

DATA-DRIVEN COMPUTATIONAL APPROACHES TO UNRAVEL AGE/SEX BIASES  
& CROSS-SPECIES ANALOGS OF COMPLEX TRAITS AND DISEASES

By

Kayla Johnson

A DISSERTATION

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

Biochemistry and Molecular Biology – Doctor of Philosophy  
Computational Mathematics, Science, and Engineering – Dual Major

2022

## **ABSTRACT**

Cellular mechanisms and genetic underpinnings of most complex diseases and traits are not well understood. Most diseases also vary in their incidence and presentation in people of different ages and sexes, yet it is still largely unclear how age and sex influence normal tissue physiology and disease at the molecular level. Additionally, while we need research organisms to experimentally study many aspects of human disease etiology, choosing the best genes and conditions in a model organism for such studies is difficult due to our incomplete knowledge of functional and phenotypic conservation across species. The goal of my research is to address these challenges towards gaining a systematic understanding of the genetic etiology of complex diseases and traits. I have worked towards this goal by developing computational frameworks capable of leveraging massive amounts of publicly-available genomic data with prior knowledge using network analysis and machine learning. These approaches have shed light on the genomic signatures, pathways, and interactions that characterize the age/sex biases and cross-species analogs of complex diseases and traits. I make all the code to reproduce these approaches available by github and have provided tools to make the results searchable by scientists investigating these important biological factors. Collectively, this research will help build infrastructure for advancing biomedical research into the era of precision medicine.



This dissertation is dedicated to the inquisitive little girl I used to be, who had bigger dreams than I could hope to live up to but would have been so proud to be a scientist.

Also to my parents, Carrie and Brogan Johnson, and my siblings, Tyler, Paige and Solei, for being supportive and such a big part of the person I turned out to be. Finally, this dissertation is dedicated to my husband, Riley Mattes, who has grown with me, cheered me on, and been my best friend since we met.

## **ACKNOWLEDGEMENTS**

I would like to thank and acknowledge the people who helped my development as a scientist to prepare me for graduate school at Central Michigan University, Dr. Ajit Sharma and Dr. Douglas Swanson, and at NanoSynthons LLC, Dr. Donald Tomalia, Dr. David Hedstrand and Linda Nixon. I would also like to thank my committee members, Drs. Amy Ralston, Jianrong Wang, David Arnosti, and Charles Hoogstraten, for their input on my research ideas, professional development opportunities, and wisdom in other aspects of obtaining a PhD. I thank Dr. Shin-Han Shiu for contributing ideas and feedback to my comprehensive proposal and Dr. Janani Ravi for mentorship and support throughout my PhD. The hundreds of labs that produce data from many experiment types have also contributed significantly to this work. I would like to thank all past and present members of the Krishnan lab for their help, feedback, and moral support as I completed this work. Specifically I want to thank Dr. Chris Mancuso for his help when I was learning to code, offers to be a sounding board, and continued support throughout my entire dissertation work; Dr. Stephanie Hickey for sharing her expertise in experimental work, writing feedback, and overall support since she joined the lab; Dr. Sarah Percival for her mathematician's perspective and encouragement; Nat Hawkins for his help with parallel computing and great moral support; Remy Liu for his thoughtful comments and support; Alex McKim for his writing feedback and supportive comments; and Hao Yuan for his enthusiasm in continuing work on coexpression networks. Finally, I would like to specifically thank my advisor, Dr. Arjun Krishnan, for always believing in me even when I doubted myself and building a fantastic lab for support.

## TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION.....	1
REFERENCES.....	11
CHAPTER 2: ROBUST NORMALIZATION AND TRANSFORMATION TECHNIQUES FOR CONSTRUCTING GENE COEXPRESSION NETWORKS FROM RNA-SEQ DATA.....	15
REFERENCES.....	60
APPENDIX.....	65
CHAPTER 3: LEVERAGING PUBLIC TRANSCRIPTOME DATA WITH MACHINE LEARNING TO INFER PAN-BODY AGE- AND SEX-SPECIFIC MOLECULAR PHENOMENA.....	87
REFERENCES.....	122
APPENDIX.....	129
CHAPTER 4: DISCOVERING ANALOGOUS GENES, PHENOTYPES, AND CONDITIONS ACROSS HUMAN AND MODEL SPECIES USING MACHINE LEARNING.....	154
REFERENCES.....	174
CHAPTER 5: SUMMARY, REFLECTIONS, LIMITATIONS, AND FUTURE DIRECTIONS.....	177
REFERENCES.....	184

# **CHAPTER 1: INTRODUCTION**

## **Overview**

Recent large-scale studies have documented hundreds of genetic variants and phenotypes associated with various diseases and complex traits in an effort to gain a population-level understanding of human health and disease [1–5]. These associations continue to be cataloged, revealing more chasms in our knowledge of the relationships between genomic variation, biological pathways, tissue physiology, and trait variation. This knowledge is critical for improving our ability to diagnose and treat complex diseases. In addition, a major method of studying particular facets of human disease is through the use of model organisms, but transferring knowledge gleaned from these organisms back to human biological insight is often challenging [6]. The goal of this PhD research is to provide insight into the genomic signatures, pathways, and interactions that characterize the age/sex biases and cross-species analogs of complex diseases and traits. This chapter will provide the necessary background and context for these research goals, followed by the questions and objectives of the study, and concluding with the significance.

## **Background**

Interest in precision medicine has soared over the past decade [1,7,8]. Precision medicine strives to approach disease prevention and treatment in a way that takes into account individual variability in genetic background, environmental factors, and lifestyle choices with the goal of providing better health outcomes for all individuals. Despite this increased interest, we still lack a population-level understanding of cellular mechanisms and genetic underpinnings of most complex diseases and traits. Without a

comprehensive framework to delineate relationships between genomic variation, expression signatures, gene interactions and pathways, cellular networks and physiological function, we lack the tools to bring precision medicine to all areas of human disease prevention and treatment.

Age and sex are two biological variables that have been tied to variation in the incidence, presentation, and treatment response of complex traits and diseases [9–11], yet it is still largely unclear how age and sex influence normal tissue physiology and disease at the molecular level. This is largely a result of age and sex effects often being historically ignored in basic and clinical studies [12,13]. Sex has been especially neglected, due to several factors. A study in the early 1970s established that fluctuating levels of ovarian hormones could differ by up to fourfold in rodents [14]. Assuming that these hormonal differences would lead to more difficulty in analyzing data, scientists largely chose to avoid the issue by choosing to use male animals in their research [15]. This issue was further compounded by the 1977 Food and Drug Administration (FDA) policy recommending that women of childbearing potential be excluded from Phase I and II drug trials. This policy was not reversed until 1993, when the FDA required data analysis to include gender effect [16]. However, the National Institutes of Health did not start requiring the use of female animals in preclinical studies until 2014 [12]. That same year, Prendergast and colleagues released a meta-analysis of almost 300 studies using mice as research subjects that showed data collected from female mice did not vary any more than data from males and sometimes even showed less variation, regardless of the estrous cycle [17]. A follow-up study in 2016 replicated this result in an investigation of rat studies [18]. These policy changes and studies dispelling the notion that

hormones are a “female problem” in animal research are steps in the right direction, but there is still work to be done to overcome years of neglect.

Although age as a biological variable has not suffered as systematic exclusion as sex, it is still underconsidered in basic studies and clinical trials [13]. For instance, older adults are vastly underrepresented in clinical drug trials in spite of their overrepresentation in consumption of prescription drugs [19], and adolescents and young adults (ages 15-39) are less likely to participate in cancer clinical trials compared to younger children and older adults [20]. In addition to the problem that age and sex are both historically understudied biological factors on their own, many studies account for one or the other, but only accounting for one can yield an incomplete understanding. An example is that women have a lower incidence of stroke than men before menopause, but afterwards prevalence of stroke is higher in women [21]. A similar trend is observed for asthma, where prevalence is higher in boys than girls as children, but more common in women than men in adulthood [22].

New studies are now beginning to uncover some of the genetic basis that underlies age and sex differences in treatment response, tissue function, phenotypes, and diseases [11,23–30], but new data is not enough to uniformly address outstanding questions about female and male biology across the entire lifespan. We also need to leverage the hundreds of thousands of existing gene expression profiles that have been generated over the past 25 years and deposited in public repositories [31–34]. These samples capture gene expression under thousands of conditions, including different stages of disease and development. It has already been well established by members in our group [35–37] and other groups [38,39] that integrating large-scale -omic data,

particularly transcriptomes, and combining them with the scattered prior knowledge that we do have, can lead to major breakthroughs in delineating gene function and interactions in specific biological contexts. It therefore stands to reason that these data can also be leveraged to provide comprehensive frameworks that will help in gaining insights into age- and sex-specific molecular pathways in various tissues as well.

The frameworks created by computational methods that can use massive amounts of transcriptomic data and limited prior knowledge provide valuable tools for hypothesis generation and studying biological processes, but these hypotheses must be experimentally tested for validation. Many of these experiments necessary to understand cellular processes and genetic interactions driving the expression of disease are impossible to perform in humans, so we must use model organisms to functionally characterize these interactions *in vivo*. The ideal *in vivo* model for studying a particular facet of human disease should, for the most part, replicate a human phenotype and share the genetic underpinnings and mechanism of action. However, choosing the best phenotype in a model organism to study any given human disease or trait is difficult due to our incomplete knowledge of the relationships between phenotypes, genes, and conditions across species [40,41].

## **Research questions**

The increased efforts to catalog genomic variants and phenotypes associated with a wide variety of complex traits and diseases offers an opportunity to use this information to inform computational models. Members of our group [35,42] and others [43,44] have recently developed approaches for combining large genomic data collections with existing functional associations to bridge some of the gaps in our understanding of how

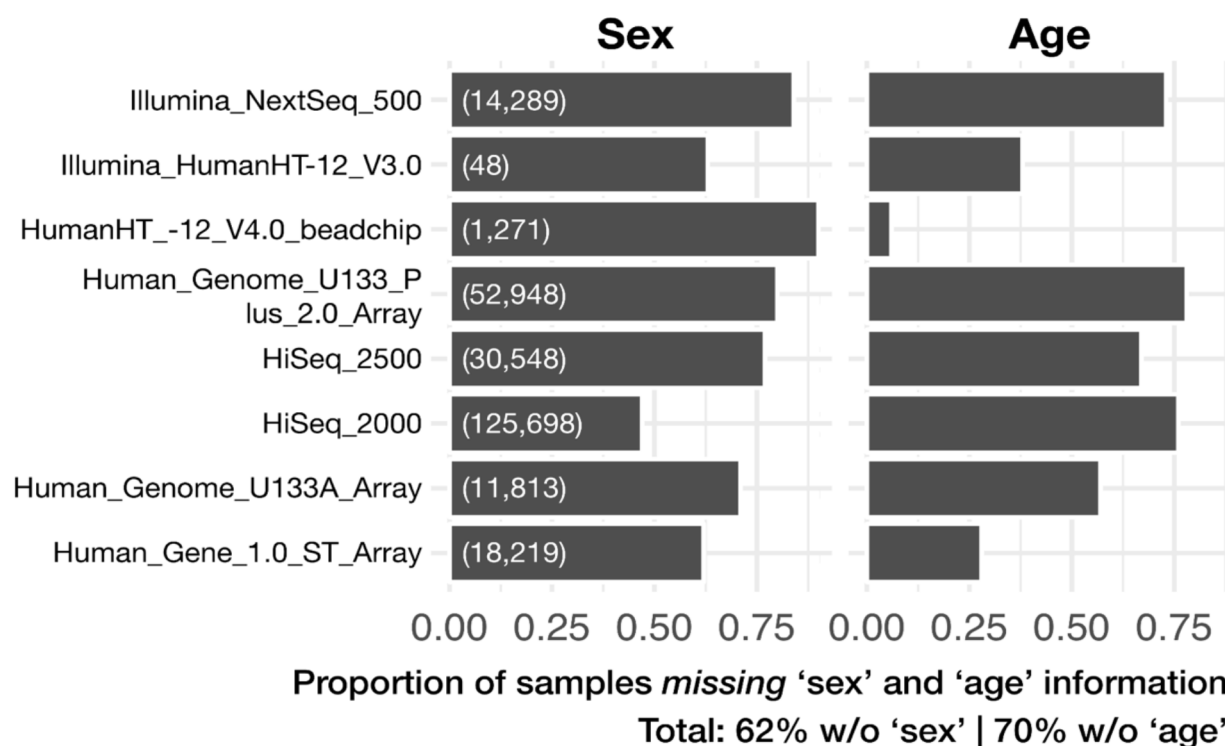
genetic variants, biological pathways, tissue function, and trait variation are related to each other. Krishnan et al demonstrated the value of leveraging computational models and large public transcriptome data for context-specific biology by using the first tissue-specific gene interaction networks to predict novel candidate genes, brain-specific pathways, and developmental stages related to autism spectrum disorder [42]. This study showed that using a tissue-specific gene interaction network for the tissue most affected by the disorder improves our ability to identify the most relevant genes and pathways for further inquiry. Some of these predictions have already been experimentally validated.

If tissue-specificity is able to enhance the insight we are able to derive from computational models, we should be able to further boost their accuracy by incorporating other biological contexts such as age and sex, which have proven to be crucial factors for prevalence and manifestation of disease as well as treatment response. However, two large obstacles stand in the way of building age- and sex-specific gene interaction networks to study how these factors influence cellular processes and the genetic etiology of complex diseases/traits.

The first obstacle is that while Greene, Krishnan, Wong and team were able to integrate transcriptomic data to build a tissue-specific gene interaction network, they did so using microarray data, without any RNA-seq data [35]. Since the time this study was published, the amount of RNA-seq data being deposited in public repositories has exponentially increased. As of October 2022, the ARCHS4 [45] repository contains over 620,000 human RNA-seq samples. This data is rich with context-specific information that should not be underused. The outstanding question is *how can we best build*



coexpression networks from heterogeneous RNA-seq data that comes from mostly small experiments generated by individual labs, with a range of sequencing depths and qualities, as well as high-quality consortium data? These coexpression networks can then be integrated into high-fidelity gene interaction networks with machine learning in the same way Greene, Krishnan, Wong and colleagues built their tissue-specific networks.



**Figure 1.1. Missing metadata.** Proportion of samples (x-axis) from eight major human gene-expression platforms (y-axis) that lack information about sex or age.

The second obstacle is that ideally we would employ all available transcriptomes for the most accurate resulting gene interaction network, but the vast majority of both microarray and RNA-seq samples in public repositories are missing information about both age and sex (**Fig. 1.1**). Lee and colleagues have demonstrated that it is possible to predict tissue of origin from gene expression data [36], and sex is quite easy to predict

given expression of sex chromosome genes. So, the remaining question is *can age or age group be predicted using only the gene expression values?* With the ability to predict age using only gene expression, we will be able to infer both age and sex for hundreds of thousands of transcriptomes, rendering all of them available for the study of age- and sex-specific processes and disease mechanisms. Further, if devised correctly, our prediction models could be biologically interpretable, *i.e.* the model will likely yield the strength and direction of importance of all the genes in the genome for each age group in each sex. This begs another question, *what do these gene signatures tell us about age- and sex-specific biological contexts?* We can use experimentally-validated genesets from different diseases, complex traits, phenotypes, tissues, and cell types to investigate this question.

Public repositories are not limited to only human data. There are also well over a million samples from model organisms covering a range of mutations, phenotypes, developmental stages, tissues, and experimental conditions [33,45]. These expression profiles represent a wide variety of biological contexts we have available to study human physiology and disease etiology. As previously mentioned, the ideal animal model for studying a specific aspect of human biology should not only display the desired phenotype but the underlying mechanism should also be as similar as possible. Although a similar expression pattern does not guarantee the same underlying mechanism, a model system that can replicate the transcriptomic landscape of a human sample is the best starting point to study a disease, trait, or treatment response of interest. So, the question is, *can we utilize massive public transcriptomic data to identify analogous samples, and therefore biological contexts and phenotypes in model species*

*that are most pertinent to human traits and diseases?* Achieving this goal would give us the ability to find genetic/experimental conditions that most closely match complex traits and diseases in humans in molecular mechanisms, improving our ability to translate functional results in model organisms back to humans. Prominent approaches for mapping related phenotypes across species rely on semantic similarity [46] of phenotypic descriptions, or consider the number of shared homologous genes that are annotated to each phenotype [47]. Semantic similarity methods ignore the genetic context of the traits and phenotypes completely by depending only on the text description of the phenotype, while methods that rely on homologous gene overlap fail in many cases due to our partial knowledge of the genes associated with any given trait or phenotype. To overcome these limitations, multiple studies have proposed directly matching samples across species based on their expression profiles [48–51]. However, as gene expression programs are shared across tissues, traits, and diseases, these methods do not place emphasis on context-specific molecular signals. The area of supervised machine learning (ML) is an enticing framework for tackling this problem. Specifically, by using not only transcriptomes from a given context (say, disease) but also transcriptomes from other contexts as a contrast, ML-based methods can automatically isolate context-specific gene expression signature, which can then be used to find samples in model organisms where this signature is active, thereby pointing to mechanistically-equivalent model systems.

## **Research summary**

In Chapter 2, I address the question: *how can we best build coexpression networks from heterogeneous RNA-seq data that comes from mostly small experiments*

*generated by individual labs, with a range of sequencing depths and qualities, as well as high-quality consortium data?* In this chapter, I elaborate on the most accurate and robust methods to build coexpression networks from RNA-seq data. I test multiple normalization and network transformation techniques and their combinations to make concrete recommendations of when and how to use these techniques.

In Chapter 3, I address two questions: (1) *can age or age range be predicted using only the gene expression values?* And (2) *what do these gene signatures tell us about age- and sex-specific biological contexts?* Here, I curate about 30,000 primary human transcriptomes and use these profiles to train machine learning (ML) models to predict age group. I also investigate age- and sex-biased gene signatures learned by these ML models using experimentally-validated genesets for determining enrichment of multiple biological contexts in different age and sex groups.

In Chapter 4, I address the question: *can we utilize massive public transcriptomic data to identify analogous samples, and therefore biological contexts and phenotypes across species?* In this chapter I describe our efforts to use ML in mapping transcriptomic landscapes and phenotypes across species to improve functional knowledge transfer.

## **Significance**

First, we address a gap in the literature involving the use of the continuously growing RNA-seq data in building coexpression networks. This work gives computational biologists clear directions for how best to integrate transcriptome-based networks into their framework. Second, to the best of our knowledge, we present the first study to utilize tens of thousands of publicly-available human gene expression profiles in the study of how biological processes change along the lifespan in both sexes. This work

will enable the prediction of age and sex labels for public expression profiles that lack this information, rendering these samples available to study age- and sex-specific genomics for the first time. Additionally, these age- and sex-labelled transcriptomic datasets enable computational researchers to make predictions about genes and pathways in age-, sex- and species-specific contexts, *even if those genes and pathways have never been functionally characterized*. Third, our current analyses of transcriptomes across species highlights key challenges in using ML to associate samples across species, pointing to both technical/experimental factors as well as biological conservation that need to be taken into account in future work by us and others.

Across all these projects, we release code to reproduce our methods and results so other computational biologists are able to verify our work and build upon and improve our methods. We will make our age- and sex-specific gene signatures, expression informations, and enriched biological genesets available for query for researchers studying these biological variables. In the near future, upon completion of the cross-species mapping work, we will also provide biomedical researchers with tools that enable them to search across species for samples with similar expression patterns to a query transcriptome from any species to aid in choosing suitable experimental settings. Collectively, completion of these aims will help build the infrastructure for advancing biomedical research into the era of precision medicine, benefitting the public that helps fund this work.

## REFERENCES

1. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, 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* [Internet]. 2015 [cited 2019 Aug 4];12. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4380465/>
2. Bernabeu E, Canela-Xandri O, Rawlik K, Talenti A, Prendergast J, Tenesa A. Sex differences in genetic architecture in the UK Biobank. *Nat Genet*. 2021;53:1283–9.
3. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics*. 2010;26:1205–10.
4. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536:285–91.
5. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res*. 2017;45:D896–901.
6. McGonigle P, Ruggeri B. Animal models of human disease: Challenges in enabling translation. *Biochem Pharmacol*. 2014;87:162–71.
7. The “All of Us” Research Program. *N Engl J Med*. Massachusetts Medical Society; 2019;381:668–76.
8. Ashley EA. Towards precision medicine. *Nat Rev Genet*. Nature Publishing Group; 2016;17:507–22.
9. Ober C, Loisel DA, Gilad Y. Sex-specific genetic architecture of human disease. *Nat Rev Genet*. Nature Publishing Group; 2008;9:911–22.
10. Khramtsova EA, Davis LK, Stranger BE. The role of sex in the genomics of human complex traits. *Nat Rev Genet*. 2018;1.
11. Cenko E, Yoon J, Kedev S, Stankovic G, Vasiljevic Z, Krljanac G, et al. Sex Differences in Outcomes After STEMI: Effect Modification by Treatment Strategy and Age. *JAMA Intern Med*. 2018;178:632–9.
12. Clayton JA, Collins FS. Policy: NIH to balance sex in cell and animal studies. *Nat News*. 2014;509:282.
13. Tannenbaum C, Day D. Age and sex in drug development and testing for adults. *Pharmacol Res*. 2017;121:83–93.

14. Shaikh AA. Estrone and Estradiol Levels in the Ovarian Venous Blood from Rats During the Estrous Cycle and Pregnancy\*. *Biol Reprod.* 1971;5:297–307.
15. Wald C, Wu C. Of Mice and Women: The Bias in Animal Models. *Science. American Association for the Advancement of Science*; 2010;327:1571–2.
16. Commissioner O of the. Gender Studies in Product Development: Historical Overview. FDA [Internet]. FDA; 2020 [cited 2022 Nov 14]; Available from: <https://www.fda.gov/science-research/womens-health-research/gender-studies-product-development-historical-overview>
17. Prendergast BJ, Onishi KG, Zucker I. Female mice liberated for inclusion in neuroscience and biomedical research. *Neurosci Biobehav Rev.* 2014;40:1–5.
18. Becker JB, Prendergast BJ, Liang JW. Female rats are not more variable than male rats: a meta-analysis of neuroscience studies. *Biol Sex Differ.* 2016;7:34.
19. Herrera AP, Snipes SA, King DW, Torres-Vigil I, Goldberg DS, Weinberg AD. Disparate Inclusion of Older Adults in Clinical Trials: Priorities and Opportunities for Policy and Practice Change. *Am J Public Health.* 2010;100:S105–12.
20. Forcina V, Vakeesan B, Paulo C, Mitchell L, Bell JA, Tam S, et al. Perceptions and attitudes toward clinical trials in adolescent and young adults with cancer: a systematic review. *Adolesc Health Med Ther.* 2018;9:87–94.
21. Haast RA, Gustafson DR, Kiliaan AJ. Sex Differences in Stroke. *J Cereb Blood Flow Metab.* SAGE Publications Ltd STM; 2012;32:2100–7.
22. Zein JG, Erzurum SC. Asthma is Different in Women. *Curr Allergy Asthma Rep.* 2015;15:28.
23. Ghosh S, Klein RS. Sex Drives Dimorphic Immune Responses to Viral Infections. *J Immunol.* 2017;198:1782–90.
24. Márquez EJ, Chung C, Marches R, Rossi RJ, Nehar-Belaid D, Eroglu A, et al. Sexual-dimorphism in human immune system aging. *Nat Commun.* 2020;11:1–17.
25. Vázquez-Martínez ER, García-Gómez E, Camacho-Arroyo I, González-Pedrajo B. Sexual dimorphism in bacterial infections. *Biol Sex Differ.* 2018;9:27.
26. Peters MJ, Joehanes R, Pilling LC, Schurmann C, Conneely KN, Powell J, et al. The transcriptional landscape of age in human peripheral blood. *Nat Commun.* 2015;6:8570.
27. Craig T, Smelick C, Tacutu R, Wuttke D, Wood SH, Stanley H, et al. The Digital Ageing Atlas: integrating the diversity of age-related changes into a unified resource. *Nucleic Acids Res.* 2015;43:D873–8.

28. Cypess AM, Lehman S, Williams G, Tal I, Rodman D, Goldfine AB, et al. Identification and Importance of Brown Adipose Tissue in Adult Humans. *N Engl J Med*. Massachusetts Medical Society; 2009;360:1509–17.
29. Stegeman R, Weake VM. Transcriptional Signatures of Aging. *J Mol Biol*. 2017;429:2427–37.
30. Costa AR, Lança de Oliveira M, Cruz I, Gonçalves I, Cascalheira JF, Santos CRA. The Sex Bias of Cancer. *Trends Endocrinol Metab*. 2020;31:785–99.
31. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res*. 2013;41:D991–5.
32. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, et al. ArrayExpress update—simplifying data submissions. *Nucleic Acids Res*. 2015;43:D1113–6.
33. Athar A, Füllgrabe A, George N, Iqbal H, Huerta L, Ali A, et al. ArrayExpress update – from bulk to single-cell expression data. *Nucleic Acids Res*. 2019;47:D711–5.
34. Leinonen R, Sugawara H, Shumway M. The Sequence Read Archive. *Nucleic Acids Res*. 2011;39:D19–21.
35. Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, et al. Understanding multicellular function and disease with human tissue-specific networks. *Nat Genet*. 2015;47:569–76.
36. Lee Y, Krishnan A, Zhu Q, Troyanskaya OG. Ontology-aware classification of tissue and cell-type signals in gene expression profiles across platforms and technologies. *Bioinformatics*. 2013;29:3036–44.
37. Zhu Q, Wong AK, Krishnan A, Aure MR, Tadych A, Zhang R, et al. Targeted exploration and analysis of large cross-platform human transcriptomic compendia. *Nat Methods*. 2015;12:211–4.
38. GTEx Consortium, Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat Commun* [Internet]. 2018 [cited 2018 Jun 1];9. Available from: <http://www.nature.com/articles/s41467-018-03621-1>
39. Oliva M, Muñoz-Aguirre M, Kim-Hellmuth S, Wucher V, Gewirtz ADH, Cotter DJ, et al. The impact of sex on gene expression across human tissues. *Science*. 2020;369:eaba3066.
40. Marian Ali J. Modeling Human Disease Phenotype in Model Organisms. *Circ Res*.



2011;109:356–9.

41. Aitman TJ, Boone C, Churchill GA, Hengartner MO, Mackay TFC, Stemple DL. The future of model organisms in human disease research. *Nat Rev Genet*. 2011;12:575–82.

42. Krishnan A, Zhang R, Yao V, Theesfeld CL, Wong AK, Tadych A, et al. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat Neurosci*. 2016;19:1454–62.

43. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype–phenotype interactions. *Nat Rev Genet*. 2015;16:85–97.

44. Krishnan A, Taroni JN, Greene CS. Integrative Networks Illuminate Biological Factors Underlying Gene–Disease Associations. *Curr Genet Med Rep*. 2016;4:155–62.

45. Lachmann A, Torre D, Keenan AB, Jagodnik KM, Lee HJ, Wang L, et al. Massive mining of publicly available RNA-seq data from human and mouse. *Nat Commun*. 2018;9:1366.

46. Köhler S, Doelken SC, Ruef BJ, Bauer S, Washington N, Westerfield M, et al. Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research. *F1000Research*. 2014;2:30.

47. McGary KL, Park TJ, Woods JO, Cha HJ, Wallingford JB, Marcotte EM. Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proc Natl Acad Sci*. 2010;107:6544–9.

48. Levin M, Anavy L, Cole AG, Winter E, Mostov N, Khair S, et al. The mid-developmental transition and the evolution of animal body plans. *Nature*. Nature Publishing Group; 2016;531:637–41.

49. Hashimshony T, Feder M, Levin M, Hall BK, Yanai I. Spatiotemporal transcriptomics reveals the evolutionary history of the endoderm germ layer. *Nature*. Nature Publishing Group; 2015;519:219–22.

50. Cardoso-Moreira M, Halbert J, Vallotton D, Velten B, Chen C, Shao Y, et al. Gene expression across mammalian organ development. *Nature*. 2019;1.

51. Le H-S, Oltvai ZN, Bar-Joseph Z. Cross-species queries of large gene expression databases. *Bioinformatics*. 2010;26:2416–23.

## **CHAPTER 2: ROBUST NORMALIZATION AND TRANSFORMATION TECHNIQUES FOR CONSTRUCTING GENE COEXPRESSION NETWORKS FROM RNA-SEQ DATA**

### **Background**

Constructing gene coexpression networks is a powerful and widely-used approach for analyzing high-throughput gene expression data from microarray and RNA-seq technologies [1]. Coexpression networks provide a framework for summarizing multiple transcriptomes of a particular species, tissue, or condition as a graph where each node is a gene and each edge between a pair of genes represents the similarity of their patterns of expression. Coexpressed genes are highly likely to be transcriptionally co-regulated and are often functionally related to each other by virtue of taking part in the same biological process or physiological trait [2–5]. Many studies have leveraged these properties to use coexpression networks in several important applications such as determining co-regulated gene groups [6] and associating genes to functions and phenotypes [7].

Nevertheless, multiple experimental factors impact the quantification of the expression of individual genes and the coexpression between pairs of genes, making it necessary to normalize and transform high-throughput gene expression data before downstream analysis. For RNA-seq data, examples of factors that affect the number of reads mapped to a gene include gene length, gene sequence, sample RNA population, and sequencing depth. Some factors have a greater effect on comparisons of gene counts within a single sample ('within-sample' effects) while others have a greater effect on comparisons of the same gene's counts in different samples ('between-sample' effects)

[8]. Many data normalization and transformation techniques have been developed to explicitly address one or more of these factors. An additional adjustment that can be considered particularly in coexpression analysis is network transformation, which is applied after calculating correlation between all gene pairs. Coexpression networks are noisy and can indiscriminately capture indirect interactions due to being estimated from noisy, steady-state gene expression data. Hence, previous studies have proposed methods to modify the raw coexpression network to upweight connections that are more likely to be real and downweight spurious correlations based on the topology of the network [9,10]. Together, appropriately normalizing and transforming RNA-seq data along with adequately transforming the coexpression strengths should yield more accurate estimates of gene-gene coexpression that best capture functional relationships between genes.

However, the best practices for normalization when building a coexpression network from a raw gene-expression dataset have been developed and compared only for data from microarrays [11,12]. Over the past decade, coexpression network analysis is being routinely applied to the exponentially increasing amount of data from RNA-seq, even though the optimal procedure for network building has not been evaluated and honed for RNA-seq data, particularly in regard to normalization and transformation. Although many normalization strategies have been developed for RNA-seq data, they have mostly been benchmarked only in the context of estimating differential gene expression [13–17]. Very little work has been done so far to comprehensively compare these strategies for normalization and network transformation (and their combinations) to

construct the most accurate coexpression networks from RNA-seq data, especially to ensure their robust application to datasets typically generated by individual research groups [1].

The most relevant prior work focuses on establishing best practices that reduce the introduction of artifacts in coexpression networks built from RNA-seq data [18]. This study includes a sequential comparison of a select number of methods for transcript assembly, normalization, and network reconstruction. However, the normalization comparison is based on 10 RNA-seq datasets, leaving considerable room for improvement. First is to increase the number and diversity of datasets studied. This is vital for finding robust procedures that work across datasets that can vary considerably in many respects, including sample size, sample variability, sequencing depth, tissue type, and other experimental factors. Further, testing on a wide range of datasets is critical both for the analysis of individual datasets as well as integrative analysis of hundreds/thousands of datasets. Second, not only do more normalization and network transformation methods need to be compared but how they might interact in combinations needs to be studied. Third, the resulting networks need to be evaluated directly on the accuracy of the coexpression between gene pairs, instead of performance in a downstream task such as gene function prediction, to ensure maximal utility of the network regardless of the subsequent biological application. Finally, the evaluation metric needs to be informative considering the fact that only a small fraction of all gene pairs in the genome are functionally related.

In this work, we present the most comprehensive benchmarking of commonly used within- and between-sample normalization strategies and network transformation methods for constructing accurate coexpression networks from human RNA-seq data. We tested every possible combination of methods from different normalization and network transformation stages. Our primary interest is in identifying robust combinations of methods that consistently result in coexpression networks that accurately capture general and tissue-aware gene relationships across a large variety of datasets. This will allow us to propose general recommendations useful for experimental research groups analyzing their own RNA-seq data as well as computational researchers seeking to build many coexpression networks from publicly available data for the purposes of data/network integration. Towards this aim, we use hundreds of datasets, generated by a consortium and by individual laboratories, covering multiple experimental factors. We then test the resulting networks on both tissue-naïve and tissue-aware prior knowledge about gene functional relationships. Based on these extensive analyses, we finally provide concrete recommendations for normalization and network transformation choices in RNA-seq coexpression analysis.

## **Results**

### **Expression data, gold standard, and benchmarking summary**

To test various within-sample normalization, between-sample normalization, and network transformation methods (and their combinations) on a large data collection, we started with gene count data from the recount2 database [19]. Recount2 contains data from both the Genotype-Tissue Expression (GTEx) project [20] and the Sequence Read

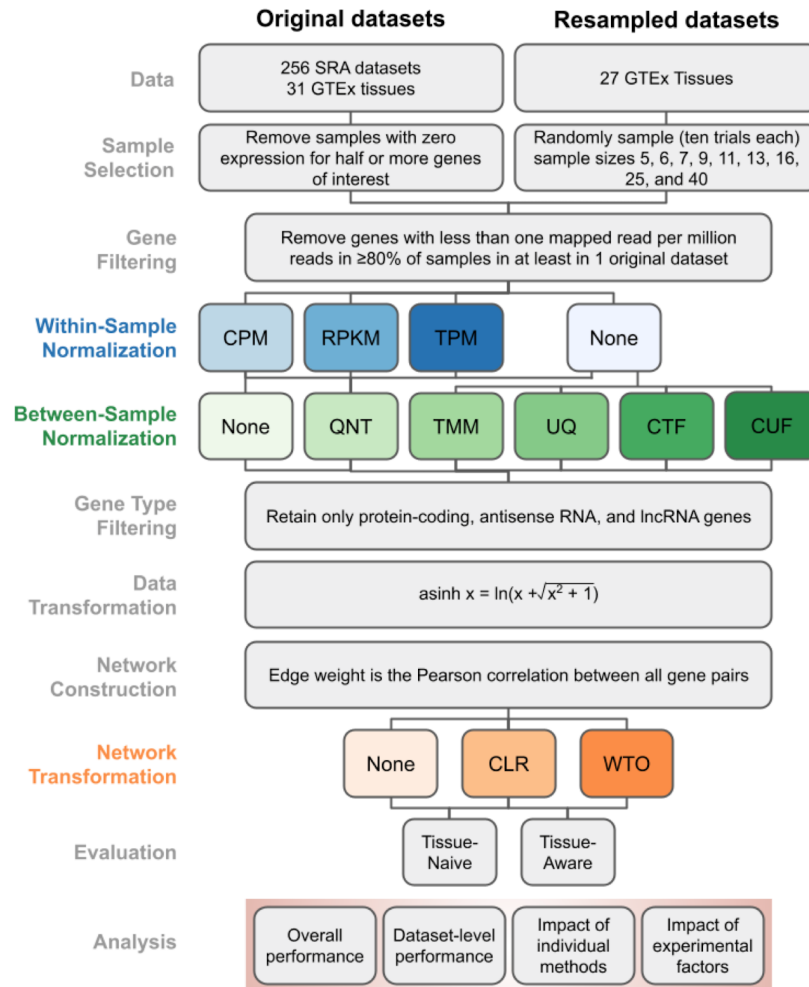
Archive (SRA) [21] repository that have been uniformly quality-controlled, aligned, and quantified to the number of reads per gene in the genome. Datasets from the GTEx project allowed us to assess method performance on large, relatively homogeneous datasets with high-sequencing depth and quality. The GTEx data was also critical for investigating the impact of experimental factors such as sample size, which we performed by doing multiple rounds of random sampling from GTEx datasets. Datasets from SRA, on the other hand, were representative of heterogeneous, mostly small experiments (median of 12 samples) that are generated by individual labs, with a range of sequencing depths and qualities. In total, we used 9,657 GTEx samples and 6,301 SRA samples from a total of 287 datasets (**Table 2.1**, Appendix: **Fig. A2.1**; see **Methods**), and processed and evaluated these two collections separately.

	<b>GTEx</b>	<b>SRA</b>
Number of samples	9,657 samples	6,301 samples
Number of datasets	31 datasets	256 datasets
Number of tissues	31 tissues	19 tissues
Median dataset size	197 samples	12 samples
Total	15,958 samples from 37 unique tissues	

**Table 2.1: Summary of data used in this study.** See *Figure A2.1* and *Methods* for more details.

After preprocessing each dataset using lenient filters in order to keep data for as many genes and samples as possible (see **Methods**), we compared methods commonly used in RNA-seq analysis to effectively construct one coexpression network per dataset (i.e. building 31 GTEx networks and 256 SRA networks). We focused on three key stages of data processing and network building: a) within-sample normalization: counts per million (CPM), transcripts per million (TPM), and reads per kilobase per million (RPKM), b)

between-sample normalization: quantile (QNT), trimmed mean of M-values (TMM), upper quartile (UQ); in addition, we tested two new variations of TMM and UQ – counts adjusted with TMM factors (CTF), counts adjusted with upper quartile factors (CUF) – that directly adjust counts by the size factors but does not correct by library size, and c) network transformation: weighted topological overlap (WTO) and context likelihood of relatedness (CLR). To systematically examine these methods and their interactions, we built 36 different workflows covering all possible combinations of choices (**Fig. 2.1**). For clarity, in the rest of the manuscript, we present individual methods in regular font (e.g. TPM normalization) and italicize workflows (e.g. *TPM*, which is TPM combined with no between-sample normalization and no network transformation, or *TPM\_CLR*, which is TPM paired with just CLR). The *Counts* workflow uses no within-sample normalization, between-sample normalization, or network transformation, but is still transformed with the hyperbolic arcsine function.



**Figure 2.1. Pipeline for benchmarking the optimal workflow for constructing coexpression networks from RNA-seq data.** The main pipeline was executed for the original GTEx and SRA datasets and a large collection of datasets of different sizes resampled from the GTEx datasets. Three key stages – within-sample normalization, between-sample normalization, and network transformation – where we tested method choices are highlighted in different colors. All the other stages were composed of standard selection, filtering, and data transformation operations. The coexpression networks resulting from all the workflows were evaluated using two gold-standards that capture generic (tissue-naive) and tissue-aware gene functional relationships. Finally, all the evaluation results were used to analyze the impact of various aspects of the workflows, methods, and datasets on the accuracy of coexpression networks. Abbreviations: CPM (Counts Per Million), RPKM (Reads Per Kilobase Million), TPM (Transcripts Per



**Figure 2.1. (cont'd)**

Million), QNT (quantile), TMM (Trimmed Mean of M-values), UQ (Upper Quartile), CTF (Counts adjusted with TMM Factors), CUF (Counts adjusted with Upper quartile Factors), CLR (Context Likelihood of Relatedness), WTO (Weighted Topological Overlap).

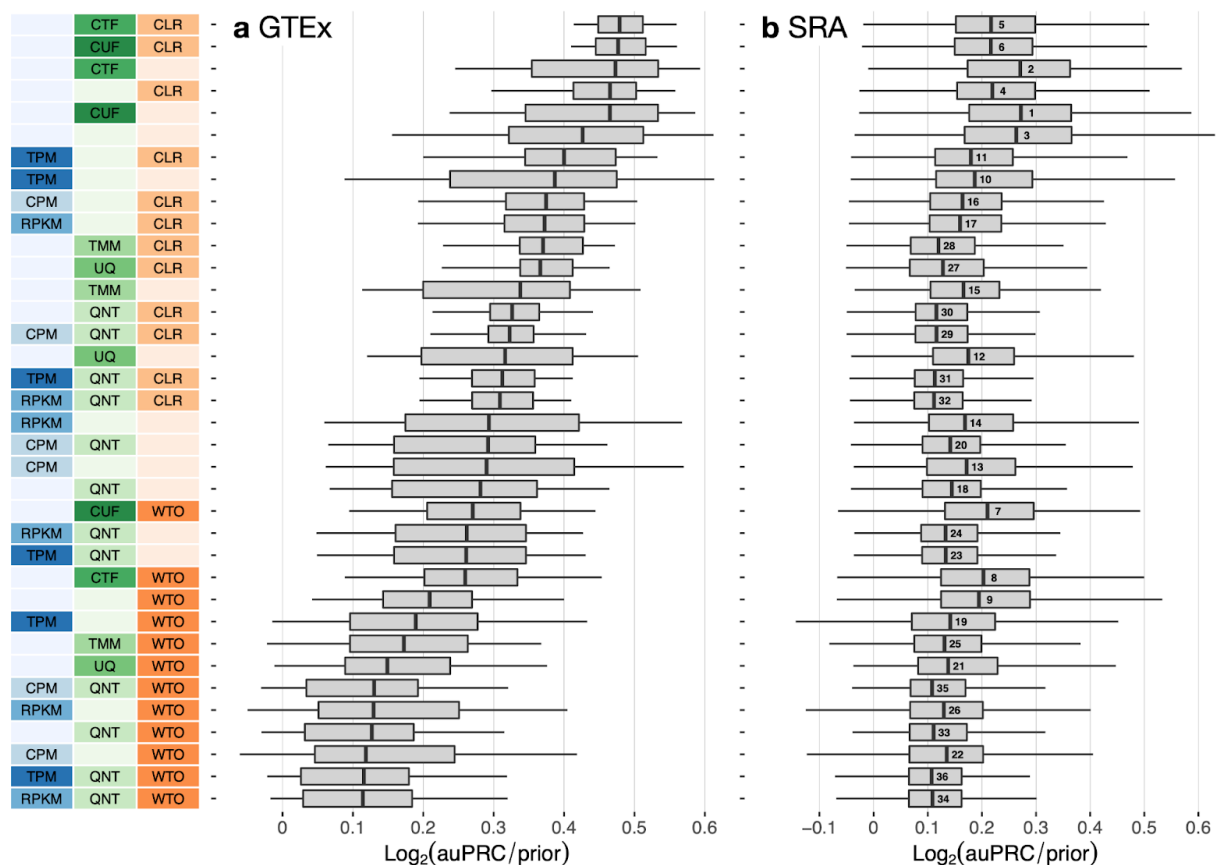
Since this entire workflow is unsupervised, i.e. not reliant on prior knowledge about gene relationships, we evaluated the resulting coexpression networks by comparing them to gold standards of known gene functional relationships. The gold standards were built using experimentally-verified co-annotations to specific biological process terms in the Gene Ontology [22]. These comparisons yielded evaluation metrics that summarize how well the patterns of coexpression captured in the network reflect known gene functional relationships (see *Network Evaluation* in **Methods** and *Supplemental Note* in Appendix). Further, gene activities and interactions vary drastically depending on cell type or tissue. Hence, we also created tissue-aware gold standards to assess whether the resulting networks were able to recapitulate tissue-aware coexpression in addition to general “tissue-naive” coexpression. Tissue-aware gold standards were created for as many tissues as possible by subsetting the naive gold standard using genes known to be expressed in a particular tissue. While area under the receiver operator curve (auROC) is frequently used to estimate network accuracy, it does not account for the fact that only a small fraction of gene pairs (out of the total possible) biologically interact. In the gold standard, this imbalance is reflected by the number of negatives (non-interactions) far outnumbering the positives (interactions) [23]. Therefore, we measured network accuracy using area under the precision recall curve (auPRC), which emphasizes the accuracy of top-ranked coexpression gene pairs [24].

In total, for each of the 287 datasets from GTEx and SRA, we built one coexpression network per dataset using each of the 36 workflows, resulting in 8,610 coexpression networks. Later on, we created 2,430 additional datasets generated by resampling GTEx that, when run through all the workflows, resulted in another 72,900 networks. Each GTEx network contains 20,418 genes while each SRA network contains 22,084 genes, and all networks are fully connected with edges weighted by their strength of correlation. Each of these networks were evaluated using the tissue-naive gold-standard and, whenever applicable, the tissue-aware gold-standard. Finally, we replicated the analysis of the top-performing workflows using as many matched SRA datasets as possible from another RNA-Seq repository, refine.bio [25], where read alignment and expression quantification were done using different methods than recount2.

### **Overall performance of workflows**

For all 36 workflows, **Figure 2.2** shows the overall performance of the networks resulting from GTEx (left) and SRA (right) recount2 datasets based on evaluation using the tissue-naive gold standard. **Figure A2.2** shows the performance of these networks based on the tissue-aware gold standards (when available). Overall, networks built from GTEx datasets are far more accurate than those built from SRA datasets (**Fig. 2.2, A2.2**). In each of the four cases – GTEx and SRA networks evaluated using tissue-naive and tissue-aware gold standards – most of the top-performing workflows contain CTF or CUF normalization. Further transforming the network with CLR (*CTF\_CLR* and *CUF\_CLR*) results in top-tier workflows for the GTEx datasets

regardless of gold standard. However, CLR transformation is only among top-performing methods for SRA datasets in recovering tissue-aware gene relationships. Though *CTF\_CLR* and *CUF\_CLR* still perform quite well on the tissue-naive standard for SRA, there is a clear gap from the top tier. Despite CTF- and CUF-containing workflows resulting in top performances, workflows that include other between-sample normalization methods are absent among the top ten workflows for both GTEx and SRA. Workflows with TMM or UQ seem to be more comparable to workflows using within-sample normalization methods.



**Figure 2.2. Overall performance of workflows.** The plots show the aggregate accuracy of all coexpression networks resulting from each individual workflow using (a) GTEx and (b) SRA datasets, evaluated based on the tissue-naive gold standard. The workflows (rows) are described in terms of the

### Figure 2.2. (cont'd)

specific method used in the within-sample normalization (blues), between-sample normalization (greens), and network transformation (oranges) stages. The performance of each workflow is presented as boxplots (without outliers) that summarizes the  $\log_2(\text{auPRC}/\text{prior})$  of each workflow where auPRC is the area under the precision recall curve (see **Methods**). The workflows are ordered by their median  $\log_2(\text{auPRC}/\text{prior})$  for the GTEx data. The numbers inside the SRA boxes indicate rank by median  $\log_2(\text{auPRC}/\text{prior})$  of the workflows for the SRA data. *Figure A2.2* contains these plots based on the tissue-aware gold standard.

The next noteworthy observation is that the top-tier workflows do not include a within-sample normalization step. Yet, workflows that do include within-sample normalization methods (CPM, RPKM, TPM) can perform better than many other workflows depending on other choices made in the pipeline, the best choice often is to be paired with no other method or CLR alone. For GTEx datasets, CLR seems to generally result in slightly improved performance, while the WTO transformation almost exclusively makes up the bottom tier of workflows. For building networks from SRA datasets, although workflows including WTO do not exclusively end up in the bottom tier (as is the case with GTEx data), adding WTO to a particular workflow always hurts performance. The worst workflows for SRA in either standard are quantile normalization (QNT) paired with CLR or WTO.

### Dataset-level performance of workflows

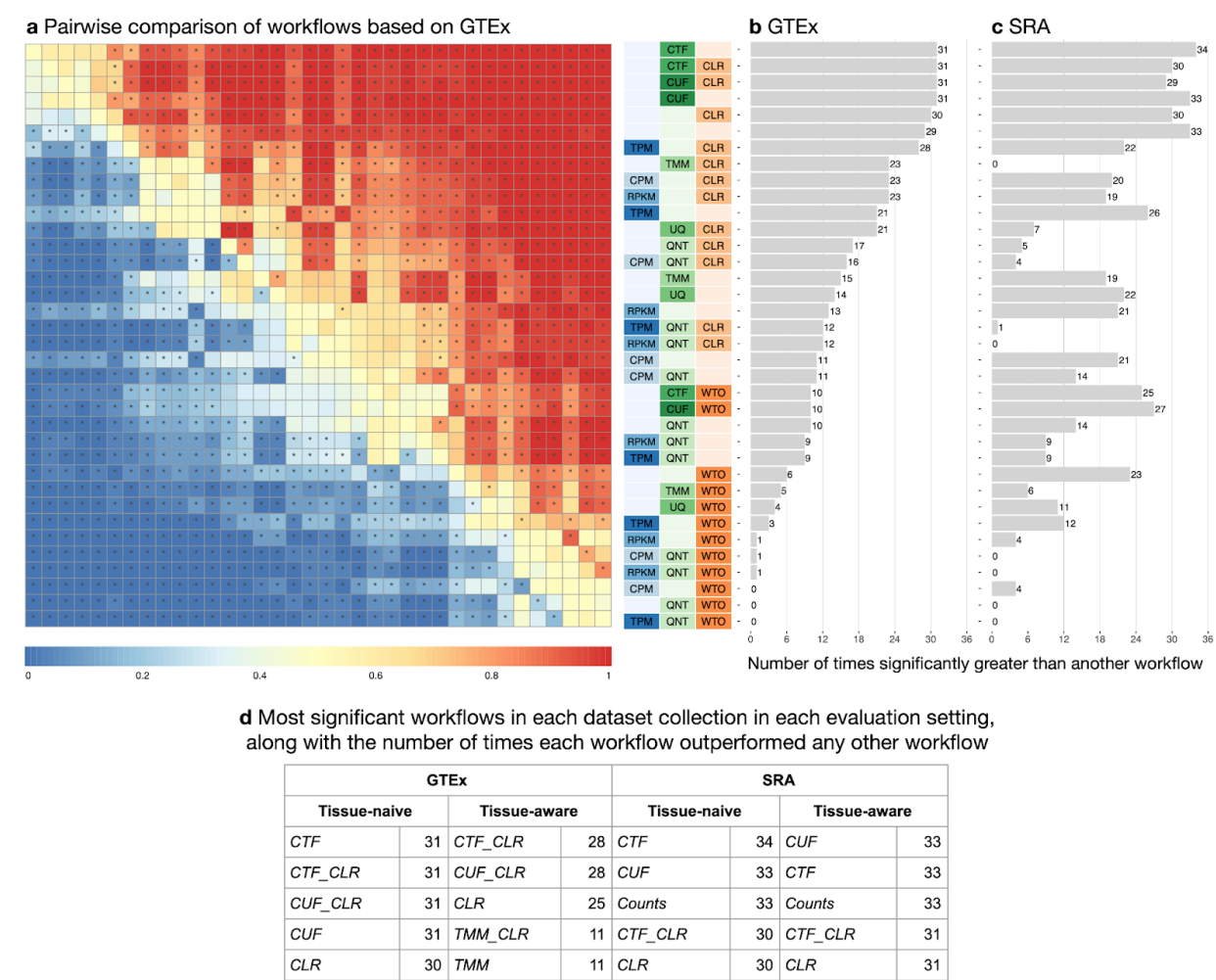
Next, we dissected the aggregated results described above for GTEx and SRA as a whole by examining the accuracy of these workflows on a per-dataset basis. First, we compared pairs of workflows to each other and determined the proportion of datasets in

which one workflow outperformed the other across all GTEx and all SRA datasets (**Fig. 2.3, A2.3–2.5**, heatmap colors). Second, we performed paired statistical tests to estimate the significance of the difference between the workflows (**Fig. 2.3, A2.3–5**, asterisks on the heatmap). Finally, we scored each workflow based on the number of other workflows it significantly outperforms (**Fig. 2.3, A2.4** barplots). Based on this analysis, in the ‘GTEx-naive’ setting (i.e. networks from GTEx data evaluated on the tissue-naive gold standard), we observed that five workflows are all significantly more accurate than 31 other workflows but not significantly different from one another (paired Wilcoxon rank-sum test; corrected p-value < 0.01; **Fig. 2.3**). Within these four workflows, *CTF* outperforms *CTF\_CLR*, *CUF*, and *CUF\_CLR* on 58%, 61%, and 58% of GTEx networks, respectively. The *CTF* workflow is also significantly better the most number of times compared to other workflows in the SRA networks using the naive standard, although *Counts* and *CUF* are only slightly behind *CTF* (**Fig. 2.3, A2.3**). These workflows tie for first place when SRA networks are evaluated on the tissue-aware gold standards (Appendix: **Fig. A2.4, A2.5**).

When the GTEx networks are evaluated on tissue-aware standards, there are much fewer significant differences between workflows overall, with the exception of *CTF\_CLR*, *CUF\_CLR*, and *CLR* being significantly greater than 28, 28, and 24 workflows, respectively (Appendix: **Fig. A2.4**). Here, *CTF\_CLR* performs better than *CUF\_CLR* on 57% of networks and better than *CLR* on 76% of networks. Despite having similar median  $\log_2(\text{auPRC}/\text{prior})$  values to *CTF\_CLR* and *CUF\_CLR* (Appendix: **Fig. A2.2**), the *CUF* and *CTF* workflows only perform significantly better than another workflow a

handful of times (Appendix: **Fig. A2.4**). This suggests that including CLR in the workflow is especially helpful in capturing tissue-aware coexpression in the GTEx networks.

Again, the impact of within-sample normalization varies depending on the choice of the other methods in the workflow. *TPM\_CLR* is generally the top-performing workflow among those including within-sample normalization across evaluation cases, though *TPM* slightly outperforms *TPM\_CLR* for the SRA networks evaluated on the naive standard (**Fig. 2.3** and **A2.3**).



**Figure 2.3. Dataset-level pairwise comparison of workflow performance.** (a) The heatmap shows the relative performance of a pair of workflows, corresponding to a row and a column, directly compared to

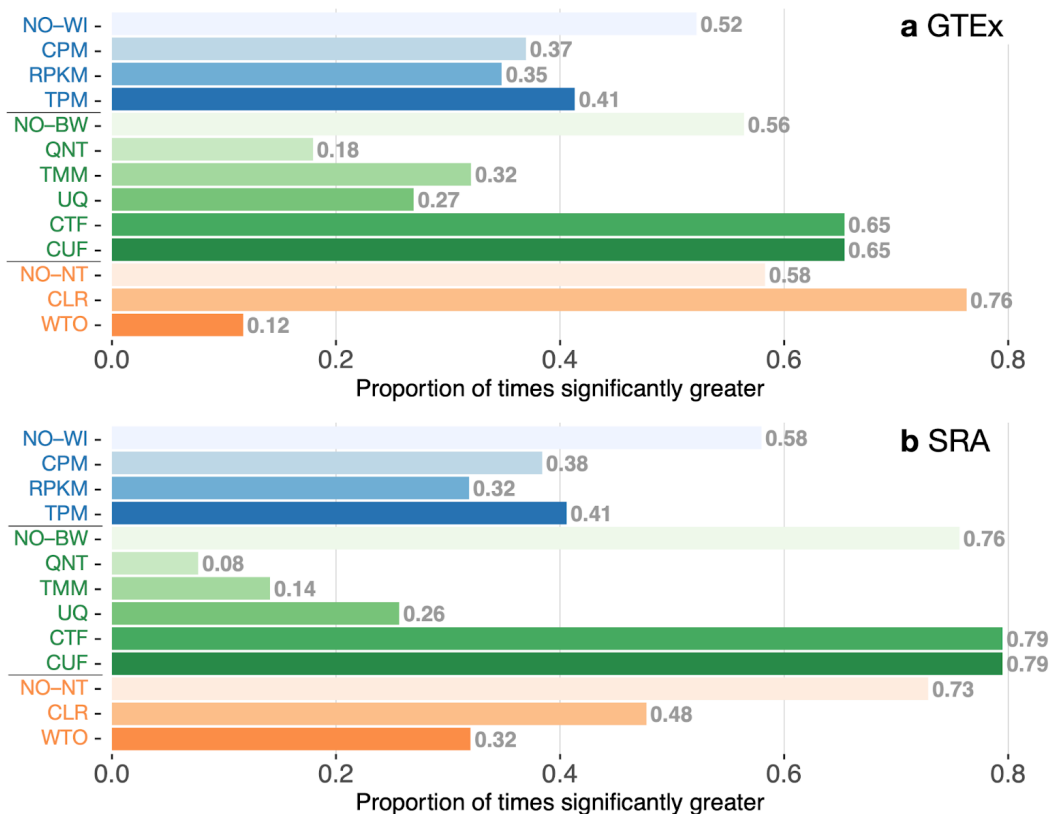
### Figure 2.3. (cont'd)

each other for the GTEx datasets based on the tissue-naive gold standard. The workflows along the rows are depicted using color swatches similar to *Figure 2.2*. The color in each cell (row, column) represents the proportion of datasets for which the workflow along the row has a higher  $\log_2(\text{auPRC}/\text{prior})$  than the workflow along the column. Comparisons that are statistically significant (corrected  $p < 0.01$ ) based on a paired Wilcoxon test are marked with an asterisk. *Figure A2.3* contains the corresponding heatmap for the SRA datasets. **(b and c)** Barplots show the number of times each workflow was significantly greater than another workflow for GTEx and SRA datasets. *Figures A2.4 and A2.5* contain these performance plots based on the tissue-aware gold standard. **(d)** The table shows the most significant workflows across evaluation cases along with the number of times a given workflow outperformed any other workflow for the GTEx and SRA datasets based on the tissue-naive and tissue-aware gold standards.

The impact of network transformation is similar between GTEx and SRA data, but there is disagreement in the very top method. With GTEx, workflows that include CLR tend to be significant the most number of times, while WTO-containing workflows tend to be the least. Not a single workflow with WTO significantly outperformed any workflow without it for GTEx based on the tissue-aware gold standard (Appendix: **Fig. A2.4**). On the other hand, CLR workflows perform well on the SRA networks, but do not constitute the workflows that were significantly greater than another the absolute most number of times (Appendix: **Fig. A2.3 and A2.5**). WTO hurts performance in every case even here. Pairing either CLR or WTO with quantile normalization (QNT) yields particularly poor performance in the SRA networks. All together, these results suggest that CTF yields the most accurate coexpression network by a very close margin and CLR can further improve the network in select cases.

## Impact of individual methods on performance of workflows

Though the previous analyses shed light on the contributions of individual methods, we wanted to more explicitly assess how choosing or not choosing a particular within-sample normalization, between-sample normalization, or network transformation affects general performance of any given workflow. To this end, for each method, we calculated the proportion of times that workflows that include a particular method performed significantly better than workflows that did not include the method (**Fig. 2.4**; see **Methods** for details).



**Figure 2.4. Impact of individual methods on performance of workflows.** Each bar in the two barplots, corresponding to a specific method, shows the proportion of times (x-axis) that workflows including that particular method (y-axis) were significantly better than other workflows. The barplots correspond to



#### **Figure 2.4. (cont'd)**

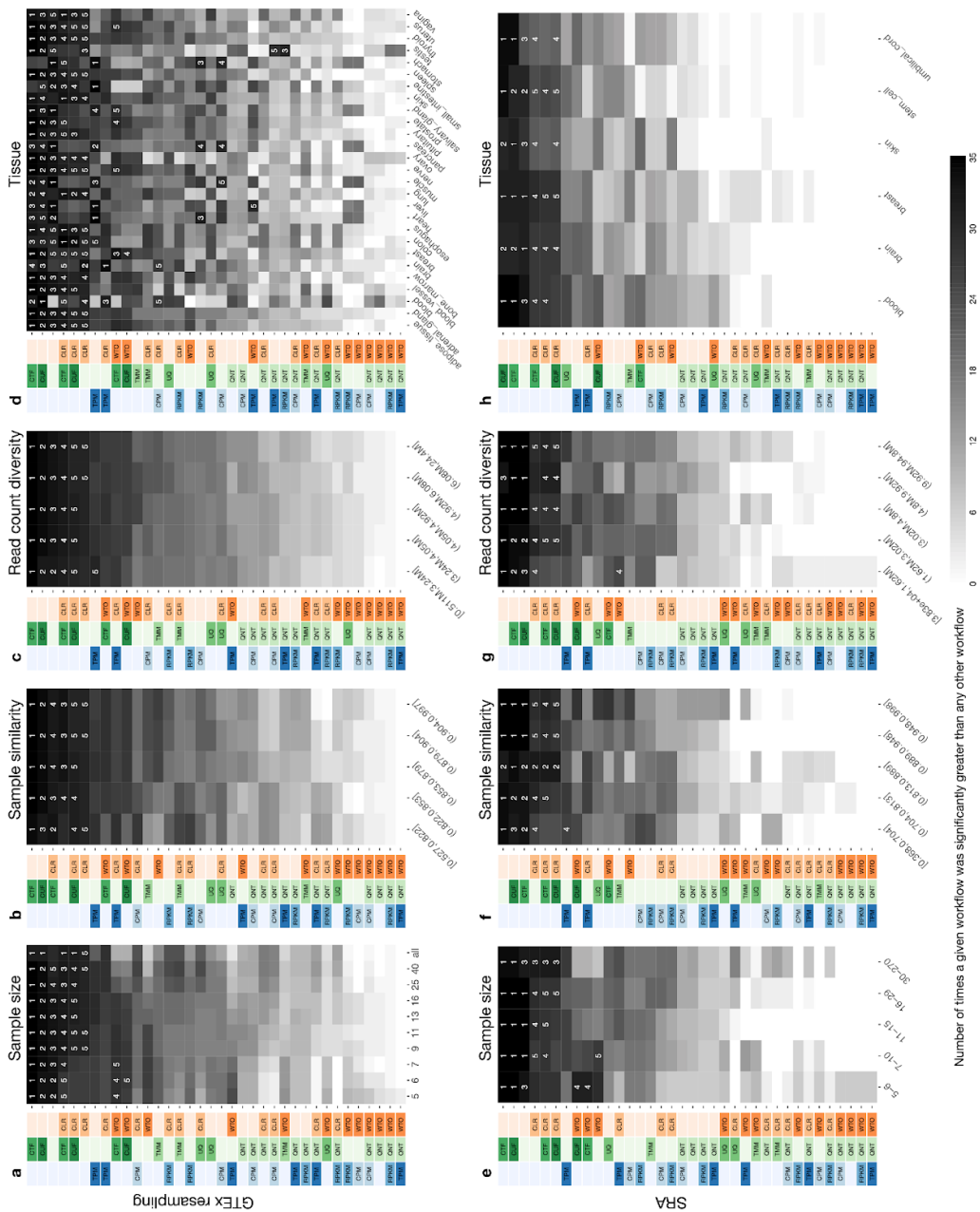
performance for the (a) GTEx and (b) SRA datasets evaluated on the tissue-naive gold standard. In order to make the comparison of between-sample normalization methods fair, workflows also including CPM, RPKM, or TPM were left out because it is not possible to pair them with CTF/TMM/CUF/UQ normalization. Similarly, CTF/TMM/CUF/UQ methods are not included for “no within-sample normalization” (NO-WI). *Figure A2.6* contains these barplots based on the tissue-aware gold standard.

This analysis clearly shows that, in all four cases (GTEx and SRA, each with tissue-naive and tissue-aware standards), utilizing any within-sample normalization method results in worse overall performance than not using it (**Fig. 2.4** and **A2.6**). Among within-sample normalization methods, TPM usually performs slightly better than CPM and RPKM. CTF and CUF are the best between-sample normalization methods. Their performances are exactly equal for GTEx data evaluated on either standard and for SRA data evaluated on the naive standard; CTF is slightly better than CUF for SRA data in the tissue-aware standards. However, doing no between-sample normalization performs quite well too, only narrowly worse than CTF or CUF. It is clear in all four cases that TMM, UQ, and quantile normalization (QNT) are vastly outperformed. Network transformation is the group most obviously different between GTEx and SRA data, with CLR being the clear winner for GTEx, while not doing any network transformation is significant many more times for SRA regardless of gold standard (**Fig. 2.4** and **A2.6**).

#### **Impact of varying experimental factors on performance of workflows**

The reason we included SRA data in this study is that SRA datasets are very representative of expression datasets typically generated by numerous individual

laboratories. Accordingly, these datasets vary considerably in terms of multiple factors including sample size, sample similarity, number of mapped reads, and tissue type. Though these factors impact the quality of coexpression networks derived from the individual datasets, it is hard to tease out the effect of each of these factors (controlling for others) on the accuracies that we observed using different workflows on SRA data. Therefore, using the large GTEx datasets, we created a collection of SRA-like datasets to more closely examine the impact of each experimental factor. First, we determined the nine sample sizes (5, 6, 7, 9, 11, 13, 16, 25, and 40) that are representative of SRA datasets. Then, from each GTEx tissue dataset with at least 70 samples, we randomly selected samples to create ten datasets for each sample size (see **Methods**). We then applied all 36 workflows to construct coexpression networks from each one of these datasets. The resulting 72,900 networks were used to investigate the effects of varying each experimental factor by counting the number of times a given workflow significantly outperformed any other workflow (**Fig. 2.5**). In addition to this analysis with these resampled data, we also examined the effect of sample similarity and number of mapped reads (see Experimental factor analysis in **Methods**) directly in the SRA data by splitting the datasets into five equal size bins based on each of these factors and determining the number of times a given workflow was significantly better than another within each bin (Appendix: **Fig. A2.7**).



**Figure 2.5. Impact of various dataset-related experimental factors on performance of workflows.**

Each heatmap shows the number of times (cell color) each workflow (row) outperforms other workflows as a particular experimental factor pertaining to the input datasets is varied (columns), when the resulting coexpression networks are evaluated based on the tissue-naïve gold standard. The darkest colors indicate workflows that are significantly better than the most other workflows. In addition, the top 5 workflows in each column are marked with their rank, with ties given minimum rank. The heatmaps on the

### Figure 2.5. (cont'd)

top (**a–d**) correspond to datasets from GTEx resampling and those on the bottom (**e–h**) correspond to SRA datasets. The heatmaps from left to right show workflow performance by sample size (**a, e**; number of samples used to make the coexpression network), sample similarity (**b, f**; median spearman correlation of 50% most variable genes between samples), read count diversity by counts (**c, g**; standard deviation of counts sums across samples), and tissue of origin (**d, h**). *Figure A2.7* contains these heatmaps based on the tissue-aware gold standard.

In the GTEx-resampled data, *CTF* was significantly better than all other workflows for sample sizes 5 through 40 when using the naive standard for assessment (**Fig. 2.5**). *CUF* is a close second, performing significantly better than all workflows other than *CTF* at sample sizes 7 through 40. Using only *Counts* (no normalization) is surprisingly effective, especially at lower sample sizes, while *CTF\_CLR* and *CUF\_CLR* improve performance with increasing sample size. In fact, when all samples from a given GTEx tissue are used ( $\geq 70$  samples), there is no significant difference between *CTF*, *CUF*, *CTF\_CLR*, and *CUF\_CLR*. *CLR* is the next best workflow after those top four. The only other workflows that are ever ranked in the top five are *CTF\_WTO* and *CUF\_WTO*, and that too only at low sample sizes (5–7). Based on the tissue-aware standards, *CTF\_CLR* is the most effective workflow on all sample sizes except 5, where *CTF* and *CUF* are the top workflows (Appendix: **Fig. A2.7**). For the highest two sample sizes (25 and 40), *CTF\_CLR* is substantially better than all other workflows. The only workflows ranked in the top five in sample sizes 5 through 40 are *CTF\_CLR*, *CUF\_CLR*, *CLR*, *CUF*, *CTF*, and *TPM\_CLR*. *CTF* and *CUF* also perform well on the SRA data evaluated on the naive standard, being the top workflows in all five sample size groups (**Fig. 2.5**).

Performance on the tissue-aware standards is slightly more variable, with *Counts*, *CTF*, and *CUF* being top ranked in lower sample size groups and *CLR*, *CUF\_CLR*, and *CTF\_CLR* performing better in the highest sample size group (Appendix: **Fig. A2.7**). Again, it is clear that *CTF* and *CUF* are superior methods, with *CLR* improving performance in select cases.

Sample similarity and read count diversity analyses show similar results to those from sample size analysis. When evaluating the GTEx-resampled data on the naive standard, *CTF* is almost always significantly better than every other workflow across all groups, while evaluating on the tissue-aware gold standards ranks *CTF\_CLR* as the top workflow most consistently (**Fig. 2.5**, Appendix: **Fig. A2.7**). In both standards, *CTF*, *CUF*, *CLR*, *CTF\_CLR*, *CUF\_CLR* and *Counts* are the workflows consistently showing up in the top five ranks. The SRA networks evaluated on either standard have *CTF*, *CUF*, and *Counts* showing up in the top three ranks across most groups, with *CLR*, *CTF\_CLR*, and *CUF\_CLR* making up most of the other workflows in the top five ranks (**Fig. 2.5**, Appendix: **Fig. A2.7**).

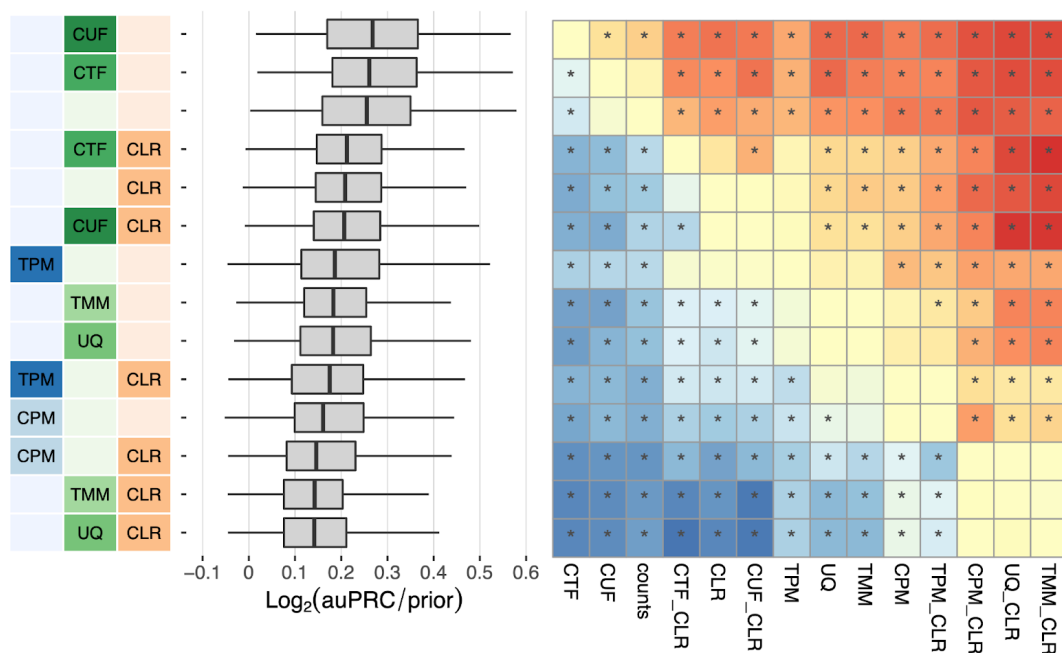
Tissue is the factor that shows the most variability in terms of what makes up the top workflows, especially when evaluating on tissue-aware gold standards. This is due in part to the fact that splitting experiments by tissue results in the smallest groups, making significance more difficult to detect. Nevertheless, the top workflows from the analyses of other factors still have the best overall performance across all tissues. In the GTEx-resampled data, *CTF* is the top-ranked workflow most often based on the naive gold standard. *CUF* and *Counts* are almost always in the top five most significant

workflows, while *CTF\_CLR*, *CUF\_CLR*, and *CLR* show up often. When evaluated on tissue-aware gold standards, *CTF*, *CLR*, and *CTF\_CLR* are ranked number one more frequently than any other workflow, but they are not as consistent as *CTF* in the naive standard. *CUF* and *CUF\_CLR* are the other top-performing workflows, but a handful of other workflows enter the top five ranks in at least a few tissues. For SRA, only tissues that had more than fifteen separate experiments were used in the significance analysis (Appendix: **Fig. A2.1**). On the naive standard, *CUF*, *CTF*, or *Counts* were always the most significant workflow in any given tissue and *CLR*, *CTF\_CLR*, and *CUF\_CLR* were usually in the top five. A similar if less consistent pattern can be observed from the tissue-aware evaluations. Taken together, these results suggest that the top-performing methods are largely robust to common experimental factors that vary from experiment to experiment. This property is critical because, to be practically beneficial, the best workflow for constructing coexpression networks should result in accurate coexpression networks irrespective of variations in these experimental factors.

### **Impact of varying alignment and counts quantification performance of workflows**

So far, our analysis has considered datasets from the recount2 database. This has allowed us to evaluate the performance of each workflow on a large, diverse set of datasets which have been uniformly aligned and transformed into gene counts. However, this begs the question of whether the observed results – especially the top performance of *CTF*, *CUF*, and *Counts* – would hold when different methods for read alignment and counts quantitation are used. To determine whether this is the case, we matched as many of our recount2 SRA datasets as possible to those from refine.bio

[25], another RNA-seq repository that uses completely different methods for alignment and quantification. This turned out to be 186 datasets in the naive evaluation and 163 of those could be evaluated with a tissue-aware standard. Unfortunately, GTEx data is not available from refine.bio. In this new analysis, we left out the worst performing methods in each tested category, i.e., RPKM, QNT, and WTO for within-sample normalization, between-sample normalization, and network transformation, respectively. This leaves us with 14 workflows to evaluate on the refine.bio datasets.



**Figure 2.6. Overall performance of workflows and pairwise-comparison using refine.bio datasets.**

The boxplots show the aggregate accuracy of all coexpression networks resulting from each individual workflow using SRA datasets in refine.bio, evaluated based on the tissue-naive gold standard. The performance of each workflow is presented as boxplots (without outliers) that summarizes the log<sub>2</sub>(auPRC/prior) of each workflow, where auPRC is the area under the precision recall curve (see **Methods**). The workflows are ordered by their median log<sub>2</sub>(auPRC/prior). The heatmap shows the relative performance of pairs of workflows (rows and columns) compared to each other for the refine.bio

### Figure 2.6. (cont'd)

SRA datasets based on the tissue-naive gold standard. The color in each cell (row, column) represents the proportion of datasets for which the workflow along the row has a higher  $\log_2(\text{auPRC}/\text{prior})$  than the workflow along the column. Comparisons that are statistically significant (corrected  $p < 0.01$ ) based on a paired Wilcoxon test are marked with an asterisk. *Figure A2.8* contains these plots based on the tissue-aware gold standard.

In the naive evaluation, *CTF*, *CUF*, and *Counts* are once again the top-tier workflows. However, the *CUF* workflow significantly outperforms the other two across all datasets (**Fig. 2.6**). The second tier consists of *CTF\_CLR*, *CUF\_CLR*, and *CLR*, though it is not quite as well separated from the remaining workflows. The tissue-aware evaluation shows much less separation between *CUF*, *CTF*, *Counts*, *CTF\_CLR*, *CUF\_CLR*, and *CLR* in terms of overall performance measured by  $\log_2(\text{auPRC}/\text{prior})$ , but *CTF* and *CUF* significantly outperform more workflows than any other (Appendix: **Fig. A2.8**). In summary, we replicated the ranking of coexpression workflows using RNA-seq data processed with an entirely different pipeline for alignment and quantification.

The general trends presented above are all based on network accuracy measured using a metric based on the area under the precision-recall curve ( $\log_2(\text{auPRC}/\text{prior})$ ). These trends also hold when network accuracy is measured using precision at low recall, which focuses on maximizing the number of functional gene pairs among the high-scoring gene pairs instead of focusing on recovering all functional gene pairs. Put another way, the trends described above hold even when a threshold is applied to the coexpression network to retain just the high-scoring gene pairs for subsequent analysis. For the sake of completion, we have also evaluated all networks using the area under



the ROC curve (auROC). All these results based on three different evaluation metrics ( $\log_2(\text{auPRC}/\text{prior})$ , precision at 20% recall, and auROC) are available as a consolidated webpage at [https://krishnanlab.github.io/RNAseq\\_coexpression](https://krishnanlab.github.io/RNAseq_coexpression) that researchers can explore to easily examine the performance of various workflows based on the properties of their RNA-seq dataset.

## Discussion

Despite the utility and growing popularity of coexpression analysis of RNA-seq data, relatively little focus has been devoted to identifying the optimal data normalization and network transformation methods that result in accurate RNA-seq-based coexpression networks. Here, we present the most comprehensive analysis of the effects of commonly-used techniques for RNA-seq data normalizations and network transformation on gene coexpression network accuracy (**Fig. 2.1**). We implemented 36 network-building workflows – one for every combination of within-sample normalization, between-sample normalization, and network transformation methods – and we ran each workflow on hundreds of RNA-seq datasets from GTEx and SRA. The resulting coexpression networks were evaluated using both known tissue-naïve and tissue-aware gene functional relationships to ensure that the networks were tested for capturing not just generic gene interactions but also interactions relevant to the tissue under consideration (Appendix: **Fig. A2.9, Fig. A2.10**). The evaluations shed light on several key aspects of the impact of within-sample normalization, between-sample normalization, and network transformation methods (and their interplay) on the accuracy of the resulting coexpression networks.

## **Impact of within-sample normalization**

Within-sample normalization – commonly executed by converting gene counts to CPM, RPKM, or TPM – corrects for factors such as library size and gene length. As gene length biases both gene counts and their downstream analysis [26], it is not very surprising that TPM usually outperforms CPM, as CPM only corrects for library size and not gene length. However, the order in which gene-length and library-size correction are combined appears to be important. For example, studies have shown that RPKM, which first corrects for library size and then for gene length, is inferior to other methods in differential expression analysis and is not recommended [13–15]. Some studies have also noted that using RPKM does not necessarily take away the length bias in gene expression and can be unduly influenced by relatively few transcripts [13,27]. TPM was proposed as an improvement over RPKM by first correcting for length and then by library size. Thus, the resulting expression values more accurately reflect the “relative molar concentration” of an RNA transcript in the sample [28]. TPM normalization scales every sample to the same total RNA abundance (i.e. the same total sum of TPM values). Thus, gene expression across samples becomes more comparable when TPM normalized than when RPKM normalized. Consistent with these previous studies, we find that RPKM generally results in lower-performing coexpression networks and that TPM consistently outperforms CPM and RPKM, and can even occasionally perform better than the general top-performers CTF and CUF. Finally, since a number of technical and biological factors affect the size and makeup of the sample library, TPM has been found to be most effective when comparing samples from the same tissue

type and experiment [29]. This observation could explain the good performance of TPM in our work wherein only samples within a dataset are compared and analyzed together to construct a coexpression network.

### **Impact of between-sample normalization**

Next, our results reinforce the expectation that between-sample normalization (using techniques such as CTF and CUF) leads to the largest improvement in coexpression accuracy. These methods are designed to make expression values across samples more comparable to one another, an aspect critical for coexpression analysis. However, QNT, a between-sample normalization method that is most commonly used with microarray data, performs very poorly for RNA-seq data. This is likely because QNT forces the distribution of samples to be exactly the same, meaning that each gene value is forced to be a particular quantile value. Consequently, it does not suit situations where there truly are different numbers of genes that are expressed outside of the typical ranges across samples [8,30], an effect that is further exacerbated in RNA-seq data because it has a larger dynamic range than microarray data. Genes with extreme values would not influence CTF or CUF normalization because they are explicitly excluded from the calculation of adjustment factors. CTF specifically finds a subset of genes that are probably not differentially expressed between samples to make gene values comparable across the entire group, while CUF uses only the upper quartile gene values to adjust samples. This makes both normalizations robust to a number of highly or lowly expressed genes. However, large-scale changes in gene expression or high amounts of asymmetry, e.g. a large difference in the number of genes expressed

above the typical range versus expressed below the typical range, violate these assumptions [8]. In our test cases, CTF and CUF performed the best, but it is possible that violation of their base assumptions may occur in specific disease conditions or external perturbations, leading to a significant decrease in their performances. The relatively lower performance of TMM and UQ, which are essentially CTF and CUF with library size correction, imply that library size correction is not the most helpful normalization strategy for building coexpression networks based on linear correlation measures. As noted below, measures such as Pearson correlation automatically include a standardization of gene expression across samples, which could explain why additional library size correction may not be needed. This implication is also supported by the *Counts* workflow outperforming within-sample normalization workflows.

### **Impact of network transformation**

Network transformation is where there is most disagreement between GTEx and SRA data. CLR was the best network transformation method for GTEx data, while doing no transformation of the coexpression values gave the best results for SRA data. The most pronounced factor that explains this difference is sample size. The median sample size of SRA datasets is 12, while that of GTEx datasets is 197. Only four GTEx datasets have less than 70 samples (Appendix: **Fig. A2.1**). Furthermore, GTEx resampling analysis showed that *CTF\_CLR* and *CUF\_CLR* improve with increasing sample size on the naive standard (**Fig. 2.5**) and to a lesser extent on the tissue-aware standards (Appendix: **Fig. A2.7**) since CLR tended to already have better performance in general on tissue-aware standards than on the naive gold standard. For each gene pair, CLR

adjusts the edge weight based on its value in relation to the distribution of edge weights for the individual genes in that pair to all other genes in the network. So, our hypothesis is that having a larger sample size results in a better estimate of each edge weight as well as the distribution of edge weights for each gene, which in turn increases CLR's accuracy. Supporting this hypothesis, other studies have noted an association between larger sample size and more accurate coexpression networks [18,27], and subsequent network transformation with CLR [31]. WTO, on the other hand, performs poorly for both GTEx and SRA data. WTO adjusts the edge weight between gene pairs based on whether they share strong connections to the same set of genes in the network. Therefore, while CLR relies on summary statistics (mean and standard deviation) of edge distributions to adjust the edge weight between each gene pair, WTO relies on the actual, likely noisy, coexpression values, which may contribute to its inferior performance. It is also possible that CLR's strategy more effectively deals with the mean-correlation relationship bias, or the observation that highly expressed genes tend to be more highly coexpressed, by capturing them as summary statistics, without relying on the fact that each of the correlation estimates are correct [32,33]. This may, in turn, explain why CLR tends to perform better on tissue-aware gold standards than on our naive gold standards, since genes that are ubiquitously expressed (and therefore involved in general, tissue-naive interactions) tend to be more highly expressed [34].

### **Impact of data transformation**

RNA-seq data analyses typically benefit from a data transformation that stabilizes the variance across mean values, i.e. renders the data more homoskedastic, because, in its

untransformed form, the expected variance grows with the mean for gene counts [35]. A standard procedure when working with RNA-seq (or even microarray) data for either differential expression analysis or coexpression analysis is to log transform gene counts. Since gene counts for several genes can be zero, the typical manner in which log transformation is applied to RNA-seq data is to add a pseudocount (of 1, for example) to every gene's count (say, 'x') before taking the log (i.e.  $\log(x + 1)$ ). However, adding a constant pseudocount to all genes is disadvantageous because low gene counts are disproportionately increased compared to high gene counts before log transformation (e.g.,  $1 + 1$  is a 100% increase for a gene count of 1, but  $941 + 1$  is almost a negligible increase). The hyperbolic arcsine (asinh) transformation –  $\log(x + \sqrt{x^2 + 1})$  – mitigates this effect [36]. The asinh function is defined along the entire real number line and circumvents the need for predefining a constant pseudocount and instead calculates a pseudocount for each gene that is proportional to that gene's original count. Therefore, it has a compression effect like the natural log function but much less so for small values of x [37]. Due to this advantage, each of our workflows uses the asinh transformation. However, since asinh has not been explicitly tested before (to the best of our knowledge), we analyzed the impact of this transformation on the coexpression network accuracy. We find that the asinh transformation yields an improvement in performance over no data transformation for our top ten workflows in GTEx and SRA datasets (Appendix: **Fig. A2.11**). It is worth noting that the *Counts* workflow performs well despite not incorporating any within- or between-sample normalization but only an asinh transformation. We speculate that this good

performance is due to the variance stabilization provided by the asinh transformation along with the across-sample normalization of gene expression vectors inherent within the calculation of the Pearson correlation coefficient.

The popular R package for differential expression analysis, DESeq2 [35], offers two other data transformations for gene counts: variance stabilizing transformation (VST) [38] and regularized logarithm transformation (rlog) [35]. Both transformations are similar to the log transformation of adjusted counts along with a pseudocount parameter that is chosen in a data-driven manner. These transformations consider between-sample effects like library size and are designed to only be used on counts data as part of calculating differential gene expression. Nevertheless, these transformations could in theory be applied to coexpression analysis. Hence, we compared asinh, VST, and rlog along with their combinations with network transformation methods and found that asinh is the best transformation for coexpression analysis in our all evaluations (Appendix: **Fig. A2.12–2.15**). The VST and rlog may perform better when supplied with sample group information. Therefore, we do not recommend the use of either transformation in DESeq2 for large-scale application to publicly-available RNA-seq datasets for coexpression analysis.

### **Recommendations for building coexpression networks from RNA-seq data**

By constructing coexpression networks for diverse datasets from both GTEx and SRA, we were able to evaluate workflows on large, homogeneous datasets as well as smaller, heterogeneous datasets to identify methods that are robust to differing technical and biological factors. Although there is some variation in performance between GTEx and

SRA data, and slightly more variation introduced by tissue-aware gold standards, many trends are consistent across datasets and evaluations. Based on all our results, we make the following recommendations for building coexpression networks from RNA-seq data using Pearson or Spearman correlation:

- If gene counts are available, use CTF or CUF to normalize the data. They consistently give the best performance regardless of various factors. Between the two, CTF seems to be slightly more consistent in yielding top performance. Even though no normalization (*Counts*) leads to good performance in our study, applying the additional normalization step is prudent to ensure robustness against variabilities specific to a new dataset.
- If data is only available after within-sample normalization, use TPM for coexpression analysis. Data in CPM and RPKM units can be easily converted to TPM. TPM outperforms CPM and RPKM and yields consistently reasonable performance.
- After normalization, perform log transformation (using `asinh`) and calculate coexpression using Pearson correlation coefficient.
- If the dataset has greater than 40 samples, use CLR to transform the pairwise gene correlations. CLR may also help certain cases where the main interest is interactions that are specific to a given tissue.
- QNT and WTO hurt performance in combination with every other method, in all cases, and should not be used.

To enable researchers to explore all our analyses in a streamlined manner and find the results most relevant to their own RNA-seq datasets of interest, we have made them



available in a rich webpage written with R Markdown. The webpage can be found at [https://krishnanlab.github.io/RNAseq\\_coexpression](https://krishnanlab.github.io/RNAseq_coexpression).

### **Potential future applications and extensions**

Going forward, we can leverage this comprehensive benchmarking framework for coexpression analysis to answer newer and subtler questions about data quality and sample composition. For example, many studies have found that removing unwanted variation, i.e. noise caused by technical rather than biological factors, in the RNA-seq data has led to improvements in downstream analysis including the calculation of coexpression networks [39,40]. Such corrections are often done using SVD-based methods, including removing the first (or the first few) principal components. However, caution must be taken when using these methods as they may easily remove biological signals from the data [41], especially in typical small-to-medium-sized datasets produced by most research labs (e.g. represented in SRA). Future work using our framework could help learn the guidelines for deciding which and how many factors to remove while carefully considering the various properties of the data and the biological objective of the analysis. For instance, one could explore if different tissues might be sensitive to different technical factors; signal from blood is often heavily influenced by the large variation in cell type composition but the brain is much more greatly affected by the post-mortem-interval [42]. Another related and open question is how cell type composition influences gene coexpression calculated from bulk tissue data. Some studies have concluded that gene coexpression networks are heavily confounded by this factor [43,44], while others have shown that coexpression derived from single-cell

data is very similar to bulk coexpression [45,46]. Finally, a similar framework could also be used to explore the best procedure for building coexpression networks from single-cell RNA-seq data, which has an entirely different set of challenges [47] that call for an entirely separate benchmarking effort.

## **Conclusions**

We have performed an extensive benchmarking and analysis of how data normalization and network transformation impact the accuracy of coexpression networks built from RNA-seq datasets. Based on this work, we have arrived at concrete recommendations on robust procedures that will generally lead to best coexpression networks. Specifically, using Counts adjustment with TMM Factors (CTF) and Counts adjustment with Upper quartile Factors (CUF) normalizations to construct coexpression networks results in the most consistently high accuracy networks, and using CLR to transform the network can further increase accuracy in select cases. All the results from this study – for GTEx, SRA, and GTEx resampling datasets, based on tissue-naïve and tissue-aware gold standard, using three different evaluation metrics – are available as a consolidated webpage at [https://krishnanlab.github.io/RNAseq\\_coexpression](https://krishnanlab.github.io/RNAseq_coexpression).

Researchers can use this website to easily examine the performance of various workflows and make appropriate choices for coexpression analysis based on the properties of their RNA-seq dataset of interest. All the scripts to reproduce our results are available at [https://github.com/krishnanlab/RNAseq\\_coexpression](https://github.com/krishnanlab/RNAseq_coexpression) [55], along with scripts that researchers can use to create coexpression networks from their datasets of

interest. Finally, all the coexpression networks constructed in this study are available at <https://doi.org/10.5281/zenodo.5510567> [56].

## **Methods**

### **Data Collection**

Read counts for both SRA and GTEx datasets were downloaded from the recount2 database [19] and processed separately. Recount2 aligns all sequenced reads using Rail-RNA, which eliminates the effect of using different alignment software on separate experiments. We obtained SRA data for any tissue with at least five separate experiments that each had at least five samples. The set of samples from each experiment (project) was considered as an individual dataset from which coexpression networks are inferred (one network per dataset). If a given experiment had samples from multiple tissues, the samples were divided into multiple datasets that each contain samples from the same tissue to yield 543 candidate SRA datasets. We downloaded all available GTEx data, which was a total of 9,657 samples from 31 tissues.

### **Preprocessing**

As a form of quality control, we excluded experiments that recount2 identified as having a misreported paired-end status. Experiments that contained “cell line”, “cell line”, “passage”, “cultured cells”, or “cell culture” in the characteristics metadata were also removed so as to retain primary tissue samples, which left 341 SRA datasets. Next, we discarded low-coverage samples that had zero expression (counts) in at least half of all genes of interest (lncRNA, antisense RNA, and protein-coding genes), and subsequently excluded entire datasets that no longer contained five or more samples.

Any dataset that had a sample removed under this criteria was not retained due to dropping below five samples. Retaining only tissues that still had at least 5 separate experiments left 256 datasets. Finally, we removed genes with very low expression across the board by filtering out those that did not have at least one read per million sample reads in at least 20% of the samples in at least one dataset. This resulted in 22,084 genes in the SRA networks and 20,418 genes in the GTEx networks. Our filtering steps are intentionally relaxed, to retain as much data as possible without keeping large amounts of completely uninformative data.

### **Calculating gene counts**

Recount2 stores quantified expression as base pair counts per gene. We converted these values into gene counts by dividing these base pair per gene counts by the average read length in the sample and accounted for paired-end read samples by further dividing by a factor of two.

### **Refine.bio data collection and processing**

To evaluate the workflows on RNA-Seq data processed with different read alignment and counts quantification methods, we matched as many SRA datasets in our final recount2 data corpus as possible to data in refine.bio. In some cases, not every sample in a recount2 dataset was available in the refine.bio database. If the number of missing samples dropped the dataset to less than 5 samples, we did not use that dataset to construct a network. This procedure brought the total number of usable refine.bio datasets to 188, most of which (120/188) contained all of the samples that were used in the recount2 datasets. These datasets were downloaded from refine.bio as

unnormalized transcript counts. Because some data in refine.bio was aligned using Ensembl release 93 and the rest was aligned using Ensembl release 96, we first subsetted all refine.bio transcripts to only the common transcripts between releases. The transcript counts were summed to gene counts (using Ensembl release 96 and the biomaRt R package [25], then subset to genes present in the recount2 data. Once gene counts are calculated, the rest of each workflow was run exactly the same as it was on the recount2 datasets.

### **Within-sample normalization**

Within-sample normalization is designed to transform the expression levels of genes within the same sample so that they can be compared to each other. Here, we considered counts per million (CPM), transcripts per million (TPM), and reads per kilobase million (RPKM) for performing within-sample normalization of the original raw gene counts [28,48]. Note that RPKM is almost the same as Fragments Per Kilobase Million (FPKM), except FPKM was introduced to accommodate paired-end RNA-seq so it accounts for the fact that two reads can map back to a single fragment. We account for paired-end samples with FPKM, but use the term “RPKM” throughout the manuscript. These three ways of normalizing counts are very commonly used in RNA-seq analysis and account for library size and gene/transcript length in different ways. CPM corrects for library size (expressed in million counts) so that each count is expressed as a proportion of the total number reads in the sample. TPM and RPKM are similar methods that correct for both library size and gene length. Each gene count is divided by both the length of the gene and the sum of counts in the sample, but these

operations are done in a different order. TPM divides counts by gene length (in kb) first to get transcript counts and then by total number of transcripts in the sample, resulting in each normalized sample having the same number of total counts. This is not guaranteed for RPKM since it corrects each gene count for the total number of reads in the sample before correcting for gene length.

### **Between-sample normalization**

Between-sample normalization transforms the expression levels of genes across a group of samples so that gene counts from the same gene in different samples can be more accurately compared to each other. We tested quantile (QNT), trimmed mean of M-values (TMM) [49], and upper quartile (UQ) normalizations [13]. In addition, we tested simple counts adjustment methods we call Counts adjusted with TMM Factors (CTF) and Counts adjusted with Upper quartile Factors (CUF). Quantile normalization is an extremely popular between-sample normalization for microarray samples. Applied to RNA-seq data, QNT forces the distribution of all gene expression values to be exactly the same in each sample. We performed quantile normalization on counts, CPM, TPM, and RPKM using the *preprocessCore* package available from Bioconductor, which implements the quantile normalization described in Bolstad *et al* [50]. TMM normalizes across samples by finding a subset of genes whose variation is mostly due to technical rather than biological factors, i.e. not differentially expressed, then using this subset to calculate a scaling factor to adjust each sample. In brief, each sample is compared to a chosen reference sample. A certain upper and lower percentage of data based on absolute intensity and log-fold-change relative to the reference sample is removed (by

default, 5% for absolute intensity and 30% for log-fold-change) and the log-fold-changes of the remaining set of genes are used to calculate a single scaling factor for the non-reference samples. UQ normalization first removes all zero-count genes and calculates a scaling factor for each sample to match the 75% quantile of the counts in all the samples. In both TMM and UQ, the scaling factors are made to multiply to one before they are used to adjust the library sizes of each sample. These adjusted library sizes are then used in place of the original library size for a calculation otherwise identical to CPM. We used the *edgeR* package [51] to calculate TMM and UQ scaling factors. These factors were also used for CTF and CUF, respectively, where they served as a divisor for each gene count in the proper sample.

### **Gene type filtering**

We chose to keep only long RNA gene types (mRNA (protein-coding), lncRNA, antisense RNA) as those are the most common gene types used in coexpression analysis and shorter reads make mapping and identification more difficult [52,53]. The excluded gene types (mostly short RNAs) are also unlikely to show up in our functional gold standard as there is very little functional information about these gene types. Therefore, relationships between genes of these types are harder to evaluate.

### **Data Transformation**

A log transformation is standard procedure when working with RNA-seq data, as the expected variance grows with the mean for gene counts [35]. A pseudocount is added to the gene count before taking the log. We use the hyperbolic arcsine (asinh) transformation, which is defined along the entire real number line and circumvents the

need for predefining a constant pseudocount and instead calculates a “pseudocount” that is proportional to the original gene count. The asinh function compresses smaller values of  $x$  less than a function like the natural logarithm [36,37].

We also compared asinh to variance stabilizing transformation (VST) [38] and regularized logarithm transformation (rlog) [35] implemented in the DESeq2 R package. These were tested on the GTEx and SRA datasets, except for the six largest GTEx datasets due to the prohibitively long running time of the rlog transformation.

### **Network construction**

A coexpression network was constructed for each individual dataset by calculating the Pearson correlation coefficient between every pair of genes in that dataset using the *Distancer* tool in the *Sleipnir* C++ library [54]. These correlations were treated as the edge weight between gene pairs. We chose Pearson correlation as it has been repeatedly shown to provide a robust measure of gene-gene correlations, especially in small-to-medium-sized datasets that are produced by individual laboratories [7,55]. Since Spearman correlation is also popular in coexpression analysis, we compared these two correlation metrics on our top ten workflows and found that Pearson correlation results in more accurate coexpression networks than Spearman correlation in both GTEx and SRA datasets, particularly in ensuring the accuracy of the top-scoring edges (Appendix: **Fig. A2.16**).

### **Network transformation**

We tested two common methods of network transformation, weighted topological overlap (WTO) [9] and context likelihood of relatedness (CLR) [10], that use different



aspects of network topology to correct the raw coexpression network. The general idea of WTO is to increase the edge weight between gene pairs that share a high number of network neighbors while diminishing edge weight between gene pairs that are tightly connected to very different sets of genes in the network. All edges in the resulting network will have normalized weighted between zero and one. CLR reweights the edge for each gene pair  $(i, j)$  based on how different the original weight of that edge is relative to all of the connections to gene  $i$  and all connections to gene  $j$  (to the rest of the genes in the network). For instance, CLR will upweight an edge between two genes if the edge weight is high compared to all of the other connections of both genes. WTO was implemented using the *wTO* function with the “sign” method in the *wTO* package [56] and CLR was implemented using the *Dat2Dab* function in the *Sleipnir* C++ library.

### **Network evaluation**

The goal of coexpression networks is to capture true functional relationships between genes in the cellular context of the original dataset. Therefore, we evaluated the accuracy of each coexpression network by comparing it to two gold standards, one representing known generic (tissue-naive) functional relationships and the other representing known tissue-aware gene functional relationships. We assembled these gold standards by beginning with a set of manually-selected Gene Ontology Biological Process (GOBP) terms [55,57] that were deemed to be specific enough to be confident that any genes co-annotated to them could be considered functionally related via experimental follow-up (see *Supplemental Note* in Appendix). Specifically, curators were considering the question “if unknown gene/protein G were predicted to be annotated to

GOBP term T, would that be enough to consider experimentally testing this relationship between G and T?” Then, any pair of genes that were co-annotated to the same specific GOBP term was set as a positive edge in the gold standard. We only used annotations based on experimental (GO evidence codes: EXP, IDA, IPI, IMP, IGI, TAS) or curated evidence (IC). We explicitly ignored gene-term annotations made based on expression (GO evidence code: IEP) to avoid circularity when comparing coexpression-derived interactions to this gold-standard. We next had to determine which pairs of genes among the ones with at least one positive edge could be declared as negative edges, i.e., gene pairs that are unlikely to be functionally related based on prior knowledge. To be clear, ‘positive’ and ‘negative’ are used here based on machine learning parlance to indicate interactions and non-interactions, respectively, and do not correspond to the sign of the relationship. This way, the terms are consistent with how we refer to true/false positive/negative edges. Following previous work, we ignored gene pairs not co-annotated to any specific term but still interact with many of the same genes in the gold standard (determined based on each being annotated to two different terms that contained very similar sets of genes; hypergeometric test; p-value <0.05). We also ignored gene pairs that were not co-annotated to any specific term but were co-annotated to certain general GOBP terms, thus introducing ambiguity in whether they are functionally related or not. All remaining gene pairs were considered negatives. We built the naive gold standard using the *Answerer* function in the *Sleipnir* C++ library. We created the tissue-aware gold standards for as many tissues as possible by subsetting the naive gold standard based on genes known to be specifically expressed

in a particular tissue. We obtained tissue-aware genes from the TISSUES 2.0 database Knowledge channel [58]. The knowledge channel contains curated manual annotations of tissue expression provided by UniProtKB. For a given tissue, a positive edge from the naive gold standard was kept in its tissue-aware standard if both genes were expressed in that tissue. Negative edges were kept if both genes were expressed in that tissue, or if one gene is expressed in the tissue and the other gene is expressed in one of the other tissues considered. Only standards containing at least 50 positive edges were used for evaluation, resulting in 24 tissue-aware gold-standards. We specifically excluded epithelium from consideration for a tissue-aware standard, as there is no straightforward way to determine the body site each sample was taken from.

We used the *DChecker* function in the *Sleipnir* C++ library to compare each coexpression network to each gold-standard and return the number of true positives, false positives, true negatives, and false negatives at various edge weight thresholds. These numbers were used to calculate the area under the precision-recall curve (auPRC) using the *trapz* function in the *pracma* package. Since gene functional relationship gold-standards of different tissues have different proportions of positives to negatives, the original auPRC scores are not directly comparable to each other. Therefore, we divided each auPRC by its “prior” – the auPRC of a random predictor, equal to the fraction of positives among all positive and negative edges – and expressed the performance as the logarithm of this ratio to enable tissue-to-tissue comparisons.

The *Supplemental Note* in the Appendix contains more details on the i) gene functional relationship gold standard, ii) additional gold standards that we explored (including spike-ins [59,60]) and their limitations, and iii) calculation of the evaluation metrics.

#### Workflow comparison and analysis by parts

To assess whether two workflows resulted in coexpression networks that were significantly different in quality, we used a paired Wilcoxon rank sum test to compare the auPRC scores across all coexpression networks generated by those two workflows. After calculating p-values, we performed a correction for multiple testing with the Benjamini-Hochberg procedure and declared workflows with  $FDR \leq 0.01$  as being significantly different. Further, each workflow is a combination of method choices at multiple stages. So, to determine the impact of including a particular method in a workflow, we across aggregated workflows to calculate the proportion of times that including a particular method in a workflow resulted in the workflow being significantly greater than one that did not include the method. As it is not possible to do within-sample normalization and then do TMM, UQ, CTF, or CUF, any workflow including CPM, TPM, or RPKM was excluded when assessing between-sample normalization methods so that method being compared to each other based on the same number of aggregated workflows. For similar reasons, workflows involving TMM, UQ, CTF, or CUF were not considered for the analysis of within-sample normalization methods.

## **GTEX resampling**

To simulate uniformly-processed datasets that have sample sizes similar to datasets from SRA, we chose nine sample sizes (5, 6, 7, 9, 11, 13, 16, 25, and 40) based on the distribution of SRA dataset sample sizes. Then, from each GTEX dataset with at least 70 samples, we randomly sampled a “dataset” of each sample size, repeating this sampling ten times to create 10 datasets per sample size from each GTEX dataset. One coexpression network was constructed and evaluated from each of these GTEX-resampled datasets in the same manner outlined above.

## **Experimental factor analysis**

In addition to dataset size (i.e. number of samples), the quality of the coexpression network reconstructed from a dataset could also depend on the similarity between the samples in that dataset as well as the total number of mapped reads. We performed an analysis to determine this impact using the GTEX-resampled datasets and the original SRA datasets. Since SRA datasets are not large enough to do resampling for sample size analysis, we split them into five groups with equal number of datasets, with datasets in each group having similar sample sizes. We define sample similarity for a given dataset as the median spearman correlation between all samples using the 50% most variable genes in the GTEX tissue they came from for the resampled GTEX datasets, or the median spearman correlation between all samples using the 50% most variable genes in each individual dataset in the case of the SRA networks. Read count diversity is calculated by summing the gene counts of each sample in a given dataset and taking their standard deviation. Based on each of these measures – sample

similarity and read count diversity – we divided the datasets into five groups of equal size while taking care to check that each group contained datasets with similar sample sizes. For the tissue analysis, we could only determine significance in SRA tissues that had at least 15 datasets.

## REFERENCES

1. van Dam S, Vösa U, van der Graaf A, Franke L, Magalhães D, Pedro J. Gene co-expression analysis for functional classification and gene–disease predictions. *Brief Bioinform.* 2018;19:575–92.
2. Allocco DJ, Kohane IS, Butte AJ. Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics.* 2004;5:18.
3. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci. National Academy of Sciences;* 1998;95:14863–8.
4. Segal E, Friedman N, Koller D, Regev A. A module map showing conditional activity of expression modules in cancer. *Nat Genet.* 2004;36:1090–8.
5. Carpenter AE, Sabatini DM. Systematic genome-wide screens of gene function. *Nat Rev Genet.* 2004;5:11–22.
6. Zhang B, Horvath S. A General Framework for Weighted Gene Co-Expression Network Analysis. *Stat Appl Genet Mol Biol. De Gruyter;* 2005; 4.
7. Zhu Q, Wong AK, Krishnan A, Aure MR, Tadych A, Zhang R, et al. Targeted exploration and analysis of large cross-platform human transcriptomic compendia. *Nat Methods.* 2015;12:211–4.
8. Evans C, Hardin J, Stoebel DM. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief Bioinform.* 2017;19:776–92.
9. Nowick K, Gernat T, Almaas E, Stubbs L. Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain. *Proc Natl Acad Sci.* 2009;106:22358–63.
10. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, et al. Large-Scale Mapping and Validation of *Escherichia coli* Transcriptional Regulation from a Compendium of Expression Profiles. *PLoS Biol.* 2007;5.
11. Reverter A, Barris W, McWilliam S, Byrne KA, Wang YH, Tan SH, et al. Validation of alternative methods of data normalization in gene co-expression studies. *Bioinformatics. Oxford Academic;* 2005;21:1112–20.
12. Lim WK, Wang K, Lefebvre C, Califano A. Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics. Oxford Academic;* 2007;23:i282–8.

13. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*. 2010;11:94.
14. Maza E, Frasse P, Senin P, Bouzayen M, Zouine M. Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments. *Commun Integr Biol*. 2013;6.
15. Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform*. Oxford Academic; 2013;14:671–83.
16. Zyprych-Walczak J, Szabelska A, Handschuh L, Górczak K, Klamecka K, Figlerowicz M, et al. The Impact of Normalization Methods on RNA-Seq Data Analysis. *BioMed Res Int*. 2015.
17. Abbas-Aghababazadeh F, Li Q, Fridley BL. Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing. *PLOS ONE*. Public Library of Science; 2018;13:e0206312.
18. Ballouz S, Verleyen W, Gillis J. Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. *Bioinformatics*. 2015;31:2123–30.
19. Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, et al. Reproducible RNA-seq analysis using recount2. *Nat Biotechnol*. 2017;35:319–21.
20. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. Nature Publishing Group; 2013;45:580–5.
21. Leinonen R, Sugawara H, Shumway M. The Sequence Read Archive. *Nucleic Acids Res*. 2011;39:D19–21.
22. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. *Nat Genet*. 2000;25:25–9.
23. Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE*. 2015;10.
24. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. *Proc 23rd Int Conf Mach Learn*. New York, NY, USA: Association for Computing Machinery; 2006. p. 233–40.
25. Greene CS, Hu D, Jones RWW, Liu S, Mejia DS, Patro R, et al. refine.bio.



Refine.bio.

26. Oshlack A, Wakefield MJ. Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct*. 2009;4:14.
27. Huang J, Vendramin S, Shi L, McGinnis KM. Construction and Optimization of a Large Gene Coexpression Network in Maize Using RNA-Seq Data. *Plant Physiol. American Society of Plant Biologists*; 2017;175:568–83.
28. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci*. 2012;131:281–5.
29. Zhao S, Ye Z, Stanton R. Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA*. 2020;rna.074922.120.
30. Hicks SC, Irizarry RA. When to use Quantile Normalization? *Genomics*; 2014 Dec.
31. Cosgrove EJ, Gardner TS, Kolaczyk ED. On the Choice and Number of Microarrays for Transcriptional Regulatory Network Inference. *BMC Bioinformatics*. 2010;11:454.
32. Wang Y, Hicks SC, Hansen KD. Co-expression analysis is biased by a mean-correlation relationship. *Genomics*; 2020 Feb.
33. Farahbod M, Pavlidis P. Differential coexpression in human tissues and the confounding effect of mean expression levels. *Bioinformatics. Oxford Academic*; 2019;35:55–61.
34. Ramsköld D, Wang ET, Burge CB, Sandberg R. An Abundance of Ubiquitously Expressed Genes Revealed by Tissue Transcriptome Sequence Data. *PLOS Comput Biol*. 2009;5:e1000598.
35. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15:550.
36. Johnson NL. Systems of Frequency Curves Generated by Methods of Translation. *Biometrika*. [Oxford University Press, Biometrika Trust]; 1949;36:149–76.
37. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods. Nature Publishing Group*; 2012;9:473–6.
38. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010;11:R106.
39. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol. Nature Publishing Group*; 2014;

32:896–902.

40. Parsana P, Ruberman C, Jaffe AE, Schatz MC, Battle A, Leek JT. Addressing confounding artifacts in reconstruction of gene co-expression networks. *Genome Biol.* 2019;20:94.

41. Jaffe AE, Hyde T, Kleinman J, Weinberg DR, Chenoweth JG, McKay RD, et al. Practical impacts of genomic data “cleaning” on biological discovery using surrogate variable analysis. *BMC Bioinformatics.* 2015;16:372.

42. Mao W, Rahimikollu J, Hausler R, Chikina M. DataRemix: a universal data transformation for optimal inference from gene expression datasets. *Bioinformatics.* 2021 Apr; 37(7): 984–991.

43. Zhang Y, Cuerdo J, Halushka MK, McCall MN. The effect of tissue composition on gene co-expression. *Brief Bioinform.* 2021; 22(1):127–139.

44. Farahbod M, Pavlidis P. Untangling the effects of cellular composition on coexpression analysis. *Genome Res.* 2020;30:849–59.

45. Crow M, Paul A, Ballouz S, Huang ZJ, Gillis J. Exploiting single-cell expression to characterize co-expression replicability. *Genome Biol.* 2016;17:101.

46. Harris BD, Crow M, Fischer S, Gillis J. Multiscale Co-Expression in the Brain. *Bioinformatics;* 2020.

47. Vieth B, Parekh S, Ziegenhain C, Enard W, Hellmann I. A systematic evaluation of single cell RNA-seq analysis pipelines. *Nat Commun.* 2019;10:1–11.

48. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods.* Nature Publishing Group; 2008;5:621–8.

49. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 2010;11:R25.

50. Bolstad BM, Irizarry RA, Åstrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics.* 2003;19:185–93.

51. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* 2010;26:139–40.

52. Garber M, Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Methods.* Nature

Publishing Group; 2011;8:469–77.

53. Łabaj PP, Leparc GG, Linggi BE, Markillie LM, Wiley HS, Kreil DP. Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. *Bioinformatics*. 2011;27:i383–91.

54. Huttenhower C, Schroeder M, Chikina MD, Troyanskaya OG. The Sleipnir library for computational functional genomics. *Bioinformatics*. 2008;24:1559–61.

55. Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, et al. Understanding multicellular function and disease with human tissue-specific networks. *Nat Genet*. 2015;47:569–76.

56. Gysi DM, Voigt A, Fragoso T de M, Almaas E, Nowick K. wTO: an R package for computing weighted topological overlap and a consensus network with integrated visualization tool. *BMC Bioinformatics*. 2018 [cited 2018 Nov 5];19.

57. Myers CL, Barrett DR, Hibbs MA, Huttenhower C, Troyanskaya OG. Finding function: evaluation methods for functional genomic data. *BMC Genomics*. 2006;7:187.

58. Palasca O, Santos A, Stolte C, Gorodkin J, Jensen LJ. TISSUES 2.0: an integrative web resource on mammalian tissue expression. *Database J Biol Databases Curation*. 2018.

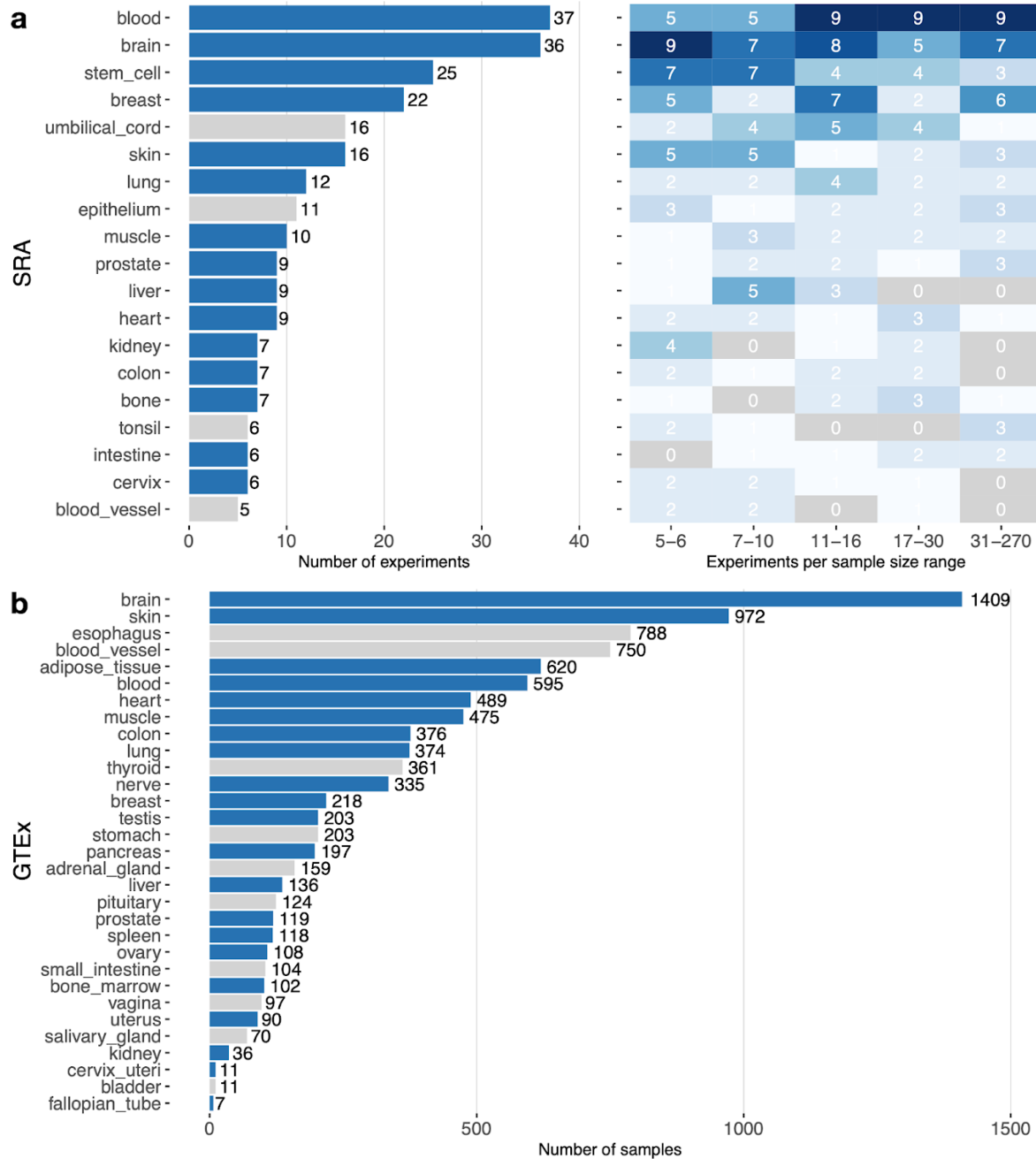
59. McCall MN, Almudevar A. Affymetrix GeneChip microarray preprocessing for multivariate analyses. *Brief Bioinform*. 2012;13:536–46.

60. Qing T, Yu Y, Du T, Shi L. mRNA enrichment protocols determine the quantification characteristics of external RNA spike-in controls in RNA-Seq studies. *Sci China Life Sci*. 2013;56:134–42.

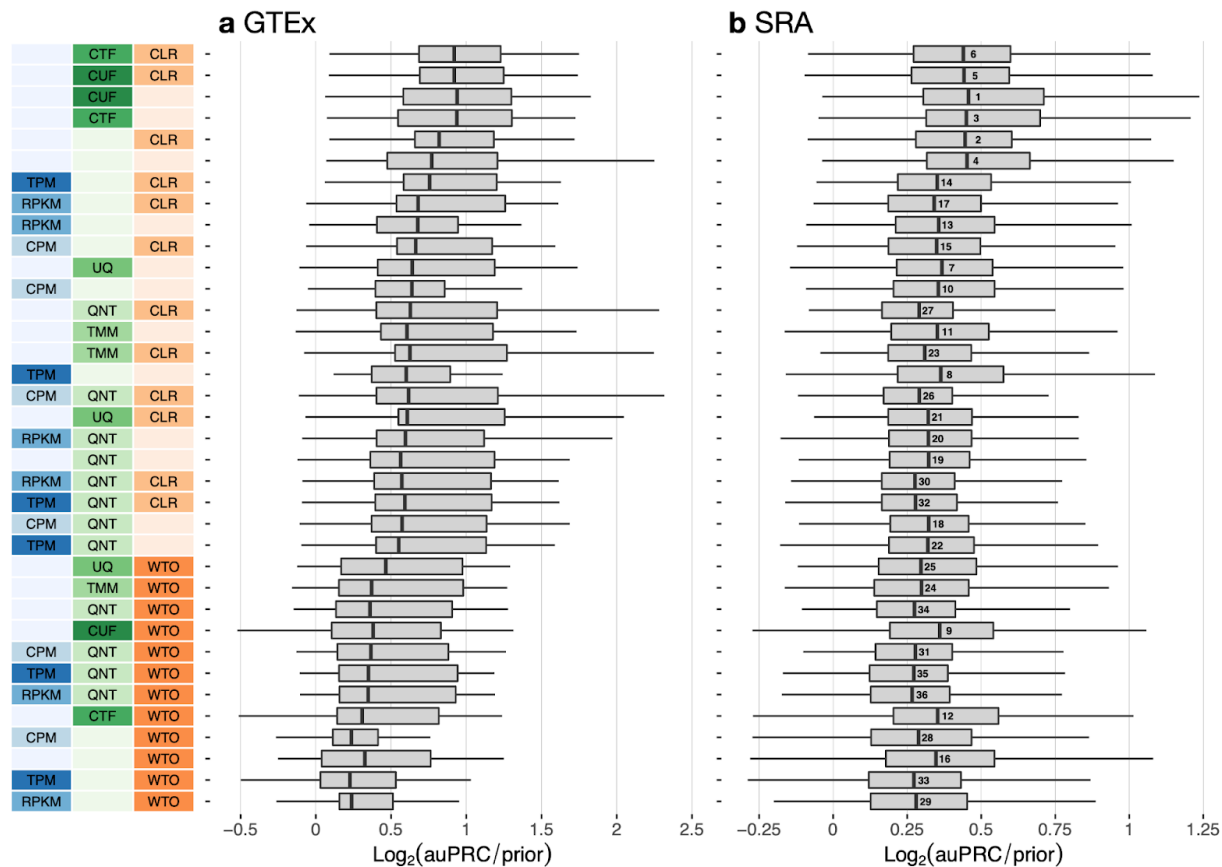
61. Johnson KA, Krishnan A. RNAseq\_coexpression. Github. [https://github.com/krishnanlab/RNAseq\\_coexpression](https://github.com/krishnanlab/RNAseq_coexpression) (2021).

62. Johnson KA, Krishnan A. Coexpression networks of 31 GTEx and 256 SRA RNA-Seq datasets. Zenodo. <https://zenodo.org/record/5510567#.YZ1lrfHMJTY> (2021).

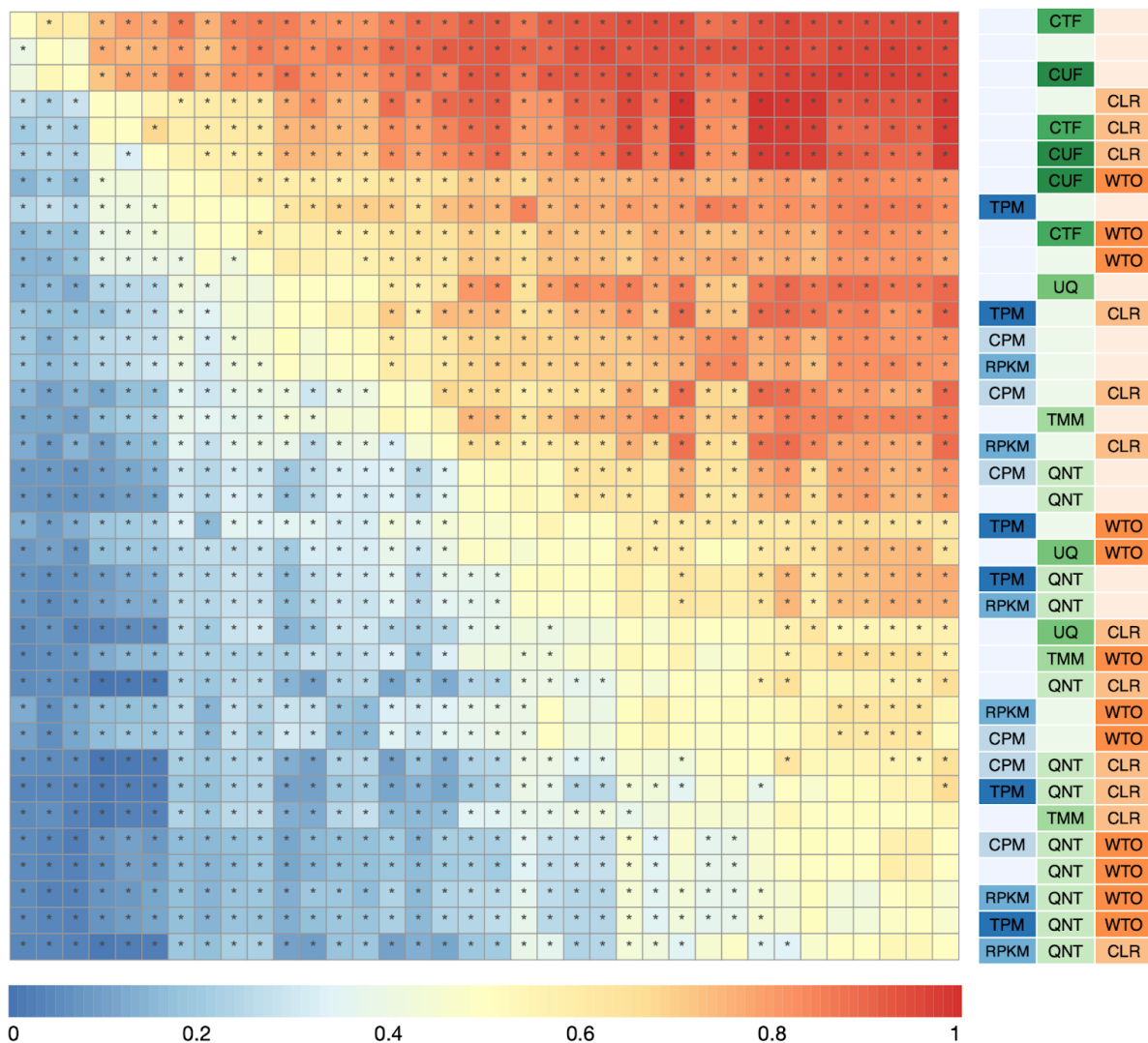
## APPENDIX



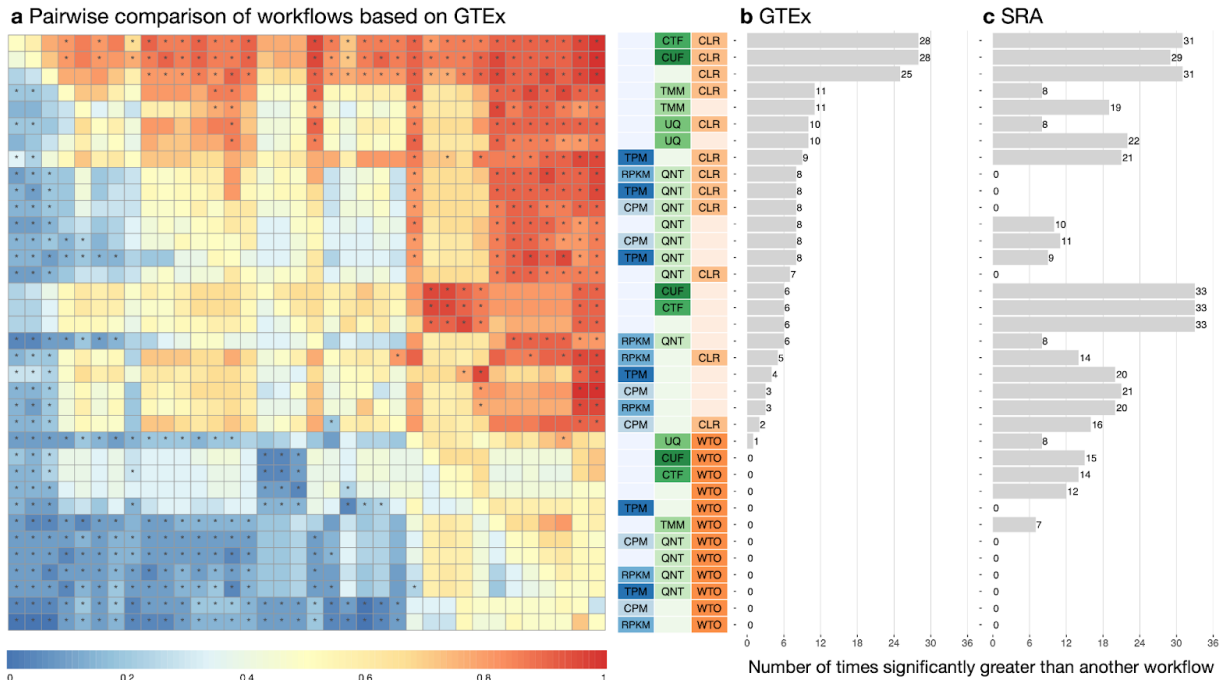
**Figure A2.1. Recount2 data used in this study.** (a) The barplot shows the number of experiments from each tissue in the SRA data. The heatmap on the right shows the number of projects/experiments that have a particular sample size for each tissue. (b) The barplot shows the number of samples for each GTEx tissue. In the barplots, blue bars indicate tissues for which we were able to create a tissue-aware gold standard. Tissues with gray bars were evaluated on the tissue-naïve standard only.



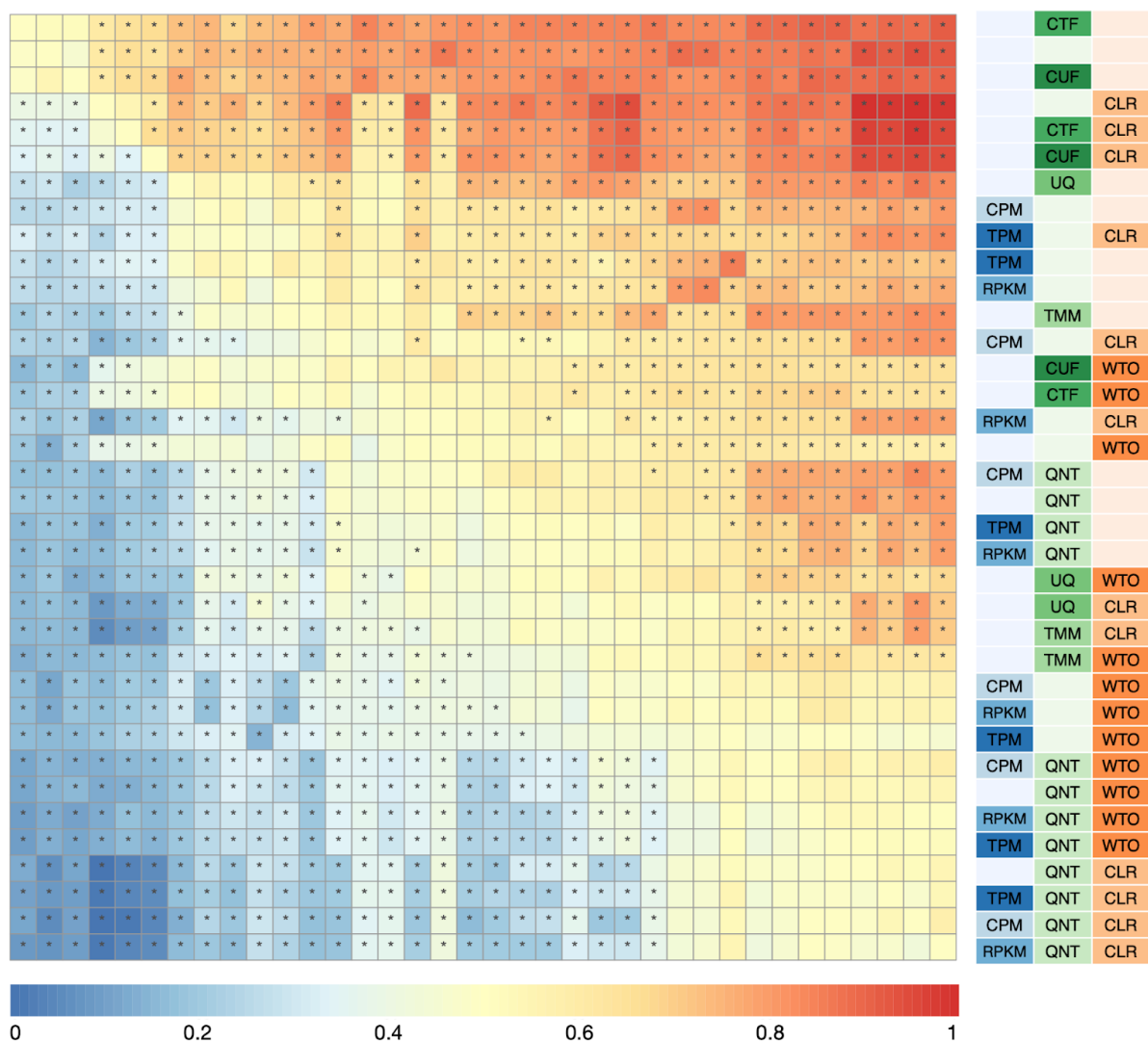
**Figure A2.2. Overall performance of workflows based on the tissue-aware gold standard.** The plots show the aggregate accuracy of all coexpression networks resulting from each individual workflow using (a) GTEx and (b) SRA datasets, evaluated based on the tissue-aware gold standard. The workflows (rows) are described in terms of the specific method used in the within-sample normalization (blues), between-sample normalization (greens), and network transformation (oranges) stages. The performance of each workflow is presented as boxplots (without outliers) that summarizes the log<sub>2</sub>(auPRC/prior) of each workflow where auPRC is the area under the precision recall curve (see **Methods**). The workflows are ordered by their median log<sub>2</sub>(auPRC/prior) for the GTEx data. The numbers inside the SRA boxes indicate rank by median log<sub>2</sub>(auPRC/prior) of the workflows for the SRA data. *Figure 2.2* contains these performance plots based on the tissue-naïve gold standard.



**Figure A2.3. Dataset-level pairwise comparison of workflow performance for SRA datasets based on the tissue-naïve gold standard.** The heatmap shows the relative performance of a pair of workflows, corresponding to a row and a column, directly compared to each other for the SRA datasets based on the tissue-naïve gold standard. The color in each cell (row, column) represents the proportion of datasets for which the workflow along the row has a higher  $\log_2(\text{auPRC}/\text{prior})$  than the workflow along the column. Comparisons that are statistically significant (corrected  $p < 0.01$ ) based on a paired Wilcoxon test are marked with an asterisk. *Figure 2.3a* contains the corresponding heatmap for GTEx datasets.

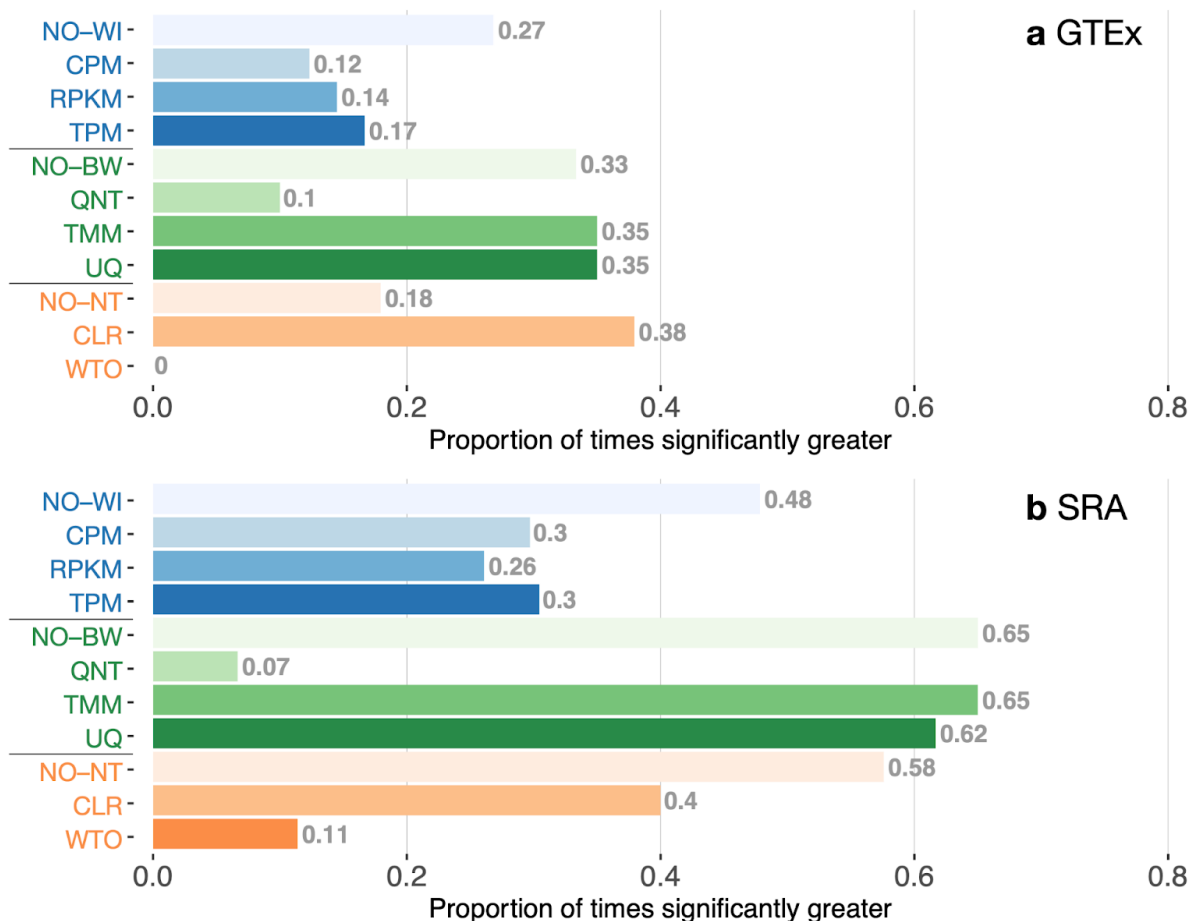


**Figure A2.4. Dataset-level pairwise comparison of workflow performance for GTEx and SRA datasets based on the tissue-aware gold standard.** (a) The heatmap shows the relative performance of a pair of workflows, corresponding to a row and a column, directly compared to each other for the GTEx datasets based on the tissue-aware gold standard. The color in each cell (row, column) represents the proportion of datasets for which the workflow along the row has a higher  $\log_2(\text{auPRC}/\text{prior})$  than the workflow along the column. Comparisons that are statistically significant (corrected  $p < 0.01$ ) based on a paired Wilcoxon test are marked with an asterisk. *Figures A2.5* contains the corresponding heatmap for the SRA datasets. (b and c) Barplots show the number of times each workflow was significantly greater than another workflow for GTEx (left) and SRA (right) datasets. Figure 2.3 and A2.3 contain these performance plots based on the tissue-naïve gold standard.

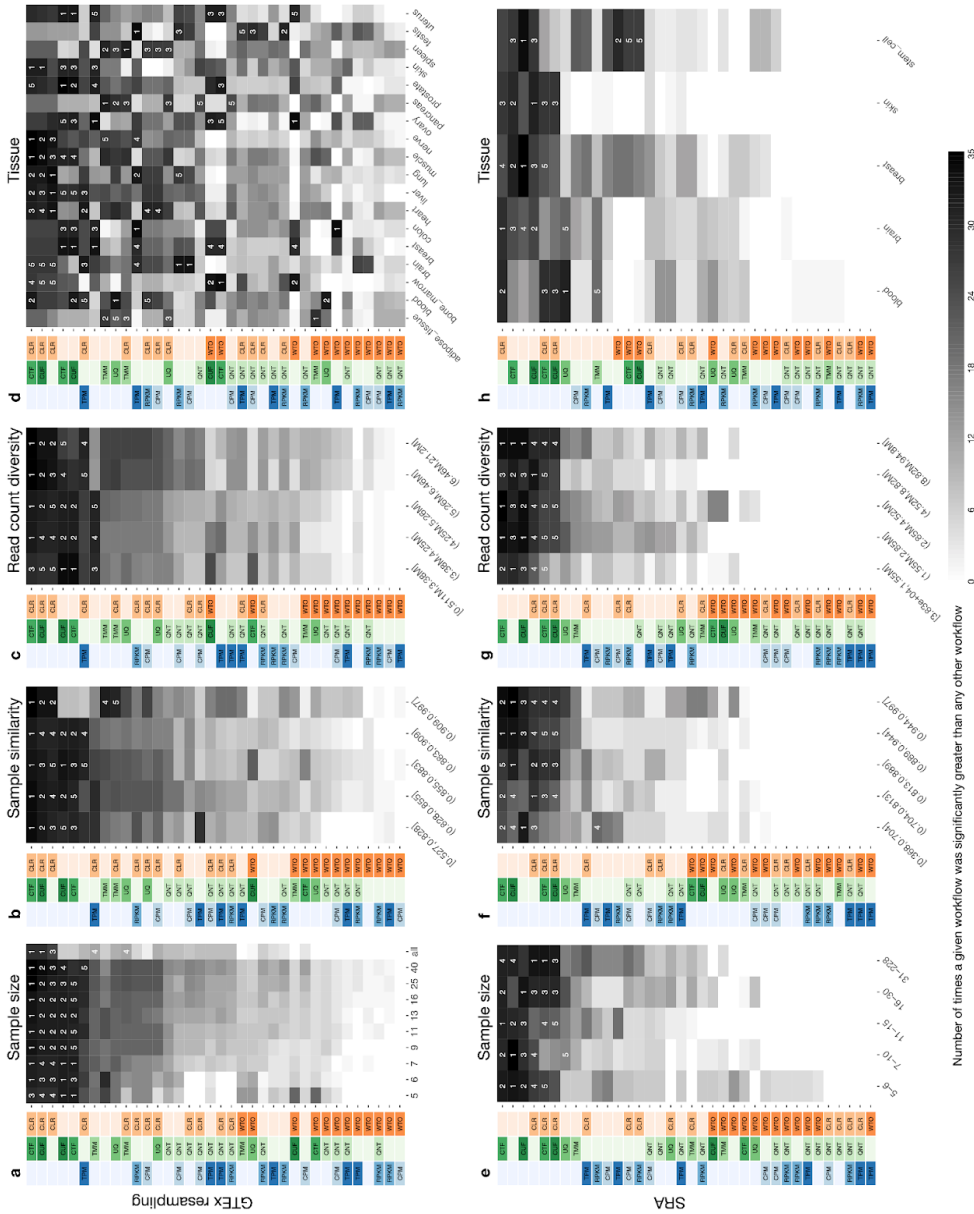


**Figure A2.5. Dataset-level pairwise comparison of workflow performance for SRA datasets based on the tissue-aware gold standard.** The heatmap shows the relative performance of a pair of workflows, corresponding to a row and a column, directly compared to each other for the SRA datasets based on the tissue-aware gold standard. The color in each cell (row, column) represents the proportion of datasets for which the workflow along the row has a higher  $\log_2(\text{auPRC}/\text{prior})$  than the workflow along the column. Comparisons that are statistically significant (corrected  $p < 0.01$ ) based on a paired Wilcoxon test are marked with an asterisk. *Figure A2.4a* contains the corresponding heatmap for GTEx datasets.

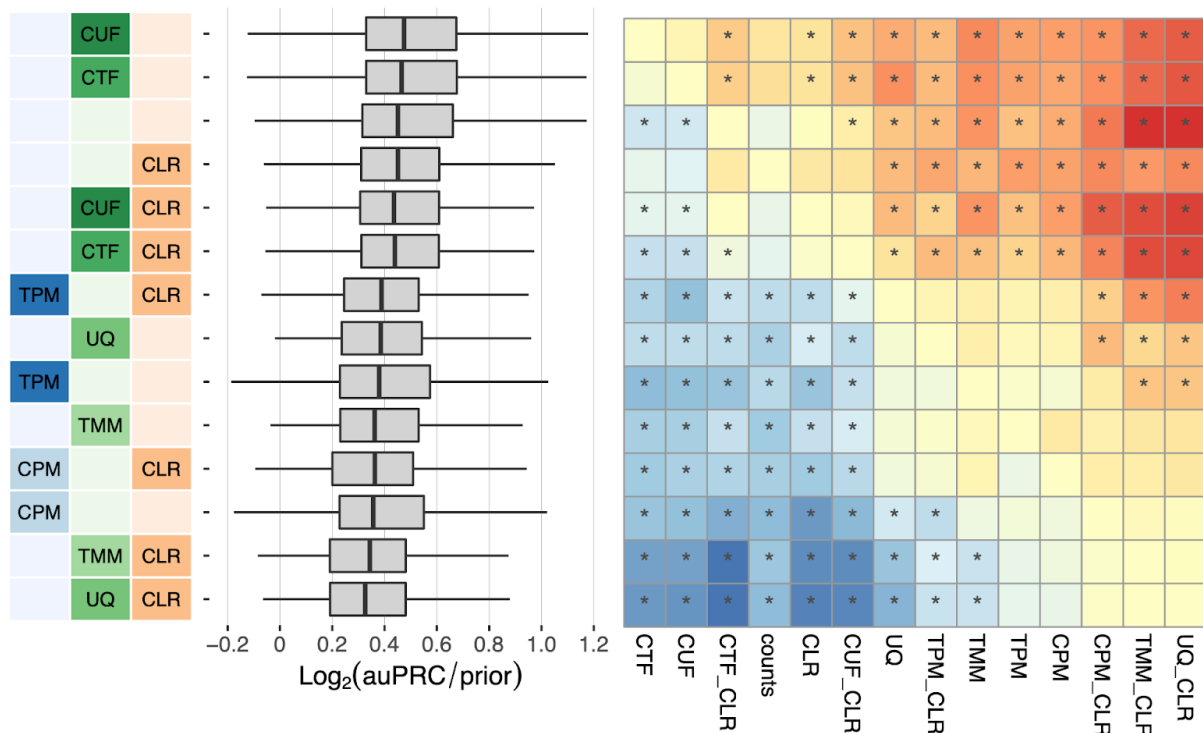




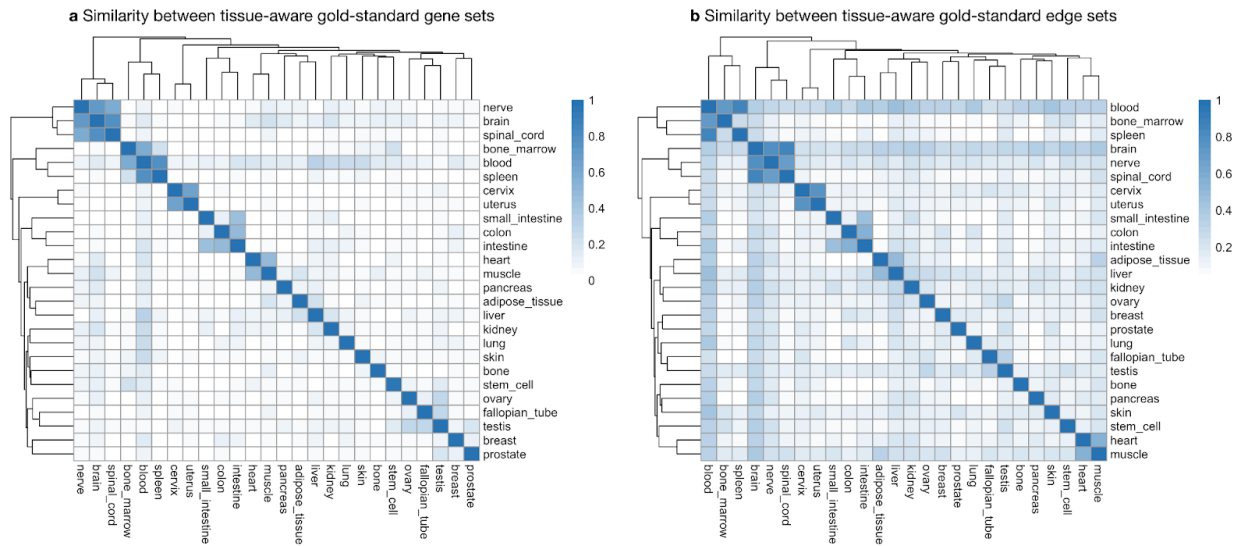
**Figure A2.6. Impact of individual methods on performance of workflows based on the tissue-aware gold standard.** Each bar in the two barplots, corresponding to a specific method, shows the proportion of times (x-axis) that workflows including that particular method (y-axis) were significantly better than other workflows. The barplots correspond to performance for the (a) GTEx and (b) SRA datasets evaluated on the tissue-naive gold standard. In order to make the comparison of between-sample normalization methods fair, workflows including CPM, RPKM, or TPM were left out because it is not possible to pair them with TMM or UQ normalization. Similarly, TMM and UQ methods are not included for “no within-sample normalization” (NO–WI). *Figure 2.4* contains these barplots based on the tissue-naive gold standard.



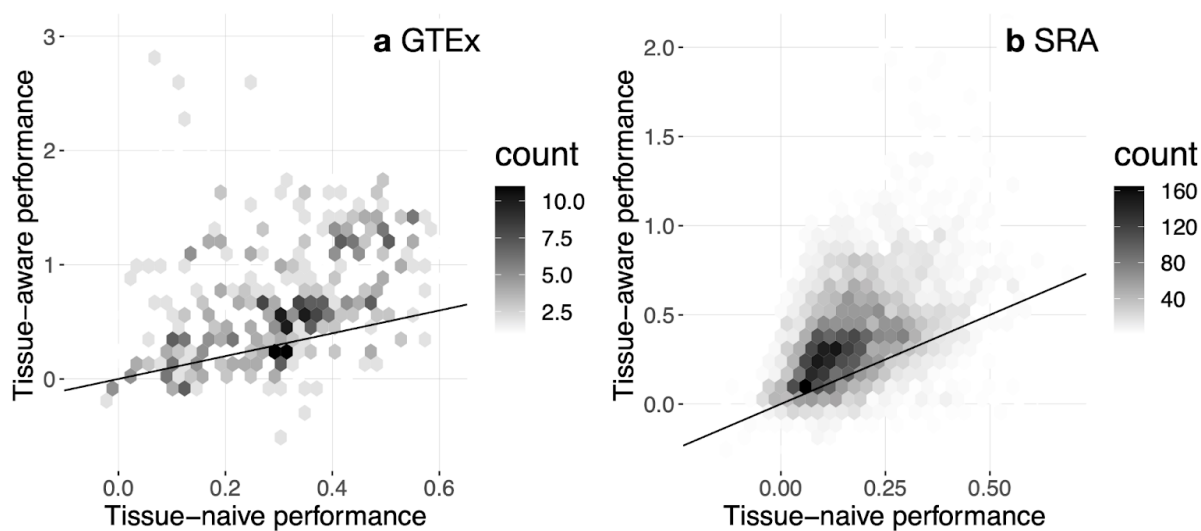
**Figure A2.7. Impact of various dataset-related experimental factors on performance of workflows based on the tissue-aware gold standard.** Each heatmap shows the number of times (cell color) each workflow (row) outperforms other workflows as a particular experimental factor pertaining to the input datasets is varied (columns), when the resulting coexpression networks are evaluated based on the tissue-naïve gold standard. The darkest colors indicate workflows that are significantly better than the most other workflows. In addition, the top 5 workflows in each column are marked with their rank, with ties given minimum rank. The heatmaps on the top (a–d) correspond to datasets from GTEx resampling and those on the bottom (e–h) correspond to SRA datasets. The heatmaps from left to right show workflow performance by sample size (a, e; number of samples used to make the coexpression network), sample similarity (b, f; median spearman correlation of 50% most variable genes between samples), library size diversity (c, g; standard deviation of counts sums across samples), and tissue of origin (d, h). Figure 2.5 contains these heatmaps based on the tissue-naïve gold standard.



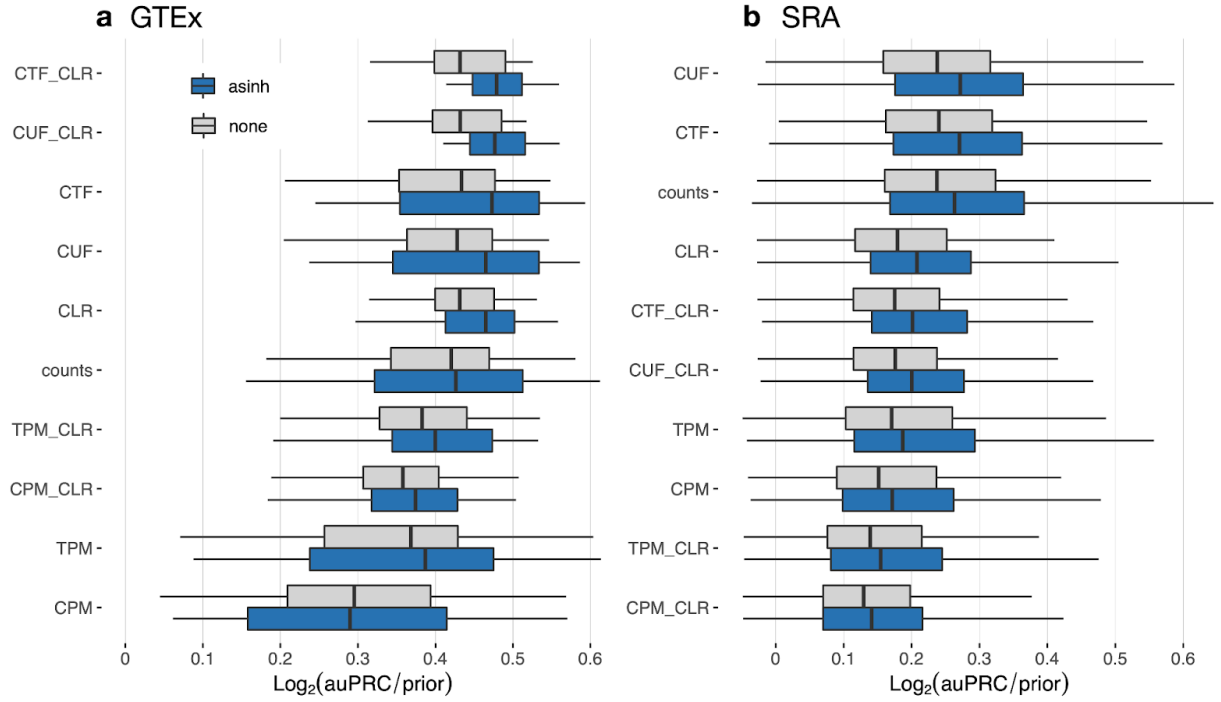
**Figure A2.8. Overall performance of workflows and pairwise-comparison using refine.bio datasets based on the tissue-aware gold standard.** The boxplots show the aggregate accuracy of all coexpression networks resulting from each individual workflow using SRA datasets in refine.bio, evaluated based on the tissue-aware gold standard. The performance of each workflow is presented as boxplots (without outliers) that summarizes the  $\text{log}_2(\text{auPRC}/\text{prior})$  of each workflow where auPRC is the area under the precision recall curve (see **Methods**). The workflows are ordered by their median  $\text{log}_2(\text{auPRC}/\text{prior})$ . The heatmap shows the relative performance of pairs of workflows (rows and columns) directly compared to each other for the refine.bio SRA datasets based on the tissue-aware gold standard. The color in each cell (row, column) represents the proportion of datasets for which the workflow along the row has a higher  $\text{log}_2(\text{auPRC}/\text{prior})$  than the workflow along the column. Comparisons that are statistically significant (corrected  $p < 0.01$ ) based on a paired Wilcoxon test are marked with an asterisk. *Figure 2.6* contains these plots based on the tissue-naive gold standard.



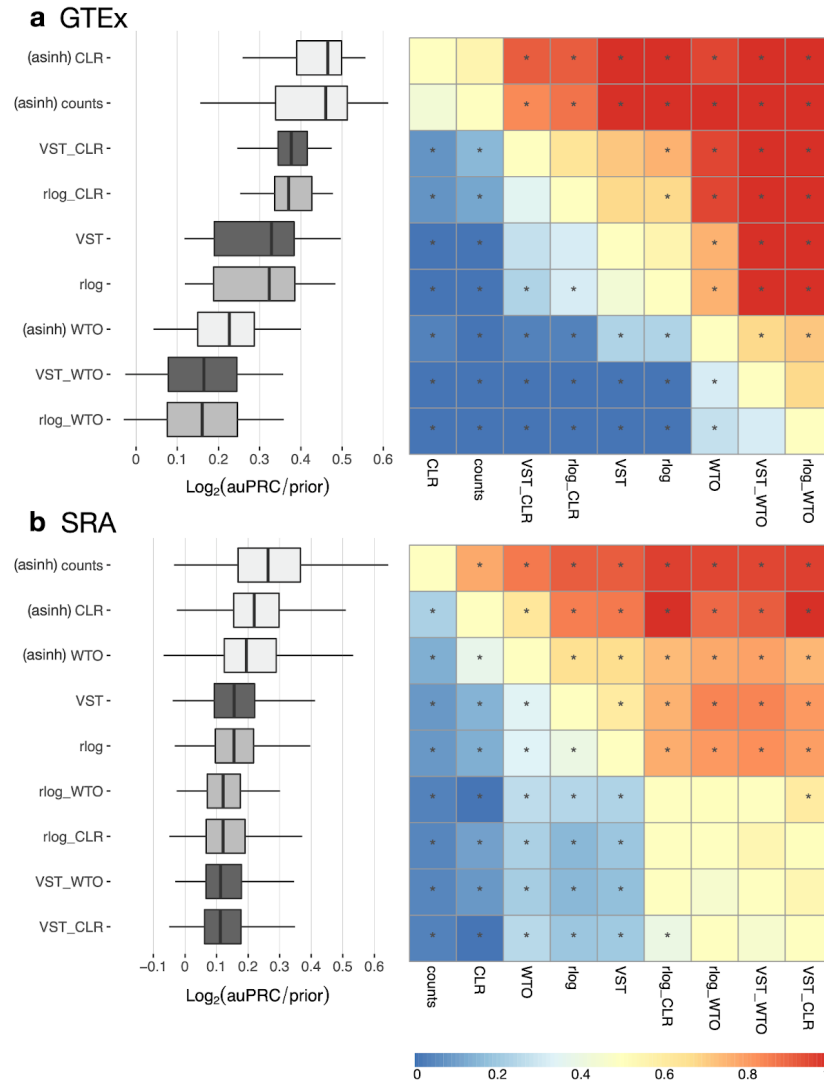
**Figure A2.9. Gene- and edge-based overlap between tissue-aware gold standards.** The heatmaps show the number of (a) genes or (b) edges that are shared between any two given tissue-aware gold standards divided by the total number of genes or edges in the smaller of the two tissue-aware gold standards. Based on the heatmaps, the proportion of shared genes and edges between unrelated tissues is small and therefore each tissue-aware gold standard is evaluating a very different set of biological relationships.



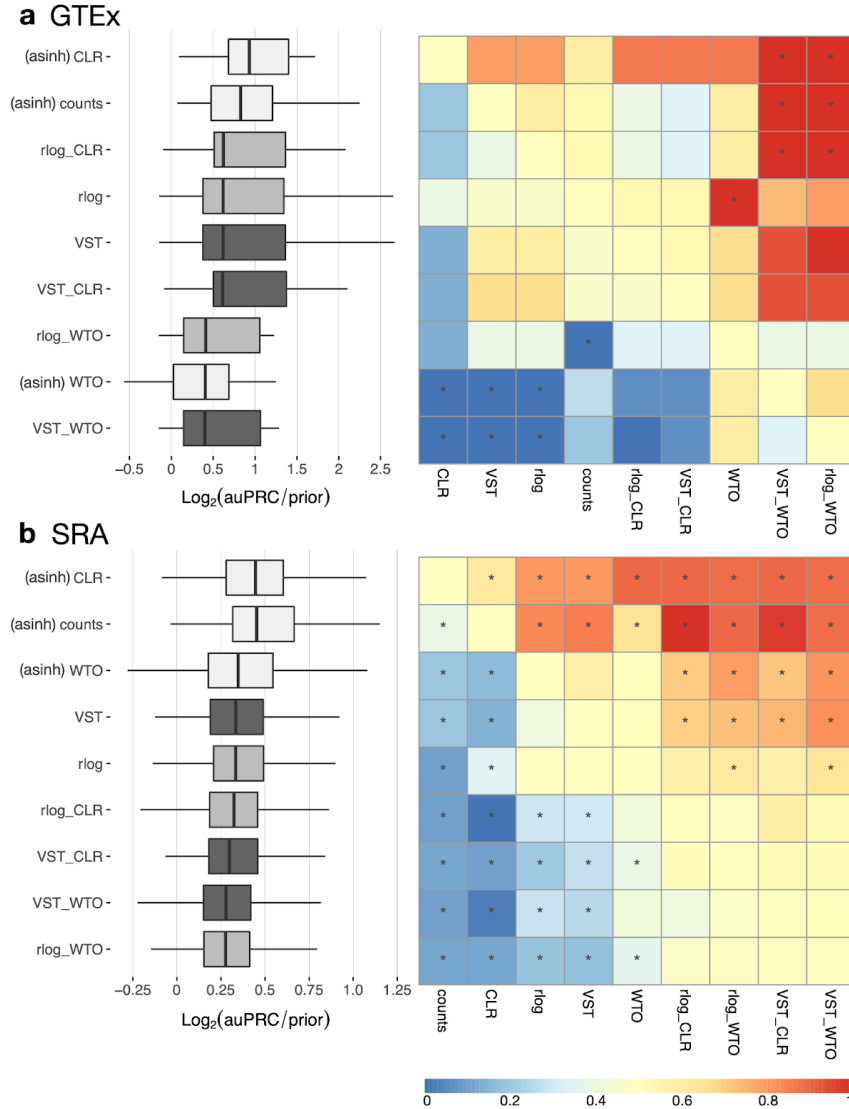
**Figure A2.10. Overall accuracy of coexpression networks when evaluated based on the tissue-naïve and tissue-aware gold standards.** Each density plot – for the (a) GTEx and (b) SRA datasets – shows the distribution of  $\log_2(\text{auPRC}/\text{prior})$  across all workflows and datasets when evaluating based on the tissue-naïve gold standard (x-axis) vs. the tissue-aware gold standard (y-axis). These distributions show that coexpression networks capture tissue-aware gene interactions and emphasizes the importance of evaluating coexpression networks using tissue-aware gold standards.



**Figure A2.11. Overall performance of top workflows with and without asinh transformation based on the tissue-naive gold standard.** The plots show the aggregate accuracy of all coexpression networks resulting from the top ten individual workflows using (a) GTEx and (b) SRA datasets with (blue) and without (gray) the asinh transformation, evaluated based on the tissue-naive gold standard. The workflows (rows) are described in terms of the specific method used in the within-sample normalization, between-sample normalization, and network transformation stages. The performance of each workflow is presented as boxplots (without outliers) that summarizes the  $\text{log}_2(\text{auPRC}/\text{prior})$  of each workflow where auPRC is the area under the precision recall curve (see **Methods**). The workflows are ordered by their median  $\text{log}_2(\text{auPRC}/\text{prior})$  in each panel.

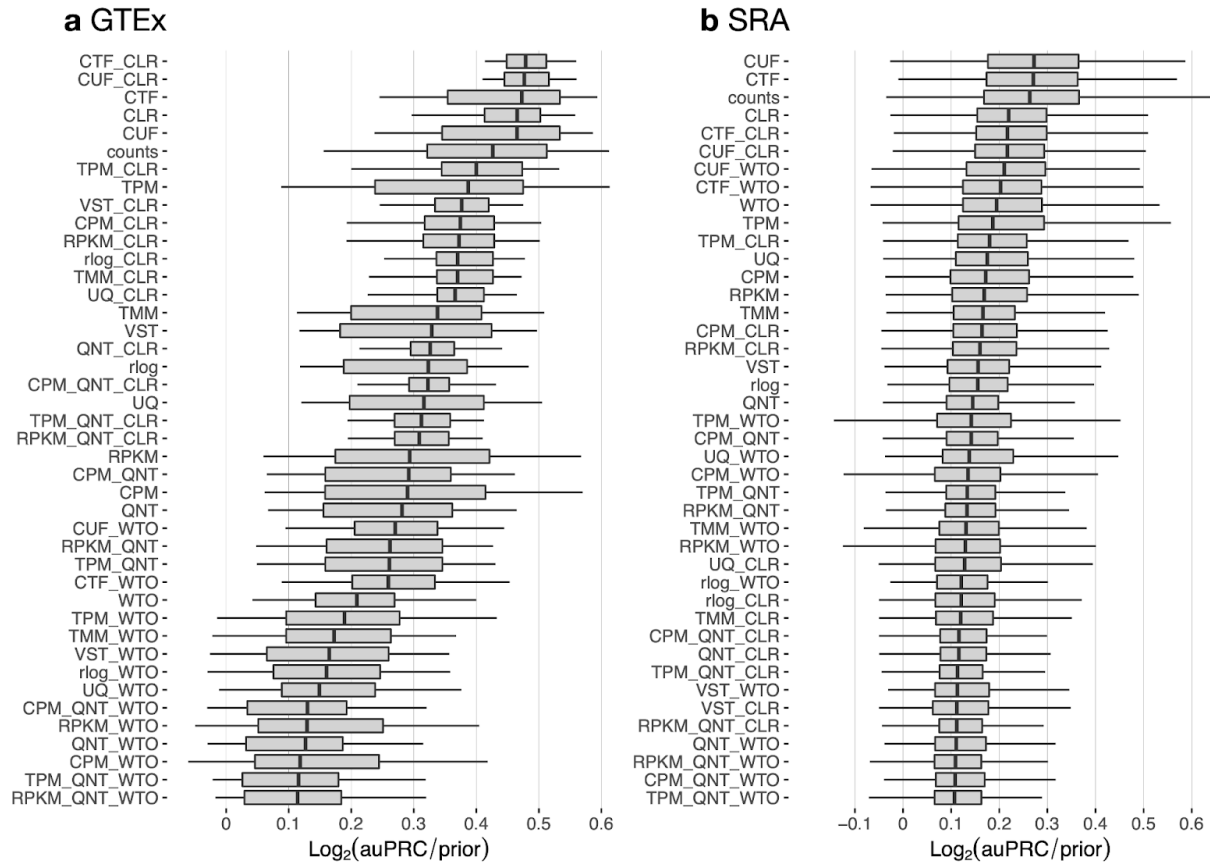


**Figure A2.12. Performance of workflows using different data transformation methods based on the tissue-naive gold standard.** The plots show the aggregate accuracy of all coexpression networks resulting from using (a) GTEx and (b) SRA datasets with different data transformations to adjust gene counts paired with the network transformation methods, evaluated based on the tissue-naive gold standard. The workflows (rows) are combinations of specific data transformations (shades of gray) and network transformations. The performance of each workflow is presented as boxplots (without outliers) that summarize the  $\text{log}_2(\text{auPRC}/\text{prior})$  of each workflow where auPRC is the area under the precision recall curve (see **Methods**). The workflows are ordered by their median  $\text{log}_2(\text{auPRC}/\text{prior})$  in each panel. The heatmaps on the right show the relative performance of a pair of workflows, corresponding to a row and a column, directly compared to each other for the GTEx (a) and SRA (b) datasets based on the tissue-naive gold standard. The color in each cell (row, column) represents the proportion of datasets for which the workflow along the row has a higher  $\text{log}_2(\text{auPRC}/\text{prior})$  than the workflow along the column. Comparisons that are statistically significant (corrected  $p < 0.01$ ) based on a paired Wilcoxon test are marked with an asterisk. The six largest GTEx datasets (adipose\_tissue, blood, blood\_vessel, brain, esophagus, and skin) are not considered in this evaluation because of the considerable amount of computing time required to use rlog transformation on large datasets. CLR and Counts significantly outperformed all other methods on GTEx datasets. For SRA datasets, Counts performed significantly better than all other workflows, and CLR and WTO both performed significantly better than all workflows incorporating VST or rlog.

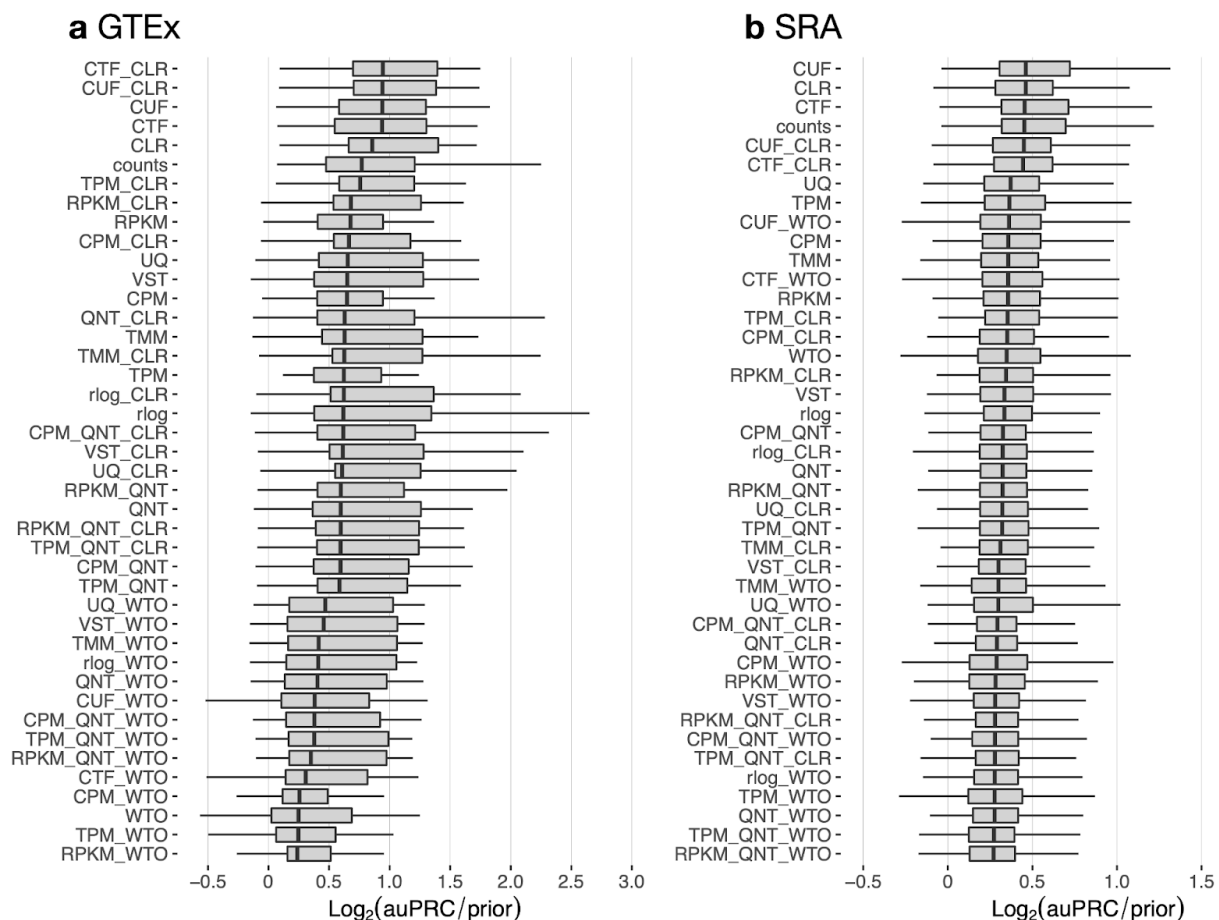


**Figure A2.13. Performance of workflows using different data transformation methods based on the tissue-aware gold standard.** The plots show the aggregate accuracy of all coexpression networks resulting from using (a) GTEX and (b) SRA datasets with different data transformations to adjust gene counts paired with the network transformation methods, evaluated based on the tissue-aware gold standard. The workflows (rows) are combinations of specific data transformations (shades of gray) and network transformations. The performance of each workflow is presented as boxplots (without outliers) that summarize the  $\text{log}_2(\text{auPRC}/\text{prior})$  of each workflow where auPRC is the area under the precision recall curve (see **Methods**). The workflows are ordered by their median  $\text{log}_2(\text{auPRC}/\text{prior})$  in each panel. The heatmaps on the right show the relative performance of a pair of workflows, corresponding to a row and a column, directly compared to each other for the GTEX (a) and SRA (b) datasets based on the tissue-aware gold standard. The color in each cell (row, column) represents the proportion of datasets for which the workflow along the row has a higher  $\text{log}_2(\text{auPRC}/\text{prior})$  than the workflow along the column. Comparisons that are statistically significant (corrected  $p < 0.01$ ) based on a paired Wilcoxon test are marked with an asterisk. The largest GTEX datasets (adipose\_tissue, blood, brain, and skin) are not considered in this evaluation because of the considerable amount of computing time required to use rlog transformation on large datasets. Fewer comparisons between workflows are statistically significant when evaluated on the tissue-aware gold standard, but CLR and Counts remain top performing methods for both GTEX and SRA datasets.

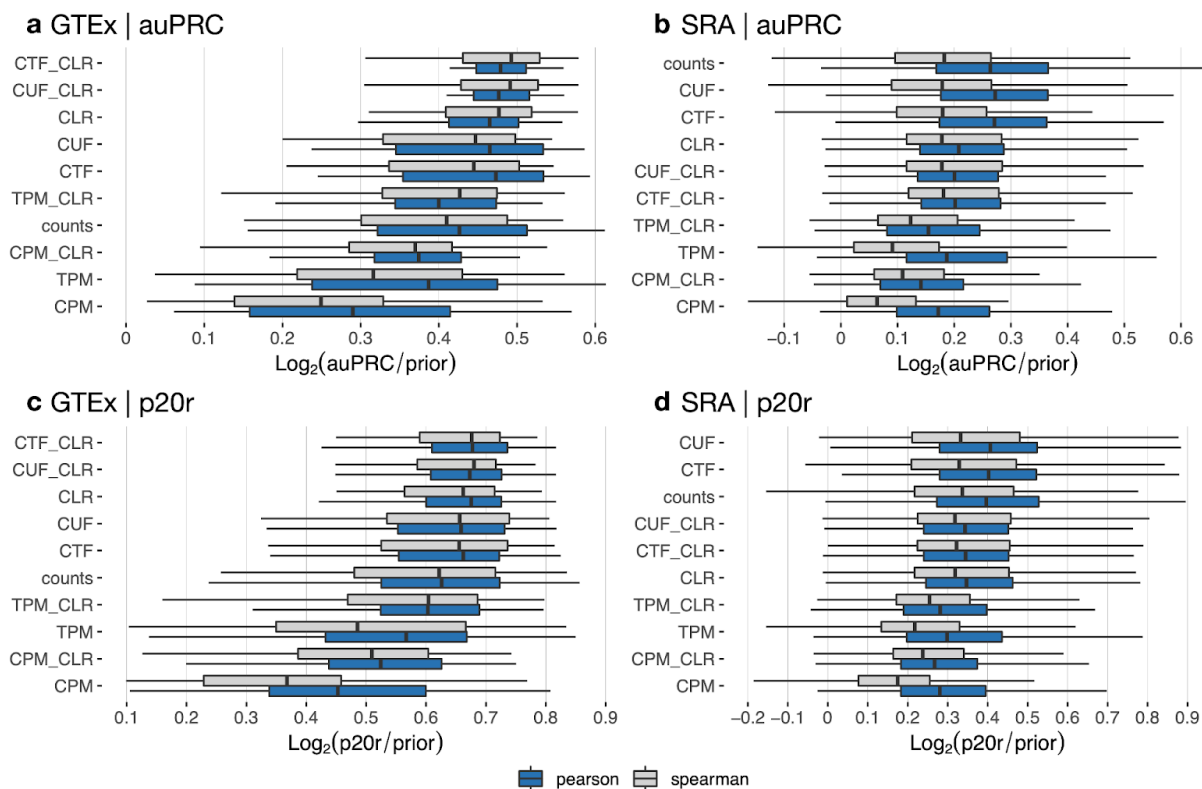




**Figure A2.14. Overall performance of workflows based on the tissue-naïve gold standard.** The plots show the aggregate accuracy of all coexpression networks resulting from each individual workflow using (a) GTEx and (b) SRA datasets, evaluated based on the tissue-naïve gold standard. The workflows (rows) are described in terms of the specific method used in the within-sample normalization, between-sample normalization, data transformation, and network transformation stages. The performance of each workflow is presented as boxplots (without outliers) that summarizes the  $\log_2(\text{auPRC}/\text{prior})$  of each workflow where auPRC is the area under the precision recall curve (see **Methods**). The workflows are ordered by their median  $\log_2(\text{auPRC}/\text{prior})$  for each panel. The six largest GTEx datasets (adipose\_tissue, blood, blood\_vessel, brain, esophagus, and skin) are not considered in this evaluation because of the considerable amount of computing time required to use rlog transformation on large datasets.



**Figure A2.15. Overall performance of workflows based on the tissue-aware gold standard.** The plots show the aggregate accuracy of all coexpression networks resulting from each individual workflow using (a) GTEX and (b) SRA datasets, evaluated based on the tissue-aware gold standard. The workflows (rows) are described in terms of the specific method used in the within-sample normalization, between-sample normalization, data transformation, and network transformation stages. The performance of each workflow is presented as boxplots (without outliers) that summarizes the  $\log_2(\text{auPRC}/\text{prior})$  of each workflow where auPRC is the area under the precision recall curve (see **Methods**). The workflows are ordered by their median  $\log_2(\text{auPRC}/\text{prior})$  in each panel.



**Figure A2.16. Overall performance of top ten workflows using Pearson and Spearman correlation based on the tissue-naïve gold standard.** The plots show the aggregate accuracy of all coexpression networks resulting from the top ten individual workflows using Pearson (blue) or Spearman (gray) correlation to build the network using (a, c) GTEx and (b, d) SRA datasets, evaluated based on the tissue-naïve gold standard. The workflows (rows) are described in terms of the specific method used in the within-sample normalization, between-sample normalization, and network transformation stages. The performance of each workflow is presented as boxplots (without outliers) that summarizes the  $\log_2(\text{auPRC}/\text{prior})$  (a, b) or the  $\log_2(\text{p20r}/\text{prior})$  (c, d) of each workflow where auPRC is the area under the precision recall curve and p20r is the precision at 20% recall (see **Methods**). The workflows are ordered by their median  $\log_2(\text{auPRC}/\text{prior})$  in each panel. Pearson correlation clearly yields better performance in all cases for the SRA data (i.e. datasets typically generated by individual research labs). Pearson also usually yields better results for the GTEx data as well, and more so when considering the accuracy of the top-scoring edges (evaluated using p20r).

## Supplemental Note

### Rationale for our functional gold standard

The definition of the “true” network structure is a crucial aspect when evaluating the accuracy of any network, including a coexpression network. Our choice and design of this ground-truth using GO biological process annotations is based on a number of factors including: A) many prior studies that link coexpression to GO co-annotation, B) the applications of coexpression networks for function prediction, and C) several previous studies that have established the strength and utility of GO-based ground-truth.

#### *A) Prior studies link coexpression to GO co-annotation.*

From the conception of high-throughput gene-expression techniques, studies have shown that coexpression between genes can be productively and accurately used to separate genes into functional modules [1, 2]. A number of other studies have explicitly tested the coexpression–co-annotation hypothesis and have shown that coexpressed genes are highly likely to be transcriptionally co-regulated and are often functionally related to each other by virtue of taking part in the same biological process or physiological trait [3, 4].

#### *B) Coexpression is commonly used to study gene function.*

Gene function prediction and gene module detection are the two major and most common applications of coexpression networks. These applications are based on the fact that functionally-related gene pairs or groups (i.e. members of a specific biological pathway or process) tend to be coexpressed with each other in high-throughput gene-expression datasets. By inverting this association, coexpression networks have often been successfully used in the literature to predict gene function and pathway membership [5]. Further, coexpression networks are frequently used to identify functional modules (i.e. entire pathways/processes) by clustering the network and performing GO-based functional enrichment on each cluster of genes [6]. Therefore, to assess the workflows examined in this study in relation to these most common applications, we chose to evaluate the accuracy of the resulting coexpression networks based on their ability to recapitulate gene functional relationships.

#### *C) Strength of GO-based ground-truth of gene functional relationships.*

Since functionally-related genes tend to be coexpressed with each other (A) and coexpression networks are routinely used to infer gene function and pathway/process membership (B), we reasoned that it would be most appropriate to evaluate the

accuracy of coexpression networks based on their ability to recapitulate gene functional relationships based on their co-annotations to GO biological processes (GOBP).

However, creating a ground-truth about gene functional relationships (gold standard) from GO BP is not straightforward and should be done very carefully. For example, the Gene Ontology has many generic terms for biological processes such as “metabolism” or “stress response”. For such terms, it would indeed be incorrect to assume that genes that are co-annotated to any of these terms should be connected in the co-expression network. Therefore, we have devised a careful procedure for constructing a gene functional gold standard based on GOBP. First, we do not use all GOBP terms to construct our gold standard. Following previous work [7], we use only 607 “specific” GOBP terms. These terms were selected by a team of seven graduate students and postdocs (with training in cell/molecular biology and genetics) based on the following procedure: To select “specific” terms, this question was considered: “if unknown gene/protein G were predicted to be annotated to GO term T, would that be enough to consider experimentally testing this relationship between G and T?”. Only terms that were declared “yes” by the majority were retained as specific terms and only gene pairs co-annotated to any of these specific terms – based on experimental evidence – are considered to have a positive relationship in our gold standard. Similarly, assuming every other gene pair is a negative (not functionally related) would be far too strong an assumption. So, the team also selected a set of 75 “intermediate” GOBP terms such as “protein folding” or “cell proliferation”. Then, to be considered as a negative in the gold standard, gene pairs must meet the following three criteria:

1. The two genes are *not* co-annotated to any intermediate term
2. The two genes are *not* co-annotated to significantly overlapping specific terms (hypergeometric test; p-value <0.05)
3. Each gene individually has at least one annotation to a specific term

Gene pairs that are co-annotated to intermediate terms (criterion 1) would be considered too close in general function to be sure that they are *not* functionally related. The hypergeometric test (criterion 2) prevents gene pairs that may share a function (due to being annotated to two overlapping terms) from being labeled as negatives. Requiring a specific term annotation for each gene ensures that no assumptions are made about genes that have not been experimentally studied before.

This procedure for creating GOBP-based gold standards of gene functional relationships has several advantages, which are outlined in detail in the paper that first used this procedure to create a functional gold standard for yeast [8]. Briefly, the advantages of this gold standard over other options include lack of substantial functional bias, lack of varying specificity problems, a thoughtful method of defining negatives, and

a more proportional ratio of positive and negative examples. Careful manual selection of specific terms covers the first two issues. Using all terms in GOBP or all pathways in a different database results in functional biases towards very large pathways which can ‘make or break’ the evaluation (see the first figure and ribosome KEGG pathway in [9]). Alternatively, defining a ‘specificity cutoff’ for these ontology structures (whether by number of annotations or depth in the ontology) results in wildly different biological specificity of terms (Figure 3, in [8]). Finally, manual selection of intermediate terms in our gold standard procedure allows negatives to be defined sensically and confidently with enough pairs to far outnumber the positive examples. This reflects the ground truth, which is that there are far more pairs of genes that do not interact with each other than gene pairs that do interact with each other.

In summary, the definition of the ground-truth network structure (i.e., the gold standard) is based on: previously established observations about the connection between coexpression and functional co-annotation; the applications of coexpression to delineate gene function; and our rigorous procedure for setting up a meaningful set of functionally-related and unrelated gene pairs based on experimental annotations of genes to specific terms in GOBP.

### **Other gold standards considered**

We spent a considerable amount of effort trying to create other gold standards without much success due to the lack of appropriate external datasets. For example, we attempted to create gold standards based on groups of genes co-bound by the same transcription factor (in ChIP-Seq experiments). However, physical binding of transcription factor does not necessarily indicate functional interaction between the transcription factor and the target gene nor does it indicate co-regulation (coexpression) between the target genes. This limitation was apparent from our observation that coexpression networks evaluated on these TF-binding-based gold standards had random performance at best and worse than random performance otherwise, regardless of workflow, sample size, or data quality.

We also attempted to create a gold standard based on groups of genes co-annotated to only tissue-specific GO biological processes. However, there was very little experimental annotation in this data to create gold standards that span tens of thousands of genes for many tissues.

At least one previous study has used spike-in data to construct ground-truth coexpression networks [9]. However, we did not choose to use spike-in data for reasons similar to the ones outlined above: limited data that prevents conducting an evaluation

on a large-enough scale to be comfortable drawing general conclusions. To our knowledge, there is not a large collection of readily-available spike-in data from multiple sources and tissues to leverage for this purpose. RNA-seq experiments are quite sensitive to technical effects when considering final quantification of each gene count in a given sample. These technical effects are a combination of the specific transcripts of interest (GC content, length, reverse transcriptase binding site sequence, etc) and the overall distribution of the population of RNA/cDNA in the sample library (whether rRNA depletion or polyA+ tail selection is used, which tissue the sample comes from, sequencing protocol, etc). All these factors can have significant effects on total read counts and, thus, gene count quantification. And, spike-in controls are not immune to these effects [10]. As we discuss throughout the paper, different normalization and network transformation techniques handle these technical biases differently, whether explicitly or implicitly. Therefore, to call a method or workflow “robust”, it must work well over a large number of datasets that encompass datasets with any number of a variety of technical biases. In our study, we check for robustness by using a large number of primary bulk RNA-seq samples (over 15,000) from over 35 tissues and 200 independent studies that were all quantified into gene counts by the same alignment software. It would be a considerable effort to collect data with spike-ins from many sources and process the raw reads into counts data for another large set of data, as we used Recount2 to take away the variability of using different alignment software. Currently, there does not seem to be a single, high-quality, large dataset for spike-in RNA-seq data that would correspond to something like the data used in our GTEx analysis where the raw reads are converted to gene counts using the same quality-control and quantification procedure (like in Recount2) to take away the variability of using different alignment software.

Furthermore, even if there is enough spike-in data available, it is not certain that it would be a useful evaluation. The purpose of spike-in experiments is often to estimate the precision and accuracy of the sequencing technology. So, typically, the concentration of the spike-in is *equal* across samples in a dataset. Even in cases where the concentration of a given spike-in probe is varied across samples, the point of these spike-ins is to find the limit of detection or to be at a detectable level so that they can be used for quantification. This means that, in many cases, we would only be able to evaluate spike-in oligos with a nominal correlation of one or zero (as was done in [9]). To be clear, this means that we would not be able to evaluate any correlation between 0 and 1. This limitation skews the assessment of workflows to an evaluation of genes that are perfectly coexpressed and highly ‘expressed’ in at least some samples. As discussed briefly in the *Discussion* section, the mean-correlation relationship bias (the observation that highly-expressed genes tend to be more highly-coexpressed in

coexpression analysis) might make this type of gold standard rather easy to achieve for all workflows. Such a gold standard does not represent a large number of genes that are never highly expressed but are nonetheless genes of great interest.

### **Evaluation procedure using our gold standard**

The gold standard contains thousands of gene pairs that either have a functional relationship (positive) or do not have a functional relationship (negative), defined based on experimental gene co-annotations to specific GOBP terms (see above). Then, we evaluate each coexpression network (derived from a single RNA-seq dataset analyzed using any one of the workflows) by comparing it to this gold standard by essentially asking the following question: do gene pairs that have very high coexpression strengths (i.e. high correlation coefficients) tend be functionally related to each other based on the gold standard?

We answer this question quantitatively by calculating the area under the precision-recall curve for that coexpression network in the following manner:

1. We rank all the gene pairs in the network from highest to lowest correlation.
2. Then, at various cutoffs of correlation strength from high to low, we calculate the number of true positives, false positives, true negatives, and false negatives.
  - a. Gene pairs with a correlation value above the cutoff and,
    - i. Functionally related in the gold standard (i.e., positive) are ‘true positives’ (TP).
    - ii. Not functionally related in the gold standard (i.e., negative) are ‘false positives’ (FP).
  - b. Gene pairs with a correlation value below the cutoff and,
    - i. Functionally related in the gold standard (i.e., positive) are ‘false negatives’ (FN).
    - ii. Not functionally related in the gold standard (i.e., negative) are ‘true negatives’ (TN).
  - c. These TP, FP, FN, and TN values are combined to calculate the precision ( $= TP / (TP + FP)$ ) and recall ( $= TP / (TP + FN)$ ) at that cutoff.
3. All the precision and recall values at the various correlation cutoffs are used together to build the precision-recall curve.
4. Finally, the area under this curve (auPRC) and the precision that corresponds to 20% recall (p20r) are used to quantify the ability of the coexpression network to recapitulate gene functional relationships in the gold standard.



## REFERENCES

1. Eisen MB, Spellman PT, Brown PO, and Botstein D. Cluster analysis and display of genome-wide expression patterns. PNAS 1998.
2. Segal E, Friedman N, Koller D, Regev. A module map showing conditional activity of expression modules in cancer. Nature Genetics 2004.
3. Allocco DJ, Kohane IS, and Butte AJ. Quantifying the relationship between co-expression, co-regulation and gene function. BMC Bioinformatics 2004.
4. Carpenter AE and Sabatini DM. Systematic genome-wide screens of gene function. Nature reviews genetics 2004.
5. Zhu Q, Wong AK, Krishnan A, Aure MR, Tadych A, Zhang R, Corney DC, Greene CS, Bongo LA, Kristensen VN, Charikar M, Li K and Troyanskaya OG. Targeted exploration and analysis of large cross-platform human transcriptomic compendia. Nature Methods 2015.
6. Zhang B, Horvath S. A General Framework for Weighted Gene Co-Expression Network Analysis. Stat Appl Genet Mol Biol. De Gruyter; 2005;4.
7. Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, Zhang R, Hartmann BM, Zaslavsky E, Sealfon SC, Chasman DI, FitzGerald GA, Dolinski K, Grosser T and Troyanskaya OG. Understanding multicellular function and disease with human tissue-specific networks. Nature Genetics 2015.
8. Myers CL, Barrett DR, Hibbs MA, Huttenhower C and Troyanskaya OG. Finding function: evaluation methods for functional genomic data. BMC Genomics 2006.
9. McCall MN and Almudevar A. Affymetrix GeneChip microarray preprocessing for multivariate analyses. Briefings in Bioinformatics 2012.
10. Qing T, Yu Y, Du T T, et al. mRNA enrichment protocols determine the quantification characteristics of external RNA spike-in controls in RNA-Seq studies. Sci China Life Sci, 2013, 56: 134–142.

## **CHAPTER 3: LEVERAGING PUBLIC TRANSCRIPTOME DATA WITH MACHINE LEARNING TO INFER PAN-BODY AGE- AND SEX-SPECIFIC MOLECULAR PHENOMENA**

### **Background**

Most complex traits and diseases have age- and sex-related differences in their incidence and manifestation [1] and yet these factors have been largely underconsidered in biomedical and clinical studies in the past [2–5]. Further, often these age- and sex-related differences are intertwined and considering one without the other will produce an incomplete understanding. For example, women have a lower prevalence of stroke before menopause, but afterwards the prevalence exceeds that of men [6]. Similarly, the peak of asthma diagnoses is between the ages of 2 and 8 years old in boys, but incidence is higher for women in adults [7]. As scientific research begins to pay closer attention to age and sex as biological factors, new studies are now beginning to uncover some of the genetic basis that underlies the processes of development and aging with or without considering sex differences in tissue physiology [8], complex traits [9], diseases [10], and treatment responses [11].

These new data alone are not enough to create holistic frameworks capable of helping biologists address questions about female and male tissue biology at specific intervals along the human lifespan (e.g., childhood, adolescence, or old age). As we aim to provide precision medicine for all, a comprehensive understanding of how age and sex influence normal physiology and in turn affect complex traits and diseases is critical [12,13]. Related differences can be quite small, producing subtle, easily-overlooked changes. For example, significant sex differentially expressed genes often have small

fold changes [12] and genes related to human longevity by GWAS have small effect sizes [14].

An opportunity to investigate these small-yet-widespread influences resides in the hundreds of thousands of publicly-available gene expression profiles generated by hundreds of labs across the world over the past 25 years and stored in databases such as NCBI GEO [15] and EBI ArrayExpress [16,17]. These transcriptomes span multiple tissues, diverse experimental, biomedical, and environmental conditions, and numerous diseases, and have been previously successfully leveraged towards gaining biological insight into molecular mechanisms of complex traits and diseases [18,19]. A few previous studies have used parts of these data to identify sex-associated genes and molecular processes with occasional minor focus on age.

One of the first large-scale endeavors to characterize human sex-biased genes was a 2016 study wherein Mayne and colleagues [20] used differential expression analysis on 22 publicly-available microarray datasets totaling about 2,500 samples from 15 tissues (**Fig. 3.1b**). Previous to that study, one of the largest sex-differential expression studies published was the 2015 study [21] from the GTEx consortium [22] which included sex as one of many biological factors considered in the expression variation between individuals. At the time, only the pilot data had been released, which included 1,641 RNA-seq samples from over 40 tissues in 175 individuals. GTEx consortium data has since grown to over 17,000 samples from 948 individuals and has been used in another handful of sex-differential expression studies. Guo and team [23] used GTEx data in addition to curated datasets from GEO and restricted themselves to healthy samples to determine sex-biased genes in 14 tissues. Gershoni and Pietrokovski [24] published

sex-differential expression results using version 6 of GTEx in 2017, and in 2019 the sex-associated gene database (SAGD) resource was published with sex-associated genes (differential expression) from 2,828 samples in 21 different species. Later that year Naqvi et al [25] used GTEx data in conjunction with data from macaque, mouse, rat, and dog to investigate conserved sex-biased expression in 12 tissues. In 2020, Lopes-Ramos and group [26] used the GTEx dataset for sex differential expression analysis and further built regulatory networks for each sample and compared female and male networks in each tissue. Finally, the GTEx consortium released a paper focused entirely on the impact of sex on gene expression in 2020. Each of these studies used publicly-available transcriptomic data, often including GTEx data, and stratified samples by tissue to find differentially expressed genes between sexes. GTEx data skews about  $\frac{2}{3}$  male, and the Mayne study had a similar ratio of male to female samples. The Guo study, SAGD, and this study are close to 50-50 sex balance. Most studies did not explicitly consider age, but the SAGD considers developmental stage when possible, and the GTEx studies incorporate age as a covariate but is focused on a dataset of mostly adult and older individuals.

Efforts to characterize age-biased genes using human transcriptomic data are also generally stratified by tissue though sometimes commonalities across tissues are investigated. One of the earliest large-scale efforts to study common aging signatures in public expression profiles was an differential expression analysis with 27 microarray datasets from mice, rats, and humans that spanned multiple tissues in 400 samples by Pedro de Magalhães and colleagues [27] in 2007. A few years later Hannum et al [28] used whole blood gene expression data of 488 individuals ranging from 20 to 75 years

of age. A large amount of focus in studying age-biased gene expression has continued to center around age prediction [28–33] and the process of aging [27,28,32,34,35] and/or development [36] through differential expression. Two studies While sometimes sex is adjusted for in the model, very little has been done to study age in a sex-dependent manner, especially that does not focus on aging specifically, though there is evidence that the processes of aging [32,37] and development [38] carry sex-dependent differences.

These important efforts have begun to characterize tissue-specific expression in each sex and in different age groups. However, there are still gaps that need to be addressed. First, most of these studies using public gene expression data are focused on age or sex, sometimes adjusting for the other, but rarely trying to delineate age and sex differences at the same time. Second, the data used in these studies (often from GTEx) skew heavily towards adults and older individuals. Third, investigating both age across all stages of life and sex specificity at the same time on a large scale has yet to be done. We are interested in not just the process of aging, but molecular processes that occur at specific stages along the human lifespan. Integration of a large body of data allows us to pick up on multi-tissue signals without only learning dataset- or tissue-specific signals. The main hurdle to leveraging the hundreds of thousands of publicly-available gene expression profiles towards this purpose is that age and sex metadata is often missing, inconsistent, or disorganized. Especially because age and sex have been historically understudied, the vast majority of these samples are not associated with any age and sex information. Sample descriptions that do contain this information often have it buried in free text and many are annotated with vague labels

that are minimally informative and imprecisely defined as, for e.g., ‘old’, ‘adult’, or ‘infant’ (i.e., without the associated age ranges), making it difficult for researchers wishing to reanalyze these datasets.

In this study, we present the largest effort to characterize age- and sex-biased genes across the entire human lifespan from public transcriptomes. First, we manually curated the largest sex- and age-annotated public transcriptome dataset containing nearly 30,000 bulk, primary human microarray and RNA-seq samples from variety of tissues. Second, to infer pan-body molecular processes impacted by age and sex, we used these transcriptomes and their labels to calculate age-stratified sex-biased gene signatures and sex-stratified age-group-specific gene signatures. Third, using existing gene annotation and association data in various databases, we associated all these age/sex gene signatures to hundreds of biomedical entities including biological processes/pathways, phenotypes, traits, and diseases. We make our sample labels, gene signatures, and associated biomedical entities available in a GitHub repository. These resources will enable scientists in studying sex-specific health and disease mechanisms in all stages of life.

## **Results**

### **Curating a large dataset of human age- and sex-annotated transcriptomes**

To characterize age- and sex-specific gene expression signatures, we first downloaded all available human microarray data in Gene Expression Omnibus (GEO) [39] measured on the same platform and all human RNA-seq data available in refine.bio [40]. We used simple text matching in downloaded sample descriptions from GEO and the Sequence Read Archive (SRA) [41] as well as labels from metaSRA to create a set of

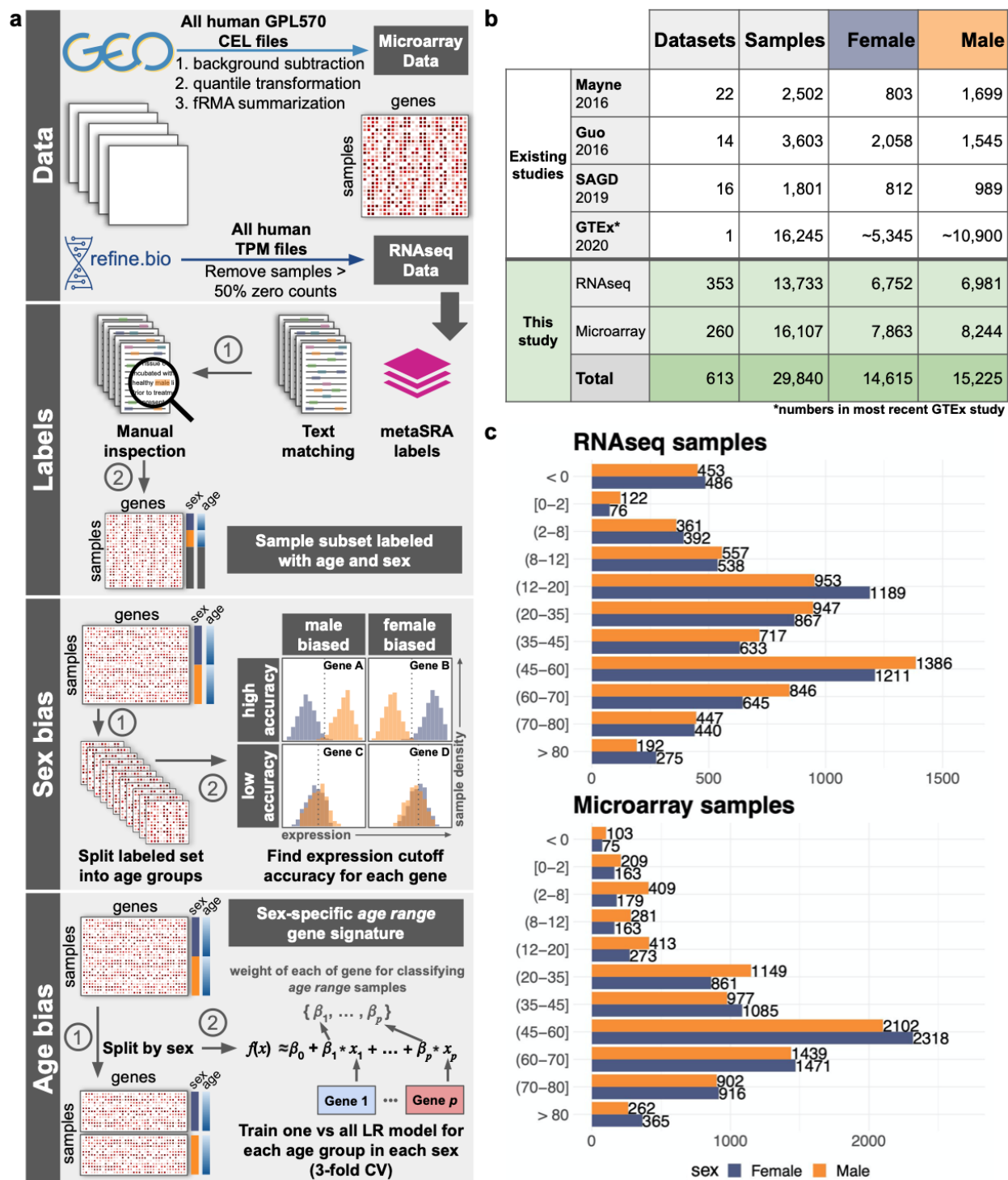
transcriptomes associated with age and sex information. We read the sample and experiment descriptions of this entire set of samples to ensure the accuracy of the age and sex labels, as well as to remove any samples that were not bulk, primary human samples.

Our goal with the manual curation step was to assemble a set of samples that would reflect age- and sex-specific molecular mechanisms as faithfully as possible. In addition to improperly assigned age and sex labels, we removed any single cell or single nuclei data, xenografts, microbiome samples, pooled samples, and cell lines. Cell lines were removed due to the tendency of many lines to lose their Y or inactive X chromosome [42], the variability in the ability of cell lines to represent biology of primary cells [43–45], and contamination issues [46]. Our final set of samples is divided into age groups based on sex hormone levels in each sex across all ages [1]. This set has a larger number of samples and datasets in the middle age groups than the youngest and oldest ranges (**Fig. 3.1c**, **Fig. A3.1**) and overall seem to be biased towards samples from blood, brain, small intestine, liver, and lung tissues (**Fig. A3.2**).

Although tissue bias should be considered, it is worth noting that even when age- or sex-biased genes are determined by restricting samples to a single tissue, cell type composition has a significant effect and will alter results if not controlled for. For example, a handful of studies previously reported breast as the most sex-differentiated tissue [21,24,26], but the most recent GTEx consortium study [47] did not observe this result after controlling for cell type composition. Pellegrino-Coppola and colleagues recently reported a similar importance for cell type correction when determining genes associated with aging from gene expression data [35]. On the task of age prediction,

Wang and team found that combining expression data from two tissues reduced the margin of error when predicting the age of GTEx (mostly adult) samples compared to using expression from one tissue [48]. Nonetheless, to investigate the effect of separating out samples by tissue, we repeated some analyses by restricting them to samples from blood only, as it was the most common tissue labeled in our set. However, not enough blood samples were annotated to the fetal ( $< 0$ ) age group or the oldest age group to include them in the blood-only analyses.





**Figure 3.1. Workflow and data.** (a) Data was obtained from Gene Expression Omnibus and refine.bio for microarray and RNA-seq, respectively. A combination of text matching and metaSRA labels were used to create a first draft for sex and age labels. These labels were then manually inspected to ensure they were correct and keep only primary human samples. Finally, the labeled data was used to assess sex and age

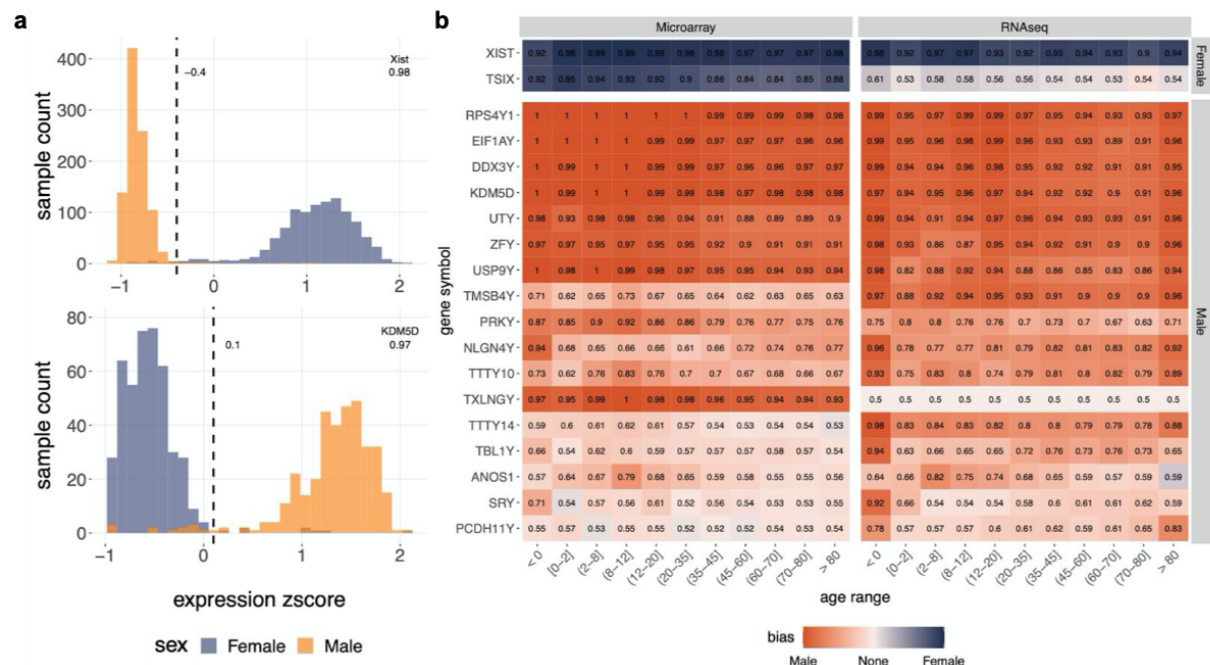
### Figure 3.1. (cont'd)

bias in the large, publicly available set. **(b)** The table contains the number of datasets and samples used in some of the largest differential expression studies across sex published in recent years as well as the number of datasets and samples used in this study. The previous studies used only RNA-seq data, while this study uses both microarray and RNA-seq. **(c)** The bar plots show the number of samples used in this study separated by age group (y axis) and sex (bar color).

### Age-stratified sex-biased genes

We first used our age- and sex-labelled transcriptome dataset to determine age-stratified sex-biased genes. Independently in microarray and RNA-seq data, we converted the expression of each gene across all samples into z-scores and, for samples within each age group, stepped through the distribution at fixed intervals to find the best expression threshold for that gene to separate ‘Female’ and ‘Male’ samples. Balanced accuracy was used as a metric to determine how well-separated samples were based on each gene’s expression while recording whether the expression was higher in the Female or Male samples (**Fig. 3.2a**; see **Methods**). Only 19 genes had a balanced accuracy  $\geq 0.8$  in at least one age group in either microarray or RNA-seq data and all these genes were on the X or Y chromosome (**Fig. 3.2b**). As sex differences tend to be quite small and often tissue-specific [12], it is not surprising that all of the strongest sex-biased genes reside in the sex chromosomes. The only Female-biased genes in this set include *XIST*, the major effector of X chromosome inactivation, and *TSIX*, the antisense RNA for *XIST*. The most surprising result is that one gene in this set, *ANOS1*, is both varyingly Male-biased across almost all age groups and is on the X chromosome. Mutations in *ANOS1* have been associated with hypogonadism and Kallmann Syndrome in men [49].

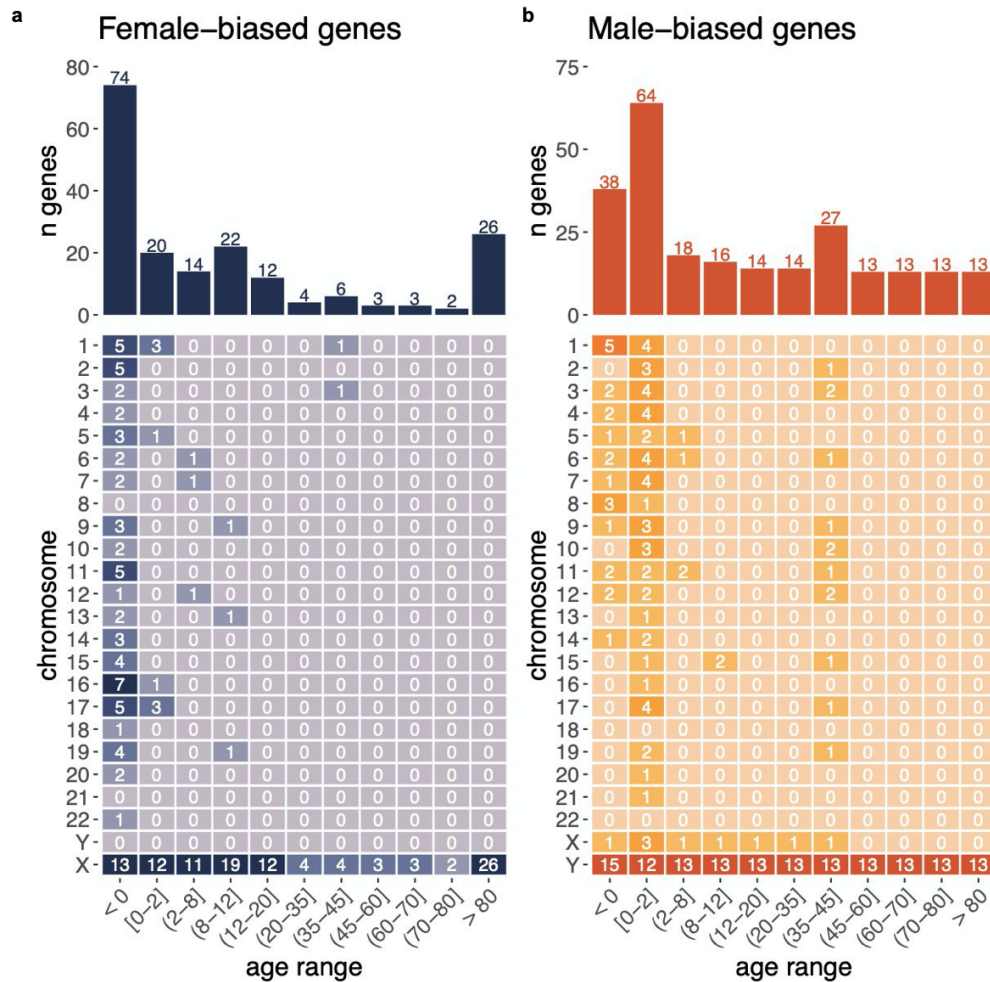
To ensure that the major signals captured here are not because of tissue-bias across females and males, we repeated this analysis using only samples that came from blood in all but the youngest and oldest age groups due to a low number of samples. Excluding genes reaching the 0.8 threshold only in these age groups from the whole sample set, the blood-only analysis recapitulated all strongly sex-biased genes except *NLGN4Y*, a membrane protein in the neuroligin family (**Fig. A3.3**). Other sex-biased genes found in blood samples were restricted to a single age group each.



**Figure 3.2. Sex bias of gene expression across age groups.** (a) Distribution of male and female samples with a given z-scored expression value of *Xist* (top) and *KDM5D* (bottom) in different age groups. (b) The heatmap displays all genes (x axis) that had a balanced accuracy of at least 0.8 in any age group (y axis) when separating male and female microarray or RNA-seq samples.

We then examined genes with weaker sex bias, but reasonably consistent across technologies. These were genes biased in the same direction in both RNA-seq and microarray data, but only reached a balanced accuracy of 0.65 in one of the

technologies (**Fig. 3.3**). No Y chromosome genes that did not reach a balanced accuracy of 0.8 in at least one age group were added in this set, but several more genes from the X chromosome were added, including 4 more genes that were Male-biased (*DUSP9*, *CD99*, *THOC2*, *SMARCA1*) in some of the younger age groups. The only Female-biased X chromosome gene common across all age groups was *KDM6A*, a lysine demethylase. The next most common Female-biased X-chromosome genes across age groups were *EIF1AX*, *PUDP*, and *ZFX*. These were all common to the seven youngest age groups along with the oldest age group. Interestingly, general sex-biased expression seems to taper off as age increases (**Fig. 3.3**), but the oldest age group contained a particularly high number of Female-biased genes (26), all on the X chromosome. These genes were not enriched for any particular function. The youngest age groups also show the most sex bias in autosomal genes; the four oldest age groups showed no sex-biased autosomal genes.

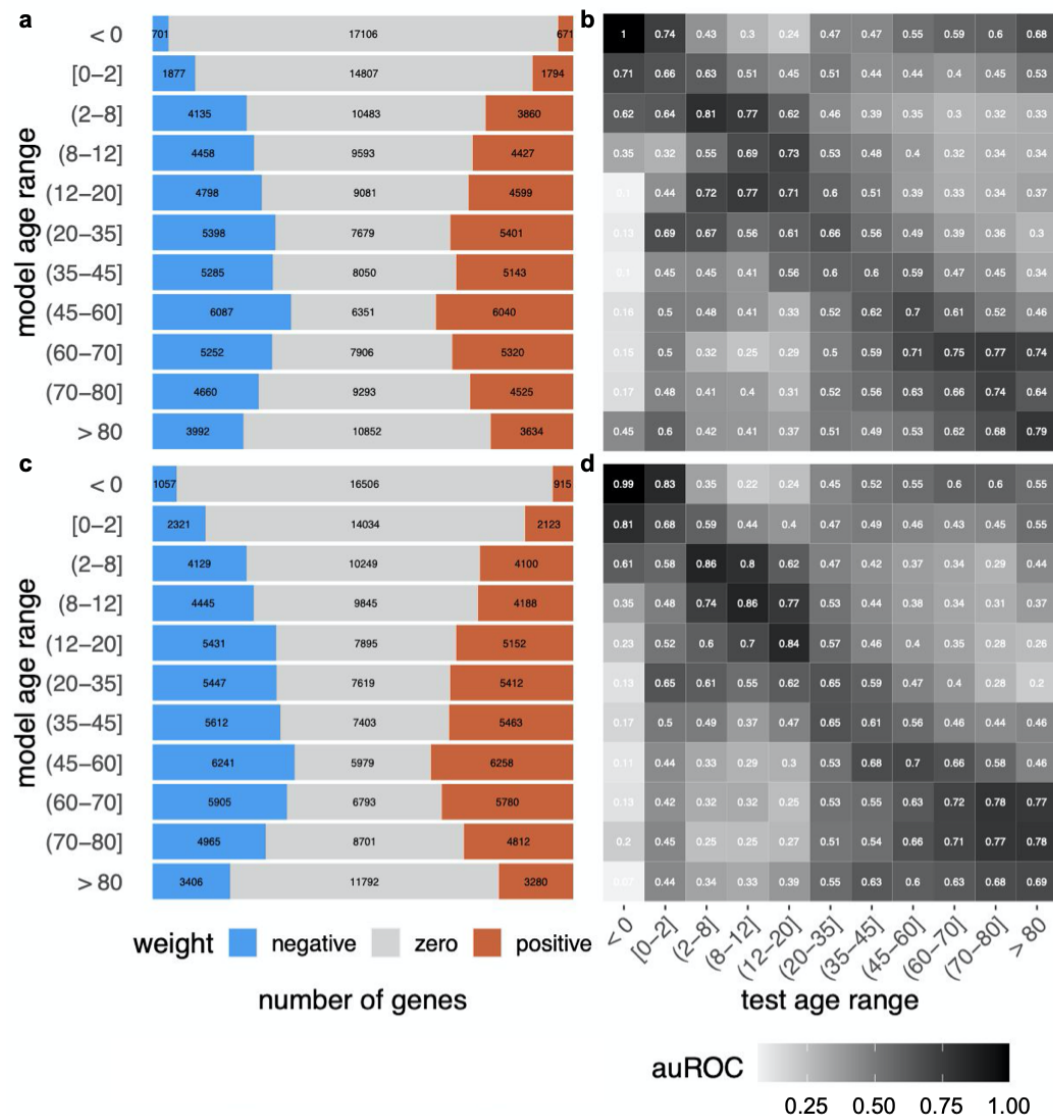


**Figure 3.3. Number of sex-biased genes across age groups.** The barplot shows the total number of (a) Female-biased and (c) Male-biased genes per age group. The tables below display the number of (b) Female-biased and (d) Male-biased genes on each chromosome per age group.

### Age group prediction stratified by sex

Next, for each sex, we used our dataset to scan for genes that were able to differentiate between age groups to find sex-stratified age-biased genes. We set this task up as a supervised machine learning problem and trained logistic models to distinguish one age group from all others, independently for each sex, in RNA-seq (**Fig. 3.4**) and microarray (**Fig. 3.5**) data separately. We used an elastic-net penalty to encourage sparsity while balancing the contributions of correlated genes. Entire datasets were always kept within

a fold to avoid rewarding the model for learning study-specific signals (**Fig. A3.4**). Across all age groups, ML models based on RNA-seq use less genes to predict sample age group (**Fig. 3.4a,c, Fig. 3.5a,c**). In both technologies, the middle age groups are harder to separate from other age groups using gene expression, as evidenced by the higher number of genes required by the model and slightly lower performance compared to the very youngest and oldest age group models (**Fig. 3.4 and 3.5**).



**Figure 3.4. Size and performance of RNA-seq age group prediction models.** The stacked barplots show the distribution of positive, zero, and negative weights for the model with the median number of

**Figure 3.4. (cont'd)**

positives across the three folds for each age group in RNA-seq for (a) Females and (c) Males. The heatmaps contain average auROC of RNA-seq models trained on the age group labeled in the rows when evaluated using the samples in the age group labeled in the column as positive examples for (b) Females and (d) Males.

In addition to measuring the performance of each model on the age group it was trained to classify, we also evaluated it on samples from each of the other age groups (**Fig. 3.4b,d, Fig. 3.5b,d**). The heatmaps in **Figure 3.4** and **Figure 3.5** contain the average area under the Receiver Operator Characteristic curve (auROC) across 3 folds for each age group model trained on a specific age group (rows) and evaluated as if the age group in the column were the positive examples. We chose to display auROC as it can be easily interpreted as a probability. Thus, the value in each cell of a heatmap is the probability that a randomly selected sample from the test age group (column) would be ranked higher than a randomly selected sample from any other age group by the age group model in the row. However, auROC is not the best measure of performance when there is high imbalance of positive and negative examples, as we have in this age-group classification task. So, we also include heatmaps of the performance measured with  $\log_2(\text{auPRC}/\text{prior})$  in the supplement (**Fig. A3.5-8**). This metric accounts for class imbalance and emphasizes the accuracy of top-ranked positive samples. Nevertheless, evaluation results with this metric largely agrees with those shown by auROC.

Overall, the classifiers show good performance across age groups in both RNA-seq and microarray samples, with more difficulty in the middle age groups (**Fig. 3.4b,d, Fig. 3.5b,d, Fig. A3.5-8**). We also see poorer performance in the [0-2] age group RNA-seq models, where we have one of the lowest number of positive examples (**Fig. 3.4b,d,**

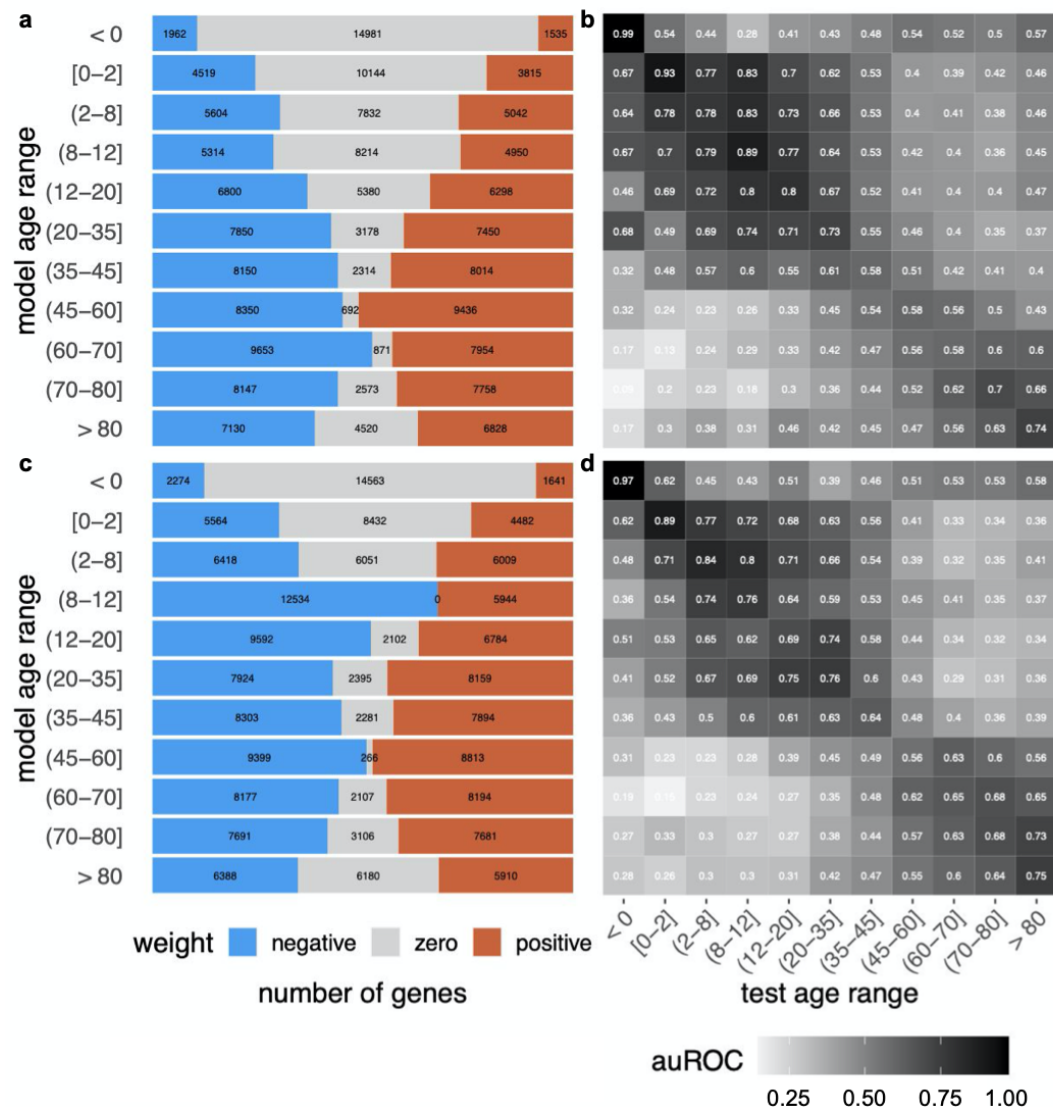


**Fig. 3.1c**). In general, however, the number of training positives does not correlate with higher performance (**Fig. A3.9**). For instance, the 45–60 age group in both sexes has the worst prediction performance with microarray samples despite having the highest number of samples. (**Fig. 3.5b,d, Fig. 3.1c**)

We repeated this analysis using only blood samples for training and evaluation using three test sets, but reusing some training data in all folds because the lower number of samples made it impossible to perform a strict three-fold cross validation for some age groups. Even with this adjustment, we still were not able to include the youngest and oldest age groups (**Fig. 3.10**). Overall, the blood-only models used less genes to classify samples (more genes had zero weight) than the models using all tissues, but had poorer prediction performance in most age groups (**Fig. A3.11-16**). This result suggests that including data from multiple tissues may improve the age signal-to-noise ratio (see *Discussion*).

To check consistency between folds and similarity between models in different age groups in both sexes and technologies, we calculated the cosine similarity of the model weights across all genes (**Fig. A3.17-21**). As expected, invariably, models trained on the same age group, sex and technology are more similar to each other than to models that differ in any of those factors. For the youngest age groups, especially fetus, we observe the models to be similar even across sex and technologies. This observation combined with the high performance of these models indicates that fetal gene expression is robust and very distinct from all other age groups. Conversely, most sex-stratified age group models are not similar across RNA-seq and microarray, indicating a substantial technology effect.

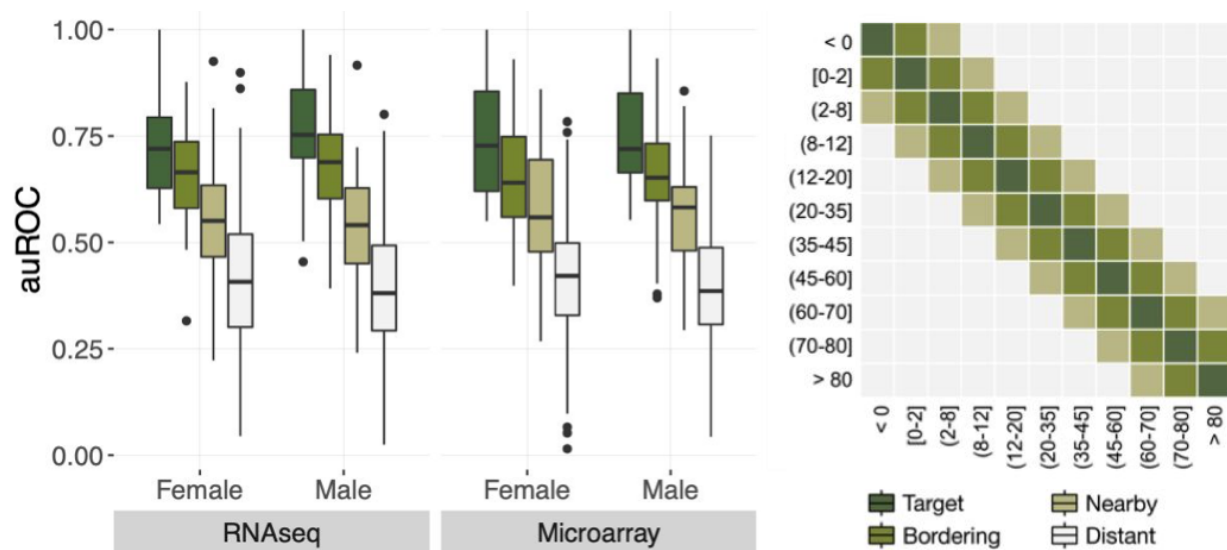




**Figure 3.5. Size and performance of microarray age group prediction models.** The stacked barplots show the distribution of positive, zero, and negative weights for the model with the median number of positives across the three folds for each age group in microarray for (a) Females and (c) Males. The heatmaps contain the average auROC of microarray models trained on the age group labeled in the rows when evaluated using the samples in the age group labeled in the column as positive examples for (b) Females and (d) Males.

Finally, the cross-age-group evaluations (**Fig. 3.4** and **3.5**) also demonstrate that the age group models capture the chronological relationships between age groups. We quantified this pattern more precisely by dividing all age groups into four categories with

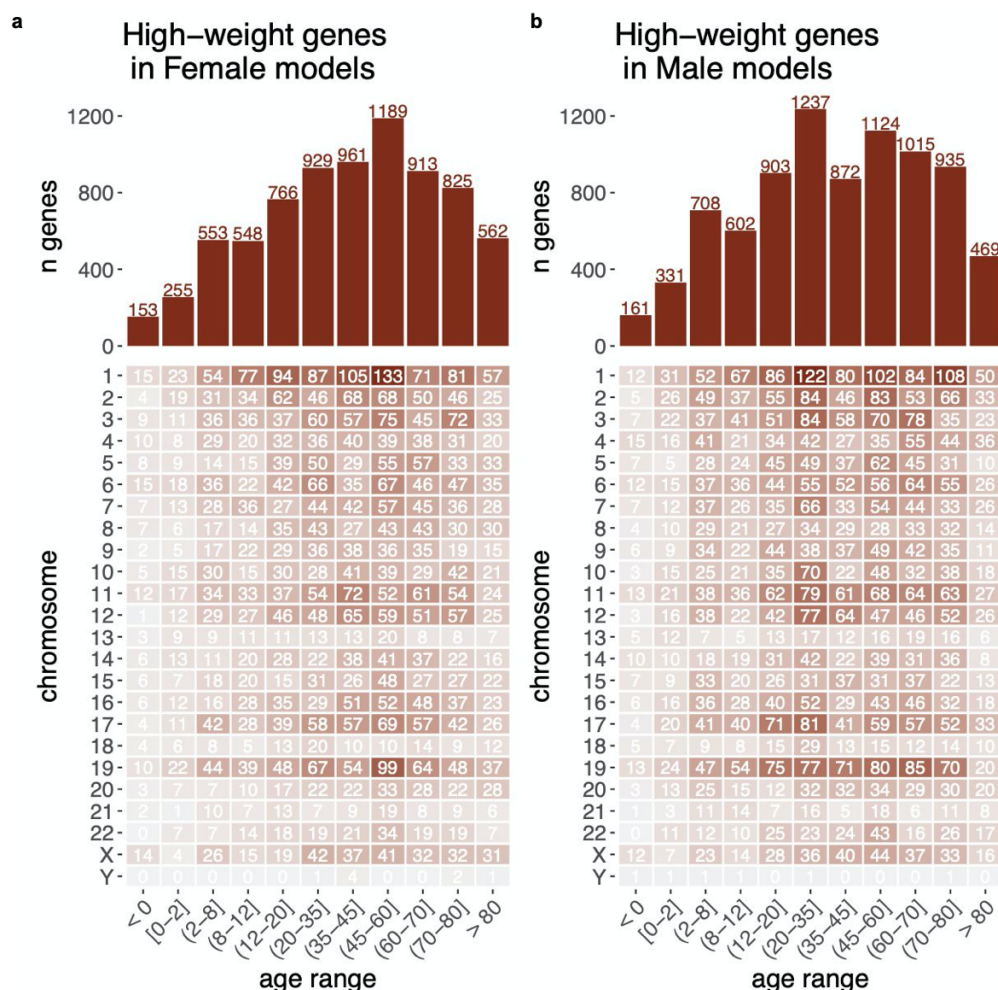
reference to each model. The ‘target’ age group is the one that the model was trained on, ‘bordering’ age groups are those directly before or after the target, and ‘nearby’ age groups are those that are one-removed from the target. Every other age group is marked as ‘distant’. As we train a one-vs-rest model for each age group, the model is designed to separate target (positive) samples from non-target (negative) samples and does not have any external information about the chronological relationships between age groups. Despite this setting, we observe that age group models assign higher probabilities to samples from neighboring age groups and lower probabilities to those from distant age groups (**Fig. 3.6**). This trend holds in both sexes for RNA-seq and microarray models and is strong evidence for our models capturing biologically-relevant age signals. A similar trend can be observed in the blood age group model performances, but it is not quite as strong, especially in Females (**Fig. A3.22**)



**Figure 3.6. Performance of models when evaluated on near and distant age groups.** Boxplot of auROC of all RNA-seq and microarray Female and Male models when considering target, bordering, nearby, and distant age groups as positive examples (key on right).

### **Sex-stratified age-biased genes**

We defined sex-stratified age-biased genes per age group by choosing genes that were assigned a positive weight in the corresponding model in at least five folds across the six folds between the microarray and RNA-seq models, with a non-negative weight in the remaining fold, if any. The middle age groups have a higher number of age-biased genes than other age groups (**Fig. 3.7, (Fig. 3.4a,c, Fig. 3.5a,c)**). Across all age groups in both sexes, the number of age-biased genes from each chromosome tends to correlate with the total number of genes on the chromosome. Similar trends are present in negatively-weighted genes from models in each sex (**Fig. A3.23**). In total, across all age groups, 6,488 genes are age-biased in Female models and 6,975 genes are age-biased in Male models, but only 2,838 of those genes are common between them. The vast majority of these age-biased genes are biased in only one age group per sex (5,447 in Females; 5,734 in Males). In each sex, about 1,000 genes are age-biased in two age groups, around 100 in three age groups, and about 10 in four or five age groups. Taken together with the good performance of our prediction models, these results suggest that each age groups has a distinct expression signature and that development and aging processes have sex-specific differences detectable in large datasets.



**Figure 3.7. Number of age-biased genes across age groups in each sex.** The table displays the number of age-biased genes on each chromosome per age group and the barplot shows the total number of age-biased genes per age group in (a) Females and (b) Males.

#### Enriched experimental genesets in age-stratified sex signatures

All the analyses presented above have resulted in several genome-wide gene signatures associated with sex and age. To enable us and other researchers efficiently interpret (and search) these signatures in relation to various biological contexts, we associated these signatures to coherent genesets annotated to biological processes [50], traits [51], diseases [52,53], phenotypes [54], and cell types [55]. First, we defined age-stratified sex gene signatures as genes across the genome along with their signed

normalized balanced accuracy scores, which indicates the extent to which each gene was able to separate Female from Male samples (or vice versa) in a particular age group (See **Methods**). We then used a permutation test to estimate the strength and direction of association (*i.e.*, ‘enrichment’) of each geneset with each signature (see **Methods**).

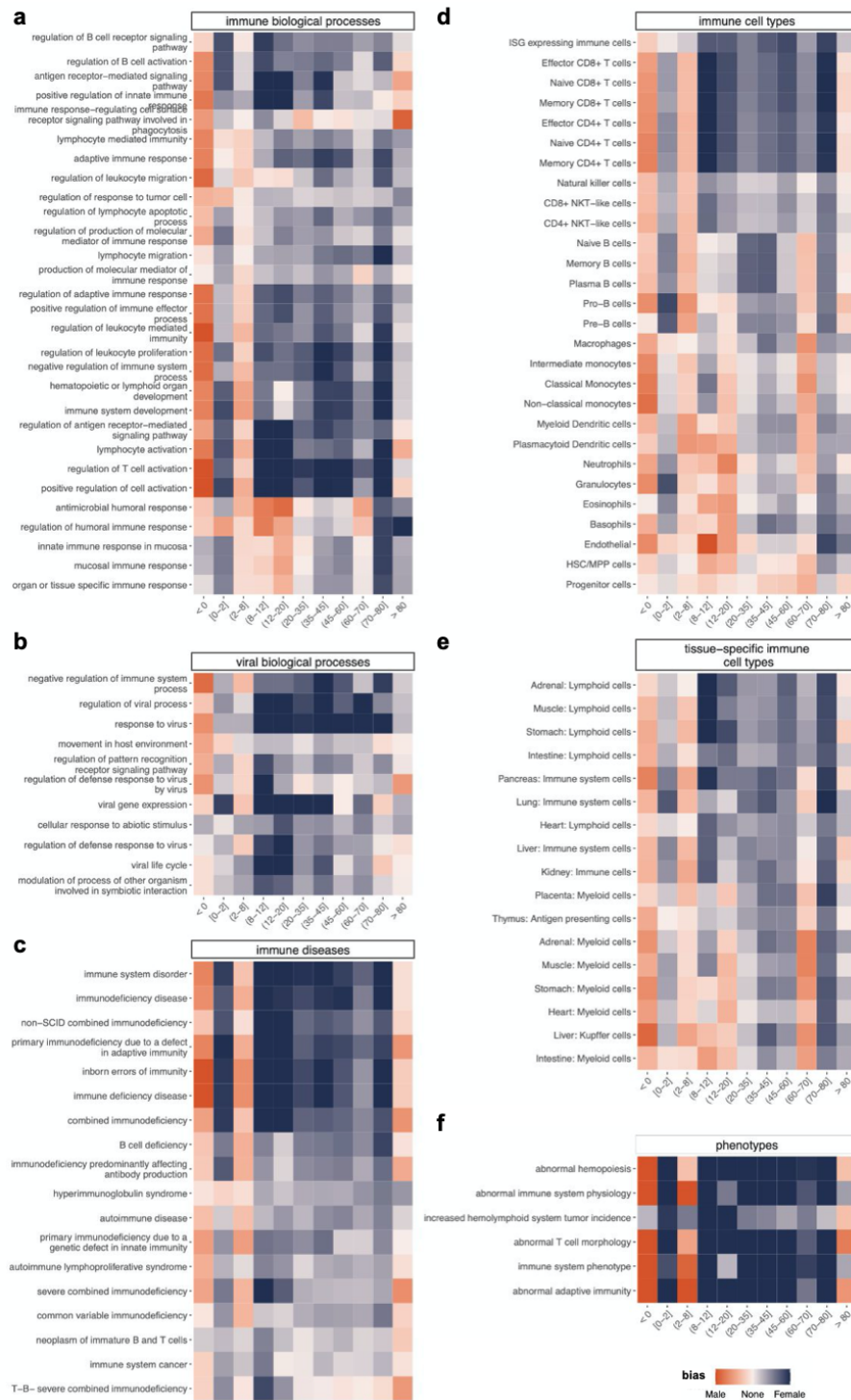
We applied this approach to compare the sex gene signatures calculated in our study and those from previous studies — Guo et al (2016) [23], SAGD (2019) [56], and GTEx (2020) [47] — estimated using differential expression analysis in multiple tissues. Overall, the agreement between studies is very high. The only prominent disagreement occurs with GTEx signatures, especially in our [0-2] sex signature (**Fig. A3.24**). However, there are no samples from children in GTEx, which makes it unsuitable for capturing what sex-biased expression should look like in children no older than 2. All the other partial and minor disagreements are likely point to sex biases in specific age groups found in our analysis that were not seen in previous analyses that did not stratify data by age. However, the allround agreement suggests that gene signatures we estimate reflect patterns of sex differences in expression across various tissues.

Application of the geneset enrichment approach to genesets from various sources resulted in hundreds of biological contexts associated with sex across age group. Notable among them is the pan-body phenomenon of immune response that plays a central role in autoimmune diseases more prevalent in females [57,58]. We found the large majority of immune response-related (**Fig. 3.8a**) and viral-related (**Fig. 3.8b**) biological processes to be Female-biased across most age groups. Notable exceptions

include the  $< 0$  and (2-8] age groups. A very similar trend can be observed with the enrichment of immune diseases (**Fig. 3.8c**) and immune phenotypes (**Fig. 3.8g**).

Consistent with previous studies [59,60], the marker genes of B cells, T cells, and lymphoid cells tend to be Female-biased in our sex signatures (**Fig. 3.8d, e**), again with the exception of the  $< 0$  and (2-8] age groups, and the (60-70] age group for B cells. We also observe that myeloid cells tend to have Male-biased enrichment in the younger age groups (consistent with a previous study [60]) and Female-biased enrichment in older age groups.

Further, we find many metabolic processes to be Male-biased in our age-stratified sex signatures, consistent with findings from a previous study [23] (**Fig. A3.25**). Together, these enrichment results indicate the potential utility of our signatures to investigate sex differences of multi-tissue processes such as immune response and metabolic processes, along with corresponding disease mechanisms in different stages of the human lifespan.



**Figure 3.8. Enrichment of experimentally-derived genes sets from in our age-stratified sex signatures.** Female- and Male-bias enrichment of experimentally-derived gene sets. Heatmaps show enrichment scores for **(a)** a representative (See **Methods**) set of immune-related GO biological processes

**Figure 3.8. (cont'd)**

(b) a representative (See **Methods**) set of viral-related GO biological processes, (c) immune-related diseases, (d) immune cell type marker genes, (e) tissue-specific immune cell type marker genes, (f) Schwann cell marker genes, and (g) immune-related phenotypes.

Enriched experimental genesets in sex-stratified age group signatures

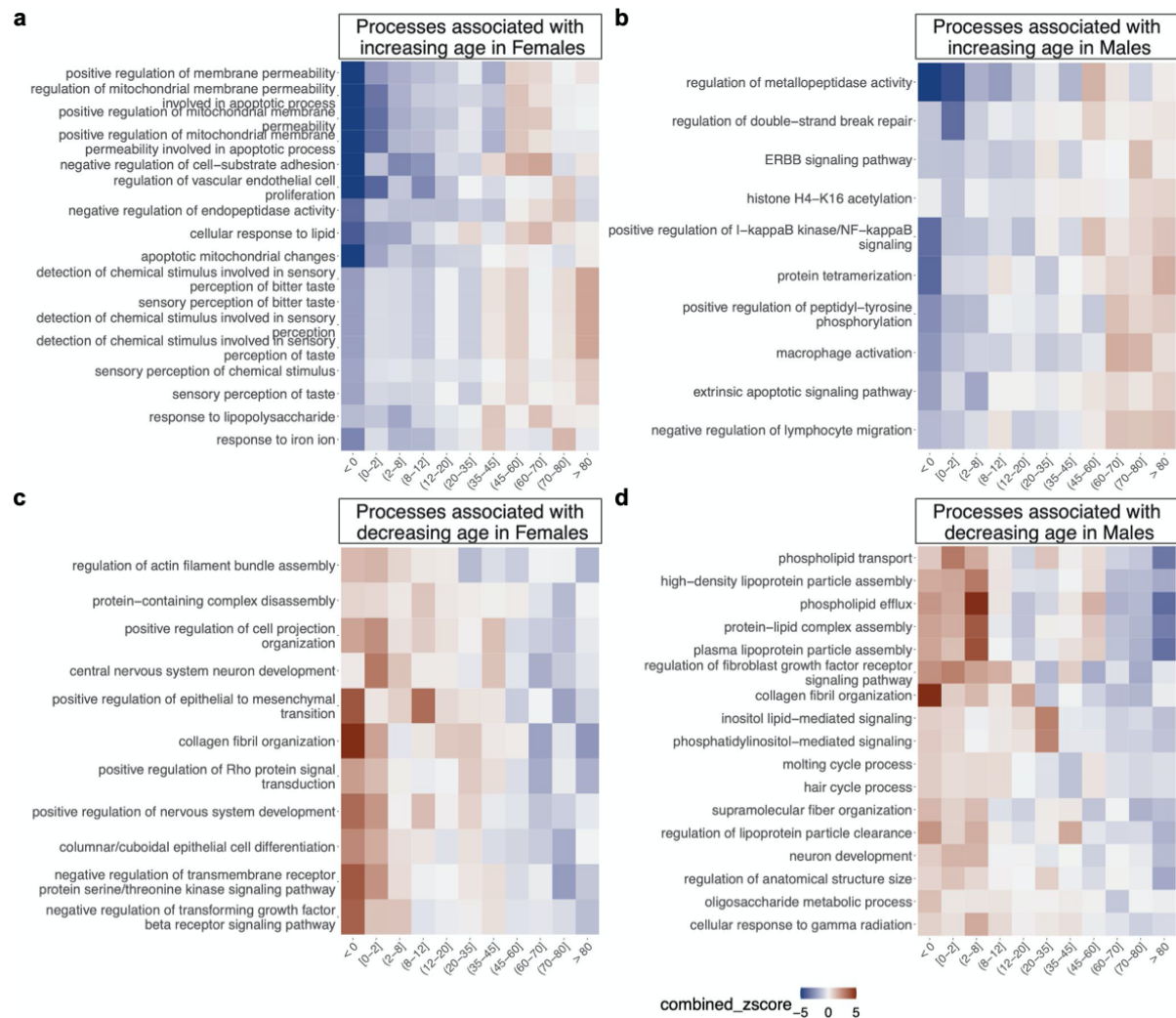
In addition to being valuable for making age group prediction of new samples, our age group models can be biologically interpreted. To do so, we first defined sex-stratified age-group gene signatures as genes across the genome along with their model coefficients. Here, the genes with positive and negative coefficients in a model correspond to genes whose relative high and low expression, respectively, are characteristic to the pertinent age group (see **Methods**). Then, we used the same permutation-based geneset enrichment strategy as above to associate hundreds of biological contexts — biological processes, traits, diseases, phenotypes, and cell types — to these signatures (see **Methods** for details).

We began our investigation of these results by focusing on biological processes that show enrichment strongly correlated with increasing (**Fig. 3.9a,b**) or decreasing (**Fig. 3.9c,d**) age. Apoptosis-related processes are associated with relatively low or non-expression in young age groups but show increased association with older age groups in age signatures in both sexes, especially in Females, reminiscent of observed aging processes [61]. In Males, positive regulation of NF-kappaB signaling and negative regulation of lymphocyte migration also increase in association with age, consistent with other studies [60,62].

On the other hand, processes that are associated with higher expression in younger age groups and lower in older age groups include developmental processes and



collagen fibril organization, which has been noted in multiple prior studies [63–65]. Together, these results suggest that our age signatures and enrichment patterns will be powerful for studying sex-specific processes of aging and development.



**Figure 3.9. Enrichment of experimentally-derived genes sets from in our sex-stratified age signatures.** Age-bias enrichment of experimentally-derived gene sets. Heatmaps show enrichment scores for GO biological processes that are associated with increasing age in (a) Females and (b) Males and GO biological processes that are associated with decreasing age in (c) Females and (d) Males. Each of these biological processes has a Spearman correlation over +0.8 or under -0.8 with age group.

## Discussion

Age and sex have historically not received the attention they deserve in biomedical research, resulting in fundamental gaps in our understanding of how these factors influence normal physiology and disease mechanisms. In this project, we make an effort to enable the biomedical community to systematically address these gaps. First, we assembled the largest set of manually-curated age- and sex-labeled bulk, primary human gene expression samples to date. Using these data, we show that it is possible to predict age group from gene expression with reasonably high accuracy using simple one-vs-rest logistic regression models. We then investigated the genes across the genome with age-stratified sex bias and sex-stratified age bias — both identified in a data-driven manner. These analyses have provided insights into several aspects of age- and sex-related human gene regulation, cellular pathways, and disease.

### Sex and age group prediction from gene expression

Though we conducted an expansive search for transcriptome samples with sex and age labels, there are several tissue and disease biases in our dataset. As blood is one of the easiest tissue to collect from humans, it is not surprising that the largest set of samples are from blood (**Fig. A3.2**). Brain, small intestine, liver, lung, and retina all have hundreds of samples each as well. Often, studies separate samples based on tissue and then determine age- and sex-biased genes through differential expression or age prediction. However, even stratifying samples by tissue is not enough to determine age- and sex-biased genes without prejudice as cell type composition affects these results [35,47]. One study that considered multi-tissue signatures in age prediction showed that prediction improves when gene expression from more than one tissue is used to predict

age [29]. Ren and Kuan [33] recently compared several different tissue-specific and across-tissue genesets to use as features for expression-based age prediction models using data from GTEx. Their results show that an across-tissue geneset derived from expression across all GTEx tissues have similar performance in prediction to using only differentially expressed genes for a given tissue. When training in one tissue and predicting age in another, the across-tissue feature set was superior. We too tested age group prediction using only blood samples in our dataset and found that prediction performance decreased in most age groups (**Fig. A3.11-16**). Combined with results from previous studies, our findings suggest that age signal-to-noise ratio improves when including expression from multiple tissues. In determining sex-biased genes, we found that subsetting to only blood samples does not meaningfully change the results (**Fig. A3.3**). In addition to tissue biases, our dataset also certainly has disease biases due to inherent differences in the incidence of disease in different age and sex groups, along with their likelihood to be studied. The National Institutes of Health lists infectious diseases, brain disorders, and cancer as amongst its top-funded in disease research in recent years [66]. As expected, these diseases make up a large number of the samples in our dataset.

Age group prediction was more difficult in the middle age groups (**Fig. 3.4b,d, Fig. 3.5b,d, Fig. A3.5-8**). This is not surprising, as environmental factors, lifestyle, and aging begin to contribute to more heterogeneity in these age groups, although nonlinearly and nonuniformly [67]. This is also reflected in the similarity between age group models (**Fig. A3.17-21**). There is more similarity between the young age group models than in older age groups, especially when comparing across sexes and technologies. The

dissimilarity between RNA-seq and microarray models is likely due to technological differences including the large difference in dynamic range between RNA-seq and microarray experiments. Nevertheless, even if the genes used in the same age group models are different across technologies, it is possible that they play a role in similar biological processes and pathways, which we can test in the future by comparing enrichment results between technologies.

### **Enrichment of experimental genesets in age and sex signatures**

Our age-stratified sex-signatures showed many immune response-related genesets to be Female-biased (**Fig. 3.8**). Females usually produce a stronger immune response than Males and this is thought to contribute to their increased susceptibility to autoimmune diseases [58,68]. This increased susceptibility is profound, as Females account for 80% of autoimmune disease occurrence [57,58]. A few autoimmune disease incidence rates are close to even between sexes, but there are no common autoimmune diseases that show a bias towards Male prevalence at the degree that autoimmune diseases like rheumatoid arthritis, lupus and Hashimoto's show towards Female prevalence [69]. The many immune response-related genesets found to be Female-biased in our signatures might help probe the molecular underpinnings of these differences.

Our age-stratified sex-biased signatures were able to recapitulate previously-observed sex differences in cell type composition. A study of individuals aged 20–35 years found that Females have a higher number and proportion of B cells [59]. Another study conducted by Márquez and colleagues with a range of individuals aged 22–93 years found a Male-specific decline in B cell proportion after the age of 65 [60]. This study

also found many lymphoid cells and T cell populations to be more abundant in Females, the latter supporting the result of another study that had found naive T cells specifically to be higher in Females [70]. These cell types are more commonly Female-biased in our age-stratified sex-biased signatures (**Fig. 3.8d,e**). On the other hand, the previously mentioned Márquez et al study also found myeloid lineage cells (particularly monocytes) to be more abundant in Males [60] (**Fig. 3.8d,e**). We observe this Male-bias in the younger age groups. Altogether, our signatures show high concordance with sex-biases observed in previous immune studies, suggesting that they will be excellent tools to study other pan-body sex-biased processes and disease mechanisms.

We noted several apoptosis and programmed cell death processes associated with increasing age in our sex-stratified age-biased signatures, which are well known to be associated with the process of aging [61] (**Fig. 3.9a,b**). Positive regulation of NF-kappaB signaling and negative regulation of lymphocyte migration is associated with increasing age in the Male signatures (**Fig. 3.9b**). These observations are consistent with studies that link NF-kappaB signaling to the aging process [62] and show that adaptive immune function decreases with age, especially in men [60]. Other processes associated with decreasing age are developmental processes and collagen fibril organization (**Fig. 9c,d**). Several studies have associated increased age with lower collagen levels and lower integrity and increased dysregulation of the collagen network [63–65]. These biologically-meaningful age associations show the utility of these data-driven signatures for investigating molecular processes related to aging and development.

## **Availability of data and code**

We make the set of ~30,000 age- and sex-associated curated transcriptome samples and code to reproduce these approaches available via GitHub so that other researchers may build upon them for their own studies. Our genome-wide sex-biased and age-biased gene signatures and the associations of hundreds of biological contexts with these signatures will be searchable by an online webserver to make these results easily accessible for biomedical researchers. The community can use these signatures and associated contexts to explore the age- and sex-biased expression patterns of one or more of their favorite genes, use the sex- and age-biased gene signatures to inform the genes prioritized in new studies, and/or search through the thousands of precalculated enrichment results using the names of pathways, cell types, phenotypes, traits, and diseases of interest to examine the association of constituent genes with sex and age. Finally, we will also make expression values of labeled transcriptomes available via the web interface for biomedical researchers to search and compare gene expression between age and sex groups of interest.

## **Methods**

### **Data collection**

We downloaded human microarray gene expression data from the Gene Expression Omnibus (GEO) [15] as raw CEL files. Due to different platforms measuring different genes, we restricted the data to samples from the *Affymetrix Human Genome U133 Plus 2.0 Array*. The CEL files were processed with background subtraction, quantile transformation, and summarization using fRMA [71] based on custom CDF [72] mapping probes to Entrez gene IDs. We downloaded Salmon [73] output files for all

human RNA-seq samples available in refine.bio [40] and removed samples with over 50% zero counts. The Salmon-calculated TPM values for the remaining RNA-seq samples were used for all further analysis. We also restricted analysis to genes measured on both platforms, for a total of 18,478 genes.

### **Curation of age and sex labels**

Age and sex labels were curated for microarray and RNA-seq data with a combination of text mining and manual curation. For the microarray data, we downloaded sample descriptions for our microarray from GEO and used simple text matching to identify samples associated with potential age and sex information. We manually checked these text-matched labels by reading the sample descriptions and verifying that the label was correct and removing erroneously labeled samples. For RNA-seq data, we downloaded metaSRA [74] version 1.8 to identify samples associated with potential age and sex information. We then used ffq [75] to fetch sample accession data from the Sequence Read Archive (SRA) [41] to match the sample identifiers used in metaSRA to the run identifiers used in refine.bio. We manually checked these labels as well by reading sample descriptions obtained from SRA. Both microarray and RNA-seq sample descriptions were also used to remove samples that are not primary human samples, and in the case of RNA-seq, to remove samples that are not bulk samples (i.e. single cell or single nuclei). Examples of removed samples include cell lines, xenografts, and pooled samples.

### **Age-stratified sex-biased genes**

Sex-biased genes were determined separately in microarray and RNA-seq data. Within each set of data, the expression values of each gene were z-scored across all samples.

Samples were divided by age group. For each gene, the z-score value that best separated the male and Female samples within an age group was found by calculating the arithmetic mean of sensitivity and specificity at each value from the minimum to maximum z score in steps of 0.2. Essentially, we considered each z score value as a simple model to predict sex. Every sample with an expression z score above the value was labeled 'Female' and every sample with an expression z score below the value was labeled 'Male'. By reconfiguring the balanced accuracy equation (below) to replace 'positives' with 'Females' and 'negatives' with Males, we create a Female-bias metric for the expression of each gene. The balanced accuracy is the arithmetic mean of sensitivity and specificity of a model and ranges from 0 to 1. In our metric, 1 is the extreme end of Female bias (all Female samples have higher expression than the cutoff value and all male samples have lower expression), 0 is the extreme end of male bias (all male samples have higher expression than the cutoff value and all Female samples have lower expression), and 0.5 is perfectly balanced. The same process was repeated using only samples from blood for the blood-only analysis. For heatmaps of strongly sex-biased genes (**Fig. 3.2, Fig. A3.3**), we subtract the Female-bias metric score from 1 to plot the balanced accuracy if Males are considered the more highly-expressed group.

$$\text{balanced accuracy} = \frac{1}{2} \left( \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} + \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \right)$$

$$\text{Female-bias metric} = \frac{1}{2} \left( \frac{\text{True Females}}{\text{True Females} + \text{False Males}} + \frac{\text{True Males}}{\text{True Males} + \text{False Females}} \right)$$

Logistic regression models



Separately for microarray and RNA-seq data, in each sex, for each age group, we trained a one-vs-rest logistic regression model with an elasticnet penalty. RNA-seq data was asinh transformed but no other scaling was used. In every sex/technology combination, three folds of all the samples were created for cross-validation by assigning entire datasets at a time, roughly by assigning the three largest remaining datasets to each of the three folds in a manner that kept the number of samples and datasets as equal as possible across all folds. In RNA-seq data, 145 fetal samples were added by predicting sex in samples without sex information to increase the number of samples to a number viable for 3-fold cross validation. Sex was predicted based on 15 genes with over 0.9 balanced accuracy in separating Female and Male samples with a simple expression cutoff. Only samples with sex agreement in at least 13 out of the 15 genes were labelled with the predicted sex and kept in our labeled set.

To create sex-stratified age signatures in blood samples only, we assigned the 3 largest datasets of each age group to one of three test folds making concessions to keep full datasets together if there was a conflict across age groups. To ensure each test fold had at least one dataset with at least 5 samples from a given age group, we excluded age groups without enough data to meet this threshold in both microarray and RNA-seq data. The remaining datasets were used in training for all three test folds. The folds sizes and age group distribution for all models are shown in Figures A3.4 and A3.10.

### **Curation of other genesets for enrichment analysis**

Genesets from previous sex differential expression studies, GTEx, SAGD, and Guo *et al*, were downloaded from the supplemental data in their respective publications. We used all genes declared significant by the authors for any defined group they had

curated. Biological Process annotations with experimental evidence codes (EXP, IDA, IPI, IMP, IGI, TAS, IC) were downloaded from the Gene Ontology [50] and propagated to all ancestor Gene Ontology Biological Process (GOBP) terms. We then subset GOBP terms to those with 10 to 200 annotated genes to remove terms with too few genes to do enrichment or terms that are too general to be practically useful. Human disease genes were downloaded from the Monarch Initiative [52,53] webpage at [https://data.monarchinitiative.org/latest/tsv/gene\\_associations/gene\\_disease.9606.tsv.gz](https://data.monarchinitiative.org/latest/tsv/gene_associations/gene_disease.9606.tsv.gz). The genes in the Monarch Initiative file are annotated to disease terms in the Mondo [76] disease ontology. We propagated these annotations to all ancestor Mondo terms and removed any terms without at least 10 genes. Phenotype genes were obtained from Mouse Genome Informatics [77] (MGI) file that can be found at [http://www.informatics.jax.org/downloads/reports/MGI\\_GenePheno.rpt](http://www.informatics.jax.org/downloads/reports/MGI_GenePheno.rpt). We converted these annotations to human genes using the MGI file that can be downloaded from [http://www.informatics.jax.org/downloads/reports/HOM\\_MouseHumanSequence.rpt](http://www.informatics.jax.org/downloads/reports/HOM_MouseHumanSequence.rpt) and propagated them to all ancestor terms in the Mammalian Phenotype Ontology [78]. We created sets of genes for GWAS Atlas [51] traits using the Release 3 metadata file at [https://atlas.ctglab.nl/#:~:text=Plain%20text%20file%3A-.gwasATLAS\\_v20191115,-.txt.gz%0AExcel](https://atlas.ctglab.nl/#:~:text=Plain%20text%20file%3A-.gwasATLAS_v20191115,-.txt.gz%0AExcel) and the MAGMA p value file that accompanies these traits at [https://atlas.ctglab.nl/#:~:text=gwasATLAS\\_v20191115\\_magma\\_P](https://atlas.ctglab.nl/#:~:text=gwasATLAS_v20191115_magma_P). The genes with the 25 lowest p values were selected as the geneset for each trait. Cell type marker genes from Ianevski et al [55] were downloaded from their github at the following link: [https://github.com/ianeovskiAleksandr/sc-type/blob/master/ScTypeDB\\_full.xlsx](https://github.com/ianeovskiAleksandr/sc-type/blob/master/ScTypeDB_full.xlsx).

## Enrichment analysis

In order to determine an enrichment score for each geneset in the age and sex signatures, we used a permutation test to calculate a z score. For each geneset, we calculated the average age-stratified sex enrichment or sex-stratified age bias using the Female-bias metric (converted to a range of  $-1$  to  $+1$  by multiplying it by 2 and subtracting 1) or weight in the logistic regression model, respectively. Then, we pulled a random sample of genes of the same size as the geneset to calculate the average bias of that set. This was repeated 100,000 times and the mean and standard deviation of this distribution was used to calculate a z score for the bias of the original geneset.

The enrichment process for age-stratified sex-biased genes was done separately in the microarray and RNA-seq data using their respective converted Female-bias metric scores for all genes. The z score obtained for each geneset in microarray and RNA-seq data is combined via Stouffer's method (equation below). For sex-stratified age enrichment, the weight of each gene in the one-vs-rest logistic regression models were used in the enrichment process. As there were six trained models for each age group (3 RNA-seq models, 3 microarray models), we used our permutation test to determine a z score for each experimentally-derived geneset for each model, averaged the z scores across the 3 folds in each technology separately, and combined the z score from RNA-seq and microarray with Stouffer's Method. This final z score was taken as the enrichment score for each age group (equation below). For visualization, extreme z scores were reduced to  $+5$  or  $-5$  if they were higher or lower than that, respectively.

$$\text{Stouffer's } Z = \frac{Z_{\text{microarray}} + Z_{\text{RNA-seq}}}{2}$$

For heatmaps with “representative” sets of genesets/ontology terms, the python package orsum [79] was used to find a nonredundant set of terms. The orsum method will discard a geneset/term if there is a more significant term that annotates at least the same genes. The remaining more significant term is representative for the discarded term.

## REFERENCES

1. Ober C, Loisel DA, Gilad Y. Sex-specific genetic architecture of human disease. *Nat Rev Genet.* Nature Publishing Group; 2008;9:911–22.
2. Clayton JA, Collins FS. Policy: NIH to balance sex in cell and animal studies. *Nat News.* 2014;509:282.
3. Tannenbaum C, Day D. Age and sex in drug development and testing for adults. *Pharmacol Res.* 2017;121:83–93.
4. Herrera AP, Snipes SA, King DW, Torres-Vigil I, Goldberg DS, Weinberg AD. Disparate Inclusion of Older Adults in Clinical Trials: Priorities and Opportunities for Policy and Practice Change. *Am J Public Health.* 2010;100:S105–12.
5. Forcina V, Vakeesan B, Paulo C, Mitchell L, Bell JA, Tam S, et al. Perceptions and attitudes toward clinical trials in adolescent and young adults with cancer: a systematic review. *Adolesc Health Med Ther.* 2018;9:87–94.
6. Haast RA, Gustafson DR, Kiliaan AJ. Sex Differences in Stroke. *J Cereb Blood Flow Metab.* SAGE Publications Ltd STM; 2012;32:2100–7.
7. Zein JG, Erzurum SC. Asthma is Different in Women. *Curr Allergy Asthma Rep.* 2015;15:28.
8. Cypess AM, Lehman S, Williams G, Tal I, Rodman D, Goldfine AB, et al. Identification and Importance of Brown Adipose Tissue in Adult Humans. *N Engl J Med.* Massachusetts Medical Society; 2009;360:1509–17.
9. Bernabeu E, Canela-Xandri O, Rawlik K, Talenti A, Prendergast J, Tenesa A. Sex differences in genetic architecture in the UK Biobank. *Nat Genet.* 2021;53:1283–9.
10. Costa AR, Lança de Oliveira M, Cruz I, Gonçalves I, Cascalheira JF, Santos CRA. The Sex Bias of Cancer. *Trends Endocrinol Metab.* 2020;31:785–99.
11. Cenko E, Yoon J, Kedev S, Stankovic G, Vasiljevic Z, Krljanac G, et al. Sex Differences in Outcomes After STEMI: Effect Modification by Treatment Strategy and Age. *JAMA Intern Med.* 2018;178:632–9.
12. Khramtsova EA, Davis LK, Stranger BE. The role of sex in the genomics of human complex traits. *Nat Rev Genet.* 2019;20:173–90.
13. Riondino S, Ferroni P, Roselli M, Guadagni F. Precision medicine in the ageing world: The role of biospecimen sciences. *Int J Biol Markers.* SAGE Publications Ltd STM; 2019;34:3–5.

14. Singh PP, Demmitt BA, Nath RD, Brunet A. The Genetics of Aging: A Vertebrate Perspective. *Cell*. 2019;177:200–20.
15. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res*. 2013;41:D991-995.
16. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, et al. ArrayExpress update—simplifying data submissions. *Nucleic Acids Res*. 2015;43:D1113–6.
17. Athar A, Füllgrabe A, George N, Iqbal H, Huerta L, Ali A, et al. ArrayExpress update – from bulk to single-cell expression data. *Nucleic Acids Res*. 2019;47:D711–5.
18. Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, et al. Understanding multicellular function and disease with human tissue-specific networks. *Nat Genet*. 2015;47:569–76.
19. Krishnan A, Zhang R, Yao V, Theesfeld CL, Wong AK, Tadych A, et al. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat Neurosci*. 2016;19:1454–62.
20. Mayne BT, Bianco-Miotto T, Buckberry S, Breen J, Clifton V, Shoubridge C, et al. Large Scale Gene Expression Meta-Analysis Reveals Tissue-Specific, Sex-Biased Gene Expression in Humans. *Front Genet*. 2016;7:183.
21. Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. The human transcriptome across tissues and individuals. *Science*. 2015;348:660–5.
22. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. Nature Publishing Group; 2013;45:580–5.
23. Guo S, Zhou Y, Zeng P, Xu G, Wang G, Cui Q. Identification and analysis of the human sex-biased genes. *Brief Bioinform*. 2018;19:188–98.
24. Gershoni M, Pietrokovski S. The landscape of sex-differential transcriptome and its consequent selection in human adults. *BMC Biol*. 2017;15:7.
25. Naqvi S, Godfrey AK, Hughes JF, Goodheart ML, Mitchell RN, Page DC. Conservation, acquisition, and functional impact of sex-biased gene expression in mammals. *Science*. American Association for the Advancement of Science; 2019;365:eaaw7317.
26. Lopes-Ramos CM, Chen C-Y, Kuijjer ML, Paulson JN, Sonawane AR, Fagny M, et al. Sex Differences in Gene Expression and Regulatory Networks across 29 Human

Tissues. *Cell Rep.* 2020;31:107795.

27. de Magalhães JP, Curado J, Church GM. Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics.* 2009;25:875–81.

28. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sada S, et al. Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Mol Cell. Elsevier*; 2013;49:359–67.

29. Wang F, Yang J, Lin H, Li Q, Ye Z, Lu Q, et al. Improved Human Age Prediction by Using Gene Expression Profiles From Multiple Tissues. *Front Genet* [Internet]. 2020 [cited 2022 Nov 11];11. Available from: <https://www.frontiersin.org/articles/10.3389/fgene.2020.01025>

30. Bulteau R, Francesconi M. Real age prediction from the transcriptome with RAPToR. *Nat Methods. Nature Publishing Group*; 2022;19:969–75.

31. Fleischer JG, Schulte R, Tsai HH, Tyagi S, Ibarra A, Shokhirev MN, et al. Predicting age from the transcriptome of human dermal fibroblasts. *Genome Biol.* 2018;19:221.

32. Shokhirev MN, Johnson AA. Modeling the human aging transcriptome across tissues, health status, and sex. *Aging Cell.* 2021;20:e13280.

33. Ren X, Kuan PF. RNAAgeCalc: A multi-tissue transcriptional age calculator. *PLOS ONE. Public Library of Science*; 2020;15:e0237006.

34. Peters MJ, Joehanes R, Pilling LC, Schurmann C, Conneely KN, Powell J, et al. The transcriptional landscape of age in human peripheral blood. *Nat Commun.* 2015;6:8570.

35. Pellegrino-Coppola D, Claringbould A, Stutvoet M, Heijmans BT, 't Hoen PAC, van Meurs J, et al. Correction for both common and rare cell types in blood is important to identify genes that correlate with age. *BMC Genomics.* 2021;22:184.

36. Cardoso-Moreira M, Halbert J, Vallotton D, Velten B, Chen C, Shao Y, et al. Gene expression across mammalian organ development. *Nature.* 2019;1.

37. Hägg S, Jylhävä J. Sex differences in biological aging with a focus on human studies. *eLife.* 10:e63425.

38. Mank JE, Rideout EJ. Developmental mechanisms of sex differences: from cells to organisms. *Development.* 2021;148:dev199750.

39. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 2002;30:207–10.

40. Greene CS, Hu D, Jones RWW, Liu S, Mejia DS, Patro R, et al. refine.bio [Internet]. Refine.bio. [cited 2021 Sep 13]. Available from: <https://www.refine.bio>
41. Leinonen R, Sugawara H, Shumway M. The Sequence Read Archive. *Nucleic Acids Res.* 2011;39:D19–21.
42. Xu J, Peng X, Chen Y, Zhang Y, Ma Q, Liang L, et al. Free-living human cells reconfigure their chromosomes in the evolution back to uni-cellularity. *eLife.* 6:e28070.
43. Kaur G, Dufour JM. Cell lines. *Spermatogenesis.* 2012;2:1–5.
44. Gillet J-P, Varma S, Gottesman MM. The Clinical Relevance of Cancer Cell Lines. *JNCI J Natl Cancer Inst.* 2013;105:452–8.
45. Flynn E, Chang A, Altman RB. Large-scale labeling and assessment of sex bias in publicly available expression data. *BMC Bioinformatics.* 2021;22:168.
46. Capes-Davis A, Theodosopoulos G, Atkin I, Drexler HG, Kohara A, MacLeod RAF, et al. Check your cultures! A list of cross-contaminated or misidentified cell lines. *Int J Cancer.* 2010;127:1–8.
47. Oliva M, Muñoz-Aguirre M, Kim-Hellmuth S, Wucher V, Gewirtz ADH, Cotter DJ, et al. The impact of sex on gene expression across human tissues. *Science.* 2020;369:eaba3066.
48. Zhang Q, Vallerga C, Walker R, Lin T, Henders A, Montgomery G, et al. Improved prediction of chronological age from DNA methylation limits it as a biomarker of ageing. 2018 [cited 2018 Jun 1]; Available from: <http://biorxiv.org/lookup/doi/10.1101/327890>
49. Gonçalves CI, Fonseca F, Borges T, Cunha F, Lemos MC. Expanding the genetic spectrum of ANOS1 mutations in patients with congenital hypogonadotropic hypogonadism. *Hum Reprod.* 2017;32:704–11.
50. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. *Nat Genet.* 2000;25:25–9.
51. Liu X, Tian D, Li C, Tang B, Wang Z, Zhang R, et al. GWAS Atlas: an updated knowledgebase integrating more curated associations in plants and animals. *Nucleic Acids Res.* 2022;gkac924.
52. McMurry JA, Köhler S, Washington NL, Balhoff JP, Borromeo C, Brush M, et al. Navigating the Phenotype Frontier: The Monarch Initiative. *Genetics.* 2016;203:1491–5.
53. Shefchek KA, Harris NL, Gargano M, Matentzoglou N, Unni D, Brush M, et al. The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.* 2020;48:D704–15.



54. Eppig JT. Mouse Genome Informatics (MGI) Resource: Genetic, Genomic, and Biological Knowledgebase for the Laboratory Mouse. *ILAR J.* 2017;58:17–41.
55. Ianevski A, Giri AK, Aittokallio T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat Commun.* Nature Publishing Group; 2022;13:1246.
56. Shi M-W, Zhang N-A, Shi C-P, Liu C-J, Luo Z-H, Wang D-Y, et al. SAGD: a comprehensive sex-associated gene database from transcriptomes. *Nucleic Acids Res.* 2019;47:D835–40.
57. Whitacre CC. Sex differences in autoimmune disease. *Nat Immunol.* Nature Publishing Group; 2001;2:777–80.
58. Klein SL, Flanagan KL. Sex differences in immune responses. *Nat Rev Immunol.* Nature Publishing Group; 2016;16:626–38.
59. Abdullah M, Chai P-S, Chong M-Y, Tohit ERM, Ramasamy R, Pei CP, et al. Gender effect on in vitro lymphocyte subset levels of healthy individuals. *Cell Immunol.* 2012;272:214–9.
60. Márquez EJ, Chung C, Marches R, Rossi RJ, Nehar-Belaid D, Eroglu A, et al. Sexual-dimorphism in human immune system aging. *Nat Commun.* 2020;11:1–17.
61. Tower J. Programmed cell death in aging. *Ageing Res Rev.* 2015;23:90–100.
62. García-García VA, Alameda JP, Page A, Casanova ML. Role of NF- $\kappa$ B in Ageing and Age-Related Diseases: Lessons from Genetically Modified Mouse Models. *Cells.* Multidisciplinary Digital Publishing Institute; 2021;10:1906.
63. Coudrillier B, Pijanka J, Jefferys J, Sorensen T, Quigley HA, Boote C, et al. Collagen Structure and Mechanical Properties of the Human Sclera: Analysis for the Effects of Age. *J Biomech Eng [Internet].* 2015 [cited 2022 Nov 12];137. Available from: <https://doi.org/10.1115/1.4029430>
64. Podolsky MJ, Yang CD, Valenzuela CL, Datta R, Huang SK, Nishimura SL, et al. Age-dependent regulation of cell-mediated collagen turnover. *JCI Insight.* 5:e137519.
65. Wang X, Shen X, Li X, Mauli Agrawal C. Age-related changes in the collagen network and toughness of bone. *Bone.* 2002;31:1–7.
66. National Institutes of Health. Research Portfolio Online Reporting Tools (RePORT) [Internet]. *Res. Portf. Online Report. Tools Rep.* 2022 [cited 2022 Nov 16]. Available from: <https://report.nih.gov/funding/categorical-spending#/>
67. Nguyen QD, Moodie EM, Forget M-F, Desmarais P, Keezer MR, Wolfson C. Health

Heterogeneity in Older Adults: Exploration in the Canadian Longitudinal Study on Aging. *J Am Geriatr Soc*. 2021;69:678–87.

68. Giefing-Kröll C, Berger P, Lepperdinger G, Grubeck-Loebenstien B. How sex and age affect immune responses, susceptibility to infections, and response to vaccination. *Aging Cell*. John Wiley & Sons, Ltd; 2015;14:309–21.

69. Why Are Women and Men So Different in Autoimmune Disease? [Internet]. [cited 2022 Nov 16]. Available from: <https://www.science.org/content/blog-post/why-are-women-and-men-so-different-autoimmune-disease>

70. Olson NC, Doyle MF, Jenny NS, Huber SA, Psaty BM, Kronmal RA, et al. Decreased Naive and Increased Memory CD4+ T Cells Are Associated with Subclinical Atherosclerosis: The Multi-Ethnic Study of Atherosclerosis. *PLOS ONE*. Public Library of Science; 2013;8:e71498.

71. McCall MN, Bolstad BM, Irizarry RA. Frozen robust multiarray analysis (fRMA). *Biostat Oxf Engl*. 2010;11:242–53.

72. Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, et al. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res*. 2005;33:e175.

73. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*. Nature Publishing Group; 2017;14:417–9.

74. Bernstein MN, Doan A, Dewey CN. MetaSRA: normalized human sample-specific metadata for the Sequence Read Archive. *Bioinformatics*. 2017;33:2914–23.

75. Gálvez-Merchán Á, Min KH (Joseph), Pachter L, Boeshaghi AS. Metadata retrieval from sequence databases with ffq [Internet]. *bioRxiv*; 2022 [cited 2022 Oct 6]. p. 2022.05.18.492548. Available from: <https://www.biorxiv.org/content/10.1101/2022.05.18.492548v2>

76. Vasilevsky NA, Matentzoglou NA, Toro S, Flack JE, Hegde H, Unni DR, et al. Mondo: Unifying diseases for the world, by the world [Internet]. *medRxiv*; 2022 [cited 2022 Nov 6]. p. 2022.04.13.22273750. Available from: <https://www.medrxiv.org/content/10.1101/2022.04.13.22273750v3>

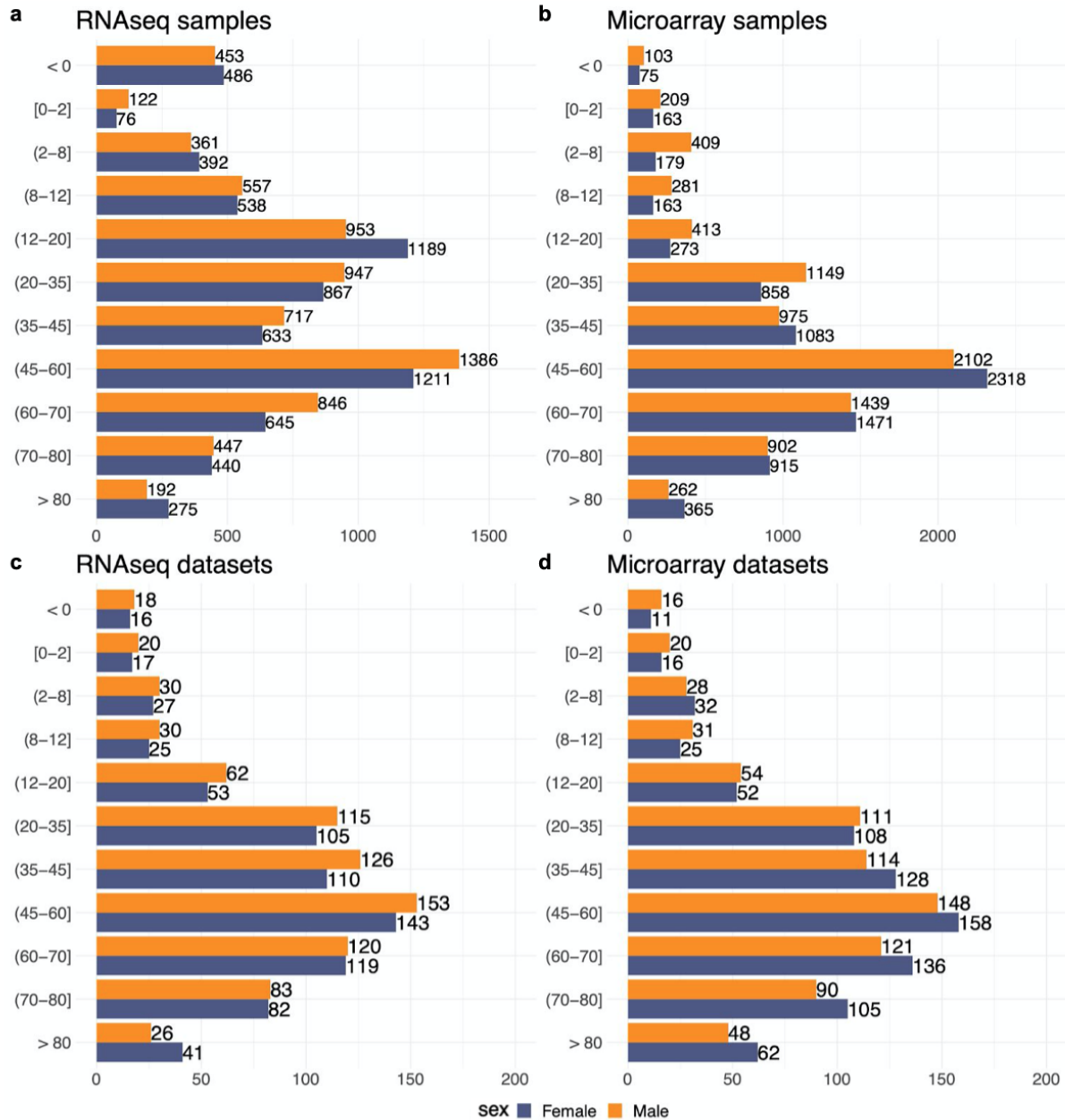
77. Blake JA, Eppig JT, Kadin JA, Richardson JE, Smith CL, Bult CJ. Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. *Nucleic Acids Res*. 2017;45:D723–9.

78. Smith CL, Eppig JT. The Mammalian Phenotype Ontology: enabling robust annotation and comparative analysis. *Wiley Interdiscip Rev Syst Biol Med*.

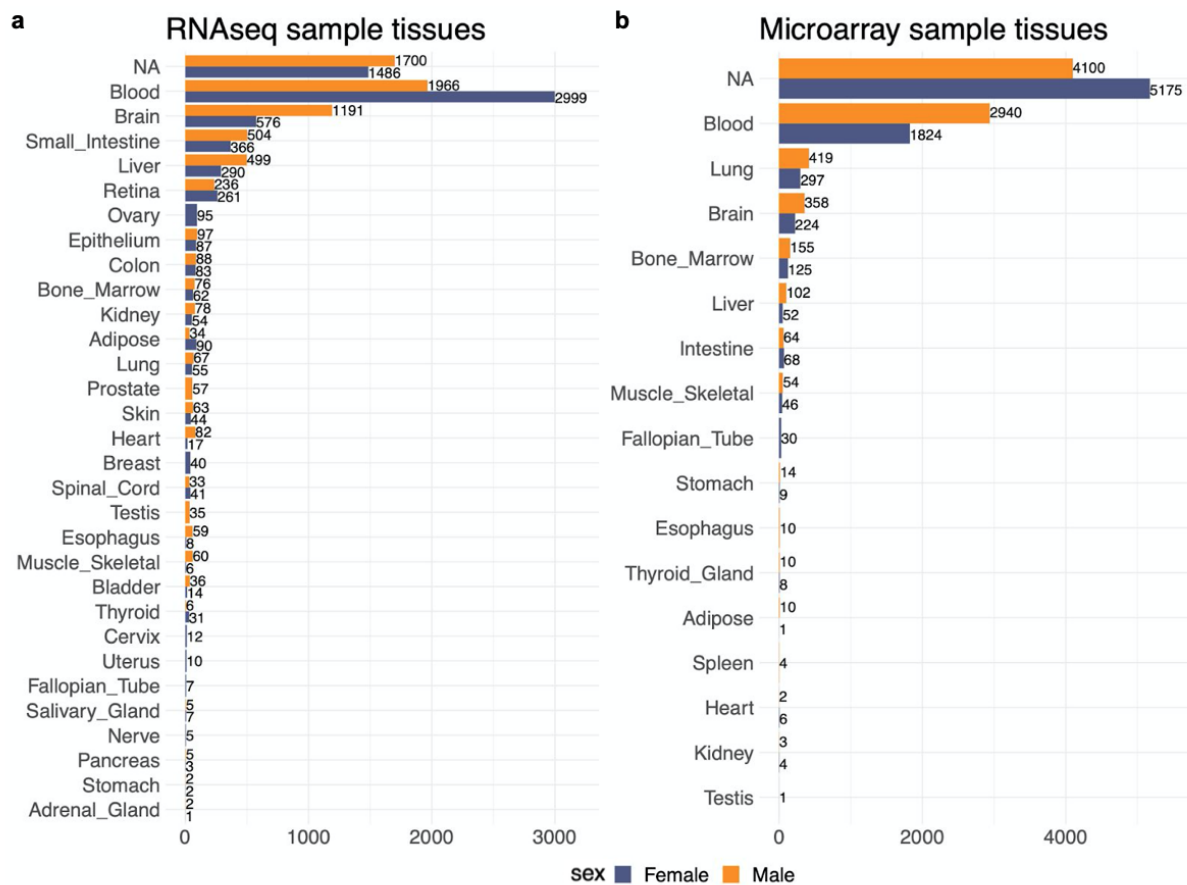
2009;1:390–9.

79. Ozisik O, Térézol M, Baudot A. orsum: a Python package for filtering and comparing enrichment analyses using a simple principle. BMC Bioinformatics. 2022;23:293.

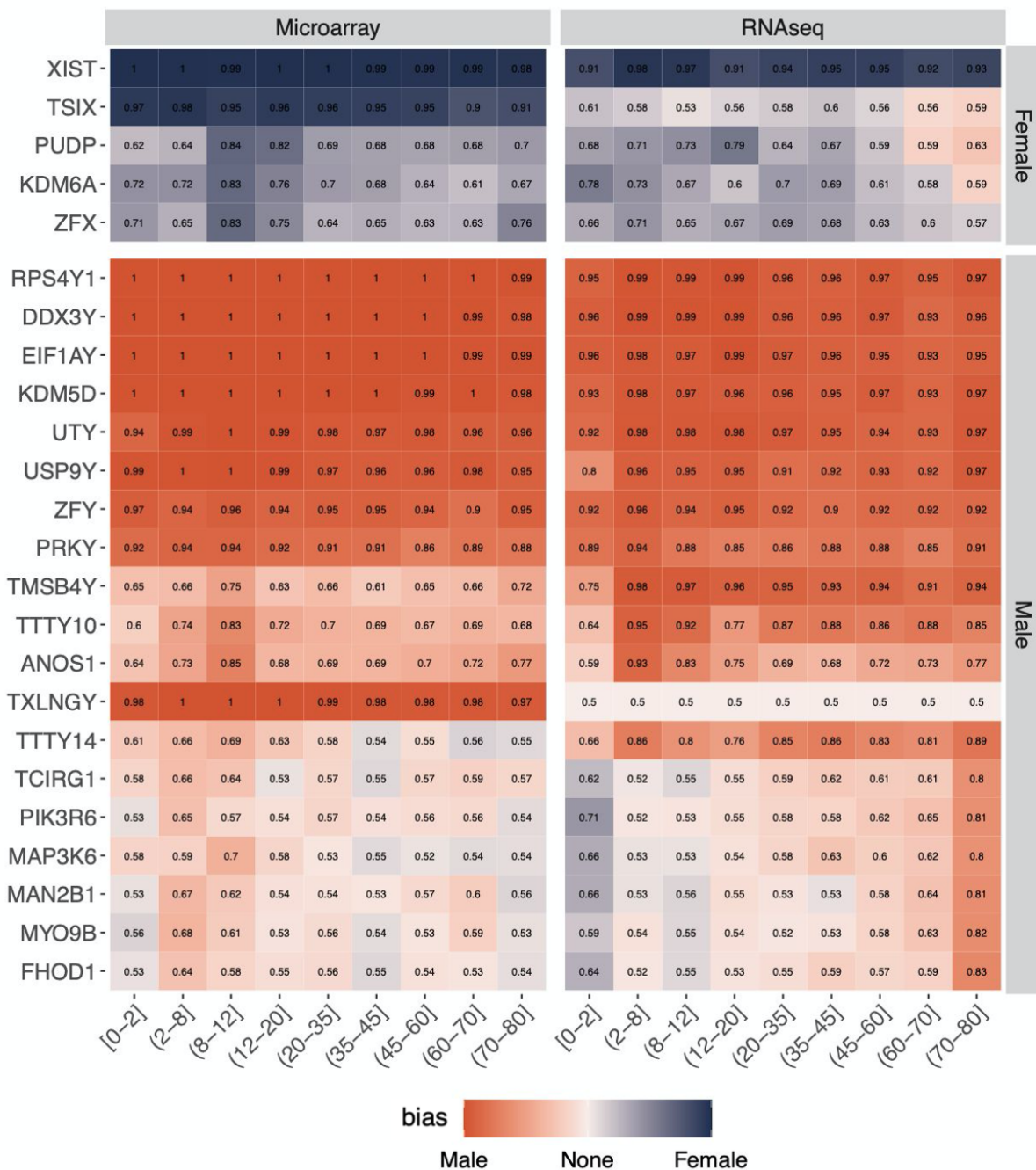
## APPENDIX



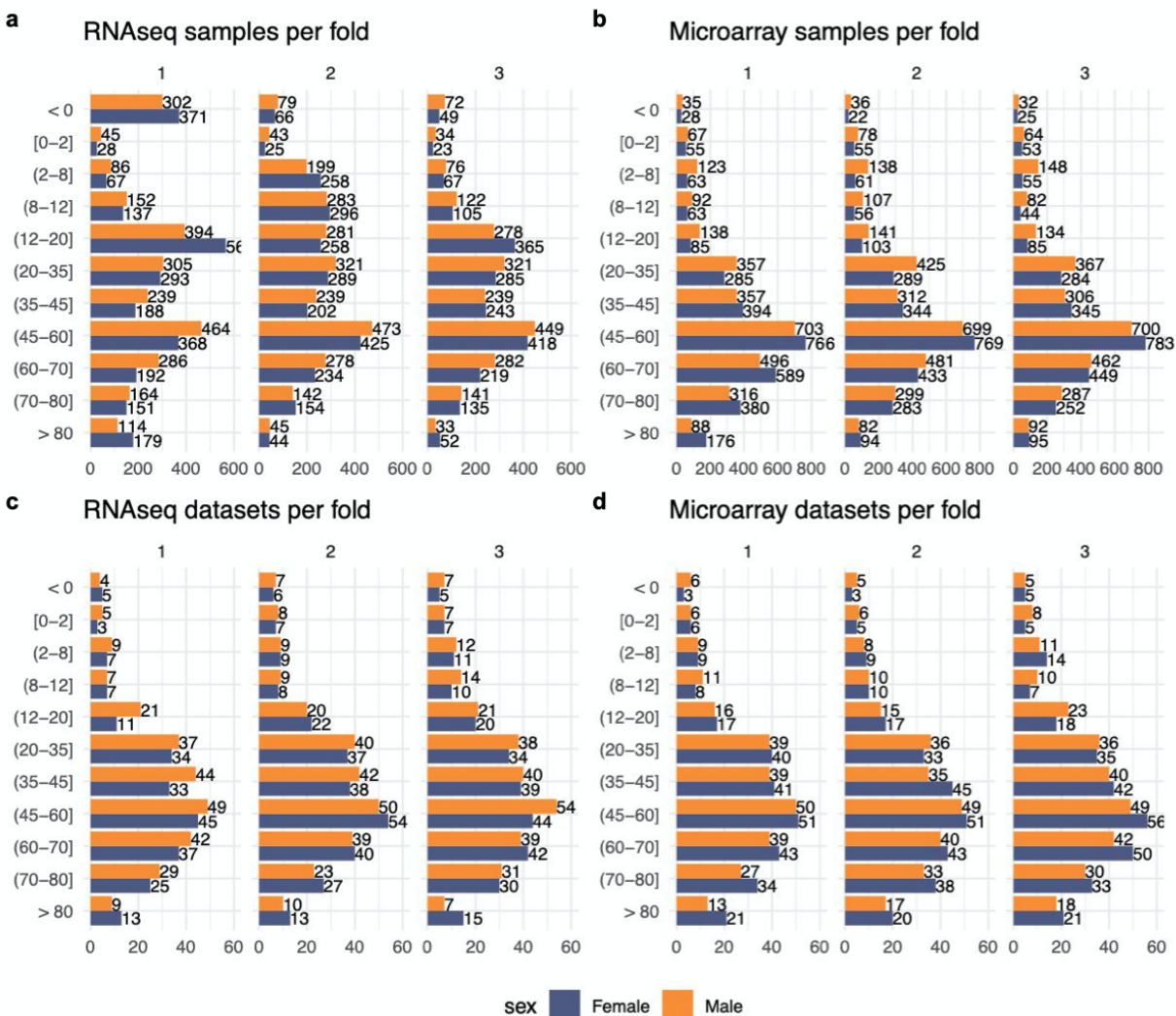
**Figure A3.1. Number of samples and datasets across age and sex groups.** Number of samples labeled with age and sex in (a) RNA-seq and (b) microarray. Number of datasets in (c) RNA-seq and (d) microarray.



**Figure A3.2. Number of samples across tissues.** Number of samples per tissue in (a) RNA-seq and (b) microarray.



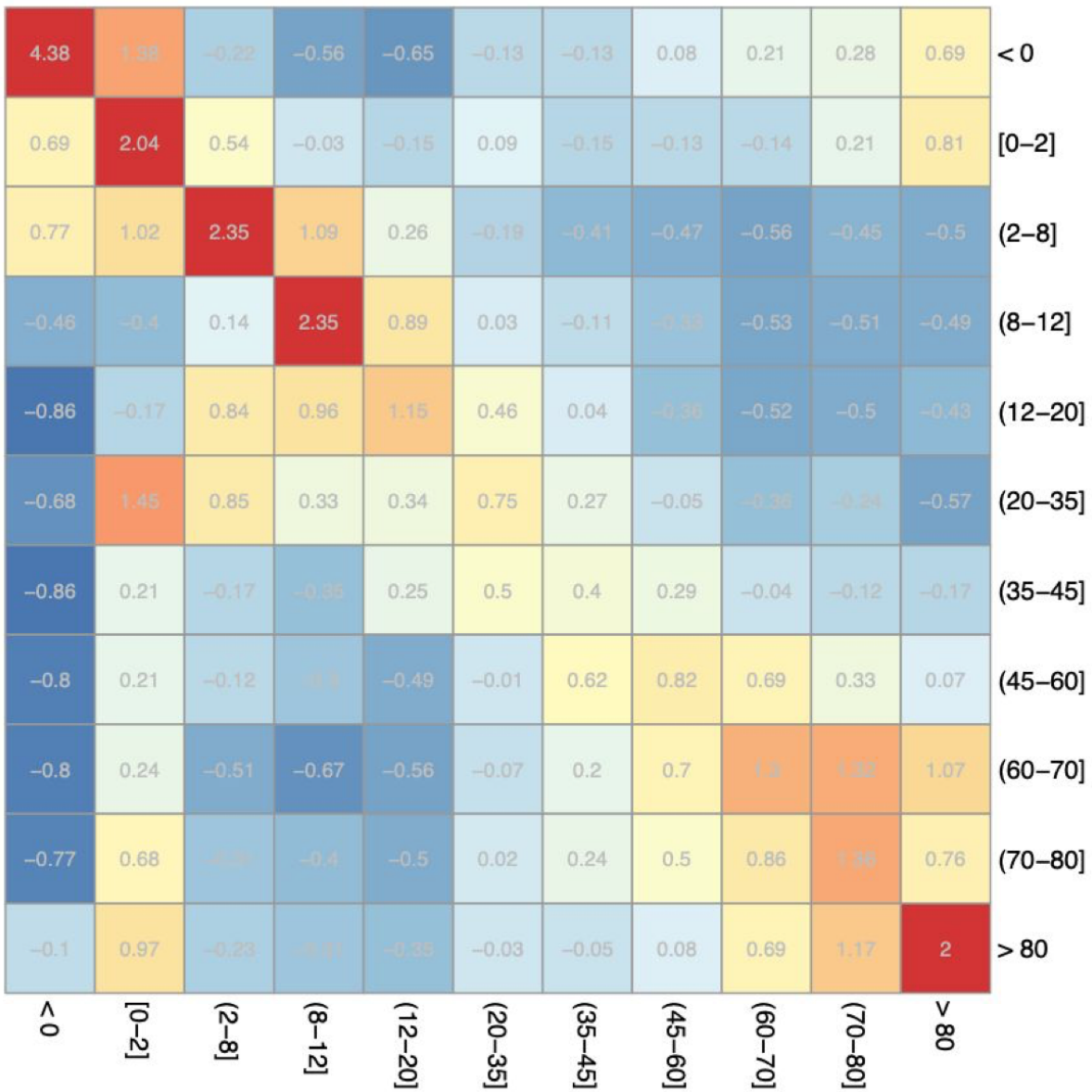
**Figure A3.3. Most strongly sex-biased genes in blood samples across age ranges.** The heatmap displays all genes (x axis) that had a balanced accuracy of at least 0.8 in any age range (y axis) when separating Female and Male microarray or RNA-seq blood samples.



**Figure A3.4. Fold sizes for age range prediction models.** The top barplots show the number of samples in each fold for models trained in (a) RNA-seq and (b) microarray. The bottom barplots show the number of datasets in each fold for models trained in (c) RNA-seq and (d) microarray.



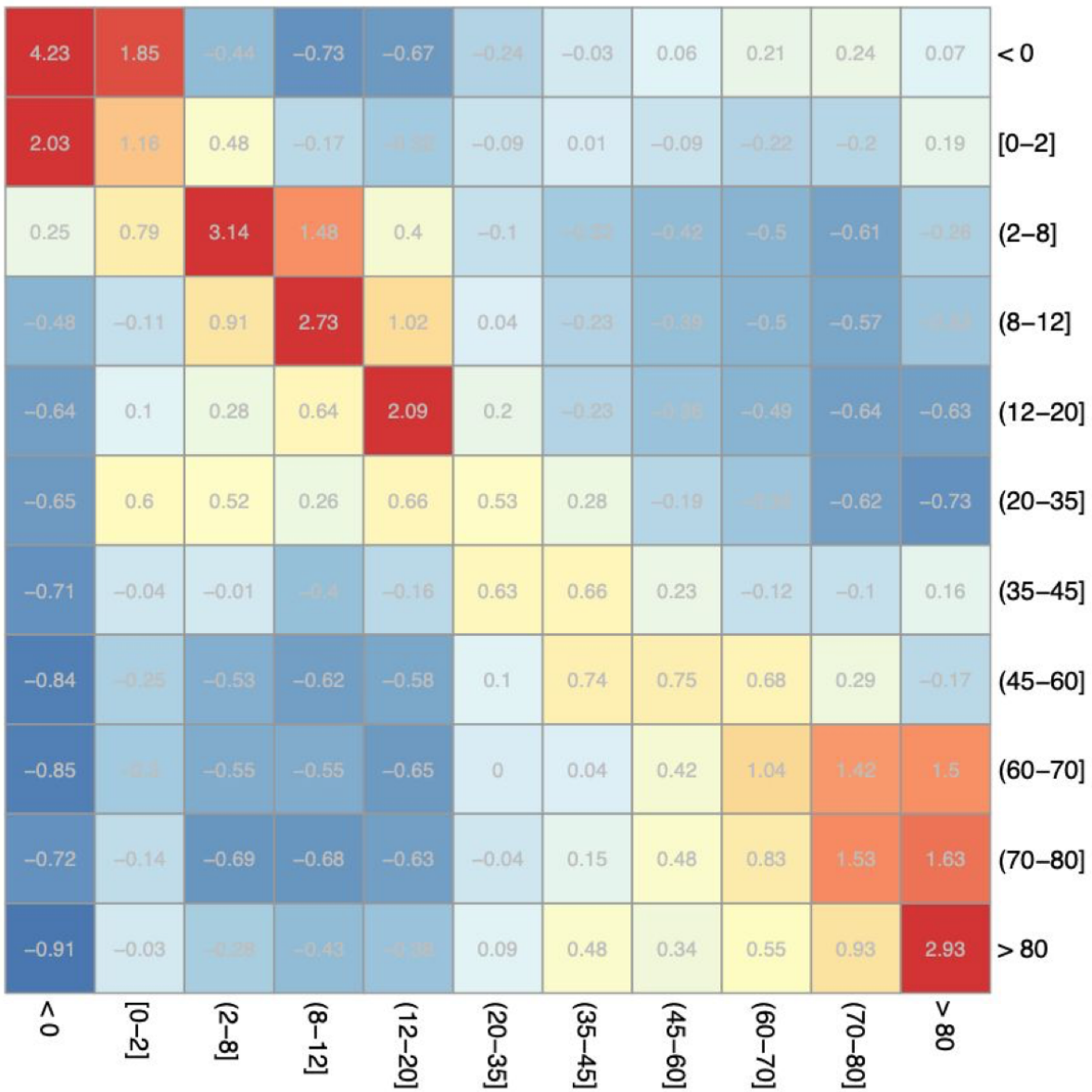
### RNAseq Female models $\log_2(\text{auPRC}/\text{prior})$



**Figure A3.5. Performance of RNA-seq Female age range prediction models.** The heatmaps contain average  $\log_2(\text{auPRC}/\text{prior})$  performance of RNA-seq models across 3 folds trained in Female samples on the age range labeled in the rows while evaluated as if the age range labeled in the column were the positive examples.

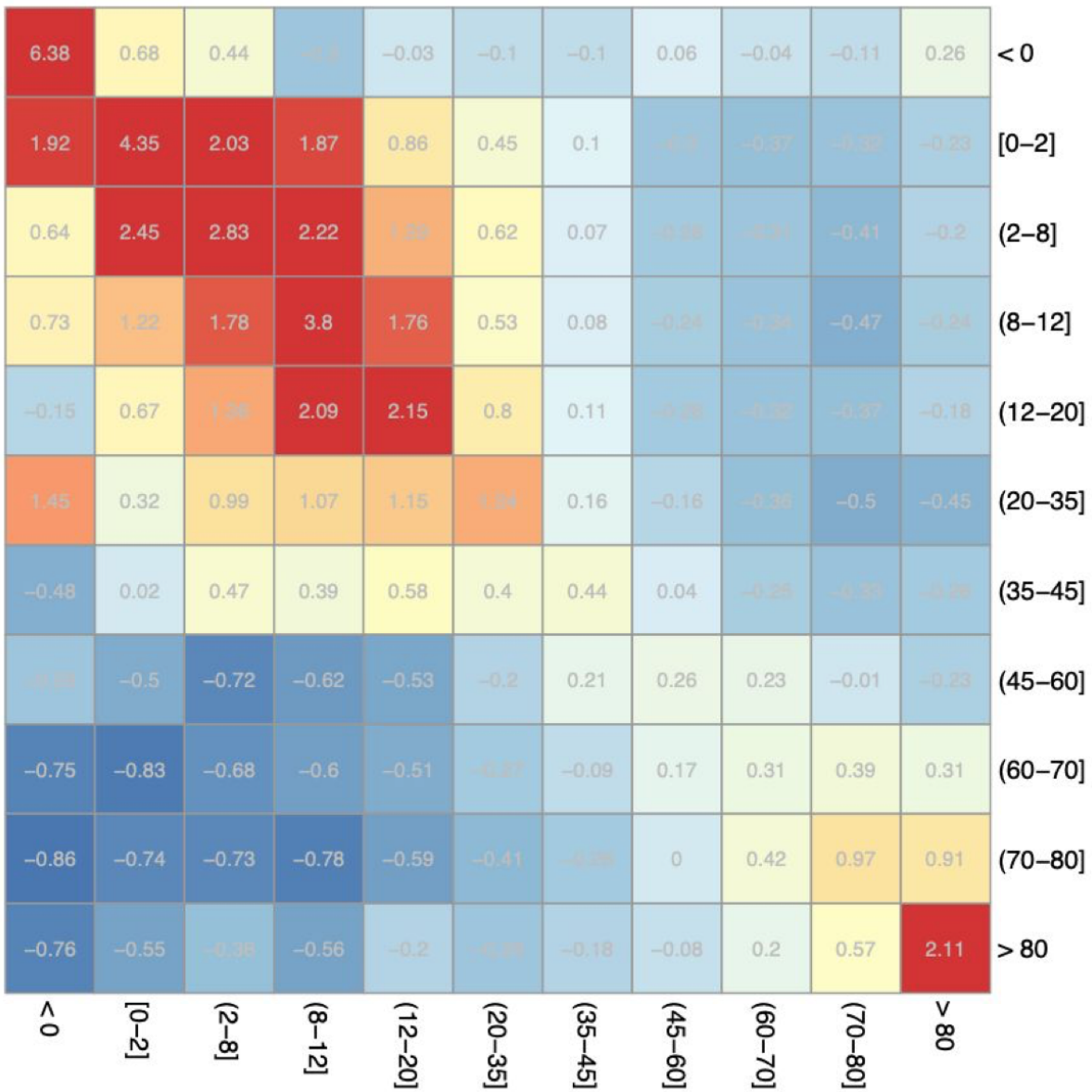


### RNAseq Male models $\log_2(\text{auPRC}/\text{prior})$



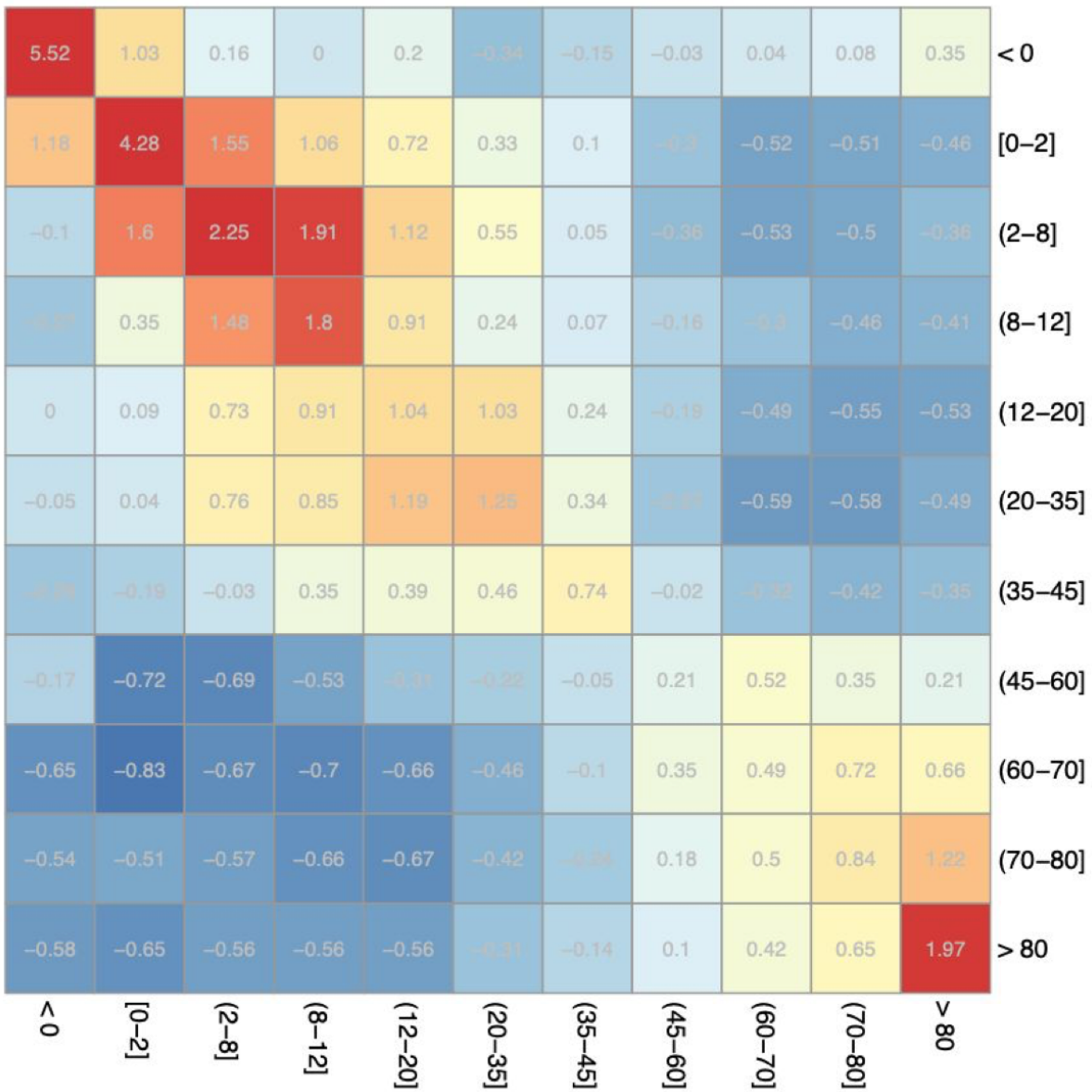
**Figure A3.6. Performance of RNA-seq Male age range prediction models.** The heatmaps contain average  $\log_2(\text{auPRC}/\text{prior})$  performance of RNA-seq models across 3 folds trained in Male samples on the age range labeled in the rows while evaluated as if the age range labeled in the column were the positive examples.

### Microarray Female models $\log_2(\text{auPRC}/\text{prior})$

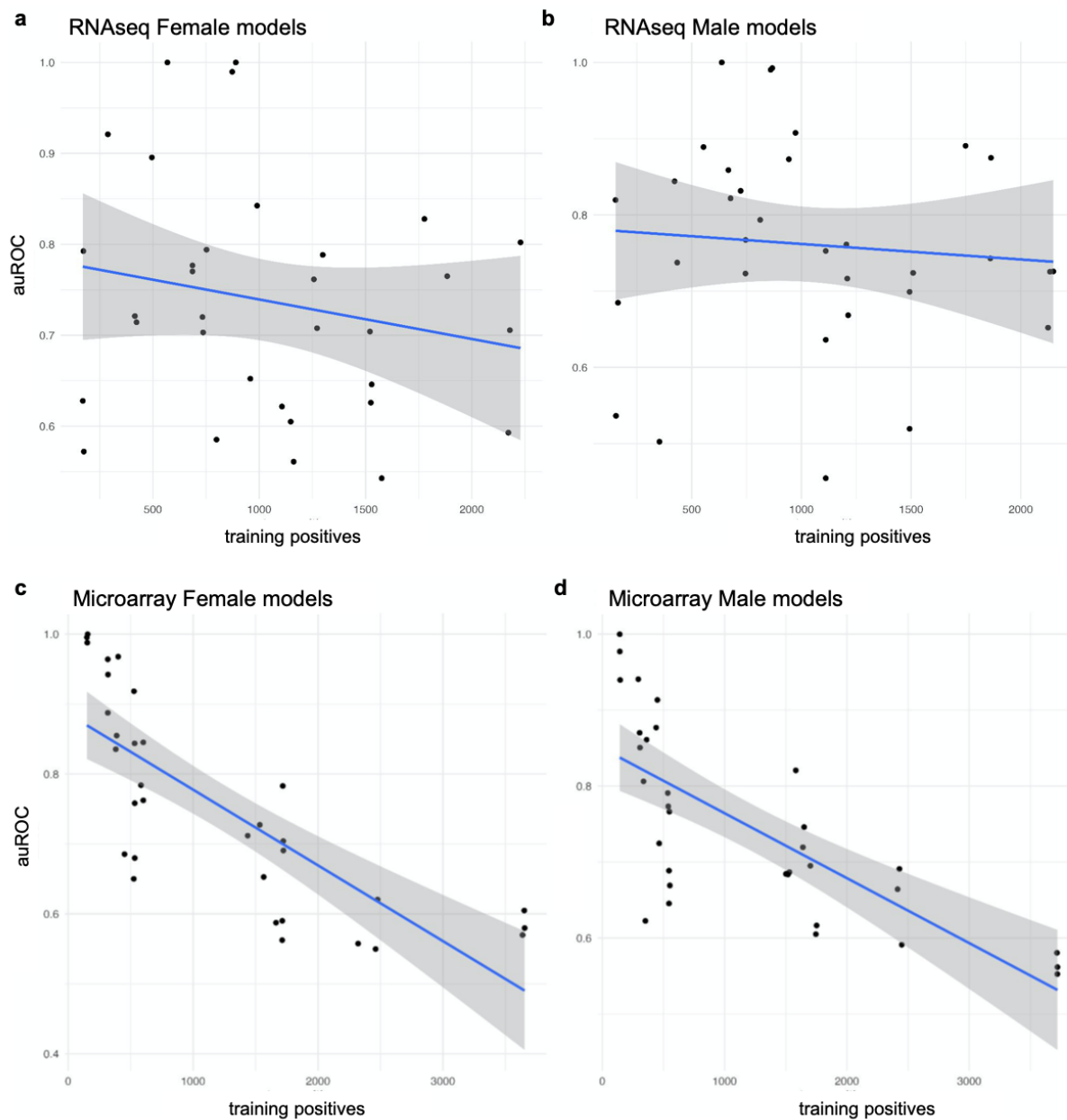


**Figure A3.7. Performance of microarray Female age range prediction models.** The heatmaps contain average  $\log_2(\text{auPRC}/\text{prior})$  performance of microarray models across 3 folds trained in Female samples on the age range labeled in the rows while evaluated as if the age range labeled in the column were the positive examples.

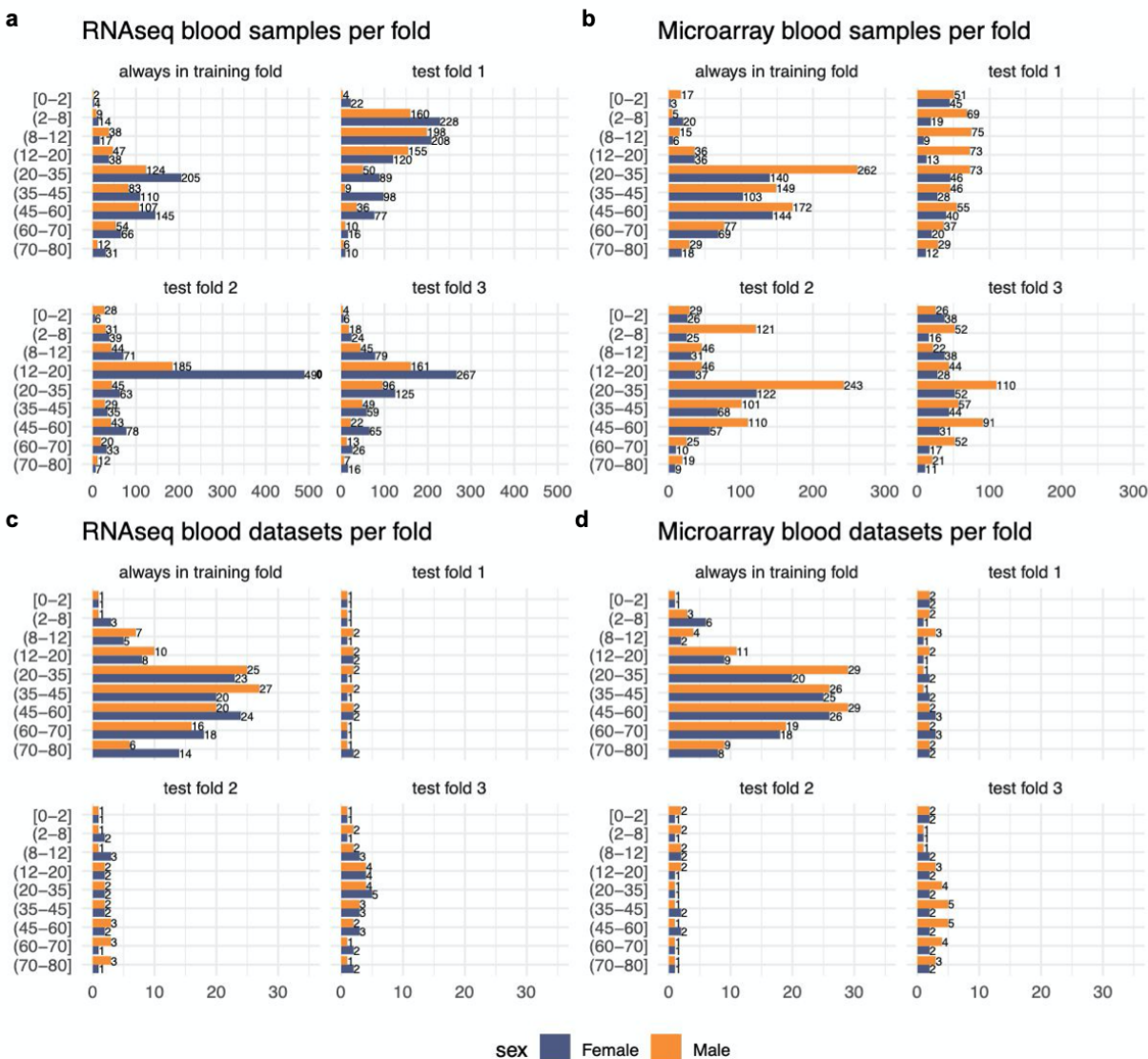
### Microarray Male models $\log_2(\text{auPRC}/\text{prior})$



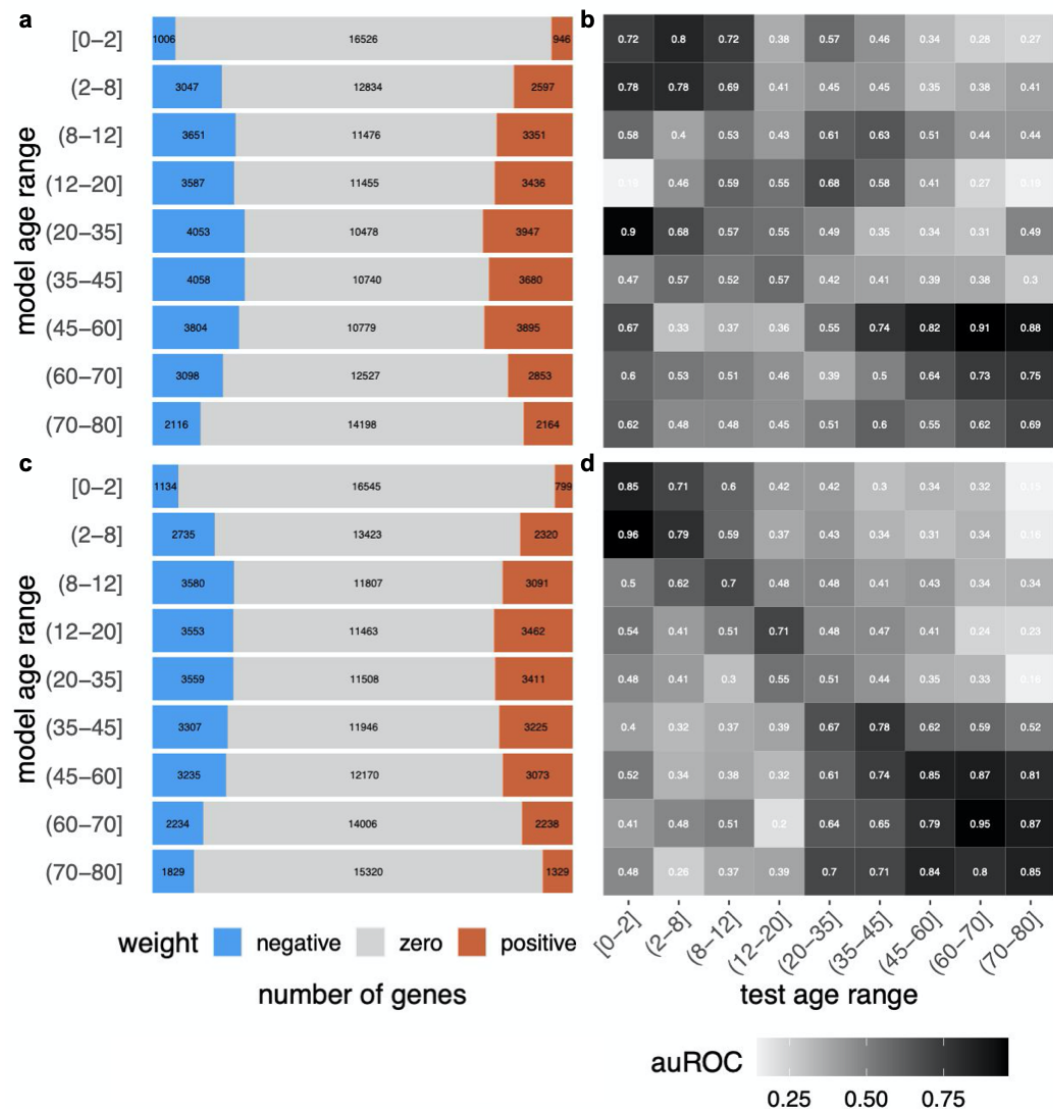
**Figure A3.8. Performance of microarray Male age range prediction models.** The heatmaps contain average  $\log_2(\text{auPRC}/\text{prior})$  performance of microarray models across 3 folds trained in Male samples on the age range labeled in the rows while evaluated as if the age range labeled in the column were the positive examples.



**Figure A3.9. Number of positive training examples vs auROC performance of age group models.** Scatterplot of the number of positive training examples vs the auROC performance of all RNA-seq and microarray Female and Male models.

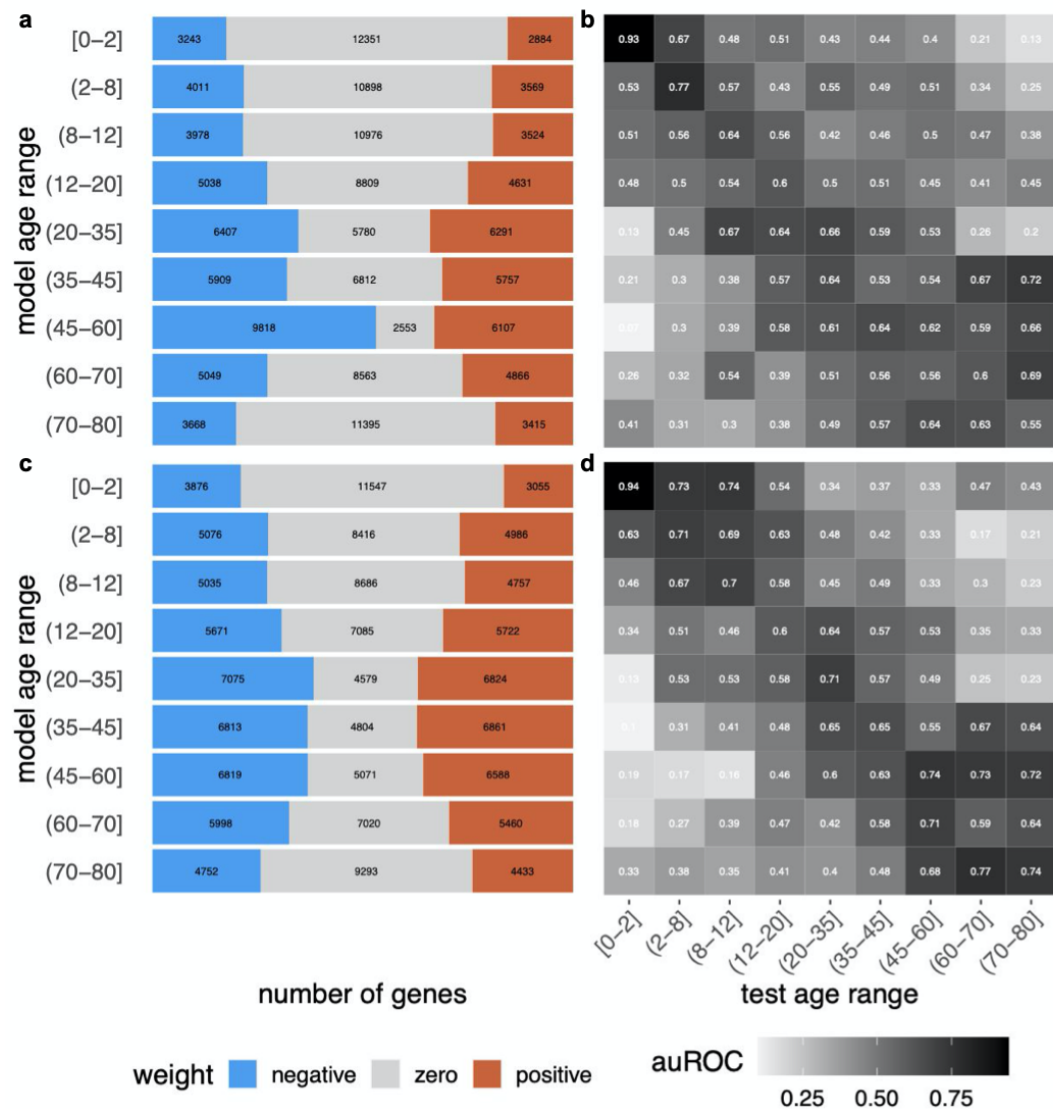


**Figure A3.10. Fold sizes for age range prediction models in blood samples.** The top barplots show the number of samples in each fold for models trained on blood samples in (a) RNA-seq and (b) microarray. The bottom barplots show the number of datasets in each fold for models trained on blood samples in (c) RNA-seq and (d) microarray.



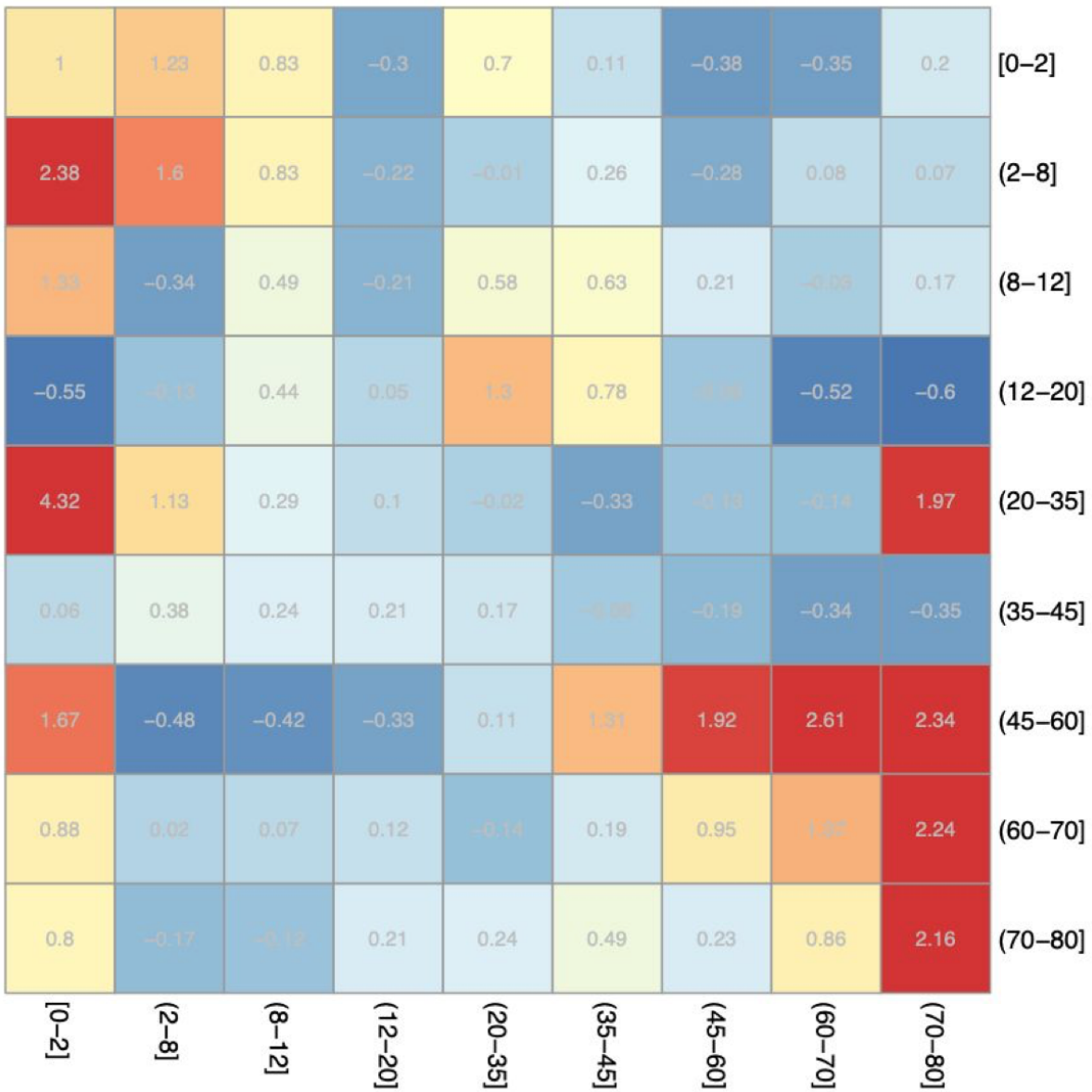
**Figure A3.11. Size and performance of RNA-seq age range prediction models in blood samples.** The stacked barplots show the distribution of positive, zero, and negative weights for the model with the median number of positives across the three folds for each age range in RNA-seq for blood samples from (a) Females and (b) Males. The heatmaps contain auROC performance of RNA-seq models trained on the age range labeled in the rows while evaluated as if the age range labeled in the column were the positive examples (c) Females and (d) Males.





**Figure A3.12. Size and performance of microarray age range prediction models in blood samples.** The stacked barplots show the distribution of positive, zero, and negative weights for the model with the median number of positives across the three folds for each age range in microarray for blood samples from (a) Females and (b) Males. The heatmaps contain auROC performance of microarray models trained on the age range labeled in the rows while evaluated as if the age range labeled in the column were the positive examples (c) Females and (d) Males.

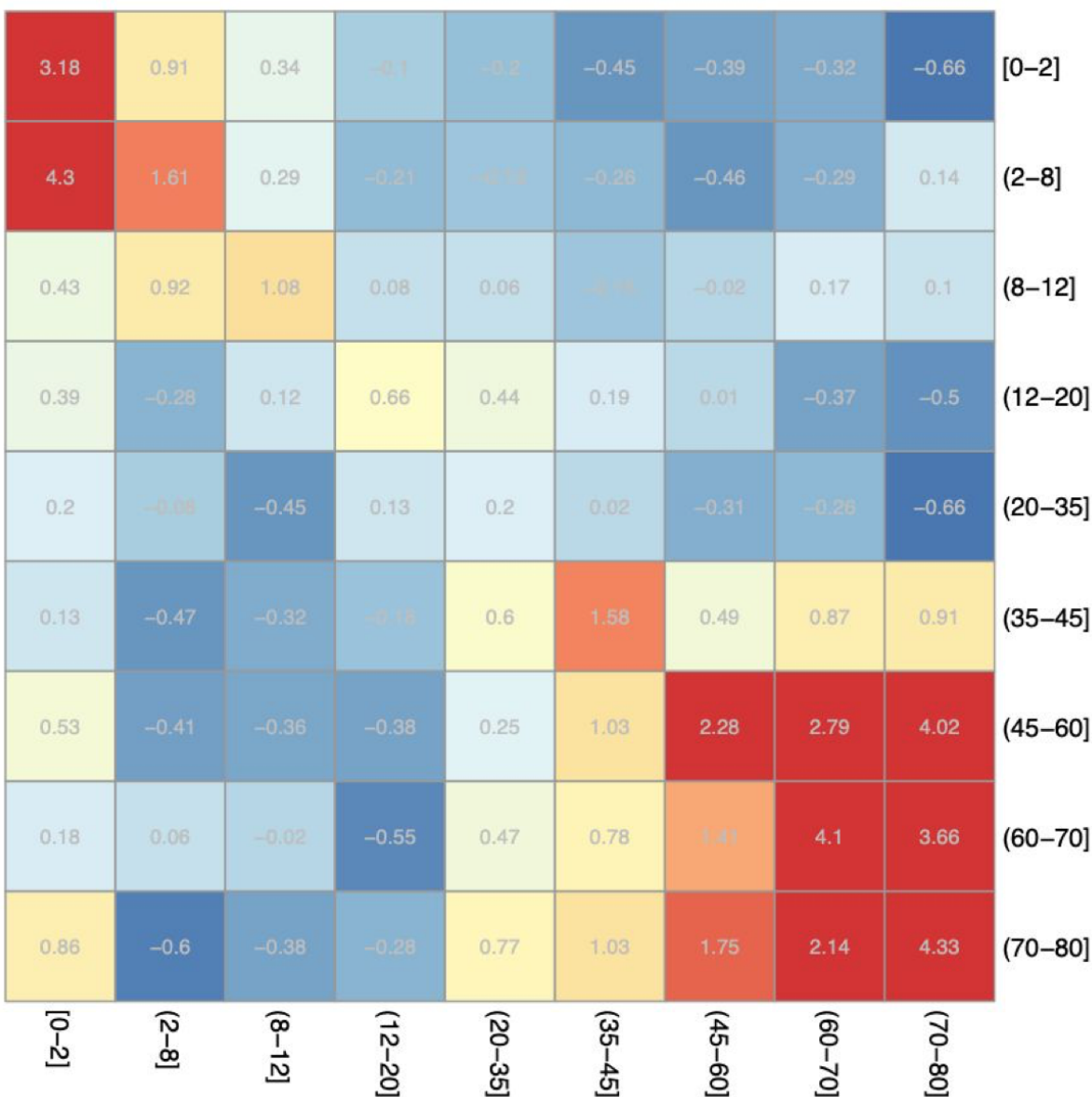
### RNAseq Female blood-only models $\log_2(\text{auPRC}/\text{prior})$



**Figure A3.13. Performance of RNA-seq Female age range prediction models for blood samples.** The heatmaps contain average  $\log_2(\text{auPRC}/\text{prior})$  performance of RNA-seq models across 3 folds trained in Female blood samples on the age range labeled in the rows while evaluated as if the age range labeled in the column were the positive examples.

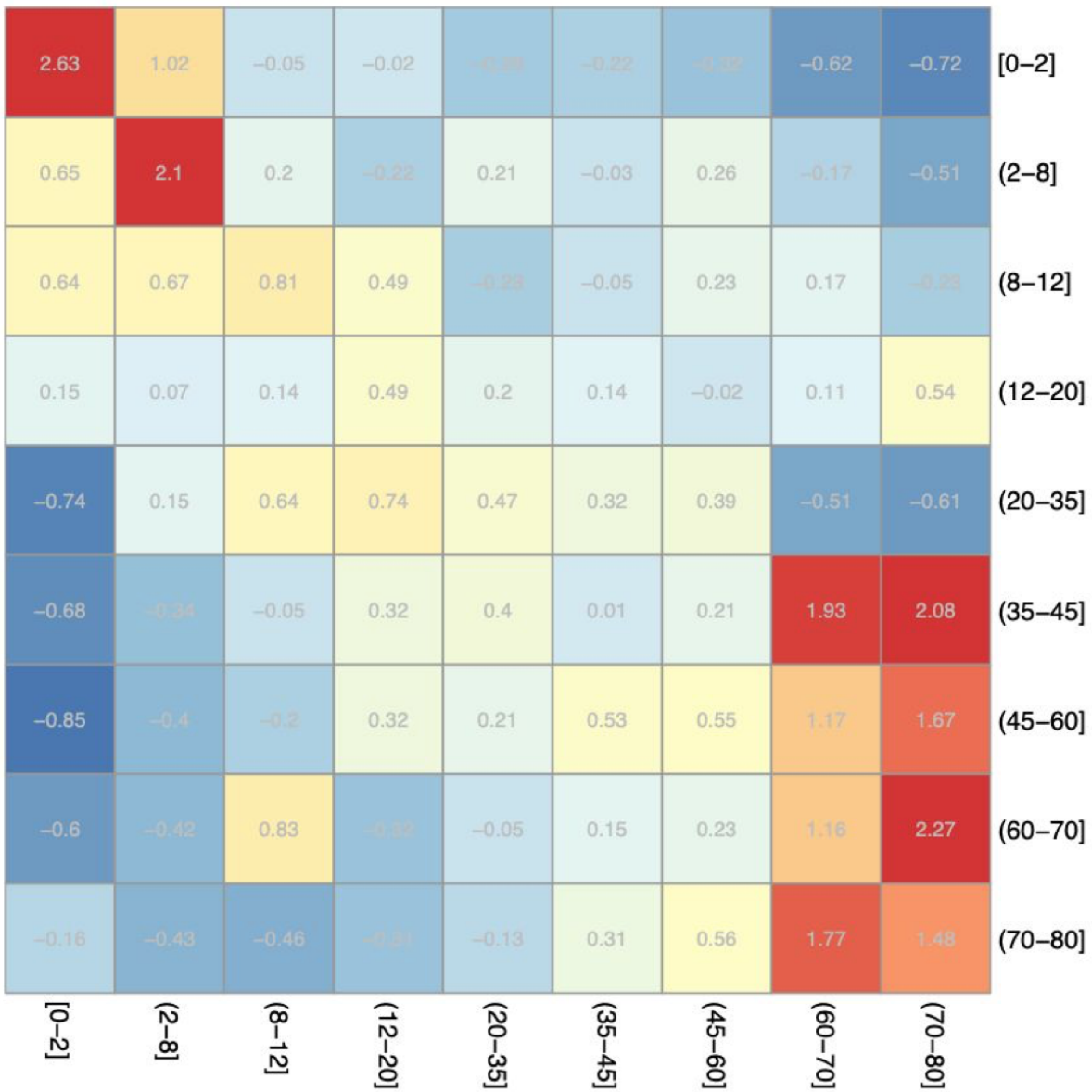


### RNAseq Male blood-only models $\log_2(\text{auPRC}/\text{prior})$



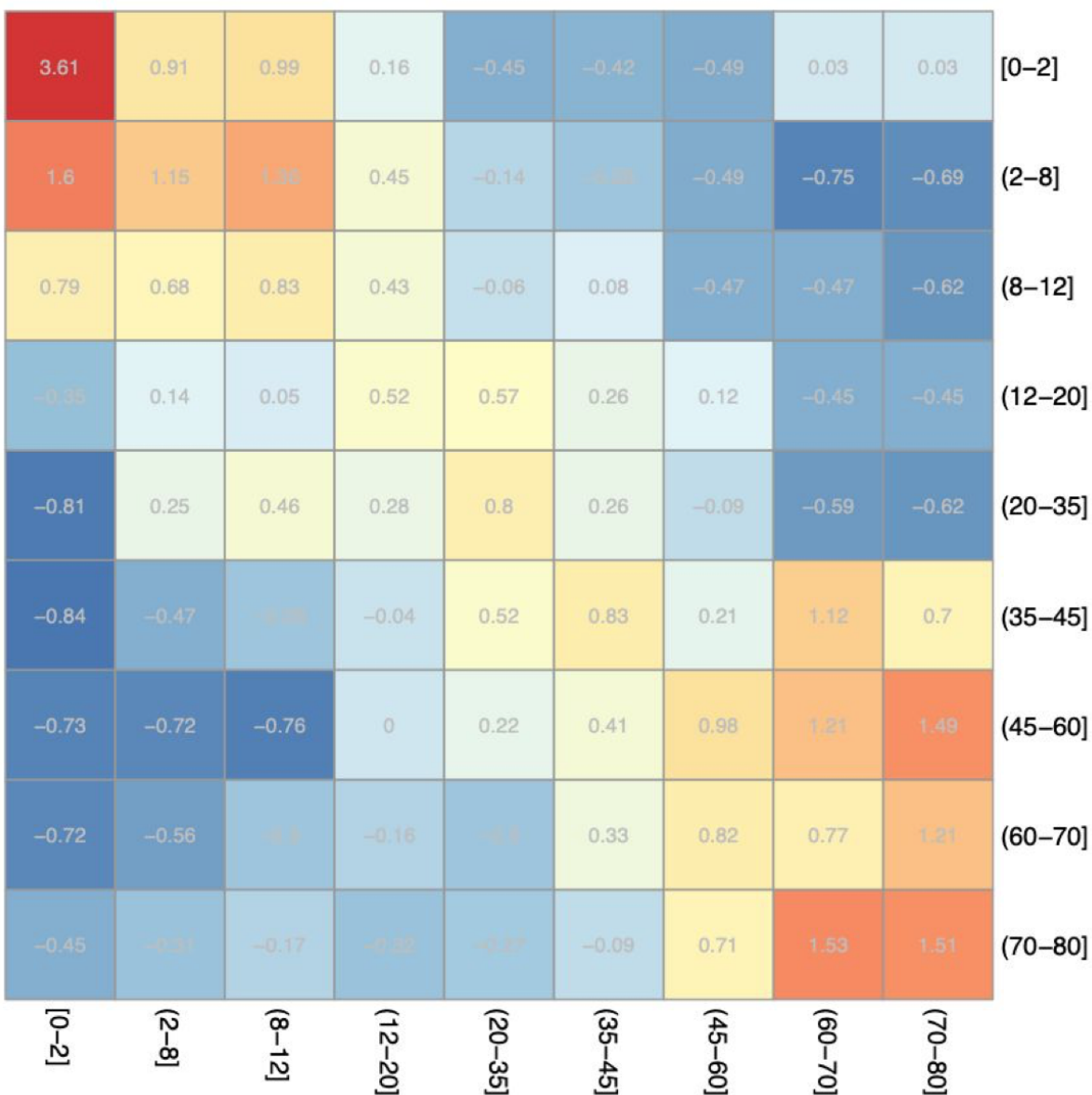
**Figure A3.14. Performance of RNA-seq Male age range prediction models for blood samples.** The heatmaps contain average  $\log_2(\text{auPRC}/\text{prior})$  performance of RNA-seq models across 3 folds trained in Male blood samples on the age range labeled in the rows while evaluated as if the age range labeled in the column were the positive examples.

### Microarray Female blood-only models $\log_2(\text{auPRC}/\text{prior})$



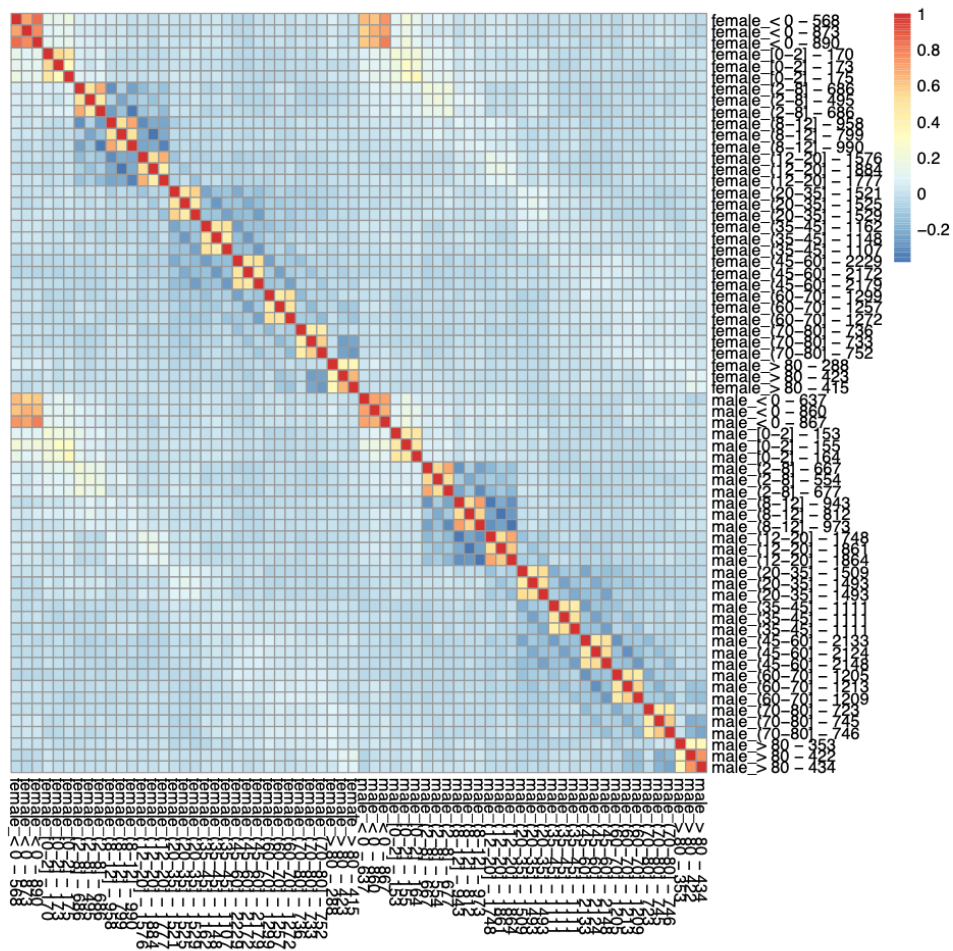
**Figure A3.15. Performance of microarray Female age range prediction models for blood samples.** The heatmaps contain average  $\log_2(\text{auPRC}/\text{prior})$  performance of microarray models across 3 folds trained in Female blood samples on the age range labeled in the rows while evaluated as if the age range labeled in the column were the positive examples.

### Microarray Male blood-only models $\log_2(\text{auPRC}/\text{prior})$



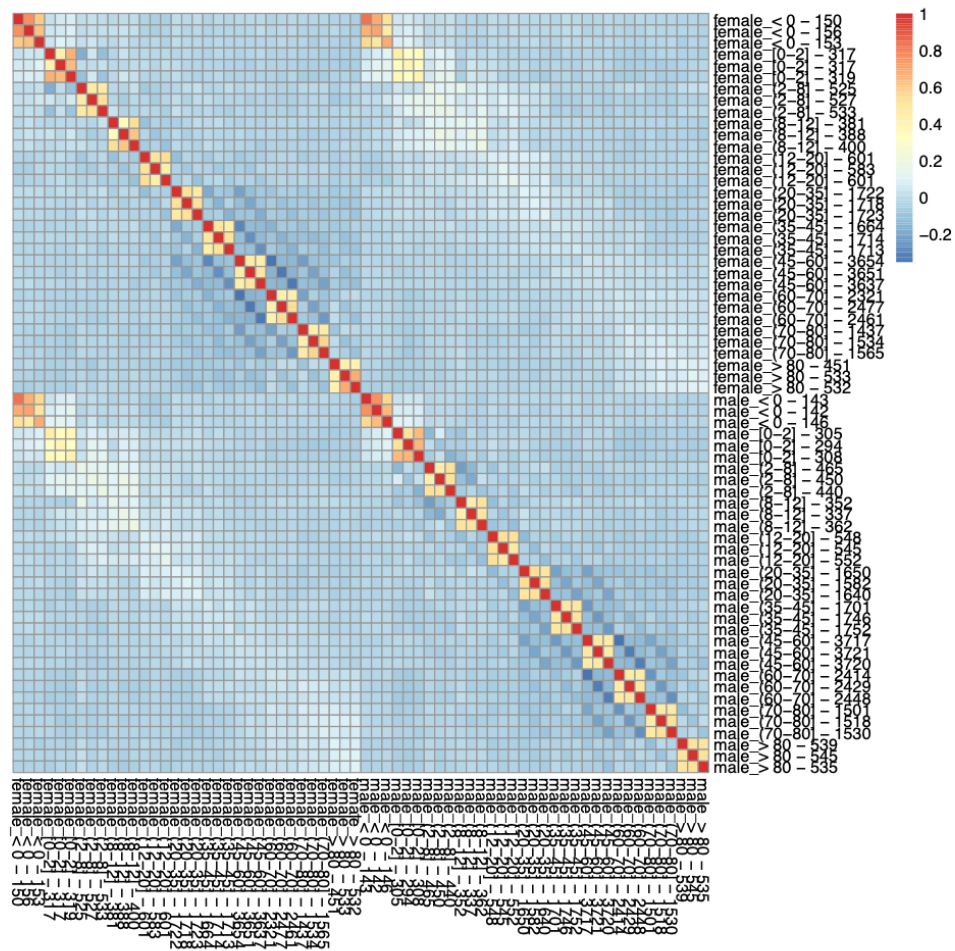
**Figure A3.16. Performance of microarray Male age range prediction models for blood samples.** The heatmaps contain average  $\log_2(\text{auPRC}/\text{prior})$  performance of microarray models across 3 folds trained in Male blood samples on the age range labeled in the rows while evaluated as if the age range labeled in the column were the positive examples.

Cosine similarity between rnaseq models



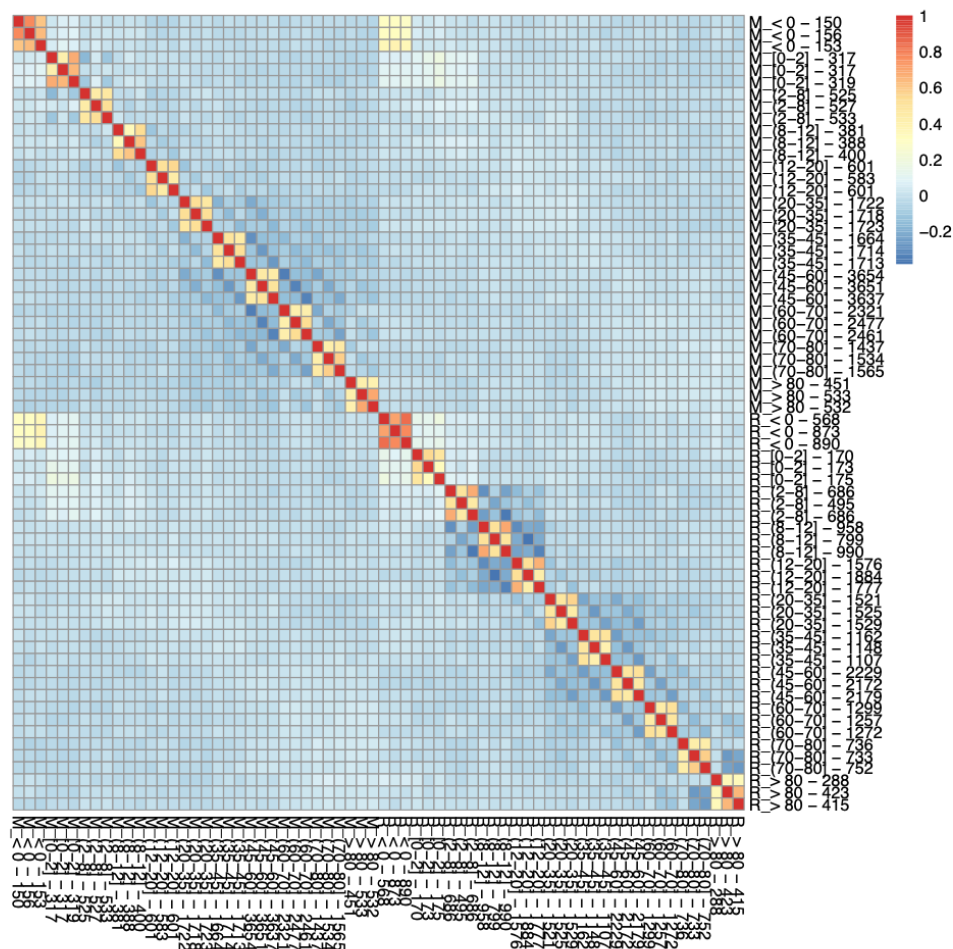
**Figure A3.17. Cosine similarity of RNA-seq model weights.** The heatmap shows the similarity between all RNA-seq models trained in both Females and Males. The number after the age range is the number of positive training examples.

Cosine similarity between microarray models



**Figure A3.18. Cosine similarity of microarray model weights.** The heatmap shows the similarity between all microarray models trained in both Females and Males. The number after the age range is the number of positive training examples.

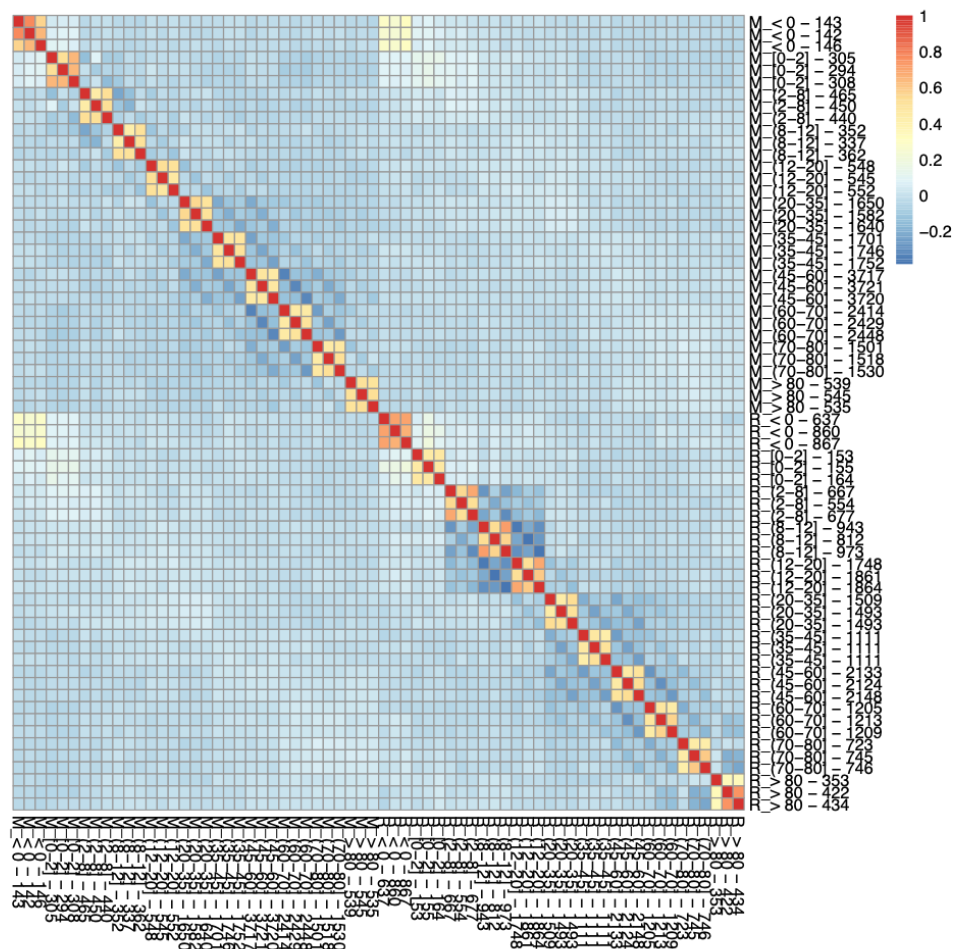
Cosine similarity between female models



**Figure A3.19. Cosine similarity of Female model weights.** The heatmap shows the similarity between all RNA-seq ('R\_' prefix) and microarray ('M\_' prefix) models trained in Females. The number after the age range is the number of positive training examples.

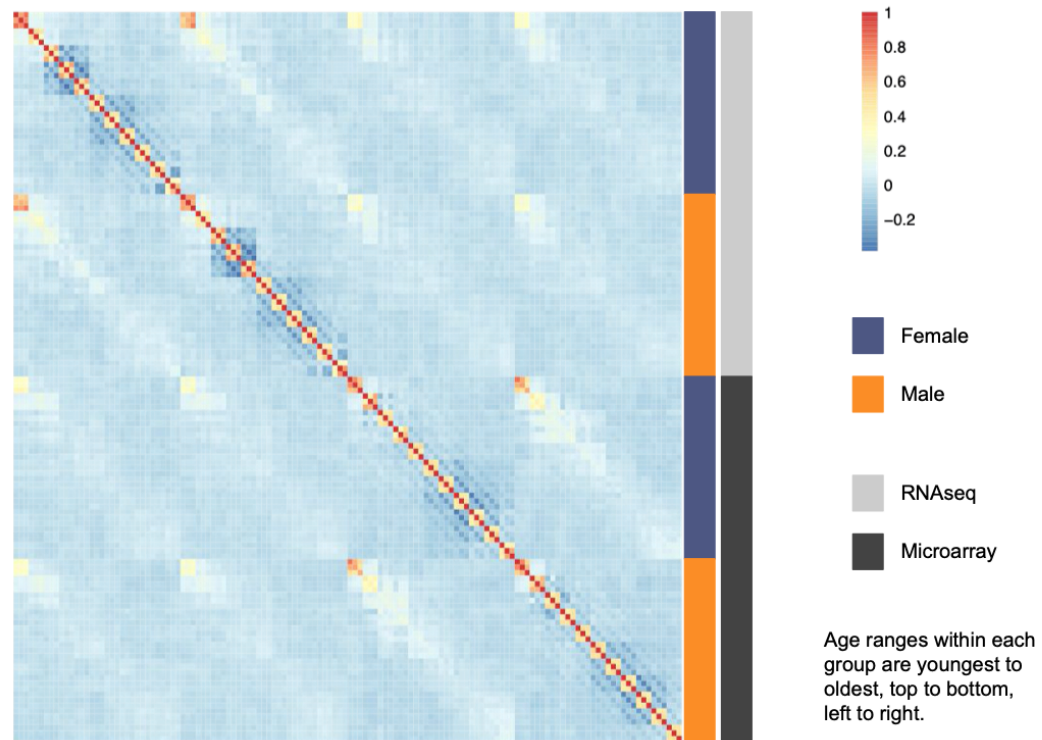


Cosine similarity between male models



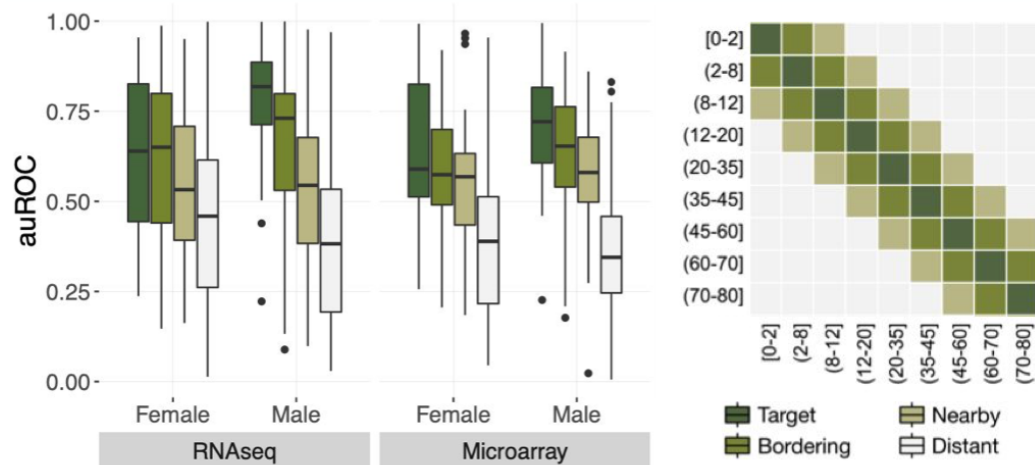
**Figure A3.20. Cosine similarity of Male model weights.** The heatmap shows the similarity between all RNA-seq ('R\_' prefix) and microarray ('M\_' prefix) models trained in Males. The number after the age range is the number of positive training examples.

Cosine similarity between all models

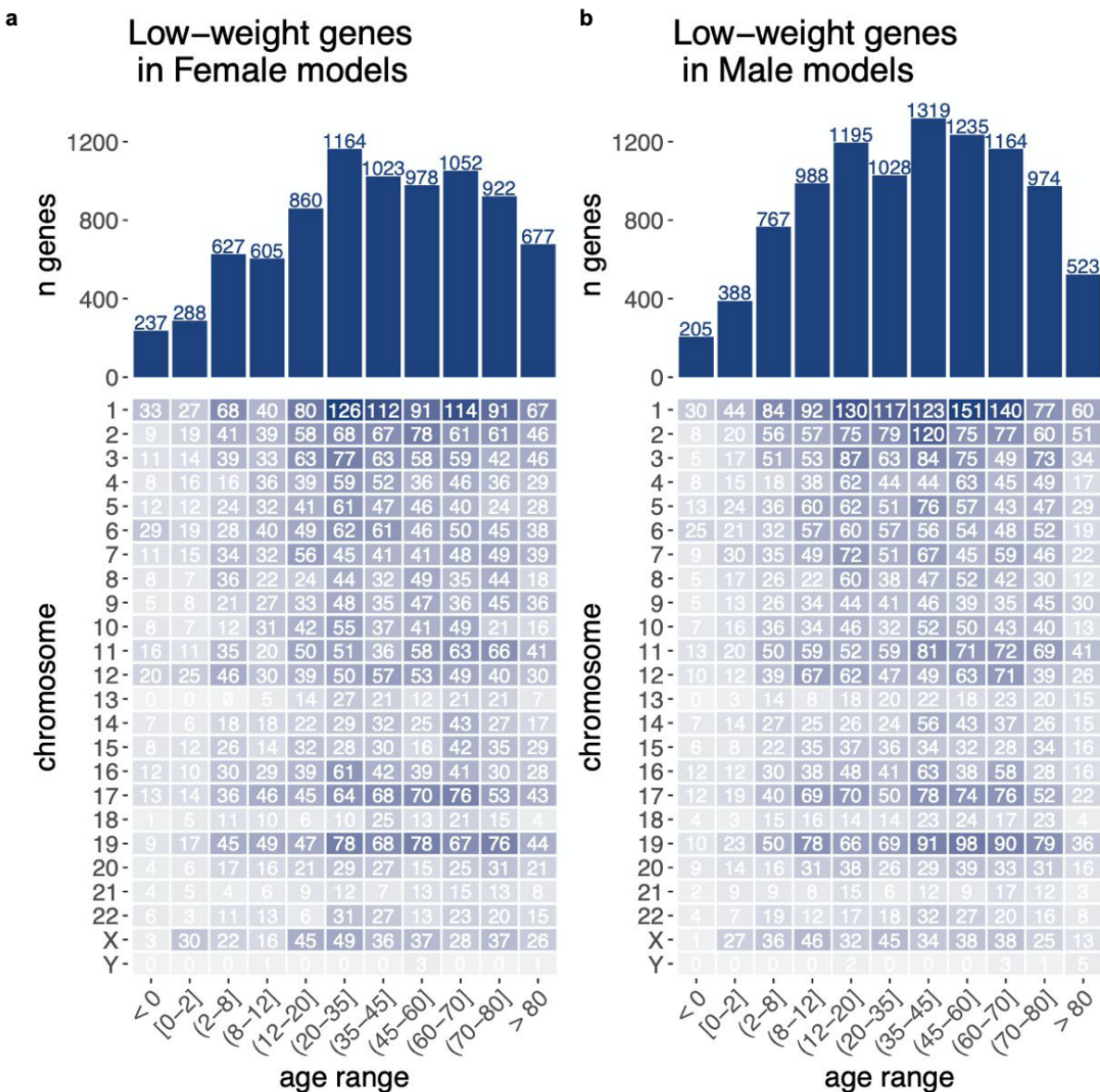


**Figure A3.21. Cosine similarity of Female and Male model weights.** The heatmap shows the similarity between all RNA-seq and microarray models trained in Males. The number after the age range is the number of positive training examples.

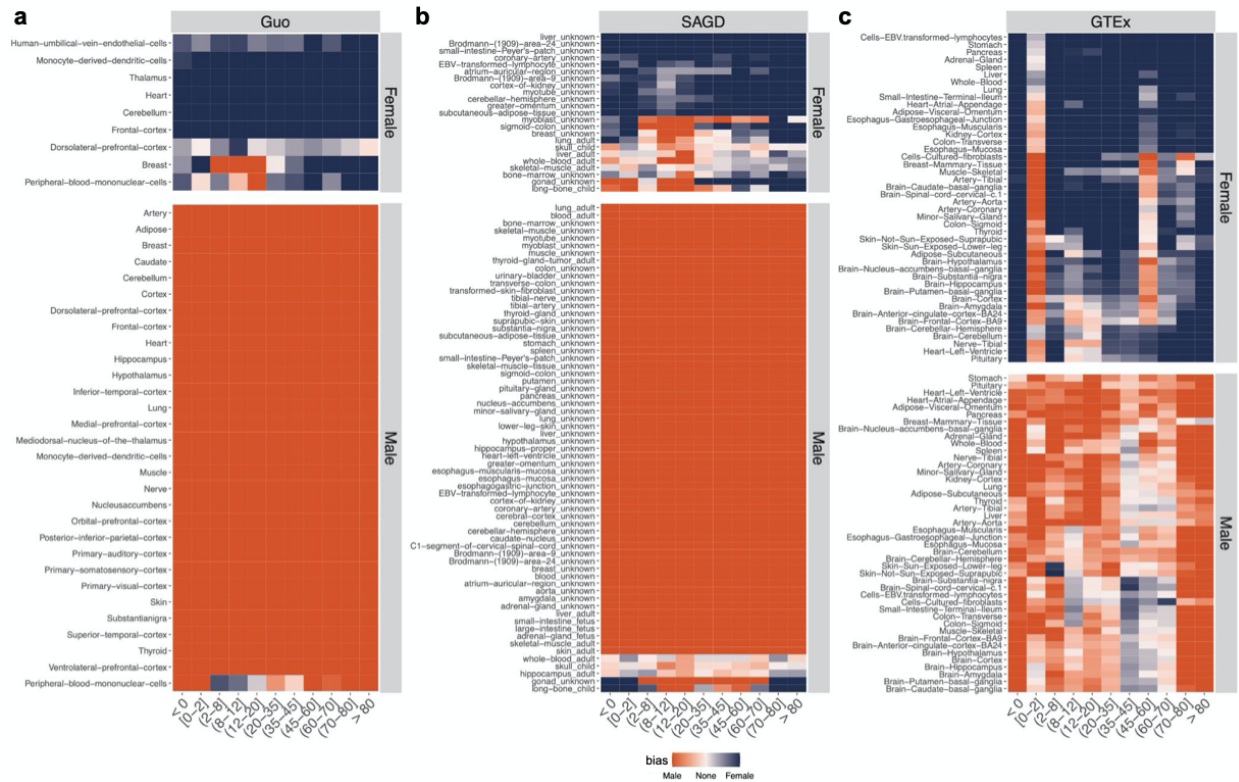




**Figure A3.22. Performance of blood sample models when evaluated on near and distant age ranges.** Boxplot of auROC performance of all RNA-seq and microarray Female and Male models when considering target, bordering, nearby, and distant age ranges as positive examples (key on right).



**Figure A3.23. Number of age-biased genes across age ranges in each sex.** (a) The table displays the number of age-biased genes on each chromosome per age range in Females and the barplot shows the total number of age-biased genes per age range. (b) The table displays the number of age-biased genes on each chromosome per age range and the barplot shows the total number of age-biased genes per age range in Males.



**Figure A3.24. Enrichment of sex-differentially expressed genes sets from previous studies in our age-stratified sex signatures.** Female- and Male-bias enrichment of gene sets from previous studies. Heatmaps show enrichment scores for (a) Guo et al gene sets, (b) SAGD gene sets and (c) GTEx gene sets.



**Figure A3.25. Enrichment of experimentally-derived genes sets from in our age-stratified sex signatures.** Female- and Male-bias enrichment of experimentally-derived gene sets. Heatmaps show enrichment scores for (a) a representative set of metabolism-related GO biological processes, (b) metabolic diseases, and (c) metabolic phenotypes.

## **CHAPTER 4: DISCOVERING ANALOGOUS GENES, PHENOTYPES, AND CONDITIONS ACROSS HUMAN AND MODEL SPECIES USING MACHINE LEARNING**

### **Background**

Model organisms are commonly used to investigate underlying mechanisms and discover therapeutic opportunities of human complex traits and diseases. However, animal models have been shown to sometimes be faithful models [1] of human biology and sometimes they are poor mimics of human biology [2,3]. Whether an animal model is a good model system depends on disease/trait and species, as changing function, regulation, and differences in redundancies [4–6] can cause unexpected divergences in biological processes and phenotypes. Drugs in development must first be tested in animal models before they are allowed to enter Phase I clinical trials, and using poor models can be especially costly. A recent study of developmental drug candidates entering Phase I in the period of 2011–2020 found the likelihood of approval to be about 8% [7]. In rare cases of disaster, compounds that showed no toxicity in other species have caused multiple organ failure and even death in humans in clinical trials, even at much lower doses than those tested in animal models [8]. This illustrates the dire need for methods that can improve the translation of functional results from one species to another.

Choosing the best model species with the correct experimental conditions is clearly quite difficult. Genetic background, tissue, developmental stage, and environmental factors are all crucial considerations. The ideal animal model for studying a specific aspect of human biology should not only display the desired phenotype but the underlying molecular mechanism should also be as similar as possible. Current

computational methods that map related phenotypes across species rely on semantic similarity of phenotypic descriptions [9], or consider the number of shared homologous genes that are annotated to each phenotype [10]. Semantic similarity methods ignore the genetic context of the traits and phenotypes completely by depending only on the text description of the phenotype, while methods that rely on homologous gene overlap fail in many cases due to our incomplete knowledge of the genes associated with any given trait or phenotype. However, there are well over a million publicly-available transcriptomes across multiple model organisms and humans that help overcome these challenges of descriptive information and incomplete knowledge. Specifically, these data can be leveraged to find expression profiles that are able to mimic the transcriptomic landscape of a given trait, disease, or treatment response captured in a human sample, which should lead to finding the ideal experimental setting for studying the human biomedical context of interest.

Many previous studies have used gene expression profiles to make comparisons across species [11]. These studies typically use differential expression [12–14] or some similarity metric over absolute expression [15] to identify analogous samples. However, many complex tissues, traits, and diseases have shared expression modules [16–18]. Hence, to truly map the most similar contexts across species, we must endeavor to use not all molecular features but just those that are specific to the context of interest. A supervised learning approach allows us to provide a machine learning classifier with positive examples of the expression profiles of a particular context, say, a disease, that need to be contrasted with negative examples of profiles from other (unrelated) diseases. The classifier is then able to automatically learn context-specific features from

the expression data, which can then be used to pick out samples from other species in which the context-specific genes are highly expressed. In this study we develop a data-driven method to prioritize transcriptomes, and predict experimental settings in model organisms for studying particular facets of human biology and disease.

## **Results**

### **Common gene feature sets across species**

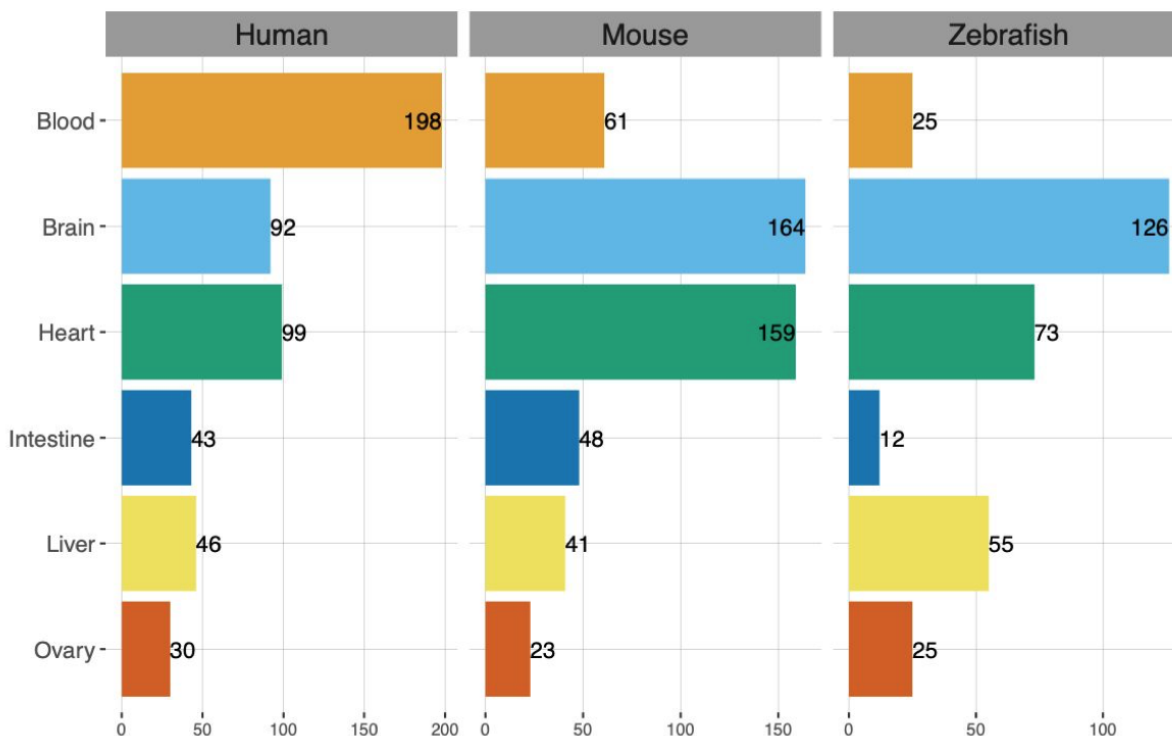
The first challenge in training machine learning models on a set of transcriptomes in one species to make predictions in another is that different species do not share the same set of genes. Therefore, a common set of features must be chosen so that a model is able to learn the weights of the features in one species and predict using features in the other. The easiest method to subset to a common set of genes across species is to retain just those that are one-to-one orthologs of each other. One-to-one orthologs are genes with a direct evolutionary relationship to only one other gene in the other species [19]. However, a large portion of genes are not part of a one-to-one orthologous relationship in any given pair of species. So, using one-to-one orthologs immediately introduces a loss in the amount of information from transcriptomes that can be used for this task. Therefore, to increase the number of genes informing the model, we developed a feature set created by two ways of combining the expression of genes within orthologous groups (OGs) in each species. One is by averaging the expression of all genes in the OG and the other is by retaining only the maximum expression value of all genes in the OG. We use the same two methods to combine expression of genes that belong to the same biological process (from Gene Ontology; GO [20]). We use GO biological processes to account for cases where, in the same biological context, the

same functional modules (biological processes) are perturbed but via different member genes in different species. Further, previous studies have reported increased performance in gene expression classification tasks by grouping gene expression into pathway expression [21,22]. The rationale for combining gene expression within groups by averaging and by retaining the maximum is that while the average is a good representation of all genes in a given group, it is not robust to our incomplete knowledge about which genes are part of which groups in any species. This yields five feature sets for evaluation: one-to-one orthologs (OnetoOne), orthologous groups averaged (OGs-avg), OGs maximum (OGs-max), GO biological processes averaged (GO-avg), and GO biological processes maximum (GO-max).

### **Tissue-labeled samples from multiple species**

Although our goal is to map a wide variety of analogous biological contexts across species, it is difficult to systematically evaluate these mappings in every biological context based on a large gold standard. The context for which enough sample information is available and easily comparable across multiple species is that of sample tissue-of-origin. Therefore, we first manually curated tissue labels for human, mouse, and zebrafish RNA-seq samples (**Fig. 4.1**). We chose the six most common tissues — blood, brain, heart, intestine, liver, and ovary — across all three species based on text mining results that were subsequently verified manually to ensure accuracy.





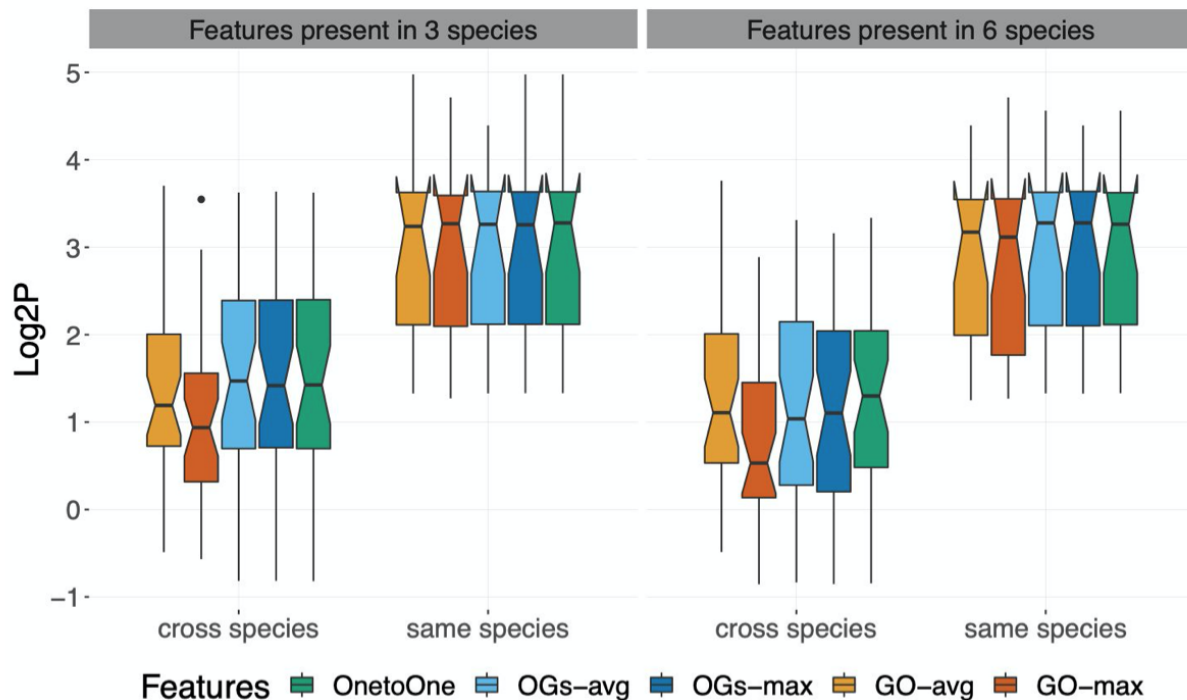
**Figure 4.1. Number of samples in each tissue across species.** The number of samples in human, mouse, and zebrafish for each of six tissues.

### Proof of concept with six tissues

Using these curated tissue-labeled samples from the three species, we trained a logistic regression model for each tissue in a given species using each of our five gene feature sets and then used each model to make predictions on held out samples from the same species and on all samples in other species. We also tested the feasibility of extending this method to include commonly-used organisms we could not get enough tissue-labeled samples for: fly, worm, and yeast. We did this by building feature sets that combine gene expression across only OGs and GO processes that are common to all six species, meaning each OG or GO feature needed to include at least one gene from each species.

Unsurprisingly, the expression-based tissue prediction models were accurate in classifying samples from the same species to the right tissue regardless of feature set (**Fig. 4.2**). When using feature sets common to 3 species, all five sets had almost identical performance when predicting on samples from the same species, but GO features performed worse than orthologous group feature sets in cross-species predictions. GO-max performance was notably lower than GO-avg.

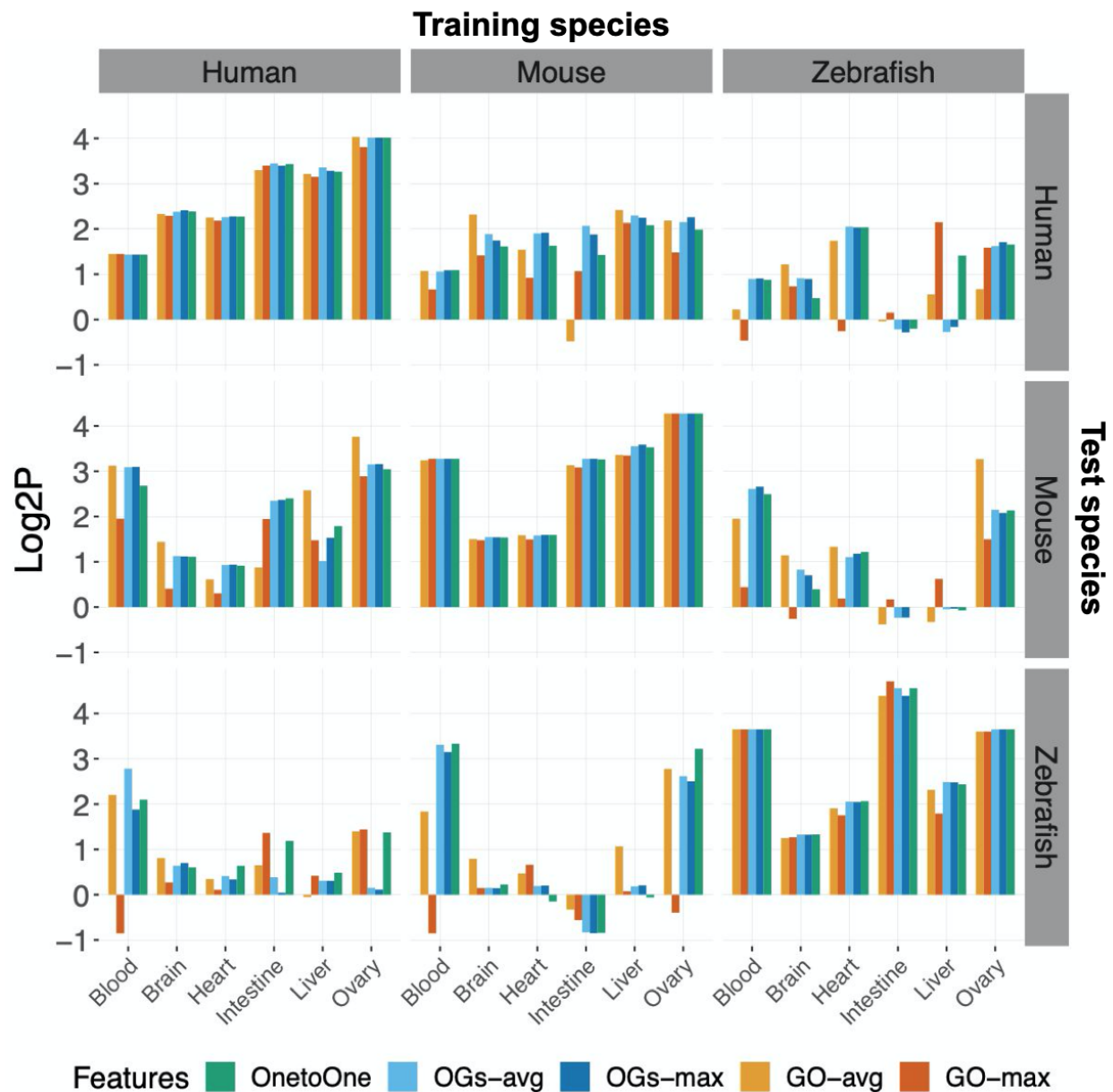
When we tested feature sets common to 6 species, performance across the board dropped slightly (**Fig. 4.2**). Predictions within species were still more accurate than predictions across species, but instead of an identical performance across feature sets, we observe that GO features have slightly poorer performance than OG-based feature sets. In cross species predictions, OnetoOne orthologs seems to have the highest performance. GO-avg shows similar performance to the OGs-avg and OGs-max, but GO-max is significantly worse. Overall, the ability to train these models using feature sets that cover six evolutionary distant species with only a small performance cost is worth it, so we use feature sets common to all six species for results in the remainder of the chapter.



**Figure 4.2. Performance of tissue classification models using each feature set built across three and six species.** The boxplots show the prediction performance ( $\text{Log2P} = \log_2(\text{auPRC}/\text{prior})$ ) of each model trained on a given tissue in each species using a different feature set common to three species (human, mouse, zebrafish) and six species (human, mouse, zebrafish, fly, worm, yeast) on samples from a different species (cross species) and the same species.

Ideally, methods to map analogous transcriptomes across species would demonstrate robust performance regardless of training or test species. We observed that the similar performance across all five feature sets within a species holds true across tissues in each species (**Fig. 4.3, diagonal plots**). When making cross-species predictions, there is more variability in the performance of different feature sets (**Fig. 4.3, off-diagonal plots**). Across all tissues, cross-species predictions involving zebrafish as the training or testing species tended to have lower performance, and had the most cases where the models had worse than random prediction performance (bars below zero). The feature set with worse-than-random performance most often was GO-max. The three

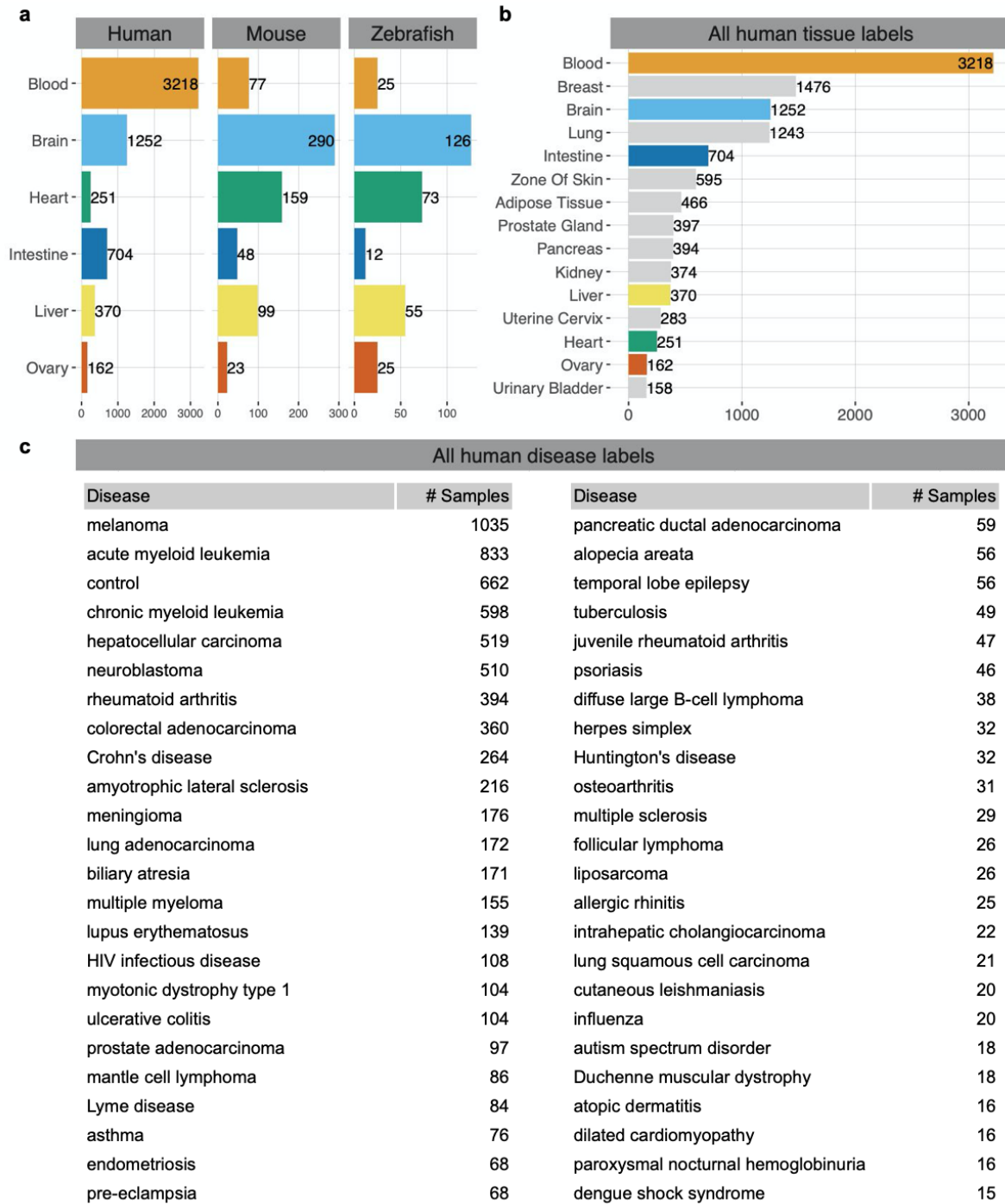
feature sets based on orthology (OnetoOne, OGs-avg, and OGs-max) performed similarly to each other in all tissues across species except in a few cases.



**Figure 4.3. Performance of tissue classification models across tissues and train/test species.** The bars show the prediction performance ( $\text{Log2P} = \log_2(\text{auPRC}/\text{prior})$ ) of each tissue classification model trained in a given species (Training species) and used to make predictions in another species (Test species). The median performance value from 5-fold cross validation is plotted for models that were trained and tested in the same species.

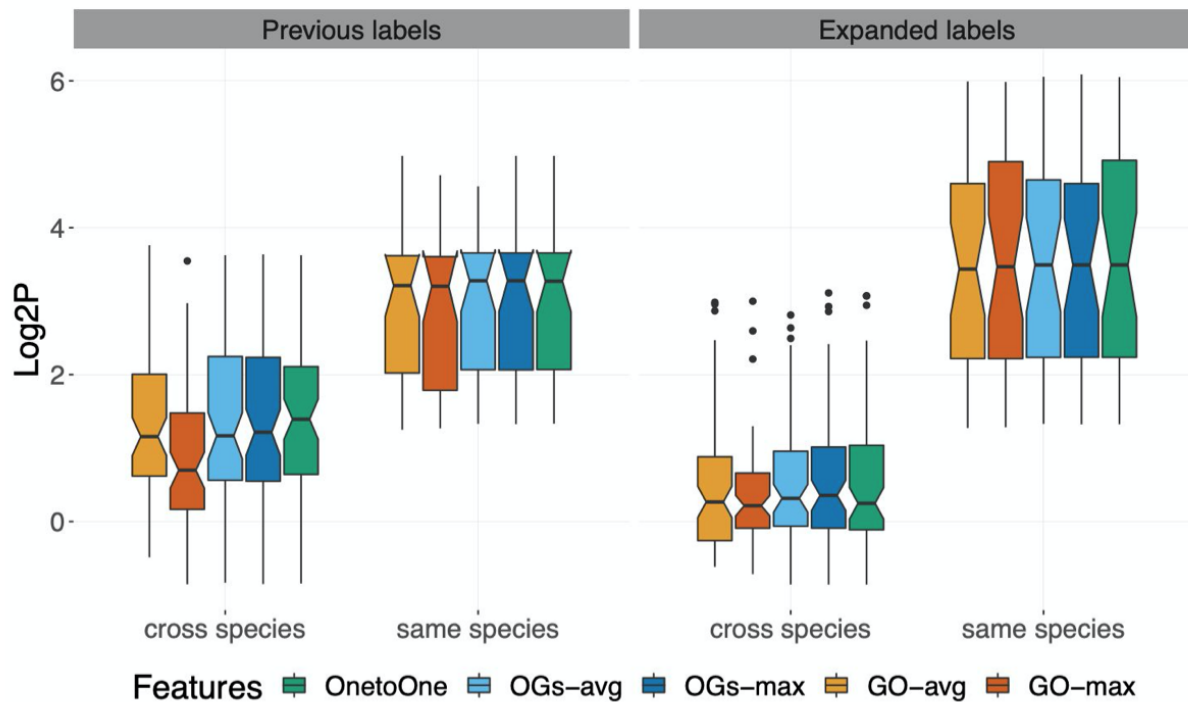
### **Performance on an expanded set of diverse tissues**

Though we observed good performance in most feature sets in most settings, the six tissues we began with — blood, brain, heart, intestine, liver, and ovary — are disparate, making for an easier classification problem compared to what we will encounter with real data from public databases. Therefore, we next expanded the human tissue labels using manually curated transcriptomes from the TissueNexus database [23]. These labels increased the number of samples from our original tissues in humans by at least 2.5 times, and added nine other tissues to evaluate performance with and use as negative examples in training (**Fig. 4.4**). We also manually curated sample disease labels for 47 diseases along with as many healthy/control samples as possible in human datasets.



**Figure 4.4. Expanded tissue labels and disease labels.** (a) Number of samples in all common tissues across species with added TissueNexus labels for human. (b) Number of human samples with tissue labels after expansion. (c) Number of human samples annotated to each disease.

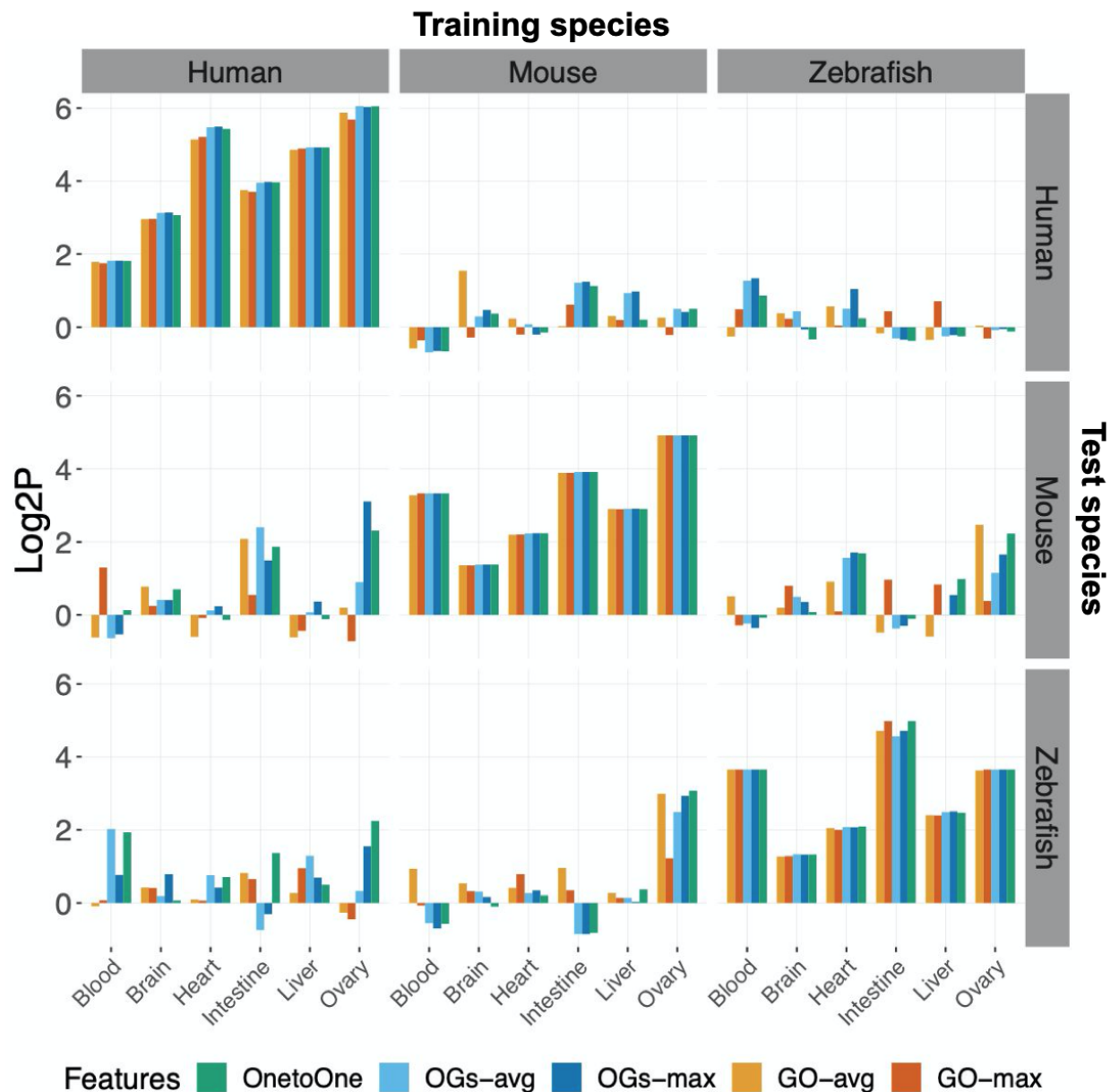
With our expanded tissue label set, we see that overall, predictions within species improve, but cross-species prediction performance lowers significantly (**Fig. 4.5**). The number of times cross-species prediction performance drops below random (less than 0) is much higher. The general pattern of all feature sets performing similarly within species is conserved using the expanded labels, but the orthology-based feature sets are all equally bad in cross-species predictions.



**Figure 4.5. Performance of tissue classification models on previous and expanded labels.** The boxplots show the prediction performance ( $\text{Log2P} = \log_2(\text{auPRC}/\text{prior})$ ) of each model trained on a given tissue in each species using previous and expanded sample labels (see Fig. 4.4) and different feature sets on samples from a different species (cross species) and the same species.

Within human samples, human tissue classification models improved prediction performance for all tissues (**Fig. 4.6**). However, they now have consistently lower performance when classifying mouse and zebrafish samples. Using the previous label set, only once did a human model perform worse than random, using the GO-max

feature set. There are several instances using the expanded tissue labels. Mouse and zebrafish tissue classification models suffered lower performance on all human tissues, with worse than random performance much more often than on the previous label set.



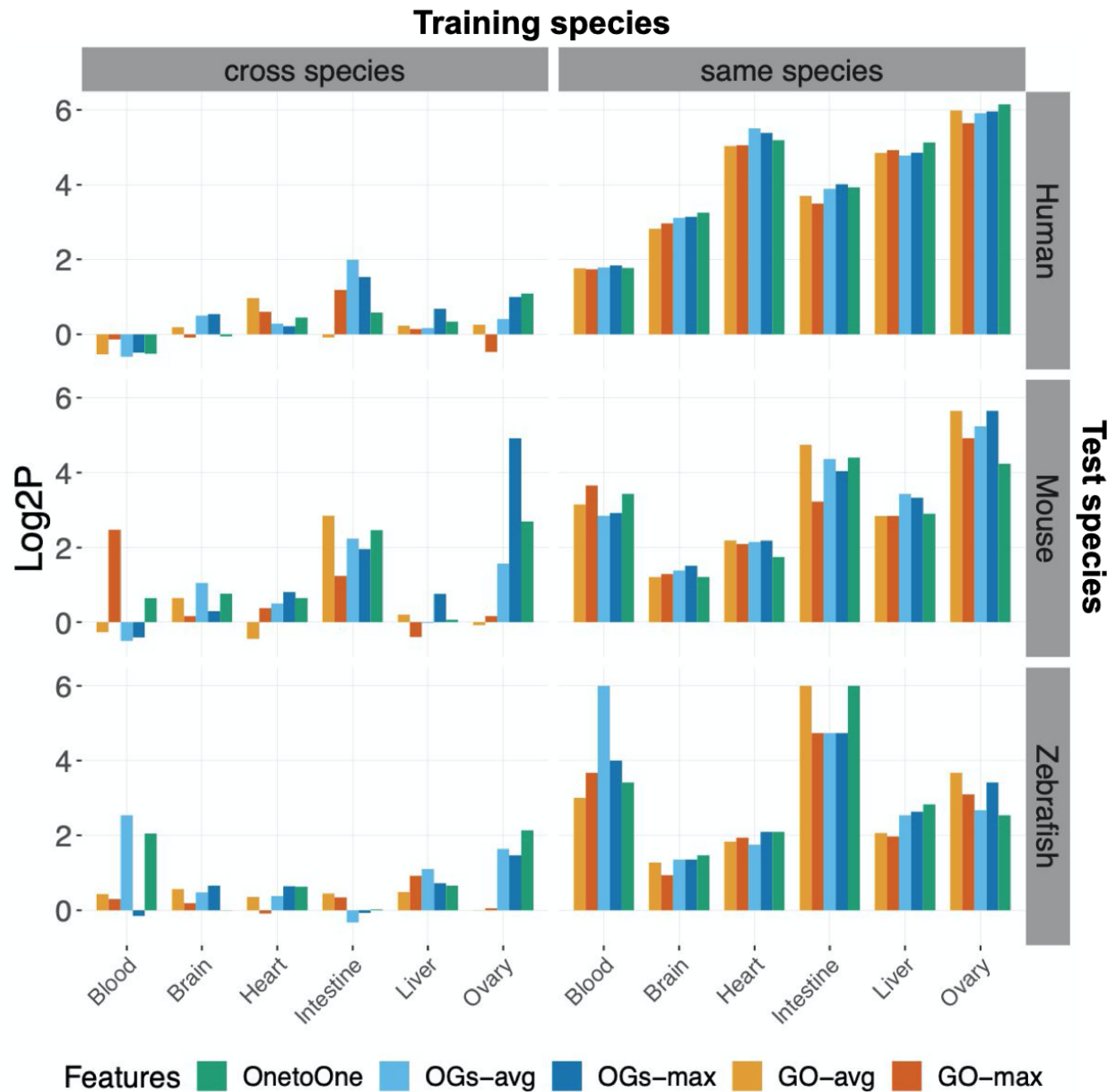
**Figure 4.6. Performance of tissue classification models across tissues and train/test species using the expanded set of tissue labels.** The bars show the prediction performance ( $\text{Log2P} = \log_2(\text{auPRC}/\text{prior})$ ) of each tissue classification model trained in a given species (Training species) and used to make predictions in another species (Test species) using the expanded tissue label set (see Fig.



**Figure 4.6. (cont'd)**

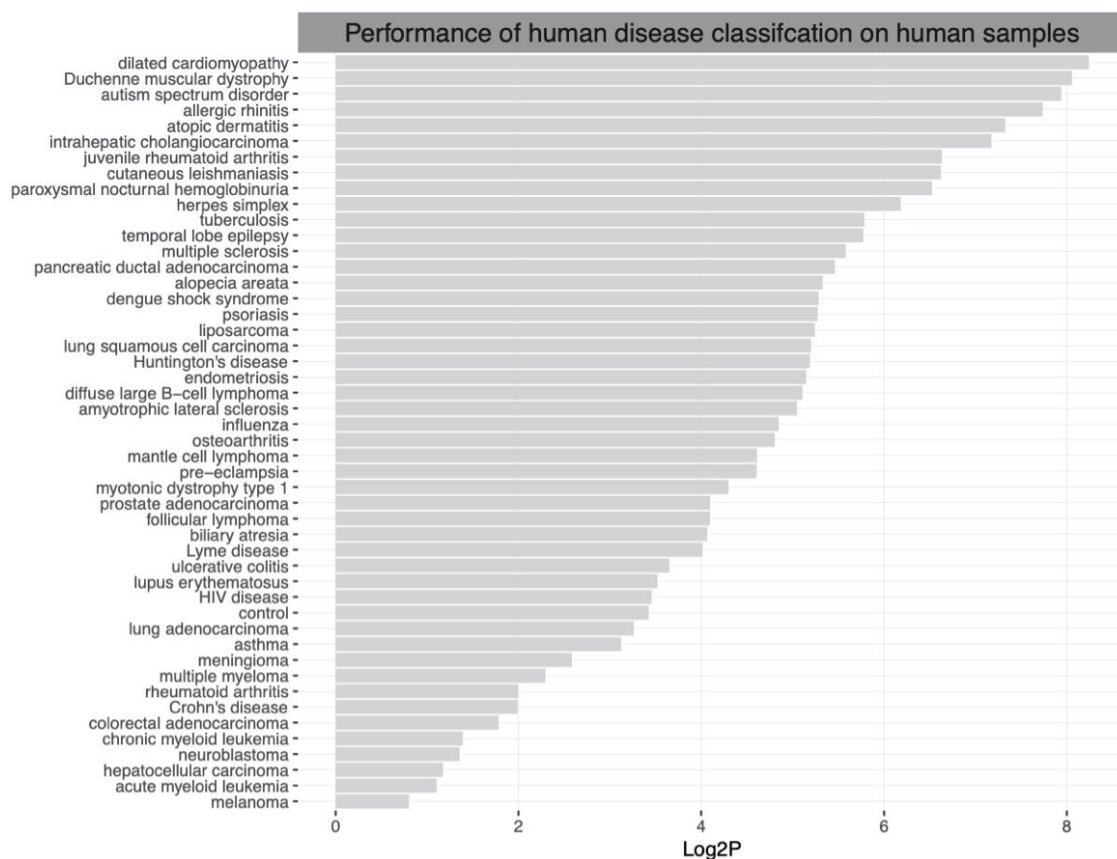
4.4). The median performance value from 5-fold cross validation is plotted for models that were trained and tested in the same species.

In an attempt to improve the significantly lower performance of cross-species classification using the label set with many more human tissues, we combined samples across two species to train tissue classification models and made predictions in the held-out species (**Fig. 4.7**). This was successful to some extent. The model performances drop below random (zero) at a lower rate, but in many cases the performance seems to be a weighted average of the combined training species' individual performances on the test species.



**Figure 4.7. Performance of combined-species tissue classification models across tissues using the expanded set of tissue labels.** The bars show the prediction performance ( $\text{Log2P} = \log_2(\text{auPRC}/\text{prior})$ ) of each tissue classification model trained in two species and used to make predictions in the held-out species (Test species) using the expanded tissue label set (see Fig. 4.4). The median performance value from 5-fold cross validation is plotted for models that were trained and tested in the same species.

Finally, we trained classification models for 47 human diseases. Within human samples, these models perform well across the board, with some showing extremely accurate performance (**Fig. 4.8**). Since we do not have disease labels for animal models, we used these models to make predictions on all other samples and manually inspected the top-ranked samples. Despite excellent performance in humans, cross-species predictions are still poor even in the best models. Many top-ranked samples do not have a recognizable connection to the disease, but worse is the number of single cell samples that are assigned high probabilities by the disease model. All of our curated labels are bulk transcriptomic samples, so the model does not train on single cell data.



**Figure 4.8. Performance of disease classification models in human samples.** The bars show the median prediction performance ( $\text{Log2P} = \log_2(\text{auPRC}/\text{prior})$ ) from 5-fold cross validation of each disease classification model trained on human disease samples.

## Discussion

Gene expression, coexpression, and regulation have been shown to vary with species, age, sex, tissue, phenotypic, and experimental factors [5,24–28]. We develop a supervised machine learning approach to map similar transcriptomes across species, thus finding samples in different species that are functionally most analogous based on their expression profiles. Many previous studies have compared expression profiles across species using differential expression and similarity metrics [12–15], however, our data-driven method allows us to provide gene expression profiles as positive examples and contrast them with negative examples so that similar transcriptomic landscapes are prioritized based on features that are specific to the trait/context of interest.

We were able to obtain some promising preliminary results for mapping samples from corresponding tissues across species, but our method still needs tuning to be robust and accurate. One immediate area for improvement is the size of the tissue label set. When we expanded to include more human tissues, performance of the human tissue classification models increased significantly when making predictions for human samples (**Fig. 4.3, 4.6**). Cross-species predictions in turn showed a significant decrease, but it is likely that the only six tissues with labeled examples in mouse and zebrafish (blood, brain, heart, intestine, liver, and ovary) are not providing enough biological context to distinguish tissues in human with higher biological similarity amongst the 15 that are labeled.

### Performance of multi-species gene feature sets

We tested five feature sets for mapping analogous samples across species: one-to-one orthologs (OnetoOne), orthologous groups averaged (OGs-avg), OGs maximum

(OGs-max), Gene Ontology [20] (GO) biological processes averaged (GO-avg), and GO biological processes maximum (GO-max). We also tested building these five feature sets for commonality across 3 species (human, mouse, and zebrafish) and 6 species (human, mouse, zebrafish, fly, worm, yeast). Although some features are lost when using features common to 6 species compared to 3 species, it only slightly lowers performance, suggesting that the most conserved features serve as a large portion of prediction power.

When classifying samples within the same species, the five feature sets (OnetoOne, OGs-avg, OGs-max, GO-avg, GO-max) generally perform very similarly across tissues in each species. However, when making cross-species predictions, feature sets based on orthology tend to perform similarly to each other, but outperform GO feature sets. This may be due to our incomplete knowledge of which genes in each species take part in specific biological processes. Orthology is based on sequence similarity, and all of the species we consider have sequenced genomes, so it is likely that this set of relationships is more complete.

### **Future directions**

From a computational perspective, most studies frame functional knowledge transfer between species as a gene classification problem wherein an approach is developed to prioritize genes in one species based on data in another. Other groups have embedded genes/proteins across species into the same vector space [29] or determined homologous genes that are most likely to be functionally similar based on the similarity of their network neighborhoods [30]. Our lab group has had recent success in combining these ideas for gene classification across species by embedding their

network structure in a shared vector space. We can incorporate these ideas into sample classification to test whether a shared space is able to boost the sample classification performance as well.

### **Availability of data and code**

We plan to make all of the code to reproduce our approach as well as the tissue and disease labels available for others to build on our methods or to use transcriptome labels in a new analysis. When we finish tuning the method to work more consistently across all tissues, diseases, and species we will make results available and develop a webserver that enables a researcher to input a query transcriptome sample and get a ranked list of similar samples (i.e. transcriptomic landscapes) to guide experimental design. With the amount of expression profiles available for comparison, the large-scale effort is a valuable tool and significant novel contribution.

## **Methods**

### **RNA-seq data collection**

We downloaded the TPM expression for all available human, mouse, and zebrafish RNA-seq samples in the ARCHS4 [31] database as of version 8. The transcripts were mapped to Entrez genes [32]. We filtered out samples that had more than 50% zero counts as lenient quality control.

### **Curation of tissue and disease labels**

We downloaded sample descriptions for all samples and used TAGGER to annotate samples with UBERON ontology tissue labels. We then manually inspected sample descriptions of tagged tissues to ensure the label was correct, removing incorrect labels

and single cell data. Later, we added human sample tissue (but not cell types not specific to a given tissue) labels from the TissueNexus database [23].

We downloaded metaSRA [33] version 1.8 and subset the disease labels to samples that we obtained from ARCHS4. Sample descriptions were again manually inspected to ensure the disease labels were correct.

### **Creating a common feature set across species for sample classification**

In order to train models that could make predictions in another species, we had to create a common set of features across samples. We tested five options for common features: one-to-one orthologs (OnetoOne), orthologous groups (OGs) averaged (OGs-avg), OGs maximum (OGs-max), Gene Ontology [20] (GO) biological processes averaged (GO-avg), and GO biological processes maximum (GO-max). We used orthologous groups that had a similarity score of at least 0.5 from WORMHOLE [34] and restricted the set of GO biological processes to terms that had 5-300 genes annotated to them. Any group that did not have at least one gene in each species was not used in the feature set. Before creating each type of feature set for each species, the TPM expression of each gene was z scored across all samples in the species. For one-to-one orthologs, we simply subset the genes to those that were common across all three species. For OGs-avg, we averaged the z score values from each gene in an OG. For OGs-max, the maximum value across all genes in an OG was retained as the feature value. For GO-avg, we used the average z score from all genes annotated to the GO term and the maximum z score for GO-max. In order to test the feasibility of extending this method to other common model organisms that we cannot obtain tissue

labels for, we also tested creating feature sets with feature groups common to six species (human, mouse, zebrafish, fly, worm, and yeast).

### **Sample classification models**

We trained a one-vs-rest logistic regression model with an L2 penalty for each tissue or disease. Tissue models were trained in each of human, mouse, and zebrafish data, but disease labels are only for human samples, so all disease models are trained on human data. Training data was standard scaled, and this scaling was applied to the test data. Each tissue classification model was trained on all samples in a given species and used to make predictions on samples from other species. Performance was evaluated on curated labels. We also performed 5-fold cross-validation within species for each tissue and disease, to get an idea of how well tissue and disease transcriptomes could be classified within a species. Performance shown in figures is the median performance value of 5 fold cross validation results. The exception is that when we are comparing tissue prediction within a species to tissue prediction with training on two species and testing in the third, prediction performance within species is based on an 80/20 train/test split.



## REFERENCES

1. Takao K, Miyakawa T. Genomic responses in mouse models greatly mimic human inflammatory diseases. *Proc Natl Acad Sci*. 2015;112:1167–72.
2. Brunner D, Balci F, Ludvig EA. Comparative psychology and the grand challenge of drug discovery in psychiatry and neurodegeneration. *Behav Processes*. 2012;89:187–95.
3. Tamaki C, Nagayama T, Hashiba M, Fujiyoshi M, Hizue M, Kodaira H, et al. Potentials and limitations of nonclinical safety assessment for predicting clinical adverse drug reactions: correlation analysis of 142 approved drugs in Japan. *J Toxicol Sci*. 2013;38:581–98.
4. Bolliger MF, Pei J, Maxeiner S, Boucard AA, Grishin NV, Südhof TC. Unusually rapid evolution of Neuroligin-4 in mice. *Proc Natl Acad Sci U S A*. 2008;105:6421–6.
5. Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, et al. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet*. Nature Publishing Group; 2007;39:730–2.
6. Kuehn MR, Bradley A, Robertson EJ, Evans MJ. A potential animal model for Lesch–Nyhan syndrome through introduction of HPRT mutations into mice. *Nature*. Nature Publishing Group; 1987;326:295–8.
7. Thomas D, Chancellor D, Micklus A, LaFever S, Hay M, Chaudhuri S, et al. Clinical Development Success Rates and Contributing Factors 2011–2020 [Internet]. Available from: <file:///Users/kayla/Downloads/2021%20Clinical%20Development%20Success%20Rates%202011-2020%20v17.pdf>
8. Attarwala H. TGN1412: From Discovery to Disaster. *J Young Pharm JYP*. 2010;2:332–6.
9. Köhler S, Doelken SC, Ruef BJ, Bauer S, Washington N, Westerfield M, et al. Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research. *F1000Research*. 2014;2:30.
10. McGary KL, Park TJ, Woods JO, Cha HJ, Wallingford JB, Marcotte EM. Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proc Natl Acad Sci*. 2010;107:6544–9.
11. Breschi A, Gingeras TR, Guigó R. Comparative transcriptomics in human and mouse. *Nat Rev Genet*. 2017;18:425–40.
12. Levin M, Anavy L, Cole AG, Winter E, Mostov N, Khair S, et al. The

mid-developmental transition and the evolution of animal body plans. *Nature*. Nature Publishing Group; 2016;531:637–41.

13. Hashimshony T, Feder M, Levin M, Hall BK, Yanai I. Spatiotemporal transcriptomics reveals the evolutionary history of the endoderm germ layer. *Nature*. Nature Publishing Group; 2015;519:219–22.

14. Cardoso-Moreira M, Halbert J, Vallotton D, Velten B, Chen C, Shao Y, et al. Gene expression across mammalian organ development. *Nature*. 2019;1.

15. Le H-S, Oltvai ZN, Bar-Joseph Z. Cross-species queries of large gene expression databases. *Bioinformatics*. 2010;26:2416–23.

16. Arneson D, Zhang Y, Yang X, Narayanan M. Shared mechanisms among neurodegenerative diseases: from genetic factors to gene networks. *J Genet*. 2018;97:795–806.

17. Pierson E, Koller D, Battle A, Mostafavi S. Sharing and Specificity of Co-expression Networks across 35 Human Tissues. *PLoS Comput Biol*. 2015;11:e1004220.

18. Rodriguez-Esteban R, Jiang X. Differential gene expression in disease: a comparison between high-throughput studies and the literature. *BMC Med Genomics*. 2017;10:59.

19. Zahn-Zabal M, Dessimoz C, Glover NM. Identifying orthologs with OMA: A primer. *F1000Research*. 2020;9:27.

20. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. *Nat Genet*. 2000;25:25–9.

21. Yousef M, Ülgen E, Uğur Sezerman O. CogNet: classification of gene expression data based on ranked active-subnetwork-oriented KEGG pathway enrichment analysis. *PeerJ Comput Sci*. 2021;7:e336.

22. Segura-Lepe MP, Keun HC, Ebbels TMD. Predictive modelling using pathway scores: robustness and significance of pathway collections. *BMC Bioinformatics*. 2019;20:543.

23. Lin C-X, Li H-D, Deng C, Guan Y, Wang J. TissueNexus: a database of human tissue functional gene networks built with a large compendium of curated RNA-seq data. *Nucleic Acids Res*. 2022;50:D710–8.

24. Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, et al. Understanding multicellular function and disease with human tissue-specific networks. *Nat Genet*. 2015;47:569–76.

25. Oliva M, Muñoz-Aguirre M, Kim-Hellmuth S, Wucher V, Gewirtz ADH, Cotter DJ, et al. The impact of sex on gene expression across human tissues. *Science*. 2020;369:eaba3066.
26. Anderson WD, Soh JY, Innis SE, Dimanche A, Ma L, Langefeld CD, et al. Sex differences in human adipose tissue gene expression and genetic regulation involve adipogenesis. *Genome Res*. 2020;30:1379–92.
27. Lopes-Ramos CM, Chen C-Y, Kuijjer ML, Paulson JN, Sonawane AR, Fagny M, et al. Sex Differences in Gene Expression and Regulatory Networks across 29 Human Tissues. *Cell Rep*. 2020;31:107795.
28. Irizar H, Goñi J, Alzualde A, Castillo-Triviño T, Olascoaga J, Lopez de Munain A, et al. Age gene expression and coexpression progressive signatures in peripheral blood leukocytes. *Exp Gerontol*. 2015;72:50–6.
29. Fan J, Cannistra A, Fried I, Lim T, Schaffner T, Crovella M, et al. Functional protein representations from biological networks enable diverse cross-species inference. *Nucleic Acids Res*. 2019;47:e51.
30. Chikina MD, Troyanskaya OG. Accurate Quantification of Functional Analogy among Close Homologs. *PLOS Comput Biol*. 2011;7:e1001074.
31. Lachmann A, Torre D, Keenan AB, Jagodnik KM, Lee HJ, Wang L, et al. Massive mining of publicly available RNA-seq data from human and mouse. *Nat Commun*. 2018;9:1366.
32. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*. 2007;35:D26–31.
33. Bernstein MN, Doan A, Dewey CN. MetaSRA: normalized sample-specific metadata for the Sequence Read Archive. *bioRxiv*. 2016;090506.
34. Sutphin GL, Mahoney JM, Sheppard K, Walton DO, Korstanje R. WORMHOLE: Novel Least Diverged Ortholog Prediction through Machine Learning. *PLoS Comput Biol*. 2016;12:e1005182.

## **CHAPTER 5: SUMMARY, REFLECTIONS, LIMITATIONS, AND FUTURE DIRECTIONS**

### **Summary**

The goal of this dissertation research was to provide insights into the genomic signatures, pathways, and interactions that characterize the age/sex biases and cross-species analogs of complex diseases and traits. These relationships are critical for improving our ability to diagnose and treat complex diseases. I have worked towards this goal by developing computational frameworks capable of leveraging massive amounts of publicly-available genomic data with prior knowledge using network analysis and machine learning. By releasing code to reproduce our approaches and providing tools for scientists to query my results, I have helped build infrastructure for advancing biomedical research into the era of precision medicine.

### **Reflections and limitations**

Although this dissertation research represents considerable progress in many areas of biology, data science, and machine learning (ML), there are still some key limitations in this work. Broadly, these encompass numerous issues faced and careful choices to be made in all computational biology studies pertaining to i) deficiencies in our curated biological ground truth databases, ii) biases in the experiments that generate the datasets, and iii) assumptions in our computational models. Even when we address these issues to the best of our abilities, all results produced by ML models using big data must be interpreted with care.

## Building RNA-seq coexpression networks

In Chapter 2, I address the question: *how can we best build coexpression networks from heterogeneous RNA-seq data that comes from mostly small experiments generated by individual labs, with a range of sequencing depths and qualities, as well as high-quality consortium data?* In this chapter, I elaborate on the most accurate and robust methods to build coexpression networks from RNA-seq data. I test multiple normalization and network transformation techniques and their combinations to make concrete recommendations of when and how to use these techniques.

We put a lot of thought and effort into the gold standard we used as ground truth, *i.e.* ‘truly’ coexpressed and non-coexpressed gene pairs (see Chapter 2 Appendix: supplemental note). Briefly, we curated pairs of genes that were functionally related based on experimental evidence in the Gene Ontology [1], but were very careful to only use specific terms in the ontology to determine these gene pairs. We were equally careful about defining pairs of genes that are unlikely to be coexpressed using another set of manually curated terms from the Gene Ontology. In this set, we kept only gene pairs that we had enough functional information about to know that based on the biological processes they play a role in, they are extremely unlikely to be functionally related. This is a much more painstaking process than that many other coexpression studies use to evaluate the set of coexpressed and non-coexpressed gene pairs in their networks [2–4].

Despite the amount of care we took to build this gold standard, it is still based on functional annotations that are not tissue-specific. We tried to account for tissue-based effects by subsetting our original gold standard gene pairs using sets of genes known to

be expressed or not expressed in any given tissue, but non-expression is not the only modulator of tissue-specific coexpression [5]. In addition, age and sex affect gene expression and coexpression as well [6–8], but we currently have no systematic way to determine the robustness of our methods to these biological factors.

### **Age and sex specificity**

In Chapter 3, I address two questions: (1) *can age or age group be predicted using only the gene expression values?* And (2) *what do these gene signatures tell us about age- and sex-specific biological contexts?* Here, I curate about 30,000 primary human transcriptomes to predict age group and investigate age- and sex-biased gene signatures. I also use experimentally-validated genesets for determining enrichment of multiple biological contexts in different age and sex groups.

I took an extraordinary amount of care to curate age and sex labels to create a dataset we could use to investigate age- and sex-biased gene expression. This set excludes cell lines, xenografts, pooled samples, and single cell data. In other words, it is about as high quality as one could hope for in a large-scale dataset of age- and sex-labeled primary, bulk transcriptomes derived from public databases. However, most samples do not have health/disease labels, and there are surely biases in how these diseases are distributed across sex and age groups. While curating samples, I noticed many more female samples with systemic lupus erythematosus (SLE), which is hardly surprising since females outnumber males somewhere in the range of 7–10 to 1 in SLE incidence [9,10]. There are many other diseases with higher prevalence in certain age and sex groups, and there is also a bias in which samples are collected due to which diseases have the most funding for study. There were also very little cancer labels (if any) in the

youngest age groups. The number of cancer-associated datasets in each age group seems to roughly correlate with the total number of samples in the group. This is not hard to believe even though cancer incidence increases with age [11], due to the fact that study of cancer is generally well-funded [12]. These may have an effect on the gene signatures and subsequent enrichment scores I calculated for experimentally-derived genesets associated with various biological processes, cell types, phenotypes, and especially diseases. I would expect it to have a greater effect on age-stratified sex signatures, due to deriving them from the TPM expression distribution across female and male samples for each gene but using machine learning to derive the sex-stratified age signatures. Machine learning has some robustness to noise due to regularization. Despite these potential biases and limitations, we captured a number of true age- and sex-biased signatures, and look forward to myself and others building upon this work.

### **Cross species analogs**

In Chapter 4, I address the question: *can we utilize mass public transcriptomic data to identify analogous samples, and therefore biological contexts and phenotypes across species?* In this chapter, I describe our efforts to use machine learning in mapping transcriptomic landscapes and phenotypes across species to improve functional knowledge transfer.

Thus far, we have only seen lukewarm success in mapping analogous samples across species, but our group will continue to test new methods for making this goal possible (see Cross species sample classification in *Future directions*). However, in the process I learned a lot about translating functional knowledge across species, model organism

biology, publicly-available experimental data, and databases curating information in individual or multiple species. There is a lot of room to improve our understanding of cross-species biology, but there are also many exciting ideas and efforts continuously advancing the field just a little bit.

## **Future directions**

### **Age- and sex-specific gene interaction networks**

The work in this dissertation has laid the groundwork to build age- and sex-specific gene interaction networks. Greene, Krishnan, Wong and team built the first genome-scale tissue-specific functional interaction networks [13] and in a follow-up study, Krishnan et al were able to show that using a brain-specific gene network to predict novel candidate genes, brain-specific pathways, and developmental stages related to autism spectrum disorder was more accurate than using a general (*i.e.* not tissue-specific) or different tissue-specific network [14]. Essentially, using a tissue-specific gene interaction network for the tissue most affected by the disorder improves our ability to make these predictions. It stands to reason that if taking tissue-specificity into account improves the relevant predictions we are able to make, accounting for other biological contexts such as age and sex should further improve our accuracy.

The sticking point is that while their study developed the method to integrate many coexpression networks into one high-fidelity tissue-specific network, they did so with only microarray data, not RNA-seq data. Very little work had been done to evaluate robust and accurate normalization and network transformation methods to build coexpression networks from RNA-seq data, so I addressed this issue in Chapter 2.



With this evaluation completed, the remaining obstacle was that the majority of publicly-available microarray and RNA-seq samples we need to build these networks are not associated with age and sex information. In Chapter 3, I manually curated nearly 30,000 bulk, primary human samples labeled with sex and age across the human lifespan. Further, I showed that age group can be predicted from gene expression. So, we can continue developing these models into more accurate ones to predict age on samples that do not have age labels. Sex is easy to predict based on expression of X and Y chromosome genes. Building on this work, in the near future, we will use these pre-trained ML models to infer age-group and sex labels for the >150,000 human public transcriptomes and then integrate these profiles into age- and sex-specific genome-scale gene networks and use these networks to study genes associated with diseases that show sex and age differences.

### **Cross species sample classification**

Our efforts to train models to match analogous samples across species were only partially successful, but we have discussed methods to improve the currently poor performance. A promising approach is one we took for successfully classifying genes across species: to first place all genes from different species into the same ‘functional’ space and then train ML models to transfer gene functions across species. In the near future, we will apply this approach for sample classification.

### **Ongoing work**

Current members of the lab have already started to build off of the work completed for this dissertation. The age, sex, tissue, and disease labels I have curated for many gene expression profiles are being used in multiple projects. Renming Liu is using disease

annotations to develop methods that can automatically identify subgroups of contrasting disease-relevant samples within transcriptome datasets. The tissue and disease labels will continue to be used to pursue accurate methods to match analogous transcriptomes, and thus biological contexts and phenotypes across species.

Hao Yuan is building coexpression networks in multiple species with RNA-seq data using the recommendations I developed in Chapter 2 for robust workflows to do so. He is also extending this work into building patient-specific networks. Stephanie Hickey is using age and sex labels, along with best practices for network building established in this dissertation, for integration of coexpression networks built using bulk and single-cell data to compare gene interactions in different regions of the brain in multiple age and sex groups.

## REFERENCES

1. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. *Nat Genet.* 2000;25:25–9.
2. Iancu OD, Kawane S, Bottomly D, Searles R, Hitzemann R, McWeeney S. Utilizing RNA-Seq data for de novo coexpression network inference. *Bioinformatics.* 2012;28:1592–7.
3. Ballouz S, Verleyen W, Gillis J. Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. *Bioinformatics.* 2015;31:2123–30.
4. Parsana P, Ruberman C, Jaffe AE, Schatz MC, Battle A, Leek JT. Addressing confounding artifacts in reconstruction of gene co-expression networks. *Genome Biol.* 2019;20:94.
5. Saha A, Kim Y, Gewirtz ADH, Jo B, Gao C, McDowell IC, et al. Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome Res.* 2017;27:1843–58.
6. Hartman RJG, Mokry M, Pasterkamp G, den Ruijter HM. Sex-dependent gene co-expression in the human body. *Sci Rep.* 2021;11:18758.
7. Irizar H, Goñi J, Alzualde A, Castillo-Triviño T, Olascoaga J, Lopez de Munain A, et al. Age gene expression and coexpression progressive signatures in peripheral blood leukocytes. *Exp Gerontol.* 2015;72:50–6.
8. Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. The human transcriptome across tissues and individuals. *Science.* 2015;348:660–5.
9. Ortona E, Pierdominici M, Maselli A, Veroni C, Aloisi F, Shoenfeld Y. Sex-based differences in autoimmune diseases. *Ann Ist Super Sanita.* 2016;52:205–12.
10. Cervera R, Khamashta MA, Font J, Sebastiani GD, Gil A, Lavilla P, et al. Systemic lupus erythematosus: clinical and immunologic patterns of disease expression in a cohort of 1,000 patients. The European Working Party on Systemic Lupus Erythematosus. *Medicine (Baltimore).* 1993;72:113–24.
11. Cronin KA, Scott S, Firth AU, Sung H, Henley SJ, Sherman RL, et al. Annual report to the nation on the status of cancer, part 1: National cancer statistics. *Cancer [Internet].* [cited 2022 Nov 16];n/a. Available from: <https://onlinelibrary.wiley.com/doi/abs/10.1002/cncr.34479>
12. National Institutes of Health. Research Portfolio Online Reporting Tools (RePORT) [Internet]. *Res. Portf. Online Report. Tools Rep.* 2022 [cited 2022 Nov 16]. Available from: <https://report.nih.gov/funding/categorical-spending#/>

13. Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, et al. Understanding multicellular function and disease with human tissue-specific networks. *Nat Genet.* 2015;47:569–76.
14. Krishnan A, Zhang R, Yao V, Theesfeld CL, Wong AK, Tadych A, et al. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat Neurosci.* 2016;19:1454–62.