SNPs with large effect sizes, what is the best way to interpret and utilize data from these studies in discovering mechanisms for complex human disease?

The observation that low effect size SNPs are likely to affect phenotypes and other genes through indirect effects is one of the primary motivations of the omnigenic model[2,3]. The omnigenic model proposes that there are two categories of genes – core and peripheral – that work together in biological networks to explain disease states or manifestation. In other words, diseases can be interpreted as a subnetwork in the human genome. The model's prefix "omni-" is because since all genes will be in a genome wide network, and all genes have at least some degree of connection,

theoretically all genes could have an influence on disease state. This interpretation of complex disease is not unique to the modern era. Ronald Fisher's infinitesimal model was motivated by reconciling ideas of Mendelian inheritance with quantitative (complex) traits like height and diseases. Fisher's model proposed that individual loci which were inherited according to Mendel's laws have a small effect by themselves that is either difficult to measure or cannot be, but the cumulative effect of the inherited loci explains how a trait manifests or is phenotypically observed. The implication is that measuring the causal effect of a loci is extremely challenging. Fisher noted in his time, that while plant and animal breeders had the capability to select for particular complex traits, this owed little to genetic and statistical analysis[4] due to not being able to explain the causal loci. In the modern era, we have proposed and determined mechanistic explanations of particular genes in complex traits, but disease-gene annotations are still incomplete and are very challenging to add to. The unique proposal of the omnigenic model is that networks explain why there are so many genes associated with complex traits, where peripheral disease genes are associated through their interactions and effects on core disease genes that are individual mechanistically interpretable entities in the complex trait.

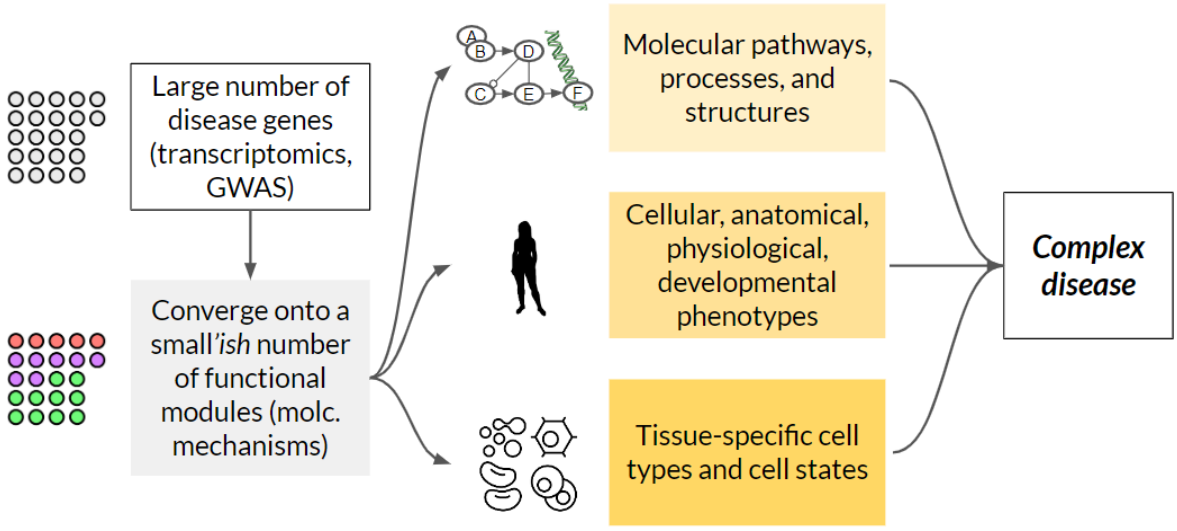

Figure 4.1: Starting with disease level data, it is possible to break it up into smaller, distinct, and biologically meaningful subsets. These subsets are used to define biological context. Understanding how pathways, phenotypes, traits, and context are

Figure 4.1 (cont'd)

annotated to a disease and relate to one another is crucial in mechanistically explaining the disease as a whole.

The observations from both the early and modern eras of genetics show why we need to use other forms of biological data to interpret genotypic data like GWAS. Alone it is very difficult to explain why loci are involved in a human disease. Therefore, post-GWAS analysis aims to provide evidence for downstream pathways and intermediary molecular traits and phenotypes through which genes contribute to diseases[5–8] (**Figure 4.1)**. Biological networks are powerful tools because they provide context about genetic relationships by integrating multiple types of biological data. Additionally, we discussed in chapter 1 and show in chapter 2 and 3 that complex diseases are modular within biological networks, and the genes in each module will correspond to different underlying traits, pathways, and phenotypes. Leveraging this disease modularity in networks can be a way to interpret and discover possible underlying phenotypes, but first we must answer (i) how do we determine disease genes using GWAS data, (ii) how do we find and utilize disease gene modules and (iii) how do we discover core genes that are directly involved in the phenotype/trait disruption that leads to disease?

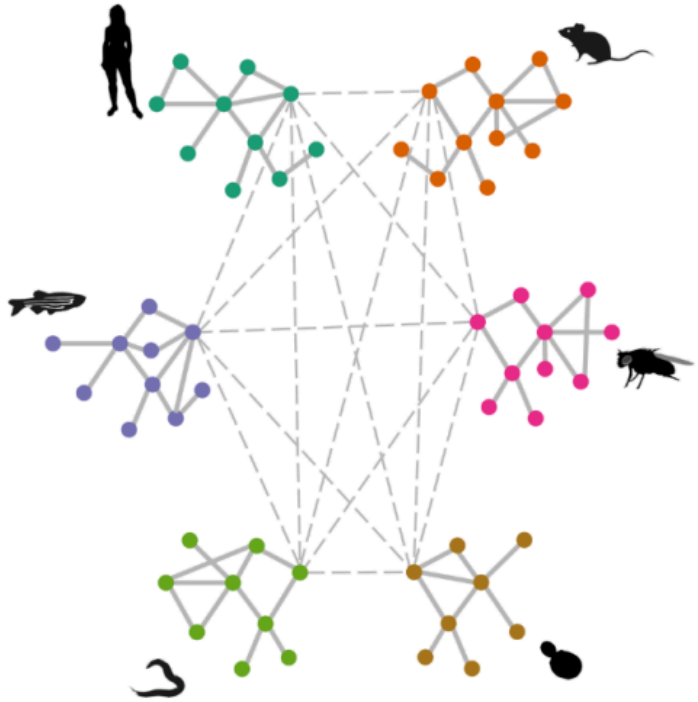

Figure 4.2: A representation of the multi-species network created for GenePlexusZoo. It is a joint representation of molecular networks of humans and five common model organisms (mouse, fish, fly, worm, and yeast) are combined using pre-calculated orthologous groups across these species for the genes at a genome-wide level.

Linking implicated variants or genes to higher-level annotations like pathways is crucial for understanding biological mechanisms. Evaluating candidate genes and pathway explanations often starts with model organisms, which have been essential in advancing our understanding of human disease, organ systems, genes, and many other biological concepts[9–12]. Model organism databases are another resource that has grown massively in recent years[13], further incentivizing their use in computational studies. Methods that allow the transfer of genetic information and biological pathways from one species to another are key in discovering model organisms genes and pathways that are relevant in highly complex human disease. Recently, our group released a modification to GenePlexus termed GenePlexusZoo[14]. It utilizes a multi-species network (**Fig 4.2;** including human, mouse, fish, fly, worm, and yeast) where a user can input genes from one of six species and retrieve results for the same species as the input set, or results

translated into one of the remaining five species. GenePlexusZoo not only demonstrated improved gene classification within a single species by leveraging the evolutionary relationships contained in the multi-species network, this multi-species network allows for cross-species correspondence of GOBPs and phenotypes to be discovered that would not have detected if evaluating relationships through a naive overlap of orthologous genes alone (**Figure A4.1, see methods**) . In chapter 2, we demonstrated that it is possible to learn meaningful phenotypes underlying modules for different human diseases, and showed how implicated inflammation phenotypes are seen across them from using gene predictions from GenePlexus. With GenePlexusZoo, we can leverage using multi-specie gene edges to discover genes relevant to model organisms, which can be tested for enrichment with phenotypes. It will be useful to implement similar methods at the module level and for cross-species analysis. If meaningful knowledge transfer occurs on a module level, we can determine which model organisms are important for particular and more specific phenotypes, rather than the disease as a whole.

In chapters 2 and 3, we have shown that using networks to give context to GWAS and highly complex human disease data helps to discover functional genes and how genes in large, experimental genesets work together in biological processes. We propose that using the network provided in GenePlexusZoo will provide additional benefit in interpreting and validating functionally relevant genes to complex disease. We propose using a modified version of ModGenePlexus (as discussed in Chapter 3) integrated with GenePlexusZoo to reveal biological insights of human disease in a multi-species and omnigenic framework. We begin by using MAGMA results to obtain an initial list of seed genes measured from GWAS for 20 complex human diseases and traits. We then apply DOMINO to discover disease gene-enriched modules, performing semi-supervised learning to recover false negatives and remove poorly evidenced hits. ModGenePlexus is then run for each module, and a final disease gene prediction list is aggregated based on the top predictions across all modules. Next, we use the predicted disease genes to define and predict a disease module, directly used to predict core and peripheral genes for the human GWAS. Our approach for classifying these core and peripheral genes exclusively uses network relationships – avoiding assumptions based on prior disease

knowledge. We further investigate various biological traits of these core genes – such as conservation, cross-disease relationships, and tissue specificity – to learn how network central genes behave. The disease genes are then translated to model species using the GenePlexusZoo framework, including determining "core genes" in other species that best reflect the human biology of the disease. Lastly, we investigate and attempt to validate the use of human disease gene modules for knowledge transfer by assessing if predicted model organism genes are related to human genetics of the GWAS. The power of these methods is that it can take disease level data, break it into parts, and then build those parts back up into a final, singular disease module for a trait for interpretation and discovery in a holistic way through computational methods with limited biological assumptions (**Figure 4.3)**.

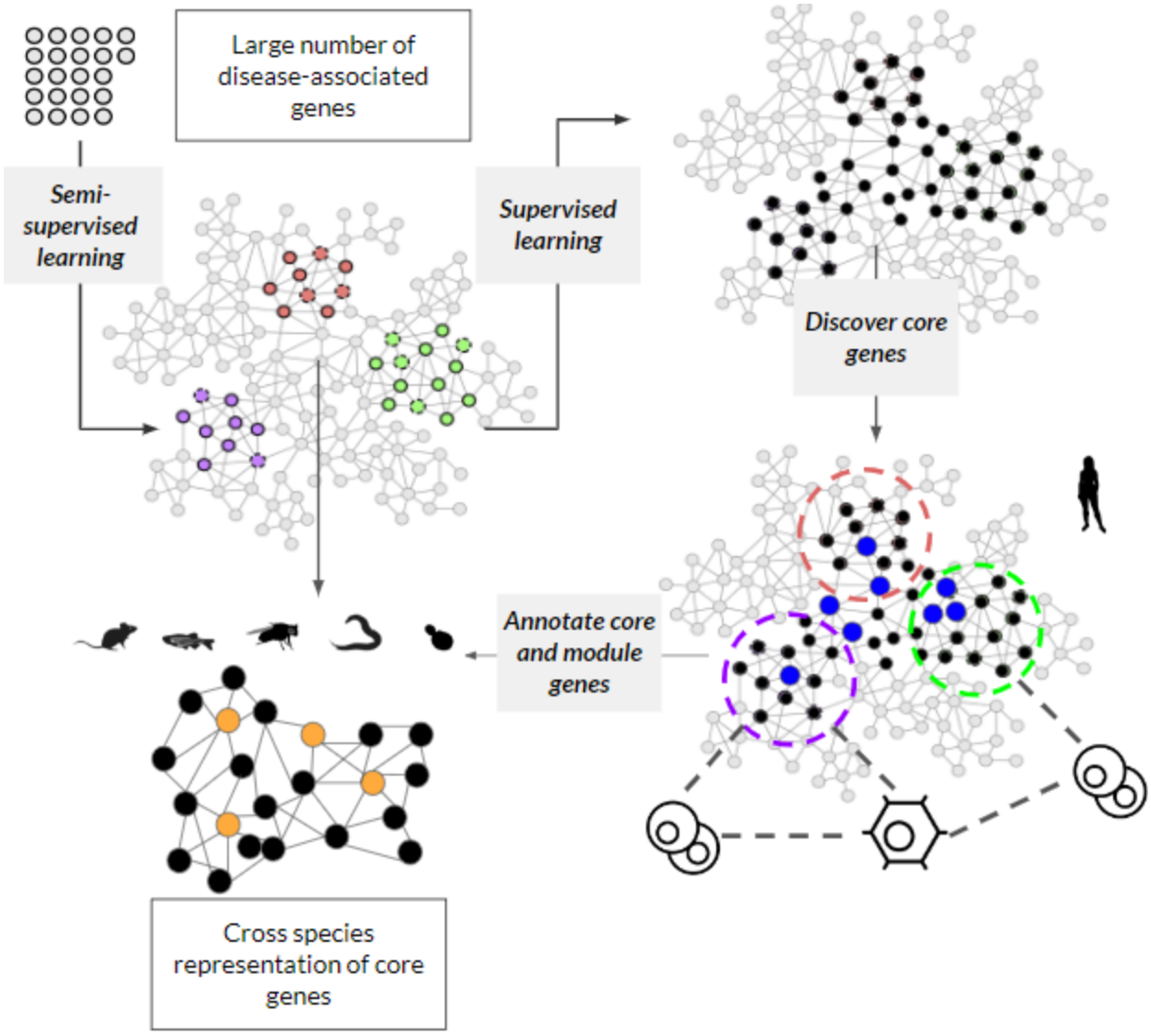

Figure 4.3: A pipeline for our methods and analysis. We start with a list of disease genes measured from genotypic data with GWAS for 20 diseases and traits. We then use ModGenePlexus to perform (semi-) and supervised learning for gene classification. The genes that have a high prediction are considered disease genes and are a predicted disease module for the underlying GWAS. To categorize genes, we use an omnigenic framework and interpretation where genes are either core – genes likely to have direct effect on disease manifestation – or peripheral whose annotation to the disease is because of their network connection on core genes. Genes are labeled core if they have the highest betweenness centrality values of the genes within the specific disease module. We return results for GenePlexusZoo predicted genes for model

Figure 4.3 (cont'd)

organisms, and we investigate if our classified human core genes appear as orthologs in non-human gene predictions. We additionally utilize large amounts of external biological data to categorize both core genes and module assignments at phenotype, tissue, cell levels and within different model-organisms.

Methods

GenePlexusZoo attributes used for creating models

GenePlexusZoo[14] is an extension of GenePlexus that utilizes a multi-species network. Networks from STRING[15] were obtained for species *Homo sapiens, Mus Musculus, Danio rerio, Drosophila melanogaster, Caenorhabditis elegans,* and *Saccharomyces cerevisiae*. The genes across networks were connected using orthology information from the eggNOG[16] database, which is a genome-scale definition of orthologous groups for many species. These cross-species edges were weighted using a function of a genes' within-species node degree and how many cross-species edges the gene has. The features used within GenePlexusZoo are derived from a low-dimensional representation of the cross-species network, generated using the PecanPy[17] software, which is a fast and scalable implementation of the node embedding method node2vec[18].

Discovering gene and disease modules

We used the DOMINO[19] software as described in chapter 3 in the same way as for ModGenePlexus. Genesets were passed in and we removed genes that fell into modules smaller than 10 genes, and added genes propagated to the original set. In addition to gene modules, we aggregate GenePlexus predictions across these modules to predict a disease module containing all genes predicted by GenenPlexus. When GenePlexusZoo returns gene classification results, it calculates a z-score that indicates how high the probability of a gene prediction is relative to the rest of the predicted genes of a single species. To aggregate gene predictions for a species of interest, we took all gene predictions in each module where the $Z > 5.0$. We use the term "disease module" for continuity based on a discussion in chapter 1 about types of modules found in biological networks. In this chapter, a "disease module" can refer to a complex trait as well depending on whether the GWAS was conducted for a complex trait like height, or a disease like atrial fibrillation.

Compiling MAGMA gene prioritization results

MAGMA[20] gene prioritization scores for summary GWAS results were compiled from gwasATLAS[21]. We created a GSC utilizing these gene predictions based on thresholds of $p < 1 \times 10^{-5}$. We then clustered these genes using DOMINO. We manually chose 20 GWAS that are diverse in terms of phenotypes, including both diseases and complex traits. A threshold of $p < 1 \times 10^{-5}$ was chosen because we wanted to use genes that met a strict threshold, however stricter thresholds would not give enough genes for ModGenePlexus, and chapter 3 shows that using looser threshold genes can improve prediction of genes that would be highly significant in GWAS results. The 20 GWAS that were chosen are: age at menarche, age at menopause, age related macular degeneration, alcohol dependence, atrial fibrillation, celiac disease, crohn's disease, educational attainment, height, inflammatory bowel disease, primary biliary cirrhosis, primary sclerosing cholangitis, rheumatoid arthritis, two systemic lupus erythematosus studies, type-1 diabetes, type-2 diabetes, ulcerative colitis, and vitiligo.

Running GenePlexusZoo and ModGenePlexus

GenePlexusZoo was run with each individual module for the disease. If there are *N* modules, then *N* models are created. Only human genes from the module assignments are used, and we return predictions for every species in the multi-species network. For each species, each gene has *N* predictions, one for each model. ModGenePlexus was run using all options – which include utilizing propagated genes from DOMINO, and creating a model for each cluster.

Finding enriched phenotypes and GOBPs using GenePlexusZoo

GenePlexusZoo utilizes GSCs of Monarch[22,23] phenotypes and Gene ontology biological processes (GOBP)[24,25], where each term in a GSC is annotated with a set of genes known to be associated with the term The GenePlexusZoo software contains files of model weights, where each set of model weights comes from models trained on each phenotype and GOBP in the GSCs. To find enriched phenotypes or GOBPs for a GWAS gene module, the model weights for the model trained using genes from a given module are compared to each phenotype/GOBP model weight vector by calculating cosine similarity. These cosine similarities are then normalized into z-scores to indicate how high the cosine similarity is relative to all other predicted genesets in the collection.

Relative to an over representation hypergeometric test, GenePlexusZoo utilizes the model weights of all genes in the genome for comparing phenotypes, acting much like a network-based gene enrichment method[26]. In this project, we use this as a way to predict non-human phenotypes for our human GWAS without needing to convert the human geneset to orthologs and then doing an enrichment test.

Determining human core and peripheral genes

Core and peripheral genes are predicted using the networks within the joint multi-species network and the predicted disease module. For human genes, the human version of string with edge weights was obtained. The network was subsetted to a disease-specific subnetwork which includes all genes that had a $Z > 5.0$. We applied the betweenness centrality metric using NetworkX[27], defined by equation:

$$c_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t \mid v)}{\sigma(s,t)}$$

Where *V* is the set of nodes, $\sigma(s,t)$ is the shortest (s,t)-paths, and $\sigma(s,t \mid v)$ is the number of those paths that pass through node *v* other than s,t. If $s = t$, $\sigma(s,t) = 1$, and if $v \in s,t, \sigma(s,t|v) = 0$. The equation that NetworkX uses is based on an algorithm first implemented by Brandes[28]. Betweenness centrality is a metric that quantifies how often nodes are in the shortest path between any pairs of nodes. For every node, the shortest paths are calculated from that node to every other node in the graph. In this iterative process, what is recorded is how often a node is within the shortest path for each pair of nodes. Higher betweenness centrality values mean the node is centrally connected to the graph as a whole. In our context, we use betweenness centrality to predict core genes because based on the omnigenic definition of core and peripheral complex disease genes, core genes are a relatively small set that are influenced by a large amount of peripheral genes through network connection. After running the betweenness centrality algorithm, the values for each gene are scaled using Z-score normalization, where high z-scores indicate betweenness centrality values that are high relative to the rest of the genes in the network. Genes with $Z > 2$ were chosen as predicted core genes for the trait.

Predicting genes for other species and a model-organism disease module

As described in **Methods 4.1**, the networks of multiple species are connected in this joint representation through use of the eggNOG database and orthology information. This allows the prediction of other species genes based on the human input data from MAGMA. Thus for each GWAS trait, we are able to obtain the top ranked genes for each species. To compile a model-organism disease module, we used the same threshold of $Z > 5.0$ for choosing genes that were highly predicted for our trait.

Compiling one-to-one human orthologs

Although GenePlexusZoo makes use of many-to-many ortholog information, to determine correspondence of core genes across species we restrict ourselves to one-to-one orthologs. We obtained these ortholog relationships from BioMart[29].

Discovering if core genes are enriched for being constrained

We determined if genes were constrained using data from gnomAD[30–32]. A gene being constrained indicates that there is strong selection against mutation of the gene. Specifically, we utilized the Loss-of-function Observed / expected upper bound fraction (LOEUF) score. A gene with a low score is one that has selection against predicted loss-of-function variation, while a high ratio means the gene does not have much selection against mutations that inactivate the gene.

Discovering core gene tissue specificity enrichment

Tissue specificity data was collected from the CONE[33] method. Using tissue expression data from GTEx[34], z-scores were calculated that indicate a geneset's tissue specificity. We calculated z-scores utilizing an average z-score method and a max z-score method, where the average z-score is based on the average score of genes across tissues, and max z-score is based on the max score of genes across tissues. Another score used is the tau value[35,36], where a small value means the gene is broadly expressed across tissues and a high value means it is specific.

Calculating module tissue and cell marker gene overlap

We used data from Jensen's TISSUE database[37] and cell marker data from CellSTAR[38]**.** For each dataset, we took genesets with at least 10 genes and calculated enrichment using a hypergeometric test between GWAS genes and each set. This test was for

every module across the GWAS.

Definition of disease modules for humans compared to model organisms

In this project, we discover disease modules for each GWAS for five species. However, the interpretation of the GWAS within human vs non-human species needs nuance. The GWAS come from human studies only, meaning the genes used as input are always human. This means that the diseases and traits are defined for humans. When we compile human gene predictions and map those genes to a network, this is what we term a predicted disease/trait module – composed of all genes thought to be possibly relevant to the disease based on network connections to genotypic data from GWAS. When we compile a module of the predicted genes of other species, it should not be thought of as the same “disease module” within another species. It should be interpreted as human knowledge transferred into the gene space of other organisms that may have similar underlying biology to the human disease module. Similarly, when the betweenness centrality method is run for these modules, these should not be thought of as “core genes” to the human disease, but as genes with high network centrality within the module. The distinction of interpretation between human/model organism data from the “same” methods is vital because while the diseases and traits are interpretable in humans, it is much less so in other species. The reason why we create modules and discover “core genes” for other species is because, while not equivalent, the information discovered is directly relevant to human genetics. What we are actually displaying are genetically similar non-human gene subnetworks discovered directly using human disease data. Our goal is to discover human core genes and see if they have meaning in other organisms. For ease of reading, if a gene is called a “mouse core gene”, that phrase is used based on the method and interpretation within the module – we are not making claims that this gene is a core gene for the respective disease in mice.

Results

Discovering core genes (Defining betweenness centrality and number of core genes compared to peripheral)

To define core genes for each of the 20 GWAS, we began with running an integration of ModGenePlexus and GenePlexusZoo (**see methods**). Models were created for each

gene module discovered using DOMINO, and GenePlexusZoo provides a multi-species network to improve classification. All GWAS used are from humans, and we first classified human genes. Every module has an initial training set determined from DOMINO and predicted genes within that module (**Figure 4.4**). The predictions from GenePlexus were compiled, and every gene with a prediction of z-score > 5 are compiled into a final disease module. This disease module is defined as containing every gene relevant to any of the determined gene modules (**Figure 4.5-6**). Core genes that are highly connected to other genes in the disease module are found using the betweenness centrality calculation (**see methods**). Genes with a high betweenness centrality are considered core, while the majority of genes across each GWAS are peripheral (**Figure 4.7**). Our method utilizes the omnigenic model to choose a small set of core genes that are defined using the context of all other genes within the disease module, including those predicted using GenePlexusZoo and ModGenePlexus. Thus, genes are core in the context of their network connections, rather than using assumptions about disease biology outside of the GWAS experiment. In **Figure 4.8,** we show that core genes are enriched for genes used in training (either implicated in GWAS or propagated) compared to those genes predicted from GenePlexus only. We see that the original experiment is enriched with network-central genes, but the network still finds those not measured in the experiment[39]. Sample results for other GWAS are in **Figure A4.2-5**.

Core genes discovered for atrial fibrillation are mechanistically meaningful

We predicted core genes using GWAS data and networks, rather than using specific knowledge about a disease to first define mechanistically important pathways, and find important gene annotations. It is still important to give biological context to our predictions to see if there is known relevant biology to the traits. In **Figure 4.9,** we list all predicted core genes for atrial fibrillation along with the knowledge source that predicted their annotation. Six genes came from the initial GWAS study, five genes came from propagation in the network, and seven came from GenePlexus predictions. All 18 predicted core genes have literature support for being related to atrial fibrillation or cardiovascular disorders in general. WASL has been linked to cardiovascular disease[40], and is in pathways enriched with known atrial fibrillation genes[41]. FXR1 is associated

with gap junction remodeling when upregulated, and this pathway is a feature of heart disease involving arrhythmia[42]. FMR1 targets in cardiac muscle[43] and fragile-X carriers have higher rates of cardiovascular issues[44]. CFL2 is directly implicated to atrial fibrillation in a recent GWAS study[45] and another recent study measured it as a possible drug target[46]. H3-7 is a putative histone gene, and histone modification has been implicated in zebrafish studies for atrial fibrillation by affecting cardiac contractile function[47]. CALML6 has been predicted to be associated with coronary artery disease[48] and it negatively regulates the NF-κB signaling pathway source[49], a pathway which has higher activity in atrial fibrillation patients[50]. UBA3 is involved in the ubiquitin mediated proteolysis pathway, which is important in atrial fibrillation[51]. It additionally is involved in Neddylation, a pathway which controls misfolded proteins and has been linked to numerous heart issues[52,53]. Some of the other predicted human core genes will be discussed later in the results section when discussing the human core gene relationship to model organisms. We see that the predicted human core genes are highly implicated in relevant organs and processes for atrial fibrillation, thus it is now worth discussing other properties of core genes across GWAS gene modules and across each GWAS.

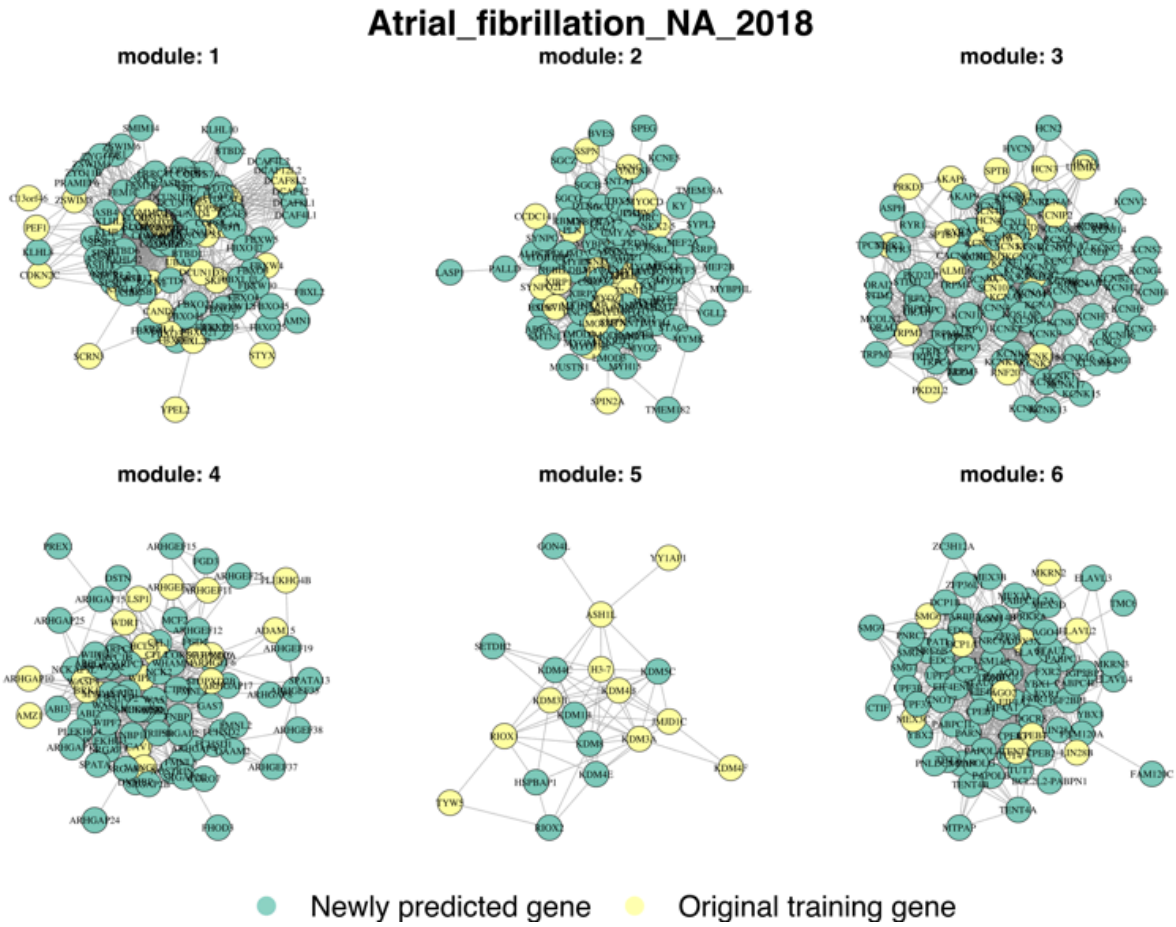

Figure 4.4: Visualizing the gene modules for atrial fibrillation. Each module with at least 10 genes was run with GenePlexusZoo and ModGenePlexus. The genes displayed here are those that had a z-score > 5.0. Original training genes (tan) are those that were implicated by GWAS or found in DOMINO propagation, and newly predicted genes (green) were predicted by GenePlexusZoo.

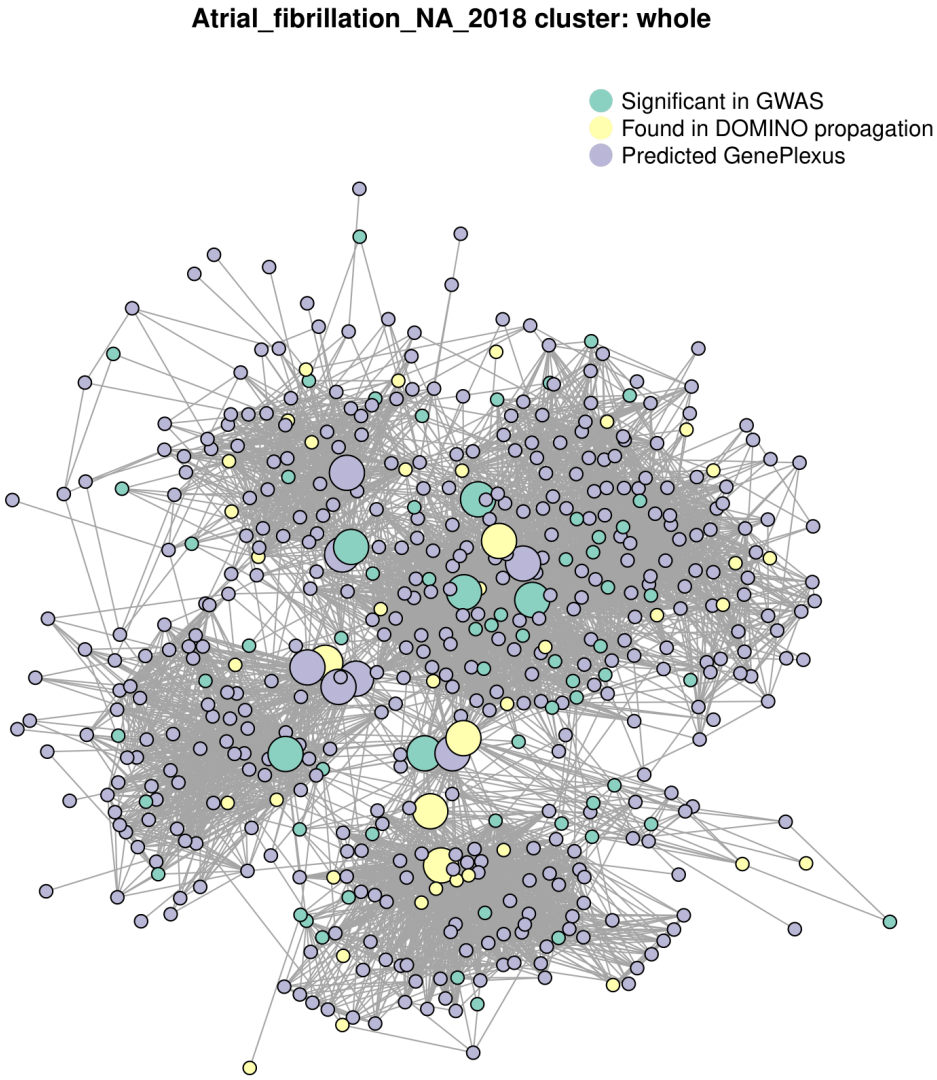

Figure 4.5: A visualization of a predicted disease module for atrial fibrillation. This module contains three types of genes, those implicated by the GWAS (green), those discovered by DOMINO (yellow) and those predicted by GenePlexus (purple). Nodes that are bigger are the predicted core genes. Core genes are come from all 3 types of gene associations and they are distributed across multiple sub-modules within the disease module.

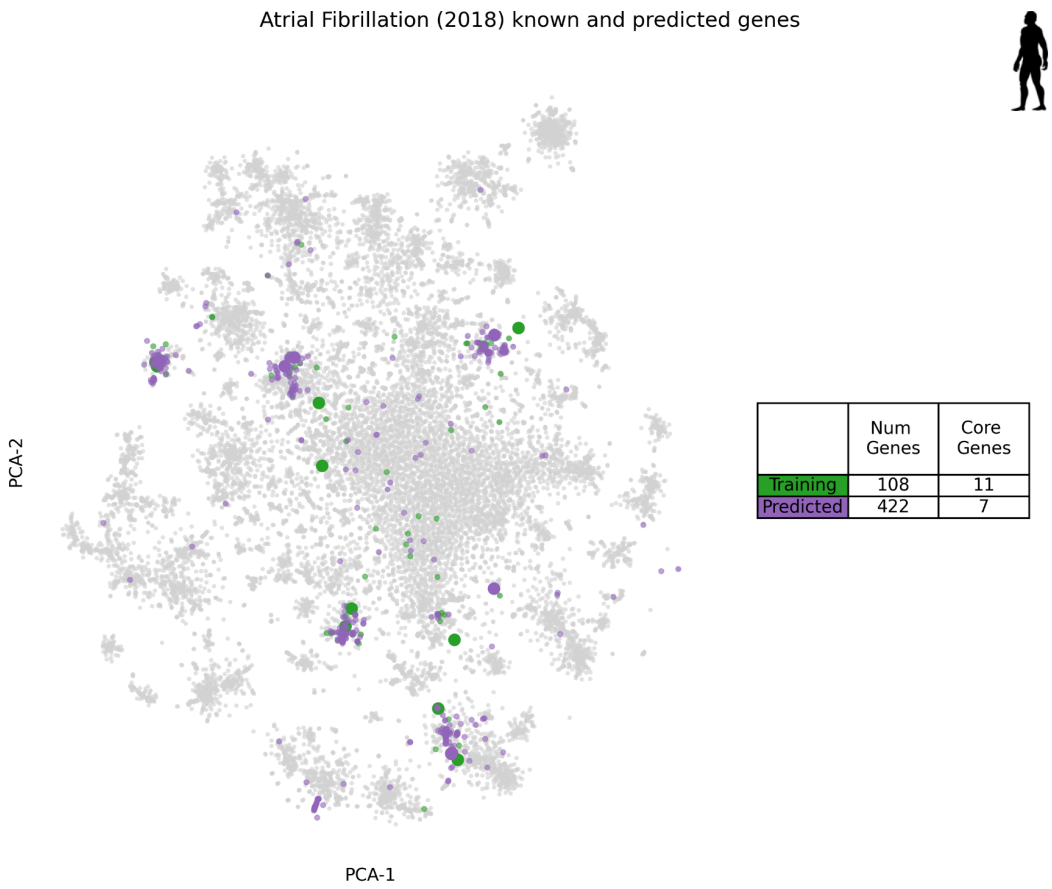

Figure 4.6: Plotting the disease module for atrial fibrillation across the embeddings of STRING in a t-SNE. Within the functional network of STRING, the disease module is not all near each other in a two-dimensional representation. Rather, the disease module is made up of multiple submodules (the gene modules) that are combined to represent a final disease module made up of all genes. Training and predicted genes can appear next to each other in the same modules, and both the training and predicted genes contain core genes.

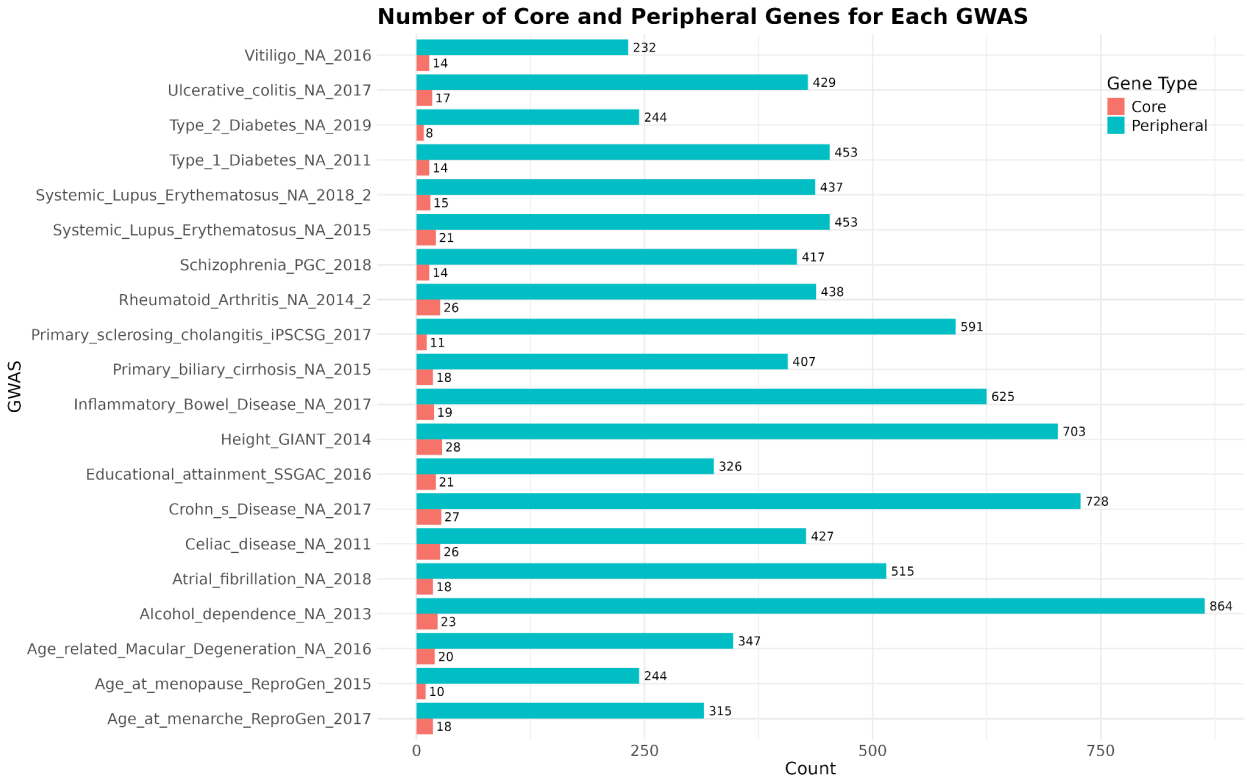

Figure 4.7: How many core and peripheral genes there are in each GWAS. The omnigenic model states that the number of core disease genes is small relative to the number of peripheral disease genes. Our method finds that a small minority of genes (around 10-30 for each GWAS) are predicted to be core, while the rest of the hundreds of implicated genes are peripheral.

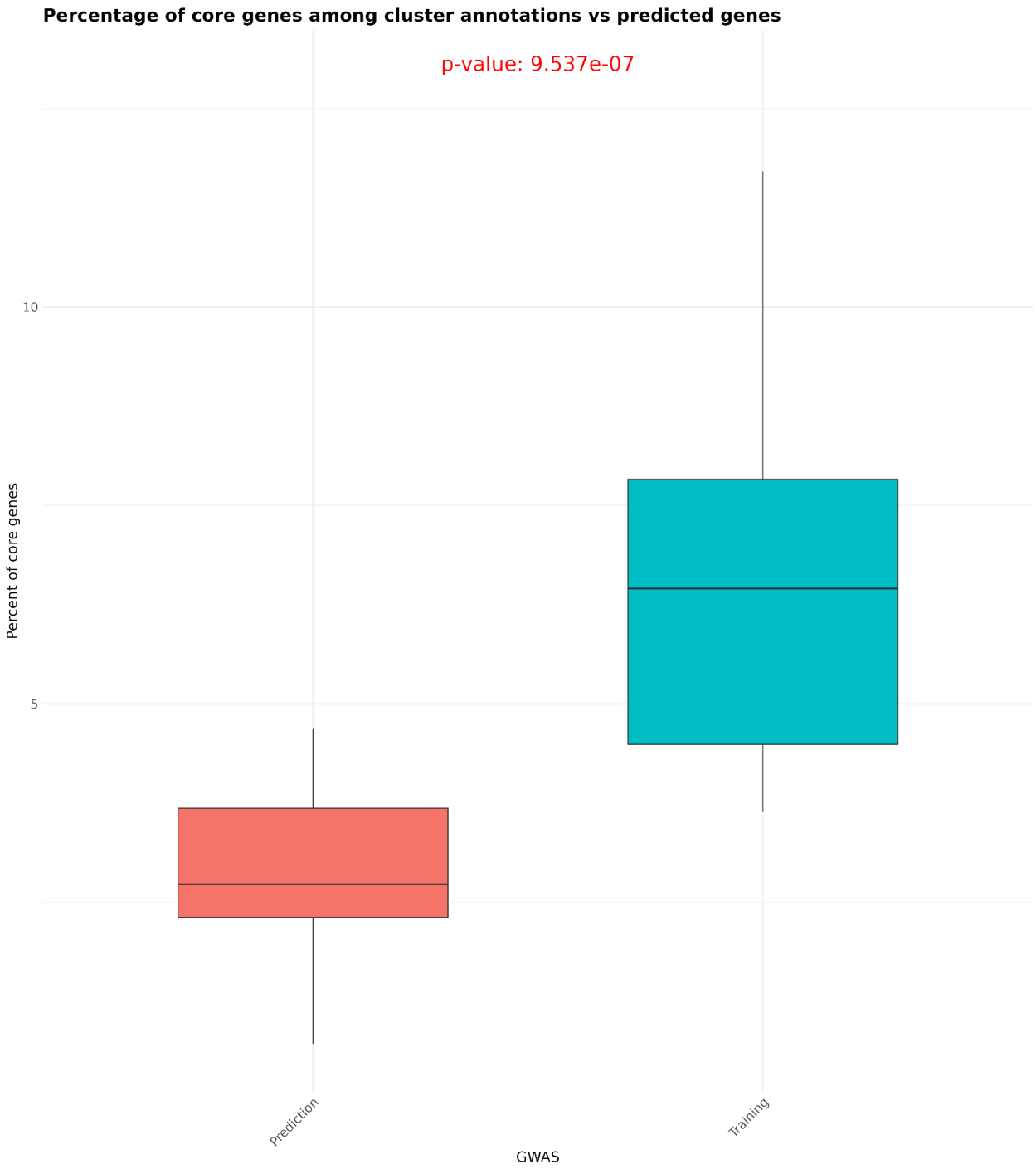

Figure 4.8: Box-plot showing that training genes have more core genes than GenePlexus predictions. The percent of core genes (y-axis) of the training genesets (green) is higher than in the GenePlexus predictions genesets (red). The initial biological data is enriched with core genes from a network perspective, and GenePlexus finds more peripheral genes.

Propagate	H3-7
Propagate	CALML6
GWAS	CAV1
Propagate	SKP1
GWAS	AGO2
GenePlexus	FMR1
GenePlexus	CFL1
GWAS	TTN
GWAS	CFL2
GenePlexus	WASL
Propagate	UBA3
GenePlexus	FXR1
GenePlexus	RYR2
GenePlexus	FXR2
GWAS	CASQ2
GWAS	FBXO32
Propagate	ELAVL2
GenePlexus	ASB5

Figure 4.9: The 18 human genes predicted to be core for atrial fibrillation. Knowledge source (**left column**) for core gene (**right column)** indicates if the gene initially came from the GWAS, network propagation, or GenePlexus.

Core genes are distributed across disease gene enriched modules

A vague part of the definition of core genes in the omnigenic model is in how many processes, phenotypes, or mechanisms they are associated with. Whether the core genes for complex traits are one coherent set or not is an essential question for applying the omnigenic model for real world disease. We decided to investigate whether core genes are seen across different modules within the GWAS. In **Figure 4.10,** we show how the core genes that were part of the original training data were distributed across modules. We see that many modules within a GWAS have at least one gene, with **Figure 4.11** showing that most of the modules have at least one core gene when considering the original module assignments. Notably, we do see some GWAS that have modules highly populated with core genes relative to others. Vitiligo module 1 has 5 core genes when the other modules have 3 in total, and Age at Menopause has only 2 modules of its 3 modules initially containing core genes, with 4 out of 5 core genes falling in module 2. Core genes are determined only after the disease module is made, and in **Figures 4.12-13** we show how each module's predicted genes from GenePlexus end up being classified as core in our method. With the predictions, 10/20 GWAS have core genes in all of its modules, and only 5/20 have more than 1 module without any core genes. For most GWAS, we see that core genes will appear across nearly all found gene modules, showing that core genes are not all contained in one coherent module. However, there will be some modules that are relatively enriched with core genes. To answer what it means for a module to be enriched with core genes, we can see if core genes are more likely to be module-specific within a GWAS, or shared across GWAS.

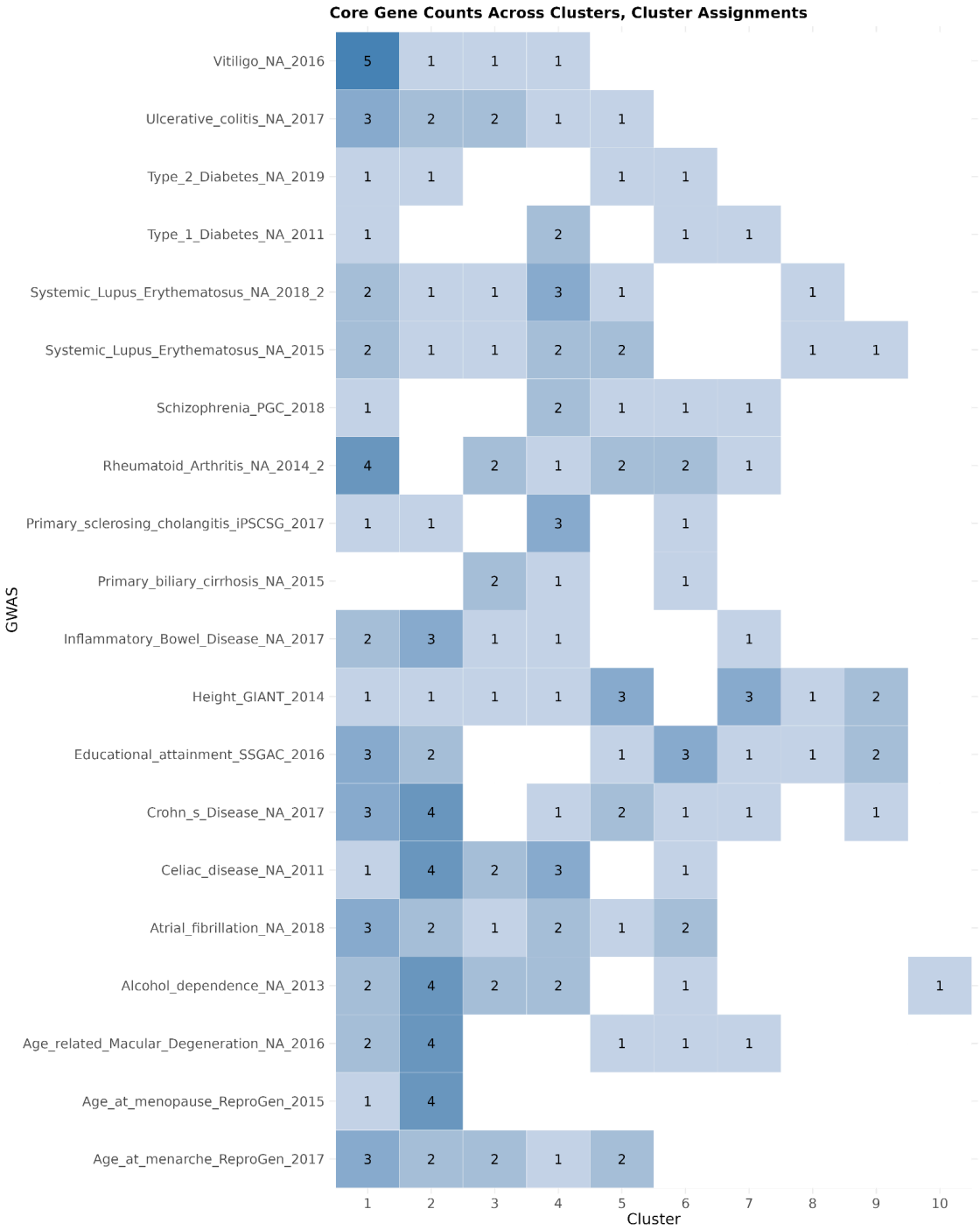

Figure 4.10: Counting the number of core genes across modules for each GWAS of training genes only. Core genes fall across modules of the GWAS. This means that the core genes are not one set of genes in a singular module, but are part of distinct processes.

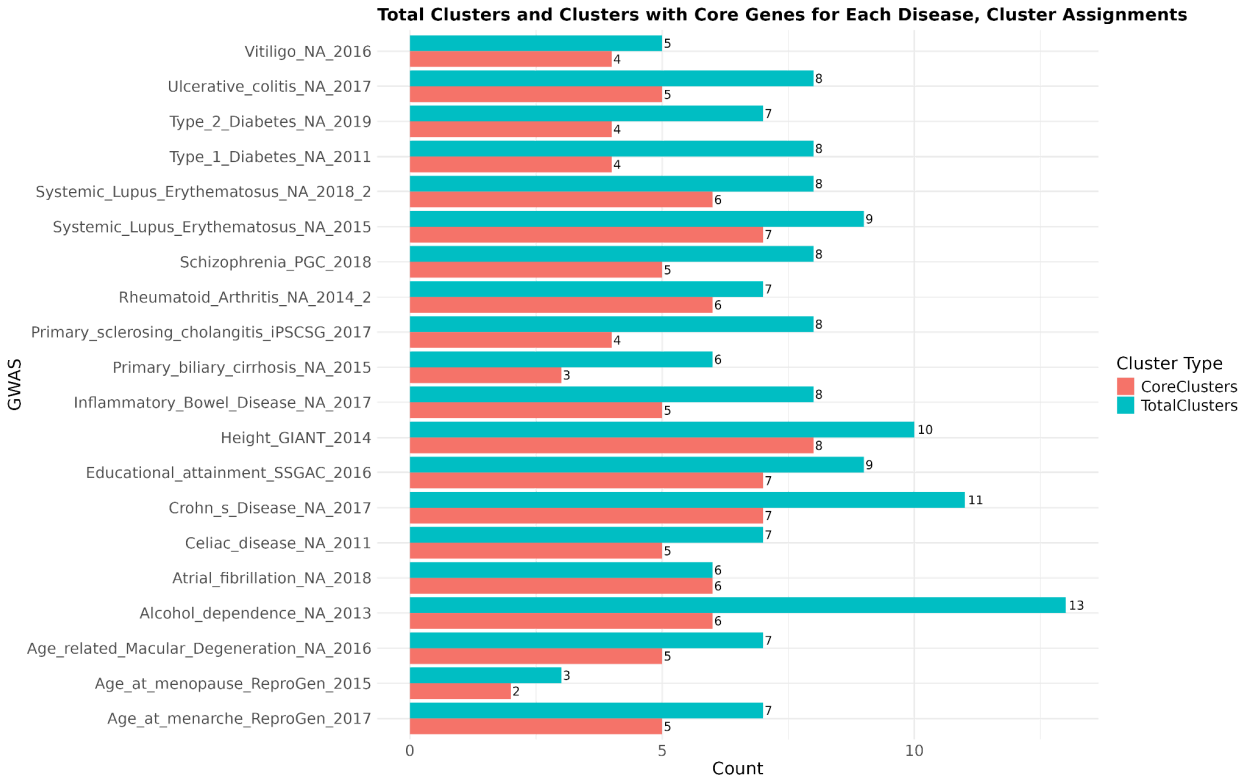

Figure 4.11: Bar plots showing the number of modules with core genes. Core genes are in multiple modules for each GWAS, and for each GWAS at least half of all modules have at least one core gene.

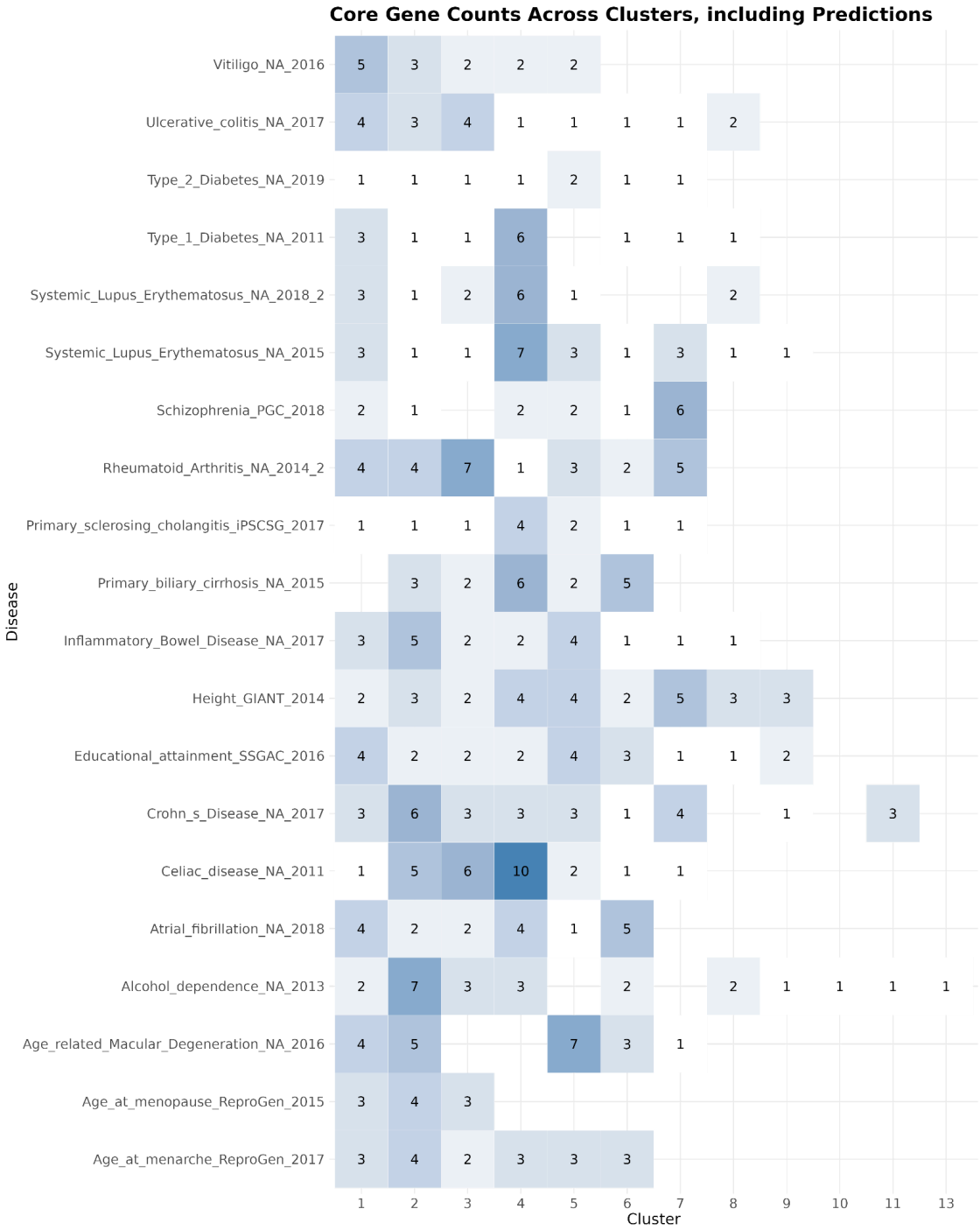

Figure 4.12: Counting the number of core genes across modules for each GWAS of training and prediction genes. Module models predict genes that are core as well. Relative to training, more modules have at least one core gene after including GenePlexuspredictions, showing GenePlexus finds genes in the network important for modules.

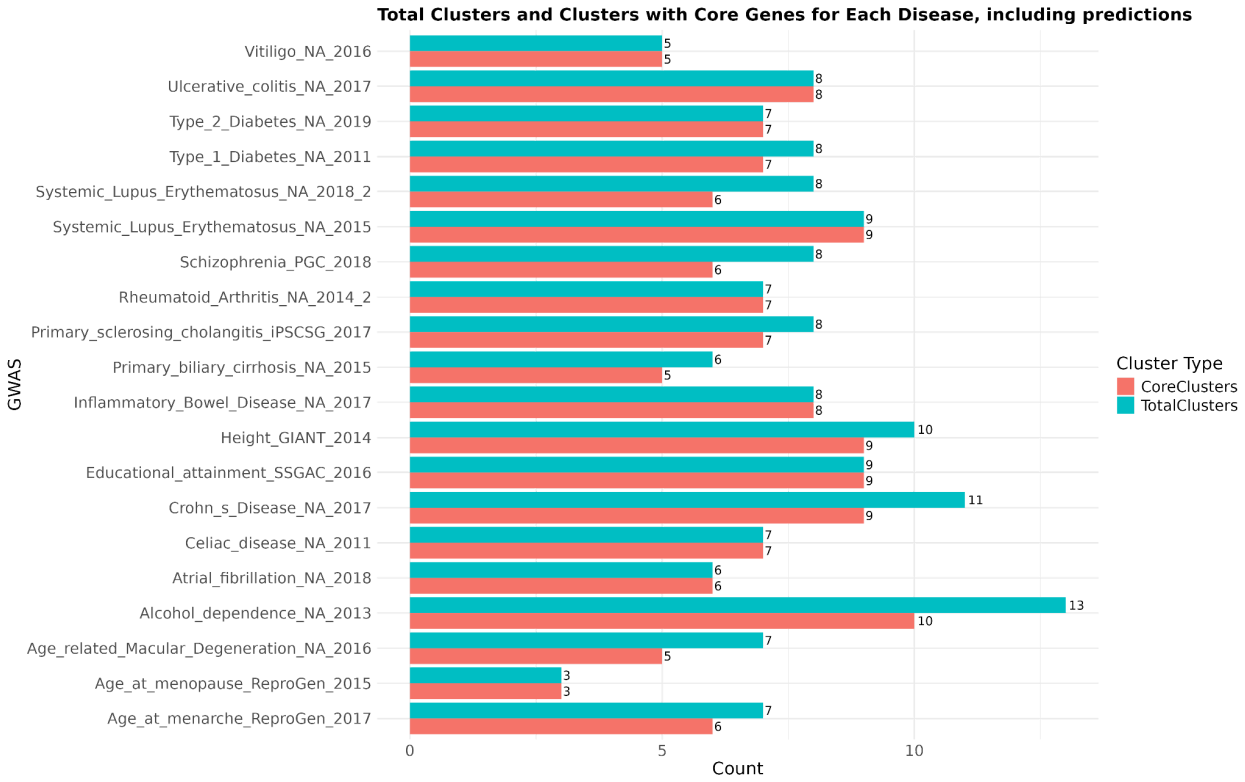

Figure 4.13: Bar plots showing the number of modules with core genes when including predicted core genes. Core genes are in multiple modules for each GWAS, and once predicted genes are considered most modules have at least one core gene in each GWAS.

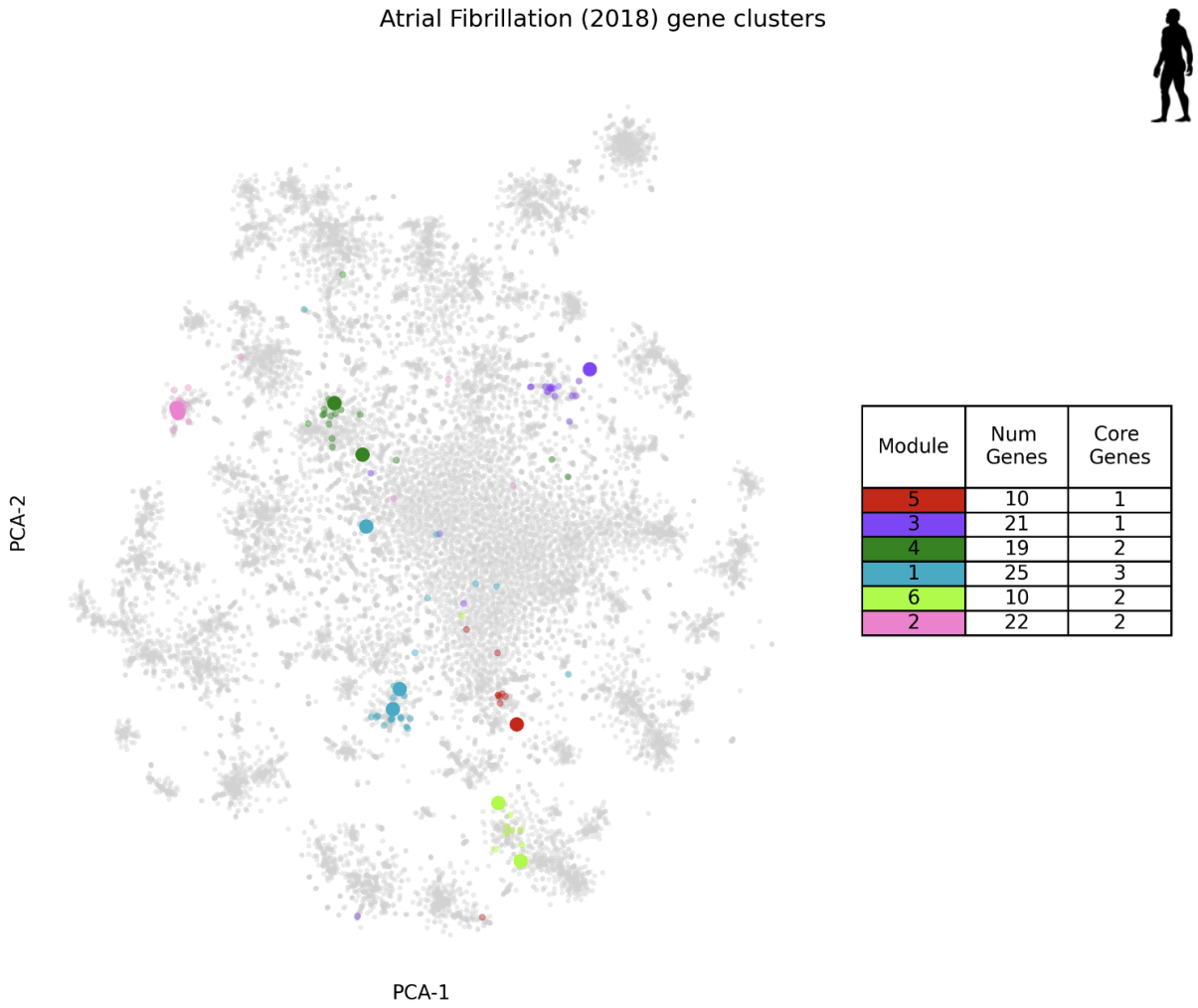

Figure 4.14: Displaying GWAS genes of atrial fibrillation within a t-SNE of the human STRING network, with each module displayed. The larger nodes are the predicted core genes, and each gene corresponds to a module (see legend). GWAS core genes are distributed across modules for a trait,and the modules with multiple core genes have those genes as neighbors in the network.

Core genes are most often module specific (but can appear as core or peripheral in other modules)

Our utilization of GenePlexusZoo means that each gene can potentially be predicted in each module. We have seen that when considering the original GWAS hits, those hits predicted to be core are distributed across modules. In the context of our predicted disease module, we asked if the core genes are predicted by multiple GenePlexus modules, i.e. have a Z > 5 for multiple modules. For most GWAS, a relatively small

number of genes are predicted by multiple modules (**Figure 4.15)**. Ten of the twenty GWAS – type 1 diabetes, both systemic lupus erythematosus GWAS, rheumatoid arthritis, primary sclerosing cholangitis, primary biliary cirrhosis, height, crohn's disease, celiac disease, and alcohol dependence – have core genes that were highly predicted by multiple modules. Fifteen GWAS had at least one peripheral gene discovered in multiple modules, all except age at menopause and menarche, age related macular degeneration, educational attainment, and schizophrenia. Gene predictions are very module specific, as seen in **Figures 4.16-17** for atrial fibrillation and height respectively. Height has the most shared genes across modules at 48, but even so the majority of the 731 genes are predicted in only one module. Overall genes are very module specific, and this provides further evidence that core genes are not one coherent set of genes but are distributed across multiple important processes that underlie complex traits and disease.

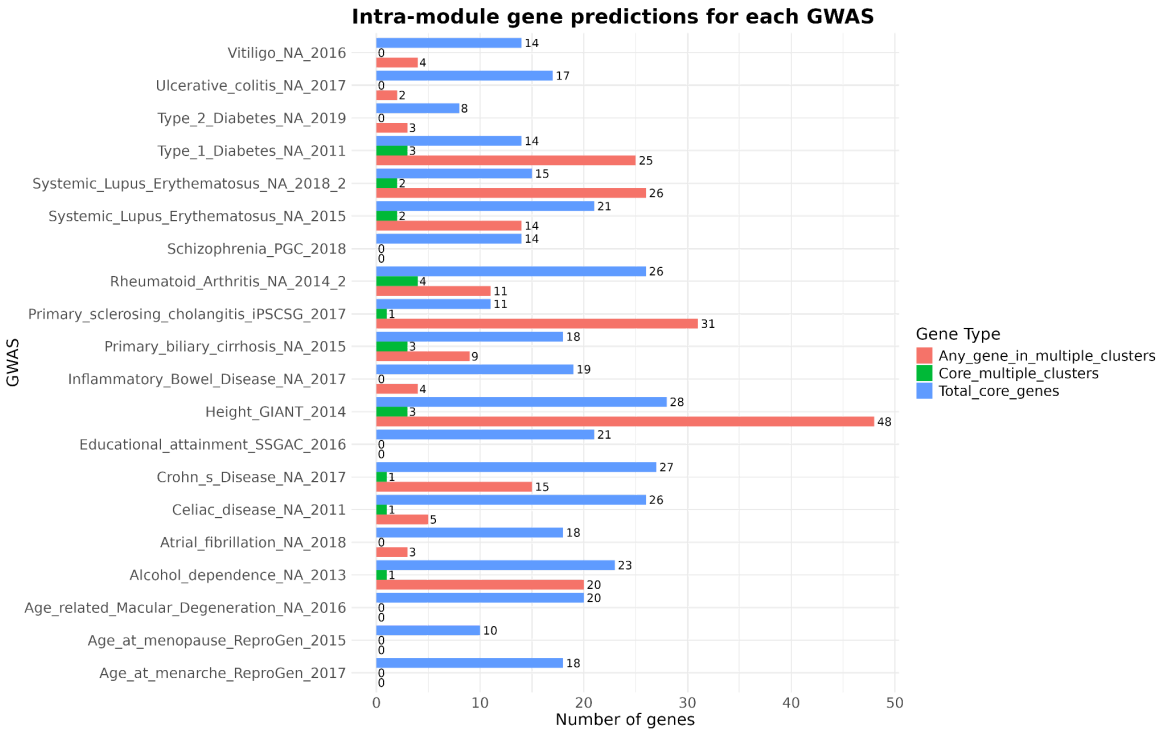

Figure 4.15: Disease gene predictions across modules for each GWAS. The blue bars refer to the total number of core genes. The green bars refer to those core genes that are predicted (z-score > 5) in multiple GenePlexus models across multiple modules.

Figure 4.15 (cont'd)

Core genes are typically module-specific and will not be predicted by other modules. The red bar refers to any gene that is predicted in multiple modules, and this varies depending on the GWAS.

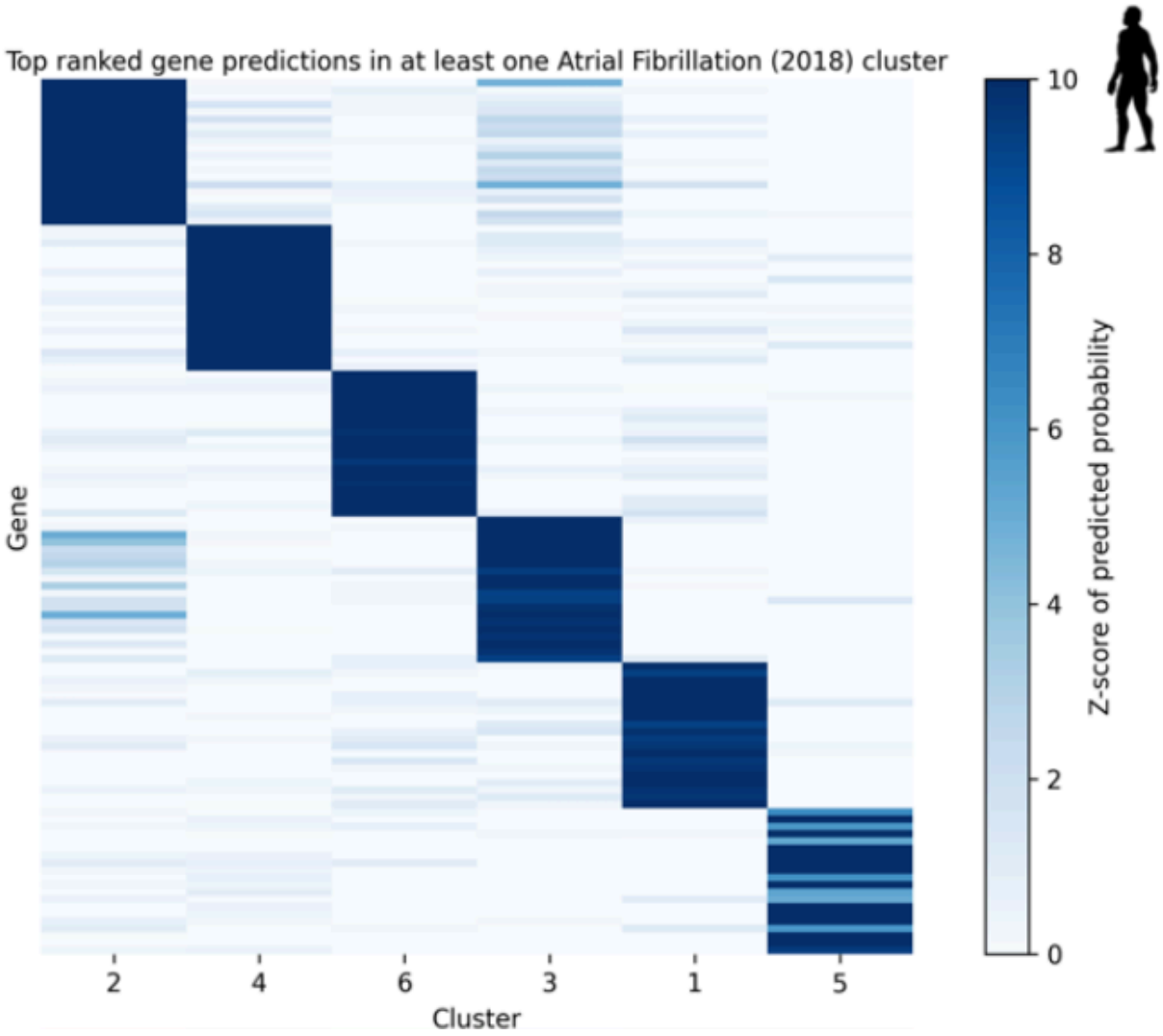

Figure 4.16: Showing the top ranked gene predictions for each atrial fibrillation modules. Gene predictions are highly clustered and most genes have a high prediction in only one module.

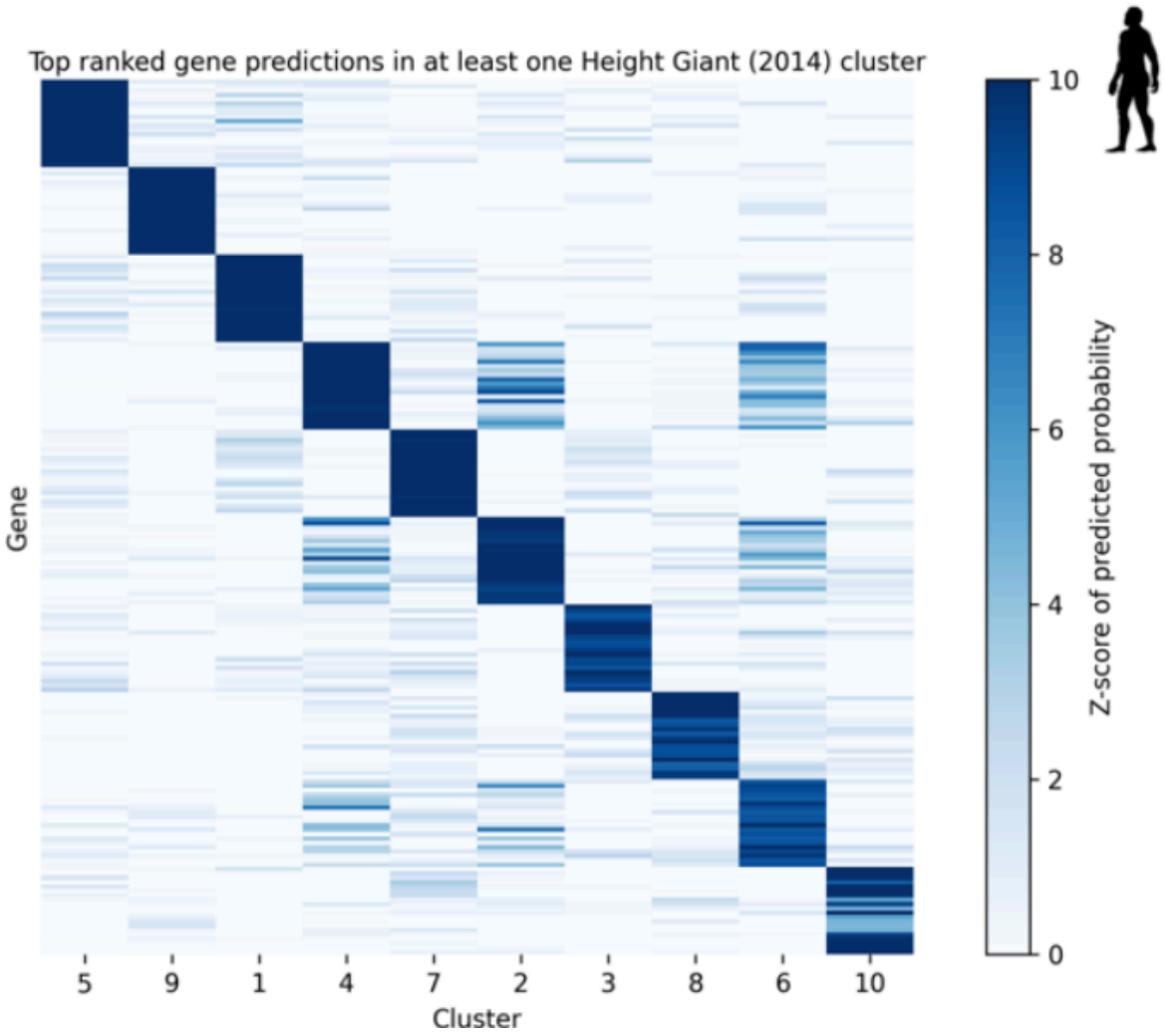

Figure 4.17: Showing the top ranked gene predictions for each height module. Gene predictions are highly clustered and most genes have a high prediction in only one module.

Predicted Core and peripheral genes are often disease specific

Every gene has a prediction score for each GWAS geneset being investigated. This makes it possible to compare gene scores of all genes across each GWAS. We first show for all genes that had a z-score > 5 for at least one GWAS, those genes z-scores for every other GWAS (**Figure 4.18**). The high gene prediction scores are highly clustered, and most genes are associated with only a single GWAS. This result is similar to the previous section where gene predictions as a whole cluster across modules. Next, we show for every gene predicted to be core in at least one GWAS the

prediction scores across all GWAS (**Figure 4.19**). Once again, there is clustering behavior here where most core genes are predicted highly in only a small number of GWAS. However, there are some core genes that have high scores in multiple GWAS. We decided to expand this analysis and see how many GWAS human core genes tend to fall in, while also showing how that knowledge transfers across species using orthologs between the human core genes and other species predicted genes.

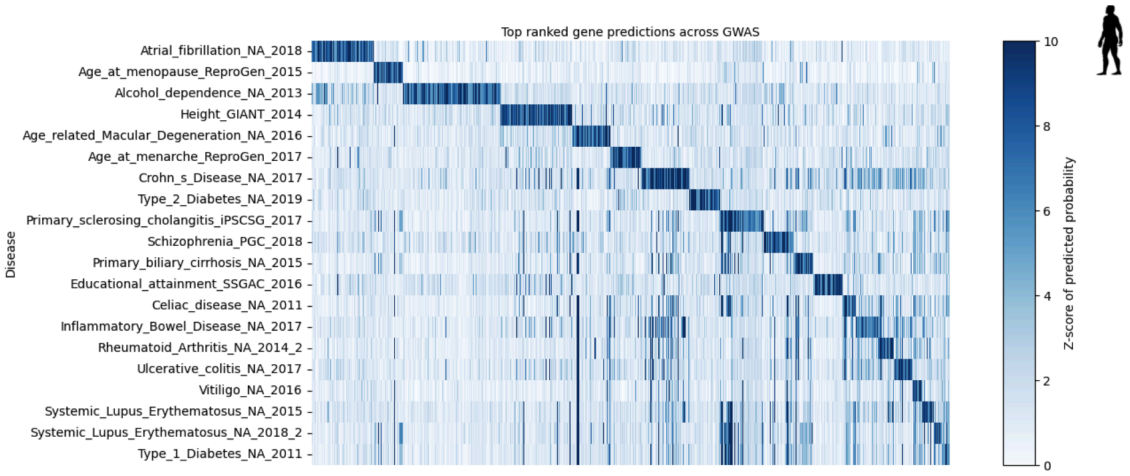

Figure 4.18: Plotting the prediction scores for all human genes that had a z-score>5 for at least one GWAS. Most genes are predicted highly in only one or a small number of GWAS.

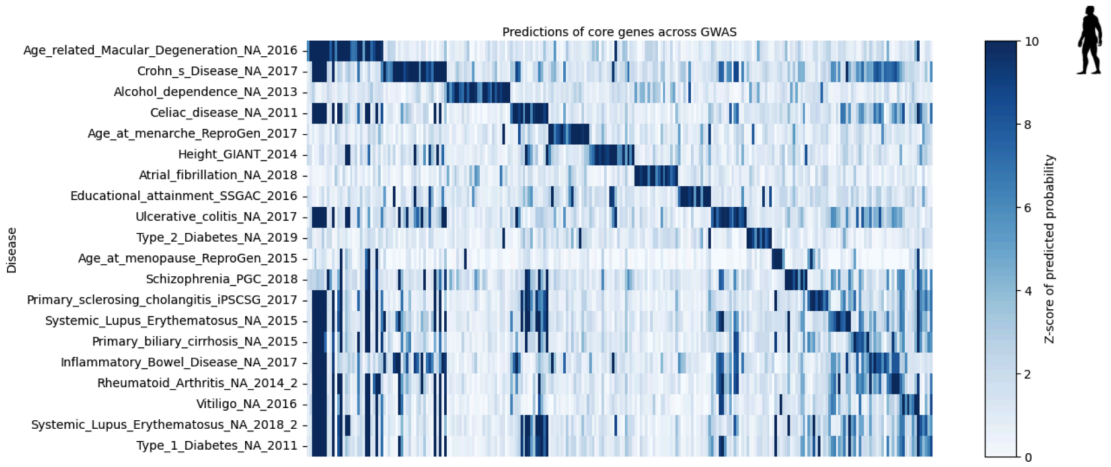

Figure 4.19: Plotting the prediction scores for genes that were predicted to be core for at least one GWAS. The core genes tend to be predicted in only one GWAS, with a minority being implicated in multiple GWAS.

Discovering orthologs across species

GenePlexusZoo allows users to make predictions for any species within the network, no matter the original species of the input. With this capability, we wanted to investigate how core and peripheral gene predictions would transfer to other species. To interpret the highly predicted genes of other species, we used one-to-one ortholog data from BioMart (**see methods)**. One-to-one orthologs have an advantage due to ease of interpretation. **Figure 4.20** shows how many one-to-one orthologs of human genes were discovered using the entire human genome, the human genes from the predicted disease module for each GWAS, and the predicted human core genes. As would be expected, zebrafish, mouse, have the most orthologs to human genes, with 7652, 16064, total orthologs respectively. Yeast and worm have noticeably less orthologs at only 722 and 1530 respectively. All predicted genes across all GWAS were mapped to orthologs, where for zebrafish there are 1826 orthologs, for yeast 139 orthologs, for worm 278 orthologs, and mouse 3866 orthologs. The relatively small minority of human core genes in these GWAS also contain orthologs to model organisms. Across all GWAS, we found for zebrafish 86 human core genes orthologs, for yeast 15 orthologs, for worm 20 orthologs, and for mouse 197 orthologs. This data is essential for interpreting how human core genes are relevant within model organisms.

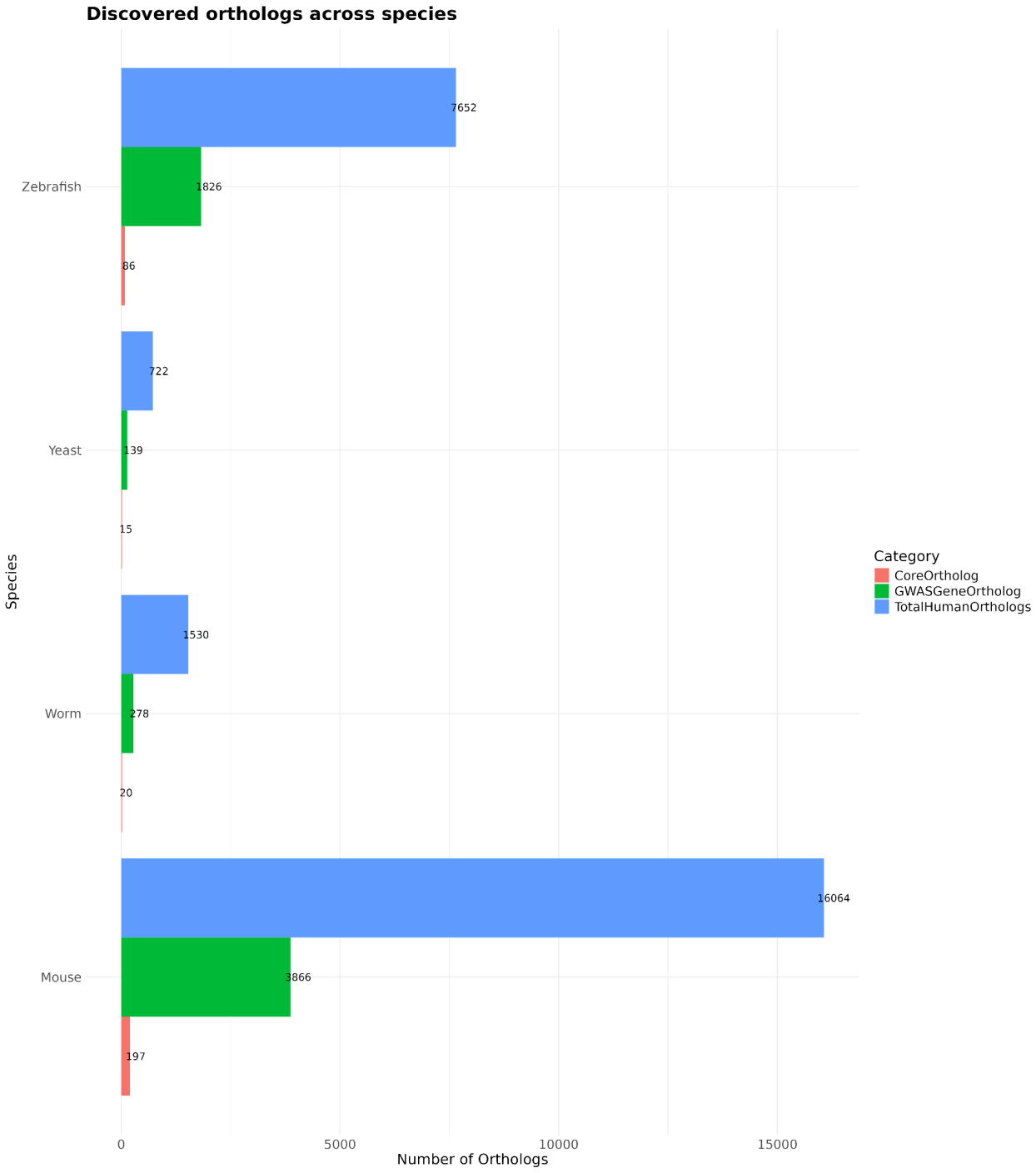

Figure 4.20: Displaying the number of orthologs discovered for each species. The total number of one-to-one orthologs discovered for all human genes (blue), the number of orthologs for genes implicated in GWAS for all diseases (green), and the number of orthologs for human core genes (red).

Core genes are GWAS-specific across species

It is interesting to know if core genes are likely to be seen across GWAS. Ideally core genes are those that are mechanistically important in a specific disease context. It is unclear whether we should expect core genes to appear across multiple diseases, so we wanted to test for two different things. First, we see for the human core genes, how many GWAS they are core in. Second, we map these human core genes to other species using orthologs. We then see how often the human core gene's ortholog is core within that species. These results are displayed in **Figure 4.21,** where each histogram corresponds to each species (human, mouse, zebrafish, worm, yeast). We see that most human core genes are only core in a singular GWAS, the GWAS it is core in. However, some core genes are implicated across multiple GWAS – where they can appear as core in up to 8 GWAS for human and zebrafish, and 7 for mouse. Next, since a gene is defined as core in the context of specific diseases, it is possible that it can be truly annotated to other diseases, but be peripheral instead. **Figure 4.22** displays these results, and we see that core genes are involved in many more GWAS if we consider their peripheral status as well. Some of the core genes appear in nearly all of the GWAS for humans in 19 out of the total 20 GWAS, and appear in other 10 GWAS as orthologs in mouse and zebrafish. We also predicted the core genes in each species (**see methods**) and show how they are distributed across the GWAS. The same trend in humans is seen in the other species; most species' core genes are core in a single GWAS (**Figure 4.23**). Lastly, we demonstrate that core genes are at higher incidence for being implicated across multiple GWAS (**Figure 4.24**) relative to peripheral genes. These results show that while most genes are only core in specific contexts, they can appear in other GWAS both as core genes and as peripheral and are more likely to appear across multiple GWAS than peripheral genes.

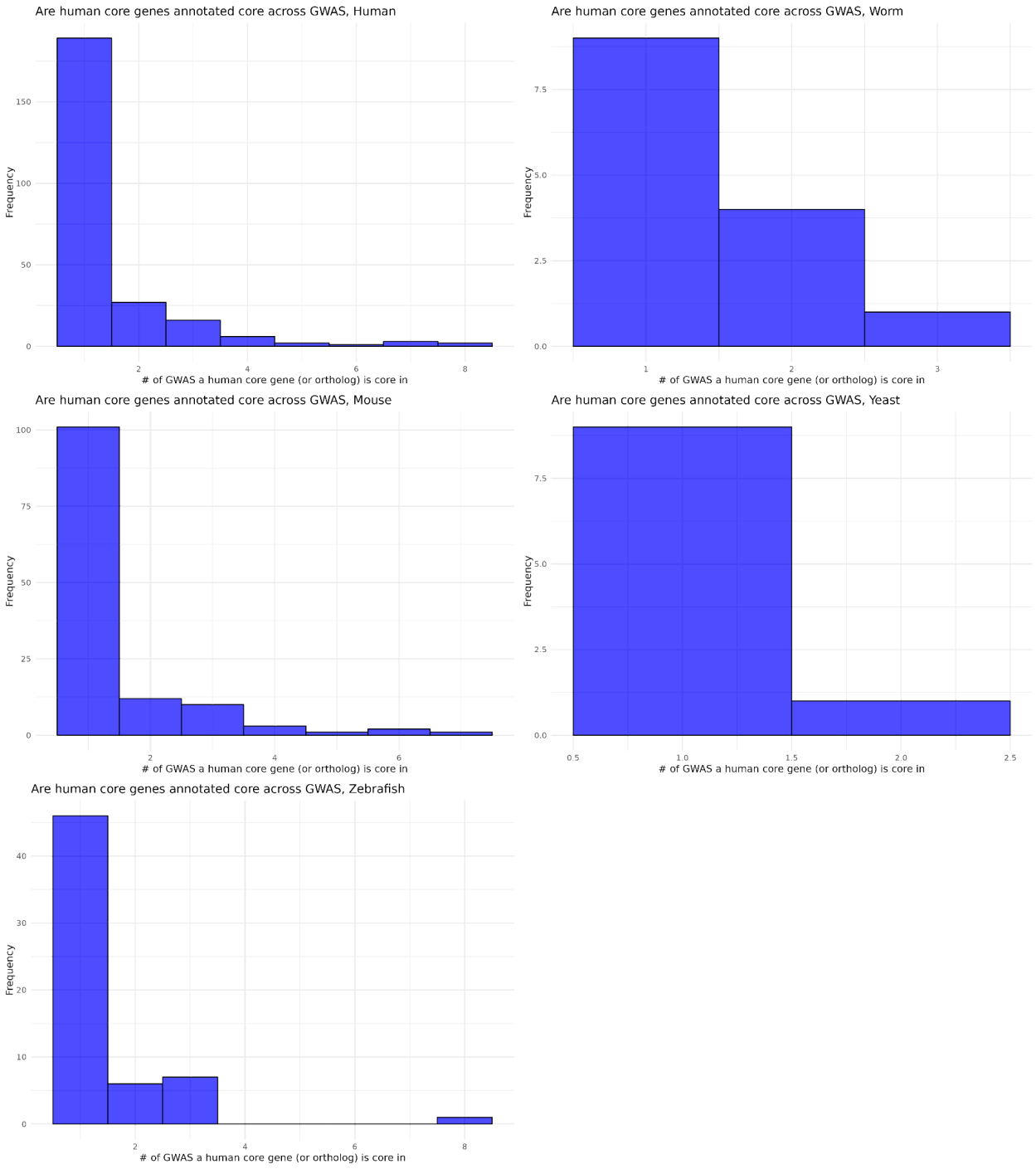

Figure 4.21: Histograms showing whether human core genes – or their respective orthologs – are labeled as core across GWAS for each species. For most species, the human core genes/orthologs are seen in only one GWAS. However, worm and yeast rarely has implicated human core gene orthologs predicted.

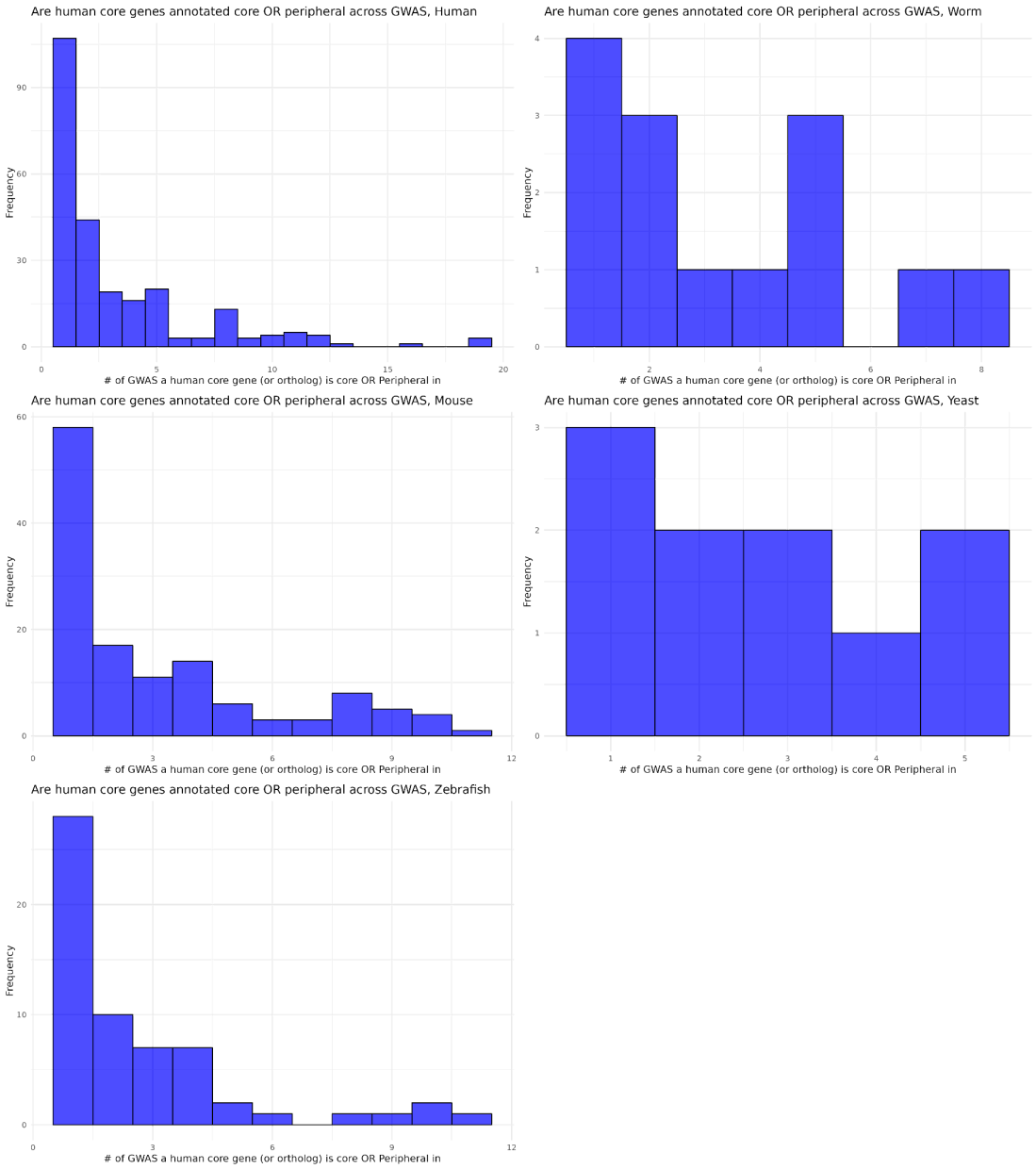

Figure 4.22: Histograms showing whether human core genes – or their respective orthologs – are labeled as core or peripheral across GWAS for each species. A core gene can be peripheral in other traits because it is discovered to be in the disease module, but is not core in both GWAS. Core genes are more likely to appear cross-GWAS as peripheral.

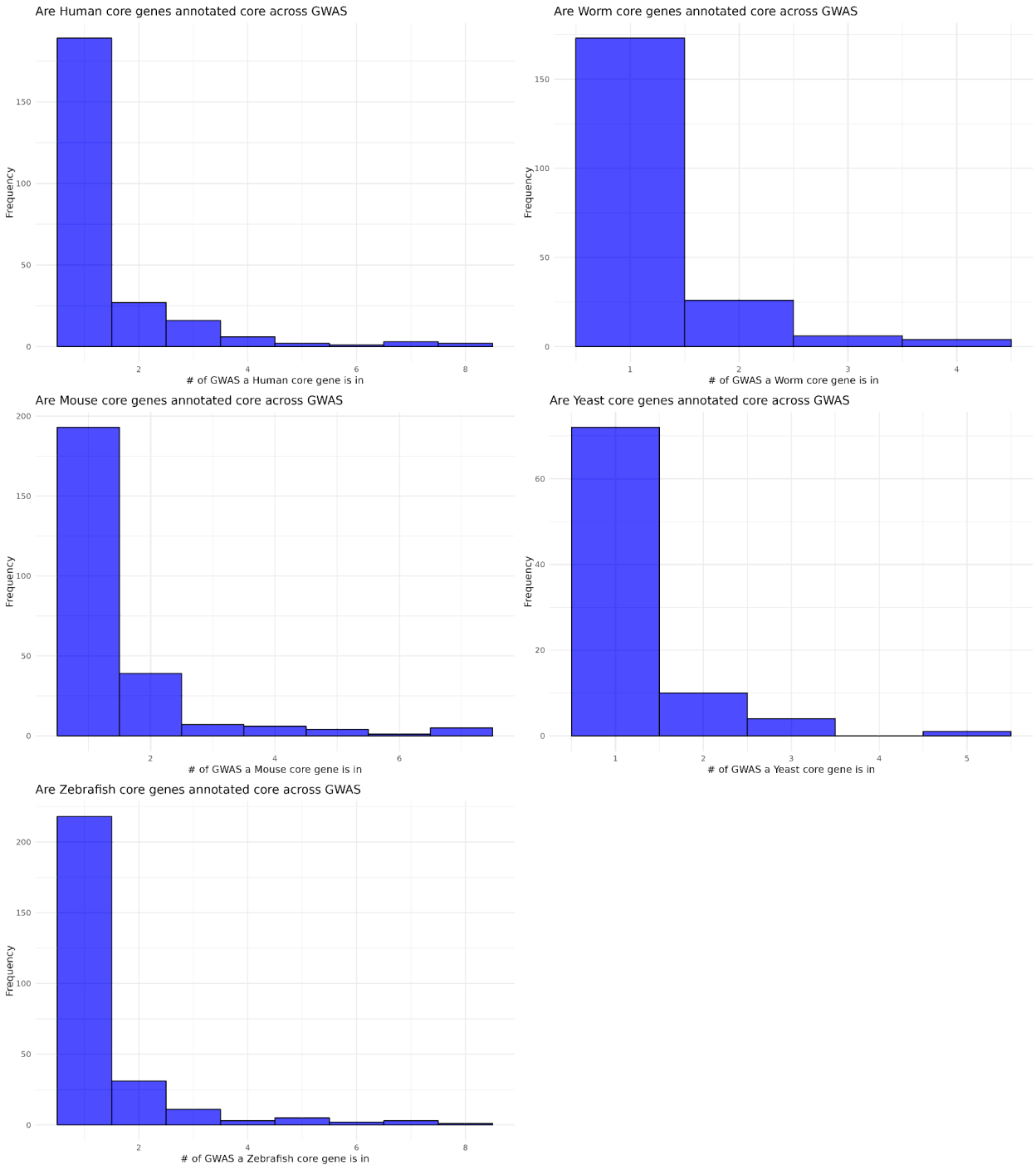

Figure 4.23: Histograms showing whether a species's core genes are labeled as core across GWAS for each species. For other species, the same pattern appears as in human core genes where most core genes are disease-specific.

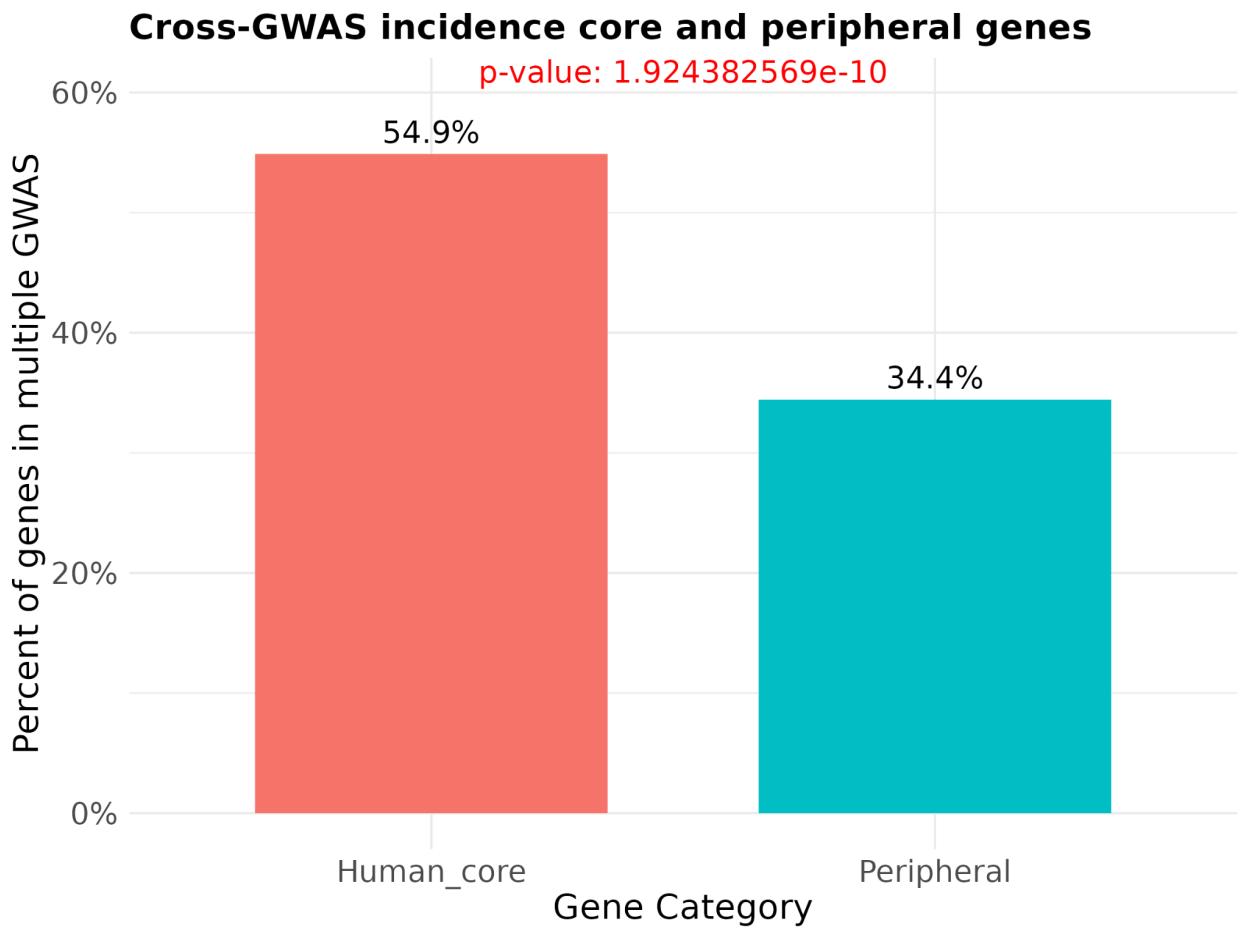

Figure 4.24: Bar chart showing percent of core and peripheral genes that are implicated in multiple GWAS. Human core genes are more likely to be seen in multiple GWAS than peripheral genes are.

Modules give distinct gene and phenotype predictions

We have already demonstrated that core and peripheral genes are likely to be module and disease specific within humans. **Figure 4.25** contrasts human, and mouse genes predicted to be associated with the human GWAS genes using GenePlexusZoo. Both species show very modular behavior for the top predicted genes. The same is true in Zebrafish (**Figure 4.26)**, worm and yeast (**Figure A4.6**). **Figure 4.27** compares model weights of the embedding feature vectors for human and mouse by looking at the top feature vectors across modules, where the top 10 feature vectors by absolute value for each module are displayed. We can see there is modular behavior here as well, with some embedding vectors having noticeably large model weights for some modules but not in others. The same observation is true for zebrafish (**Figure 4.28**) and worm and

yeast (**Figure A4.7**). Lastly, we utilized GenePlexusZoo to find the top three enriched phenotypes and GOBPs in **Figure 4.29** based on these module model weights and comparing them to model weights built for the GSCs obtained from GO and Monarch (**see methods**). Firstly, there is again clustering behavior at the module level within the GWAS for atrial fibrillation when comparing and contrasting enrichment of GOBPs and phenotypes for each species. We also see multiple meaningful GOBPs and phenotypes within the modules. Using module 2 as a test case, we see it has directly implicated pathways in cardiac muscle tissue morphogenesis[54,55] and myofibril assembly[56]. Some of the Monarch phenotypes implicated in module 2 include ST-segment depression, which was implicated in a recent study where during atrial fibrilation rhythm, ST-segment depression is associated with subsequent heart failure risk for afflicted patients[57]. Septal hypertrophy was also an enriched human phenotype from Monarch, and this is a predictor for patients that start with the disease hypertrophic cardiomyopathy to eventually also have atrial fibrillation[58]. Some of the enriched GOBPS in module 2 for mice includes sex differentiation and male sex differentiation specifically. This is interesting as there are many sex differences in atrial fibrillation in terms of symptoms and severity[59,60]. A study[61] was done in mice specifically to determine if sex differences in intracellular Ca^{2+} homeostatsis in atrial myocytes might explain increased incidence of atrial fibrillation in males, and this pathway was found to have major sex differences in mice. It's also involved in cell types like myocytes also implicated in atrial fibrillation[62]. Overall we see multiple meaningful predicted human and mouse GOBPs for atrial fibrillation within this module, and for the monarch human phenotypes that are backed up well in the literature. The GOBP and phenotype enrichment comparisons in zebrafish, worm, and yeast are in **Figures A4.8-10.**

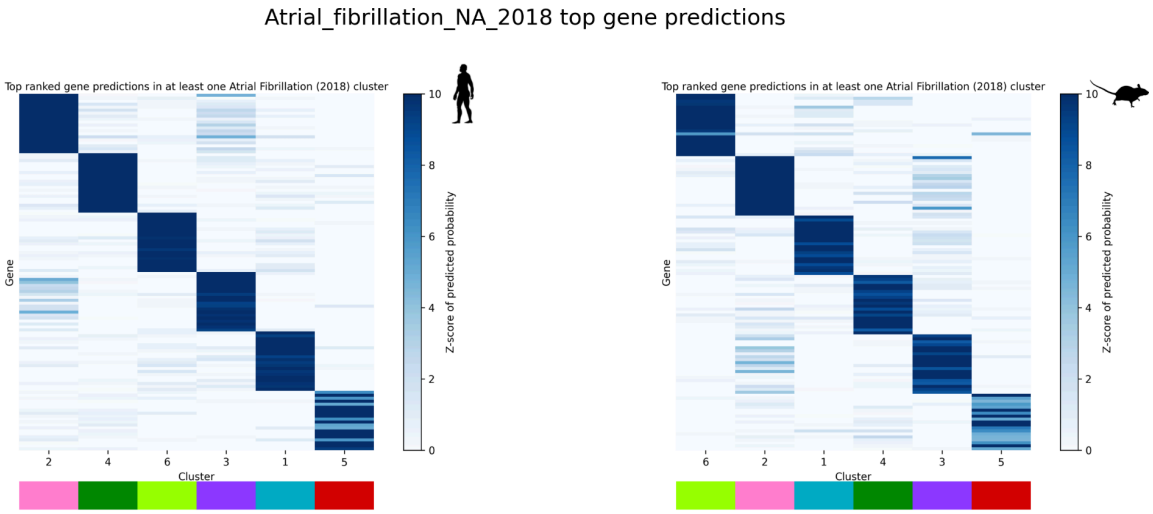

Figure 4.25: Heatmaps showing highly predicted genes in human and mouse across modules for atrial fibrillation. The genes are not the same across heatmaps, as the one on the left are human gene predictions across modules and the one on the right are mouse gene predictions for the modules of atrial fibrillation. The gene scores are highly clustered across modules for both species.

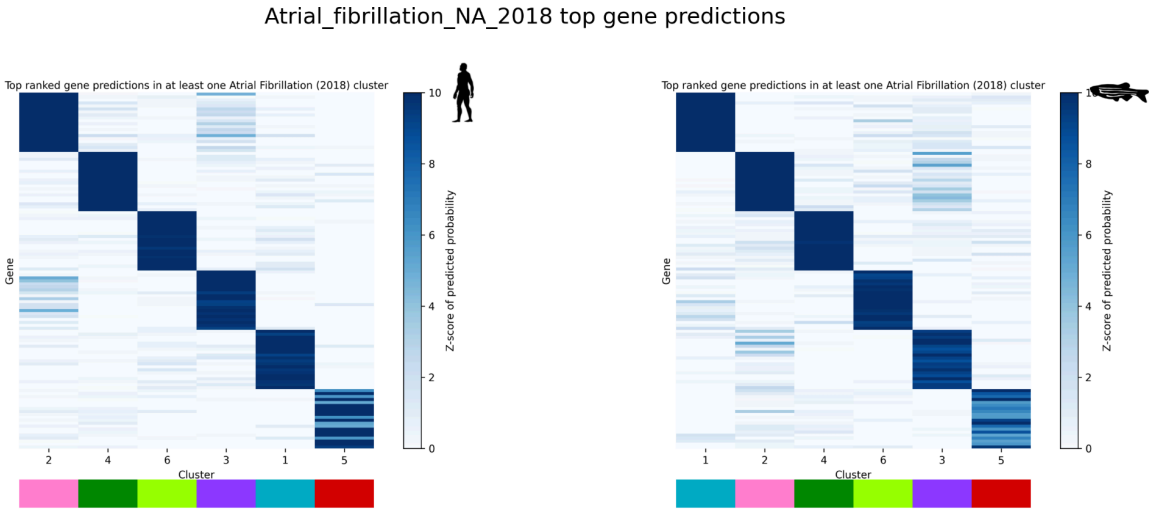

Figure 4.26: Heatmaps showing highly predicted genes in human and zebrafish across modules for atrial fibrillation. The genes are not the same across heatmaps, as the one on the left are human gene predictions across modules and the one on the right are zebrafish gene predictions for the modules of atrial fibrillation. The gene scores

Figure 4.26 (cont'd)

are highly clustered across modules for both species. The color annotation bar for the heatmaps clarify the module order.

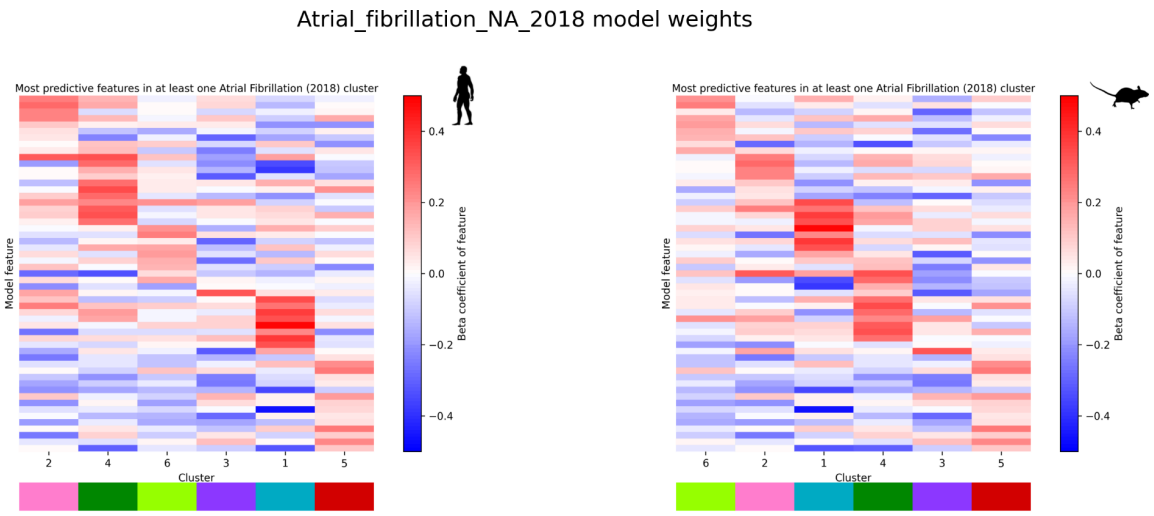

Figure 4.27: Heatmaps showing model weights of feature embedding dimensions in human and mouse across modules for atrial fibrillation. There is clustering behavior between the feature vectors and the modules within the GWAS. Meaning that feature embedding dimension vectors are distinct across modules for both human and mouse. The color annotation bar for the heatmaps clarify the module order.

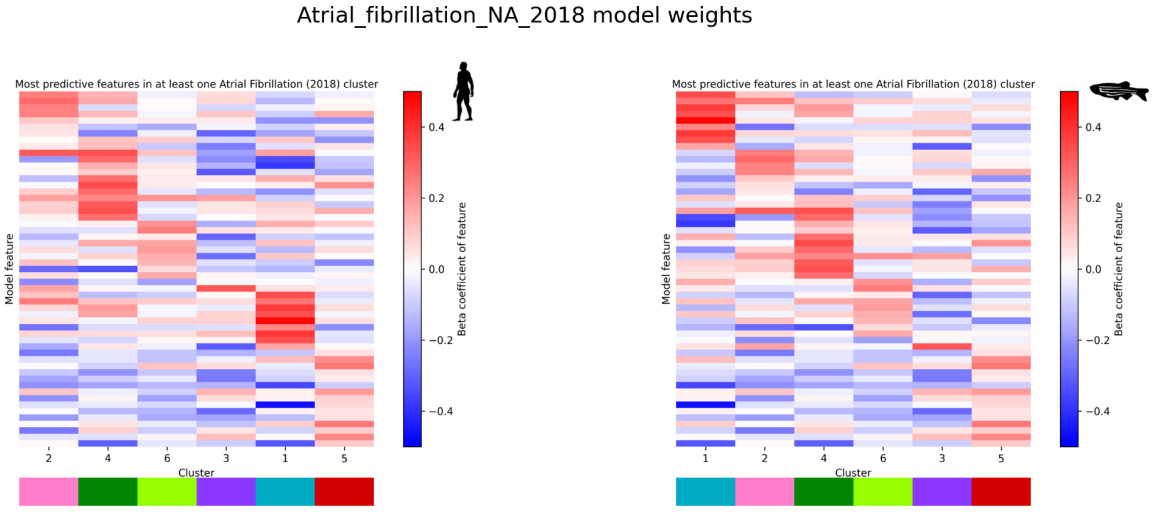

Figure 4.28: Heatmaps showing model weights of feature embedding dimensions in human and zebrafish across modules for atrial fibrillation. There is clustering behavior between the feature vectors and the modules within the GWAS. Meaning that feature embedding dimension vectors are distinct across modules for both human and zebrafish. The color annotation bar for the heatmaps clarify the module order.

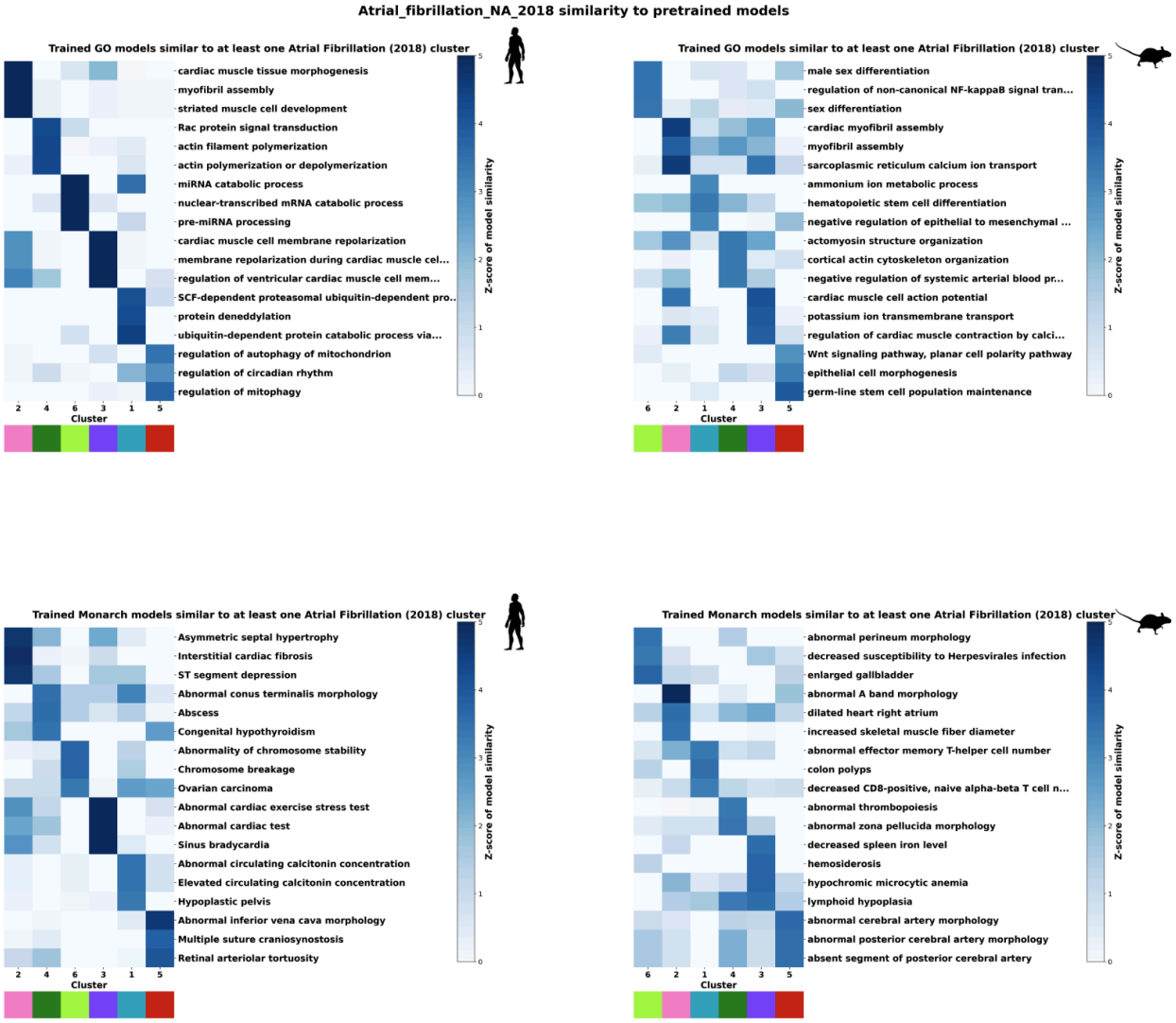

Figure 4.29: Comparing human GOBP and Monarch phenotype predictions across modules with mouse GOBP and phenotype predictions. There is clustering behavior for both species between GOBP and phenotype predictions, and we see multiple relevant GOBPs for atrial fibrillation in humans and mouse.

Analyzing core gene and significant GWAS gene biology

Significant GWAS loci tend to be near genes that are mechanistically important[39] and that have functional annotations using many diverse methods and data mentioned in chapter 1. Additionally, GWAS has revealed loci that are in highly conserved genomic regions[63,64] that are near human orthologs[65], with evidence suggesting GWAS finds them at higher rates when compared to e-QTL studies[66]. This implies that non-coding

regions implicated in GWAS are more likely to have functional annotation than specific e-QTL studies. In Pritchard et. al[39], they indeed show that the top hits of three GWAS complex traits do correspond to “known” core genes – genes involved in known implicated pathways for the traits. Given that GWAS results are enriched with mechanistically important genes, we contrast comparing the significant results discovered from MAGMA to nominal p-values, with our core genes to peripheral genes, if GWAS experimental results or our predicted disease module are enriched with biological statistics.

We tested if the core genes are more enriched for genes in terms of their age, constraint, in the amount of one-to-one orthologs contained in **Figures 4.30-31**. We decided to test for two sets of core genes in this analysis, one being the usual definition of core genes where the z-score of the betweenness centrality (BC) values is above 2, and one where it is simply positive (Z-score > 0). In both sets we see a notable number of GWAS (9 and 13 sets respectively) that have a significant number of constrained genes relative to the small betweenness centrality values. In contrast, only 4 of the GWAS significant genes are enriched for constrained genes relative to nominal p-values. For comparing the number of orthologs, the set where the z-score of BC values above 0 has many sets enriched for containing orthologs, implying that orthologs heavily populate those genes with high network connections in the disease module. Only 1 GWAS set had its significant genes enriched for orthologs, and only 1 of the normal core gene set (Z-score > 2) did.

There is rarely enrichment for age in either the GWAS significant genes or in either core gene set. It is unclear whether it should be expected that significant GWAS genes or mechanistically important genes are more likely to be old or young. It is commonly assumed that essential genes are more likely to be old because they correspond to conserved processes across species. However, there is conflicting evidence that young genes can quickly become essential in important pathways[67,68] and that they are essential at the same rate as old genes in knockout studies[69]. Additionally, just because a gene is old does not mean that function has been conserved – as functional divergence has been observed over time[70]. We lastly tested if core genes are more likely to have high tissue expression using z-scores obtained from GTEx (**see**

methods). We found that while some GWAS significant gene sets were enriched for tissue specific genes, the core genes were not. This result was surprising, so we expanded to using multiple metrics to determine tissue specificity (**Figure 4.32**) to see if any core gene sets were tissue specific. Using the tau scores and average z-score, there are some sets that have tissue specific genes, but only for 2 or 3 GWAS depending on the metric.These results show that using network-defined core genes rather than only top GWAS results better represents some expected properties of the mechanistically important genes within a trait – particularly in their selection pressure and correspondence to orthologs in other species. It is worth investigating further tissue and cell enrichments to see if there is any meaning within the modules as a whole.

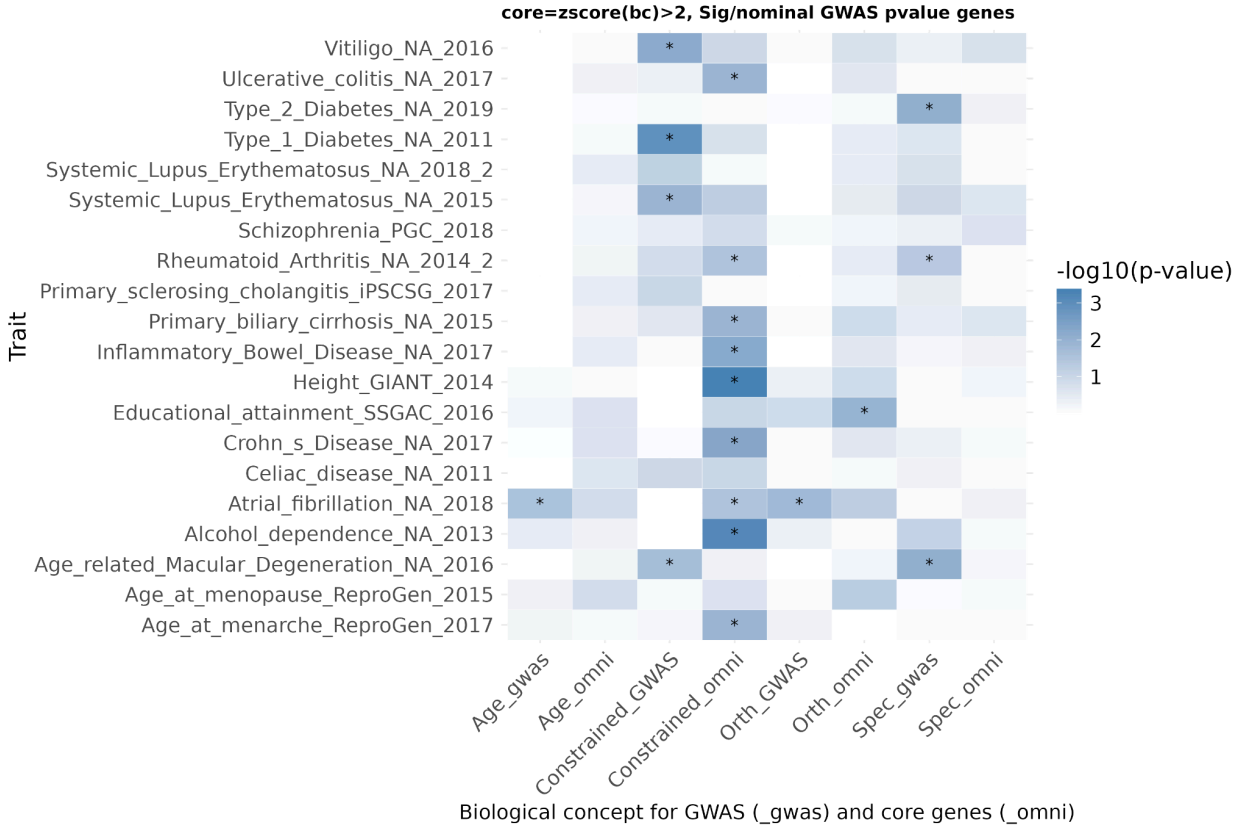

Figure 4.30: Heatmap showing enrichment of predicted core genes – defined here as the zscore of the betweenness centrality (bc) values being above 2 – contrasted with peripheral genes, and significant GWAS genes contrasted with non-significant GWAS genes that were predicted to be in the module. For disease genes, we show whether core genes and significant MAGMA GWAS hits are enriched

Figure 4.30 (cont'd)

for biological annotations of gene age, conservation, the number of 1-1 orthologs to other species, and tissue specificity.

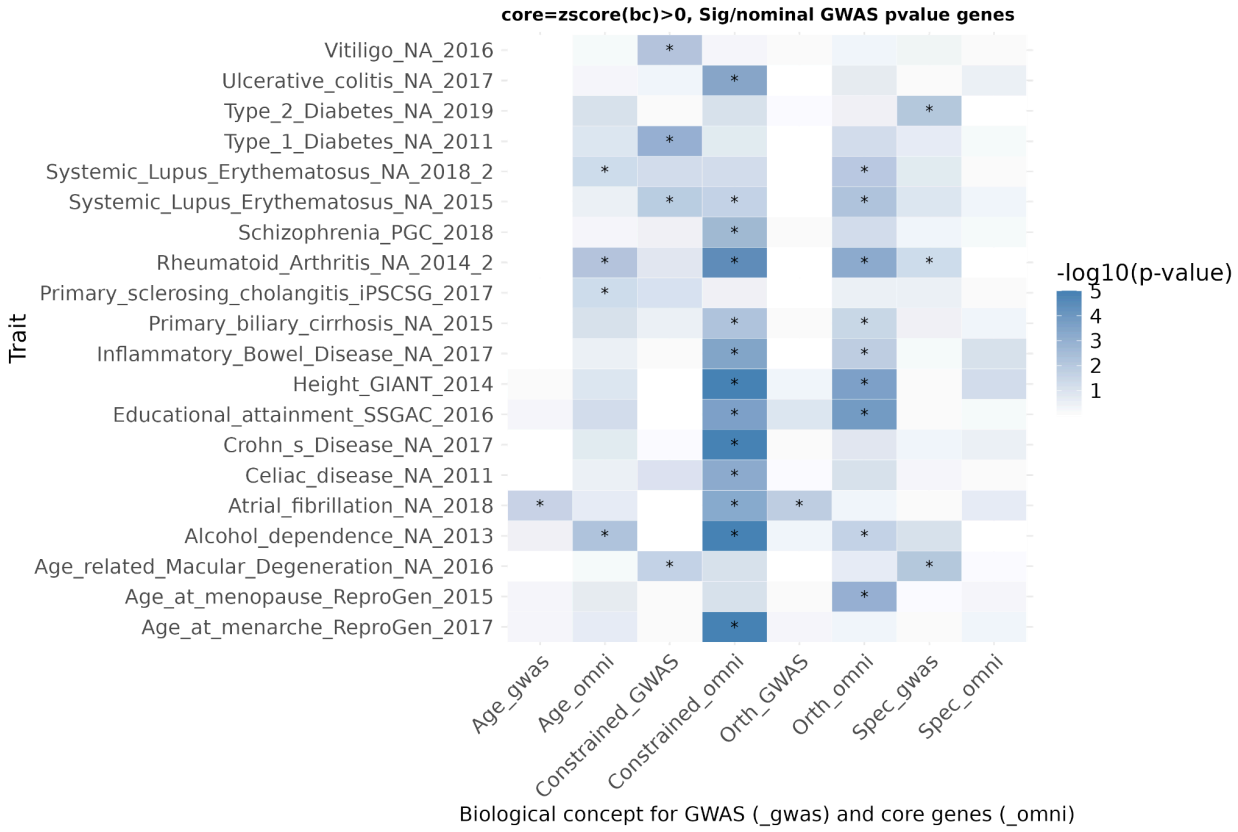

Figure 4.31: Heatmap showing enrichment of genes with a z-score of betweenness centrality values (bc) above 0, contrasted with genes with a negative z-score. For disease genes, we show whether core genes significant MAGMA GWAS hits are enriched for biological annotations of gene age, conservation, the number of 1-1 orthologs to other species, and tissue specificity.

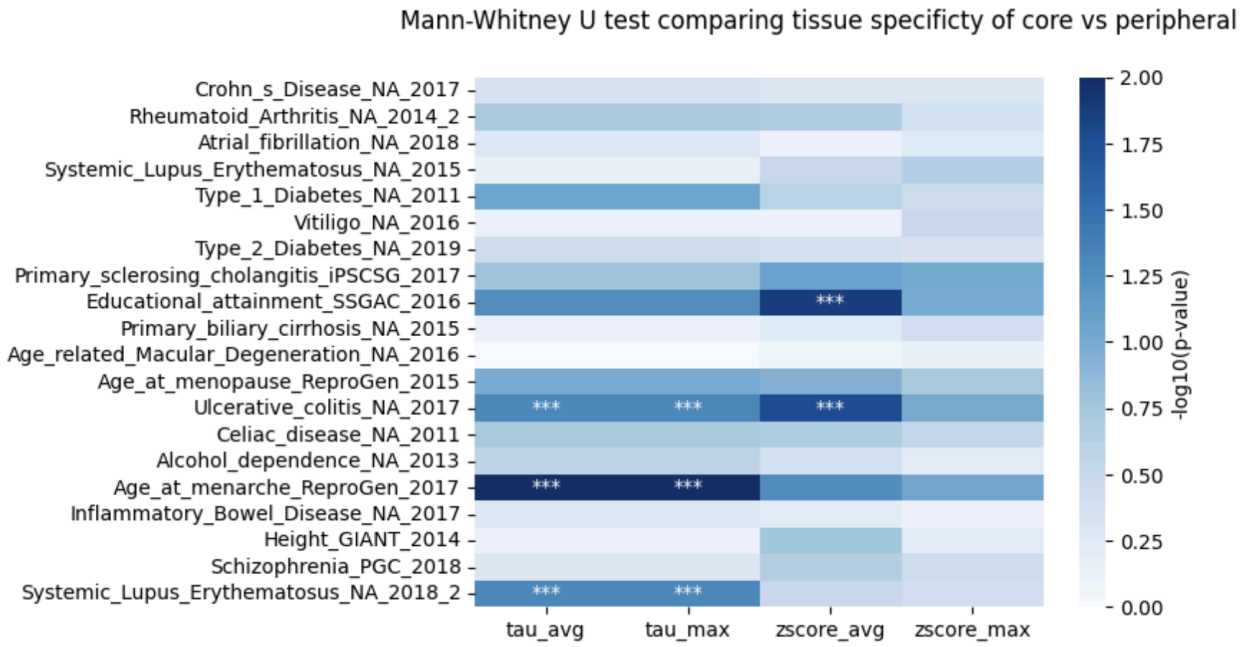

Figure 4.32: Heatmap displaying enrichment of core genes for tissue specific genes for each GWAS. The values are -log10(p-value) of the results of a Mann-Whitney U test comparing the tissue specificity scores of core vs peripheral genes. *** indicates a $pval < 0.05$. Tau score and z-score were taken from GTEX for every gene/tissue pair. Both a maximum and averaging aggregation strategy were used when determining a final tissue specificity score for a gene. Each column uses a different metric to determine if core genes are more tissue specific than peripheral genes for a given GWAS.

Enriched tissues across modules for atrial fibrillation shows module-specific enrichment

We next ran enrichment for tissues and cell types from Jensen TISSUES and CellSTAR databases **(see methods)**. This was done for each module. For atrial fibrillation specifically, we see in **Figure 4.33** that there is a mixture of module specificity and shared tissues/cells across modules from the TISSUES database. Modules 2 and 3 share enrichment results for tissues related to muscle, cardiovascular systems, and heart, vital systems for explaining atrial fibrillation manifestation and phenotypes[71]. Module 4 is uniquely enriched for lymphocytes and leukocytes and inflammation is known to be relevant to atrial fibrillation[62,72,73]. Leukocytes are immune system cells that are a predisposing factor for atrial fibrillation[74], where an increased cell count means

there is inflammation which affects atrial fibrillation severity. The neutrophil-lymphocyte ratio is an inflammatory biomarker that has been used in predicting outcomes for atrial fibrillation[75–78]. Module 4 is additionally enriched for adipose and fat tissues, and specific fat tissues such as epicardial fat – a visceral fat that is near the heart – have been implicated in atrial fibrillation manifestation[79–81]. For enrichment with CellStar in **Figure 4.34**, we see less clustering behavior across modules but still have meaningful biological results, which is not surprising given that cell types can often appear in related but distinct tissues[82]. Module 2 relates to two different types of cardiac muscle cells (cardiac muscle cell and CL:0000746 both are types of cardiac muscle cells), along with smooth muscle cells and myocytes. Myocytes are cells in muscles, and myocytes such as cardiomyocytes and other atrial myocytes are implicated in atrial fibrillation[83,84]. Multiple of the modules are enriched for relevant tissue, cell, and organ systems for atrial fibrillation.

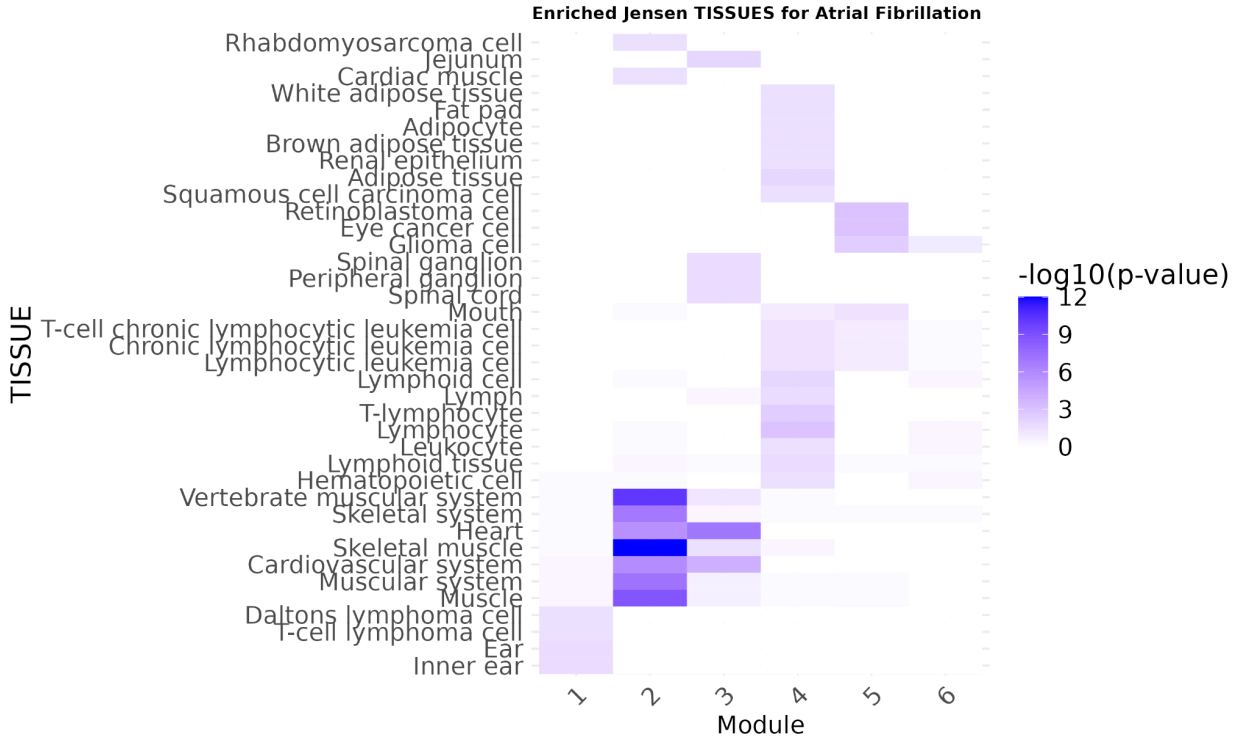

Figure 4.33: Enrichment of atrial fibrillation modules for the Jensen TISSUE database. There is clustering behavior for the modules of enrichment for the TISSUE database.

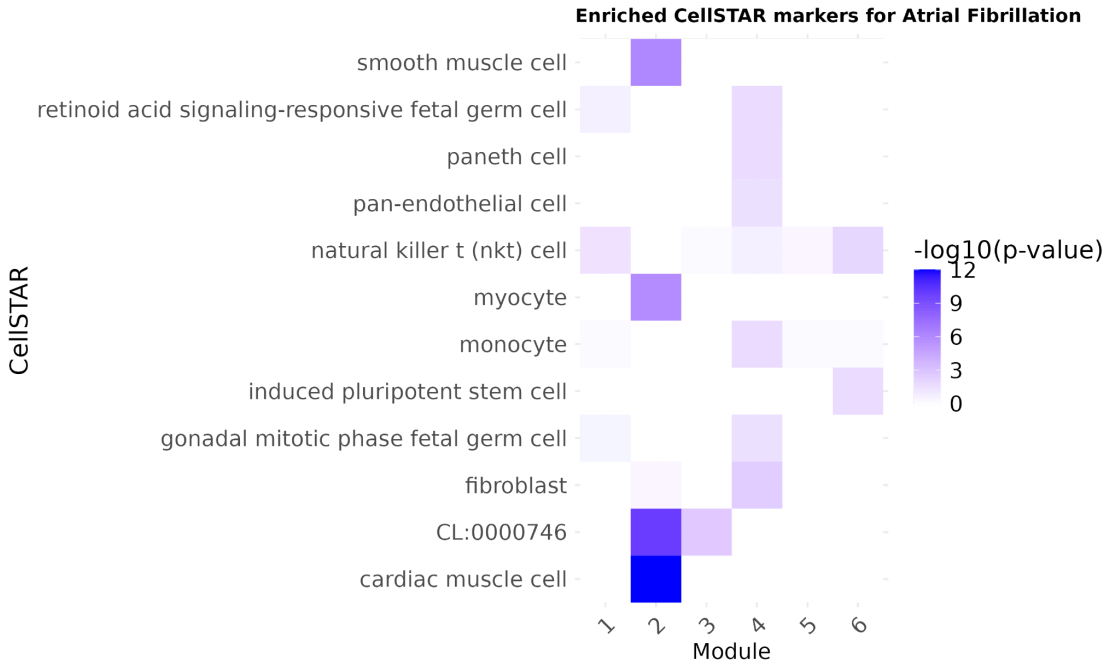

Figure 4.34: Enrichment of atrial fibrillation modules for the CellStar database. Enrichment was performed for the module genesets and cell marker genes.

One-to-one orthologs of human core genes in other species reveals biologically meaningful genes at the human and model organism levels

We have seen in previous sections that some diseases contain a relatively large number of orthologs to human core genes in some species (mouse, zebrafish) compared to others (worm, yeast). A reason this could be the case is that model organisms are not relevant to all parts of a human complex disease, but to only certain biological concepts like molecular pathways that are involved. We investigated whether human core genes are predicted as orthologs in model organisms. In **Figure 4.35-36**, we see the number of atrial fibrillation human core genes whos' orthologs are either core or peripheral in each model organism. There are a total of 13 unique of the total 18 (**Figure 4.9**) human core genes of atrial fibrillation that are orthologs in other species. Some genes only appear a single time, and others appear across multiple species (**Figure 4.37**). **Figure 4.38** shows from what source – the initial GWAS, propagation, or from GenePlexus – the core genes were discovered from. A recent GWAS study[45] predicted candidate genes of TTN, CFL2, CASQ2, FBXO32, and FXR2 for human atrial fibrillation. We predicted these genes as human core genes, and in addition their

orthologs are calculated to have a high BC value among the predicted mouse genes, with the exception of FXR2 which is core in human but it's ortholog is peripheral among mouse genes. FBXO32 and CASQ2 orthologs are core in Zebrafish, while CFL2 and FXR2 orthologs are peripheral. Multiple other studies implicate these genes[85–88] and have been used in animal models for heart diseases[89–91]. We predict the FXR2 orthologs as peripheral in mouse. A recent study has predicted FXR2 predicted this gene as a target of miR-10a using mouse studies, a pathway which has been implicated in atrial structure remodeling[92]. For Zebrafish, atrogin-1 (FBXO32) has been implicated directly with heart failure when deficient[93]. This observation is important as atrogin-1 is a muscle-specific E3 ubiquitin ligase that is involved in protein degradation and autophagy in zebrafish. The gene SKP1 is interesting as it's orthologs were predicted as relevant in three of the model organisms – mouse, zebrafish, and yeast. SKP1 is an assembly factor of a family of E3 ligases in mammals, and has been implicated in regulating the switch between protein secretion to autophagy[94]. RYR2 is a gene predicted from GenePlexusZoo and has been implicated in multiple very recent studies for atrial fibrillation[95]. Notably, we predict this gene's ortholog as being core in mouse, and mouse studies have shown that this gene is relevant in atrial fibrillation. RyR2-mediated Ca^{2+} triggers paroxysmal atrial fibrillation. This shows that in mice, mutation is RyR2 is directly implicated in mouse atrial fibrillation. This gene is the subject of numerous studies in humans as well, showing that mutation leads to atrial fibrillation and other atrial issues[96–98]. This brief overview has shown direct evidence of the biology behind our predicted core genes for atrial fibrillation for multiple genes. These genes spanned the sources they came from, whether from the initial GWAS data or from gene classification, and have direct evidence of being involved with atrial fibrillation across multiple studies, including human GWAS and model organism experiments. In **Figures 4.39-4.42.**, we plot how the orthologs fall within the disease modules of the other species. This overview shows that network relationships within other species allow for isolation of genes that have important relationships to the disease within the organism and are orthologous to important human disease genes. Additionally, it allows the discovery of genes like RYR2 that are directly and mechanistically explainable for atrial fibrillation – which is important for defining human disease core genes.

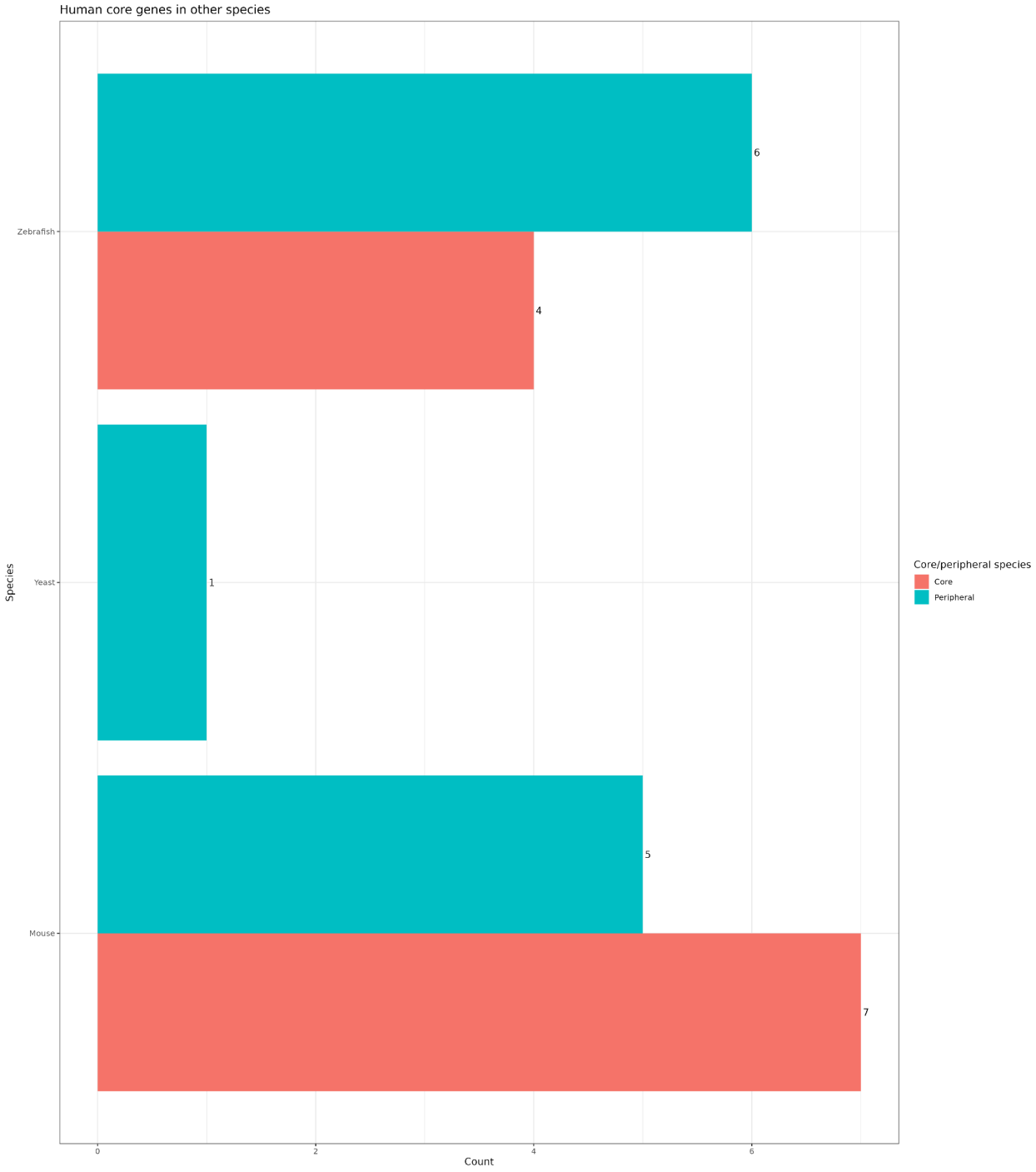

Figure 4.35: Count of how many human core genes are within model organism disease modules, and whether they are core or peripheral within that organism.

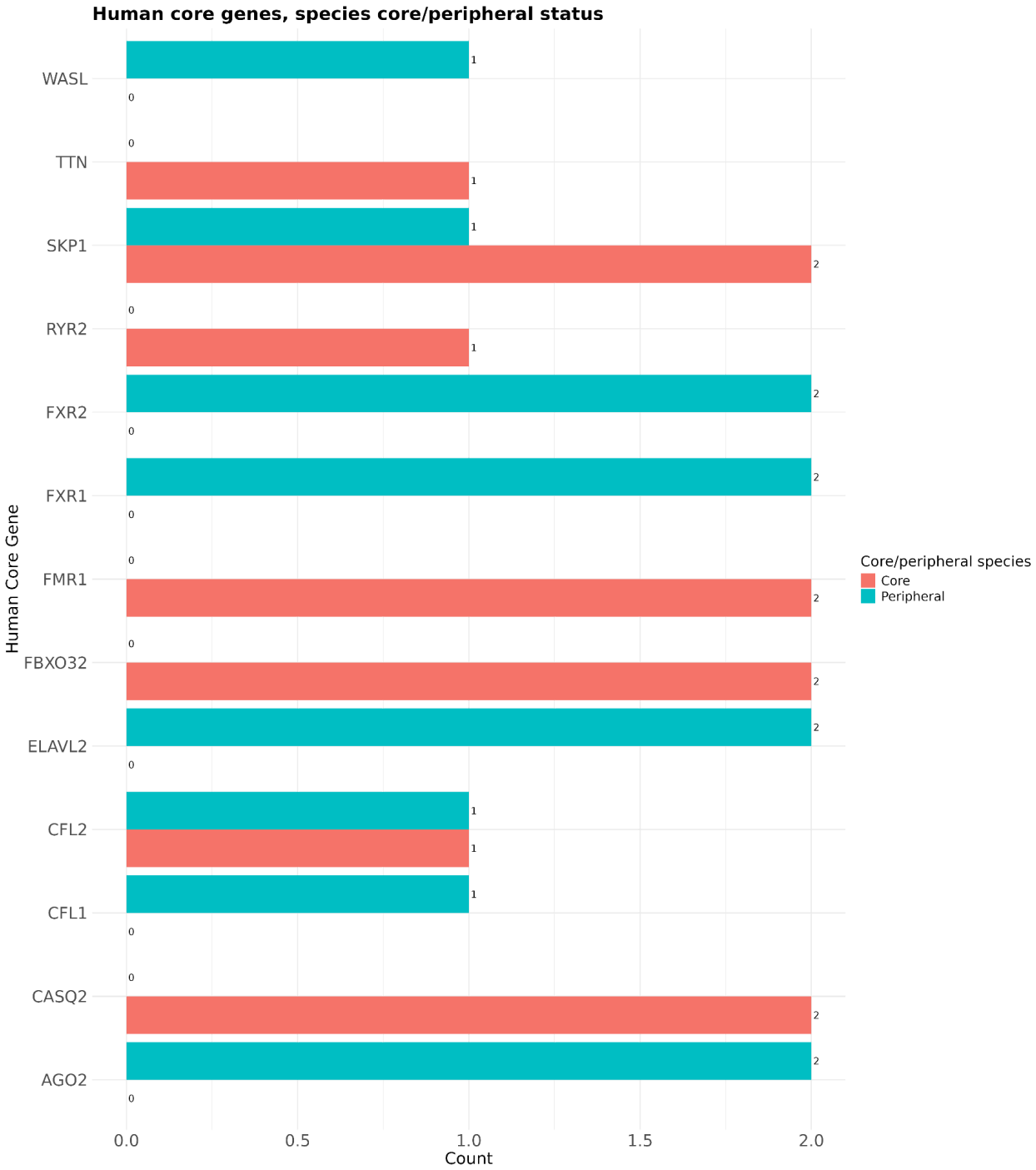

Figure 4.36: Showing the human core genes that have orthologs in other species for atrial fibrillation and how many times they are core/peripheral in five model organisms. We indicated whether the core gene is core/peripheral in the context of the other species it is in, and how many other diseases it is an ortholog for. There are 13 human core genes with orthologs in the model organisms.

Human core genes orthologs predicted in other species

Species	Human Core Gene	Omni type
Mouse	SKP1	Core
Mouse	FMR1	Core
Mouse	AGO2	Peripheral
Mouse	TTN	Core
Mouse	CFL2	Core
Mouse	WASL	Peripheral
Mouse	RYR2	Core
Mouse	CASQ2	Core
Mouse	FXR1	Peripheral
Mouse	FBXO32	Core
Mouse	FXR2	Peripheral
Mouse	ELAVL2	Peripheral
Yeast	SKP1	Peripheral
Zebrafish	SKP1	Core
Zebrafish	CFL1	Peripheral
Zebrafish	CFL2	Peripheral
Zebrafish	AGO2	Peripheral
Zebrafish	FMR1	Core
Zebrafish	FBXO32	Core
Zebrafish	FXR1	Peripheral
Zebrafish	FXR2	Peripheral
Zebrafish	CASQ2	Core
Zebrafish	ELAVL2	Peripheral

Figure 4.37: Displaying all human core genes with orthologs in each species. Some genes only appear in a single species such as ELAVL2 in Zebrafish, while others like SKP1 appear in all species.

Propagate	SKP1
GWAS	AGO2
GenePlexus	FMR1
GenePlexus	CFL1
GWAS	TTN
GWAS	CFL2
GenePlexus	WASL
GenePlexus	FXR1
GenePlexus	RYR2
GenePlexus	FXR2
GWAS	CASQ2
GWAS	FBXO32
Propagate	ELAVL2

Figure 4.38: Showing the source for each human core gene predicted in other species of how it was implicated with atrial fibrillation. Core genes (right column) come from each source (left column), where GWAS is the original GWAS from GWAS Atlas, Propagate are genes from DOMINO propagation, and GenePlexus means the genes were predicted with the GenePlexus model.

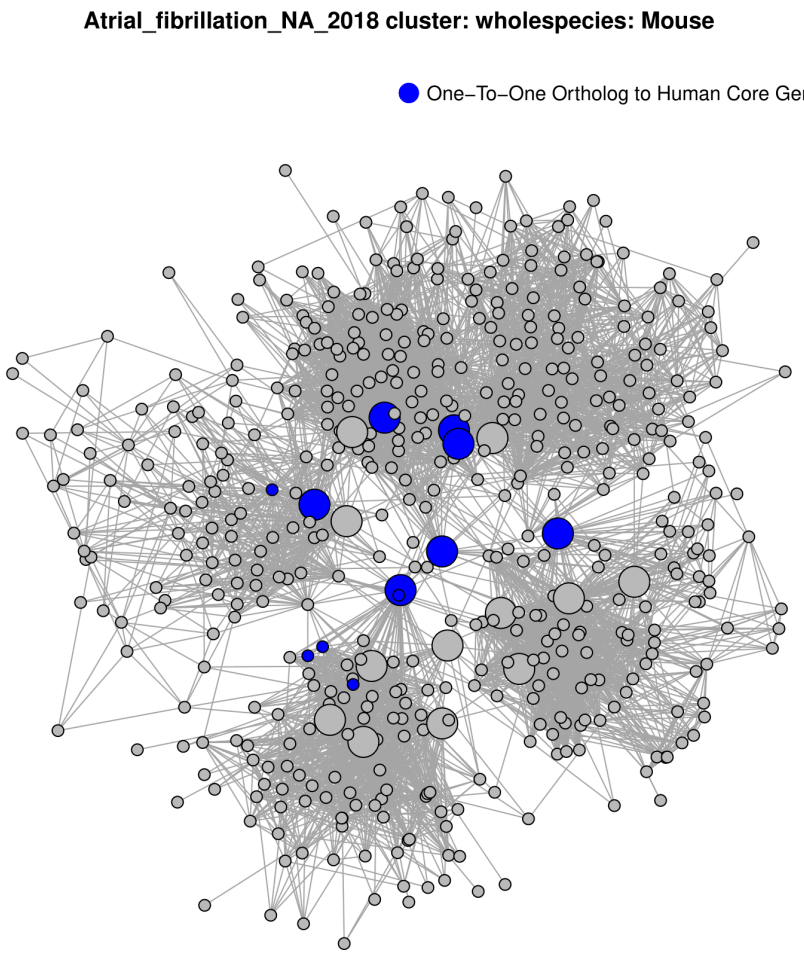

Figure 4.39: The mouse genes predicted from transferred human disease genes, with human core genes highlighted. Many of the genes that are core within the mouse module are also human core genes.

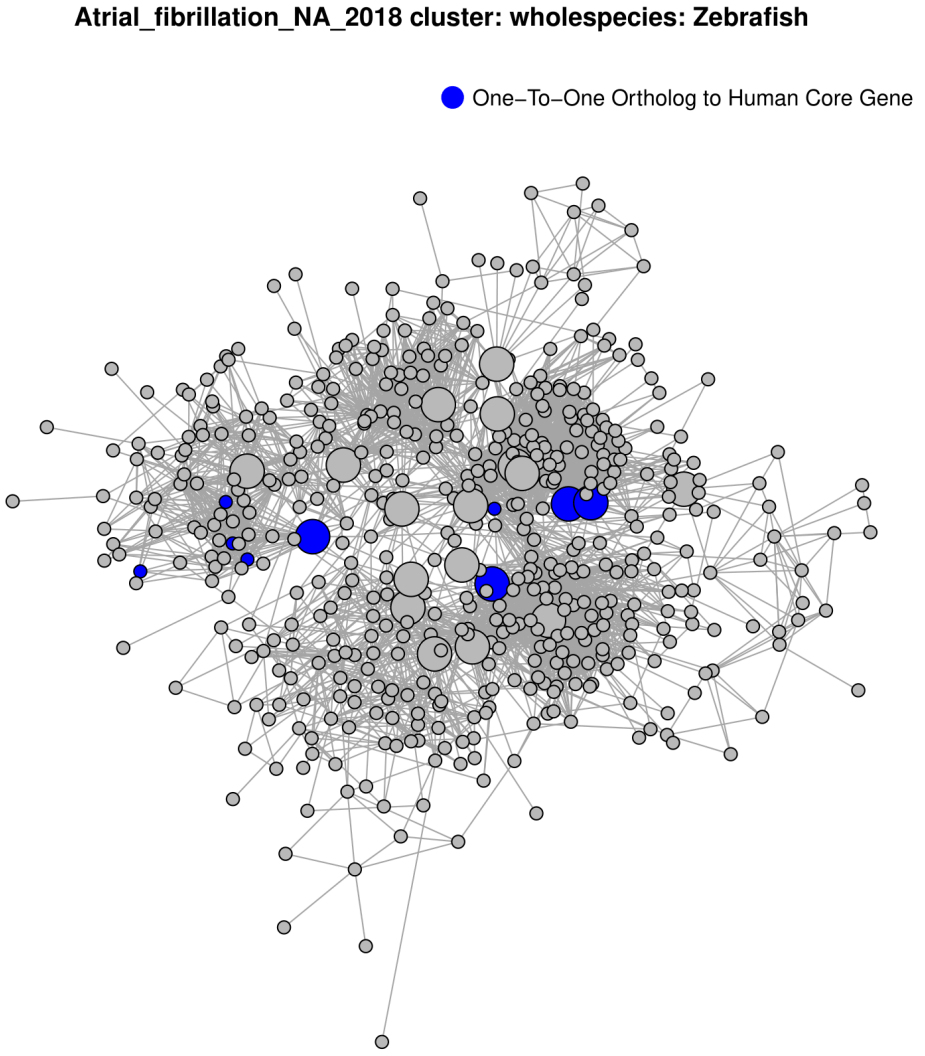

Figure 4.40: The zebrafish genes predicted from transferred human disease genes, with human core genes highlighted. Human core genes are both core and peripheral in the zebrafish module.

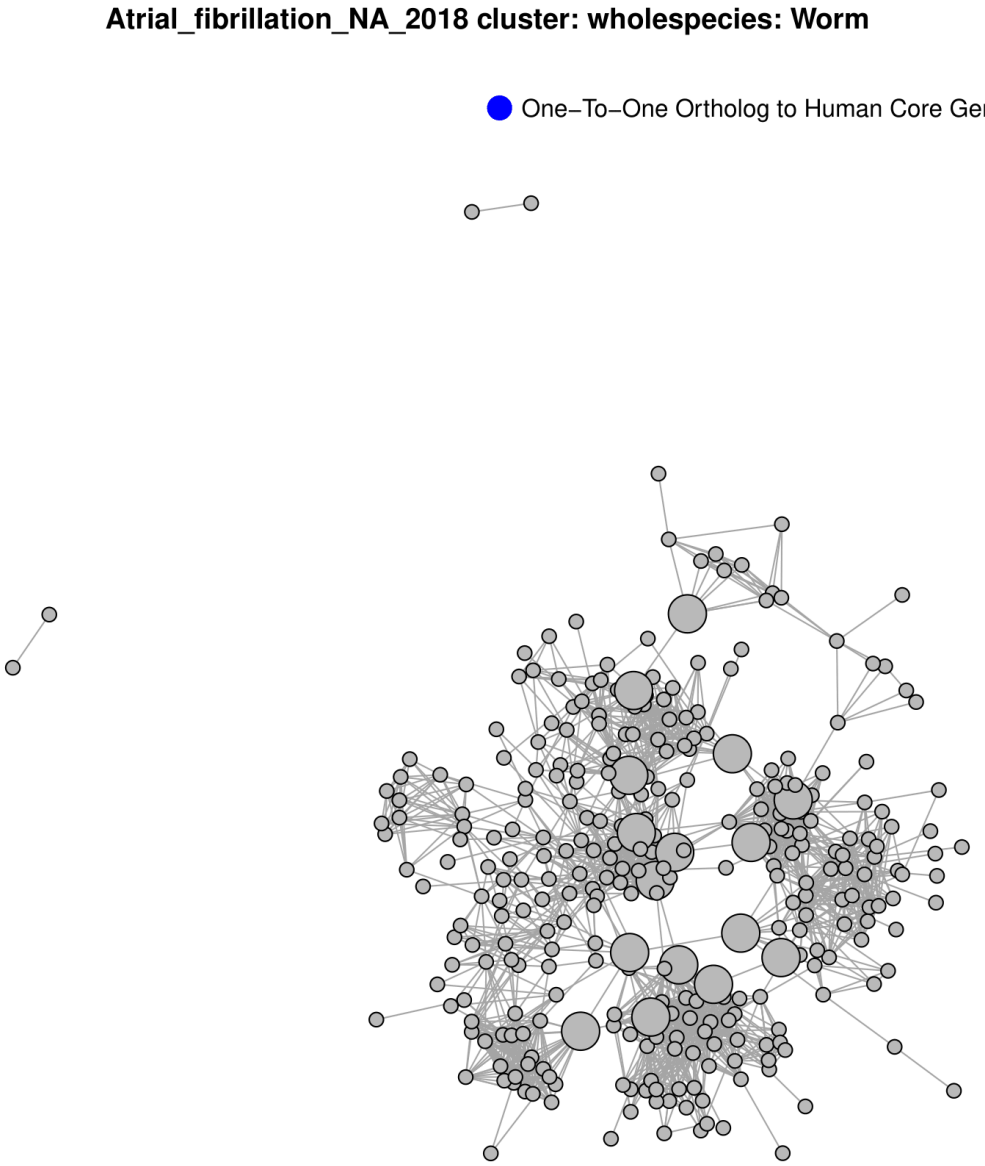

Figure 4.41: The worm genes predicted from transferred human disease genes, with human core genes highlighted. The worm module contains no human core gene orthologs.

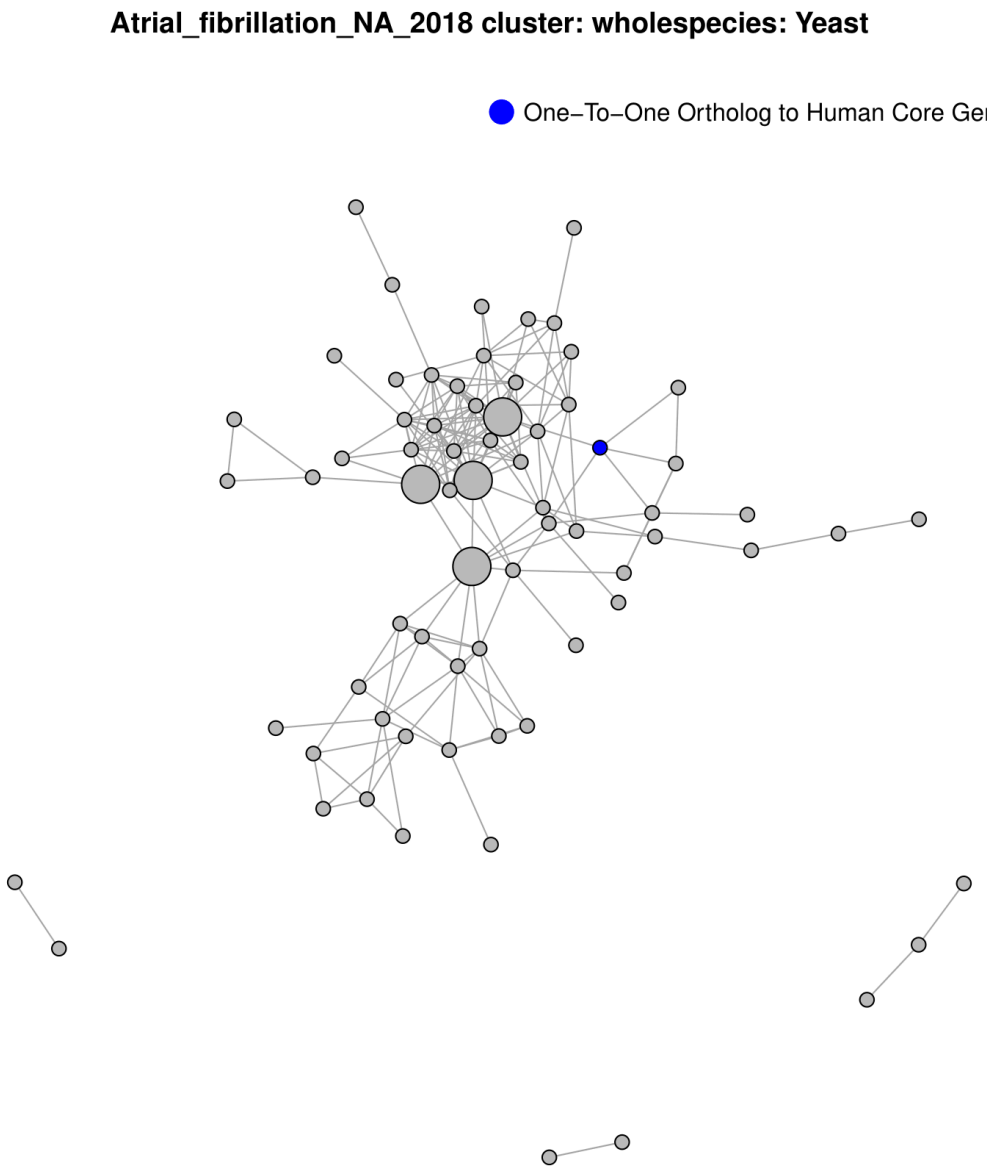

Figure 4.42: The yeast genes predicted from transferred human disease genes, with human core genes highlighted. The yeast module has a single human core gene that is peripheral within the yeast genes.

Discussion

The contribution of this study is an implementation of a method that hypothesizes core genes not based on prior knowledge of gene importance, but exclusively through networks and experimental data. We also show the ability to directly use networks to answer if human core genes are relevant in model organisms through GenePlexusZoo. We demonstrated this by focusing on atrial fibrillation across multiple biological levels, including discovering a disease module from which we predicted core genes that were shown to be biologically meaningful in heart disease and often specifically atrial fibrillation (**Figure 4.9**), contextualizing gene modules by discovering GOBPs and phenotypes related to each module that are important for atrial fibrillation (**Figure 4.29**), and in showing core genes that have relevant orthologs across species, with literature support showing how these genes can be used in model organism studies to provide insight into atrial fibrillation mechanisms (**Figure 4.37**). We also show how core genes tend to be core only in specific contexts, in this case specific diseases or traits (**Figure 4.21-23**), and are enriched for containing orthologs and highly constrained genes (**Figure 4.30-4.31**). We created a pipeline for analyzing experimental results of any complex trait or disease by performing gene classification using ModGenePlexus and GenePlexusZoo, discovering meaningful biological information about the initial geneset at the gene and module levels, and then transfering that knowledge meaningfully across species.

ModGenePlexus and GenePlexusZoo independently improve hypothesis generation

ModGenePlexus and GenePlexusZoo are two additions to GenePlexus that improve gene classification results as a whole. We demonstrate the improvement of results ModGenePlexus provides in chapter 3, and also demonstrate that it leads to discovering more underlying biology of particular diseases like type-2 diabetes. GenePlexusZoo also improves gene classification through utilizing the multi-species network. However, improving gene classification is only one of the main motivations for using these methods or for combining them like we did in this chapter. ModGenePlexus allows large genesets, like GWAS, to be usable in gene classification and be split into multiple, meaningful subsets of gene modules. These gene modules are what allows us to

discover biology about the trait that using all genes as a whole would not. GenePlexusZoo is powerful in allowing human genes to be transferred into any other species the user desires. This allows for the discovery of model organism genes that are relevant to the human trait or disease of interest. When we combine these two aspects, we are able to recover multiple meaningful genes and biological processes across species (**Figure 4.29**) using atrial fibrillation as an example. The importance of combining these two methods is that they allow us to interpret specific modules in other species' networks, allowing for a more refined way of generating hypotheses for model organism experiments that explain a specific aspect of the disease biology. When designing model organism studies where the goal is to learn more about human diseases, there are three main challenges. First is that model organisms do not perfectly capture every aspect of the human disease, but rather may be a model for particular mechanisms and molecular pathways. Using module relationships can allow the discovery and isolation of important orthologs. The second challenge is that when running parallel analyses of the same trait in GWAS between humans and model organisms, it is likely that orthologous genes will not be discovered[99]. Third, transferring ortholog knowledge across species through knockout studies of orthologous pairs does not necessarily give rise to the same observed phenotypes[100]. This is because orthologous genes do not necessarily have the same level of importance due to the different polygenic architecture underlying each species[101–103]. Since ortholog connections make it difficult to map function on its own, it is vital to use other data to give context on whether an ortholog is a suitable candidate for predicting a particular human function. Combining ModGenePlexus and GenePlexusZoo lets us use networks to solve this problem, as we integrate large amounts of information from the datasets used to construct the STRING networks of each species to interpret human genes in biologically meaningful gene modules. One more additional motivation of using GenePlexusZoo is that it has implemented a method that allows for finding enriched GOBP and Monarch phenotypes using pre-calculated model weights, rather than using a traditional enrichment software like ClusterProfiler, which we used in chapter 3. This method is powerful because ongoing work in our group has shown the GenePlexus framework can provide network-based gene set enrichment, achieving similar

performance to the widely used overrepresentation method. Additionally, the GenePlexus software offers a way to seamlessly obtain enrichments in another species without the use of converting gene lists using orthology. An example is seen in the Bardet-Biedl Syndrome where GenePlexusZoo uncovered enriched GOBPs that are biologically relevant, but could not be found from doing ortholog overlap in an ORA because there are no direct one-to-one orthologs implicated. Rather, network connections between the two species were used to provide biological insight.

Defining human core genes based on the disease module rather than at the gene module level

Multiple decisions had to be made to decide an ideal definition for core genes. We have already discussed why we used betweenness centrality to define core genes within the predicted disease module. This definition is useful as it does not require assumptions of gene knowledge about any particular disease – meaning it works generally across genesets – and because we are using an entire disease module, it also allows us to find peripheral disease genes, as those disease genes that are not core must be peripheral in an omnigenic framework. One question was whether it would be better to define core genes at the module level, rather than at the disease level. This was motivated through seeing that core genes are not one meaningful biological set in themselves, but are split across modules for each GWAS (**Figure 4.10-13**). This was decided against for multiple reasons. The primary reason is that modules themselves are difficult to define biologically and work together to give rise to disease manifestations. Interpreting what a core gene is in this context would be very challenging and is a leap from thinking about complex traits omnigenically. It is useful to use modules to give biological context to a disease and discover new biological relationships, but defining them as a biological entity itself is very challenging. As such, we chose to use our method to define the omnigenic mapping of the genes holistically. We take a whole disease, break it into modules, and then use those modules as tools for superior gene classification. Those gene predictions are then aggregated into a final list that we predict is the disease module – all genes that could be considered relevant based on genome-wide network connections – and defines core and peripheral genes in the context of the entire complex trait. It is valid to interpret disease genes all at once because they are all

involved in the disease, and as such do have relevance to one another in biological systems. Good core genes need to be determined from that observation because the scale of complexity is what motivates core genes being a concept in the first place.

Human core genes are not one meaningful biological set

Traditional discussions of core genes, such as in the original paper introducing the omnigenic model[2], mention that there could be multiple “core pathways” involved with a complex disease. However, it's unclear how the core genes would be distributed across a disease module or a genome-wide network, and what the implications of the answer would be. In this study, we show for the 20 GWAS that the core genes are divided amongst the discovered gene modules (**Figure 4.10-13**). This means that core genes must be considered in terms of how they work in distinct processes or phenotypes. Core genes being seen across modules raises questions for how to best determine causal loci and genetic relationships across diseases. Pritchard et. al. mention that an assumption behind methods like mendelian randomization is that pleiotropy between traits that are not causally related is rare[105]. If there is intersection of core genes across multiple, not related diseases, then those diseases may have similar underlying modules due to the presence of similar genes, and thus have a form of pleiotropy where mutations on the same genes affect multiple, unrelated complex diseases. In chapter 2 we showed that complex diseases can share modules that are enriched for more general pathways like inflammation. It will be vital to implement methods like that in omnigenically interpreting disease and core gene relationships.

The omnigenic model can provide context to GWAS results

We discussed in the introduction for this chapter that as GWAS get bigger, the number of significant loci discovered increases, but the effect size of loci that were previously implicated in smaller studies are decreasing over time. This observation shows the necessity of post-GWAS analysis, as traditional methods have assumptions in place that will no longer work well with the genotype level data. This is the motivation behind the omnigenic model and behind the method we implemented. We integrated the GWAS hits with vast amounts of biological data through the multi-species network to predict genes that the GWAS missed and provide biological context using GOBPs and phenotypes across species. The GWAS genes successfully discovered other

mechanistically important genes for atrial fibrillation. The omnigenic model is not just a useful framework for thinking about the scale of complexity of human complex diseases and traits, but we provided evidence that the concept of core genes has validity within biological networks.

Future directions

A major future direction for this project will involve doing phenotype interpretation at the cross-species level. We are implementing a double blind study where we asked graduate students to determine if a set of model organism phenotypes were relevant to a human disease and a set of human phenotypes. For each module we randomly provided positive module examples (where the GOBP or phenotype was determined to be related to the module using GenePlexusZoo), positive disease examples (where the GOBP or phenotype was determined to be related to a different module for the same disease) and negative disease examples (where the GOBP or phenotype was determined to be related to a different disease). This study will help us determine how relevant our annotated model organism phenotypes are to the GWAS and how biologically distinct modules are within a GWAS. A second direction to take this project is to focus on cross-GWAS comparison of core genes. We have shown that the core genes are meaningful for the GWAS, but many insights into the omnigenic model can be obtained by seeing how core genes relate across traits and within the network. This can answer questions in the original paper, such as how pleiotropy could affect studies of causality and how distinct traits really are from each other in terms of implicated loci. We began investigating this by seeing that core genes tend to be unique within GWAS and in seeing similar patterns where core genes are distributed across modules for each GWAS.

REFERENCES

1. Crouch, D. J. M. & Bodmer, W. F. Polygenic inheritance, GWAS, polygenic risk scores, and the search for functional variants. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 18924–18933 (2020).

2. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).

3. Liu, X., Li, Y. I. & Pritchard, J. K. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell* **177**, 1022-1034.e6 (2019).

4. Fisher, R. A., Immer, F. R. & Tedin, O. THE GENETICAL INTERPRETATION OF STATISTICS OF THE THIRD DEGREE IN THE STUDY OF QUANTITATIVE INHERITANCE. *Genetics* **17**, 107–124 (1932).

5. Van Der Sijde, M. R., Ng, A. & Fu, J. Systems genetics: From GWAS to disease pathways. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **1842**, 1903–1909 (2014).

6. Zhu, Z. *et al.* Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat Commun* **9**, 224 (2018).

7. Kao, P. Y. P., Leung, K. H., Chan, L. W. C., Yip, S. P. & Yap, M. K. H. Pathway analysis of complex diseases for GWAS, extending to consider rare variants, multi-omics and interactions. *Biochimica et Biophysica Acta (BBA) - General Subjects* **1861**, 335–353 (2017).

8. Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z. & Bergmann, S. Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. *PLoS Comput Biol* **12**, e1004714 (2016).

9. McGary, K. L. *et al.* Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 6544–6549 (2010).

10. Li, Z. *et al.* Understanding autism spectrum disorders with animal models: applications, insights, and perspectives. *Zoological Research* **42**, 800–823 (2021).

11. Alghamdi, S. M., Schofield, P. N. & Hoehndorf, R. Contribution of model organism phenotypes to the computational identification of human disease genes. *Disease Models & Mechanisms* **15**, dmm049441 (2022).

12. Ganz, J., Melancon, E. & Eisen, J. S. Zebrafish as a model for understanding enteric nervous system interactions in the developing intestinal tract. *Methods Cell Biol* **134**, 139–164 (2016).

13. Howe, D. G. *et al.* Model organism data evolving in support of translational medicine. *Lab Anim* **47**, 277–289 (2018).

14. Mancuso, C. A., Johnson, K. A., Liu, R. & Krishnan, A. Joint representation of

molecular networks from multiple species improves gene classification. *PLoS Comput Biol* **20**, e1011773 (2024).

15. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research* **47**, D607–D613 (2019).

16. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* **47**, D309–D314 (2019).

17. Liu, R. & Krishnan, A. PecanPy: a fast, efficient and parallelized Python implementation of node2vec. *Bioinformatics* **37**, 3377–3379 (2021).

18. Grover, A. & Leskovec, J. node2vec: Scalable Feature Learning for Networks. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* 855–864 (ACM Press, San Francisco, California, USA, 2016). doi:10.1145/2939672.2939754.

19. Levi, H., Elkon, R. & Shamir, R. *DOMINO: A Novel Algorithm for Network-Based Identification of Active Modules with Reduced Rate of False Calls*. http://biorxiv.org/lookup/doi/10.1101/2020.03.10.984963 (2020) doi:10.1101/2020.03.10.984963.

20. Leeuw, C. A. de, Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Computational Biology* **11**, e1004219 (2015).

21. Tian, D. *et al.* GWAS Atlas: a curated resource of genome-wide variant-trait associations in plants and animals. *Nucleic Acids Res* **48**, D927–D932 (2020).

22. Shefchek, K. A. *et al.* The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Research* **48**, D704–D715 (2020).

23. Mungall, C. J. *et al.* The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Research* **45**, D712–D722 (2017).

24. The Gene Ontology Consortium *et al.* The Gene Ontology knowledgebase in 2023. *GENETICS* **224**, iyad031 (2023).

25. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat Genet* **25**, 25–29 (2000).

26. Geistlinger, L. *et al.* Towards a gold standard for benchmarking gene set enrichment analysis. *bioRxiv* 674267 (2019) doi:10.1101/674267.

27. Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring network structure, dynamics, and function using NetworkX. *Proceedings of the 7th Python in Science*

Conference (SciPy2008) 11–15 (2008).

28. Brandes, U. A faster algorithm for betweenness centrality*. *The Journal of Mathematical Sociology* **25**, 163–177 (2001).

29. Smedley, D. *et al.* BioMart – biological queries made easy. *BMC Genomics* **10**, 22 (2009).

30. Chen, S. *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).

31. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

32. Exome Aggregation Consortium *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

33. Liu, R., Yuan, H., Johnson, K. A. & Krishnan, A. CONE: COntext-specific Network Embedding via Contextualized Graph Attention. Preprint at https://doi.org/10.1101/2023.10.21.563390 (2023).

34. GTEx Consortium *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* **47**, 1091–1098 (2015).

35. Kryuchkova-Mostacci, N. & Robinson-Rechavi, M. A benchmark of gene expression tissue-specificity metrics. *Brief Bioinform* bbw008 (2016) doi:10.1093/bib/bbw008.

36. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).

37. Palasca, O., Santos, A., Stolte, C., Gorodkin, J. & Jensen, L. J. TISSUES 2.0: an integrative web resource on mammalian tissue expression. *Database* **2018**, (2018).

38. Zhang, Y. *et al.* CellSTAR: a comprehensive resource for single-cell transcriptomic annotation. *Nucleic Acids Research* **52**, D859–D870 (2024).

39. Sinnott-Armstrong, N., Naqvi, S., Rivas, M. & Pritchard, J. K. GWAS of three molecular traits highlights core genes and pathways alongside a highly polygenic background. *eLife* **10**, e58615 (2021).

40. Domingo-Relloso, A. *et al.* Arsenic Exposure, Blood DNA Methylation, and Cardiovascular Disease. *Circulation Research* **131**, (2022).

41. Liu, H. *et al.* Atrial fibrillation alters the microRNA expression profiles of the left atria of patients with mitral stenosis. *BMC Cardiovasc Disord* **14**, 10 (2014).

42. Chu, M. *et al.* Increased Cardiac Arrhythmogenesis Associated With Gap Junction Remodeling With Upregulation of RNA-Binding Protein FXR1. *Circulation* **137**, 605–618 (2018).

43. Whitman, S. A. *et al.* Desmoplakin and Talin2 Are Novel mRNA Targets of Fragile X–Related Protein-1 in Cardiac Muscle. *Circulation Research* **109**, 262–271 (2011).

44. Tassanakijpanich, N., Cohen, J., Cohen, R., Srivatsa, U. N. & Hagerman, R. J. Cardiovascular Problems in the Fragile X Premutation. *Front. Genet.* **11**, 586910 (2020).

45. Wass, S. Y. *et al.* Novel functional atrial fibrillation risk genes and pathways identified from coexpression analyses in human left atria. *Heart Rhythm* **20**, 1219–1226 (2023).

46. Ning, Z. *et al.* Novel Drug Targets for Atrial Fibrillation Identified Through Mendelian Randomization Analysis of the Blood Proteome. *Cardiovasc Drugs Ther* (2023) doi:10.1007/s10557-023-07467-8.

47. Cai, W. *et al.* asb5a/asb5b Double Knockout Affects Zebrafish Cardiac Contractile Function. *IJMS* **24**, 16364 (2023).

48. Shapiro, D., Lee, K., Asmussen, J., Bourquard, T. & Lichtarge, O. Evolutionary Action–Machine Learning Model Identifies Candidate Genes Associated With Early‐Onset Coronary Artery Disease. *JAHA* **12**, e029103 (2023).

49. Sheng, C. *et al.* CALML6 Controls TAK1 Ubiquitination and Confers Protection against Acute Inflammation. *The Journal of Immunology* **204**, 3008–3018 (2020).

50. Qu, Y.-C. *et al.* Activated nuclear factor-κB and increased tumor necrosis factor-α in atrial tissue of atrial fibrillation. *Scandinavian Cardiovascular Journal* **43**, 292–297 (2009).

51. Liu, Y. *et al.* Identification of atrial fibrillation-associated lncRNAs and exploration of their functions based on WGCNA and ceRNA network analyses. *gpb* **40**, 289–305 (2021).

52. Zou, J. *et al.* Neddylation mediates ventricular chamber maturation through repression of Hippo signaling. *Proc. Natl. Acad. Sci. U.S.A.* **115**, (2018).

53. Maejima, Y. & Sadoshima, J. SUMOylation: A Novel Protein Quality Control Modifier in the Heart. *Circulation Research* **115**, 686–689 (2014).

54. Martins, I. L. F. *et al.* Reviewing Atrial Fibrillation Pathophysiology from a Network Medicine Perspective: The Relevance of Structural Remodeling, Inflammation, and the Immune System. *Life (Basel)* **13**, 1364 (2023).

55. Zou, R. *et al.* Analysis of Genes Involved in Persistent Atrial Fibrillation: Comparisons of ‘Trigger’ and ‘Substrate’ Differences. *Cell Physiol Biochem* **47**, 1299–1309 (2018).

56. Li, C. Y., Zhang, J. R., Hu, W. N. & Li, S. N. Atrial fibrosis underlying atrial fibrillation (Review). *Int J Mol Med* **47**, 9 (2021).

57. Kawaji, T. *et al.* Clinical significance of ST-segment depression during atrial fibrillation rhythm for subsequent heart failure events. *European Heart Journal Open* **3**, oead060 (2023).

58. Park, K.-M. *et al.* Atrial Fibrillation in Hypertrophic Cardiomyopathy: Is the Extent of Septal Hypertrophy Important? *PLoS One* **11**, e0156410 (2016).

59. Westerman, S. & Wenger, N. Gender Differences in Atrial Fibrillation: A Review of Epidemiology, Management, and Outcomes. *Curr Cardiol Rev* **15**, 136–144 (2019).

60. Siddiqi, H. K. *et al.* Sex Differences in Atrial Fibrillation Risk: The VITAL Rhythm Study. *JAMA Cardiol* **7**, 1027 (2022).

61. Thibault, S., Long, V. & Fiset, C. Higher Na+-Ca2+ Exchanger Function and Triggered Activity Contribute to Male Predisposition to Atrial Fibrillation. *IJMS* **23**, 10724 (2022).

62. Dobrev, D., Heijman, J., Hiram, R., Li, N. & Nattel, S. Inflammatory signalling in atrial cardiomyocytes: a novel unifying principle in atrial fibrillation pathophysiology. *Nat Rev Cardiol* **20**, 145–167 (2023).

63. Wangler, M. F., Hu, Y. & Shulman, J. M. *Drosophila* and genome-wide association studies: a review and resource for the functional dissection of human complex traits. *Disease Models & Mechanisms* **10**, 77–88 (2017).

64. Duncan, L. E., Ostacher, M. & Ballon, J. How genome-wide association studies (GWAS) made traditional candidate gene studies obsolete. *Neuropsychopharmacol.* **44**, 1518–1523 (2019).

65. Madelaine, R. *et al.* A screen for deeply conserved non-coding GWAS SNPs uncovers a MIR-9-2 functional mutation associated to retinal vasculature defects in human. *Nucleic Acids Research* **46**, 3517–3531 (2018).

66. Mostafavi, H., Spence, J. P., Naqvi, S. & Pritchard, J. K. Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nat Genet* **55**, 1866–1875 (2023).

67. Chen, S., Zhang, Y. E. & Long, M. New Genes in *Drosophila* Quickly Become Essential. *Science* **330**, 1682–1685 (2010).

68. Flintoft, L. Young genes are essential too. *Nat Rev Genet* **12**, 79–79 (2011).

69. Chen, S., Krinsky, B. H. & Long, M. New genes as drivers of phenotypic evolution. *Nat Rev Genet* **14**, 645–660 (2013).

70. Andersson, D. I., Jerlström-Hultqvist, J. & Näsvall, J. Evolution of new functions de novo and from preexisting genes. *Cold Spring Harb Perspect Biol* **7**, a017996 (2015).

71. Iwasaki, Y., Nishida, K., Kato, T. & Nattel, S. Atrial Fibrillation Pathophysiology:

Implications for Management. *Circulation* **124**, 2264–2274 (2011).

72. Nso, N., Bookani, K. R., Metzl, M. & Radparvar, F. Role of inflammation in atrial fibrillation: A comprehensive review of current knowledge. *Journal of Arrhythmia* **37**, 1–10 (2021).

73. Zhou, X. & Dudley, S. C. Evidence for Inflammation as a Driver of Atrial Fibrillation. *Front. Cardiovasc. Med.* **7**, 62 (2020).

74. Lessomo, F. Y. N., Fan, Q., Wang, Z.-Q. & Mukuka, C. The relationship between leukocyte to albumin ratio and atrial fibrillation severity. *BMC Cardiovasc Disord* **23**, 67 (2023).

75. Chen, Y. *et al.* Association between circulating leukocytes and arrhythmias: Mendelian randomization analysis in immuno-cardiac electrophysiology. *Front. Immunol.* **14**, 1041591 (2023).

76. Paquissi, F. C. The Predictive Role of Inflammatory Biomarkers in Atrial Fibrillation as Seen through Neutrophil-Lymphocyte Ratio Mirror. *Journal of Biomarkers* **2016**, 1–14 (2016).

77. Ertaş, G. *et al.* Neutrophil/lymphocyte ratio is associated with thromboembolic stroke in patients with non-valvular atrial fibrillation. *Journal of the Neurological Sciences* **324**, 49–52 (2013).

78. Gibson, P. H. *et al.* Usefulness of Neutrophil/Lymphocyte Ratio As Predictor of New-Onset Atrial Fibrillation After Coronary Artery Bypass Grafting. *The American Journal of Cardiology* **105**, 186–191 (2010).

79. Conte, M. *et al.* Epicardial Adipose Tissue and Cardiac Arrhythmias: Focus on Atrial Fibrillation. *Front Cardiovasc Med* **9**, 932262 (2022).

80. Chahine, Y. *et al.* Epicardial adipose tissue is associated with left atrial volume and fibrosis in patients with atrial fibrillation. *Front. Cardiovasc. Med.* **9**, 1045730 (2022).

81. Wong, C. X. *et al.* Associations of Epicardial, Abdominal, and Overall Adiposity With Atrial Fibrillation. *Circ: Arrhythmia and Electrophysiology* **9**, e004378 (2016).

82. Karlsson, M. *et al.* A single–cell type transcriptomics map of human tissues. *Sci. Adv.* **7**, eabh2169 (2021).

83. Weil, B. R. & Ozcan, C. Cardiomyocyte Remodeling in Atrial Fibrillation and Hibernating Myocardium: Shared Pathophysiologic Traits Identify Novel Treatment Strategies? *Biomed Res Int* **2015**, 587361 (2015).

84. Hao, H. *et al.* Atrial myocyte-derived exosomal microRNA contributes to atrial fibrosis in atrial fibrillation. *J Transl Med* **20**, 407 (2022).

85. Patel, K. K. *et al.* Genomic approaches to identify and investigate genes associated

with atrial fibrillation and heart failure susceptibility. *Hum Genomics* **17**, 47 (2023).

86. Faggioni, M. *et al.* Suppression of Spontaneous Ca Elevations Prevents Atrial Fibrillation in Calsequestrin 2-Null Hearts. *Circ: Arrhythmia and Electrophysiology* **7**, 313–320 (2014).

87. Nakano, Y. Genome and atrial fibrillation. *J Arrhythm* **39**, 303–309 (2023).

88. Roselli, C. *et al.* Multi-ethnic genome-wide association study for atrial fibrillation. *Nat Genet* **50**, 1225–1233 (2018).

89. Faggioni, M. & Knollmann, B. C. Calsequestrin 2 and arrhythmias. *American Journal of Physiology-Heart and Circulatory Physiology* **302**, H1250–H1260 (2012).

90. Zhang, J.-C. *et al.* Calcium-Mediated Oscillation in Membrane Potentials and Atrial-Triggered Activity in Atrial Cells of Casq2R33Q/R33Q Mutation Mice. *Front Physiol* **9**, 1447 (2018).

91. Kalyanasundaram, A. *et al.* Functional consequences of stably expressing a mutant calsequestrin (CASQ2 ^{D307H}) in the CASQ2 null background. *American Journal of Physiology-Heart and Circulatory Physiology* **302**, H253–H261 (2012).

92. Li, P.-F. *et al.* Modulation of miR-10a-mediated TGF-β1/Smads signaling affects atrial fibrillation-induced cardiac fibrosis and cardiac fibroblast proliferation. *Bioscience Reports* **39**, BSR20181931 (2019).

93. Bühler, A. *et al.* Atrogin-1 Deficiency Leads to Myopathy and Heart Failure in Zebrafish. *Int J Mol Sci* **17**, 187 (2016).

94. Li, J. *et al.* A noncanonical function of SKP1 regulates the switch between autophagy and unconventional secretion. *Sci Adv* **9**, eadh1134 (2023).

95. Boehm, B. M., Gaa, J., Hoppmann, P., Martens, E. & Westphal, D. S. The Role of RYR2 in Atrial Fibrillation. *Case Rep Cardiol* **2023**, 6555998 (2023).

96. Di Pino, A., Caruso, E., Costanzo, L. & Guccione, P. A novel RyR2 mutation in a 2-year-old baby presenting with atrial fibrillation, atrial flutter, and atrial ectopic tachycardia. *Heart Rhythm* **11**, 1480–1483 (2014).

97. Yoneda, Z. T. *et al.* Early-Onset Atrial Fibrillation and the Prevalence of Rare Variants in Cardiomyopathy and Arrhythmia Genes. *JAMA Cardiol* **6**, 1371 (2021).

98. Gillis, A. M. & Dobrev, D. Targeting the RyR2 to Prevent Atrial Fibrillation. *Circ: Arrhythmia and Electrophysiology* **15**, (2022).

99. Wright, S. N. *et al.* Genome-wide association studies of human and rat BMI converge on synapse, epigenome, and hormone signaling networks. *Cell Rep* **42**, 112873 (2023).

100. Liao, B.-Y. & Zhang, J. Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc Natl Acad Sci U S A* **105**, 6987–6992 (2008).

101. Beura, L. K. *et al.* Normalizing the environment recapitulates adult human immune traits in laboratory mice. *Nature* **532**, 512–516 (2016).

102. Mestas, J. & Hughes, C. C. W. Of mice and not men: differences between mouse and human immunology. *J Immunol* **172**, 2731–2738 (2004).

103. Stamboulian, M., Guerrero, R. F., Hahn, M. W. & Radivojac, P. The ortholog conjecture revisited: the value of orthologs and paralogs in function prediction. *Bioinformatics* **36**, i219–i226 (2020).

104. Chen, H. *et al.* New insights on human essential genes based on integrated analysis and the construction of the HEGIAP web-based platform. *Brief Bioinform* **21**, 1397–1410 (2020).

105. Wagner, G. P. & Zhang, J. The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. *Nat Rev Genet* **12**, 204–213 (2011).

APPENDIX A4: CORE GENES

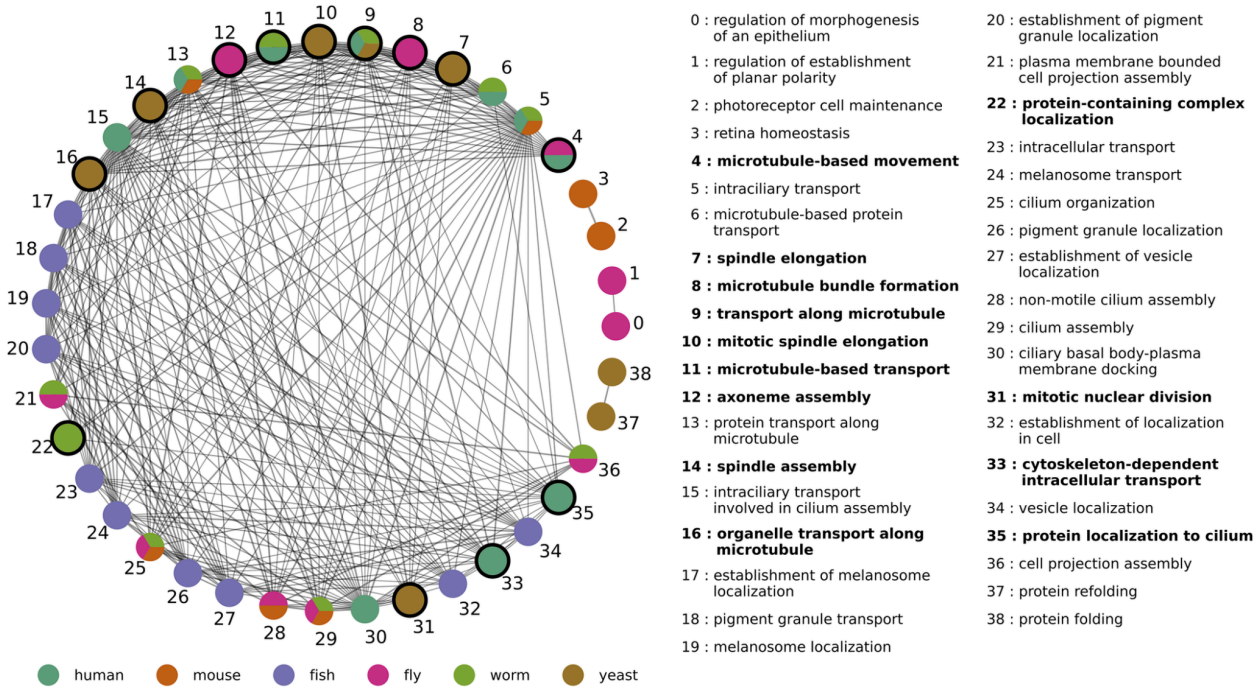

Figure A4.1: The ten most enriched biological processes associated with the top genes in each species predicted to be related to Bardet-Biedel Syndrome (BBS): This figure show results from the original GenePlexusZoo paper[14]. A classifier trained using human BBS genes was used to predict the BBS-related genes in model organisms within the multi-species network. In this graph, nodes represent the ten most enriched biological processes of each species and are colored by species they are identified in. Edges represent semantically similar processes. Biological process nodes with thick borders (and bolded labels) represent those processes in which at least one species in which none of the annotated genes are orthologous to any human BBS gene. The process was instead implicated through network connections.

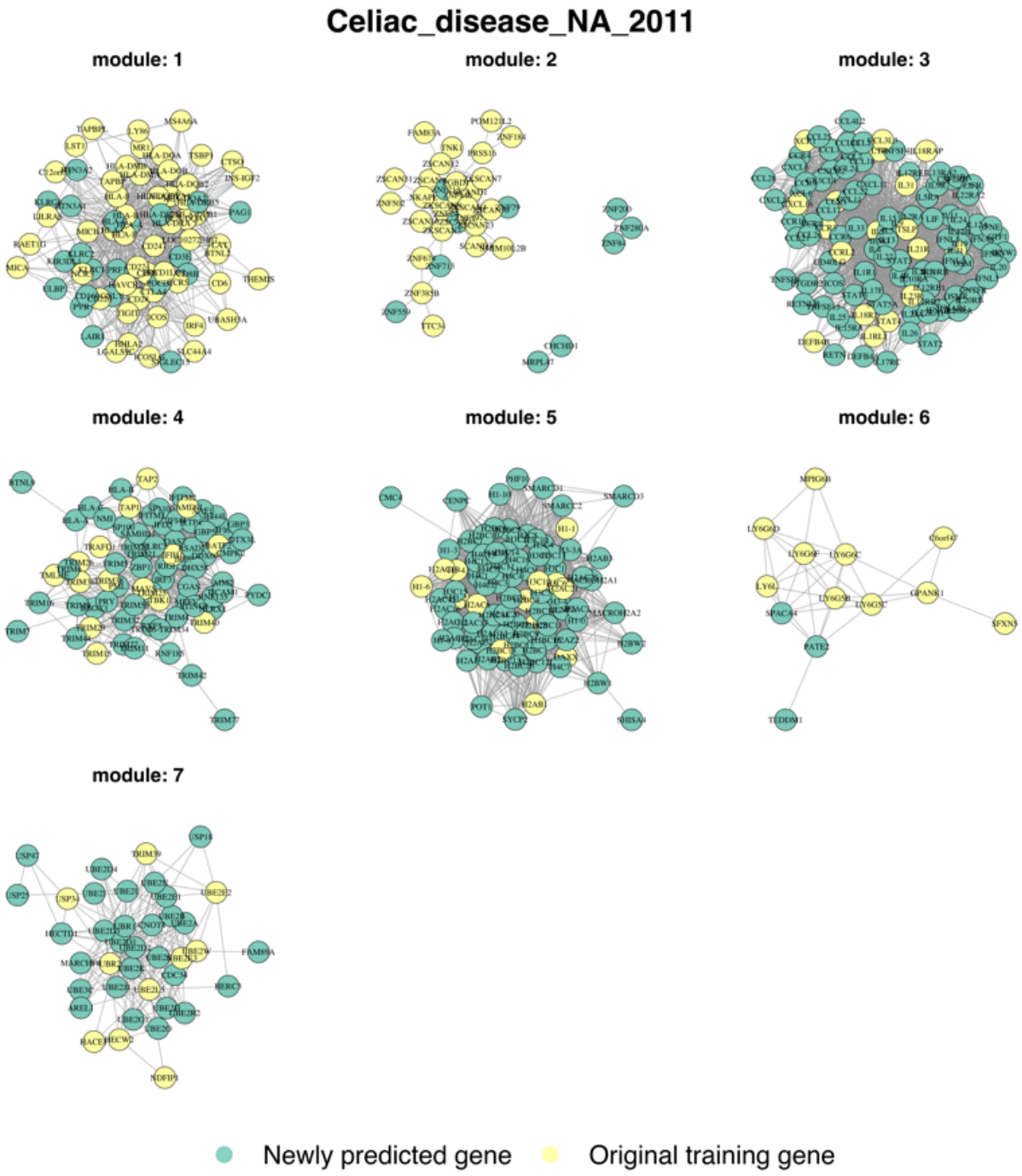

Figure A4.2: Genes in gene modules for celiac disease. Tan nodes are the original training genes used in GenePlexusZoo, and green nodes are predicted genes from GenePlexusZoo.

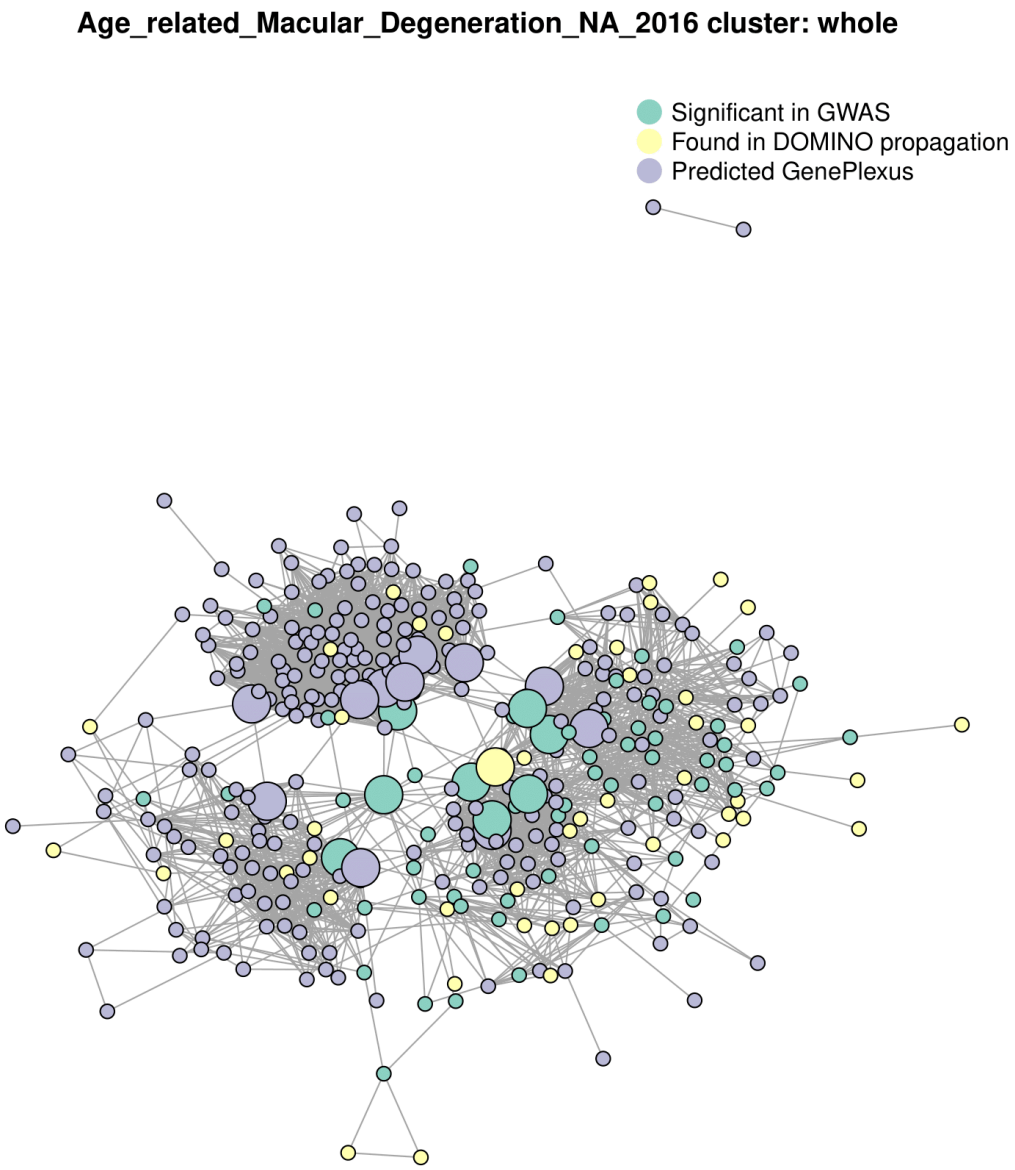

Figure A4.3: Predicted disease module for GWAS age related macular degeneration.

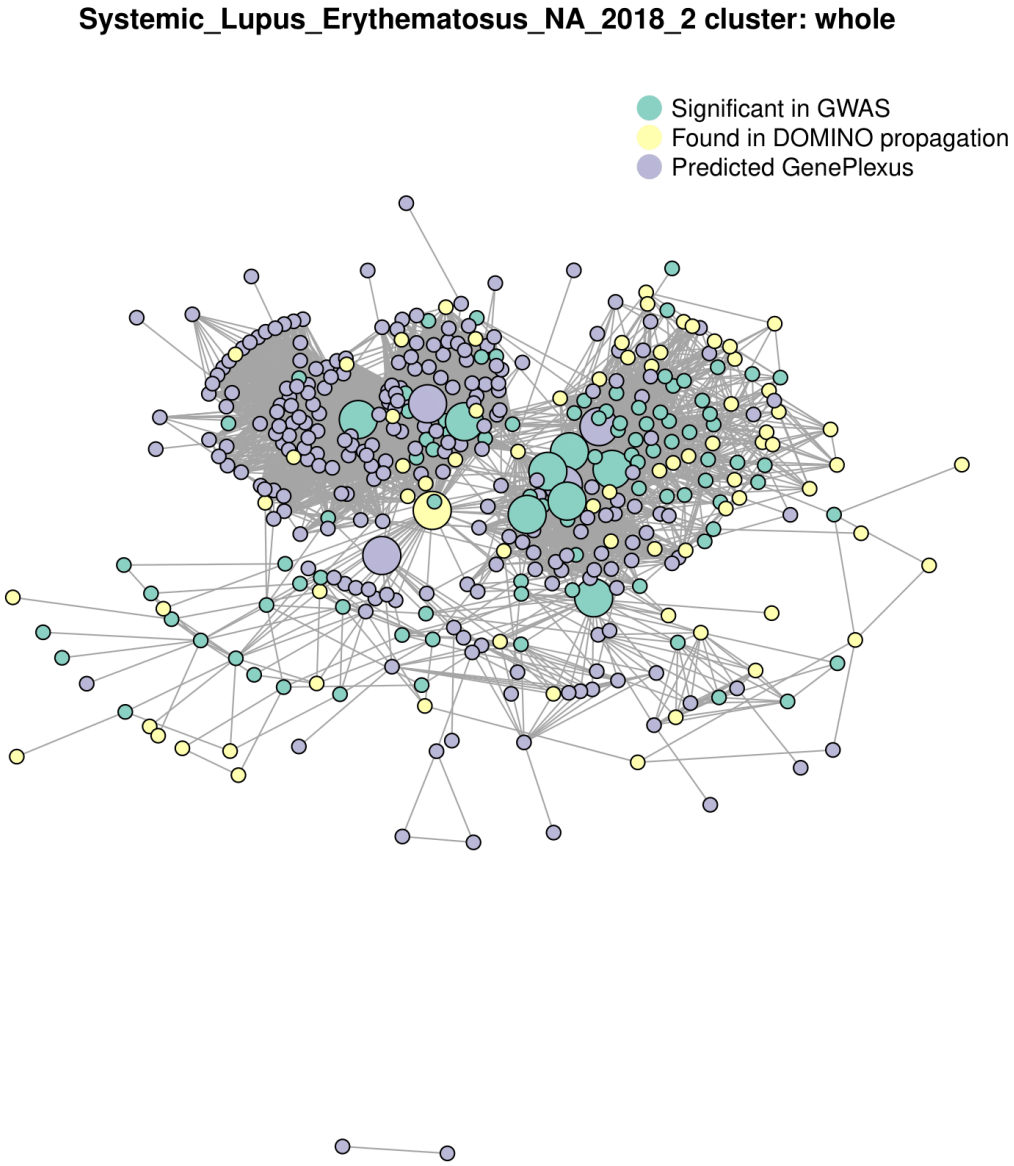

Figure A4.4: A disease module for GWAS systemic lupus erythematosus.

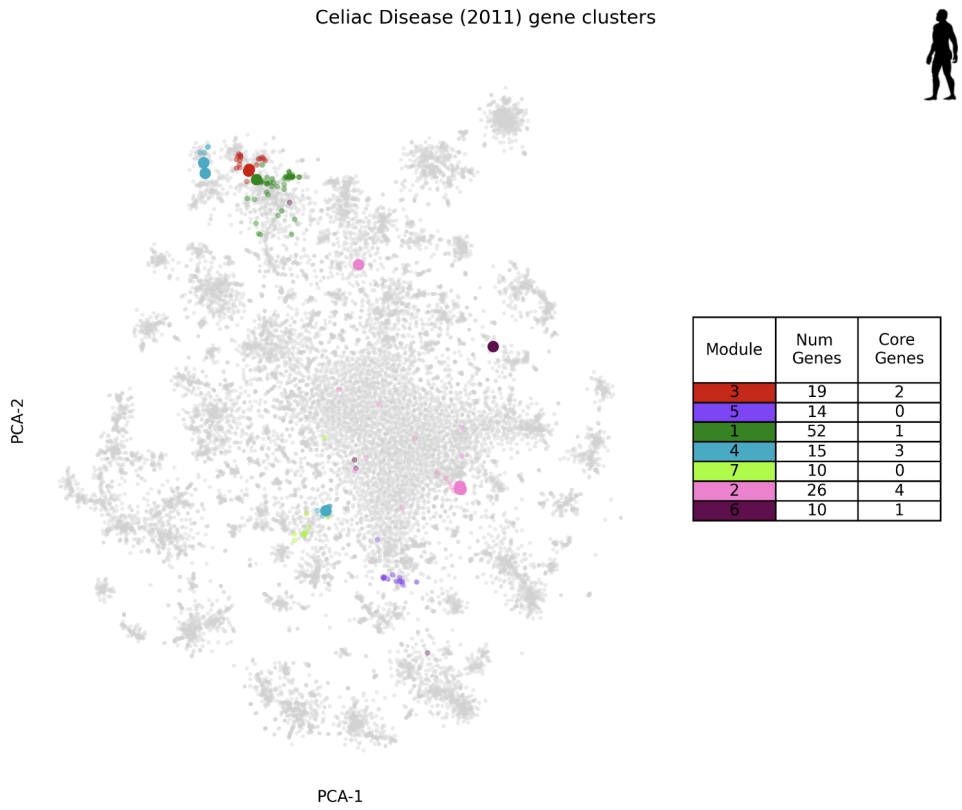

Figure A4.5: Displaying GWAS genes of celiac disease within a t-SNE of the human STRING network, with each module displayed. The larger nodes are the predicted core genes, and each gene corresponds to a module (see legend). GWAS core genes are distributed across modules for a trait,and the modules with multiple core genes have those genes as neighbors in the network.

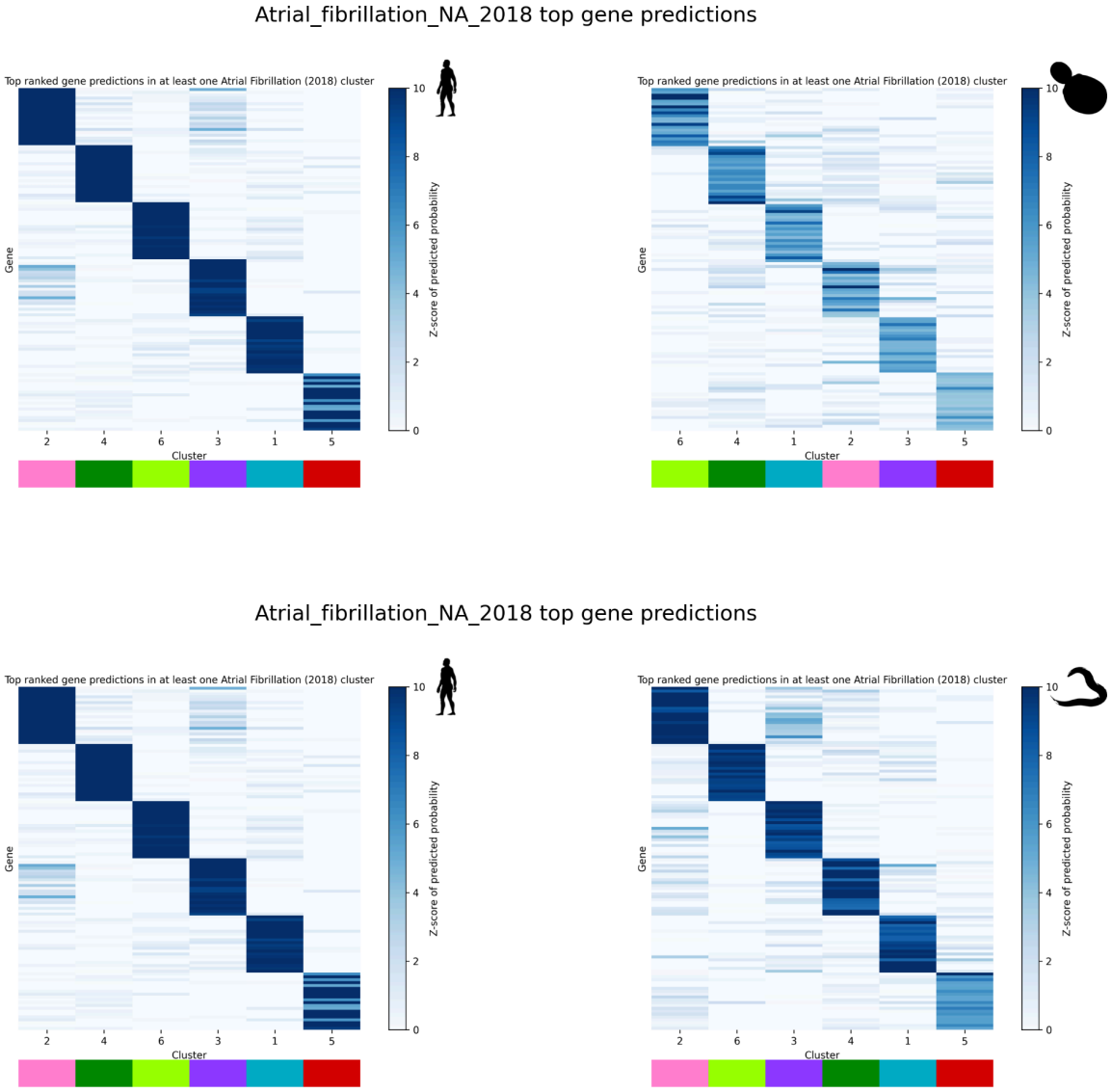

Figure A4.6: Gene predictions for yeast (top) and worm (bottom) for atrial fibrillation.

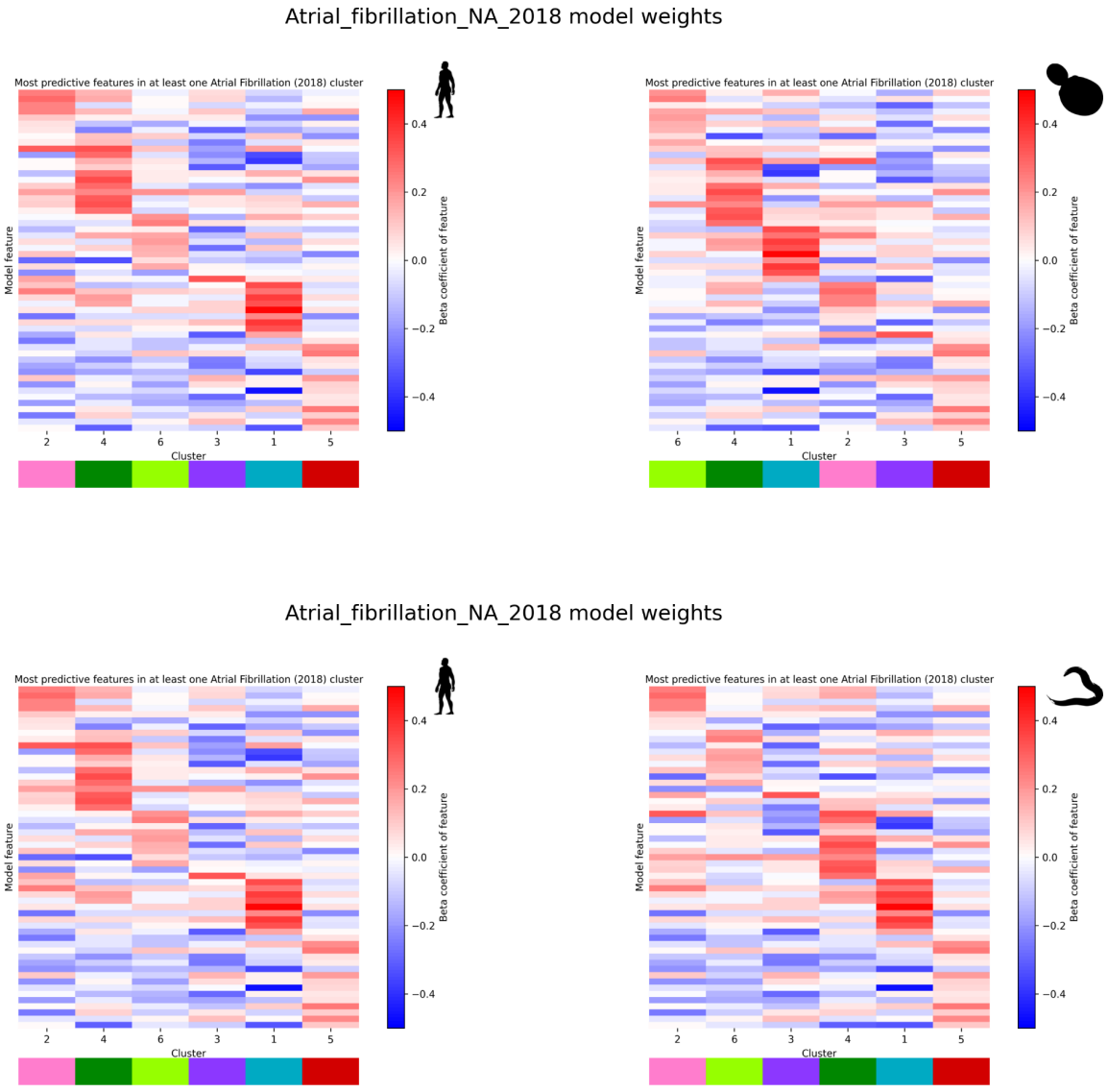

Figure A4.7: Comparing model weights for yeast (top) and worm (bottom) across modules for atrial fibrillation.

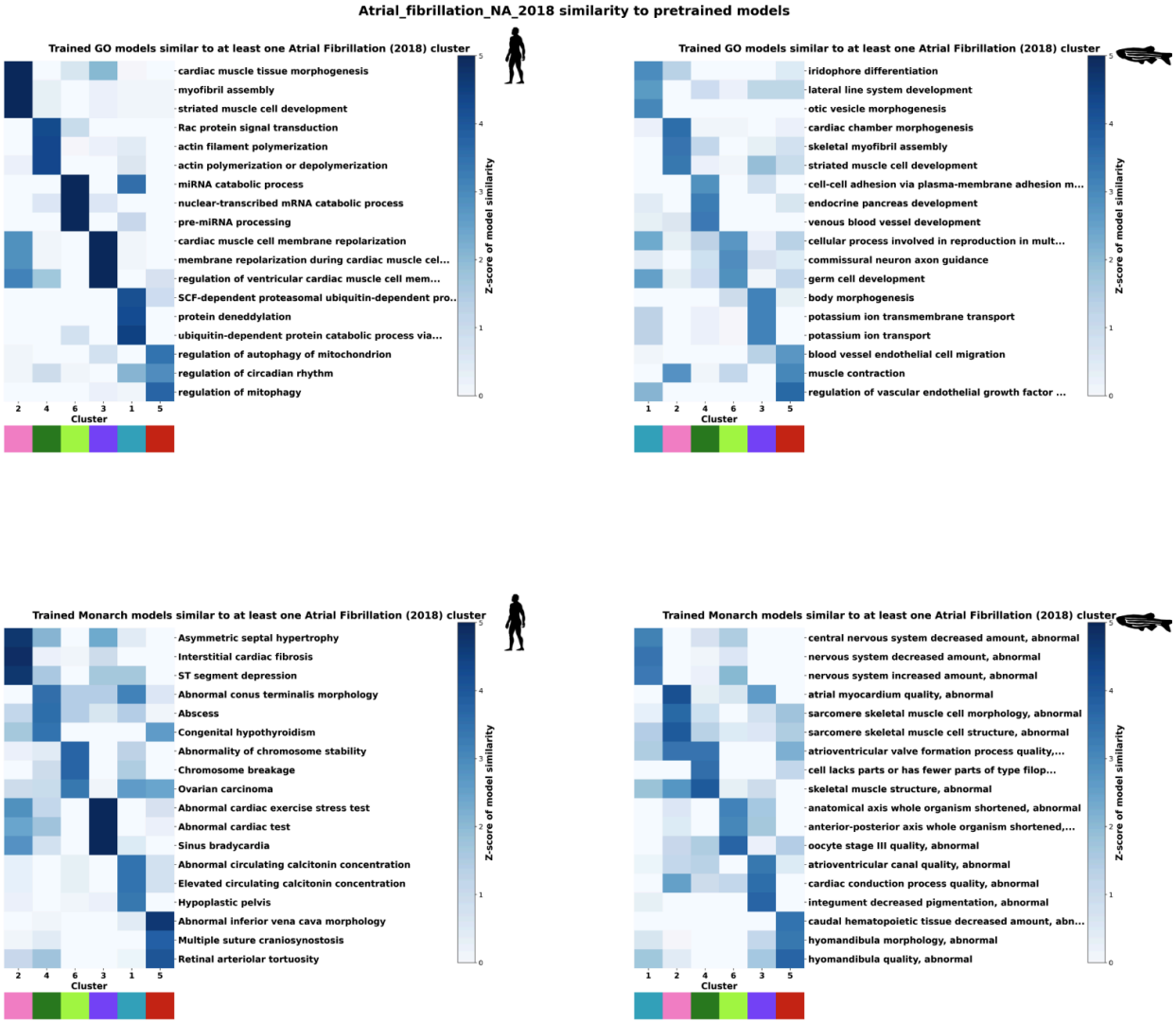

Figure A4.8: GOBP and phenotype enrichments for zebrafish for atrial fibrillation across modules.

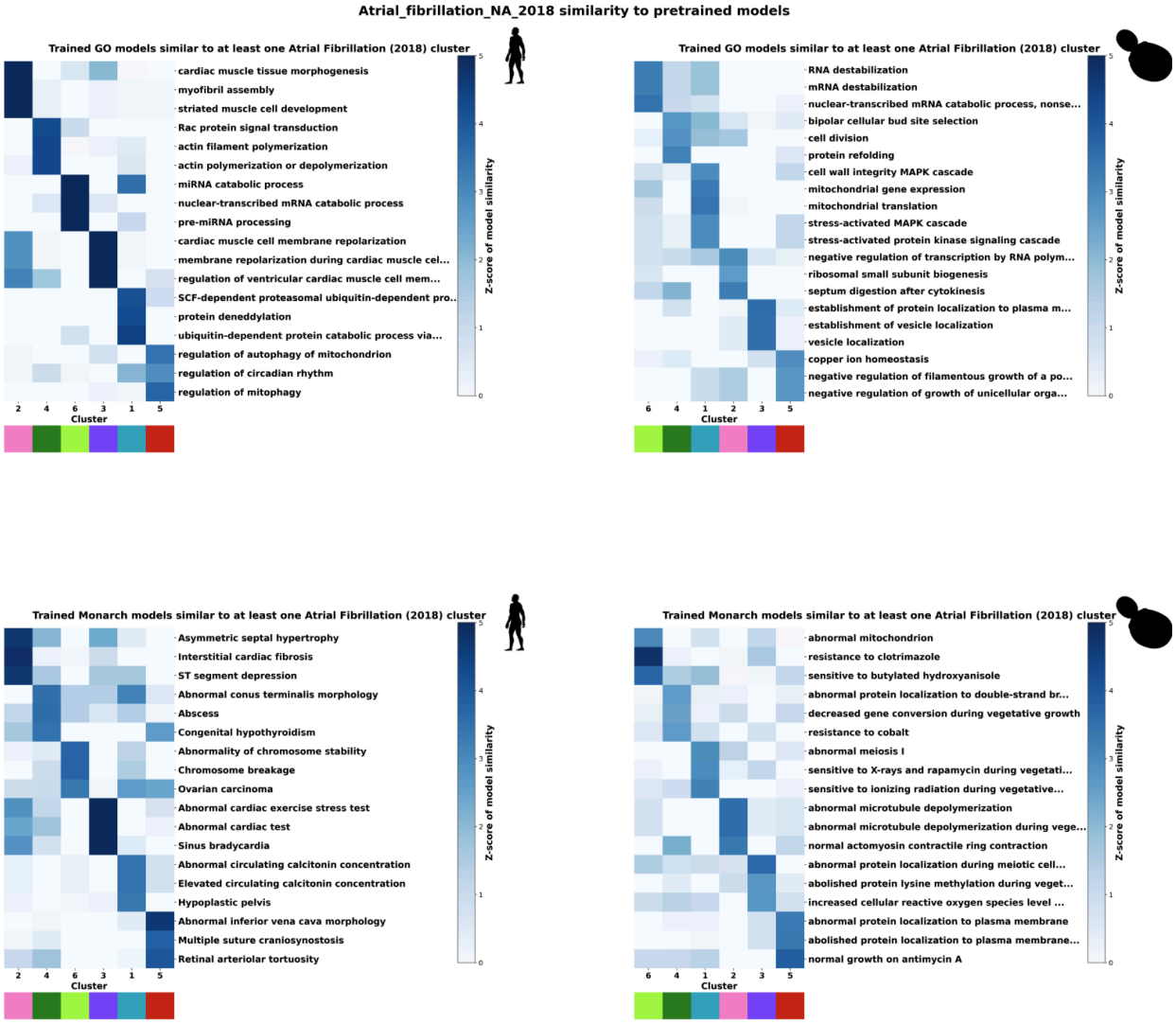

Figure A4.9: GOBP and phenotype enrichments for yeast for atrial fibrillation across modules.

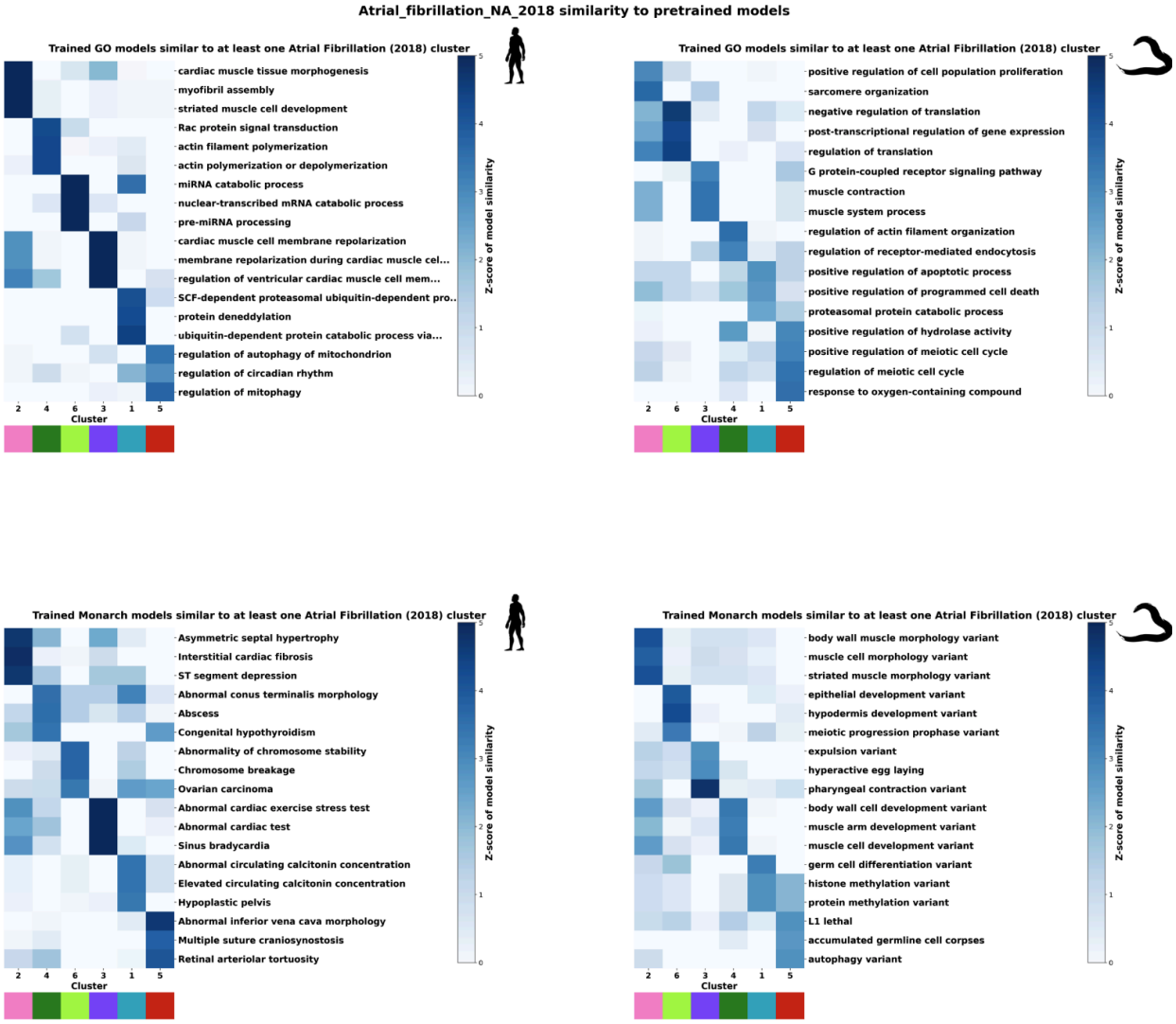

Figure A4.10: GOBP and phenotype enrichments for worm for atrial fibrillation across modules.

CHAPTER 5: SUMMARY, LIMITATIONS, REFLECTION, AND FUTURE DIRECTIONS

Summary

A primary goal of this dissertation was to propose methods to unravel the complexity of human disease by providing insights into the validity of interpreting diseases as multiple distinct subsets. These projects show how applying our methods to experimental data that is very complex and noisy can lead to discovery of relevant biology. These results are crucial because we know that diseases are modular and composed of multiple subsets of genes that contribute to distinct aspects of disease biology, but explaining how this biology applies to human disease remains extremely challenging. The discussed projects are united by using network biology to analyze complex human disease and traits to create general-purpose methods that provide biological insight across many different diseases. In chapter 2 we validate that we can use modules to discover relevant chronic inflammation genes, processes, and drug repurposing candidates for multiple common human complex diseases which are known to have inflammatory responses, but for which the genetic basis and mechanism had not been unraveled. By comparing modules across diseases we can learn which diseases manifest inflammatory phenotypes in a similar manner. The next projects in the dissertation delve deeply into aspects of the pipeline used in chapter 2. In chapter 3, we implemented a new method to utilize module information to improve disease gene classification by refinement of an initial geneset to remove false positives and reclassify false negatives. This method allows for easier discovery of enriched, biologically relevant phenotypes and processes with increased statistical power. Using modules, we uncovered biology relating to experimental datasets that were not found when considering the dataset as a whole. Lastly, in chapter 4 we demonstrated a pipeline where genotype-derived GWAS data for diseases can be broken up into smaller network-driven modules to improve performance of gene classification, then aggregated back into a final predicted disease module. This technique -- validated and justified in chapter 3 – allowed us to use the omnigenic framework to explain the biology of complex disease through classifying functional relevant core genes, relevant tissues and cell types, relevant enriched GOBPs, and genes that have important orthologs in

model organisms. Our method for classifying core genes is general-purpose and does not require assumptions about gene knowledge of specific diseases. These core genes are functionally relevant, and definitionally have many relationships to other disease genes. Chapter 2 has been published and the code released, while the code for chapter 3 is soon to be ready for release after publication. We implemented general-purpose methods that work with large amounts of diverse gene sets spanning multiple types of experiments for multiple complex diseases and traits. Interpreting diseases in the contexts of modules improves the performance of common computational methods, and is key for discovering how mechanistically relevant genes contribute to disease from highly heterogeneous data.

Limitations

This dissertation makes strides in implementing and improving general-purpose computational methods and delving deep into considering how to best learn more about disease biology. There are notable limitations that we had to work around to obtain meaningful results. Broadly, a clear issue is the limited knowledge of complex disease biology in the form of gold standards. Two examples of this limitation are for modules and for core/peripheral genes in an omnigenic framework. We have justified our use of modules for computational methods and biological insight despite there being no ground truth. True modules for disease are undiscovered because the question of what a module is is controversial in the first place. There is no correct answer on what is an ideal sized module, how many processes it would be enriched for, what types of phenotypes or how many, and many more biological questions. Despite their ambiguity, we showed that modules are useful tools for the methods we implemented. We are not claiming that the modules used for each geneset are the final modules that would ever be discovered for any disease. Modules will always change as both geneset annotations and networks advance as more information is discovered, but this is beneficial as it will make our methods even more useful when more biological data is annotated.

A second place where gold standards are highly limited is in core and peripheral genes. No genesets of these exist for disease. In fact, one of the critiques of the omnigenic model is that the biology behind core genes is not clearly articulated[1]. In chapter 4,

since there is no gold standard, we had to choose our own definition of core genes. First we derive a disease module by implementing ModGenePlexus to predict genes that are likely missing and adding these to our original disease gene module. Then, we define core/peripheral genes using this module, where the genes with the highest betweenness centrality values are considered core genes. This definition utilizes only network connections to determine what genes are core, with the idea being that there will be a small subset of genes that are highly connected to the rest of the hundreds of disease genes and that these highly connected genes are more likely to be directly involved in an important disease mechanism. These genes will have the highest betweenness centrality because they are the most connected to the other disease genes. This was done to create a general purpose method for predicting core genes that can be investigated, rather than relying on prior assumptions about biology to define potential core genes.

Reflection: Discovering chronic inflammation processes for complex disease

In chapter 2, we asked the question of whether we can isolate inflammation processes for complex diseases that are known to have a relationship to inflammation, but the biology behind that relationship is unclear. We introduce inflammation as an endophenotype for complex disease, meaning it is an intermediate phenotype that underlies disease. We are not the first to consider inflammation as an endophenotype in general or for specific complex diseases[2–5], and we consider the concept invaluable because it is known that multiple distinct phenotypes underlie complex disease. Answering how inflammation underlies each complex disease we investigated in chapter 2 would not only give a genetic explanation for the inflammation, but also could allow the development of hypotheses of whether inflammation is a good indicator of a disease being manifested in the first place. The motivation for the drug repurposing method is to find drugs that target the specific inflammation component of complex diseases. We validate that this is possible, but the usefulness depends on whether inflammation is considered a highly important process underneath the disease. For non-autoimmune disorders, these answers are still unknown. We put a lot of thought into how to relate chronic inflammation to disease, and this is where the idea of using network-based gene classification came from. We originally tried using DIAMOnD[6], but

this failed because we would get very large genesets that did not mean anything. Using GenePlexus allowed for a refinement of the genelist to one that passed our designed permutation tests for finding biological associations between actual discovered disease clusters and inflammation. This chapter resulted in the completion of a general pipeline that answered specific questions about inflammation in complex disease. We showed that using modules is valid in discovering meaningful biological information about a particular phenotype and predicted drugs that targeted that phenotype at the module level. The future projects in this dissertation modify and improve aspects of this pipeline.

Reflection: Implementing ModGenePlexus as an extension of GenePlexus

In chapter 2 we used a simple process for discovering disease clusters. We subsetted the network to include disease genes and clustered this disease-specific subnetwork using the Leiden Clustering algorithm[7]. Two considerations of this method are (i) we cluster after expanding the entire disease list, and (ii) this algorithm is run on a disease subnetwork, which means it does not take the rest of the genome into account. A major theme of this dissertation is that diseases are made up of multiple meaningful subsets of genes where not all genes interact equally together. Because of this, running GenePlexus on an entire disease geneset at once is questionable. If we want to gain insight into the inflammation component of a specific disease, why would it make sense to expand on other processes that have nothing to do with inflammation? This is the original motivation behind ModGenePlexus. Implementing this method ran into numerous issues. This can be seen in the simulation where fake traits were created from multiple GOBP. ModGenePlexus here re-created the performance of simply creating one model at once. This result was initially surprising but is ultimately a reflection of the fact that GenePlexus is a very good method optimized for predicting well-annotated genesets, and improving a good method is a challenging task. These challenges are what led us to integrate multiple computational methods that all work together to improve results. Most importantly, this was our motivation for focusing on real-world experimental data for which GenePlexus performance is typically worse than on well-annotated genesets. The primary goal of the GenePlexus software is to make it more usable for tasks scientists actually perform, like downstream analysis of experimentally determined genesets. ModGenePlexus directly improves GenePlexus

through additional geneset refinement. For the analysis where we train on nominal p-value genes of MAGMA predictions and evaluate on stringent p-value understudied genes, this was born out of the observation that the stringent genesets performed poorly with ModGenePlexus. ModGenePlexus does well the more complex and large the initial input is. This is because larger genesets allow for more dense and complete clusters to be discovered. This is an ideal result – where our method improves for data that performs particularly poorly.

Reflection: Interpreting diseases omnigenically through geneset refinement of noisy experimental data and geneset refinement

GWAS data is useful because it has become relatively straightforward and cheap to sequence participants. However, an increased number of studies with increased sample sizes has created many problems for interpretation and disease etiology. Our reliance on biological networks in these projects, and our interest in the vast complexity of many human traits means the omnigenic model is a very important concept to us. We considered for a long time how to create a general-purpose method that can predict core genes across multiple diseases. Rather, we leveraged the network and utilized it directly in defining core genes as those that are very important in the context of all disease gene annotations. The methods we implemented in this dissertation were all essential to be able to do this final chapter. We used GenePlexus to do gene classification on the entire genome, ModGenePlexus for improved performance on large scale experimental datasets like GWAS, and modules to investigate how core genes relate to underlying disease biology. Integrating GenePlexusZoo introduces the ability to transfer this knowledge to other species and discover meaningful orthologs that can be used in hypothesis generation to explain relevant aspects of disease biology . Atrial fibrillation is the perfect complex disease to use as a test case as it is very complex, common, has hundreds of gene annotations, and is involved in multiple diverse tissue and cell types. We provide ample evidence that our method for calculating core genes finds genes relevant to disease function.

Overall reflection: Leveraging complexity and gene relationships is what will uncover new breakthroughs in understanding human complex trait etiology

A generation ago the scale of complexity of human traits was still unknown[8]. GWAS is what revealed that the number of relevant loci was not in the dozens, but much larger[9]. Many methods are based on trying to isolate highly significant, large effect size genes and focusing on them, rather than focusing on using the underlying polygenic architecture because of the challenges it presents. However, we show in this dissertation that utilizing large amounts of genes is useful in multiple computational contexts. We find relevant inflammation phenotypes for diseases where inflammation is harder to define when considering the hundreds of gene annotations, we perform gene classification better for datasets containing thousands of genes – uncovering truly relevant biology, and we discover functionally relevant core genes. Genome-wide networks are a tool built from countless studies and datasets of vast amounts of biology, and the gene relationships are meaningful. By leveraging this data, we can provide context to hundreds of relevant genes recovered in experimental datasets. We saw how actually using less significant genes led to better prediction of highly significant genes in GWAS in chapter 3, and in chapter 4 we show that using GenePlexus to create a disease module, rather than using only the very few top predictions for hypothesis generation, allowed us to recover functionally relevant core genes and define them as core within a network. These methods and the positive results only make sense with the knowledge that diseases are highly complex and contain hundreds of true gene annotations. Understanding this polygenic environment will be vital in the future for being able to interpret mutations that are very rare or are novel. Networks explain how unique mutations can cause the same disease across patients, because a module within a network is being disrupted. The technology to sequence people is here, but it is determining how a disease maps in biological networks that will explain disease genetics broadly and for specific individuals.

Future directions

In chapter 2, we would like to expand on the work done to investigate endophenotypes further. Specifically, we would like to use GWAS of molecular, directly measurable traits that are “proxies" for higher level complex disease that is difficult to diagnose genetically

and phenotypically. A classic example of a disease with clinically useful endophenotypes is schizophrenia, where an example is prepulse inhibition, the inhibition of the startle reflex in response to weak prestimulus[10]. A vital question is whether molecular traits that underlie inflammation have significant enrichment with complex diseases with inflammation components. The challenge of using the inflammasome or other large entities like the thrombosome and fibrosome as endophenotypes[5] is that it is unclear how to use that knowledge in clinical settings, as there are multiple biomarkers for inflammation[11]. One way to do this proposal is to use GWAS Atlas as a source because it contains many molecular phenotype summary statistics. We can determine which phenotypes relate to chronic inflammation through either gene overlap or GenePlexus, and then see if these phenotypes also relate to complex disease using methods proposed in chapter 2-4, integrating what we have learned about how to utilize the polygenic architecture of diseases to our advantage.

For chapter 3, the primary future directions for this project are to integrate it into the GenePlexus webserver[12]. This is a technical challenge in itself given the space requirements needed to create many models for a disease rather than a single one. A biological direction to take this project is to directly compare the results of multiple GWAS that were conducted on the same disease, and see how modules and results differ. This would be especially interesting as we could quantify how modules change over time in terms of gene assignments and size when comparing older GWAS to newer ones.

For chapter 4, the primary goal is to do cross-GWAS analysis at the core gene level and validate making model organism phenotype predictions at the module level. These questions are vital in terms of being able to answer and state implications about the omnigenic model in the context of networks for human disease, and in being able to transfer this knowledge to other organisms using our methods. This knowledge transfer is important for generating hypotheses for further biological research.

REFERENCES

1. Sinnott-Armstrong, N., Naqvi, S., Rivas, M. & Pritchard, J. K. GWAS of three molecular traits highlights core genes and pathways alongside a highly polygenic background. *eLife* **10**, e58615 (2021).

2. Barber, R. *et al.* P3‐267: An inflammatory endophenotype of Alzheimer's disease. *Alzheimer's & Dementia* **6**, (2010).

3. Guglielmo, R., Miskowiak, K. W. & Hasler, G. Evaluating endophenotypes for bipolar disorder. *Int J Bipolar Disord* **9**, 17 (2021).

4. Genkel, V. V. & Shaposhnik, I. I. Conceptualization of Heterogeneity of Chronic Diseases and Atherosclerosis as a Pathway to Precision Medicine: Endophenotype, Endotype, and Residual Cardiovascular Risk. *Int J Chronic Dis* **2020**, 5950813 (2020).

5. Ghiassian, S. D. *et al.* Endophenotype Network Models: Common Core of Complex Diseases. *Sci Rep* **6**, 27414 (2016).

6. Ghiassian, S. D., Menche, J. & Barabási, A.-L. A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome. *PLoS Comput Biol* **11**, e1004120 (2015).

7. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* **9**, 5233 (2019).

8. Baxter, R. C., Martin, J. L. & Beniac, V. A. High molecular weight insulin-like growth factor binding protein complex. Purification and properties of the acid-labile subunit from human serum. *J Biol Chem* **264**, 11843–11848 (1989).

9. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).

10. Greenwood, T. A., Shutes-David, A. & Tsuang, D. W. Endophenotypes in Schizophrenia: Digging Deeper to Identify Genetic Mechanisms. *J Psychiatr Brain Sci* **4**, e190005 (2019).

11. Liu, C. H. *et al.* Biomarkers of chronic inflammation in disease development and prevention: challenges and opportunities. *Nat Immunol* **18**, 1175–1180 (2017).

12. Mancuso, C. A. *et al.* GenePlexus: a web-server for gene discovery using network-based machine learning. *Nucleic Acids Research* **50**, W358–W366 (2022).

CHAPTER 6: INTEGRATING GENE PRIORITIZATION METHODS FOR SUMMARY GWAS STATISTICS

Predicting (prioritizing) which genes are targeted by SNPs in GWAS is a difficult challenge because most discovered SNPs fall within non-coding regions of the genome[1]. Multiple methods have been implemented to perform the task. In chapter 1 we discuss methods that utilize physical distance of implicated SNPs to genes in the genome. These are methods such as MAGMA[2] and Pascal[3]. Chapter 1 additionally discusses that these methods are quite good at finding the genes most SNPs target. However, they are quite naive and other methods have been implemented that integrate other biological data to better predict genes certain SNPs target.

One such method is known as Transcriptome Wide Association Studies (TWAS). Briefly, what makes TWAS unique is that it uses gene expression data – specifically eQTL associations – to prioritize the genes that SNPs target. If a SNP is associated with expression of a particular gene, and if that SNP is implicated in a GWAS, then the gene it is a eQTL for is the gene that should be associated, not simply the genes nearby. A primary example of a TWAS method is the PrediXcan family[4–6]. PrediXcan uses transcriptome data sets from GTEx[7,8] to build models that relate SNPs to gene expression within each tissue. Each tissue in GTEx had a model built for it because gene expression is different depending on biological context. Next, these models are used to associate the genotypes of participants in the GWAS cohort to predict expression for those individuals. The expression is predicted for the specific tissue model used. Lastly, this predicted expression is associated with GWAS case/controls. If a SNP is seen in the cases quite often, then the model chooses targets based on how that SNP is related to gene expression data.

A third method for prioritization is to integrate Hi-C data to discover interactions. TWAS studies use expression data to relate implicated SNPs to genes, and Hi-C has a similar motivation by investigating how genes in enhancer regions interact with promoter regions of genes. When the DNA is unwrapped in the nucleus, the DNA is moving around in the cell environment, and this can lead to interaction between regions of DNA that can be far away from each other. This is how proteins bound to enhancer regions can target proteins that are far away from a kilobase perspective, but interact in the

3-Dimensional cell environment. Using Hi-C data with GWAS data means that if an implicated SNP falls within an enhancer region, the regions on the genome the enhancer are known to interact with are considered as targets for those SNPs. Multiple methods have been implemented that integrate Hi-C data[9,10].

At the start of this PhD program, the first project that I worked on was not any of the chapters presented in this dissertation. It was to build a method that works across many GWAS summary statistics to improve gene prioritization from SNP data using multiple types of data and methodology including locality base-methods, TWAS, a novel implementation to integrate Hi-C data in prioritization, and a novel network-based method to integrate the predictions. Specifically, we designed a project to create a general purpose method that would, for each GWAS, perform gene prioritization using a physical distance method (Pascal), a TWAS method (PrediXcan), and a new implementation for integrating Hi-C data to prioritize summary GWAS statistics from UK Biobank[11].

The first implementation for this project was modification to the Pascal method that would integrate Hi-C data indicating enhancer/target regions within the genome. We modified the Java source code for Pascal to integrate and use this data. The window for considering genes was defined as the target regions of enhancer SNPs. If they targeted gene promoters, these genes were considered for prioritization. This method was termed 3DPascal. This method was implemented and results were compiled, but there were numerous difficulties. A major challenge with this method is that it also requires tissue and cell specific data, as gene regulation and 3D contacts are context specific[5,12]. We ran 3DPascal for every GWAS, for 21 Hi-C datasets of tissues and cell lines[13]. A major bottleneck here was computational scalability, as running this in parallel for thousands of GWAS would take days.

In 2019 and 2020, the software H-MAGMA[9] was released and then published. H-MAGMA modified the original MAGMA method for gene prioritization based on physical distance to integrate Hi-C information. A motivation of 3DPascal was to create a user friendly method for doing this by modifying a well validated gene prioritization method with open source code, and H-MAGMA accomplishes that.

After the publication of this method, we focused on integrating results of different prioritization methods into one final prediction score for the gene. We wanted to use networks to accomplish this. Nodes in a network would be weighted by how many gene prioritization modalities implicated them, and we would do a random walk to see how often genes are in the paths between other implicated nodes. We first investigated how often modalities predicted genes, noticing that if genes were ever predicted, 3DPascal would often find them (**Figure 6.1**). This could mean that there are a lot of false positive predictions in 3DPascal and would further motivate using a network to provide context in which predicted genes are good. While implementing this method, a new study[14] was released that integrates 13 different methods to link SNPs to genes, including physical distance, enhancer gene linking, Hi-C data and scATAC-seq data. They show increased performance for prioritization with their method compared to other prioritization aggregation methods, and do analysis on dozens of GWAS from UK Biobank. In other words, this method and study was an expanded version of what we were planning to release.

While these projects ultimately did not finish, it was still foundational in influencing our views of complex disease and traits in humans. Gene prioritization is a challenge with GWAS because of the polygenic architecture complex traits have. It is not obvious how to relate non-coding variants to genes, and to improve results when multiple types of biological data are being integrated. Giving context to entities like SNPs and genes allows for the discovery of novel biology. The idea to utilize biological networks as a centerpiece in our work also came about with trying to integrate gene prioritization scores across methods. The idea is that if gene predictions are meaningful, they should be well connected within a biological network. This was the first project that made us consider the nuances in integrating experimental results like GWAS with a network to denoise and better refine the results. Using a network specifically to remove false positives from our 3DPascal motivation is very analogous to our using a network to remove false positives of experimental genesets in chapter 3. This project was where that idea initially came from.

Additionally, many of the methods used in these first projects were directly used in chapter 2-4 of this dissertation. In chapter 2, we utilized the non-disease GWAS from UK Biobank to compare them to complex disease gene classification predictions, and if diseases had more enrichment to inflammation relative to traits we know are not related. In chapters 2-4, we used gene prioritization results for summary statistics to do some methods or to analyze method validity with genotypic data. In chapter 2, we used Pascal[3], which uses physical distance of SNPs to genes to predict which genes are relevant based on the significant SNPs found in the GWAS study. MAGMA is a similar method that we used for chapters 3 and 4.

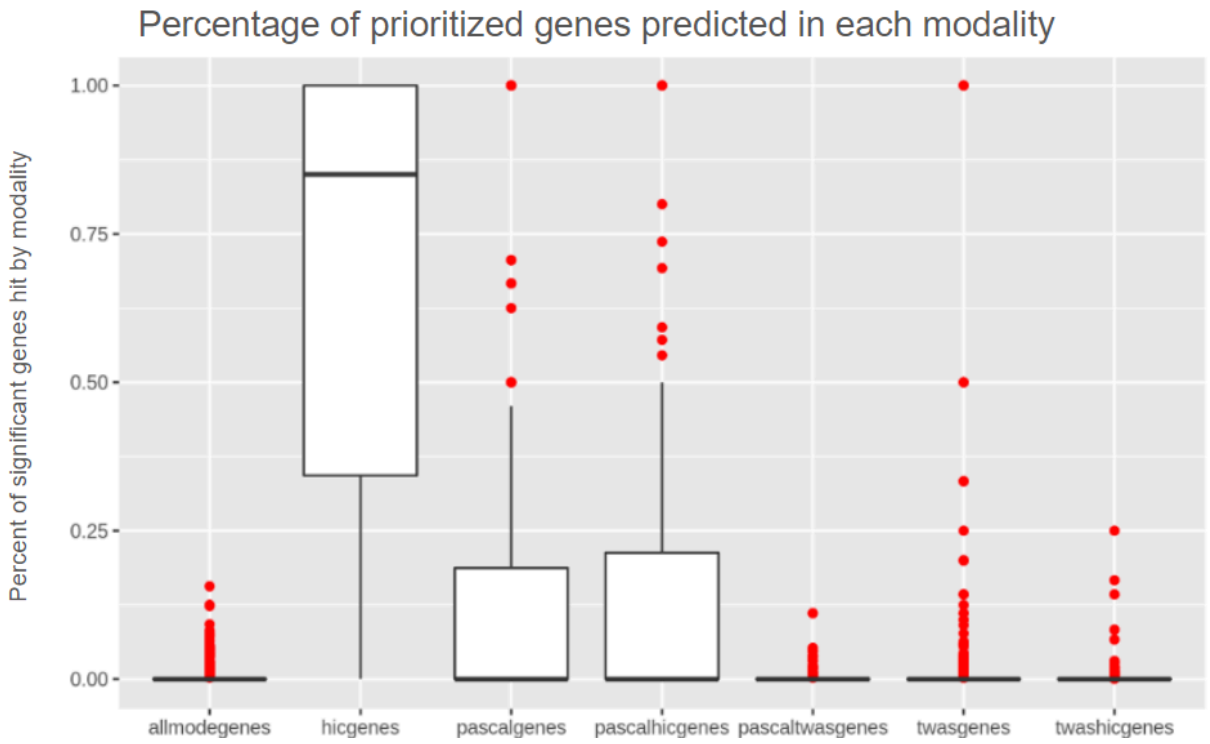

Figure 6.1: The percentage of gene hits for GWAS studies by each modality or combination of modalities. Notably, 3DPascal (hicgenes) hits most of the genes that are predicted by any method, and Pascal and TWAS only predict a small percentage of genes because 3DPascal predicts so many. This was our motivation for using networks to integrate the results, as it would hopefully remove false positive 3DPascal results that do not relate well to the other predictions.

REFERENCES

1. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 9362–9367 (2009).

2. Leeuw, C. A. de, Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Computational Biology* **11**, e1004219 (2015).

3. Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z. & Bergmann, S. Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. *PLoS Comput Biol* **12**, e1004714 (2016).

4. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* **47**, 1091–1098 (2015).

5. Barbeira, A. N. *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat Commun* **9**, 1825 (2018).

6. Barbeira, A. N. *et al.* Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS Genet* **15**, e1007889 (2019).

7. GTEx GWAS Working Group *et al.* Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *Genome Biol* **22**, 49 (2021).

8. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580–585 (2013).

9. Sey, N. Y. A. *et al.* A computational tool (H-MAGMA) for improved prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles. *Nat Neurosci* **23**, 583–593 (2020).

10. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826 (2017).

11. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine* **12**, (2015).

12. Pal, K., Forcato, M. & Ferrari, F. Hi-C analysis: from data generation to integration. *Biophys Rev* **11**, 67–78 (2019).

13. Schmitt, A. D. *et al.* A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Reports* **17**, 2042–2059 (2016).

14. Gazal, S. *et al.* Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. *Nat Genet* **54**, 827–836 (2022).