Pan-tissue pan-cancer characterization of novel human orphan genes via analysis of RNA-Sequencing data

by

Urminder Singh

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Eve Syrkin Wurtele, Co-major Professor
Xiaoqiu Huang, Co-major Professor
Karin Dorman
Andrew Severin
Marna Yandeau-Nelson

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2021

Copyright © Urminder Singh, 2021. All rights reserved.

DEDICATION

To my Mother, Father, Brother, and Sister

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

ACKNOWLEDGMENTS

I cannot thank enough my advisor, Dr. Eve Wurtele, for her support and guidance. I am incredibly thankful for the efforts and dedication she has put continuously to improve my research and me as a researcher. I am grateful to Dr. Karin Dorman for helping me understand several concepts in statistics, and providing critical feedback on my first paper. My special thanks to the rest of my committee members: Dr. Xiaoqiu Huang, Dr. Andrew Severin, and Dr. Marna Yandeau-Nelson for their helpful comments and feedback. I was fortunate to have Dr. Arun Seetharam as a mentor, and as a friend, from whom I have learned a great deal about *Bioinformatics*. Many thanks to all my fellow lab members and friends for providing valuable suggestions regarding my work. I am thankful for the assistance and support I received from BCB and GDCB administrative staff members Trish Stauble, Carla Harris, Diane Jepsen, and Danise Jones. I thank all the co-authors of my papers for their valuable contributions. I thank all the anonymous reviewers for critiquing my manuscripts. I am thankful for the financial aid I received from Iowa State University. The work presented in this dissertation was completed with financial support from National Science Foundation grant IOS 1546858, "Orphan Genes: An Untapped Genetic Reservoir of Novel Traits".

ABSTRACT

The recently emerged, young orphan genes provide an organism with a cadre of unique species-specific proteins. Since first described 25 years ago, several functionally important orphan genes from diverse species have been characterized. However, there remain significant lacunae in our knowledge about the origins and functions of orphan genes.

In a bid to decipher the "dark transcriptome", a result of pervasive transcription, researchers are exploring high-throughput sequencing-based gene annotation methods. Recent studies have made efforts to compile a comprehensive human transcriptome using data from experimental approaches such as RNA-Seq, Ribo-Seq, and proteomics. A number of these studies continue to ignore orphan genes because of their non-canonical features, such as short length, and lack of introns.

There is a growing interest to catalog the unannotated novel transcripts and ORFs and understand their roles in context of human physiology and diseases like cancer. These novel transcripts include unannotated genes, small Open Reading Frames (smORFs) of < 100 codons, novel ORFs encoded by lncRNAs and other non-coding RNAs, and other regulatory non-coding RNAs.

This dissertation presents methods and tools to efficiently process, and analyze large RNA-Seq datasets for the purpose of identifying and characterizing orphans and other yet unannotated genes. First, I developed MetaOmGraph a Java tool for interactive exploratory analysis of large expression datasets. MetaOmGraph provides an easy framework to explore expression patterns of genes and transcripts and build hypotheses about their functional roles. Next, I developed orfipy a fast and flexible Open Reading Frame (ORF) finder for quick and accurate annotation of Coding Sequences (CDS) in large transcriptomic datasets. Third, I present pyrpipe, a python package for straightforward and reproducible analysis of RNA-Seq datasets.

Using these tools and methods, a reproducible and scalable pipeline for annotating the human "dark transcriptome" is proposed. Leveraging terabytes of tumor and non-diseased RNA-Seq data,

the pipeline identified thousands of tissue- and tumor-specific transcripts coding for novel peptides. Phylostratigraphy and synteny analysis revealed the majority of novel genes are orphans encoding a human-specific protein. The expression and translation status of these novel transcripts are validated using independent RNA-Seq and Ribo-seq data. Pan-cancer analysis of these novel genes reveals their differential expression and association with overall patient survival suggesting their potential to be utilized for novel diagnostic and therapeutic interventions.

CHAPTER 1. GENERAL INTRODUCTION

The rise of new genes in a genome provides the opportunity for evolutionary novelty. A special class of functional genes, orphan genes, provides an organism with a unique set of proteins (Singh and Wurtele, 2020; Blevins et al., 2021; Van Oss and Carvunis, 2019; Arendsee et al., 2014; Vakirlis et al., 2020; Ruiz-Orera et al., 2015). Coding for completely novel proteins, these orphan genes can play a central role in adaptation to new biological niches (Singh and Wurtele, 2020; Ruiz-Orera et al., 2015; Van Oss and Carvunis, 2019; Li et al., 2021; Singh and Wurtele, 2021).

However, the annotation, and functional characterization of these young genes is limited by several challenges. First, existing gene annotation pipelines and tools are biased against orphan genes as they have different features compared to the "canonical" protein-coding genes (Arendsee et al., 2014; Singh and Wurtele, 2020; Li et al., 2021). Secondly, orphan genes are ignored because they lack homology to known functional proteins. Thus, more direct approaches like using RNA-Seq data to identify expressed transcripts are more suitable for orphan gene annotation (Li et al., 2021; Singh and Wurtele, 2021, 2020; Blevins et al., 2021; Ruiz-Orera et al., 2015; Pertea et al., 2018; Hon et al., 2017). Even when using high-throughput sequencing data such as RNA-Seq, many orphan genes remain undetected as they are very sparsely expressed in a tissue-specific manner or only under limited environmental factors. Thus, discovery of these genes requires RNA-data that encompass diverse tissues and environmental conditions (Li et al., 2021). It is also essential and challenging to distinguish the "real" novel transcripts from transcriptional noise resulting due to pervasive transcription of the genome (Pertea et al., 2018; Hangauer et al., 2013).

Examining thousands of RNA-Seq samples to annotate and study expression patterns of orphan genes is limited by existing tools, which are inefficient for analysis of very large datasets. The main contributions of this dissertation are bioinformatics tools, methods and workflows to

annotate and functionally characterize young orphan and other unannotated genes. These tools solve the challenges associated with: reproducible RNA-Seq processing, fast and accurate CDS annotation in transcriptomic data, and efficient downstream analysis of massive expression datasets.

This dissertation is organized in eight chapters. The **second chapter** is a short review of the evolutionary mechanisms by which new genes arise in genomes (Singh and Wurtele, 2020). Particularly, it summarizes and compares the mechanisms of rapid divergence and *de novo* emergence for evolution of orphan genes (Vakirlis et al., 2020).

The **third chapter** presents MetaOmGraph, an efficient tool for exploratory analysis of large expression datasets (Singh et al., 2020). MetaOmGraph provides several novel features for exploration of gene expression patterns such as coexpression, differential expression or differential coexpression. MetaOmGraph's interactive visualizations and fully interactive interface enables easy exploratory or confirmatory analysis. Such expression and coexpression based analyses are central for inferring the functional roles of orphan genes as they lack homology to known proteins (Singh et al., 2020).

The **fourth chapter** presents a study investigating population-specific gene expression differences among African American and European American populations and their role in COVID-19 disease severity and outcome (Singh et al., 2021a). It is known that the COVID-19 disease has affected people of African-American ancestries disproportionately (Millett et al., 2020; Singh et al., 2021a). This study used MetaOmGraph for the exploratory analysis of existing human expression data from multiple tissues and tumors. We report a number of immune related genes are differentially regulated among European and African populations. Further, single-cell RNA-Seq data analysis reveals cell-specific expression of the differentially expressed genes. This study significantly contributes to the ongoing efforts to understand the COVID-19 pandemic and provides insights into approaching personalized treatment of COVID-19 for people with different genetic ancestries.

The **fifth chapter** introduces orfipy, a fast and flexible tool for finding Open Reading Frames (ORFs) in massive transcriptomic datasets (Singh and Wurtele, 2021). orfipy was developed to provide an easy, fast and flexible solution for annotation of Coding Sequences (CDS) in transcriptomic data in order to annotated novel protein-coding transcripts.

The **sixth chapter** describes pyrpipe, a python library for reproducible RNA-Seq analysis (Singh et al., 2021b). pyrpipe presents a novel object-oriented framework to write RNA-Seq processing pipelines in pure python. pyrpipe provides a number of features for reproducible, easy-to-implement and easy-to-modify workflows. A dedicated module for NCBI-SRA access provides straightforward implementation of harmonized re-analysis pipelines for publicly available RNA-Seq datasets.

Finally, the **seventh chapter** describes a best-practices RNA-Seq based pipeline to annotated novel orphan and other genes. Using this pipeline, thousands of novel human genes are identified by utilizing terabytes of RNA-Seq data from tens of thousands of non-tumor and tumor samples across multiple diverse tissues. Phylostratigraphy (Arendsee et al., 2019; Tautz and Domazet-Lošo, 2011) analysis classifies the majority of novel genes as orphans encoding a human-specific protein. The expression and translation status of these novel genes is further validated using multiple independent strand-specific RNA-Seq data, single-cell RNA-Seq data and Ribo-Seq datasets. We identified thousands of novel genes that are differentially expressed and associated with patient survival making them promising candidates for diagnostic and therapeutic applications in multiple cancer types.

1.1 References

Arendsee, Z., Li, J., Singh, U., Seetharam, A., Dorman, K. S., and Wurtele, E. S. (2019). phylostratr: a framework for phylostratigraphy. *Bioinformatics*, 35(19):3617–3627.

Arendsee, Z. W., Li, L., and Wurtele, E. S. (2014). Coming of age: orphan genes in plants. *Trends in plant science*, 19(11):698–708.

Blevins, W. R., Ruiz-Orera, J., Messeguer, X., Blasco-Moreno, B., Villanueva-Cañas, J. L., Espinar, L., Díez, J., Carey, L. B., and Albà, M. M. (2021). Uncovering de novo gene birth in yeast using deep transcriptomics. *Nature Communications*, 12(1):1–13.

Hangauer, M. J., Vaughn, I. W., and McManus, M. T. (2013). Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet*, 9(6):e1003569.

Hon, C.-C., Ramilowski, J. A., Harshbarger, J., Bertin, N., Rackham, O. J., Gough, J., Denisenko, E., Schmeier, S., Poulsen, T. M., Severin, J., et al. (2017). An atlas of human long non-coding RNAs with accurate 5 ends. *Nature*, 543(7644):199–204.

Li, J., Singh, U., Bhandary, P., Campbell, J., Arendsee, Z., Seetharam, A. S., and Wurtele, E. S. (2021). Foster thy young: Enhanced prediction of orphan genes in assembled genomes. *bioRxiv*, pages 2019–12.

Millett, G. A., Jones, A. T., Benkeser, D., Baral, S., Mercer, L., Beyrer, C., Honermann, B., Lankiewicz, E., Mena, L., Crowley, J. S., et al. (2020). Assessing differential impacts of COVID-19 on Black communities. *Annals of Epidemiology*, 47:37–44.

Pertea, M., Shumate, A., Pertea, G., Varabyou, A., Breitwieser, F. P., Chang, Y.-C., Madugundu, A. K., Pandey, A., and Salzberg, S. L. (2018). CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome biology*, 19(1):208.

Ruiz-Orera, J., Hernandez-Rodriguez, J., Chiva, C., Sabidó, E., Kondova, I., Bontrop, R., Marqués-Bonet, T., and Albà, M. M. (2015). Origins of de novo genes in human and chimpanzee. *PLoS Genetics*, 11(12):e1005721.

Singh, U., Hernandez, K. M., Aronow, B. J., and Wurtele, E. S. (2021a). African americans and european americans exhibit distinct gene expression patterns across tissues and tumors associated with immunologic functions and environmental exposures. *Scientific Reports*, 11(1):1–14.

Singh, U., Hur, M., Dorman, K. S., and Wurtele, E. S. (2020). Metaomgraph: a workbench for interactive exploratory data analysis of large expression datasets. *Nucleic Acids Research*, 48(4):e23–e23. gkz1209.

Singh, U., Li, J., Seetharam, A., and Wurtele, E. S. (2021b). pyrpipe: a Python package for RNA-Seq workflows. *NAR Genomics and Bioinformatics*, 3(2). lqab049.

Singh, U. and Wurtele, E. S. (2020). Genetic novelty: How new genes are born. *Elife*, 9:e55136.

Singh, U. and Wurtele, E. S. (2021). orfipy: a fast and flexible tool for extracting ORFs. *Bioinformatics*. btab090.

Tautz, D. and Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nature Reviews Genetics*, 12(10):692–702.

Vakirlis, N., Carvunis, A.-R., and McLysaght, A. (2020). Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *Elife*, 9:e53500.

Van Oss, S. B. and Carvunis, A.-R. (2019). De novo gene birth. *PLoS Genetics*, 15(5):e1008160.

CHAPTER 2. GENETIC NOVELTY: HOW NEW GENES ARE BORN

Urminder Singh and Eve Syrkin Wurtele

Modified from a manuscript published in *eLife*

2.1 Abstract

Analysis of yeast, fly and human genomes suggests that sequence divergence is not the main source of orphan genes. For half a century, most scientists believed that new protein-coding genes arise as a result of mutations in existing protein-coding genes. It was considered impossible for anything as complex as a functional new protein to arise from scratch. However, every species has certain genes, known as 'orphan genes', which code for proteins that are not homologous to proteins found in any other species. What do these orphan genes do, and how are they formed?

2.2 Main

To date the roles of hundreds of orphan genes have been characterized. Although this is just a tiny fraction of the total, it is known that most of them code for proteins that bind to conserved proteins such as transcription factors or receptors. Some of these proteins are toxins, some are involved in reproduction, some integrate into existing metabolic and regulatory networks, and some confer resistance to stress (Carvunis et al., 2012; Li et al., 2009; Arendsee et al., 2014; Xiao et al.; Belcaid et al., 2019). However, none of them are enzymes (Arendsee et al., 2014). Orphan genes arise quickly, so they may provide a disruptive mechanism that allows a given species to survive changes to its environment. Thus, the study of how orphan genes arise (and fall) is central to understanding the forces that drive evolution (Figure 1).

One possible mechanism is the "de novo" appearance of a gene from an intergenic region or a completely new reading frame within an existing gene (Tautz and Domazet-Lošo, 2011). An

alternative mechanism is that the coding sequence of the orphan gene arises by rapid divergence from the coding sequence of a preexisting gene: this would mean that an entire set of regulatory and structural elements would be available to the gene as it evolves. Now, in eLife, Nikolaos Vakirlis and Aoife McLysaght (both from Trinity College Dublin) and Anne-Ruxandra Carvunis (University of Pittsburgh) report how they have studied yeast, fly and human genes to compare the contributions of these two mechanisms to the emergence of orphan genes (Vakirlis et al., 2020).

Previous studies have used simulations to estimate the number of orphan genes that appear by divergence; until now, no one had relied on actual genomics data to study this phenomenon. Vakirlis et al. use a new approach to analyze orphan genes that have originated through divergence. They examine regions of the genome that correspond to each other (so-called syntenic regions) in related species to determine whether a gene exists in both regions and, if so, whether the proteins are non-homologous. If the genes have no homology, they may have originated by rapid divergence from the coding sequence of a preexisting gene.

Using this method, Vakirlis et al. infer that at most 45% of *S. cerevisiae* (yeast) orphan genes, 25% of *D. melanogaster* (fruit fly) orphan genes, and 18% of human orphan genes arose by rapid divergence, but this is an upper estimate. For example, it is possible that a new coding sequence might have arisen de novo within an existing gene, rather than the existing coding sequence having been modified beyond recognition.

But how can a protein sequence continue to be selected for as it rapidly diverges? Vakirlis et al. suggest that divergence might occur by a process of partial pseudogenation: the existing gene becomes non-functional, and then, with no selection pressure to retain the old protein, it diverges to form an orphan gene.

Many orphan genes may not have been identified yet, because they do not have homologs in other species, and have few recognizable sequence features. This means that up to 80% of orphan genes can be missed when a new genome is annotated (Seetharam et al., 2019). The approach detailed by Vakirlis, Carvunis and McLysaght evaluates specifically those annotated orphan genes for which a similar gene exists in a related species (which is 50% of them; (Arendsee et al.,

2019)). As high-quality genomes from more species become available, and as more orphan genes are annotated, the approach will provide yet deeper insights into the origin of these genes.

2.3 Conclusion

One of the many open questions in this field deals with genes of 'mixed age'. Some such genes have incorporated 'chunks' of orphans into their coding sequences. A gene that has done this is (somewhat arbitrarily) considered to be the age of its most ancient segment, but we know little about the mechanism of this process or its significance. Another question involves the unique strategies and rates of evolution of each gene (Revell et al., 2018). How might the abundance and mechanisms of orphan gene origin vary among species? And how do different environments affect the emergence of orphan genes?

2.4 References

Arendsee, Z., Li, J., Singh, U., Seetharam, A., Dorman, K. S., and Wurtele, E. S. (2019). phylostratr: A framework for phylostratigraphy. *Bioinformatics*, 35(19):3617–3627.

Arendsee, Z. W., Li, L., and Wurtele, E. S. (2014). Coming of age: orphan genes in plants. *Trends in Plant Science*, 19(11):698–708.

Belcaid, M., Casaburi, G., McAnulty, S. J., Schmidbaur, H., Suria, A. M., Moriano-Gutierrez, S., Pankey, M. S., Oakley, T. H., Kremer, N., Koch, E. J., Collins, A. J., Nguyen, H., Lek, S., Goncharenko-Foster, I., Minx, P., Sodergren, E., Weinstock, G., Rokhsar, D. S., McFall-Ngai, M., Simakov, O., Foster, J. S., and Nyholm, S. V. (2019). Symbiotic organs shaped by distinct modes of genome evolution in cephalopods. *Proceedings of the National Academy of Sciences*, 116(8):3030–3035.

Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., Charloteaux, B., Hidalgo, C. A., Barbette, J., Santhanam, B., et al. (2012). Proto-genes and de novo gene birth. *Nature*, 487(7407):370.

Li, L., Foster, C. M., Gan, Q., Nettleton, D., James, M. G., Myers, A. M., and Wurtele, E. S. (2009). Identification of the novel protein qqs as a component of the starch metabolic network in arabidopsis leaves. *The Plant Journal*, 58(3):485–498.

Palmieri, N., Kosiol, C., and Schlötterer, C. (2014). The life cycle of drosophila orphan genes. *elife*, 3:e01311.

Revell, L. J., González-Valenzuela, L. E., Alfonso, A., Castellanos-García, L. A., Guarnizo, C. E., and Crawford, A. J. (2018). Comparing evolutionary rates between trees, clades and traits. *Methods in Ecology and Evolution*, 9(4):994–1005.

Seetharam, A. S., Singh, U., Li, J., Bhandary, P., Arendsee, Z., and Wurtele, E. S. (2019). Maximizing prediction of orphan genes in assembled genomes. *BioRxiv*.

Tautz, D. and Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nature Reviews Genetics*, 12(10):692–702.

Vakirlis, N., Carvunis, A.-R., and McLysaght, A. (2020). Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *Elife*, 9:e53500.

Van Oss, S. B. and Carvunis, A.-R. (2019). De novo gene birth. *PLoS Genetics*, 15(5):e1008160.

Xiao, W., Liu, H., Li, Y., Li, X., Xu, C., Long, M., and Wang, S. A rice gene of de novo origin negatively regulates pathogen-induced defense response. 4(2):e4603.

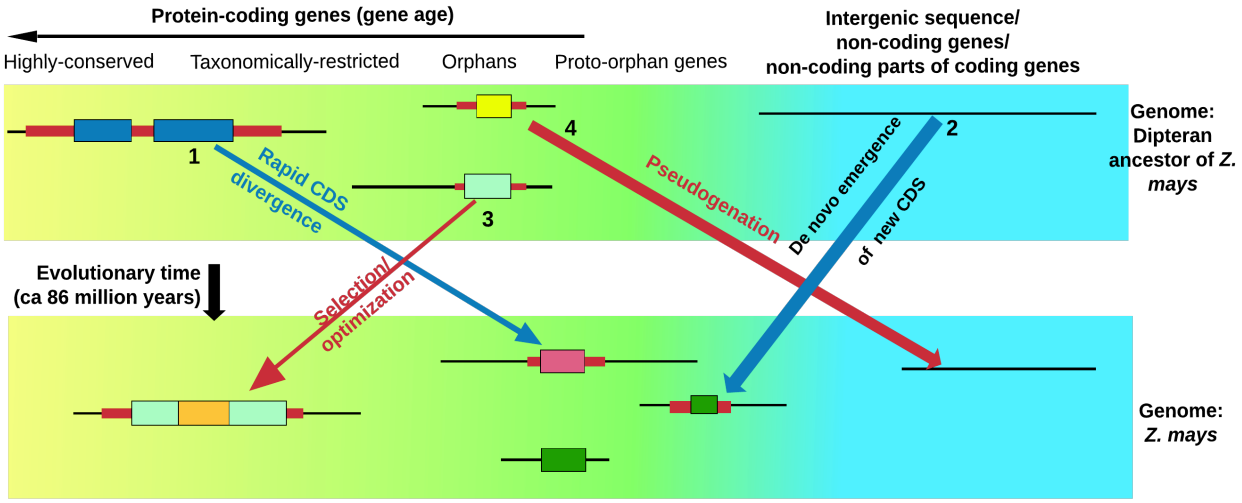

Figure 2.1 Life cycle of orphan genes. Every species has orphan genes that have no homologs in other species. This schematic shows the genome of the fruit fly (bottom) and the genome of an ancestor of the fruit fly (top). Each panel also shows (from left to right): genes that are highly conserved and can be traced back to prokaryotic organisms (yellow background); genes that are found in just a few related species (taxonomically restricted genes), orphan genes and potential orphan genes that are not currently expressed and are thus free from selection pressure (proto-orphan genes); and regions of the genome that do not code for proteins (blue background) (Van Oss and Carvunis, 2019; Palmieri et al., 2014). An orphan gene can form through the rapid divergence of the coding sequence (CDS) of an existing gene (1), or arise de novo from regions of the genome that do not code for proteins (including the non-coding parts of genes that evolve to code for proteins; 2). Some orphan genes will be important for survival, and will thus be selected for and gradually optimized (3). This means that the genes in a single organism will have a gradient of ages (Tautz and Domazet-Lošo, 2011). Many proto-orphan genes will undergo pseudogenation (that is, they will not be retained; 4). Coding sequences (shown as thick colored bars) with detectable homology are shown in similar colors. Vakirlis et al. estimate that a minority of orphan genes have arisen by divergence of the coding sequence of existing genes.

CHAPTER 3. METAOMGRAPH: A WORKBENCH FOR INTERACTIVE EXPLORATORY DATA ANALYSIS OF LARGE EXPRESSION DATASETS

Urminder Singh [1,2,3], Manhoi Hur [2], Karin Dorman [1,3,4] and Eve Syrkin Wurtele [1,2,3]

[1]Bioinformatics and Computational Biology Program, [2]Center for Metabolic Biology, [3]Department of Genetics Development and Cell Biology, [4]Department of Statistics, Iowa State University, Ames, IA 50011, USA

Modified from a manuscript published in *Nucleic acids research*

3.1 Abstract

The diverse and growing omics data in public domains provide researchers with tremendous opportunity to extract hidden, yet undiscovered, knowledge. However, the vast majority of archived data remain unused. Here, we present MetaOmGraph, a free, open-source, standalone software for exploratory analysis of massive datasets. Researchers, without coding, can interactively visualize and evaluate data in the context of its metadata, honing-in on groups of samples or genes based on attributes such as expression values, statistical associations, metadata terms, and ontology annotations. Interaction with data is easy via interactive visualizations such as line charts, box plots, scatter plots, histograms and volcano plots. Statistical analyses include coexpression analysis, differential expression analysis, and differential correlation analysis, with significance tests. Researchers can send data subsets to R for additional analyses. Multithreading and indexing enable efficient big data analysis. A researcher can create new MOG projects from any numerical data; or explore an existing MOG project. MOG projects, with history of explorations, can be saved and shared. We illustrate MOG by case studies of large curated datasets from human cancer RNA-Seq, where we identify novel putative biomarker genes in different tumors, and microarray and metabolomics data from Arabidopsis thaliana. MOG

executable and code: http://metnetweb.gdcb.iastate.edu/ and https://github.com/urmi-21/MetaOmGraph/.

3.2 Introduction

Petabytes of raw and processed data generated with microarray, RNA-Seq (bulk and single-cell), and mass spectrometry for small molecules and proteins are available through public data repositories (Brazma et al., 2003; Kodama et al., 2011; Haug et al., 2012; Martens et al., 2005). These data represent multiple species, organs, genotypes, and conditions; some are the results of groundbreaking research. Buried in these data are biological relationships among molecules that have not yet been explored. Integrative analysis of data from the multiple studies representing diverse biological conditions is the key to fully exploit these vast data resources for scientific discovery (Lazar et al., 2012; Rhodes and Chinnaiyan, 2005). Such analysis allows efficient reuse and recycling of these available data and metadata (Lazar et al., 2012; Brazma et al., 2003; Li et al., 2019). Higher statistical power can be attained with bigger datasets, and the wide variety of biological conditions can reveal the complex regulatory structure of genes. Yet, despite the availability of such vast data resources, most bioinformatic studies use only a limited amount of the available data.

A common goal of analyzing omics data is to infer functional roles of particular features (genes, proteins, metabolites, or other biomolecules) by investigating coexpression and differential expression patterns. A wide variety of R-based (Ihaka and Gentleman, 1996) tools can provide specific analyses (Rau et al., 2014; Ritchie et al., 2015; Love et al., 2014; Robinson et al., 2010). Such tools are based upon rigorous statistical frameworks and produce accurate results when the model assumptions hold. Several tools avoid the need to code by providing "shiny" interfaces (Chang et al., 2018) to various subsets of R's functionalities (Ma et al., 2018; Zhu et al., 2018; Choi and Ratner, 2019; Monier et al., 2019; Rue-Albrecht et al., 2018; Kucukural et al., 2019; Marini, 2018). Such R tools based on the "shiny" interface have the general limitations that they are not well suited for very large datasets and can have limited interactivity.

Increasing the usability of the vast data resources by enabling efficient exploratory analysis would provide a tremendous opportunity to probe the expression of transcripts, genes, proteins, metabolites and other features across a variety of different conditions. Such exploration can generate novel hypotheses for experimentation, and hence improving the fundamental understanding of the function of genes, proteins, and their roles in complex biological networks (Wang et al., 2018; Rhodes and Chinnaiyan, 2005; Brazma and Vilo, 2000; Mentzen and Wurtele, 2008; Almeida-de Macedo et al., 2013; Trevino et al., 2012; Li et al., 2019).

Presently, there are very limited options for researchers to interact with expression datasets using the fundamental principles of exploratory data analysis (Tukey, 1977). Exploratory data analysis is a technique to gain insight into a dataset, often using graphical methods which can reveal complex associations, patterns or anomalies within data at different resolutions. By adding interactivity for visualizations and statistical analyses, researchers with little or no programming experience are able to directly explore the underlying, often complex and multidimensional, data themselves. Researchers in diverse domains (e.g., experts in Parkinson's disease, malaria, or nitrogen metabolism) can mine and re-mine the same data, extracting information and deriving testable hypotheses pertinent to their particular areas of expertise. These hypotheses can inform the design of new laboratoryexperiments. Being able to explore and interact with data becomes even more critical as datasets become larger. The information content inherent in the vast stores of public data is enormous. Due to the sheer size and complexity of such big data, there is a pressing requirement for effective interactive analysis and visualization tools (Kelder et al., 2010; Shannon et al., 2013).

In this paper, we present MetaOmGraph (MOG), a Java software, to interactively explore and visualize large expression datasets. MOG overcomes the challenges posed by the size and complexity of big datasets by efficient handling of the data files. Further, by incorporating metadata, MOG adds extra dimensions to the analyses and provides flexibility in data exploration. At any stage of the analysis, a researcher can save her/his progress. Saved MOG projects can be shared, reused, and included in publications. MOG is user-centered software,

designed for exploring diverse types of numerical data and their metadata, but specialized for expression data.

3.3 Materials and Methods

3.3.1 Overview

MOG is an interactive software that can run on any operating system capable of running Java (Linux, Mac and Windows). MOG's Graphical User Interface (GUI) is the central component through which all the functionality is accessed (Figure 3.1). Access to MOG is easy. MOG is a standalone program and runs on the researcher's computer; thus, the researcher does not need to rely on internet accessibility for computations, and is not slowed down by the data transfer latency. Furthermore, the data in a researcher's project is secure, remaining on the researcher's computer, particularly important for confidential data such as human RNA-Seq.

3.3.1.1 Interactive data exploration.

MOG displays all the data in interactive tables and trees, providing a flexible and structured view of the data. The user can interactively filter or select data for analysis. This ability is particularly important for aggregated datasets, as users may wish to split data into groups of studies, treatments, or organs. A novel aspect of MOG is its capability of producing *interactive* visualizations. The researcher can visualize data via line charts, histograms, box plots, volcano plots, scatter plots and bar charts, each of which is programmed to allow real-time interaction with the data and the metadata. Users can group, sort, filter, change colors and shapes, zoom, and pan interactively, via the GUI. At any point in the exploration, the researcher can look-up external databases: GeneCards (Safran et al., 2010), Ensembl (Hubbard et al., 2002), EnsemblPlants (Kersey et al., 2015), RefSeq (Pruitt et al., 2006), TAIR (Lamesch et al., 2011) and ATGeneSearch (`http://metnetweb.gdcb.iastate.edu/MetNet_atGeneSearch.htm`) for additional information about the genomic features in the dataset. Researchers can also easily access SRA and GEO databases using the accessions present in the study metadata.

3.3.1.2 Efficient, multithreaded and robust.

A key advantage of MOG is its minimal memory usage, enabling datasets to be analyzed that are too large for other available tools. Researchers with a laptop/desktop computer can easily run MOG with data files containing thousands of samples and fifty thousands of transcripts. MOG achieves computational efficiency via two complementary approaches. First, MOG indexes the data file, rather than storing the whole data in main memory. This enables MOG to work with very large files using a minimal amount of memory. Second, MOG speeds up the computations using multithreading, optimizing the use of multi-core processors. MOG is robust and can cope with most of the errors and exceptions (such as missing values or forbidden characters) that can occur when handling diverse data types. Bug reports can be submitted with a single click, if encountered.

3.3.1.3 Data-type agnostic.

Although specifically created for the analysis of 'omics data, which is the focus of this paper, MOG is designed to be flexible enough to generally handle numerical data. A user can supplement a MOG project with any type of metadata about the features, and about the studies. Thus, a MOG user can interactively analyze and visualize voluminous data on any topic. For example, a user could create a project on: transmission of mosquito-borne infectious diseases world-wide; public tax return data for world leaders over the past 40 years; daily sales at Dimo's Pizza over five years; player statistics across all Women's National Basketball Association (WNBA) teams; climate history and projections since 1900.

3.3.1.4 Leverage of third party Java libraries.

In addition to the functionality we have programmed into MOG, MOG borrows some functionality from freely available and extensively tested third-party Java libraries (JFreeChart, Apache Commons Math, Nitrite, and JDOM). We have combined these to create a highly modular system that is amendable to changes and extensions and developers can easily implement

new statistical analyses and visualizations in the future. MOG is an open source project and we plan to expand and develop it further through community driven efforts. Information on how to contribute to MOG, and who to contact with further questions, is provided at \hl{https://github.com/urmi-21/MetaOmGraph/blob/master/CONTRIBUTING.md}.

3.3.1.5 Interface to R.

Based on the utility and popularity of R for data analysis, we have implemented a GUI to facilitate execution of R scripts through MOG. MOG's GUI enables a user to interactively select or filter data using MOG; these data are then passed to R. This avoids the need to constantly write new R code to specify different genes and samples for analyses. For example, a user can write an R script for hierarchical clustering of genes based on the expression levels, interactively select or filter data using MOG, and execute the R script. More details on how to use MOG for executing R scripts are provided in the user manual available from (https://github.com/urmi-21/MetaOmGraph/tree/master/manual).

3.3.2 Creating a new MOG project or using an existing one.

A user can quickly create a new MOG project using two delimited files: 1) a file with unique identifiers (IDs) for each feature (e.g, gene), metadata about that feature, and numerical data quantifying each feature across multiple conditions (e.g., multiple samples and studies), and 2) a file containing unique identifiers for each sample and metadata about the samples and studies in the datafile. These are virtually combined by MOG, using the unique identifier in each file (Supplementary Figure 1). Selecting appropriate methods for data normalization, batch correction, and vetting are important considerations for a user when creating a new project (Supplementary File 1).

New MOG projects, as well as those from well-vetted datasets, including the human and *A. thaliana* datasets described herein, can be re-opened, analyzed, modified or shared. Ongoing exploration results, such as correlations, lists, and other interactive analyses, can be saved in any

MOG project, regardless of whether it was obtained from our website or created from custom data.

3.3.3 Detecting statistical association within data

Measures of statistical association between a pair of features in a dataset quantify the similarity in their expression patterns across the samples that comprise that dataset (Eisen et al., 1998; Kumari et al., 2012; Langfelder and Horvath, 2008; Faith et al., 2007). Genes with significant statistical association may participate in common biological processes and pathways (van Dam et al., 2017; Mentzen and Wurtele, 2008; Eisen et al., 1998). Genes with significant association only under specific conditions may reveal their functional significance under those conditions (Vandenbon et al., 2016; McKenzie et al., 2016).

MOG provides the researcher with several statistical measures to estimate associations/coexpression among the features. It can also compute association between samples, which reflects similarity between the samples. Choosing appropriate statistical measures and interpretations for each dataset is left to the user.

3.3.3.1 Correlation, mutual information and relatedness.

We have incorporated four key methods that measure association among pairs of features. Each has its own advantages and disadvantages, depending on the types of relationships the researcher wishes to detect, and the characteristics of the dataset being explored.

MOG can compute pairwise Pearson and Spearman correlation for pairs of selected features across all samples or conversely, between selected samples across all features. The Pearson correlation coefficient measures the extent of a linear relationship between two random variables, X and Y, whereas, the Spearman correlation coefficient measures monotonic relationships between the two variables. Both excel at detecting linear relationships, however, Spearman is less sensitive to outliers (Wang and Huang, 2014). Pearson and Spearman correlations are often used

to find coexpressed genes and generate matrices used for inferring gene expression networks (Kumari et al., 2012; Mentzen and Wurtele, 2008; Vandenbon et al., 2016).

MOG also computes pairwise mutual information (MI) between selected features across samples. MI quantifies the amount of information shared between two random variables. Let (X, Y) be a pair of discrete random variables over the space $\mathcal{X} \times \mathcal{Y}$. Then, the MI for X and Y is defined as:

$$I(X;Y) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right),$$

where, $p(x, y)$ is the joint probability mass function of X and Y, and $p(x)$ and $p(y)$ are the marginal probability mass functions for X and Y, respectively. Compared to correlation measures, MI is a more general approach that can detect complex, non-linear associations. The interpretation of the MI value is different than that of correlation values: an MI value of zero, $I(X;Y) = 0$, implies statistical independence of X and Y, whereas a correlation value of zero need not imply statistical independence (Wang and Huang, 2014). MI has been applied to detect non-linear associations in gene expression datasets (Daub et al., 2004; Song et al., 2012; Trevino et al., 2012; Singh et al., 2017). MOG computes MI using B-splines density estimation, as described in Daub et. al. 2004 (Daub et al., 2004).

MOG can also determine the context likelihood of relatedness (CLR) (Faith et al., 2007). CLR determinations aim to identify biologically relevant associations by discounting features (e.g., genes) that have promiscuous associations. Specifically, the CLR compares the MI value between each pair of features to the background distribution of MI values that include either of these features (Faith et al., 2007).

3.3.3.2 Meta-analysis of correlation coefficients.

MOG can perform meta-analysis of Pearson correlations. Studies using microarray data showed that meta-analysis and analysis of pooled normalized samples each bring out meaningful, but different, relationships among genes (Almeida-de Macedo et al., 2013). For meta-analysis of correlation coefficients, MOG calculates a weighted average of the individual Pearson correlation

coefficients computed from each study. The weights are proportional to the sample size, i.e., correlations estimated from larger studies are more trusted (Hedges and Vevea, 1998; Field, 2001). Meta-analysis can be useful when multiple studies run a similar experiment (e.g., effect of heat-stress on *A. thaliana*), but may not control ancillary sources of variation (e.g., coverage variation in RNA-Seq data). MOG provides a choice between a fixed effects model (FEM) or a random effects model (REM) (Hedges and Vevea, 1998; Field, 2001) for the meta-analysis. The FEM combines the estimated effects by assuming that all studies probe the same correlation in the same population, i.e., studies are homogeneous. In contrast, the REM allows studies to be heterogeneous, with additional, uncontrolled sources of variation (Hedges and Vevea, 1998; Field, 2001). The FEM does not account for all heterogeneities, thus the researcher should choose a model and interpret the results with appropriate caution.

3.3.4 Differential expression between groups

Determining differentially-expressed features from aggregated datasets provides direction for further data exploration. In MOG, we have incorporated several popular statistical methods to evaluate differential expression between two groups of samples. For analysis of groups with independent samples, we have implemented: Mann-Whitney U test (a non-parametric test that makes no assumptions about data distribution); Student's t-test (assumes equal variance and normally distributed data); Welch's t-test (does not assume equal variance, assumes a normal distribution of data); and a permutation test (makes no assumptions about data distribution; computes null distribution empirically using the data). For analysis of groups with paired samples, we have implemented: a Paired t-test (assumes normal distribution of data); a Wilcoxon signed-rank test (a non-parametric test; no assumption of data distribution); and a permutation test for paired samples (makes no assumptions about data distribution but computes null distribution empirically using the data).

MOG's methods to identify differentially expressed genes are general statistical methods which are designed for large sample sizes (30 or more samples for gene expression data).

Computation of these methods via MOG permits interactivity, which promotes data exploration. A limitation of the interactive differential expression analysis methods implemented in MOG is that they are designed for large sample sizes and use normalized data as input. For smaller sample sizes, a user can apply specialized model-based methods, accessible through R, to infer differentially expressed genes in RNA-Seq or microarray datasets. For example, methods like edgeR (Robinson et al., 2010), DESeq2 (Love et al., 2014) and limma (Ritchie et al., 2015) require raw counts as input and can provide more reliable differential expression analysis (Soneson and Delorenzi, 2013) for smaller sample sizes. Tools like ideal (Marini, 2018) and DEBrowser (Kucukural et al., 2019) provide interactive interface for accessing these popular differential expression analysis methods (Ritchie et al., 2015; Robinson et al., 2010; Love et al., 2014).

3.3.5 Differential correlation between groups

Features whose correlation with other features is significantly different only under particular environmental, genetic or developmental conditions are designated as differentially correlated. Such *shifting* biological interactions among these genes or their regulators (McKenzie et al., 2016; Fukushima, 2013) reflect the context-dependency of gene expression.

MOG can find the features whose Pearson correlation to a user-selected feature differs significantly between two groups of samples. To do this, MOG applies a Fisher transformation (Fisher, 1915) and performs a hypothesis test for equality of Pearson correlation coefficients from the two groups. (The difference of the two Fisher transformed Pearson correlation coefficients follows a normal distribution (McKenzie et al., 2016)). The researcher can choose to conduct a test for statistical significance on the Fisher transformed Pearson correlation coefficients or on the raw Pearson correlation coefficients.

3.3.6 Statistical significance determinations

For each statistical test, MOG provides a non-parametric option (a permutation test) and parametric options (calculations under distributional assumptions) to estimate p-values.

Empirical p-values are calculated by a permutation test that estimates the null distribution of a test statistic by randomly permuting the labels of the observed data points (assuming that the labels are exchangeable under the null hypothesis) (Edgington, 1980). Because permutation tests do not rely on any data distribution, they are applicable even if parametric assumptions are not met. More permutations yield more precise estimates of the null distribution and p-values, but at the cost of longer computation times. MOG accelerates computation of permutation tests by multithreading, and processing the permuted datasets in parallel (Supplementary File 1).

MOG provides three popular parametric methods to adjust the p-values for multiple comparisons: the Bonferroni method (Weisstein, 2004), the Holm method (Holm, 1979) and the Benjamini—Hochberg (BH) method (Benjamini and Hochberg, 1995). Bonferroni and Holm methods are applied to control the family-wise error rate (FWER), whereas the BH method controls the false discovery rate (FDR). Controlling the FWER limits the total number of false positives; the Holm method is less conservative as compared to the Bonferroni method. In contrast, controlling the FDR controls the proportion of false positives among the significant tests.

3.3.7 Datasets

To create case-studies with MOG, we assembled MOG projects based on three technical platforms.

3.3.7.1 Human cancer RNA-Seq dataset (7,142 samples).

We created a new MOG project based on the well-vetted dataset from Wang et al., 2018 (Wang et al., 2018). This dataset combines RNA-Seq data from The Cancer Genome Atlas (TCGA, tumor and non-tumor samples) (https://cancergenome.nih.gov/) and Genotype Tissue Expression (GTEX, non-tumor samples) (Lonsdale et al., 2013).

To create the MOG project, we excluded from the dataset any organ types in which the number of tumor or non-tumor samples was < 30. To ensure statistical independence among the samples, we removed all non-tumor samples from TCGA and included only one TCGA

Table 3.1 Tumor and non-tumor samples in the *Hu-cancer-RNASeq-dataset* and the number of *up*regulated and *down*regulated genes in each tumor type with respect to the corresponding normal samples, as calculated by MOG.

TCGA disease	**GTEX organ**	**#TCGA samples**	**#GTEX samples**	**Total**	**#Up**	**#Down**
Breast invasive carcinoma (BRCA)	Breast	965	89	1,054	1,093	2,827
Colon adenocarcinoma (COAD)	Colon	277	339	616	1,401	3,036
Esophageal carcinoma (ESCA)	Esophagus	182	659	841	1,989	2,229
Kidney Chromophobe (KICH)	Kidney	60	32	92	986	4,214
Kidney renal clear cell carcinoma (KIRC)	Kidney	470	32	502	1,877	2,263
Kidney renal papillary cell carcinoma (KIRP)	Kidney	236	32	268	1,152	2,737
Liver hepatocellular carcinoma (LIHC)	Liver	295	115	410	1,527	1,485
Lung adenocarcinoma (LUAD)	Lung	491	313	804	1,361	2,753
Lung squamous cell carcinoma (LUSC)	Lung	486	313	799	2,210	3,734
Prostate adenocarcinoma (PRAD)	Prostate	426	106	532	577	1,633
Stomach adenocarcinoma (STAD)	Stomach	380	192	572	1,527	1,631
Thyroid carcinoma (THCA)	Thyroid	441	318	759	993	1,525
Uterine Corpus Endometrial Carcinoma (UCEC)	Uterus	141	82	223	2,135	3,250
Uterine Carcinosarcoma (UCS)	Uterus	47	82	129	2,419	2,491

tumor-sample per patient (Supplementary File 2). We also excluded an outlier sample with very low expression values, based on a preliminary exploration of sample replicates using MOG (See Results).

We then compiled metadata for the studies/samples and for the genes and integrated this metadata into the dataset. We downloaded the study and sample metadata (TCGA metadata from TCGAbiolinks (Colaprico et al., 2015); GTEX metadata from GTEX's website (https://gtexportal.org/home)). We were unable to locate metadata for 17 of the TCGA samples and excluded these samples from the dataset (Supplementary File 2). We extracted metadata about the genes from the HGNC (https://www.genenames.org/), NCBI Gene (https://www.ncbi.nlm.nih.gov/gene), Ensembl (Hubbard et al., 2002), Cancer Gene Census (Sondka et al., 2018) and OMIM (Amberger et al., 2014) databases, and added these information to the gene metadata in our dataset. We also eliminated the 1,870 genes that were not reported for all studies resulting in a dataset, called herein, "*Hu-cancer-RNASeq-dataset*".

We generated the MOG project (*Hu-Cancer-18212-7412-RNASeq.mog*) from the *Hu-cancer-RNASeq-dataset* and its metadata. The MOG project contains expression values for

18,212 genes, 30 fields of metadata detailing each gene, across 7,142 samples representing 14 different cancer types and associated non-tumor tissues (Table 3.1); it also has 23 fields of metadata describing each study and sample in the dataset. We used MOG to log_2 transform the data for subsequent analyses.

3.3.7.2 *A. thaliana* microarray dataset (424 samples).

We created a new MOG project, *AT-Affy-22746-424-microarray.mog*, based on the *A. thaliana* curated microarray dataset (*"AT-microarray-dataset"*) from Mentzen and Wurtele, 2008 (Mentzen and Wurtele, 2008). This dataset had been compiled using 963 Affymetrix ATH1 chips with 22,746 probes from 70 diverse studies encompassing different conditions of development, stress, genotype, and environment. All chips in the dataset were individually normalized and scaled to a common mean using MAS 5.0 algorithm. Only chips with good quality biological replicates were kept and all the biological replicates were averaged to yield 424 samples. Finally, median absolute deviation (MAD)-based normalization (Yang et al., 2002) was applied to the data. We compiled new metadata for the genes from TAIR gene annotations (Lamesch et al., 2011) and added phylostratal inferences (Arendsee et al., 2019). The sample metadata were obtained from Mentzen and Wurtele, 2008 (Mentzen and Wurtele, 2008).

3.3.7.3 *A. thaliana* metabolomics GC-MS dataset (656 samples).

The small molecule composition (metabolomics) data that we used to create a MOG project were from 656 GM-MS samples describing the effect of 50 knock out (or knock down) mutations of genes of mostly unknown functions on the accumulation of metabolites in *A. thaliana* (Fukushima et al., 2014) (called herein, *"AT-metab-dataset"*). We downloaded these data from the Plant/Eukaryotic and Microbial Resource (PMR) (Hur et al., 2013). We created the MOG project *AT-Mutation-242-656-metab.mog* with this dataset.

3.4 Results

We illustrate MOG's usability and flexibility by exploring three diverse datasets from different perspectives. The statistical analyses and visualizations shown were generated exclusively using MOG. Often, our exploration led us to conclusions consistent with prior experimental or *in silico* results. In other cases, the exploration led us to completely novel predictions that could be tested experimentally.

3.4.1 Preliminary exploration of the *Hu-cancer-RNASeq-dataset*

Determining that a dataset is valid, properly normalized and free of batch effects is a critical preliminary step in the analysis. To verify that samples from similar biological conditions exhibit similar expression patterns for all the genes, we used MOG to compute pairwise Pearson correlations among samples from the same biological condition (tumor/non-tumor and organ type). All the samples had high Pearson correlations (> 0.70) with others taken from the same organ and tumor status, except one sample from lung adenocarcinoma (LUAD), which we removed from the dataset (Additional File 1).

We visualized the distribution of Pearson correlation values for non-tumor samples. For homogeneous samples, such distributions should appear unimodal. However, several organs show multimodal distributions (Supplementary Figure 2). This finding led us to conjecture that samples might have been taken from different anatomical sites within these organs. By exploring further with MOG, we were able to identify additional metadata on sub-locations in the colon and esophagus that support this conjecture (Supplementary Figure 2). However, the stomach sample metadata does not further specify location (or any other obvious factor, such as gender, race, or age) that might distinguish subgroups of samples. Because the stomach samples are of several distinct types, a researcher might want to consider analyzing them as such.

3.4.2 Using MOG to identify a catalog of differentially expressed genes in cancers

We wanted to identify *key genes* that are regulated by, or implicated in, the molecular and cellular processes driving cancer, and to further explore the processes in which these genes are involved. For this task, we used MOG first to identify the differentially expressed genes in samples from each tumor type vs. corresponding non-tumor samples, and then to examine the expression patterns of these genes. We define a gene as differentially expressed between two groups if it meets each of the following criteria:

1. Estimated fold change in expression of 2-fold or more (log fold change, $|logFC| \geq 1$ where $logFC$ is calculated as in limma (Ritchie et al., 2015).)

2. Mann-–Whitney U test, on the $\log_2$ transformed data, is significant between the two groups (BH corrected p-value $< 10^{-3}$)

In each type of cancer in the *Hu-cancer-RNASeq-dataset*, between 2,000-5,000 of the 18,212 genes are differentially expressed (Table 3.1; Supplementary File 3). Thirty-five of these genes are consistently differentially expressed in all of the cancers (Table 3.2).

Table 3.2 MOG identifies 35 genes as differentially expressed in all of the 14 tumor types. (Mann-Whitney U test, $|FC| \geq 2$, BH corrected p-value $< 10^3$).

Upregulated in each cancer	**Downregulated in each cancer**
BIRC5, BUB1, CDC45	ADH1B, C7, CHRDL1
CDKN2A, CENPF, DLGAP5	CMTM5, DCN, DES
FAM111B, KIF4A, KIF20A	DPT, GPM6A, GSTM5
MELK, MKI67, PBK	HPD, HSPB6, MRGPRF
PKMYT1, TOP2A, TPX2	NKAPL, PEG3, PI16
UBE2C	PTGDS, SCN7A, TCEAL2
	TGFBR3

Several genes that are deeply implicated in cancer are not differentially expressed in any of the tumor types we analyzed. One example is tumor suppressor protein 53 (TP53) (Figure 3.2 A and 3.2 B). (TP53 is differentially expressed in colorectal tumors (Slattery et al., 2018); colorectal tumors are not included in the *Hu-cancer-RNAseq-dataset*).

Fifteen of the 16 genes upregulated across all tumor types are coexpressed across the tumor samples, across the non-tumor samples, and across the combined tumor plus non-tumor samples (Figure 3.2 C , Supplementary File 4). Cyclin dependent kinase inhibitor (CDKN2A) is an outlier (Spearman correlation < 0.50) (Figure 3.2 D; Supplementary File 4). This coexpression might imply that these 15 genes function together as a module in both tumor and non-tumor cells.

In contrast, there is no coexpression cluster among the 19 genes that are downregulated across all cancer types; 62 individual gene pairs are correlated across all the samples (Spearman correlation ≥ 0.60) (Supplementary File 4). Expression of seven of these gene pairs is strongly correlated only among tumor samples but is not correlated among non-tumor samples; conversely, 18 gene pairs are strongly correlated among non-tumor samples but not among the tumor samples (e.g., Figure 3.2 E)– this finding indicates a context-dependent coordination of these gene pairs. Four gene pairs are strongly correlated among both tumor and in non-tumor samples (e.g, Figure 3.2 F.

3.4.2.1 Functional analysis of differentially expressed genes.

To determine whether the genes that are differentially expressed in cancers are involved in known biological processes, we performed gene ontology (GO) enrichment analysis using GO::TermFinder (Boyle et al., 2004) and Revigo (Supek et al., 2011) on the genes that are upregulated, downregulated or not significantly changed across all the cancer types. Consistent with the behavior of cancer cells, upregulated genes are significantly enriched in GO terms related to cell proliferation: cell cycle, cell division, organelle organization, regulation of cellular component organization and regulation of cell cycle (Supplementary File 4, Supplementary Figure 3). The 5,784 genes that did not change expression were enriched in GO terms RNA processing, mRNA metabolic process, nucleic acid metabolic process, and gene expression (Supplementary Figure 4; Supplementary File 4). The downregulated genes show no significant GO term enrichment.

3.4.3 Using MOG for gene-level exploration

With the aim to use MOG from the vantage point of an individual gene, we selected the glypican 3 (GPC3) gene as an interesting candidate for a case study. GPC3, encoding a glycosylphosphatidylinositol-linked heparan sulfate proteoglycan, is located on the X chromosome and has been implicated as a critical regulator of tissue growth and morphogenesis (Kaur and Cummings, 2019). GPC3 inhibits cell proliferation and hedgehog signaling during embryonic development (Capurro et al., 2008). In tumors, GPC3's role is complex and not well understood. It can promote or inhibit cell growth depending on the cancer type (Gao and Ho, 2011; Filmus and Capurro, 2008). Mutations in GPC3 have been linked to Wilm's tumour as well as Simpson-Golabi-Behmel syndrome (SGBS) (Blackhall et al., 2001; Davoodi et al., 2007).

3.4.3.1 GPC3 Expression patterns.

We explored expression patterns of GPC3 with regards to the 14 tumor types. Differential expression of GPC3 in non-tumor versus tumor samples varies by organ. GPC3 expression is 30 fold higher in the LIHC samples than in the non-tumor liver samples, and eight fold higher in the UCS samples compared to the non-tumor uterus samples (Supplementary File 5). In contrast, GPC3 is downregulated in nine tumor types (BRCA, COAD, ESCA, KIRC, KIRP, LUAD, LUSC, THCA, and UCEC) and unchanged in three tumor types (KICH, STAD and PRAD) (Figure 3.3 A and 3.3 B; Supplementary File 5).

These results are consistent with targeted studies of liver, breast, and lung tumors. GPC3 expression is upregulated in liver cancer (Ho and Kim, 2011; Anatelli et al., 2008; Gao and Ho, 2011), and has been suggested as a diagnostic biomarker and as a potential target for cancer immunotherapy in hepatocellular carcinoma (Ho and Kim, 2011; Anatelli et al., 2008; Capurro et al., 2003). GPC3 is downregulated in breast (Xiang et al., 2001), lung (Sasisekharan et al., 2002) and ovarian cancers (Kim et al., 2003), and it may act as a tumor suppressor in lung and renal cancer (Kim et al., 2003; Valsechi et al., 2014).

3.4.3.2 GPC3 Coexpression patterns.

We then investigated coexpression patterns of GPC3 in the tumor and non-tumor tissues from different organs (Additional File 3). GPC3 coexpression patterns differ between tumor and non-tumor samples according to the organ sampled (Figure 3.3 C), moreover, the genes whose expression is correlated with GPC3 are distinct according to organ types, all reflecting the complex role of this gene (Supplementary File 5). For example, 4,219 genes are coexpressed with GPC3 in non-tumor esophagus samples, whereas no gene is coexpressed with GPC3 in non-tumor samples from prostate and stomach (Supplementary File 5). Coexpressed genes also differed according to whether disease was present. For seven organs, fewer genes were coexpressed with GPC3 in tumor samples than in non-tumor samples (Supplementary File 5). For example, 192 genes were coexpressed with GPC3 in non-tumor liver samples, whereas no genes were significantly coexpressed with GPC3 in LIHC tumor samples (Figure 3.3 D and 3.3 E).

We analyzed GO term enrichment for those organs with more than 10 GPC3-coexpressed genes: colon, esophagus, kidney and liver. The term cell adhesion is enriched in GPC3-coexpressed genes from colon, esophagus, kidney and liver. The terms cell development, extracellular matrix organization and multicellular organism development are enriched among GPC3-coexpressed genes in colon, esophagus, and kidney. Other GO terms are overrepresented in a organ specific manner (Supplementary File 5).

3.4.3.3 GPC3-associated clusters in tumor versus non-tumor samples from liver.

To further explore potential interactions of GPC3 with other genes, we used MOG to build two gene coexpression networks from the 3,012 genes that are differentially expressed in LIHC – one network from non-tumor liver samples, and a second from LIHC samples (Additional File 3). Then, we imported each network into Cytoscape (Shannon et al., 2003) and identified the tightly connected modules using MCODE (Bader and Hogue, 2003).

In the network built from non-tumor liver samples, MCODE ranked the GPC3-containing cluster second most significant (73 nodes (genes); MCODE score 30.7). GPC3 was directly

connected with 21 genes in this cluster (Supplementary Figure 5), which is most enriched in GO terms: sulfur compound catabolic process, aminoglycan catabolic process, and extracellular matrix organization (Supplementary Figure 6; Supplementary File 5).

In contrast, in the LIHC samples, GPC3 was not significantly coexpressed with *any* other genes, and thus was absent from the entire LIHC network. However, the LIHC network does contain a module with 114 genes (MCODE score 94.3), 33 of which are in the GPC3-containing cluster identified from the non-tumor network (17 of these genes are directly connected with GPC3 in the non-tumor network) (Supplementary Figure 5). This cluster is enriched in GO terms: extracellular matrix organization, blood vessel development, and vasculature development (Supplementary File 5; Supplementary Figure 7).

3.4.4 Stage-wise analysis of *Hu-cancer-RNASeq-dataset*

3.4.4.1 Identifying new candidate biomarkers for cancers.

To identify potential biomarkers for tumors, we used MOG to distinguish genes whose expression is associated with the disease progression. We used MOG to separate samples by organ, and then by early stage (stage I or stage II) and late stage (stage III and later), based on the study metadata. Finally, we performed a Mann-Whitney U test on those genes that are upregulated in tumor versus non-tumor samples (Supplementary File 3) to reveal the genes that are upregulated in late stage compared to early stage (expression increase 2-fold or more, and BH corrected p-value < 0.05). These genes have increasing expression with cancer progression. We similarly identified the genes that have a decreasing pattern of expression with cancer progression.

ESCA, KIRP, KIRC THCA all included metadata and had sufficient numbers per stage to detect differentally expressed genes. (Full results in Additional file 4.) MOG reveals 221 genes that increase expression during tumor progression (gene numbers for each tumor type are: ESCA:91, KIRP:89, THCA:25, KIRC:24), and 227 genes that decrease expression (gene numbers for each tumor type: ESCA:89, KIRP:68, LIHC:64, KIRC:13) (Supplementary File 6; Additional File 4). Of these 448 genes, 122 are flagged as prognostic markers by The Human Protein Atlas

(THPA), which identifies prognostic markers by survival analysis (Uhlen et al., 2017). For example, figure 3.4 B and 3.4 C shows the expression pattern of two such genes, Phosphoenolpyruvate Carboxykinase 1 (PCK1, known to be downregulated in KIRC (Sun et al., 2016) and general marker of renal failure (Swe et al., 2019)) and Chromosome 10 Open Reading Frame 99 (C10orf99, a known colon cancer inhibitor (Pan et al., 2014), and positive marker of KIRC (Tian et al., 2019)), in KIRC and KIRP.

Three hundred and twenty-seven genes that were identified in our study as differentially expressed in at least one tumor type were *not* labeled as prognostic in THPA (Supplementary File 6). For example, out of the 111 genes that increase during progression of KIRC or KIRP, only 56 were flagged as unfavourable prognostic for renal cancer by THPA. Of the 79 genes MOG identifies as decreasing with cancer progression in KIRC or KIRP, 39 were labeled as prognostic favourable for renal cancer by THPA. Twenty-seven genes out of the 64 that we identified by MOG as decreasing with cancer progression in LIHC were labeled by THPA as prognostic favourable for liver cancer. Out of 25 genes identified as having increasing pattern in THCA, none were labeled as prognostic by THPA. We propose that these genes may provide new candidates as biomarkers for prognosis of these tumor types (Supplementary File 6).

A number of the 327 genes identified as differentially expressed in MOG but not listed in THPA have been experimentally evaluated for their potential as prognostic markers (Table 3.3). For example, ARG1, CYP2C8, CYP3A4, CYP3A7 and CYP4A11, which we identified using MOG as decreasing expression with LIHC progression, have each been recently studied as prognostic markers for hepatocellular carcinoma (You et al., 2018; Ren et al., 2018; Yu et al., 2018; Eun et al., 2019). MOG analysis provides additional support for use of these genes as biomarkers.

Using MOG to analyze and visualize the results by tumor type can reveal more nuanced information. For example, the Cluster of Differentiation 70 (CD70) gene is flagged by THPA and high CD70 expression is prognostic unfavourable for renal cancer. MOG analysis shows CD70 expression is higher in two types of renal tumors, KIRC and KIRP, and increases with disease progression (Figure 3.4A), but CD70 levels in another renal tumor type, KICH, have slightly

lower expression than in non-tumor samples; thus, specifically in the case of KICH, *low* CD70 levels might be an unfavorable prognosis.

For prognosis and personalized medicine (de Vries et al., 2018; Lightbody et al., 2018) *exceptions* can be extremely important, because specific tumor sub-types might respond differently to a particular treatment. By using MOG to explore RNA-Seq from large numbers of conditions and organs, a researcher can visualize data for individual samples or groups of that show changed expression of a prognostic marker or sets of markers, and compare these to those that do not.

Such exploration could suggest statistical analyses to try out in other, independent datasets to determine whether subsets of non-canonical samples might have a biologically distinct signature, revealing a different modality for a particular cancer. This in turn could be followed up by targeted experimental approaches or clinical studies.

3.4.5 Exploring genes of unknown functions in *AT-microarray-dataset*

Our aim in the case study of *AT-microarray-dataset* was to explore patterns of expression of genes with little or nothing known about them. The well-vetted dataset we used (Mentzen and Wurtele, 2008), encompasses expression values for 22,746 genes across 424 *A. thaliana* samples, representing 71 diverse studies and a wide variety of environmental, genetic and developmental conditions (Mentzen and Wurtele, 2008). We updated the gene metadata to the current TAIR annotations (Lamesch et al., 2011) and added phylostrata designations (obtained from phylostratr (Arendsee et al., 2019)).

We sought to identify genes of unknown or partially-known function that might be involved in photosynthesis, the process that gave rise to the earth's oxygenated atmosphere and the associated evolution of extant complex eukaryotic species. We focused particularly on regulation of the assembly and disassembly of the photosystem I and II light harvesting complexes; these dynamic processes respond sensitively to shifts in light and other environmental factors (Chen et al., 2018; Ruban and Johnson, 2015; Nosek et al., 2017; Bhuiyan et al., 2015). In particular,

Met1 (AT1G55480) is a 36 Kda protein that regulates the assembly of the photosystem II (PSII) complex (Bhuiyan et al., 2015). To explore genes that might be involved in PSII assembly, we calculated Spearman correlation of Met1 expression with that of the 22,746 genes represented on the Affymetrix chip (Figure 3.5). This analysis finds 104 genes whose expression is highly correlated to Met1 (Spearman's coefficient > 0.9) across all conditions (Supplementary File 7).

We examined whether genes of photosynthesis were over-represented in this Met1-coexpressed cohort. Among the Met1 coexpressed genes, the Gene Ontology (GO) Biological Functional terms most highly over-represented (p-value $< 10^{-5}$) are integral to the light reactions of photosynthesis: generation of precursor metabolites and energy; photosynthetic electron transport in photosystem I (PSI); reductive pentose-phosphate cycle; response to cytokinin; and PS2 assembly (Supplementary File 7). For example, the gene most highly correlated with Met1 is At2g04039, a gene encoding the NdhV protein, which is thought to stabilize the nicotinamide dehydrogenase (NDH) complex of PS1 (Fan et al., 2015); phylostratal analysis (Arendsee et al., 2019) indicates that NdhV has homologs across the photosynthetic organisms, streptophyta (land plants and most green algae). Eighteen of the Met1 coexpressed genes are designated as "unknown function" or "uncharacterized"; six are restricted to Viridiplantae. These genes would be good candidates to evaluate experimentally for a possible function in photosynthetic light reaction.

Our next aim was to use MOG to directly explore an orphan gene (a gene encoding a protein unrecognizable by homology to those of other species) (Arendsee et al., 2014; Gollery et al., 2006; Arendsee et al., 2019), and to determine potential processes that it might be involved in. First, we filtered each gene's target description to retain "unknown". From these, we filtered to retain only the phylostratigraphic designation "*Arabidopsis thaliana*". From this gene list, we identified genes that had an expression value greater than 100 in at least five samples. We selected the orphan gene of unknown function, At2G04675, for exploratory analysis. At2G04675 encodes a predicted protein of 67 aa with no known sequence domains (domains searched using CDD (Marchler-Bauer et al., 2014)). A Pearson correlation analysis of the expression pattern of At2G04675 with the other genes represented on the Affymetrix chip showed 48 genes had a

Pearson correlation of higher than 0.95 (Supplementary File 7); these genes are expressed almost exclusively in pollen (the male gametophyes of flowering plants) (Figure 3.6). The exploration implicates At2G04675 as a candidate for involvement in some aspect of pollen biology.

Using MOG to further explore genes that are associated with pollen, we identified sets of leaf and pollen samples (Supplementary Figure 8; Additional File 5), and then calculated genes that are differentially expressed in the leaf samples versus the pollen samples using a Mann-Whitney U test (fold change of 2-fold or more; BH corrected p-value $< 10^{-3}$) (Additional File 5). The GO terms most highly enriched (p-value $< 10^{-20}$) among genes upregulated in pollen are processes of cell cycle, mitosis, organellar fission, chromosome organization and DNA repair (Additional File 5). This reflects and emphasizes the critical role of these processes in male gametophyte development, particularly sperm biogenesis. Each angiosperm pollen grain must produce two viable sperm each used in the double fertilization of the ovule. Above all else, proper mitogenesis is essential to the function of a pollen grain. We visualized the *leaf vs. pollen* differential analysis by volcano plot (Figure 3.7; Supplementary Figure 9), this time to explore genes upregulated in leaf. Among these is At1G67860, an Arabidopsis specific gene encoding a protein of "unknown function". We used MOG to correlate expression of this gene versus all genes across all samples. One hundred sixteen genes, dispersed across all five chromosomes, are coexpressed with At1G67860 (Spearman correlation ≥ 0.65) (Supplementary File 7). The genes are expressed almost exclusively in mature leaf (Supplementary Figure 10). Most have no known function; a GO enrichment test indicates that GO biological processes overrepresented (p-value $< 10^{-3}$) among the genes are: defense response, response to stress, response to external biotic stimulus and response to other organism (Supplementary File 7).

3.4.6 Identifying coexpressed metabolites in *AT-metab-dataset*

Metabolomics is providing a growing resource for understanding metabolic pathways and identifying the structural and regulatory genes that shape these pathways and their interconnected lattice (Sumner et al., 2015; Quanbeck et al., 2012; Hur et al., 2013). Here, we use

the *AT-metab-dataset* metabolomics dataset that represents a comprehensive study of 50 mutants with a normal morphological phenotype but altered metabolite levels, and 19 wild type control lines (Fukushima et al., 2014). There are 8-16 biological replicates for each genetic line; data is corrected for batch effects. Data and metadata were retrieved from PMR (Hur et al., 2013). The aim of this case study was to tease out coexpressed metabolites that are affected by genetic perturbations. We identified a group of four highly coexpressed metabolites (Pearson correlation > 0.8): the amino acid arginine, its precursor L-ornithine, cyclic ornithine (3-amino-piperidine-2-one), and one unidentified metabolite. Plots across the means of the biological replicates of each sample (Supplementary Figure 11), shows accumulation of these metabolites is upregulated over four-fold in four mutant lines: *mur9*, mutants have altered cell wall constituents; *vtc1*, encodes GDP-mannose pyrophosphorylase, required for synthesis of manose, major constituent of cell walls, upregulated upon bacterial infection; *cim13*, gene of unknown function associated with disease resistance, *eto1*, negative regulator of biosynthesis of the plant hormone ethylene. An arginine-derived metabolite, nitrous oxide, has been widely implicated in signaling pathways in plants (del Rıo et al., 2004). MOG analysis might suggest to a researcher a potential relationship between arginine and the cell wall defense response, providing a suggestion for future experimentation.

3.4.7 Comparison to other software

Few tools that do not require coding are available for on-the-fly exploration of expression data. Most are "shiny" (Chang et al., 2018) apps (Monier et al., 2019; Price et al., 2019; Zhu et al., 2018; Choi and Ratner, 2019; Rue-Albrecht et al., 2018) providing a web interface to a limited number of R packages for data visualization, batch correction, differential expression analysis, PCA analysis (among samples) and gene enrichment analysis. Although shiny (Chang et al., 2018) is constantly improving, existing tools written in R (Monier et al., 2019; Zhu et al., 2018; Choi and Ratner, 2019) must rely on R's present capabilities for interactive applications (Furtună and Vinte, 2016). In contrast to R, Java, MOG's platform, has been used to develop

numerous software with interfaces that are interactive and user-friendly (e.g., (Shannon et al., 2003; López-Fernández et al., 2018; Ignatchenko et al., 2015; Kirov et al., 2017)), and MOG provides the researcher with specialized GUIs and methods for exploratory data analysis. MOG's GUI allows direct interactivity with the data through interactive tables, trees and visualizations, so that a researcher can easily explore data from different perspectives.

Most available R-based tools read all data directly into the main memory. Thus, on a laptop/desktop computer, analysis of a big dataset is slow (or crashes) if the available memory is not sufficiently large. For example, a dataset of 100,000 human transcripts over 5,000 samples (500,000,000 expression values) requires at least 4GB (8 byte for each value) of free memory to be loaded into memory at once. To circumvent this problem, R developers can use the new DelayedArray (Pagès et al., 2019) framework together with DelayedMatrixStats (Hickey, 2019) which can enable efficient handling of big datasets with R. For example, iSEE's (Rue-Albrecht et al., 2018) code is compatible with using DelayedArray (Pagès et al., 2019) objects.

In contrast, MOG uses an indexing strategy to read data only when it is needed, which drastically reduces the total memory consumption of the system. Table 3.4 compares five of the most recent tools for exploratory analysis of expression data to MOG. (More details are provided in Supplementary File 8.)

3.4.7.1 Benchmarking

We benchmarked MOG's performance with the *Hu-cancer-RNASeq-dataset* (18,212 genes over 7,142 samples) using a laptop with 64 bit Windows 10, 8 GB RAM and Intel(R) Core(TM) i5-7300HQ CPU; the system's resource utilization was monitored by Windows Performance Monitor tool (WPMT) (per). During benchmarking, only the software being tested was running. MOG's efficiency was compared to that of one of the R-based "shiny" web-app (Chang et al., 2018) (choosing PIVOT, because it permits loading normalized data).

PIVOT repeatedly crashed and failed to load the full *Hu-cancer-RNASeq-dataset* (Additional File 6), but was able to load a subset of data consisting only of 410 tumor and non-tumor liver

samples. We measured the execution time (time taken to compute and display output) of the Mann-Whitney U test for differentially-expressed genes in tumor vs. non-tumor samples. The test completed in 21 minutes with PIVOT, but only seven seconds with MOG (Figure 3.8). We kept MOG running idle until total runtime reached 30 minutes and compared memory and processor usage (Supplementary File 8); average memory usage of PIVOT (1,869 MB) was about twice that of MOG (995 MB) (Supplementary File 8). Peak % processor time (CPU) was greater for MOG; however, MOG completed its task much more quickly, and over the 30 minutes, the *average* % processor time was 64% for PIVOT but only 2% for MOG (Figure 3.8 A).

We benchmarked MOG's performance on datasets of different sizes, created by splitting the *Hu-cancer-RNASeq-dataset* by organ type (tumor and non-tumor samples). For each dataset, we performed a Mann-Whitney U test on all the genes for tumor vs. non-tumor groups. MOG took only 31 seconds to compute a Mann-Whitney U test on 18,212 genes over 1,054 samples (Figure 3.8 B; Additional File 6). We then measured the execution time for calculating Pearson correlations of one gene with all others. MOG took only a couple of seconds to compute a Pearson correlation over 1,000 samples and 16 seconds to compute over 7,142 samples (Figure 3.8 C).

3.5 Discussion and Conclusion

We demonstrated MOG's functionalities by exploring three different well-validated datasets: a human RNA-Seq dataset from non-tumor and tumor samples (*Hu-cancer-RNASeq-dataset*), an *A. thaliana* microarray dataset (*AT-microarray-dataset*), and an *A. thaliana* metabolomics dataset (*AT-metab-dataset*). In each case, known information was recapitulated in the MOG analysis, and new potential relationships became apparent.

During exploration of the *Hu-cancer-RNASeq-dataset* by MOG, we created a catalog of genes that are differentially expressed in different types of tumors, identifying in this process 35 genes that are consistently upregulated or downregulated in every type of cancer in the dataset. GPC3 (Ho and Kim, 2011; Anatelli et al., 2008; Gao and Ho, 2011) was identified by MOG as a biomarker gene for liver cancer. Gene-level resolution analysis by MOG revealed that the cadre of

genes that are coexpressed with the GPC3 gene change drastically among the individual organs, and between tumor samples and corresponding non-tumor samples. By mining the sample and study metadata, we identified genes that showed regulation with cancer progression. Many of the genes we identified have been reported previously in the literature and in THPA to be prognostic biomarkers for different cancers. Many other genes that MOG identified as differentially expressed genes are *not marked as prognostic in THPA*. These genes present potential new biomarkers for disease progression. Because each tumor type has many variations, investigating multiple candidate prognostic markers in individual tumors can provide critical information for personalized medicine (Cieślik and Chinnaiyan, 2018).

Using the *AT-microarray-dataset*, we explored expression patterns of genes with unknown functions including orphan genes, identifying 18 mostly plant-restricted genes that are tightly coexpressed with genes central to photosystem assembly. We also identified an Arabidopsis-specific gene, At2G04675, to be highly expressed in pollen development, suggesting a potential involvement of this gene in gametogenesis. With the *AT-metab-dataset*, we identified a potential relationship between arginine and the cell wall defense response. Such exploratory analyses provide clues as to how to approach experimentally testing the function of these genes or metabolites.

Processing multiple heterogeneous RNA-Seq data is a formidable and unsolved challenge. We have intentionally not added capabilities for data processing (e.g., alignment, normalization, and batch-correction to minimize unwanted technical and biological effects) into MOG for two reasons. First, the selection of appropriate statistical and computational methods depends on the data structure and the biological questions to be asked. Different types of data have different characteristics (Chawade et al., 2014; Hicks et al., 2017; Evans et al., 2017), and if statistical methods are misapplied during normalization and batch-correction, especially when the data are from multiple heterogeneous studies, the resultant dataset may be misleading. Much as if using R or MATLAB statistical software, a MOG user must consider these technicalities. Second, the data science field is far from unsettled (Evans et al., 2017; Paulson et al., 2017; Hicks et al., 2017;

Schmidt et al., 2018) with new approaches and variations being developed each year. (GoogleScholar retrieved over 10,000 journal articles from the first half of 2019 for "RNA-Seq normalization methods"). Potentially a researcher could use MOG as a tool to compare the results of different methods of processing the same raw data. Such interactive comparisons would enable biologists to gain insight as to which processing methods best reflect experimentally-established "ground truths". This approach would provide a complement to the more typical validation of a dataset by determining GO term enrichment in gene clusters.

Analyses performed while exploring and statistically analyzing datasets on MOG can be saved; by clicking "save", all the analyses that have been performed are added as objects to the MOG project file. Results obtained with MOG can be shared by sharing the saved MOG project file. If a user wishes to document the information to reproduce the analysis, she/he needs to manually specify the parameters and methods used. In the future, we plan to implement automated report generation for each analysis.

MOG is a novel Java software for interactive exploratory analysis of big 'omics datasets or other datasets. By using an indexing strategy to read data only when it is needed, the total memory consumption of the system is minimized, enabling MOG to perform much more efficiently than the available R-based software. Visualizations produced by MOG are fully interactive, and enable researchers to detect and mine interesting data points and probe the relationships among them. The statistical methods implemented in MOG help to guide exploration of hidden patterns in a user-friendly manner. By integrating metadata, MOG affords an opportunity to extract new insights into the relationships between gene expression and gene structure, gene location, or any of the diverse information entered by scientists about the biology and experimental conditions.

Taken together these features can aid a researcher in developing new, experimentally testable hypotheses.

3.6 Data Availability

We subscribe to FAIR data and software practices (Wilkinson et al., 2016). MOG is free and open source software published under the MIT License. MOG software, user guide, and all compiled datasets in this article are freely downloadable from http://metnetweb.gdcb.iastate.edu/MetNet_MetaOmGraph.htm. MOG's source code and user guide is available at https://github.com/urmi-21/MetaOmGraph/. MOG's source code (version 1.8.0) at the time of submission is archived and can be accessed using the DOI:10.5281/zenodo.3520986. Additional files are available at https://github.com/urmi-21/MetaOmGraph/tree/master/MOG_SupportingData.

3.7 Supplementary Data

Supplementary Data are available at NAR online.

3.8 Funding

This work is funded in part by National Science Foundation grant IOS 1546858, Orphan Genes: An Untapped Genetic Reservoir of Novel Traits, and by the Center for Metabolic Biology, Iowa State University.

3.9 Acknowledgements

We especially thank Nick Ransom for his formative role in MOG's early development. We are grateful to our collaborators, Kevin Bassler, Pramesh Singh, and Ling Li, for their help and feedback. We much appreciate the efforts of Jing Li, Priyanka Bhandhary, Arun Seetharam, and the early-adopters who beta-tested MOG and provided valuable feedback.

3.10 References

Overview of windows performance monitor. *Microsoft Docs*.

Almeida-de Macedo, M. M., Ransom, N., Feng, Y., Hurst, J., and Wurtele, E. S. (2013). Comprehensive analysis of correlation coefficients estimated from pooling heterogeneous microarray data. *BMC bioinformatics*, 14(1):1–14.

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., and Hamosh, A. (2014). Omim. org: Online mendelian inheritance in man (omim®), an online catalog of human genes and genetic disorders. *Nucleic acids research*, 43(D1):D789–D798.

Anatelli, F., Chuang, S.-T., Yang, X. J., and Wang, H. L. (2008). Value of glypican 3 immunostaining in the diagnosis of hepatocellular carcinoma on needle biopsy. *American journal of clinical pathology*, 130(2):219–223.

Arendsee, Z., Li, J., Singh, U., Seetharam, A., Dorman, K. S., and Wurtele, E. S. (2019). phylostratr: a framework for phylostratigraphy. *Bioinformatics*, 35(19):3617–3627.

Arendsee, Z. W., Li, L., and Wurtele, E. S. (2014). Coming of age: orphan genes in plants. *Trends in plant science*, 19(11):698–708.

Bader, G. D. and Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics*, 4(1):2.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Bhuiyan, N. H., Friso, G., Poliakov, A., Ponnala, L., and van Wijk, K. J. (2015). Met1 is a thylakoid-associated tpr protein involved in photosystem ii supercomplex formation and repair in arabidopsis. *The Plant Cell*, 27(1):262–285.

Blackhall, F. H., Merry, C. L., Davies, E., and Jayson, G. C. (2001). Heparan sulfate proteoglycans and cancer. *British journal of cancer*, 85(8):1094.

Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M., and Sherlock, G. (2004). Go:: Termfinder—open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics*, 20(18):3710–3715.

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G. G., et al. (2003). Arrayexpress—a public repository for microarray gene expression data at the ebi. *Nucleic acids research*, 31(1):68–71.

Brazma, A. and Vilo, J. (2000). Gene expression data analysis. *FEBS letters*, 480(1):17–24.

Capurro, M., Wanless, I. R., Sherman, M., Deboer, G., Shi, W., Miyoshi, E., and Filmus, J. (2003). Glypican-3: a novel serum and histochemical marker for hepatocellular carcinoma. *Gastroenterology*, 125(1):89–97.

Capurro, M. I., Xu, P., Shi, W., Li, F., Jia, A., and Filmus, J. (2008). Glypican-3 inhibits hedgehog signaling during development by competing with patched for hedgehog binding. *Developmental cell*, 14(5):700–711.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2018). *shiny: Web Application Framework for R*. R package version 1.2.0.

Chawade, A., Alexandersson, E., and Levander, F. (2014). Normalyzer: a tool for rapid evaluation of normalization methods for omics data sets. *Journal of proteome research*, 13(6):3114–3120.

Chen, S.-T., Liu, D.-W., Lin, J.-D., Chen, F.-W., Huang, Y.-Y., and Hsu, B. R.-S. (2012). Down-regulation of matrix metalloproteinase-7 inhibits metastasis of human anaplastic thyroid cancer cell line. *Clinical & experimental metastasis*, 29(1):71–82.

Chen, Y.-E., Su, Y.-Q., Mao, H.-T., Nan, W., Zhu, F., Yuan, M., Zhang, Z.-W., Liu, W.-J., and Yuan, S. (2018). Terrestrial plants evolve highly-assembled photosystem complexes in adaptation to light shifts. *Frontiers in plant science*, 9:1811.

Choi, K. and Ratner, N. (2019). igeak: an interactive gene expression analysis kit for seamless workflow using the r/shiny platform. *BMC genomics*, 20(1):177.

Cieślik, M. and Chinnaiyan, A. M. (2018). Cancer transcriptome profiling at the juncture of clinical translation. *Nature Reviews Genetics*, 19(2):93.

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T. S., Malta, T. M., Pagnotta, S. M., Castiglioni, I., et al. (2015). Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data. *Nucleic acids research*, 44(8):e71–e71.

Daub, C. O., Steuer, R., Selbig, J., and Kloska, S. (2004). Estimating mutual information using b-spline functions–an improved similarity measure for analysing gene expression data. *BMC bioinformatics*, 5(1):118.

Davoodi, J., Kelly, J., Gendron, N. H., and MacKenzie, A. E. (2007). The simpson–golabi–behmel syndrome causative glypican-3, binds to and inhibits the dipeptidyl peptidase activity of cd26. *Proteomics*, 7(13):2300–2310.

de Vries, J. K., Levin, A., Loud, F., Adler, A., Mayer, G., and Pena, M. J. (2018). Implementing personalized medicine in diabetic kidney disease: Stakeholders' perspectives. *Diabetes, Obesity and Metabolism*, 20:24–29.

del Rıo, L. A., Corpas, F. J., and Barroso, J. B. (2004). Nitric oxide and nitric oxide synthase activity in plants. *Phytochemistry*, 65(7):783–792.

Edgington, E. S. (1980). Validity of randomization tests for one-subject experiments. *Journal of Educational Statistics*, 5(3):235–251.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868.

Eun, H. S., Cho, S. Y., Lee, B. S., Kim, S., Song, I.-S., Chun, K., Oh, C.-H., Yeo, M.-K., Kim, S. H., and Kim, K.-H. (2019). Cytochrome p450 4a11 expression in tumor cells: A favorable prognostic factor for hepatocellular carcinoma patients. *Journal of Gastroenterology and Hepatology*, 34(1):224–233.

Evans, C., Hardin, J., and Stoebel, D. M. (2017). Selecting between-sample rna-seq normalization methods from the perspective of their assumptions. *Briefings in bioinformatics*, 19(5):776–792.

Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J., and Gardner, T. S. (2007). Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biology*, 5(1):e8.

Fan, X., Zhang, J., Li, W., and Peng, L. (2015). The ndhv subunit is required to stabilize the chloroplast nadh dehydrogenase-like complex in arabidopsis. *The Plant Journal*, 82(2):221–231.

Field, A. P. (2001). Meta-analysis of correlation coefficients: a monte carlo comparison of fixed-and random-effects methods. *Psychological methods*, 6(2):161.

Filmus, J. and Capurro, M. (2008). The role of glypican-3 in the regulation of body size and cancer. *Cell Cycle*, 7(18):2787–2790.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521.

Fukushima, A. (2013). Diffcorr: an r package to analyze and visualize differential correlations in biological networks. *Gene*, 518(1):209–214.

Fukushima, A., Kusano, M., Mejia, R. F., Iwasa, M., Kobayashi, M., Hayashi, N., Watanabe-Takahashi, A., Narisawa, T., Tohge, T., Hur, M., et al. (2014). Metabolomic characterization of knockout mutants in arabidopsis: development of a metabolite profiling database for knockout mutants in arabidopsis. *Plant physiology*, 165(3):948–961.

Furtună, T. F. and Vinte, C. (2016). Integrating r and java for enhancing interactivity of algorithmic data analysis software solutions. *Rom. Stat. Rev*, 64:29–41.

Gao, W. and Ho, M. (2011). The role of glypican-3 in regulating wnt in hepatocellular carcinomas. *Cancer reports*, 1(1):14.

Gollery, M., Harper, J., Cushman, J., Mittler, T., Girke, T., Zhu, J.-K., Bailey-Serres, J., and Mittler, R. (2006). What makes species unique? the contribution of proteins with obscure features. *Genome biology*, 7(7):R57.

Haug, K., Salek, R. M., Conesa, P., Hastings, J., de Matos, P., Rijnbeek, M., Mahendraker, T., Williams, M., Neumann, S., Rocca-Serra, P., et al. (2012). Metabolights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic acids research*, 41(D1):D781–D786.

Hedges, L. V. and Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological methods*, 3(4):486.

Hickey, P. (2019). *DelayedMatrixStats: Functions that Apply to Rows and Columns of 'DelayedMatrix' Objects*. R package version 1.6.0.

Hicks, S. C., Okrah, K., Paulson, J. N., Quackenbush, J., Irizarry, R. A., and Bravo, H. C. (2017). Smooth quantile normalization. *Biostatistics*, 19(2):185–198.

Ho, M. and Kim, H. (2011). Glypican-3: a new target for cancer immunotherapy. *European journal of cancer*, 47(3):333–338.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 6(2):65–70.

Huang, Y., Prasad, M., Lemon, W. J., Hampel, H., Wright, F. A., Kornacker, K., LiVolsi, V., Frankel, W., Kloos, R. T., Eng, C., et al. (2001). Gene expression in papillary thyroid carcinoma reveals highly consistent profiles. *Proceedings of the National Academy of Sciences*, 98(26):15044–15049.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., et al. (2002). The ensembl genome database project. *Nucleic acids research*, 30(1):38–41.

Hur, M., Campbell, A. A., Almeida-de Macedo, M., Li, L., Ransom, N., Jose, A., Crispin, M., Nikolau, B. J., and Wurtele, E. S. (2013). A global approach to analysis and interpretation of metabolic data for plant natural product discovery. *Natural product reports*, 30(4):565–583.

Ignatchenko, V., Ignatchenko, A., Sinha, A., Boutros, P. C., and Kislinger, T. (2015). Venndis: A javafx-based venn and euler diagram software to generate publication quality figures. *Proteomics*, 15(7):1239–1244.

Ihaka, R. and Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, 5(3):299–314.

Kaur, S. P. and Cummings, B. S. (2019). Role of glypicans in regulation of the tumor microenvironment and cancer progression. *Biochemical Pharmacology*.

Kelder, T., Conklin, B. R., Evelo, C. T., and Pico, A. R. (2010). Finding the right questions: exploratory pathway analysis to enhance biological discovery in large datasets. *PLoS biology*, 8(8):e1000472.

Kersey, P. J., Allen, J. E., Armean, I., Boddu, S., Bolt, B. J., Carvalho-Silva, D., Christensen, M., Davis, P., Falin, L. J., Grabmueller, C., et al. (2015). Ensembl genomes 2016: more genomes, more complexity. *Nucleic acids research*, 44(D1):D574–D580.

Kim, H., Xu, G.-L., Borczuk, A. C., Busch, S., Filmus, J., Capurro, M., Brody, J. S., Lange, J., D'armiento, J. M., Rothman, P. B., et al. (2003). The heparan sulfate proteoglycan gpc3 is a potential lung tumor suppressor. *American journal of respiratory cell and molecular biology*, 29(6):694–701.

Kirov, I., Khrustaleva, L., Van Laere, K., Soloviev, A., Meeus, S., Romanov, D., and Fesenko, I. (2017). Drawid: user-friendly java software for chromosome measurements and idiogram drawing. *Comparative cytogenetics*, 11(4):747.

Kodama, Y., Shumway, M., and Leinonen, R. (2011). The sequence read archive: explosive growth of sequencing data. *Nucleic acids research*, 40(D1):D54–D56.

Kucukural, A., Yukselen, O., Ozata, D. M., Moore, M. J., and Garber, M. (2019). Debrowser: interactive differential expression analysis and visualization tool for count data. *BMC genomics*, 20(1):6.

Kumari, S., Nie, J., Chen, H.-S., Ma, H., Stewart, R., Li, X., Lu, M.-Z., Taylor, W. M., and Wei, H. (2012). Evaluation of gene association methods for coexpression network construction and biological knowledge discovery. *PloS one*, 7(11):e50411.

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., et al. (2011). The arabidopsis information resource (tair): improved gene annotation and new tools. *Nucleic acids research*, 40(D1):D1202–D1210.

Langfelder, P. and Horvath, S. (2008). Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559.

Lazar, C., Meganck, S., Taminau, J., Steenhoff, D., Coletta, A., Molter, C., Weiss-Solís, D. Y., Duque, R., Bersini, H., and Nowé, A. (2012). Batch effect removal methods for microarray gene expression data integration: a survey. *Briefings in bioinformatics*, 14(4):469–490.

Li, J., Arendsee, Z., Singh, U., and Wurtele, E. S. (2019). Recycling rna-seq data to identify candidate orphan genes for experimental analysis. *bioRxiv*, page 671263.

Lightbody, G., Haberland, V., Fiona, B., Taggart, L., Zheng, H., Parks, E., and Blayney, J. (2018). Review of applications of high-throughput sequencing in personalised medicine: Barriers and facilitators of future progress in research and clinical application. *Briefings in bioinformatics*, page bby051.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580.

López-Fernández, H., Reboiro-Jato, M., Glez-Peña, D., Laza, R., Pavón, R., and Fdez-Riverola, F. (2018). Gc4s: A bioinformatics-oriented java software library of reusable graphical user interface components. *PloS one*, 13(9):e0204474.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550.

Luo, D., Chen, H., Lu, P., Li, X., Long, M., Peng, X., Huang, M., Huang, K., Lin, S., Tan, L., et al. (2017). Chi3l1 overexpression is associated with metastasis and is an indicator of poor prognosis in papillary thyroid carcinoma. *Cancer Biomarkers*, 18(3):273–284.

Ma, T., Huo, Z., Kuo, A., Zhu, L., Fang, Z., Zeng, X., Lin, C.-W., Liu, S., Wang, L., Liu, P., et al. (2018). Metaomics: analysis pipeline and browser-based software suite for transcriptomic meta-analysis. *Bioinformatics*, 35(9):1597–1599.

Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., Geer, R. C., He, J., Gwadz, M., Hurwitz, D. I., et al. (2014). Cdd: Ncbi's conserved domain database. *Nucleic acids research*, 43(D1):D222–D226.

Marini, F. (2018). *ideal: Interactive Differential Expression AnaLysis*.

Martens, L., Hermjakob, H., Jones, P., Adamski, M., Taylor, C., Gevaert, K., Vandekerckhove, J., Apweiler, R., et al. (2005). Pride: the proteomics identifications database. *Proteomics*, 5(13):3537–3545.

McKenzie, A. T., Katsyv, I., Song, W.-M., Wang, M., and Zhang, B. (2016). Dgca: a comprehensive r package for differential gene correlation analysis. *BMC systems biology*, 10(1):106.

Mentzen, W. I. and Wurtele, E. S. (2008). Regulon organization of arabidopsis. *BMC plant biology*, 8(1):99. MCL clustering.

Monier, B., McDermaid, A., Wang, C., Zhao, J., Miller, A., Fennell, A., and Ma, Q. (2019). Iris-eda: An integrated rna-seq interpretation system for gene expression data analysis. *PLoS computational biology*, 15(2):e1006792.

Nandi, P., Lim, H., Torres-Garcia, E. J., and Lala, P. K. (2018). Human trophoblast stem cell self-renewal and differentiation: role of decorin. *Scientific reports*, 8(1):8977.

Nosek, L., Semchonok, D., Boekema, E. J., Ilík, P., and Kouřil, R. (2017). Structural variability of plant photosystem ii megacomplexes in thylakoid membranes. *The Plant Journal*, 89(1):104–111.

Pagès, H., with contributions from Peter Hickey, and Lun, A. (2019). *DelayedArray: A unified framework for working transparently with on-disk and in-memory array-like datasets*. R package version 0.10.0.

Pan, W., Cheng, Y., Zhang, H., Liu, B., Mo, X., Li, T., Li, L., Cheng, X., Zhang, L., Ji, J., et al. (2014). Csbf/c10orf99, a novel potential cytokine, inhibits colon cancer cell growth through inducing g1 arrest. *Scientific reports*, 4:6812.

Paulson, J. N., Chen, C.-Y., Lopes-Ramos, C. M., Kuijjer, M. L., Platig, J., Sonawane, A. R., Fagny, M., Glass, K., and Quackenbush, J. (2017). Tissue-aware rna-seq processing and normalization for heterogeneous and sparse data. *BMC bioinformatics*, 18(1):437.

Price, A., Caciula, A., Guo, C., Lee, B., Morrison, J., Rasmussen, A., Lipkin, W. I., and Jain, K. (2019). Devis: an r package for aggregation and visualization of differential expression data. *BMC bioinformatics*, 20(1):110.

Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2006). Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, 35(suppl_1):D61–D65.

Qiu, J., Zhang, W., Zang, C., Liu, X., Liu, F., Ge, R., Sun, Y., and Xia, Q. (2018). Identification of key genes and miRNAs markers of papillary thyroid cancer. *Biological research*, 51(1):45.

Quanbeck, S. M. M., Brachova, L., Campbell, A. A., Guan, X., Perera, A., He, K., Rhee, S. Y., Bais, P., Dickerson, J., Dixon, P., et al. (2012). Metabolomics as a hypothesis-generating functional genomics tool for the annotation of arabidopsis thaliana genes of "unknown function". *Frontiers in plant science*, 3:15.

Rau, A., Marot, G., and Jaffrézic, F. (2014). Differential meta-analysis of rna-seq data from multiple studies. *BMC bioinformatics*, 15(1):91.

Ren, X., Ji, Y., Jiang, X., and Qi, X. (2018). Downregulation of cyp2a6 and cyp2c8 in tumor tissues is linked to worse overall survival and recurrence-free survival from hepatocellular carcinoma. *BioMed research international*, 2018.

Rhodes, D. R. and Chinnaiyan, A. M. (2005). Integrative analysis of the cancer transcriptome. *Nature genetics*, 37(6s):S31.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

Ruban, A. V. and Johnson, M. P. (2015). Visualizing the dynamic structure of the plant photosynthetic membrane. *Nature plants*, 1(11):15161.

Rue-Albrecht, K., Marini, F., Soneson, C., and Lun, A. T. (2018). isee: interactive summarizedexperiment explorer. *F1000Research*, 7.

Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., Nativ, N., Bahir, I., Doniger, T., Krug, H., et al. (2010). Genecards version 3: the human gene integrator. *Database*, 2010.

Sasisekharan, R., Shriver, Z., Venkataraman, G., and Narayanasami, U. (2002). Roles of heparan-sulphate glycosaminoglycans in cancer. *Nature Reviews Cancer*, 2(7):521.

Schmidt, F., List, M., Cukuroglu, E., Köhler, S., Göke, J., and Schulz, M. H. (2018). An ontology-based method for assessing batch effect adjustment approaches in heterogeneous datasets. *Bioinformatics*, 34(17):i908–i916.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504.

Shannon, P. T., Grimes, M., Kutlu, B., Bot, J. J., and Galas, D. J. (2013). Rcytoscape: tools for exploratory network analysis. *BMC bioinformatics*, 14(1):217.

Singh, P., Chen, T., Arendsee, Z., Wurtele, E. S., and Bassler, K. E. (2017). A Regulatory Network Analysis of Orphan Genes in Arabidopsis Thaliana. In *APS March Meeting Abstracts*, page V6.005.

Slattery, M. L., Mullany, L. E., Wolff, R. K., Sakoda, L. C., Samowitz, W. S., and Herrick, J. S. (2018). The p53-signaling pathway and colorectal cancer: Interactions between downstream p53 target genes and mirnas. *Genomics*.

Sondka, Z., Bamford, S., Cole, C. G., Ward, S. A., Dunham, I., and Forbes, S. A. (2018). The cosmic cancer gene census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer*, page 1.

Soneson, C. and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of rna-seq data. *BMC bioinformatics*, 14(1):91.

Song, L., Langfelder, P., and Horvath, S. (2012). Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC bioinformatics*, 13(1):328.

Sumner, L. W., Lei, Z., Nikolau, B. J., and Saito, K. (2015). Modern plant metabolomics: advanced natural product gene discoveries, improved technologies, and future prospects. *Natural product reports*, 32(2):212–229.

Sun, X., Zhang, H., Luo, L., Zhong, K., Ma, Y., Fan, L., Fu, D., and Wan, L. (2016). Comparative proteomic profiling identifies potential prognostic factors for human clear cell renal cell carcinoma. *Oncology reports*, 36(6):3131–3138.

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). Revigo summarizes and visualizes long lists of gene ontology terms. *PloS one*, 6(7):e21800.

Swe, M. T., Pongchaidecha, A., Chatsudthipong, V., Chattipakorn, N., and Lungkaphin, A. (2019). Molecular signaling mechanisms of renal gluconeogenesis in nondiabetic and diabetic conditions. *Journal of cellular physiology*, 234(6):8134–8151.

Tian, Z.-H., Yuan, C., Yang, K., and Gao, X.-L. (2019). Systematic identification of key genes and pathways in clear cell renal cell carcinoma on bioinformatics analysis. *Annals of translational medicine*, 7(5).

Trevino, S., Sun, Y., Cooper, T. F., and Bassler, K. E. (2012). Robust detection of hierarchical communities from *Escherichia coli* gene expression data. *PLoS Comput. Biol.*, 8(2):e1002391.

Tukey, J. W. (1977). *Exploratory data analysis*, volume 2. Reading, Mass.

Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., Benfeitas, R., Arif, M., Liu, Z., Edfors, F., et al. (2017). A pathology atlas of the human cancer transcriptome. *Science*, 357(6352):eaan2507.

Valsechi, M. C., Oliveira, A. B. B., Conceição, A. L. G., Stuqui, B., Candido, N. M., Provazzi, P. J. S., de Araújo, L. F., Silva, W. A., de Freitas Calmon, M., and Rahal, P. (2014). Gpc3 reduces cell proliferation in renal carcinoma cell lines. *BMC cancer*, 14(1):631.

van Dam, S., Vosa, U., van der Graaf, A., Franke, L., and de Magalhaes, J. P. (2017). Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in bioinformatics*, 19(4):575–592.

Vandenbon, A., Dinh, V. H., Mikami, N., Kitagawa, Y., Teraguchi, S., Ohkura, N., and Sakaguchi, S. (2016). Immuno-navigator, a batch-corrected coexpression database, reveals cell type-specific gene networks in the immune system. *Proceedings of the National Academy of Sciences*, 113(17):E2393–E2402.

Vander Ark, A., Cao, J., and Li, X. (2018). Tgf-β receptors: In and beyond tgf-β signaling. *Cellular signalling*, 52:112–120.

Wang, Q., Armenia, J., Zhang, C., Penson, A. V., Reznik, E., Zhang, L., Minet, T., Ochoa, A., Gross, B. E., Iacobuzio-Donahue, C. A., et al. (2018). Unifying cancer and normal rna sequencing data from different sources. *Scientific data*, 5:180061.

Wang, Y. R. and Huang, H. (2014). Review on statistical methods for gene network reconstruction using expression data. *Journal of theoretical biology*, 362:53–61.

Weisstein, E. W. (2004). Bonferroni correction.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

Xiang, Y.-Y., Ladeda, V., and Filmus, J. (2001). Glypican-3 expression is silenced in human breast cancer. *Oncogene*, 20(50):7408.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. (2002). Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic acids research*, 30(4):e15–e15.

You, J., Chen, W., Chen, J., Zheng, Q., Dong, J., and Zhu, Y. (2018). The oncogenic role of arg1 in progression and metastasis of hepatocellular carcinoma. *BioMed research international*, 2018.

Ysuhiro, I., Hiroshi, Y., Kennichi, K., Yasushi, N., Kanji, K., and Akira, M. (2006). Inverse relationships between the expression of mmp-7 and mmp-11 and predictors of poor prognosis of papillary thyroid carcinoma. *Pathology*, 38(5):421–425.

Yu, T., Wang, X., Zhu, G., Han, C., Su, H., Liao, X., Yang, C., Qin, W., Huang, K., and Peng, T. (2018). The prognostic value of differentially expressed cyp3a subfamily members for hepatocellular carcinoma. *Cancer management and research*, 10:1713.

Zhan, S., Li, J., Wang, T., and Ge, W. (2018). Quantitative proteomics analysis of sporadic medullary thyroid cancer reveals fn1 as a potential novel candidate prognostic biomarker. *The oncologist*, 23(12):1415–1425.

Zhang, H., Cai, Y., Zheng, L., Zhang, Z., Lin, X., and Jiang, N. (2018a). Long noncoding rna neat1 regulate papillary thyroid cancer progression by modulating mir-129-5p/klk7 expression. *Journal of cellular physiology*, 233(10):6638–6648.

Zhang, Y., Hu, J., Zhou, W., and Gao, H. (2018b). Lncrna foxd2-as1 accelerates the papillary thyroid cancer progression through regulating the mir-485-5p/klk7 axis. *Journal of cellular biochemistry*.

Zhu, Q., Fisher, S. A., Dueck, H., Middleton, S., Khaladkar, M., and Kim, J. (2018). Pivot: platform for interactive analysis and visualization of transcriptomics data. *BMC bioinformatics*, 19(1):6.

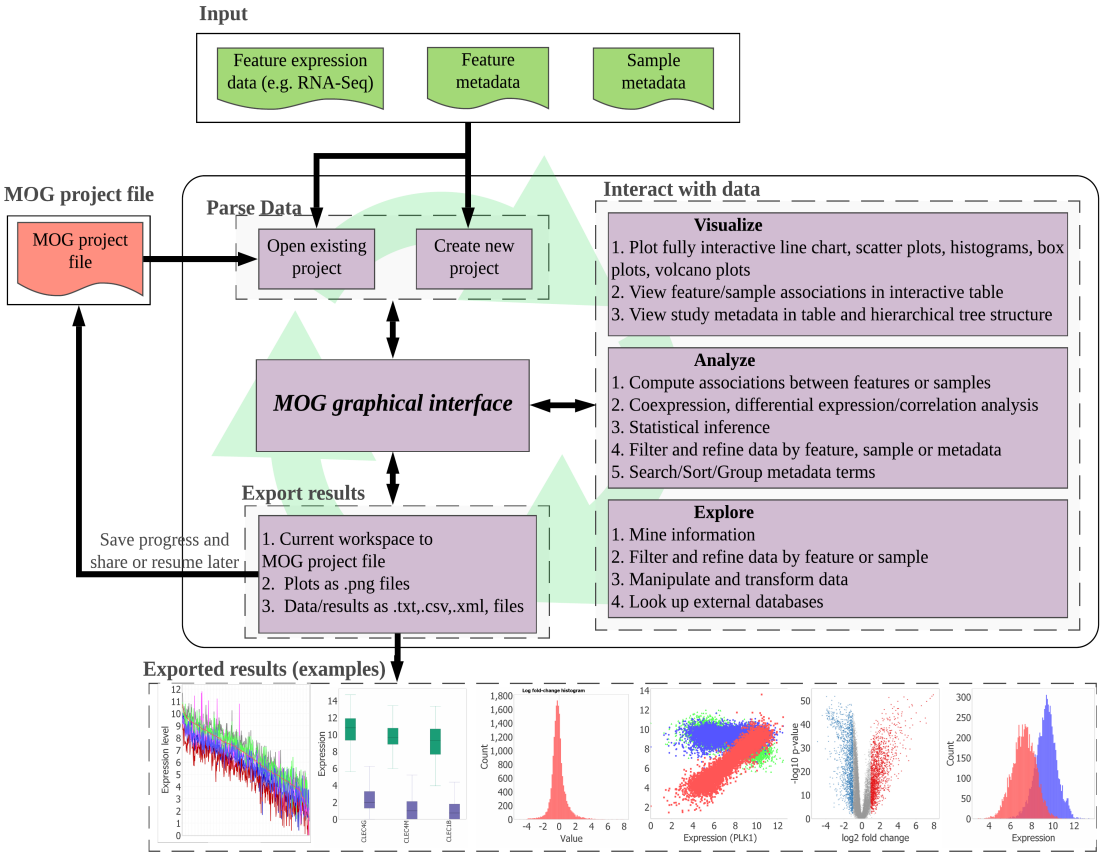

Figure 3.1 An overview of MOG's modules. All functionality is accessed through MOG's graphical user interface (GUI). First, the researcher selects an existing MOG project or creates a new MOG project (.mog) with input data files. Once the project is open in MOG, the workflow is non-linear. The GUI enables interactive exploration of data through a choice of statistical analyses and data visualizations. The researcher can export visualizations and results throughout the analysis, and can save her/his feature lists and statistical analyses in the MOG project file for future exploration. Saved MOG projects can be shared and further analyzed by new researchers.

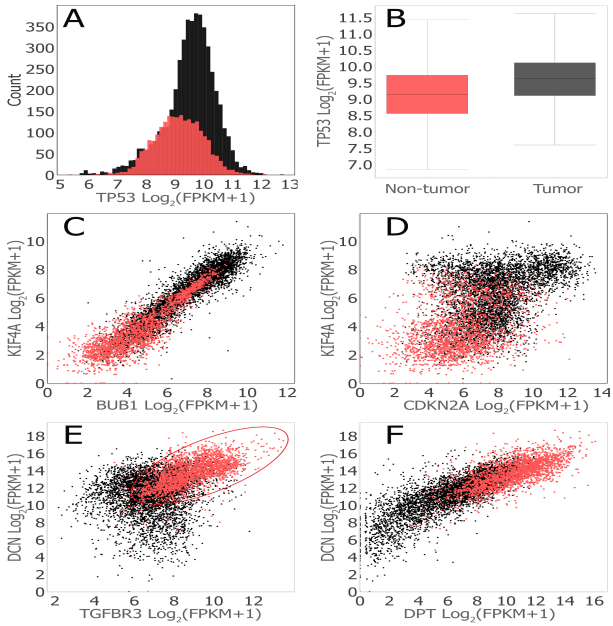

Figure 3.2 MOG visualizations of expression of selected genes across all tumor types and non-tumor samples. Tumor samples, black dots; non-tumor samples, red dots. Correlations and differential expression analyses were performed using MOG (Mann-Whitney U test, $|FC| \geq 2$, BH corrected p-value $< 10^3$). **(A)** Histogram showing the distribution of Tumor Protein P53 (TP53) expression (number of bins set to 50). **(B)** Box plot summarizing the expression of TP53 over all tumor vs. all non-tumor samples. The horizontal line represents the median log expression, which is 9.1 for non-tumor samples and 9.6 for tumor samples. **(C)** Scatter plot visualizing coexpression of mitotic checkpoint serine/threonine kinase (BUB1) and kinesin family member 4A (KIF4A) (both are upregulated across all tumor types). **(D)** Scatter plot visualizing coexpression of genes cyclin dependent kinase inhibitor 2A (CDKN2A) and KIF4A. Both are upregulated across all tumor types, but they are are not co-expressed. **(E)** Scatter plots visualizing transforming growth factor beta receptor3 (TGFBR3), which has a complex role as regulator of angiogenesis (Vander Ark et al., 2018), decorin (DCN), autophagy, mitophagy and embryonic cell development including endovascular differentiation (Nandi et al., 2018). TGFBR3 and DCN are downregulated across all tumor types and are coexpressed in non-tumor samples (Spearman correlation= 0.64) but not in tumor cells (Spearman correlation= 0.14). The coexpression of TGFBR3 and DCN in only the non-tumor samples suggests that the processes in which each gene participates are associated under normal conditions. **(F)** Scatter plot visualizing coexpression of genes dermatopontin (DPT) and DCN Spearman correlations are 0.82 (tumor samples), 0.69 (non-tumor samples), and 0.84 (combined samples) (Both gene are downregulated across all tumor types.)

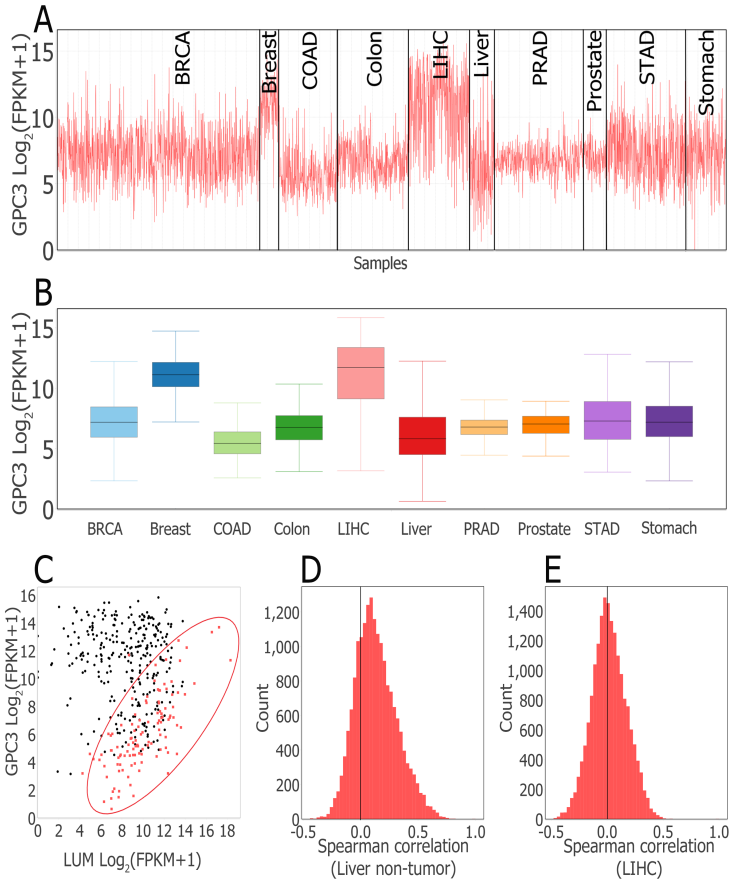

Figure 3.3 MOG visualizations of glypican 3 (GPC3) expression pattern in tumor and non-tumor organs. **(A)** Line chart generated by interactively filtering by study metadata to retain 3,184 samples from 5 tumor types and corresponding non–tumor organs, and grouping the chart by organ/tumor type. **(B)** Box plot summary of data in **(A)**. Generated by interactively splitting box plot according to organ/tumor type. **(C)** Scatter plot showing co-expression of GPC3 and Lumican (LUM) in liver non-tumor and LIHC samples. In non-tumor liver (red), GPC3 and LUM expression are strongly correlated (Spearman correlation≤ 0.7). In LIHC samples (black), GPC3 and LUM expression show no association (Spearman correlation$= -0.1$). **(D and E)** Histograms of distribution of Spearman correlation coefficients of expression of GPC3 with all other genes. Non-tumor liver samples (D), LIHC samples (E). The longer right tail of non-tumor liver samples indicates Spearman correlation coefficients of GPC3 expression with selected genes are higher in non-tumor than LIHC samples.

Table 3.3 Genes identified by MOG as showing changing expression with cancer progression (B-H corrected p-value < 0.05) that had been identified in experimental studies as potential prognostic biomarkers but were not marked as prognostic for the given cancer type in The Human Protein Atlas (THPA) (Uhlen et al., 2017).

Disease	**Gene**	**Gene name**	**Pattern**	**Ref.**
LIHC	ARG1	arginase 1	Decreasing	(You et al., 2018)
LIHC	CYP2C8	cytochrome P450 family 2 subfamily C member 8	Decreasing	(Ren et al., 2018)
LIHC	CYP3A4	cytochrome P450 family 3 subfamily A member 4	Decreasing	(Yu et al., 2018)
LIHC	CYP3A7	cytochrome P450 family 3 subfamily A member 7	Decreasing	(Yu et al., 2018)
LIHC	CYP4A11	cytochrome P450 family 4 subfamily A member 11	Decreasing	(Eun et al., 2019)
THCA	CHI3L1	chitinase 3 like 1	Increasing	(Luo et al., 2017)
THCA	SFTPB	surfactant protein B	Increasing	(Huang et al., 2001)
THCA	CD207	CD207 molecule	Increasing	(Qiu et al., 2018)
THCA	MUC21	mucin 21, cell surface associated	Increasing	(Qiu et al., 2018)
THCA	MMP7	matrix metallopeptidase 7	Increasing	(Ysuhiro et al., 2006; Chen et al., 2012)
THCA	IGFL2	IGF like family member 2	Increasing	(Qiu et al., 2018)
THCA	KLK7	kallikrein related peptidase 7	Increasing	(Zhang et al., 2018a,b)
THCA	FN1	fibronectin 1	Increasing	(Zhan et al., 2018)

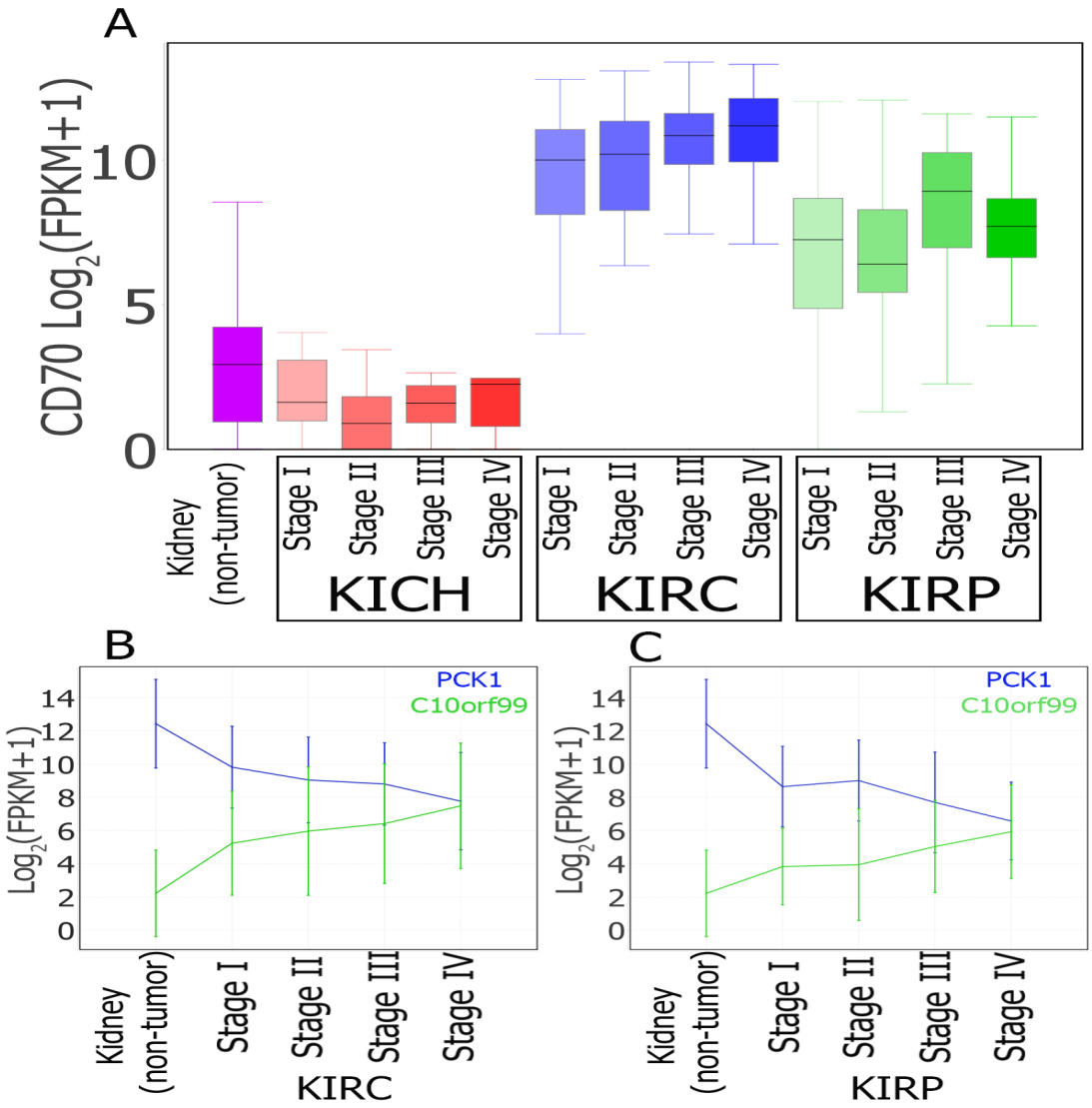

Figure 3.4 MOG visualization of expression of selected genes during progression of three types of renal cancer. **(A)** Box plots summarizing CD70 expression in non–tumor kidney and in different stages of KICH, KIRC, and KIRP cancer progression. CD70 is designated as prognostic unfavourable for renal cancer by THPA (Uhlen et al., 2017). However, although CD70 levels in tumor samples increase 93-fold in KIRC and 14-fold in KIRP, CD70 levels *decrease* in KICH by 3-fold ($logFC = -1.56$; B-H corrected p-value = 0.004). **(B and C)** Line charts showing average expression of PCK1 (blue) and C10orf99 (green) over different stages of KIRC (B) and KIRP (C). The vertical lines are error bars. THPA designates PCK1 as prognostic favourable and C10orf99 as prognostic unfavourable for renal cancer.

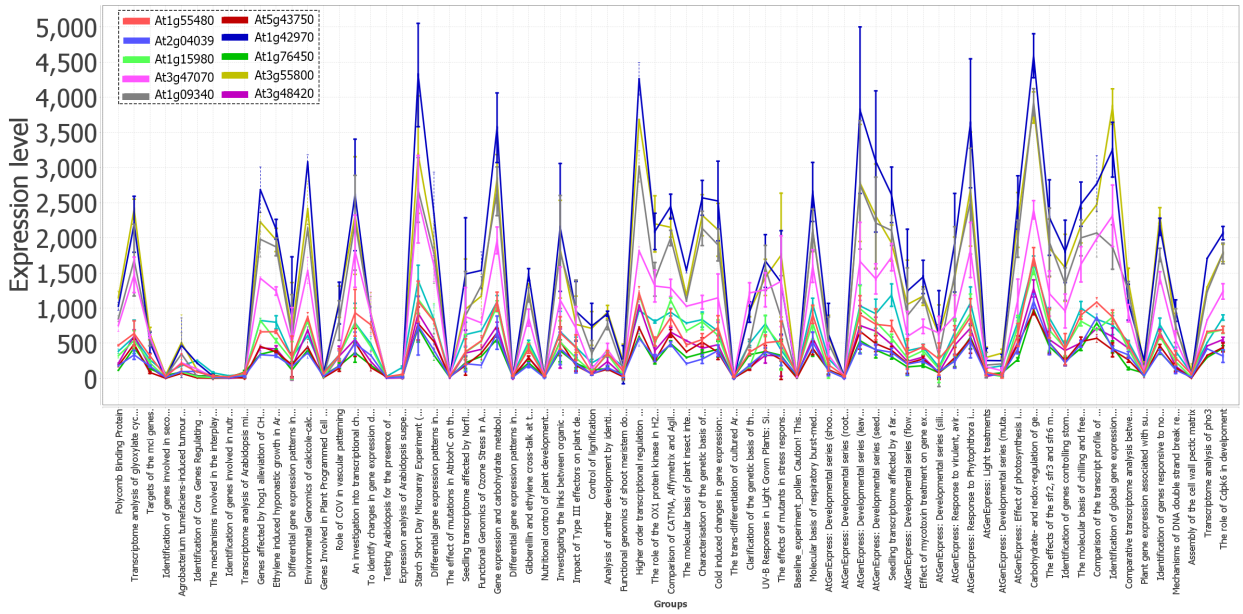

Figure 3.5 Spearman correlation followed by line-plot visualization, using MOG, shows that Met1 (At1G55480) is highly expressed in photosynthetic organs and highly correlated with several genes of unknown function. The "peaks" of expression are all leaf samples; the "troughs" of expression are predominantly root and cell culture samples. *AT-microarray-dataset* representing 71 diverse studies and a wide variety of environmental, genetic and developmental conditions (Mentzen and Wurtele, 2008). Several genes of unknown function are closely coexpressed in this cluster.

Table 3.4 MOG compared to existing tools for exploratory analysis of expression data. MOG's GUI, designed with Java swing, is fully interactive; in contrast, other available tools are based on R and provide limited or no interactivity. A MOG user can execute any R package/script with interactively-selected subsets of data if s/he wishes to perform additional analysis, whereas only a limited number of R-packages are available in the other tools. The last row compares the Man-Whitney U test's execution time for MOG and PIVOT using the liver tumor and non-tumor datasets (18,212 genes over 410 samples). A more detailed comparison of the tools is available in Supplementary File 8.

	MOG	**PIVOT**	**iSEE**	**iGEAK**	**IRIS-EDA**	**DEvis**
Reference	This paper	(Zhu et al., 2018)	(Rue-Albrecht et al., 2018)	(Choi and Ratner, 2019)	(Monier et al., 2019)	(Price et al., 2019)
Year	2019	2018	2018	2019	2019	2019
Platform/GUI	Java/Swing	R/Shiny	R/Shiny	R/Shiny	R/Shiny	R/None
Interactive tables and trees	Yes	No	No	No	No	No
Interactive drag and drop operations	Yes	No	No	No	No	No
Interactive visualizations	Yes	Partial	Partial	Partial	Partial	No
Interactively subset data	Yes	Partial	Partial	Partial	No	No
Save progress	Yes	Yes	Partial (if user saves R code)	No	No	No
Use any R package	Yes	No	No	No	No	No
Supported data types	Omics or other numerical data	RNA-Seq/scRNA-Seq	Omics	RNA-Seq/microarray	RNA-Seq/scRNA-Seq	RNA-Seq
M-W U test (sec.)	7	1,260	NA	NA	NA	NA

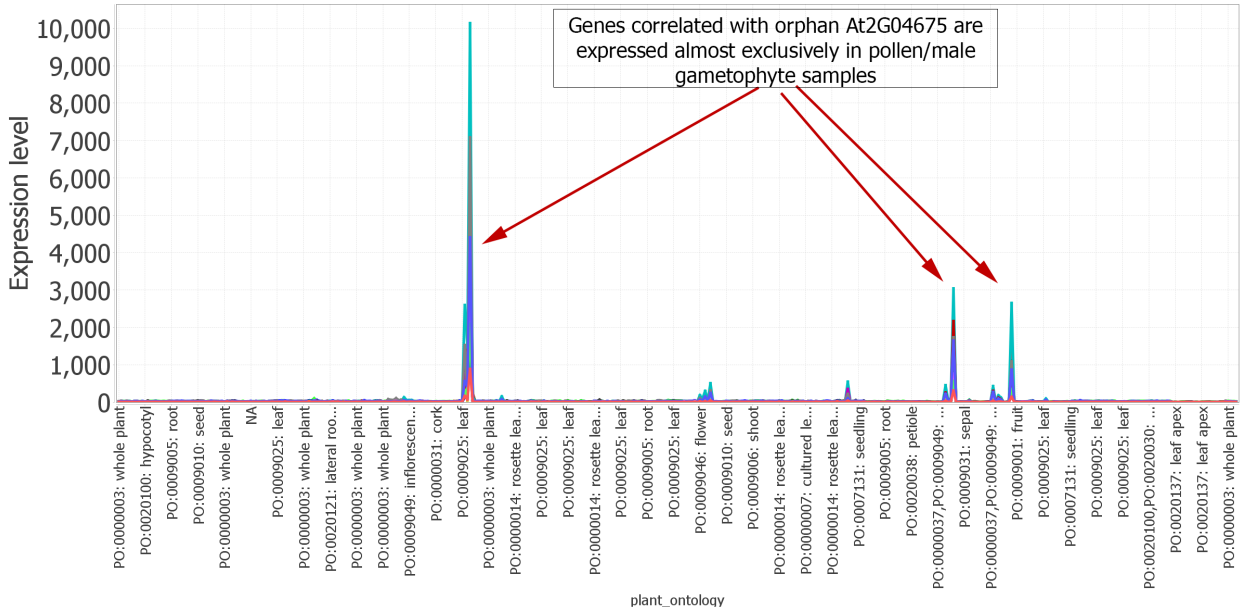

Figure 3.6 MOG line chart visualization shows the expression of orphan gene At2G04675 over the *AT-microarray-dataset* representing 71 diverse studies and a wide variety of environmental, genetic and developmental conditions (Mentzen and Wurtele, 2008). X-axis are samples, and Y-axis indicates their expression value. The orphan gene At2G04675 is of no known function, and genes highly correlated with At2G04675 are expressed almost exclusively in pollen/male gametophyte samples. Each line represents a gene. (Lines in this visualization are for clarity and the connections from sample to sample do not imply a relationship).

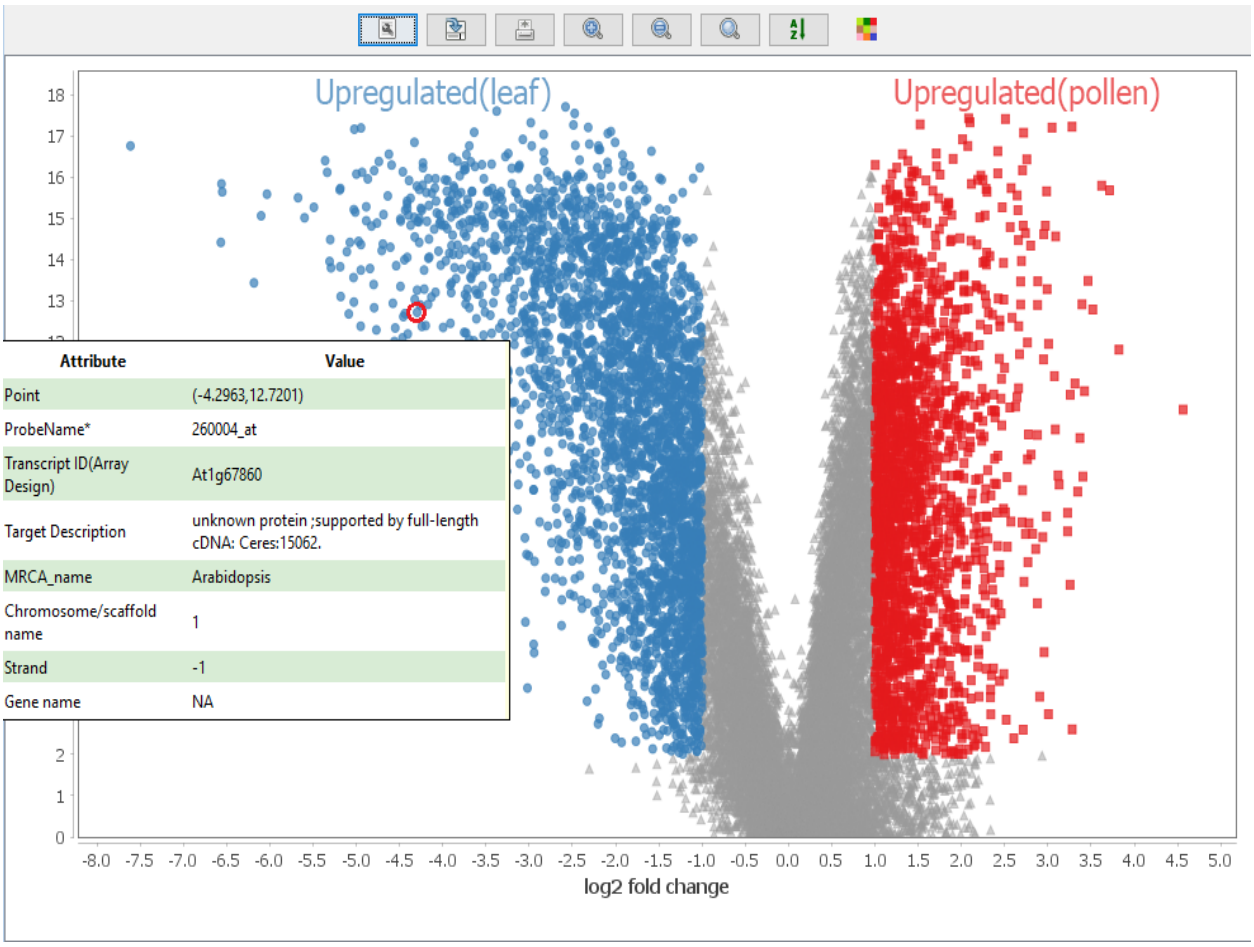

Figure 3.7 Using MOG for differential expression analysis of leaf and pollen samples, followed by volcano plot visualization (Y axis: -$\log_{10}$(p-value)). At1G67860, an Arabidopsis specific gene with no known function, is 16-fold more highly accumulated in leaves relative to pollen (Mann-Whitney U test; B-H corrected p-value$< 10^{-3}$). The gene metadata is revealed upon hovering the mouse over a data point.

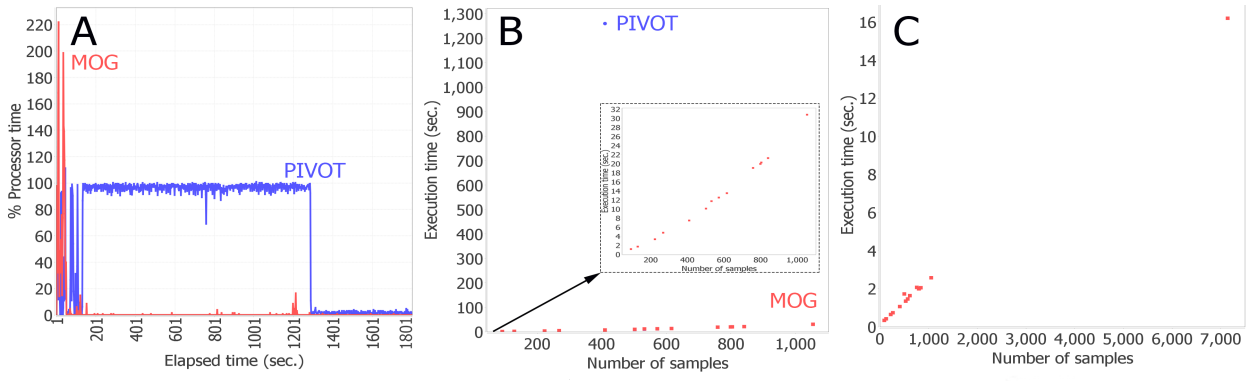

Figure 3.8 MOG performance benchmarks. MOG was benchmarked using the entire *Hu–cancer-RNASeq-dataset* (18,212 genes over 7142 samples), and using chunks of this dataset.**(A)** Comparison of MOG to R-based (PIVOT). Dataset size was limited to the amount of data that could be loaded in PIVOT (410 samples). % processor time (% CPU utilization) was calculated over 30 minutes; theoretical maximum value = total processors in computer x 100 (400 in this case). **(B)** Execution times for computing differentially-expressed genes using Mann-Whitney U test. Red dots, MOG; blue dot, PIVOT (410 samples). Inset, expanded scale to display MOG execution times. **(C)** MOG execution times for pairwise computations of Pearson correlation of a gene (BIRC5) with all other genes in the datasets. (Other tools cannot perform this computation). Execution times are linear with data size; full dataset analysis took 16 seconds.

CHAPTER 4. AFRICAN AMERICANS AND EUROPEAN AMERICANS EXHIBIT DISTINCT GENE EXPRESSION PATTERNS ACROSS TISSUES AND TUMORS ASSOCIATED WITH IMMUNOLOGIC FUNCTIONS AND ENVIRONMENTAL EXPOSURES

Urminder Singh [1,2,3], Kyle M. Hernandez [4,5], Bruce J. Aronow [6], and Eve Syrkin Wurtele [1,2,3]

[1]Bioinformatics and Computational Biology Program, Iowa State University, Ames, IA 50011.

[2]Center for Metabolic Biology, Iowa State University, Ames, IA 50011.

[3]Department of Genetics Development and Cell Biology, Iowa State University, Ames, IA 50011.

[4]Department of Medicine, University of Chicago, Chicago, IL, 60637

[5]Center for Translational Data Science, University of Chicago, Chicago, IL, 60637

[6]Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229

Modified from a manuscript published in *Scientific Reports*

4.1 Abstract

The COVID-19 pandemic has affected African American populations disproportionately with respect to prevalence, and mortality. Expression profiles represent snapshots of combined genetic, socio-environmental (including socioeconomic and environmental factors), and physiological effects on the molecular phenotype. As such, they have potential to improve biological understanding of differences among populations, and provide therapeutic biomarkers and environmental mitigation strategies. Here, we undertook a large-scale assessment of patterns of gene expression between African Americans and European Americans, mining RNA-Seq data from 25 non-diseased and diseased (tumor) tissue-types. We observed the widespread enrichment of pathways implicated in COVID-19 and integral to inflammation and reactive oxygen stress. Chemokine CCL3L3 expression is up-regulated in African Americans. GSTM1, encoding a glutathione S-transferase

that metabolizes reactive oxygen species and xenobiotics, is upregulated. The little-studied F8A2 gene is up to 40-fold more highly expressed in African Americans; F8A2 encodes HAP40 protein, which mediates endosome movement, potentially altering the cellular response to SARS-CoV-2. African American expression signatures, superimposed on single cell-RNA reference data, reveal increased activity of esophageal glandular cells and lung ACE2-positive basal keratinocytes. Our findings establish *basal prognostic signatures* that can be used to refine approaches to minimize risk of severe infection and improve precision treatment of COVID-19 for African Americans. To enable dissection of *causes* of divergent molecular phenotypes, we advocate routine inclusion of metadata on genomic and socio-environmental factors for human RNA-sequencing studies.

4.2 Introduction

The COVID-19 pandemic has infected over 31 million people and killed over 970,000 worldwide as of September, 2020 (https://coronavirus.jhu.edu/map.html). Its causative agent, the novel SARS-CoV-2, is an enveloped single stranded RNA virus that infects tissues including epithelial cells in the upper respiratory tract, lung alveoli, GI tract, vasculature endothelium, renal tubules, central nervous system, and myocardium (Jeyanathan et al., 2020; Bansal, 2020; Varga et al., 2020; Snell, 2020; Frithiof et al., 2020; Jarrahi et al., 2020). The complex combinations and severities of symptoms caused by SARS-CoV-2 include fever, cough, fatigue, dyspnea, diarrhea, thrombosis, stroke, acute respiratory failure, renal failure, cardiac failure; in some individuals these may lead to long-term disability or death (Frithiof et al., 2020; Bansal, 2020; Jarrahi et al., 2020). Differing patterns of disease may result from direct cellular infection, secondary inflammatory repercussions, and circulating immune and necrotic complexes from distal sites of infection and response (Tay et al., 2020; Teuwen et al., 2020; Del Valle et al., 2020; Chua et al., 2020). Individuals who suffer the most severe sets of symptoms are more likely to be over 65 years of age, and/or have obesity or preexisting comorbidities such as diabetes, hypertension and heart disease (Zhou et al., 2020). How these attributes confer risk of increased disease severity to individuals is not well understood (Teuwen et al., 2020; Millett et al., 2020;

Battagello et al., 2020; Snell, 2020; Chua et al., 2020). Identifying individuals most at-risk for severe COVID-19 infection, and determining the molecular and physiological basis for this risk, is critical to enable more informed public health decisions, and improving our identification and use of precision interventions.

COVID-19 cases and deaths are disproportionately higher among African Americans in the US relative to European Americans (Millett et al., 2020). This disparity is caused in part by complex combinations of socio-economic factors, including underlying comorbidities, air quality, population density, and health care access (Millett et al., 2020); heritable factors in the human host also influence COVID-19 symptoms (Williams et al., 2020; Ellinghaus et al., 2020; Warren and Birol, 2020; Devaux et al., 2020; Woo et al., 2020). To date, several genetic determinants of COVID-19 severity have been partially elucidated. Genetic variants of Angiotensin-Converting Enzyme2 (ACE2), a major human host receptor for the SARS-CoV-2 spike protein, may be linked to increased infection by COVID-19 (Devaux et al., 2020). Human Leukocyte Antigen (HLA) gene alleles have been associated with susceptibility to diabetes and SARS-CoV-2 (Warren and Birol, 2020). A COVID-19 association at locus 9q34.2 spans several genes related to COVID-19, including blood type (Ellinghaus et al., 2020). The genetic propensity in southern European populations for mutations in the pyrin-encoding Mediterranean Fever gene (MEFV) has been proposed to be associated with elevated levels of pro-inflammatory molecules, a cytokine storm, and greater severity of COVID-19 (Woo et al., 2020). Multiple GWAS associations based on ancestry are beginning to emerge (https://grasp.nhlbi.nih.gov/Covid19GWASResults.aspx) (Ellinghaus et al., 2020).

Gene expression is a reflection of a cell's composition and its spatial and developmental context in an organism. Modifying factors that determine gene expression span genetics, and physiological, environmental, and socio-environmental influences. In this study we seek to investigate potential differential expression of genes and pathways that may impact the severity of COVID-19 infection in African Americans. Research with macrophage cell lines has identified ancestry-related differences in innate immune response to bacterial pathogens, with cell lines

isolated from individuals with African ancestry more likely to exhibit stronger inflammatory responses (Nédélec et al., 2016). However, studies on the impact of Covid-19 mostly lack in sufficient numbers of individuals of different populations to achieve a high resolution analysis of differential expression responses.

Here, we utilize diverse, publicly-available datasets from 25 tissue-types to explore gene expression differences between African American and European American individuals. Specifically, we analyze -Seq data of "non-diseased" tissues from the Genotype Tissue Expression (GTEx, https://gtexportal.org/home/) project. And, as representative of highly perturbed systems, we analyze tumor samples from The Cancer Genome Atlas (TCGA, https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga); tumor tissue-types are particularly important because cancer confers an increased risk for severe outcomes from COVID-19 (Robilotti et al., 2020). Further, we seek to unravel the cellular origins of ancestry-associated differential gene expression through the use of Human Cell Atlas single-cell datasets from esophagus and lung tissues.

Taken together, our analyses reveal consistent differences between European Americans and African Americans in pathways, genes, and cell types likely to impact the severity of COVID-19. In esophagus and lung, two tissues critical to early SARS-CoV-2 infections, differential gene signatures between African-American and European American populations implicate specific cell lineages that are likely to alter viral disease severity. The results provide a critical baseline in the context of cellular and organismal health and resilience to disease from which to assess COVID-19 gene expression studies from a population perspective. Finally, we highlight the importance of evaluating population-related impacts on gene expression in the combined light of socio-environmental and genetic factors.

4.3 Results

In order to identify genes differentially-expressed (DE) between African American and European Americans, we constructed an aggregated dataset of 7,142 RNA-Seq samples

encompassing non-diseased tissues from GTEx and tumors from TCGA (Singh et al., 2020; Wang et al., 2018). The batch-corrected and processed data (Wang et al., 2018) enable comparison across samples, and the large sample-size increases statistical power of the analysis. Race assignments are self-reported in the metadata; however, many of the individuals identifying as a single race may be from an admixed population (Baharian et al., 2016; Zhong et al., 2020). We analyzed data and metadata using MetaOmGraph (MOG) (Singh et al., 2020), software that supports interactive exploratory analysis of large data to identify and distinguish patterns across multiple dimensions (Table 1 and Supplementary Table S1).

4.3.1 Multiple genes are DE between populations in a tissue- and tumor-specific manner

DE genes were identified for each tissue-type, as well as for pooled TCGA and GTEx data (Supplementary Tables S2-S28). To test for potential confounding factors that might explain gene expression pattern differences, we scrutinized differences between African American and European Americans populations controlling for biologically-relevant factors (sex, age, tissue-type, Body mass index (BMI) (as available in metadata), and cancer sub-type (as available in metadata)); under these analysis, DE genes from each Mann-Whitney (MW) analysis retained statistical significance in the corresponding limma model (Supplementary Tables S29-S55; Additional File 1). We used Hartigans' dip test to each gene to evaluate bi- or multi-modality in gene expression distributions (Additional File 2). For a given gene and tissue-type, a bimodal structure could imply presence of underlying hidden variables that affect expression of that gene, such as unreported sub-population structure or environmental factors.

These analyses indicate there are numerous genes DE more than 2-fold between African American and European American populations (Table 1 and Supplementary Tables S2-S2). The analyses cannot distinguish as to whether these differences in expression are associated with socio-environmental factors or genetic factors, because this information is not included in the available metadata. Only tissue-types with over 15 African American individuals sampled showed

DE genes >2-fold difference in expression based on Mann-Whitney U test (BH-corrected p-value < 0.05).

Table 4.1 **Number of DE genes in African Americans (AA) compared to European Americans (EA) in nine non-diseased tissue types and eight tumor types.** Only tissue-types with AA sample size 12 or greater are shown. Samples are sorted first by project, and then by the number of upregulated genes. The number of samples affects the power of the DE test. Criteria for DE:>2-fold difference in expression based on Mann-–Whitney U test (BH-corrected p-value < 0.05).

Project	Tissue-type	#AA samples	#EA samples	#Upreg.	#Downreg.
GTEx	Breast	12	75	0	0
GTEx	Prostate	13	89	0	0
GTEx	Uterus	13	68	0	0
GTEx	Liver	15	97	0	0
GTEx	Stomach	29	159	4	6
GTEx	Colon	41	292	13	9
GTEx	Esophagus	80	564	19	11
GTEx	Thyroid	43	267	25	30
GTEx	Lung	39	269	45	20
TCGA	Lung squamous cell carcinoma (LUSC)	28	337	2	0
TCGA	Thyroid carcinoma (THCA)	25	292	3	3
TCGA	Lung adenocarcinoma (LUAD)	48	368	16	5
TCGA	Kidney renal papillary cell carcinoma (KIRP)	49	166	19	13
TCGA	Uterine Corpus Endometrial Carcinoma (UCEC)	54	70	28	5
TCGA	Colon adenocarcinoma (COAD)	54	188	30	21
TCGA	Kidney renal clear cell carcinoma (KIRC)	46	410	68	94
TCGA	Breast invasive carcinoma (BRCA)	142	674	83	164
GTEx	pooled GTEx samples	292	1,905	12	11
TCGA	pooled TCGA samples	497	3,238	13	21

4.3.2 Expression differences between populations are enriched for the broad network of infection, inflammation, endosomal development, and ROS metabolism

GO terms related to the interrelated biological processes of inflammation/cytokines, endosomal development, and ROS metabolism are overrepresented among those genes that are DE between African Americans and European Americans (Supplementary Table S56).

Similarly, Gene Set Enrichment Analysis (GSEA) of all of the 25 GTEx and TCGA tissue-types shows KEGG pathways (Kanehisa et al., 2017) of immune- and inflammation-related

processes are highly enriched (Supplementary Table S57-S58); the single most commonly-enriched pathway (found in 19 of the 25 tissue-types) is "cytokine-cytokine receptor interaction"; glutathione-oxidative processes of ROS and xenobiotic metabolism are enriched in nine tissue-types (Figure 4.1A and Additional File 3). For example, analysis of pooled GTEx data detects coordinated changes between African Americans and European Americans associated with four cytokine-related pathways and oxidative drug metabolism (Figure 4.1B and Supplementary Table S57).

Multiple genes are DE between African American and European American populations (Supplementary Table S2-S25). However, seven genes are highly and consistently DE between African Americans vs European Americans across all or most tissue-types. These are: C-C Motif Chemokine Ligand, CCL3L3; mitochondrial Glutathione-S-Transferase, GSTM1; Nuclear Pore Complex Interacting Protein Family Member, NPIPB15; Coagulation Factor VIII Associated genes, F8A3 and F8A2; FAM21B; and serine protease, PRSS21. Of these, four, C-C Motif Chemokine Ligand, CCL3L3; mitochondrial Glutathione-S-Transferase, GSTM1, F8A3 and F8A2, are directly related to the interrelated processes of infection, inflammation, endosomal motility, and ROS metabolism.

4.3.2.1 Cytokines, ROS and the storm

Among the DE cytokines, the small inducible chemokine, CCL3L3, is more highly expressed in African Americans by up to 7-fold in most diseased and non-diseased tissue-types (Figure 4.2) (Supplementary Table S2-S25). Genes involved in common biological processes that are DE < 1.3-fold change in one or more tissue-types include: CCL4L1, CCL4L2, CCL3L1, CXCL9, CXCL13, CXCL17, CXCL10, GRK1, VAV3, CCL21, CCL8, and CCL15 (Supplementary Table S2-S25).

Several genes that mitigate oxidative stress, an inducer of cytokines, are DE between African American and European American populations. In particular, GSTM1, a key enzyme of oxidative stress, is more highly expressed in African Americans than European Americans across multiple

tissue-types, including over 9-fold higher expression in lung (Figure 4.2)). Functionally-related genes that are DE $<$ 1.3-fold change in expression based on Mann-–Whitney U test in one or more tissue-types include: GSTM3, GTTT1, GSTT2, GSTT2B, GSTM4, FMO2, GSTM5, and CYP2A46; the CCL3L3 chemokine receptor proteins CCR1, CCR3, and CCR5 are not significantly DE (Supplementary Table S2-S25).

4.3.2.2 F8As and endosome motility

Endosomal function and autophagy are implicated in COVID-19 and intimately intertwined with cytokine and ROS signaling (Ayres, 2020; Carmona-Gutierrez et al., 2020; Yang and Shen, 2020). One little-studied player implicated in early endosome motility (Pal et al., 2006) is the F8A/HAP40 (HAP40) protein, encoded by three genes (F8A1, F8A2, and F8A3) in humans (Perez-Riba and Itzhaki, 2019). The three F8A proteins are identical in sequence, and thus likely have the same molecular function.

F8A1 is more highly expressed by about 2-fold in European Americans in almost every tissue-type analyzed (Figure 4.3). Conversely, F8A2 and F8A3 are more highly expressed in African Americans. Expression of F8A2 in African Americans is up to 40-fold greater; expression of F8A3 is up to 6.6-fold greater. In LUSC, F8A2 and F8A3 are the only genes DE $>$ 2-fold (Supplementary Table S7). F8A2 and F8A3 follow a similar trend, being more highly expressed in African Americans (**Figure** 4.3, Supplementary Figure 1, and Supplementary Table S2-S25).

Distribution of F8A2 and F8A3 expression is bimodal in European Americans for most cancers, and part of the difference in levels of F8A2 and F8A3 expression between the two populations is due to their extremely low/undetectable levels of expression in a large proportion of the European American population.

Because of the vast differences in expression levels of the three HAP40-encoding genes between African Americans and European Americans, the paucity of literature on HAP40 (Furr-Stimming et al., 2020), and the unclear relationships among F8A1, F8A2, and F8A3 genes, we further investigated the sequences, sequence variants, and the expression patterns of these genes.

The sequences of the HAP40-encoding proteins from F8A1, F8A2, and F8A3 are identical to each other in human reference genome GRCh38.p13 (https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.39). We searched for potential allele variants of HAP40 proteins encoded by F8A1, F8A2, and F8A3 in The Genome Aggregation Database (gnomAD) (Karczewski et al., 2020). gnomAD assigns individuals to populations, by clustering of genetic features. Our search identified only very rare sequence variants in the HAP40s encoded by F8A1, F8A2, or F8A3 (gnomAD v3). No structural variants were identified for HAP40 of F8A1 or F8A3; a duplication of 54 aa is, very rarely, present in F8A2 (gnomAD SVs v2.1).

To our knowledge, F8A1, F8A2 and F8A3 gene expression has never been compared. This may be because expression of F8A2 and F8A3 genes is relatively low in most European Americans, and European Americans are the predominant population studied. Furthermore, most RNA-Seq studies report expression of *only* F8A1 or F8A3 (and not F8A2), presumably aligning all reads to one or the other gene.

We analyzed coexpression of the three F8A genes relative to the other 18,212 genes represented in the full TCGA-GTEx dataset using two statistical measures: Pearson's correlation and Mutual Information (MI) (Daub et al., 2004). Although the three F8A genes are proximately located on the X chromosome, their expression patterns are *not* correlated. F8A2 and F8A3 have a Pearson's correlation of ($r = 0.40$), while both are negatively correlated with F8A1. Indeed, of all 18,212 genes represented in the data, the expression pattern of F8A1 is most *negatively* (anti-) correlated with that of F8A2 ($r = -0.45$) and F8A3, ($r = -0.24$) (Supplementary Table S60). MI analysis indicates that F8A2 and F8A3 genes are more closely associated with F8A1 than with any other gene, consistent with the negative Pearson correlation (Supplementary Table S60). Also of note, F8A1 expression is not correlated with the F8 (Coagulation Factor FVIII) gene, although it resides with intron 22 of this gene.

4.3.3 Signatures of DE genes correspond to specific cell types in esophagus and lung

We sought to determine whether genes differentially expressed between African Americans and European Americans corresponded to distinct cell populations present in the whole-tissue GTEx samples. This would provide information on cell-type representation across the two populations. To do this, we evaluated single cell datasets from two tissues highly relevant for SARS-CoV-2 infection: esophagus and lung (Jimenez et al., 2020; Travaglini et al., 2020).

Genes upregulated in African Americans in the esophagus map predominantly to two cell lineages, glandular epithelial cells of esophagus glands, and hematolymphoid lineage-associated dendritic cells (Figure 4.4). In proximal and distal airway cells of the lung, the signature of DE genes in African Americans versus European Americans corresponds to basal differentiating and proliferating keratinocytes (Figure 4.5).

Discussion

Human genetics contribute to the propensity and severity of diseases (Zhong et al., 2020; Amorim et al., 2017; Kruzel-Davila et al., 2017; Paulucci et al., 2017; Wu et al., 2019; Barrow et al., 2019; Chi et al., 2019; O'Brien et al., 1971; Burt et al., 1995). Sometimes the contribution is straightforward; a single allele variation found in Ashkenazi Jews, causes the vast majority of Tay-Sachs disease (O'Brien et al., 1971). Sometimes it is more complex; for example, hypertension is more prevalent in African American than European American populations (Burt et al., 1995) in part due to detrimental APOL1 mutations that are more frequent in West African populations (Kruzel-Davila et al., 2017). Despite the paucity of studies focused on Western African populations, the propensity and severity of other diseases among this population have been attributed to genetics (Need and Goldstein, 2009; Kruzel-Davila et al., 2017; Zhong et al., 2020; Backer, 2020).

In this study, we describe the molecular phenotypes, as revealed by differences in gene expression, in African Americans and European Americans across multiple non-diseased and

diseased tissues. These distinct molecular phenotypes are likely caused by complex combinations of socio-environmental and genetic factors.

The predominant differences in gene expression, pathway enrichment, and cell-types between African Americans and European Americans are implicated in biological processes that highly impact COVID morbidity and mortality. These genes and pathways are not specific to COVID-19, but also would impact other diseases. Many COVID-19 deaths have been attributed to a cyclic over-excitement of the innate immune system (Tay et al., 2020; Woo et al., 2020; Jeyanathan et al., 2020). This process, often termed a cytokine storm, results in a massive production of cytokines, and the body attacking itself rather than specifically destroying the pathogen-containing cells (Tay et al., 2020; Jeyanathan et al., 2020). People with comorbidities, the elderly, and immunosuppressed individuals, may be at a greater risk for COVID-19 morbidity and mortality either because they may not respond to infection with a sufficient immune response (Ahmadpoor and Rostaing, 2020) and/or because they may be more likely to develop a cytokine storm (Tay et al., 2020; Jeyanathan et al., 2020). Many cytokines and other immunomodulatory molecules are DE, and cytokine-related KEGG pathways are enriched, between African Americans and European Americans in one or more tissue-type.

The chemokine CCL3L3, upregulated in African Americans relative to European Americans under almost every diseased and non-diseased tissue-type we tested, and notably by 3.8-fold in lungs, is also upregulated in COVID-19-diseased human bronchoalveolar lavage fluid (Didangelos, 2020). CCL3L3 encodes the CCL3 protein (also called MIP-1), a member of the functionally-diverse C-C motif chemokine family. A neutrophil chemotaxis protein, CLL3 acts as ligand for CCR1, CCR3, and CCR5, recruiting and activating neutrophils (Struyf et al., 2001; Chua et al., 2020). Neutrophils themselves are highly implicated in the severity of COVID-19 (Veras et al., 2020; Didangelos, 2020; Zhai et al.). CCL3 expression is upregulated in severe COVID-19 (Zhai et al.; Chua et al., 2020). Increase in accumulation of the CCL3 protein has been strongly associated with severe (but not mild) COVID-19 disease (Chevrier et al., 2020).

GSTM1, more highly expressed in African Americans compared to European Americans in almost every tissue-type evaluated, is a key enzyme of mitochondrial ROS metabolism (Sies and Jones, 2020). Mitochondrially-generated ROS induce expression of proinflammatory cytokines and chemokines, and are considered to play a key role in modulating innate immune responses against RNA viruses (Sies and Jones, 2020) including SARS-COV-2 (Schönrich et al., 2020). GSTM1 itself is induced by nuclear factor erythroid 2-related factor 2 (Nrf2), a transcription factor that integrates cellular stress signals (Matsuyama et al., 2020). Increased expression of GSTM1, could lead to increased mitochondrial ROS, which might ultimately trigger inflammation and a cytokine storm (Sies and Jones, 2020). Alternatively, increased GSTM1 expression might cause ROS to be metabolized rapidly, and prevent ROS from initiating a sufficient immune response. GSTM1 has a second critical function– in metabolism of xenobiotics, including many toxins and pharmaceuticals (Sies and Jones, 2020). In the latter case, pharmaceuticals may be more rapidly metabolized and rendered inactive.

The most dramatic differences in gene expression in African Americans compared to European Americans are associated with the highly-conserved but little-studied F8A genes, which each encode the HAP40 protein. F8A1 is upregulated about 2-fold in European Americans. In contrast, F8A2 and F8A3 are even more highly upregulated in African Americans, and in over half of the samples from European American individuals, levels of expression of F8A2 and F8A3 were negligible.

Although coagulation factor VIII has a high frequency of mutations across populations (Graw et al., 2005), we found the F8A1, F8A2 and F8A3 genes and CDSs to be highly conserved across populations (gnomAD v3). This conservation is consistent with the three genes having a similar and specific molecular function. However, despite their proximity and encoding the identical protein, F8A1, F8A2 and F8A3 each have highly distinct patterns of expression across the thousands of samples of tissues and cancers in the TCGA/GTEx dataset, indicating they may participate in different or overlapping biological scenarios.

HAP40 function has been researched mostly in the context of F8A1 and the critical role of that gene in slowing early endosome mobility in Huntington's disease (Perez-Riba and Itzhaki, 2019). In Huntington's, HAP40 forms a bridge between the huntingtin protein and the regulatory small guanosine triphosphatase, RAB5; formation of this complex reduces endosomal motility by shifting endosomal trafficking from the microtubule to the actin cytoskeleton (Pal et al., 2006). F8A1 overexpression in striatal neuron cell lines from mice resulted in increased ROS and mitochondrial dysfunction (Huang et al., 2017). Knockouts of F8A1 in human HeLa and HEK293 cells yield altered/reduced autophagy and shorter life spans (Huang et al., 2017). Knockouts of the single F8A gene in Drosophila similarly show reduced activity, altered/reduced autophagy, and shorter lifespan (Xu et al., 2020).

F8A1 expression is increased under several conditions, including Huntington's disease (Peters and Ross, 2001), presence of a SNP variant for type 1 diabetes risk (Brumpton and Ferreira, 2016), cytotrophoblast-enriched placental tissues in women with severe preeclampsia (Gormley et al., 2017), and mesenchymal bone marrow cells as women age (Roforth et al., 2015). Its potential roles in the latter conditions has not been investigated.

Altered endosome motility would play an important but complex role in infection and the innate immune response, and might either promote or hinder the battle between SARS-CoV-2 and its human host (Tao and Drexler, 2020; Carmona-Gutierrez et al., 2020). Coronaviruses including SARS-CoV-2 mainly enter host cells via binding to the ACE2 receptor followed by endocytosis (Chowdhury and Maranas, 2020; Tay et al., 2020). Nascent early endosomes are moved along the microtubule cytoskeleton, fusing with other vesicles; varied molecules can be incorporated into the endosomal membrane or its interior (Tao and Drexler, 2020; Carmona-Gutierrez et al., 2020). This regulated development enables diverse fates. For example, in the context of SARS-CoV-2, endosomes might release viral RNA or particles; they might merge with lysosomes and digest their viral cargo; or they might fuse with autophagosomes (autophagy) and subsequently with lysosomes that digest the cargo (Tao and Drexler, 2020; Carmona-Gutierrez et al., 2020). SARS-CoV-2 might reprogram cellular metabolism to suppress autophagy and promote viral

replication (Gassen et al., 2020); conversely, the cell might modify autophagy machinery to decorate viral invaders with ubiquitin for eventual destruction, activate the immune system by displaying parts of the virus, or catabolize excess pro-cytokines. Autophagy might induce cytokine signaling, which could promote protective immune response or engender a destructive storm of cytokines, inflammation and tissue damage (Carmona-Gutierrez et al., 2020). Because of its function in early endosome motility, HAP40 has implications as a potential molecular target in therapy of endosomal and autophagy-related disorders such as COVID-19.

Our results regarding differentially-expressed genes and biological processes are consistent with those of a study using cultured primary macrophages that had been isolated from individuals of African and European ancestry. This study identified thousands of genes with ancestry-associated differences in expression in response to bacterial infection, and additional evidence of underlying genetic control and population-specific signatures of adaptation (Nédélec et al., 2016). Despite the disparity between the biological systems analyzed, the differentially expressed genes were similar (See Supplementary File 5).

Our study using single cell reference data indicate several cell type-specific associations of the signatures of DE genes in African Americans versus European Americans in esophagus and lung. This interrogation reveals enrichment of DE genes in immune-related cell-types. One model by which this might occur is that individuals of one population tend to have different proportions of a given cell type or histological structure. An alternative model is that individuals of one population might tend to maintain some of their cell types in a state of relatively higher activation. Either explanation would lead bulk RNA-Seq analyses, such as tissue-types from GTEX or TCGA, to demonstrate elevated expression of those transcripts in that population.

Although at a population level, major differences exist in expression of inflammation-related genes and cell-type-specific associations between African Americans and European Americans, when considered on the basis of each individual within each population, gene expression differences are more complex. Individuals within a population may exhibit all, no, or some portion of the prevailing differences in a population. That some genes show bimodal expression

distribution in some tissue-types African American and/or European American populations further emphasizes this variation.

Thus, the significance of these patterns and their relationship to differential susceptibility or risk of severity from COVID-19 (or another disease) must be considered from nuanced perspectives. Importantly, it may be that only a fraction of the signature and a fraction of the individuals in a population are at elevated risk of more severe disease. In addition, different mechanisms of risk may be operative within different individuals within an population. For example, elevated abundance or activity of cells that are the target of COVID-19 (e.g., ACE2-positive basal keratinocytes) could lead to a greater infection burst during initial phases with a larger number of virions being released systemically. If, as it appears from the alignment of the DE genes in African Americans compared to European Americans to the lung single cell data, this is the case for African American-individuals, then they might be more readily taken over by infecting SARS-Cov2 virions.

The differential expression of genes implicated in COVID-19 morbidity and mortality between African Americans and European Americans reported herein emphasizes the importance of integrating gene expression data into the genetic and socio-environmental factors at a population level. Further, RNA-Seq data has been shown useful in clinical practice for pediatric cancers (Vaske et al., 2019), and this practice could be extended to other diseases. Our analysis, in concurrence with those of (Nédélec et al., 2016; Quintana-Murci, 2019), supports the concept that processes of disease and stress are enriched in comparisons of African American and European American populations, and this may be in part because ancestral selection pressures such as pathogens, temperature stress, and toxins, were very strong, and there were very different complements of these stresses in the regions where these two populations lived. To survive, humans living in Europe and those living in Western Africa would have had to evolve the ability to resist the diverse prevalent local pathogens and stresses. Other differences would be due to a difference in socio-environmental factors, such as stress, commorbidity, or exposure to pollution (Cole, 2014).

Expression data provided a tremendous wealth of information from which researchers can model the factors that predict and determine disease. However, the utility of these data is reliant on adequate *representation of cohorts* and on sufficient *metadata* describing the individuals sampled. For example, ethnic bias, practical factors (such as subject availability), as well as a paucity of molecular medical research in many regions of the world often result in insufficient numbers of subjects from many populations being represented in medical studies (Friedman et al., 2019; McGuire et al., 2020). This lack of representation greatly impedes the development of precision prognosis and therapy based on genetics (Barrow et al., 2019; Friedman et al., 2019). For example, here, we were limited to comparison of differences between gene expression in African American and European American populations because even in the large GTEx and TCGA studies, sample sizes for the other three major population groups (Asian, Native American, and Pacific Islanders) were generally too low for robust statistical assessment (Supplementary Table S1).

In addition, even if sample sizes for race are sufficient, information on the ancestry of each individual sampled is needed. Self-reported metadata on race is often not publicly available for individual samples. However, methods of assigning ancestry to individuals sampled for RNA-Seq are being developed and applied (Yuan et al., 2018; Barral-Arca et al., 2019).

Finally, current pipelines for RNA-Seq analysis often represent only the more highly or consistently expressed annotated genes (Morillon and Gautheret, 2019; Martinez et al., 2020). Population-specific genes may be missed in the analysis unless they are in the predominant population being studied. The same is true for members of genes families that are preferentially-expressed in particular populations. An example brought out by our study is the F8A2 gene, which is DE-up in African Americans compared to European Americans; however, F8A2 is not even represented in the processed data of many RNA-Seq studies.

Combined information on socio-environmental factors and genomics of individuals sampled is critical in dissecting the determinants of gene expression in that individual. Yet for humans, a dichotomy exists between socio-environmental and genomic investigations. Among the vast body

of human RNA-Seq data deposited, not only are metadata on the ancestry of the sampled individuals often unavailable, but socio-environmental metadata are almost never present. Thus, apart from the pioneering sociogenomics research of (Cole, 2014; Dieckmann et al., 2020) and studies such as (Favé et al., 2018; Quintana-Murci, 2019; Hooten and Evans, 2019), socio-environmental information are rarely considered in 'omics analyses. Indeed, because of the scant metadata on socio-environmental determinants it is not even possible to determine possible skewness of representation of socio-environmental groups among the individuals sampled; thus, socio-environmental factors represent high-impact complex hidden covariates that would be challenging to model.

Conversely, sociological studies rarely incorporate 'omics information. For example, the U.S.-based Robert Wood Johnson Foundation (https://www.rwjf.org/en/library/interactives/whereyouliveaffectshowlongyoulive.html) cites research that "your zip code can be more important than your genetic code" for your health; however, the analyses were done without actually evaluating genetic codes. Because socio-environmental data was absent in these studies, they were unable to distinguish genetic effects from socio-environmental causes.

In the current study, because of the lack of socio-environmental metadata, we are limited to reporting population-based differences (rather than ancestry-based differences or socio-environmental associations) in gene expression. The very real health benefits that can be gained from metadata access need to be more carefully balanced against privacy concerns. Without routine inclusion and availability of diverse metadata for human 'omics samples, data mining is hampered, and important medical information is lost.

4.4 Conclusion

We have found that genes whose expression differs between African American and European American populations across multiple biological sample types and tissues are deeply associated with multiple pathways and cell types associated with infection, inflammation, environmental

exposures, and immunologic and mucosal cell types that are central to targets-of and defenses-against COVID-19. These differences are evident despite the fact that race is self-reported in the metadata, and many Americans are racially admixed (Zhong et al., 2020). By highlighting the wide-ranging differences in expression of genes implicated in the morbidity and mortality of COVID-19 across populations, and by revealing apparent cell-type differences between populations, we provide baseline signatures that could factor genomics, environmental, and immunologic parameters to improve preventives and therapeutics essential to fight diseases such as COVID-19.

4.5 Methods

4.5.1 Datasets

We selected bulk RNA-Seq data for this study from Genotype Tissue Expression (GTEx, https://gtexportal.org/home/) and The Cancer Genome Atlas (TCGA, https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga). GTEx provides data representing "non-diseased" samples from diverse tissues. Non-diseased refers to the tissue itself, however, in some cases the individual sampled was postmortem and the causes of death are varied. TCGA project is the largest project available on different diseased samples (tumors) of multiple tissue origins. Both projects have metadata on the (self-reported) races of the individuals who contributed samples. These two projects provide a unique opportunity to evaluate differences in gene expression across populations in multiple tissue-types that vary by cite of collection and disease status. Tissue-types were selected for downstream analysis based largely on having sufficient numbers of individuals from each ancestry. (Even between African American and European American populations, not every "non-diseased" tissue or cancer tissue had sufficient samplings of African Americans for robust statistical assessment (Supplementary Table S1)). We refer to those self-reporting as "Black or African American" as "African Americans" and "White" as "European Americans".

The data files and the precompiled MOG project, *MOG_HumanCancerRNASeqProject*, were downloaded from http://metnetweb.gdcb.iastate.edu/MetNet_MetaOmGraph.htm (Singh et al., 2020). This project uses batch-corrected and processed data to enable comparison across samples (Wang et al., 2018). *MOG_HumanCancerRNASeqProject* contains expression values for 18,212 genes, 30 fields of metadata detailing each gene, across 7,142 samples representing 14 different cancer types and associated non-tumor tissues (TCGA and GTEX samples) integrated with 23 fields of metadata describing each study and sample (Singh et al., 2020).

4.5.2 Statistical and correlation analyses

The MOG tool was used to interactively explore, visualize and perform differential expression and correlation analysis of genes.

The Mann-Whitney (MW) test was used to identify DE genes between two groups; we chose this non-parametric analysis as it makes no assumptions about the data distribution. We define a gene as DE 2-fold or more between two groups if it meets each of the following criteria:

1. Estimated fold-change in expression of 2-fold or more (log fold change, $|logFC| \geq 1$), where $logFC$ is calculated as in limma (Ritchie et al., 2015).)

2. Mann-–Whitney U test is significant between the two groups (Benjamini-Hochberg (BH) corrected p-value < 0.05)

Pearson correlation values and Mutual Information values were computed after data was log_2 transformed within MOG, in MOG's statistical analysis module. R scripts were written to create the violin plots; these scripts were executed interactively via MOG.

4.5.3 Covariate evaluation

To check for potential sampling differences between populations that might confound the analysis, we fit linear models using limma (Ritchie et al., 2015) in R, to adjust for for biologically relevant, potential confounding factors of race, gender, tissue/tumor type, age, and as metadata was available, BMI, and cancer subtypes. (Supplementary Table S29-S55).

Because ratios of cancer subtypes may differ between races (as reported for breast cancer in African American women) (Parada et al., 2017; Barrow et al., 2019)), we evaluated the RNA-Seq data from African Americans and European Americans in BRCA samples for potential confounding effects due to different ratios of four breast cancer subtypes: basal-like (BAS), human epidermal growth factor receptor-2 positive/estrogen receptor negative (Her2), luminal A (LumA), and luminal B (LumB) (subtype information was collected using TCGABiolinks (Colaprico et al., 2015)); all genes DE with >2-fold change in MW analysis retained statistical significance in limma analysis of BRCA data, although the fold-change levels varied (Supplementary Table S41). Similarly, we included BMI, where it was available in the metadata, in the limma analysis (Additional File 1).

To assess whether a given distribution shows bi- or multi-modality we applied the Hartigans' dip (Dip) test, using the R package diptest (`https://cran.r-project.org/package=diptest`) (Additional File 2).

4.5.4 Gene expression enrichment

Overrepresentation of biological processes and other functional analysis was assessed at https://toppgene.cchmc.org/. Geneset enrichment analyses (GSEA) were performed using the clusterProfiler library in R (Yu et al., 2012).

4.5.5 Cell-type analysis

African American vs European American gene signatures were compared to cell type and compartment-specific gene signatures using the newly developed cell type specific gene modules available in the ToppGene tool (Chen et al., 2009). The corresponding gene lists in ToppGene were derived from large-scale gene expression signature mining in this case of human cell atlas reference datasets from human esophagus and lung (Jimenez et al., 2020; Travaglini et al., 2020) hosted in ToppCell (http://toppcell.cchmc.org. Heat map visualization of genes differentially-expressed by African Americans versus European Americans in each cell type module

in the selected tissues was done using Morpheus (https://software.broadinstitute.org/morpheus/) using ToppCell's "super binned" gene expression for each cell type within each single cell dataset.

4.5.6 Availability of data and materials

We subscribe to an open data model (https://www.go-fair.org/fair-principles/). MOG is free and open source software published under the MIT License. MOG software, user guide, and the *MOG_HumanCancerRNASeqProject* project datasets and metadata described in this article are freely downloadable from http://metnetweb.gdcb.iastate.edu/MetNet_MetaOmGraph.htm. MOG's source code is available at https://github.com/urmi-21/MetaOmGraph/. Detailed information and code on how to reproduce the results, along with Additional files, are available at https://github.com/urmi-21/COVID-DEA.

4.6 Supplementary data

Supplementary data are available at https://github.com/urmi-21/COVID-DEA and from *Scientific Reports* online.

4.7 Acknowledgements

We are grateful to Mashette Syrkin-Nikolau, Diane Bassham, Karin Dorman and Judy Syrkin-Nikolau for valuable discussions and comments on the manuscript. We thank Afshin Beheshti and Shawn Brown for providing leadership and encouragement in combating COVID. We are grateful for the collaborate spirit of the COVID-19 International Research Team (COVIRT) (https://covirt19.org/) and the International COVID-19 Knowledge Exchange group, and the many helpful interactions we have with researchers in these and other consortia.

4.8 Author contributions statement

US conceived the study, analyzed the data, and wrote the paper.

KH performed GSEA analysis and commented on the paper.

BJA developed and performed the cell-specific profile analysis, wrote the parts of the paper he contributed to, and commented on the paper.

ESW conceived the study, analyzed the data, and wrote the paper.

4.9 Funding

This work is funded in part by the National Science Foundation award IOS 1546858, "Orphan Genes: An Untapped Genetic Reservoir of Novel Traits" and by the Center for Metabolic Biology, Iowa State University. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. In particular, it used the Bridges HPC environment through allocations TG-MCB190098 and TG-MCB200123 awarded from XSEDE and the HPC Consortium.

4.10 Competing interests

The authors declare that they have no competing interests

4.11 References

Ahmadpoor, P. and Rostaing, L. (2020). Why the immune system fails to mount an adaptive immune response to a COVID-19 infection. *Transplant International*, 33(7):824–825.

Amorim, C. E. G., Gao, Z., Baker, Z., Diesel, J. F., Simons, Y. B., Haque, I. S., Pickrell, J., and Przeworski, M. (2017). The population genetics of human disease: The case of recessive, lethal mutations. *PLoS Genetics*, 13(9):e1006915.

Ayres, J. S. (2020). A metabolic handbook for the COVID-19 pandemic. *Nature metabolism*, 2(7):572–585.

Backer, A. (2020). Why COVID-19 may be disproportionately killing african americans: Black overrepresentation among COVID-19 mortality increases with lower irradiance, where ethnicity is more predictive of COVID-19 infection and mortality than median income. *Where Ethnicity Is More Predictive of COVID-19 Infection and Mortality Than Median Income (April 8, 2020)*.

Baharian, S., Barakatt, M., Gignoux, C. R., Shringarpure, S., Errington, J., Blot, W. J., Bustamante, C. D., Kenny, E. E., Williams, S. M., Aldrich, M. C., et al. (2016). The great migration and african-american genomic diversity. *PLoS Genetics*, 12(5):e1006059.

Bansal, M. (2020). Cardiovascular disease and COVID-19. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 14(3):247–250.

Barral-Arca, R., Pardo-Seco, J., Bello, X., Martinon-Torres, F., and Salas, A. (2019). Ancestry patterns inferred from massive rna-seq data. *RNA*, 25(7):857–868.

Barrow, M. A., Martin, M. E., Coffey, A., Andrews, P. L., Jones, G. S., Reaves, D. K., Parker, J. S., Troester, M. A., and Fleming, J. M. (2019). A functional role for the cancer disparity-linked genes, cryβb2 and cryβb2p1, in the promotion of breast cancer. *Breast Cancer Research*, 21(1):1–13.

Battagello, D. S., Dragunas, G., Klein, M. O., Ayub, A. L., Velloso, F. J., and Correa, R. G. (2020). Unpuzzling COVID-19: tissue-related signaling pathways associated with sars-cov-2 infection and transmission. *Clinical Science*, 134(16):2137–2160.

Brumpton, B. M. and Ferreira, M. A. (2016). Multivariate eqtl mapping uncovers functional variation on the x-chromosome associated with complex disease traits. *Human genetics*, 135(7):827–839.

Burt, V. L., Whelton, P., Roccella, E. J., Brown, C., Cutler, J. A., Higgins, M., Horan, M. J., and Labarthe, D. (1995). Prevalence of hypertension in the us adult population: results from the third national health and nutrition examination survey, 1988-1991. *Hypertension*, 25(3):305–313.

Carmona-Gutierrez, D., Bauer, M. A., Zimmermann, A., Kainz, K., Hofer, S. J., Kroemer, G., and Madeo, F. (2020). Digesting the crisis: autophagy and coronaviruses. *Microbial Cell*, 7(5):119.

Chen, J., Bardes, E. E., Aronow, B. J., and Jegga, A. G. (2009). Toppgene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic acids research*, 37(suppl_2):W305–W311.

Chevrier, S., Zurbuchen, Y., Cervia, C., Adamo, S., Raeber, M. E., de Souza, N., Sivapatham, S., Jacobs, A., Bächli, E., Rudiger, A., et al. (2020). A distinct innate immune signature marks progression from mild to severe COVID-19. *bioRxiv*, 2(1):100166.

Chi, C., Shao, X., Rhead, B., Gonzales, E., Smith, J. B., Xiang, A. H., Graves, J., Waldman, A., Lotze, T., Schreiner, T., et al. (2019). Admixture mapping reveals evidence of differential multiple sclerosis risk by genetic ancestry. *PLoS Genetics*, 15(1):e1007808.

Chowdhury, R. and Maranas, C. D. (2020). Biophysical characterization of the sars-cov2 spike protein binding with the ace2 receptor explains increased COVID-19 pathogenesis. *bioRxiv*.

Chua, R. L., Lukassen, S., Trump, S., Hennig, B. P., Wendisch, D., Pott, F., Debnath, O., Thürmann, L., Kurth, F., Völker, M. T., et al. (2020). COVID-19 severity correlates with airway epithelium–immune cell interactions identified by single-cell analysis. *Nature biotechnology*, 38(8):970–979.

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T. S., Malta, T. M., Pagnotta, S. M., Castiglioni, I., et al. (2015). Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data. *Nucleic acids research*, 44(8):e71–e71.

Cole, S. W. (2014). Human social genomics. *PLoS Genetics*, 10(8):e1004601.

Daub, C. O., Steuer, R., Selbig, J., and Kloska, S. (2004). Estimating mutual information using b-spline functions–an improved similarity measure for analysing gene expression data. *BMC bioinformatics*, 5(1):118.

Del Valle, D. M., Kim-Schulze, S., Huang, H.-H., Beckmann, N. D., Nirenberg, S., Wang, B., Lavin, Y., Swartz, T. H., Madduri, D., Stock, A., et al. (2020). An inflammatory cytokine signature predicts COVID-19 severity and survival. *Nature medicine*, 26(10):1–8.

Devaux, C. A., Rolain, J.-M., and Raoult, D. (2020). ACE2 receptor polymorphism: Susceptibility to SARS-CoV-2, hypertension, multi-organ failure, and COVID-19 disease outcome. *Journal of Microbiology, Immunology and Infection*, 53(3):425–435.

Didangelos, A. (2020). COVID-19 hyperinflammation: What about neutrophils? *mSphere*, 5(3):e00367–20.

Dieckmann, L., Cole, S., and Kumsta, R. (2020). Stress genomics revisited: gene co-expression analysis identifies molecular signatures associated with childhood adversity. *Translational psychiatry*, 10(1):1–11.

Ellinghaus, D., Degenhardt, F., Bujanda, L., Buti, M., Albillos, A., Invernizzi, P., Fernández, J., Prati, D., Baselli, G., Asselta, R., et al. (2020). Genomewide association study of severe Covid-19 with respiratory failure. *New England Journal of Medicine*, 383(16):1522–1534.

Favé, M.-J., Lamaze, F. C., Soave, D., Hodgkinson, A., Gauvin, H., Bruat, V., Grenier, J.-C., Gbeha, E., Skead, K., Smargiassi, A., et al. (2018). Gene-by-environment interactions in urban populations modulate risk phenotypes. *Nature communications*, 9(1):1–12.

Friedman, P. N., Shaazuddin, M., Gong, L., Grossman, R. L., Harralson, A. F., Klein, T. E., Lee, N. H., Miller, D. C., Nutescu, E. A., O'Brien, T. J., et al. (2019). The acco u nt consortium: A model for the discovery, translation, and implementation of precision medicine in african americans. *Clinical and translational science*, 12(3):209–217.

Frithiof, R., Bergqvist, A., Järhult, J. D., Lipcsey, M., and Hultström, M. (2020). Presence of SARS-CoV-2 in urine is rare and not associated with acute kidney injury in critically ill COVID-19 patients. *Critical Care*, 24(1):1–3.

Furr-Stimming, E., Shiyu, X., Ye, X., Zhang, S., et al. (2020). Hap40 is a conserved partner and regulator of huntingtin and a pathogenic modifier of huntington's disease (2817).

Gassen, N. C., Papies, J., Bajaj, T., Dethloff, F., Emanuel, J., Weckmann, K., Heinz, D. E., Heinemann, N., Lennarz, M., Richter, A., et al. (2020). Analysis of sars-cov-2-controlled autophagy reveals spermidine, mk-2206, and niclosamide as putative antiviral therapeutics. *bioRxiv*.

Gormley, M., Ona, K., Kapidzic, M., Garrido-Gomez, T., Zdravkovic, T., and Fisher, S. J. (2017). Preeclampsia: novel insights from global RNA profiling of trophoblast subpopulations. *American journal of obstetrics and gynecology*, 217(2):200–e1.

Graw, J., Brackmann, H.-H., Oldenburg, J., Schneppenheim, R., Spannagl, M., and Schwaab, R. (2005). Haemophilia a: from mutation analysis to new therapies. *Nature Reviews Genetics*, 6(6):488.

Hooten, N. N. and Evans, M. K. (2019). Age and poverty status alter the coding and noncoding transcriptome. *Aging (Albany NY)*, 11(4):1189.

Huang, Z.-N., Chung, H. M., Fang, S.-C., and Her, L.-S. (2017). Adhesion regulating molecule 1 mediates hap40 overexpression-induced mitochondrial defects. *International journal of biological sciences*, 13(11):1420.

Jarrahi, A., Ahluwalia, M., Khodadadi, H., Salles, E. d. S. L., Kolhe, R., Hess, D. C., Vale, F., Kumar, M., Baban, B., Vaibhav, K., et al. (2020). Neurological consequences of COVID-19: what have we learned and where do we go from here? *Journal of Neuroinflammation*, 17(1):1–12.

Jeyanathan, M., Afkhami, S., Smaill, F., Miller, M. S., Lichty, B. D., and Xing, Z. (2020). Immunological considerations for COVID-19 vaccine strategies. *Nature Reviews Immunology*, 20(10):1–18.

Jimenez, L., Codo, A. C., Sampaio, V. S., Oliveira, A. E., Ferreira, L. K., Davanzo, G. G., Monteiro, L. B., Virgilio-da Silva, J. V., Borba, M. G., Souza, G. F., et al. (2020). The influence of ph on sars-cov-2 infection and COVID-19 severity. *medRxiv*.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, 45(D1):D353–D361.

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *bioRxiv*, page 531210.

Kruzel-Davila, E., Wasser, W. G., and Skorecki, K. (2017). Apol1 nephropathy: a population genetics and evolutionary medicine detective story. In *Seminars in nephrology*, volume 37, pages 490–507. Elsevier.

Madissoon, E., Wilbrey-Clark, A., Miragaia, R., Saeb-Parsy, K., Mahbubani, K., Georgakopoulos, N., Harding, P., Polanski, K., Huang, N., Nowicki-Osuch, K., et al. (2020). scrna-seq assessment of the human lung, spleen, and esophagus tissue stability after cold preservation. *Genome biology*, 21(1):1–16.

Martinez, T. F., Chu, Q., Donaldson, C., Tan, D., Shokhirev, M. N., and Saghatelian, A. (2020). Accurate annotation of human protein-coding small open reading frames. *Nature chemical biology*, 16(4):458–468.

Matsuyama, S., Nao, N., Shirato, K., Kawase, M., Saito, S., Takayama, I., Nagata, N., Sekizuka, T., Katoh, H., Kato, F., et al. (2020). Enhanced isolation of sars-cov-2 by tmprss2-expressing cells. *Proceedings of the National Academy of Sciences*, 117(13):7001–7003.

McGuire, A. L., Gabriel, S., Tishkoff, S. A., Wonkam, A., Chakravarti, A., Furlong, E. E., Treutlein, B., Meissner, A., Chang, H. Y., López-Bigas, N., et al. (2020). The road ahead in genetics and genomics. *Nature Reviews Genetics*, 21(10):1–16.

Millett, G. A., Jones, A. T., Benkeser, D., Baral, S., Mercer, L., Beyrer, C., Honermann, B., Lankiewicz, E., Mena, L., Crowley, J. S., et al. (2020). Assessing differential impacts of COVID-19 on Black communities. *Annals of Epidemiology*, 47:37–44.

Morillon, A. and Gautheret, D. (2019). Bridging the gap between reference and real transcriptomes. *Genome biology*, 20(1):1–7.

Nédélec, Y., Sanz, J., Baharian, G., Szpiech, Z. A., Pacis, A., Dumaine, A., Grenier, J.-C., Freiman, A., Sams, A. J., Hebert, S., et al. (2016). Genetic ancestry and natural selection drive population differences in immune responses to pathogens. *Cell*, 167(3):657–669.

Need, A. C. and Goldstein, D. B. (2009). Next generation disparities in human genomics: concerns and remedies. *Trends in Genetics*, 25(11):489–494.

O'Brien, J. S., Okada, S., Fillerup, D. L., Veath, M. L., Adornato, B., Brenner, P. H., and Leroy, J. G. (1971). Tay-sachs disease: prenatal diagnosis. *Science*, 172(3978):61–64.

Pal, A., Severin, F., Lommer, B., Shevchenko, A., and Zerial, M. (2006). Huntingtin–hap40 complex is a novel rab5 effector that regulates early endosome motility and is up-regulated in huntington's disease. *The Journal of cell biology*, 172(4):605–618.

Parada, H., Sun, X., Fleming, J. M., Williams-DeVane, C. R., Kirk, E. L., Olsson, L. T., Perou, C. M., Olshan, A. F., and Troester, M. A. (2017). Race-associated biological differences among luminal a and basal-like breast cancers in the carolina breast cancer study. *Breast Cancer Research*, 19(1):131.

Paulucci, D. J., Sfakianos, J. P., Skanderup, A. J., Kan, K., Tsao, C.-K., Galsky, M. D., Hakimi, A. A., and Badani, K. K. (2017). Genomic differences between black and white patients implicate a distinct immune response to papillary renal cell carcinoma. *Oncotarget*, 8(3):5196.

Perez-Riba, A. and Itzhaki, L. S. (2019). The tetratricopeptide-repeat motif is a versatile platform that enables diverse modes of molecular recognition. *Current opinion in structural biology*, 54:43–49.

Peters, M. F. and Ross, C. A. (2001). Isolation of a 40-kda huntingtin-associated protein. *Journal of Biological Chemistry*, 276(5):3188–3194.

Quintana-Murci, L. (2019). Human immunology through the lens of evolutionary genetics. *Cell*, 177(1):184–199.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47.

Robilotti, E. V., Babady, N. E., Mead, P. A., Rolling, T., Perez-Johnston, R., Bernardes, M., Bogler, Y., Caldararo, M., Figueroa, C. J., Glickman, M. S., et al. (2020). Determinants of COVID-19 disease severity in patients with cancer. *Nature medicine*, 26(8):1218–1223.

Roforth, M. M., Farr, J. N., Fujita, K., McCready, L. K., Atkinson, E. J., Therneau, T. M., Cunningham, J. M., Drake, M. T., Monroe, D. G., and Khosla, S. (2015). Global transcriptional profiling using rna sequencing and dna methylation patterns in highly enriched mesenchymal cells from young versus elderly women. *Bone*, 76:49–57.

Schönrich, G., Raftery, M. J., and Samstag, Y. (2020). Devilishly radical network in COVID-19: Oxidative stress, neutrophil extracellular traps (nets), and t cell suppression. *Advances in Biological Regulation*, 77:100741.

Sies, H. and Jones, D. P. (2020). Reactive oxygen species (ros) as pleiotropic physiological signalling agents. *Nature Reviews Molecular Cell Biology*, 21(7):1–21.

Singh, U., Hur, M., Dorman, K. S., and Wurtele, E. S. (2020). Metaomgraph: a workbench for interactive exploratory data analysis of large expression datasets. *Nucleic Acids Research*, 48(4):e23–e23. gkz1209.

Snell, J. (2020). SARS-CoV-2 infection and its association with thrombosis and ischemic stroke: A review COVID-19, thrombosis, and ischemic stroke. *The American Journal of Emergency Medicine*, 40:188–192.

Struyf, S., Menten, P., Lenaerts, J.-P., Put, W., D'Haese, A., De Clercq, E., Schols, D., Proost, P., and Van Damme, J. (2001). Diverging binding capacities of natural ld78β isoforms of macrophage inflammatory protein-1α to the cc chemokine receptors 1, 3 and 5 affect their anti-hiv-1 activity and chemotactic potencies for neutrophils and eosinophils. *European journal of immunology*, 31(7):2170–2178.

Tao, S. and Drexler, I. (2020). Targeting autophagy in innate immune cells: Angel or demon during infection and vaccination? *Frontiers in Immunology*, 11:460.

Tay, M. Z., Poh, C. M., Rénia, L., MacAry, P. A., and Ng, L. F. (2020). The trinity of COVID-19: immunity, inflammation and intervention. *Nature Reviews Immunology*, 20(6):1–12.

Teuwen, L.-A., Geldhof, V., Pasut, A., and Carmeliet, P. (2020). COVID-19: the vasculature unleashed. *Nature Reviews Immunology*, 20(7):1–3.

Travaglini, K. J., Nabhan, A. N., Penland, L., Sinha, R., Gillich, A., Sit, R. V., Chang, S., Conley, S. D., Mori, Y., Seita, J., et al. (2020). A molecular cell atlas of the human lung from single cell rna sequencing. *BioRxiv*, 587(7835):742320.

Varga, Z., Flammer, A. J., Steiger, P., Haberecker, M., Andermatt, R., Zinkernagel, A. S., Mehra, M. R., Schuepbach, R. A., Ruschitzka, F., and Moch, H. (2020). Endothelial cell infection and endotheliitis in COVID-19. *The Lancet*, 395(10234):1417–1418.

Vaske, O. M., Bjork, I., Salama, S. R., Beale, H., Shah, A. T., Sanders, L., Pfeil, J., Lam, D. L., Learned, K., Durbin, A., et al. (2019). Comparative tumor rna sequencing analysis for difficult-to-treat pediatric and young adult patients with cancer. *JAMA network open*, 2(10):e1913968–e1913968.

Veras, F. P., Pontelli, M., Silva, C., Toller-Kawahisa, J., de Lima, M., Nascimento, D., Schneider, A., Caetite, D., Rosales, R., Colon, D., et al. (2020). Sars-cov-2 triggered neutrophil extracellular traps (nets) mediate COVID-19 pathology. *medRxiv*, 217(12):e20201129.

Wang, Q., Armenia, J., Zhang, C., Penson, A. V., Reznik, E., Zhang, L., Minet, T., Ochoa, A., Gross, B. E., Iacobuzio-Donahue, C. A., et al. (2018). Unifying cancer and normal rna sequencing data from different sources. *Scientific data*, 5:180061.

Warren, R. L. and Birol, I. (2020). Hla predictions from the bronchoalveolar lavage fluid samples of five patients at the early stage of the wuhan seafood market COVID-19 outbreak. *arXiv preprint arXiv:2004.07108*.

Williams, F. M., Freydin, M., Mangino, M., Couvreur, S., Visconti, A., Bowyer, R. C., Le Roy, C. I., Falchi, M., Sudre, C., Davies, R., et al. (2020). Self-reported symptoms of COVID-19 including symptoms most predictive of sars-cov-2 infection, are heritable. *medRxiv*, 23(6):316–321.

Woo, Y.-L., Kamarulzaman, A., Augustin, Y., Staines, H., Altice, F., and Krishna, S. (2020). A genetic predisposition for cytokine storm in life-threatening COVID-19 infection.

Wu, M., Miska, J., Xiao, T., Zhang, P., Kane, J. R., Balyasnikova, I. V., Chandler, J. P., Horbinski, C. M., and Lesniak, M. S. (2019). Race influences survival in glioblastoma patients with kps 80 and associates with genetic markers of retinoic acid metabolism. *Journal of neuro-oncology*, 142(2):375–384.

Xu, S., Li, G., Ye, X., Chen, D., Chen, Z., Xu, Z., Ye, L., Stimming, E. F., Marchionini, D., and Zhang, S. (2020). Hap40 is a conserved central regulator of huntingtin and a specific modulator of mutant huntingtin toxicity. *bioRxiv*.

Yang, N. and Shen, H.-M. (2020). Targeting the endocytic pathway and autophagy process as a novel therapeutic strategy in COVID-19. *International journal of biological sciences*, 16(10):1724.

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterprofiler: an r package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, 16(5):284–287.

Yuan, J., Hu, Z., Mahal, B. A., Zhao, S. D., Kensler, K. H., Pi, J., Hu, X., Zhang, Y., Wang, Y., Jiang, J., et al. (2018). Integrated analysis of genetic ancestry and genomic alterations across cancers. *Cancer cell*, 34(4):549–560.

Zhai, B., He, Y., Zhou, S., Wang, X., and Wang, R. Characterization of lung bronchoalveolar humoral immunity in patients with COVID-19.

Zhong, Y., De, T., Alarcon, C., Park, C. S., Lec, B., and Perera, M. A. (2020). Discovery of novel hepatocyte eqtls in african americans. *PLoS Genetics*, 16(4):e1008662.

Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., Xiang, J., Wang, Y., Song, B., Gu, X., et al. (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in wuhan, china: a retrospective cohort study. *The lancet*, 395(10229):1054–1062.

A

tissue	infection/immune-response-related (# pathways enriched)	CK-CK interaction pathway	glutathione-oxidative-metabolic-processes (# pathways enriched)
kidney	4	yes	2
breast	6	no	3
esophagus	8	no	0
lung	8	no	2
uterus	8	yes	1
stomach	9	yes	3
thyroid	10	no	0
prostate	13	yes	0
colon	15	yes	0
liver	17	yes	0
pooledGTEx	10	yes	3

B

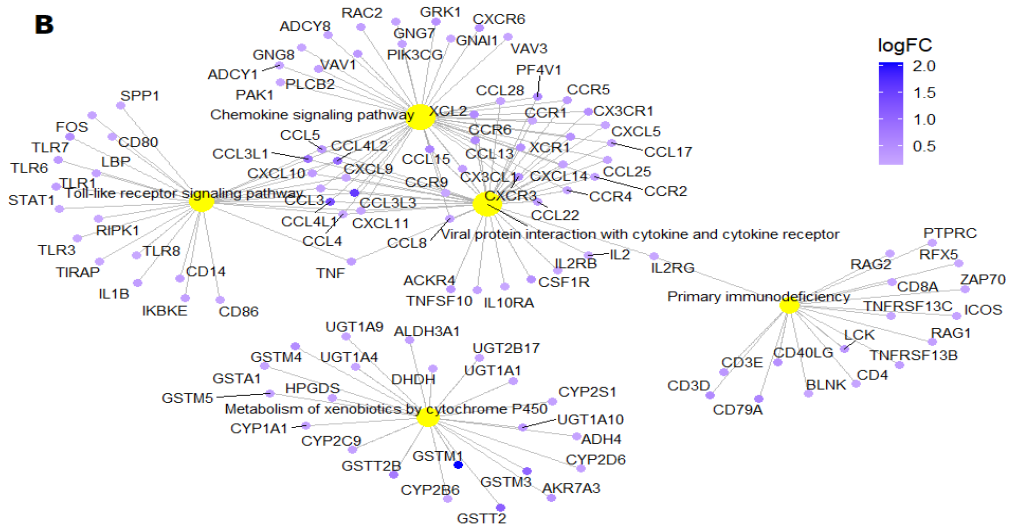

Figure 4.1 Gene Set Enrichment Analysis (GSEA) enrichment of KEGG pathways in African Americans compared to European Americans in pooled GTEx data. GSEA comprehensively analyses data for expression of all genes, rather than only the DE genes. **A.** The most common pathways enriched among upregulated genes in African Americans for tissue-types in GTEx. See Additional File 3 for complete list of enriched pathways in 25 tissue-types. CK-CK, cytokine-cytokine receptor interaction; glutathione-oxidative metabolism includes (oxidative) metabolism of xenobiotics. The full enrichment analysis for each tissue-type is shown in Supplementary Table S57-S59. **B.** The five most highly enriched pathways among upregulated genes of pooled samples from all tissue–types in GTEx are: Tol-like receptor signaling; chemokine signaling; primary immunodeficiency; viral protein interaction with cytokine and cytokine receptor; metabolism of xenobiotics by cytochrome P450.

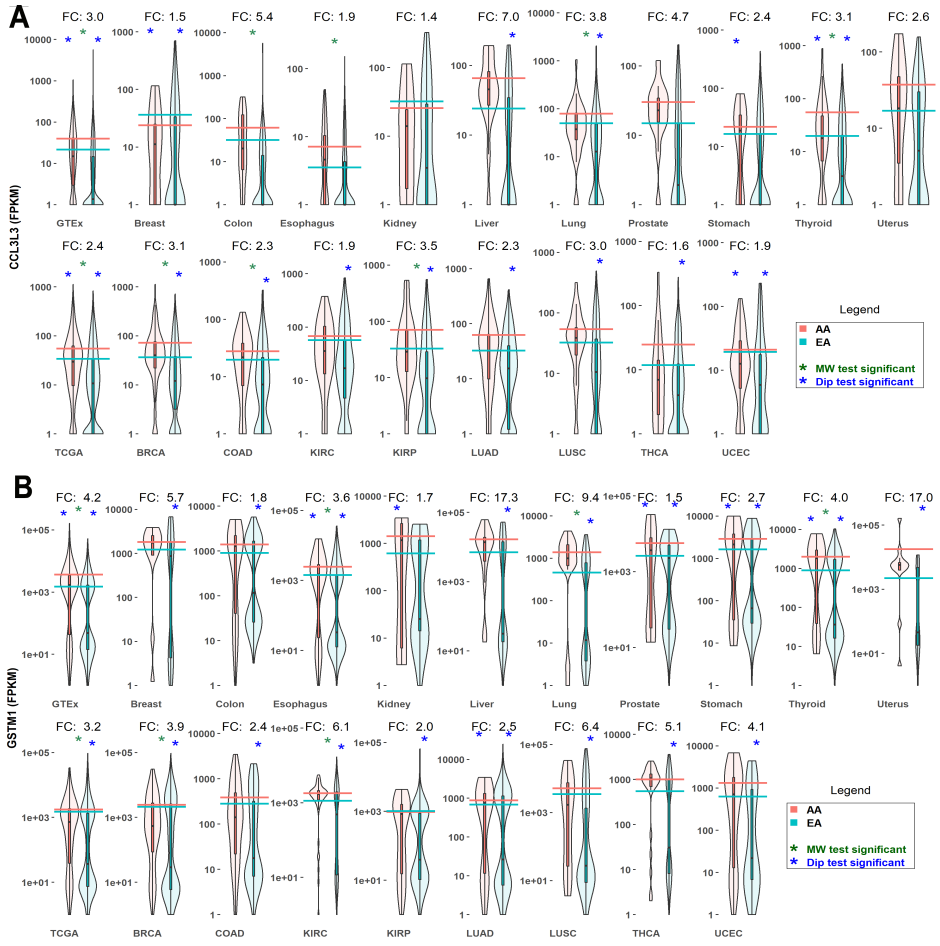

Figure 4.2 Upregulated expression of chemokine **CCL3L3** and mitochondrial glutathione-S-transferase **GSTM1** in African Americans compared to European Americans across multiple conditions. **A. CCL3L3** is more highly expressed in African Americans over a wide range of tissue-types. CL3L3 binds to chemokine receptor proteins CCR1, CCR3, and CCR5. **B. GSTM1** is more highly expressed in African Americans over a wide range of tissue-types. GSTM1 is a key player in metabolism of ROS and xenobiotics. (See Supplementary Tables S2-S28 for complete DE analysis). Violin plots summarize expression over each sample across the two populations. AA, African American; EA, European American. Horizontal lines represent mean log expression. $*$, Mann-Whitney (MW) test for DE significant (Benjamini-Hochberg (BH) corrected p-value < 0.05). $*$, Hartigans' dip test. Expression distribution is influenced by differences in population sizes (significant p-value < 0.05). FC, fold change AA/EA. GTEx and TCGA violin plots represent the pooled samples from each project. DE were computed within MetaOmGraph (MOG) (Singh et al., 2020), in MOG's statistical analysis module; R scripts were executed interactively via MOG to generate the violin plots.

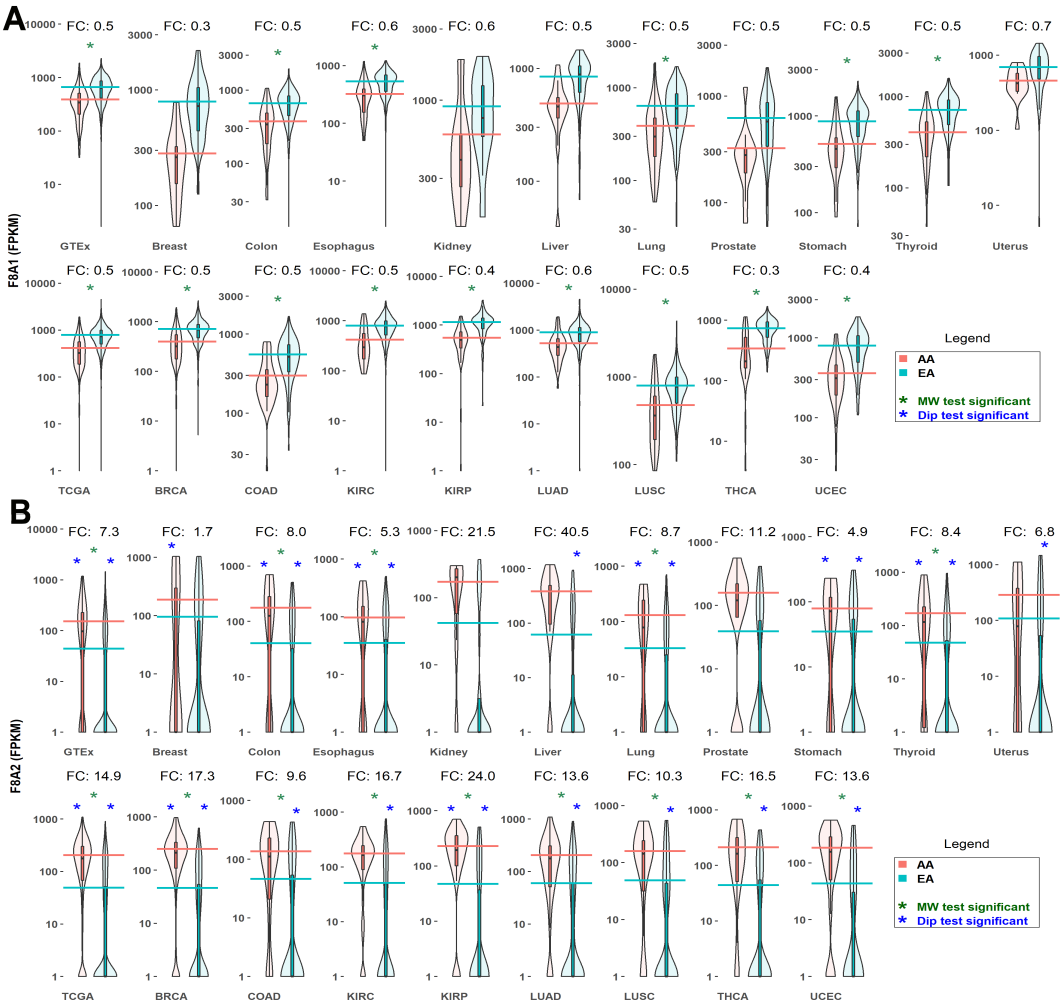

Figure 4.3 Differential expression of the HAP40 genes **F8A1** and **F8A2** in African Americans and European Americans across multiple tissue-types. HAP40 is a key molecular component of Huntington's Disease, and shifts endosomal trafficking from the microtubules to actin fibers (Pal et al., 2006). **A. F8A1** expression is upregulated in European Americans. **B. F8A2** expression is upregulated in African Americans. Violin plots summarize expression over each sample across the two populations. AA, African American; EA, European American. Horizontal lines represent mean log expression. $*$, MW test for DE significant (BH corrected p-value < 0.05). $*$, Hartigans' dip test. Expression distribution is influenced by differences in population sizes (significant p-value < 0.05). FC, fold change AA/EA. GTEx and TCGA violin plots represent the pooled samples from each project. DE were computed within MetaOmGraph (MOG) (Singh et al., 2020), in MOG's statistical analysis module; R scripts were executed interactively via MOG to generate the violin plots. (See Supplementary Figure 2 for line plot comparison across individuals)

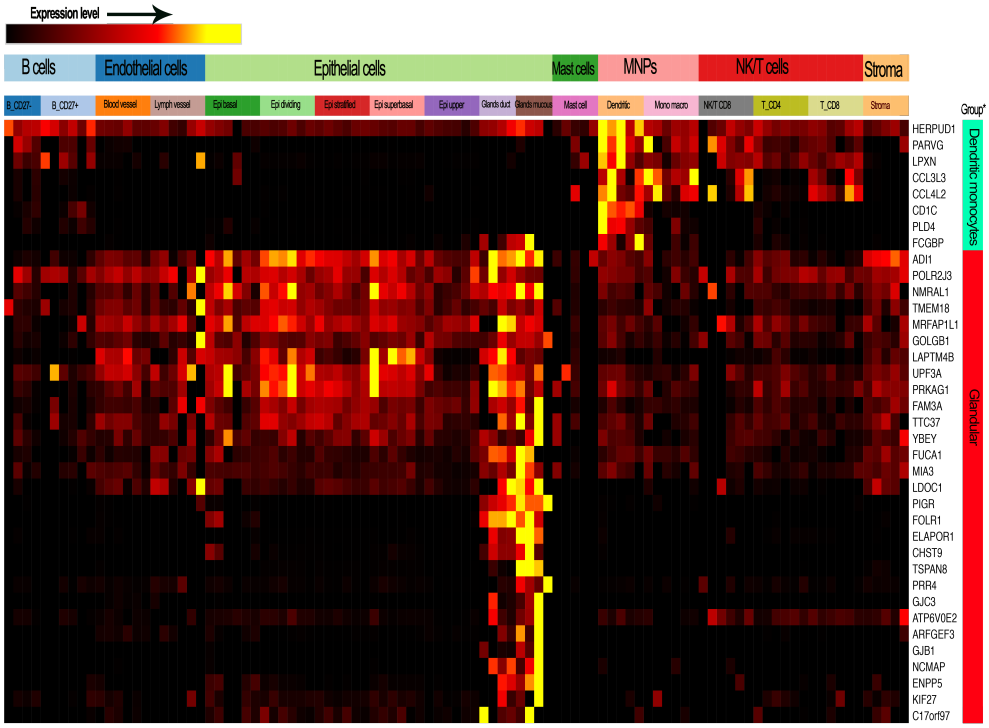

Figure 4.4 Esophageal genes that are differentially expressed in African Americans and European American samples correspond to genes known to be expressed in specific cell types. Genes upregulated in African American versus European American esophagus mapped to two cell lineages with prominent presence in the esophageal tissue stability dataset of the human cell atlas (https://data.humancellatlas.org/). One significant fraction of the African American-upregulated gene signature maps to glandular mucous epithelial cells of esophageal glands (genes marked by red, far right bar). Expression of several of the genes upregulated in African Americans is highly restricted to the mucous epithelial cells (TSAPN8, PRR4, ELAPOR1), whereas FOLR1, for example, is more highly expressed in the ductal epithelial cells of mucosal glands. A second, smaller, signature corresponds to hematolymphoid/myeloid lineage dendritic cells, as shown by CDC1C, PLD4, HERPUD1, and LPXN (genes marked by green, far right bar). In addition the genes that are most strongly expressed by those cell types, additional genes of the AA vs EA esophageal signature included several genes that are essentially exclusively expressed by those cell types. Toppcell-constructed gene modules (http://toppcell.cchmc.org) for each of the cell types reported to be present in the large scRNA-Seq dataset from esophagus (Madissoon et al., 2020)

.

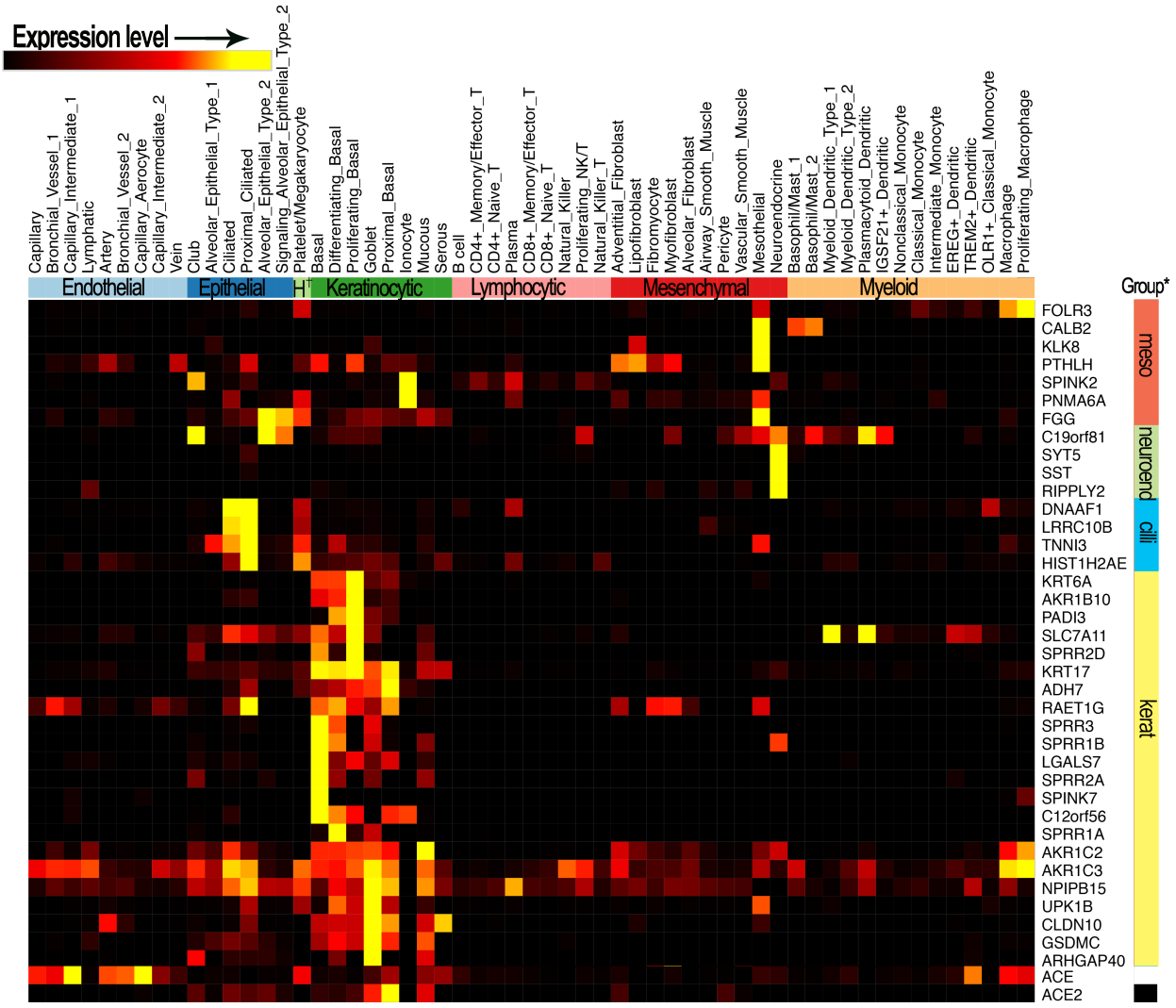

Figure 4.5 Lung gene signatures upregulated in African Americans versus European Americans map to proximal airway keratinocytic epithelial lineage, and to mesenchymal mesothelial and neuroendocrine cells. Marker genes for kerotinocytes (genes marked by yellow, far right bar); ciliated epithelial cells (genes marked by turquoise, far right bar); mesothelial mesenchymal cells (genes marked by red, far right bar); and neuroendocrine mesenchymal cells (mesenchymal). Note that the keratinocytic proximal basal epithelial cell is the cell subtype with the highest expression of ACE2 receptor, a major target of COVID-19 (ACE2 marked by black on bar at right). ToppCell-constructed gene modules (http://toppcell.cchmc.org) for each of the cell types reported to be present in the large scRNA-Seq dataset from lung (Travaglini et al., 2020).

CHAPTER 5. ORFIPY: A FAST AND FLEXIBLE TOOL FOR EXTRACTING ORFS

Urminder Singh [1,2,3], and Eve Syrkin Wurtele [1,2,3]

[1]Bioinformatics and Computational Biology Program, Iowa State University, Ames, IA 50011.

[2]Center for Metabolic Biology, Iowa State University, Ames, IA 50011.

[3]Department of Genetics Development and Cell Biology, Iowa State University, Ames, IA 50011.

Modified from a manuscript published in *Bioinformatics*

5.1 Abstract

Summary: Searching for open reading frames is a routine task and a critical step prior to annotating protein coding regions in newly sequenced genomes or *de novo* transcriptome assemblies. With the tremendous increase in genomic and transcriptomic data, faster tools are needed to handle large input datasets. These tools should be versatile enough to fine-tune search criteria and allow efficient downstream analysis. Here we present a new python based tool, `orfipy`, which allows the user to flexibly search for open reading frames in genomic and transcriptomic sequences. The search is rapid and is fully customizable, with a choice of FASTA and BED output formats. **Availability and implementation:** `orfipy` is implemented in python and is compatible with python v3.6 and higher. Source code: https://github.com/urmi-21/orfipy. Installation: from the source, or via PyPi (https://pypi.org/project/orfipy) or bioconda (https://anaconda.org/bioconda/orfipy).

5.2 Introduction

Open reading frames (ORFs) are sequences that have potential to be translated into proteins. They are delineated by start sites, at which translation is initiated by assembly of a ribosome

complex, and stop sites, at which translation is terminated and the ribosome complex disassembles (Sieber et al., 2018).

Accurate annotation of the protein coding regions in sequenced genomes remains a challenging task in bioinformatics. For simpler prokaryotic genomes, ORFs correspond to the potential coding sequences (CDS) (Sieber et al., 2018). In eukaryotes, where gene splicing is prevalent, eukaryotic CDS prediction a much more challenging task (Sieber et al., 2018; Seetharam et al., 2019).

Transcriptomic data is critical in addressing this challenge, where presence of an ORF in a mature tranrscript may indicate a potential protein coding gene (Seetharam et al., 2019; Martinez et al., 2020; Mahmood et al., 2020). These data are key to identifying potential orphan genes (Seetharam et al., 2019), young genes unique to a species (Tautz and Domazet-Lošo, 2011; Singh and Wurtele, 2020; Vakirlis et al., 2020); standard *ab initio* gene-prediction models are trained on canonical gene features and do not work well for identifying orphan genes, which are often sparse in canonical gene features (Ruiz-Orera et al., 2015; Seetharam et al., 2019; Heames et al., 2020).

Depending on data (genomic, transcriptomic or metagenomic) and researcher interest, the computational problem of ORF prediction may be stated in multiple ways (Sieber et al., 2018), yet existing tools lack the flexibility to allow users to fine-tune or customize the search for ORF sequences. Here we present `orfipy`, an efficient tool for extracting ORFs from nucleotide sequences. `orfipy` provides rapid, flexible searches in multiple output formats to allow easy downstream analysis of ORFs.

5.3 Implementation

`orfipy` is written in python, with the core ORF search algorithm implemented in cython to achieve faster execution times. `orfipy` uses the pyfastx library (Du et al., 2020) for efficient parsing of input FASTA/FASTQ file. `orfipy` can leverage multiple cpu-cores to process FASTA sequences in parallel, based on available memory and cpu cores (Supplementary Data).

5.3.1 Input, flexible search and output

`orfipy` takes nucleotide sequences in a multi-FASTA/FASTQ, plain or gz-compressed, file as input. Users can provide input parameters that include minimum and maximum size of ORFs, list of start and stop codons, and/or a user-defined codon table (Supplementary Data). For efficient and flexible downstream analysis (Figure 6.1A, B), `orfipy` provides multiple output types including BED format. BED files reduce disk space use by storing only the coordinates of the ORFs, and are useful in developing more scalable, flexible downstream analysis pipelines. `orfipy` also adds relevant information about codon use and ORF types, and can group the output by longest ORF contained in each transcript, or can list each reading frame in each transcript.

`orfipy` enables researchers to fully fine-tune ORF searches using a variety of options (Figure 6.1A). For example, users can limit ORF searching to a specific start codon or choose to output ORFs without an inframe start codon. `orfipy` labels each ORF for users to easily comprehend results (Supplementary Data).

5.3.2 Comparison with existing tools

We compared `orfipy` with two popular ORF searching tools, `getorf` (Rice et al., 2000) and `OrfM` (Woodcroft et al., 2016). What sets `orfipy` apart is its flexibility and the options to fine-tune ORF searches and output (Figure 6.1A, B). Runtimes (Figure 6.1C, D) depend on software, environment, input (FASTA input is shown) and output-type. In all scenarios except using a PC to analyse the *A. thaliana* genome, `orfipy` is much faster than `getorf`, and comparable to `OrfM`, with `OrfM` being faster for FASTQ input (Supplementary Data).

5.4 Acknowledgements

We are grateful to all reviewers for careful reviews and helpful suggestions. We thank the developer of `OrfM`, Ben Woodcroft, for suggestions regarding extending `orfipy` functionality. We thank all the members of Wurtele lab.

5.5 Supplementary Data

Supplementary Data are available at Bioinformatics online.

5.6 Funding

This work is funded in part by National Science Foundation grant IOS 1546858, Orphan Genes: An Untapped Genetic Reservoir of Novel Traits, and by the Center for Metabolic Biology, Iowa State University. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562, through allocations TG-MCB190098 and TG-MCB200123 awarded from XSEDE.

5.7 References

Du, L., Liu, Q., Fan, Z., Tang, J., Zhang, X., Price, M., Yue, B., and Zhao, K. (2020). Pyfastx: a robust python package for fast random access to sequences from plain and gzipped fasta/q files. *Briefings in Bioinformatics*.

Heames, B., Schmitz, J., and Bornberg-Bauer, E. (2020). A continuum of evolving de novo genes drives protein-coding novelty in drosophila. *Journal of molecular evolution*, pages 1–17.

Mahmood, K., Orabi, J., Kristensen, P. S., Sarup, P., Jørgensen, L. N., and Jahoor, A. (2020). De novo transcriptome assembly, functional annotation, and expression profiling of rye (secale cereale l.) hybrids inoculated with ergot (claviceps purpurea). *Scientific reports*, 10(1):1–16.

Martinez, T. F., Chu, Q., Donaldson, C., Tan, D., Shokhirev, M. N., and Saghatelian, A. (2020). Accurate annotation of human protein-coding small open reading frames. *Nature chemical biology*, 16(4):458–468.

Rice, P., Longden, I., and Bleasby, A. (2000). Emboss: the european molecular biology open software suite.

Ruiz-Orera, J., Hernandez-Rodriguez, J., Chiva, C., Sabidó, E., Kondova, I., Bontrop, R., Marqués-Bonet, T., and Albà, M. M. (2015). Origins of de novo genes in human and chimpanzee. *PLoS Genetics*, 11(12):e1005721.

Seetharam, A. S., Singh, U., Li, J., Bhandary, P., Arendsee, Z., and Wurtele, E. S. (2019). Maximizing prediction of orphan genes in assembled genomes. *BioRxiv*.

Sieber, P., Platzer, M., and Schuster, S. (2018). The definition of open reading frame revisited. *Trends in Genetics*, 34(3):167–170.

Singh, U., Li, J., Seetharam, A., and Wurtele, E. S. (2021). pyrpipe: a Python package for RNA-Seq workflows. *NAR Genomics and Bioinformatics*, 3(2). lqab049.

Singh, U. and Wurtele, E. S. (2020). Genetic novelty: How new genes are born. *Elife*, 9:e55136.

Tautz, D. and Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nature Reviews Genetics*, 12(10):692–702.

Vakirlis, N., Carvunis, A.-R., and McLysaght, A. (2020). Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *Elife*, 9:e53500.

Woodcroft, B. J., Boyd, J. A., and Tyson, G. W. (2016). OrfM: a fast open reading frame predictor for metagenomic data. *Bioinformatics*, 32(17):2702–2703.

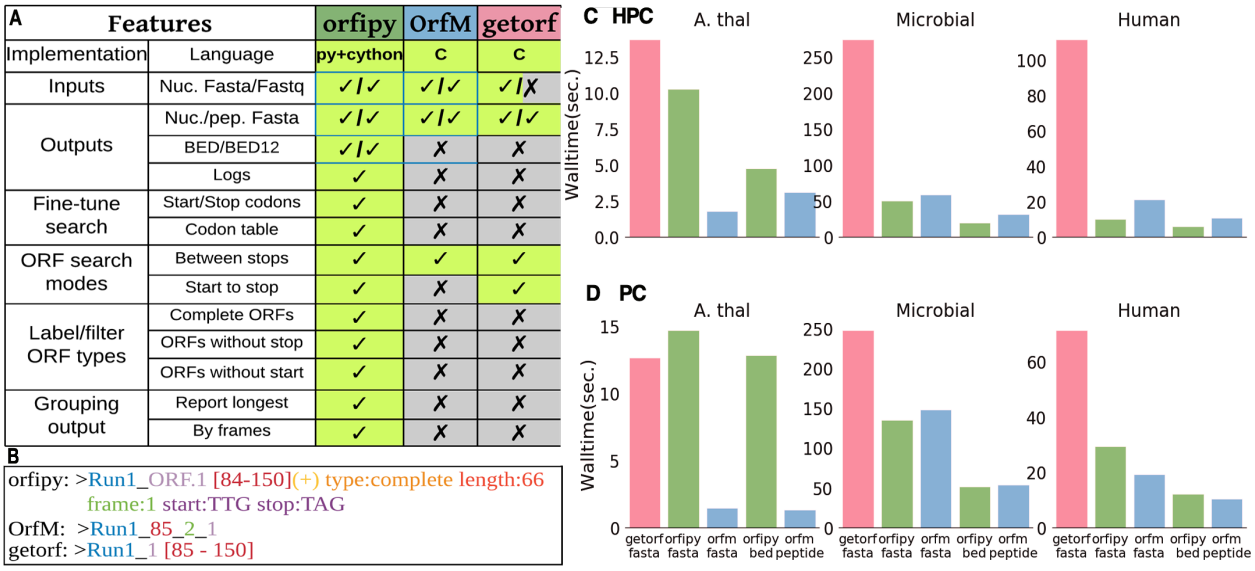

Figure 5.1 **Comparison of `orfipy` features and performance with `getorf` and `OrfM`.** We compared attributes of `orfipy` with two commonly-used tools for ORF identification. **A.** Comparison of `orfipy` features with `getorf` and `OrfM`. `orfipy` provides a number of options to fine-tune ORF search, this includes labeling the ORF type, reporting only the longest ORF, and reporting ORFs by translation frame. To allow reproducible analysis, `orfipy` logs the commands. **B.** Example of FASTA headers written to output files by each tool. `orfipy` output provides information about each ORF that can be readily used in downstream analyses. **C and D.** Runtimes, using plain FASTA input, on HPC (128 GB RAM; 28 cores) (**C**) and PC (16 GB RAM; 8 cores) (**D**) environments (Supplementary Data). Each analysis was run three times, via pyrpipe (Singh et al., 2021), and the mean runtime is reported. `orfipy` runtimes are comparable to `OrfM` for the large microbial and human transcriptome data. `orfipy` is fastest when ORFs are saved to a BED file; `OrfM` is fastest when ORFs are saved to peptide FASTA. Data sizes: *A. thaliana* genome 120 MB; microbial sequences 1.5 GB; human transcriptome 370 MB. fasta, output ORFs to nucleotide *and* peptide FASTA; bed, output ORFs to BED file; peptide, output ORFs to peptide-only FASTA.

CHAPTER 6. PYRPIPE: A PYTHON PACKAGE FOR RNA-SEQ WORKFLOWS

Urminder Singh [1,2,3], Jing Li [2,3], Arun Seetharam [4] and Eve Syrkin Wurtele [1,2,3]

[1]Bioinformatics and Computational Biology Program, Iowa State University, Ames, IA 50011.

[2]Center for Metabolic Biology, Iowa State University, Ames, IA 50011.

[3]Department of Genetics Development and Cell Biology, Iowa State University, Ames, IA 50011.

[4]Genome Informatics Facility, Iowa State University, Ames, IA 50011.

Modified from a manuscript published in *NAR Genomics and Bioinformatics*

6.1 Abstract

The availability of terabytes of RNA-Seq data and continuous emergence of new analysis tools, enable unprecedented biological insight. There is a pressing requirement for a framework that allows for fast, efficient, manageable, and reproducible RNA-Seq analysis. We have developed a Python package, (`pyrpipe`), that enables straightforward development of flexible, reproducible and easy-to-debug computational pipelines purely in Python, in an object-oriented manner. `pyrpipe` provides access to popular RNA-Seq tools, within Python, via high-level APIs. Pipelines can be customized by integrating new Python code, third-party programs, or Python libraries. Users can create checkpoints in the pipeline or integrate `pyrpipe` into a workflow management system, thus allowing execution on multiple computing environments, and enabling efficient resource management. `pyrpipe` produces detailed analysis, and benchmark reports which can be shared or included in publications. `pyrpipe` is implemented in Python and is compatible with Python versions 3.6 and higher. To illustrate the rich functionality of `pyrpipe`, we provide case studies using RNA-Seq data from GTEx, SARS-CoV-2-infected human cells, and *Zea mays*. All source code is freely available at https://github.com/urmi-21/pyrpipe; the package can be

installed from the source, from PyPI (https://pypi.org/project/pyrpipe), or from bioconda (https://anaconda.org/bioconda/pyrpipe). Documentation is available at (http://pyrpipe.rtfd.io).

6.2 Introduction

Since its inception, RNA-Seq has become the most widely used method to quantify transcript levels (Mortazavi et al., 2008; Stark et al., 2019). A researcher can leverage the now-massive RNA-Seq data in public databases, encompassing samples from multiple species, organs, genotypes, and conditions (Kodama et al., 2011). Integrated reanalysis of aggregations of these diverse RNA-Seq samples enables exploration of changes in gene expression over time and across different biological conditions (Singh et al., 2020).

A major challenge in analysis of RNA-Seq datasets is implementing data processing pipelines in an efficient, modular, and reproducible manner (Grüning et al., 2018a; Lampa et al., 2019; Köster and Rahmann, 2012; Di Tommaso et al., 2017). Most bioinformatics tools are executable programs, executed via a command-line interface (CLI), that must be specified inside a scripting language for automated execution. Thus, writing bioinformatics pipelines as Perl, Bash or Python scripts is a common practice among bioinformaticians. Scripting is powerful and flexible. However, plain scripting has several significant downsides (Leipzig, 2017). First, especially for complicated pipelines, bash scripts can be difficult to develop or maintain. Second, for beginners it is hard to write bash scripts in a robust manner that can handle exceptions or resolve errors dynamically. Writing multiple commands along with all the parameters in a single bash script often becomes hard to read, understand, and modify. Third, bash scripts do not provide an easy-to-use framework for building modular pipelines. Fourth reproducibility of methods is best-practice in computing, being required by more and more journals (Wittenburg, 2021), and scripting alone has significant limitations for reproducible bioinformatics (Leipzig, 2017). Fifth, scripts often contain significant "boilerplate code" as the user repeats commands and parameters in the script. This results in challenges, particularly for complex pipelines. Managing the tool's

parameters and making changes becomes difficult and error prone. Controlling parameters is essential for reproduciblity but are difficult to document, and hard to track if not well documented. Moreover, no straightforward framework exists to define tools and parameters and modify them at runtime.

Here we present Python RNA-Seq pipeliner, `pyrpipe`, a lightweight Python package for users to code and execute computational pipelines in an object-oriented manner, in pure Python. No new *workflow* syntax that is specific to `pyrpipe` is required. `pyrpipe` delivers an intuitive framework to easily *import* any Linux/macOS executable command or third-party tool as reusable Python objects.

Using the `pyrpipe` framework, users can implement RNA-Seq downloading and processing pipelines in a single go, quickly and intuitively. In addition, we have designed APIs to popular RNA-Seq tools and incorporated these into pyrpipe to enable coherent RNA-seq processing – from managing the raw data, to trimming, alignment, and assembly or quantification. pyrpipe's simple API design allows for automated access to publicly available NCBI-SRA RNA-Seq data (Sherry and Xiao, 2012) allowing users to quickly implement pipelines for harmonized re-analysis of these datasets (Figure 6.1).

`pyrpipe` will be helpful for users looking for a robust approach to write pipelines in pure Python. Compared to plain Bash, Perl, or Python scripting, `pyrpipe` provides many helpful features for building reproducible and easy-to-share pipelines. These features include: extensive logging and reports; loading tool options from YAML files to easily modify and document tool parameters; a dry-run mode to check dependencies and targets before implementing large-scale analysis; resuming of jobs if they are interrupted; and saving pyrpipe sessions.

`pyrpipe` can be used for quick prototyping of RNA-Seq processing pipelines because of the ease in swapping out pyrpipe objects, such as substituting Stringtie (Pertea et al., 2015) for Cufflinks (Trapnell et al., 2010) for transcript assembly. `pyrpipe` pipelines can be easily scaled using a workflow manager, including the popular Snakemake (Köster and Rahmann, 2012), NextFlow (Di Tommaso et al., 2017) or Toil (Vivian et al., 2017). The workflow management

system then can scale and manage jobs on clusters and schedule independent jobs for parallel processing, facilitating scalable pipelines and optimizing resource usage. Meanwhile, `pyrpipe`, whether used independently or as part of a workflow management system, facilitates ease-of-implementation, reproducibility, understandability, and modification of the RNA-Seq processing pipeline.

6.3 Materials and Methods

6.3.1 Overview

We developed `pyrpipe` to provide a light-weight Python framework for implementing bioinformatics or other computational analysis pipelines. The `pyrpipe` framework include: 1. high-level APIs to popular RNA-Seq tools; 2. a general API to import any executable command/tool into Python, enabling use of any bioinformatics tool; and 3. extensive monitoring and logging details of the commands that are executed. Thus, `pyrpipe` allows users to *import* any Linux/macOS executable command/tool into the Python ecosystem and implement pipelines in pure-Python incorporating their own Python code, existing Python libraries and third-party programs. To execute the commands, `pyrpipe` uses Python's *subprocess* library but adds many useful features and options. The commands executed via `pyrpipe` are automatically logged, monitored, and can be flexibly controlled using `pyrpipe` options. `pyrpipe` is packaged as a Python library and can be installed via PyPI or conda. An advantage of using the Python platform is that it is widely used, free, flexible, object-oriented, has high-level data structures (Suarez et al., 2018; Mariano et al., 2020; Kossaifi et al., 2019; Kanterakis et al., 2019), and a growing repository of $> 200,000$ packages and tools.

6.3.2 The `pyrpipe` framework

`pyrpipe` enables users to code pipelines in an object-oriented manner, using specialized API "classes" provided by `pyrpipe`. Each class in `pyrpipe` is designed to work with a particular processing tool, for example, the *Star* class implements the necessary functionality to use the

STAR tool (Dobin et al., 2013) for RNA-Seq alignment via pyrpipe. Users can create specific *objects* of these classes and use the *objects* in their Python scripts (Figure 6.1). Each RNA-Seq processing tool is fully accessible via these *objects* and the user is not required to remember the full usage syntax of that tool, hence promoting *abstraction*. Instead, the data and parameters required by these tools are *encapsulated* within the respective *objects*. For example, when creating a *Star* object, its index and other parameters are saved with the object (Figure 6.1). `pyrpipe` provides flexible parameter management (Figure 6.1). If no parameters are provided by the user, the tool is executed with its default parameters. We recommend that users fully understand and use the best parameters for their pipelines.

Tools performing similar types of RNA-Seq processing steps are grouped together in a single `pyrpipe` module, and are designed to have identical APIs. This enables their *objects* to be easily interchangeable in pipelines, promoting reusability and modification. For example, the classes "Star" and "Hisat2", both in the `pyrpipe` "mapping" module, implement the *build_index* and *perform_alignment* functions. Thus, changing a *Star* object with a *Hisat2* object is straightforward. See Supplementary Data for implementation details.

6.3.3 APIs for RNA-Seq processing

`pyrpipe` provides high-level APIs, to access full functionality of 11 popular RNA-Seq analysis tools that expedite and enhance implementation of RNA-Seq pipelines that can be readily shared, modified, or reused, including a dedicated module to facilitate access and management of the extensive RNA-Seq data available from the National Center for Biotechnology Information Research Sequence Read Archives (NCBI-SRA) database (Kodama et al., 2011).

These API classes are implemented inside several highly cohesive modules: (*sra*, *mapping*, *alignment*, *quant*, *qc*, *tools*). Each module has been designed to capture steps integral to RNA-Seq analysis: 1) access NCBI-SRA and manage raw RNA-Seq data; 2) quality control; 3) read alignment; 4) transcript assembly; and 5) transcript quantification. (Supplementary Table 1 and Supplementary Figures 1, 2). We have built and integrated these APIs into the `pyrpipe` package,

such that any RNA-Seq processing pipeline can be intuitively executed by the researcher while writing minimal code (Figure 6.1).

By default, all output files are consistently named and managed by `pyrpipe`, and put in the same directory as the RNA-Seq data files. Users can provide a different output directory.

6.3.4 Flexibility in pipeline execution, debugging, and pipeline sharing

`pyrpipe` flexibility extends to enabling the user to choose how to execute and handle exceptions and errors to modify their pipeline's behavior.

Users can create checkpoints in the pipeline, save the current `pyrpipe` *session*, and resume later. This is particularly useful for running different blocks of a workflow in different environments that can optimize resource usage. For example, on a typical high performance computing (HPC) cluster, a researcher might use a dedicated data-transfer node to retrieve data from SRA and then use compute nodes for data processing.

`pyrpipe` allow users to *dry run* the pipeline, during which commands are printed to screen, but not executed; thus, any potential error in the pipeline can be detected and fixed before using it to process large amounts of data. In addition, `pyrpipe` can skip execution of commands for which the output files are already present, saving computer time. Users can deploy the *–force* option to re-execute these commands (See Supplementary Data).

`pyrpipe`'s logging features enable efficient error detection and reports (Fig. 6.1). Errors and extensive environment information, such as operating system and Python version, along with version and path information for each program used within the pipeline, are all logged. `pyrpipe` logs are saved in JavaScript Object Notation (JSON) format for parsing by `pyrpipe` and other software (Supplementary Table 2).

The *pyrpipe_diagnostic* command can be invoked to generate comprehensive reports about the analysis, benchmark comparisons (Supplementary Figure 3), shell scripts and MultiQC reports (Ewels et al., 2016). These reports, along with the Python scripts, can be shared or included with publications to ensure reproducibility.

The default pyrpipe behaviour for logging, dry-run, and reports, can be modified by supplying pyrpipe with specific options via command-line or by specifying these in a *pyrpipe_conf.yaml* file.

6.3.5 Reproducible analysis

Reproducibility can be a major challenge in bioinformatics studies because of heavy computational intensive tasks that depend on a number of software and system libraries. Reproducibility can be ensured by controlling execution environments via environment managers such as Anaconda, container systems such a Docker, or isolated virtual machines (Grüning et al., 2018a).

`pyrpipe` is a Python package available through bioconda (Grüning et al., 2018b) and can be installed and managed within conda environments, containers or VMs. We have included in pyrpipe documentation the recommended way of installing the required tools, with version information (Supplementary Table 1), for RNA-Seq analysis via bioconda (Grüning et al., 2018b).

Besides the user controlling the execution environment, `pyrpipe` adds several layers to enhance reproducibility of analysis. `pyrpipe` creates a local copy of the pipeline script so that user has access to the exact pipeline code later. `pyrpipe` logs the MD5 checksums of the pipeline script and any input files provided as arguments. Thus, the user can verify which scripts and input files were used in the analysis. We recommend users to use a version control software such as Git to keep a track of the changes to the scripts.

`pyrpipe` allows and encourages users to define separate YAML files for the tool parameters. This enables the user to modify, manage, share and reproduce computational analysis on different data and platforms. Further, `pyrpipe` logs contain detailed information about all the tools/commands used and their versions, which can be utilized to re-build the environments.

6.4 Results

We evaluated `pyrpipe` by three case studies, each illustrating a different aspect of what the tool can accomplish and how new functionality can be added.

6.4.1 Case Study 1: Scaling up `pyrpipe` to process 17,328 RNA-Seq samples from non-diseased human tissues

This case study demonstrates the ability of `pyrpipe` to process large amounts of data – 17,328 human RNA-Seq samples from the Genotype-Tissue Expression (GTEx V8) (Consortium et al., 2017). We developed and implemented our pipeline to identify expressed human orphan genes, as well as annotated genes, in diverse tissues using `pyrpipe`. This pipeline cohesively automated the steps of RNA-Seq processing into a single. It : 1) downloaded data from 17,328 raw GTEx RNA-Seq samples via AnVil (anvilproject.org); 2) aligned the reads of each sample to the human reference genome using STAR (Dobin et al., 2013); 3) assembled transcripts using Stringtie (Pertea et al., 2015); 4) merged transcriptomes from individual samples into a consistent assembly using orfipy (Singh and Wurtele, 2021), Mikado (Venturini et al., 2018), and Taco (Niknafs et al., 2017), and 5) quantified the annotated and unannotated transcripts using Salmon (Patro et al., 2017) (See Supplementary Figure 4 for details). This pipeline was run on the PSC Bridges HPC system (`https://www.psc.edu/resources/bridges/`). The pipeline was scaled to run multiple batches of RNA-Seq samples in parallel on multiple nodes. Code and data for this project is available `https://github.com/urmi-21/pyrpipe/tree/master/case_studies/GTEx_processing`.

To assess the results of our pipeline, we have compared the expression of annotated genes identified by the `pyrpipe` pipeline with those reported in GTEx (a pipeline that only quantifies the annotated genes). This comparison showed good accordance between expression values from the two pipelines. We compare the median TPMs of annotated genes for two types of adipose tissue, as processed by `pyrpipe` and by the GTEx portal (Figure 6.2).

6.4.2 Case Study 2: Integrating `pyrpipe` within a workflow manager to quantify gene expression in COVID-19 samples for exploratory analysis

We implemented `pyrpipe` within two workflow management systems, Snakemake (Köster and Rahmann, 2012) and NextFlow (Di Tommaso et al., 2017), selecting these specifically because

they are widely used by the bioinformatics community (Jackson et al., 2021). Snakemake and NextFlow were independently used to implement, manage and execute the pipeline for multiple RNA-Seq samples in parallel on a single cluster.

We used this pipeline to quantify RNA-Seq data from a COVID-19 study of circulating monocytes (Rother et al., 2020), and provide output that can be directly analyzed by biologists, using the versatile Java software for exploratory analysis of large datasets, MetaOmGraph (MOG) (Singh et al., 2020).

Specifically, we used `pyrpipe` to seamlessly download 29 RNA-Seq samples from NCBI-SRA (accession SRP287810) and quantify expression of annotated transcripts using Salmon's selective alignment approach (Patro et al., 2017; Srivastava et al., 2020). This study analyzed RNA-Seq data from circulating monocytes derived from individuals with COVID-19 and healthy individuals, treated and untreated with hydroxychloroquine. The final transcript and gene level TPMs from each sample are merged into a single file to create a MetaOmGraph (Singh et al., 2020) project (*MOGproject-monocytes-60241genes-29samples-HCQtreat-2021-1-17*) for exploratory data analysis.

Using MetaOmGraph for rapid exploration of the data, we identified genes that show differential expression patterns between COVID-19-diseased individuals (n=20) vs healthy individuals (n=9).

Nine of the fourteen genes most highly overexpressed in monocytes from healthy individuals are involved in the biological process, neutrophil chemotaxis (GO:0030593; Bonferoni corrected P-value of 5.229E-10); these include six CCL- and CXL-type chemokines. Interestingly, in lung tissues, CCL2 and other chemokine expression are decreased by ACE2, but up-regulated during a COVID-19-induced cytokine storm (Merad and Martin, 2020). Of the 14 genes highly expressed in monocytes from COVID-19-diseased individuals but not in healthy indivduals (Figure 3C), 12 participate in immune effector (GO:0002252; Bonferoni corrected P-value of 3.161E-12), including nine defensins or immunoglobins. These biological processes are also associated strongly with

COVID-19 in neutrophils (Aschenbrenner et al., 2021). Functional designations were obtained using ToppGene (toppgene.cchmc.org) (Supplementary File 1).

The code, data, and MetaOmGraph project are available at https://github.com/urmi-21/pyrpipe/tree/master/case_studies/Covid_RNA-Seq.

6.4.3 Case Study 3: Use of `pyrpipe` for *de novo* transcriptome assembly

We used a new, high-quality genome of *Zea mays* B73 cultivar (https://doi.org/10.1101/2021.01.14.426684) as reference genome, and gathered RNA-Seq data from ten diverse samples (B73 cultivar), representing different tissue and development stages, for *de novo* transcriptome assembly (Supplementary Figure 5). Our pipeline identified a total of 57,916 distinct transcripts. Of these, 38,881 transcripts were homologous to UniProt proteins (Consortium, 2019; Altschul et al., 1990). These transcripts could be non-coding RNAs (ncRNAs), low-level "noise" (Pertea et al., 2018), or pseudogenes; others might represent as yet unannotated genes encoding conserved proteins. The remaining 6,306 transcripts, with no similarity to any protein in the database, could be ncRNAs, "noise"; others are likely to be as yet unannotated species-specific ("orphan") genes (Singh and Wurtele, 2020). The transcript length and GC content distribution for transcripts with conserved CDS and transcripts are shown in Supplementary Figure 6. The mean length of non-homologous transcripts (1,290 nt) is shorter than conserved transcripts (1,981 nt); mean GC content is indistinguishable (50.8% vs 50.6%). The median expression of non-homologous transcripts across the 10 RNA-Seq samples analyzed is lower than the median expression of conserved transcripts; however, in each sample, hundreds of non-homologous transcripts are more highly expressed than the mean of the conserved genes. These characteristics follow the same trend as those of the conserved and orphan genes in the well-characterized *Arabidopsis thaliana* genome (Arendsee et al., 2014). Pipeline scripts, downstream analysis code and data are available at https://github.com/lijing28101/maize_pyrpipe.

Table 6.1 Comparison of `pyrpipe` features with Ruffus and Pypiper. *For parallel execution support, `pyrpipe` easily can be integrated with a workflow management system, e.g. Case Study 2.

Feature	`pyrpipe`	**Ruffus**	**Pypiper**
Latest version	0.0.5	2.8.4	0.12.1
Latest update	2021	2020	2019
API to RNA-Seq tools	Yes	No	No
Import tools as objects	Yes	No	No
Auto-load tool parameters	Yes	No	No
Dry run mode	Yes	No	No
Resume Interrupted	Yes	Yes	Yes
Exception handling	Yes	Yes	Yes
Parallel execution support*	No	Yes	No
Logs/Reports	Yes/Yes	Yes/No	Yes/Yes

6.4.4 Comparison of `pyrpipe` to existing Python libraries that can be used for RNA-Seq analysis

Several Python libraries enable workflows to be specified. However, they do not provide a dedicated API suite for RNA-Seq data analysis. Instead, these frameworks depend on the user to explicitly write the commands and provide data.

We compared `pyrpipe` with two such Python libraries that allow specifying bioinformatics pipeline - Ruffus (Goodstadt, 2010) and Pypiper (`http://code.databio.org/pypiper/`). Ruffus is a Python library for specifying and executing workflows. Ruffus allows users to specify pipeline tasks using several *"decorator"* functions. Pypiper is a Python package for coding pipelines in Python. It provides the "PipelineManager" class which a user can employ to execute commands in a serial manner. Pypiper has a built-in toolkit, NGSTk, to allow users to generate commonly used bioinformatics shell commands. These functions return commands as *string* objects that can be passed to "PipelineManager" for execution. Table 6.1 compares `pyrpipe` features with Ruffus and Pypiper.

6.5 Discussion

The `pyrpipe` package allows users to code and implement RNA-Seq workflows in an object-oriented manner, purely using Python. `pyrpipe` is intended for any user who analyzes RNA-Seq data – beginner or advanced. APIs to RNA-Seq tools make it straightforward to code RNA-Seq processing pipelines. Access to NCBI-SRA is automated, such that users can readily retrieve raw read RNA-Seq data. The downloaded raw RNA-Seq data and data files are automatically managed, and consistently accessed through *SRA* objects. Users need not keep track of data files or paths, as these are integrated with `pyrpipe` objects. `pyrpipe` enhances the re-usability of the code-blocks, cutting down development time for new pipelines from existing code base. It also improves re-usability of the workflows, because all the parameters that needs to be adjusted for new analyses could be read from YAML files. `pyrpipe` workflows can be modified using Python's control flow abilities and a user can create complex, reproducible, workflow structures. Any third party tool, executable command, or script can be integrated into `pyrpipe` for additional data processing capability. `pyrpipe` logs and reports enable debugging and reproducibility.

Analysis of the 17,328 GTEx RNA-Seq samples was easily scaled using `pyrpipe` alone, by creating smaller *batches* of samples and submitting the processing jobs in parallel, on an HPC system with a slurm job scheduler.

When building more complex and scalable workflows, it may be more efficient to integrate `pyrpipe` into a workflow management system. This can easily be done, as shown in our second case study. Workflow management systems are developed for robust implementation of computational pipelines; nevertheless, they differ significantly in terms of workflows, definitions, job scheduling, and features (Lampa et al., 2019; Köster and Rahmann, 2012; Vivian et al., 2017; Di Tommaso et al., 2017). For example, Snakemake uses a "pull-based" strategy to check for specific output files and schedule jobs accordingly (Köster and Rahmann, 2012; Lampa et al., 2019), whereas Nextflow uses a "push-based" scheme in which a "process" defined in the workflow pushes its outputs to downstream "processes" (Di Tommaso et al., 2017; Strozzi et al., 2019).

The SciPipe (Lampa et al., 2019) workflow library is written in the GO language; similar to Nextflow it implements dataflow based task scheduling. Toil (Vivian et al., 2017) provides explicit application programming interfaces (APIs) for defining static or dynamic tasks and supports common workflow language (CWL) and multiple cloud environments (Vivian et al., 2017; Leipzig, 2017). Hence, users need to make informed decisions if choosing a workflow management system (Jackson et al., 2021) for `pyrpipe`.

The modular design of `pyrpipe` permits users to write *pythonic* code, which is designed to read, manage, and share. Because of the rapid emergence of new bioinformatics tools, this design feature is particularly important. From a developer's perspective, `pyrpipe`'s modularity facilitates reuse and extensibility; new tools/APIs can be easily integrated into `pyrpipe` and promotes sustainability.

Compared to plain Bash or Python scripting, `pyrpipe` provides many handy features for writing reproducible, robust and flexible pipelines in pure Python. Being simple, powerful, and easy to learn, Python has become one of the most popular languages among biologists and bioinformaticians; furthermore, a wide variety of tools are available in Python (Suarez et al., 2018; Mariano et al., 2020). Keeping this in mind, we designed a general framework such that `pyrpipe` is fully extendable to include any third-party tool, while writing minimal code. This design lets Python users integrate their own customized APIs into the `pyrpipe` ecosystem and thus incorporate diverse functionality into their pipelines. We have provided an example in the documentation (`https://pyrpipe.readthedocs.io/en/latest/tutorial/api.html`).

`pyrpipe` will appeal to users who are looking for a simple way to deploy small or large scale RNA-Seq processing pipelines, and to make these pipelines accessible to the community. `pyrpipe` supports the model of reproducible open science. Straightforward and seamless integration, execution, and sharing of RNA-Seq workflows make it an ideal choice for users with less computational expertise, as well as seasoned bioinformaticians. Writing Python code using `pyrpipe` is intuitive and maintainable. Leveraging Python language's flow control and exception-handling abilities, users can quickly create complex and dynamic pipelines. Moreover,

downstream analysis and data manipulation steps can be directly integrated into `pyrpipe` pipelines via Python.

6.6 Data Availability

We subscribe to FAIR data and software practices (Wilkinson et al., 2016). `pyrpipe` source code is available at https://github.com/urmi-21/pyrpipe. `pyrpipe` source code (v0.0.5) can be accessed via DOI: 10.5281/zenodo.4448373. The `pyrpipe` package can be installed from the source, from PyPi (https://pypi.org/project/pyrpipe) or from bioconda (https://anaconda.org/bioconda/pyrpipe). Extensive documentation to guide users on how to use `pyrpipe` and the APIs implemented within it is available on Read the Docs (http://pyrpipe.rtfd.io). We encourage contributions from the bioinformatics community (Contribution guide along with a Code of Conduct to guide new contributors is available at https://github.com/urmi-21/pyrpipe We hope to see pyrpipe evolve as a community driven project.

6.7 Supplementary Data

Supplementary Data are available at https://github.com/urmi-21/pyrpipe and at NARGAB Online.

6.8 Funding

This work is funded in part by National Science Foundation grant IOS 1546858, Orphan Genes: An Untapped Genetic Reservoir of Novel Traits, and by the Center for Metabolic Biology, Iowa State University. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562, in particular the Bridges HPC environment through allocations TG-MCB190098 and TG-MCB200123.

6.9 References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410.

Arendsee, Z. W., Li, L., and Wurtele, E. S. (2014). Coming of age: orphan genes in plants. *Trends in plant science*, 19(11):698–708.

Aschenbrenner, A. C., Mouktaroudi, M., Kraemer, B., Oestreich, M., Antonakos, N., Nuesch-Germano, M., Gkizeli, K., Bonaguro, L., Reusch, N., Baßler, K., et al. (2021). Disease severity-specific neutrophil signatures in blood transcriptomes stratify COVID-19 patients. *Genome Medicine*, 13(1):1–25.

Consortium, G. et al. (2017). Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213.

Consortium, U. (2019). Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, 47(D1):D506–D515.

Daub, C. O., Steuer, R., Selbig, J., and Kloska, S. (2004). Estimating mutual information using b-spline functions–an improved similarity measure for analysing gene expression data. *BMC bioinformatics*, 5(1):118.

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature biotechnology*, 35(4):316.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21.

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). Multiqc: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048.

Goodstadt, L. (2010). Ruffus: a lightweight python library for computational pipelines. *Bioinformatics*, 26(21):2778–2779.

Grüning, B., Chilton, J., Köster, J., Dale, R., Soranzo, N., van den Beek, M., Goecks, J., Backofen, R., Nekrutenko, A., and Taylor, J. (2018a). Practical computational reproducibility in the life sciences. *Cell systems*, 6(6):631–635.

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., and Köster, J. (2018b). Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature methods*, 15(7):475–476.

Jackson, M., Kavoussanakis, K., and Wallace, E. W. (2021). Using prototyping to choose a bioinformatics workflow management system. *PLoS Computational Biology*, 17(2):e1008622.

Kanterakis, A., Iatraki, G., Pityanou, K., Koumakis, L., Kanakaris, N., Karacapilidis, N., and Potamias, G. (2019). Towards reproducible bioinformatics: the openbio-c scientific workflow environment. In *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 221–226. IEEE Computer Society.

Kodama, Y., Shumway, M., and Leinonen, R. (2011). The sequence read archive: explosive growth of sequencing data. *Nucleic acids research*, 40(D1):D54–D56.

Kossaifi, J., Panagakis, Y., Anandkumar, A., and Pantic, M. (2019). Tensorly: Tensor learning in python. *The Journal of Machine Learning Research*, 20(1):925–930.

Köster, J. and Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522.

Lampa, S., Dahlö, M., Alvarsson, J., and Spjuth, O. (2019). Scipipe: A workflow library for agile development of complex and dynamic bioinformatics pipelines. *GigaScience*, 8(5):giz044.

Leipzig, J. (2017). A review of bioinformatic pipeline frameworks. *Briefings in bioinformatics*, 18(3):530–536.

Mariano, D., Ferreira, M., Sousa, B. L., Santos, L. H., and de Melo-Minardi, R. C. (2020). A brief history of bioinformatics told by data visualization. In *Brazilian Symposium on Bioinformatics*, pages 235–246. Springer.

Merad, M. and Martin, J. C. (2020). Pathological inflammation in patients with COVID-19: a key role for monocytes and macrophages. *Nature Reviews Immunology*, 20(6):355–362.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621–628.

Niknafs, Y. S., Pandian, B., Iyer, H. K., Chinnaiyan, A. M., and Iyer, M. K. (2017). Taco produces robust multisample transcriptome assemblies from rna-seq. *Nature methods*, 14(1):68–70.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, 14(4):417.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L. (2015). Stringtie enables improved reconstruction of a transcriptome from rna-seq reads. *Nature biotechnology*, 33(3):290.

Pertea, M., Shumate, A., Pertea, G., Varabyou, A., Breitwieser, F. P., Chang, Y.-C., Madugundu, A. K., Pandey, A., and Salzberg, S. L. (2018). CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome biology*, 19(1):208.

Rother, N., Yanginlar, C., Lindeboom, R. G., Bekkering, S., van Leent, M. M., Buijsers, B., Jonkman, I., de Graaf, M., Baltissen, M., Lamers, L. A., et al. (2020). Hydroxychloroquine inhibits the trained innate immune response to interferons. *Cell Reports Medicine*, page 100146.

Sherry, S. and Xiao, C. (2012). Ncbi sra toolkit technology for next generation sequence data. In *Plant and Animal Genome XX Conference (January 14-18, 2012). Plant and Animal Genome*.

Singh, U., Hur, M., Dorman, K. S., and Wurtele, E. S. (2020). Metaomgraph: a workbench for interactive exploratory data analysis of large expression datasets. *Nucleic Acids Research*, 48(4):e23–e23. gkz1209.

Singh, U. and Wurtele, E. S. (2020). Genetic novelty: How new genes are born. *Elife*, 9:e55136.

Singh, U. and Wurtele, E. S. (2021). orfipy: a fast and flexible tool for extracting ORFs. *Bioinformatics*. btab090.

Srivastava, A., Malik, L., Sarkar, H., Zakeri, M., Almodaresi, F., Soneson, C., Love, M. I., Kingsford, C., and Patro, R. (2020). Alignment and mapping methodology influence transcript abundance estimation. *Genome biology*, 21(1):1–29.

Stark, R., Grzelak, M., and Hadfield, J. (2019). Rna sequencing: the teenage years. *Nature Reviews Genetics*, 20(11):631–656.

Strozzi, F., Janssen, R., Wurmus, R., Crusoe, M. R., Githinji, G., Di Tommaso, P., Belhachemi, D., Möller, S., Smant, G., de Ligt, J., et al. (2019). Scalable workflows and reproducible data analysis for genomics. In *Evolutionary Genomics*, pages 723–745. Springer.

Suarez, C. G. H., Burbano, M. E. G., Guerrero, V. A. B., and Tovar, P. A. M. (2018). Bioinformatics software for genomic: a systematic review on github.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511.

Venturini, L., Caim, S., Kaithakottil, G. G., Mapleson, D. L., and Swarbreck, D. (2018). Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *GigaScience*, 7(8):giy093.

Vivian, J., Rao, A. A., Nothaft, F. A., Ketchum, C., Armstrong, J., Novak, A., Pfeil, J., Narkizian, J., Deran, A. D., Musselman-Brown, A., et al. (2017). Toil enables reproducible, open source, big biomedical data analyses. *Nature biotechnology*, 35(4):314.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

Wittenburg, P. (2021). Open Science and Data Science. *Data Intelligence*, 3(1):95–105.

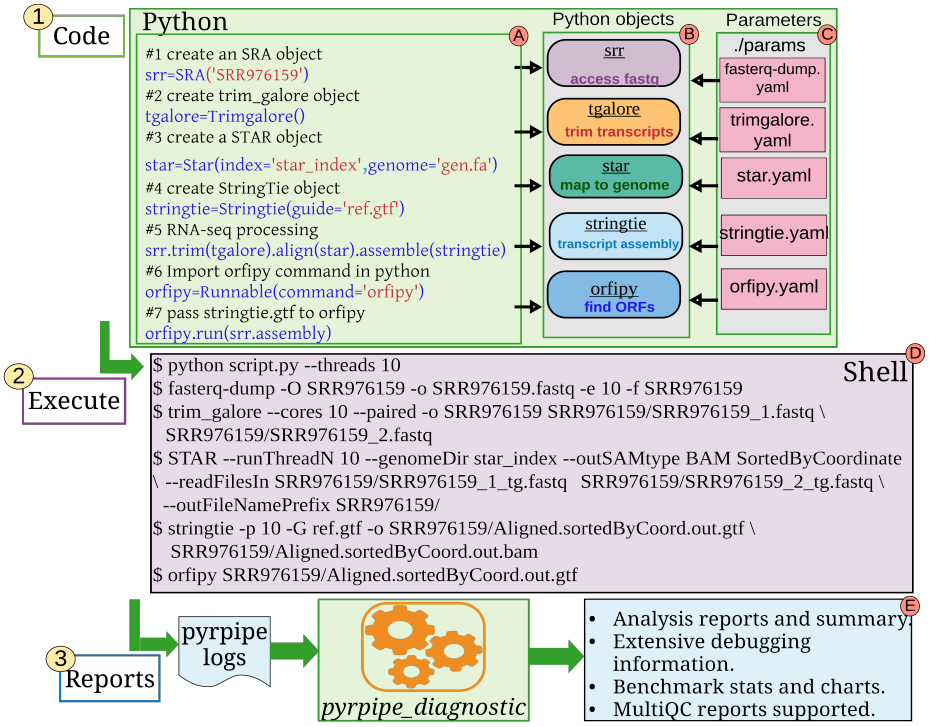

Figure 6.1 The `pyrpipe` framework. This simple example illustrates the relationship between the Python code that the user writes for `pyrpipe`, the corresponding Python objects, the YAML parameter files, the corresponding shell script, and the output. The user need only define the NCBI-SRA Run Accessions and the tools to be used, the rest is automatic. A key advantage of `pyrpipe` is that it can be used to easily create complex workflows that are intuitive, understandable, reproducible, and modifiable. `pyrpipe` can automatically load and resolve tool parameters from YAML files. `pyrpipe` is represented by the green boxes. The user writes the code in Python (blue text), creating Python *objects* of specific `pyrpipe` classes that provide APIs to RNA-Seq tools. To execute the full pipeline, the user need to run only the Python file, e.g. "python script.py –threads 10", to designate executing the pipeline using 10 threads (Box A). Each *object* encapsulates specific *methods* and *data* (Box B). For example, each *SRA* object stores the directory path for the associated raw RNA-Seq data that is used as the default directory by `pyrpipe` to output files from different RNA-Seq processing steps, i.e., trimming, alignment, assembly or quantification. Tool parameters, if supplied in YAML files, are automatically loaded and stored in the corresponding `pyrpipe` object (Box C). During processing, shell commands are automatically constructed and executed by the `pyrpipe` APIs; (Box D). After execution, the *pyrpipe_diagnostic* tool generates extensive data analyses and diagnostic reports from the logs. These enable users to summarize, share, benchmark or debug their pipelines (Box E).

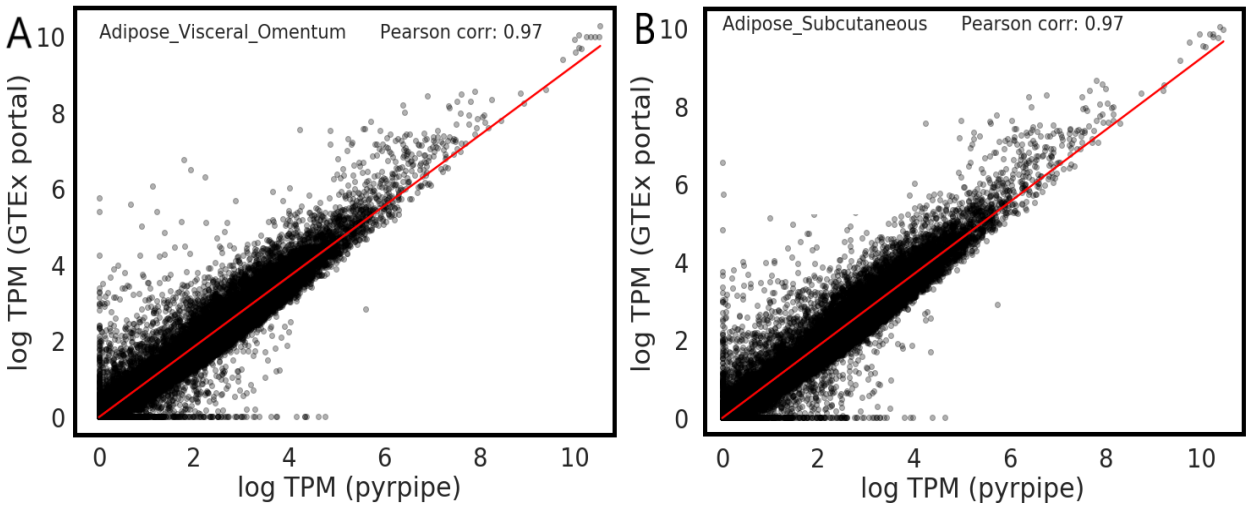

Figure 6.2 Comparison of median TPMs for two tissue types **A.** Visceral Adipose and **B.** Subcutaneous Adipose. Y-axis shows the logged Median TPMs computed by the GTEx portal pipeline. X-axis shows the logged Median TPMs computed by our pipeline implemented with `pyrpipe`. Pearson correlations are 0.97%. Differences in quantification of several 100 genes are likely due differences in reference annotations. Code and data to reproduce this plot and, to compare other tissue types are available at `github.com/urmi-21/pyrpipe/tree/master/case_studies/GTEx_processing`.

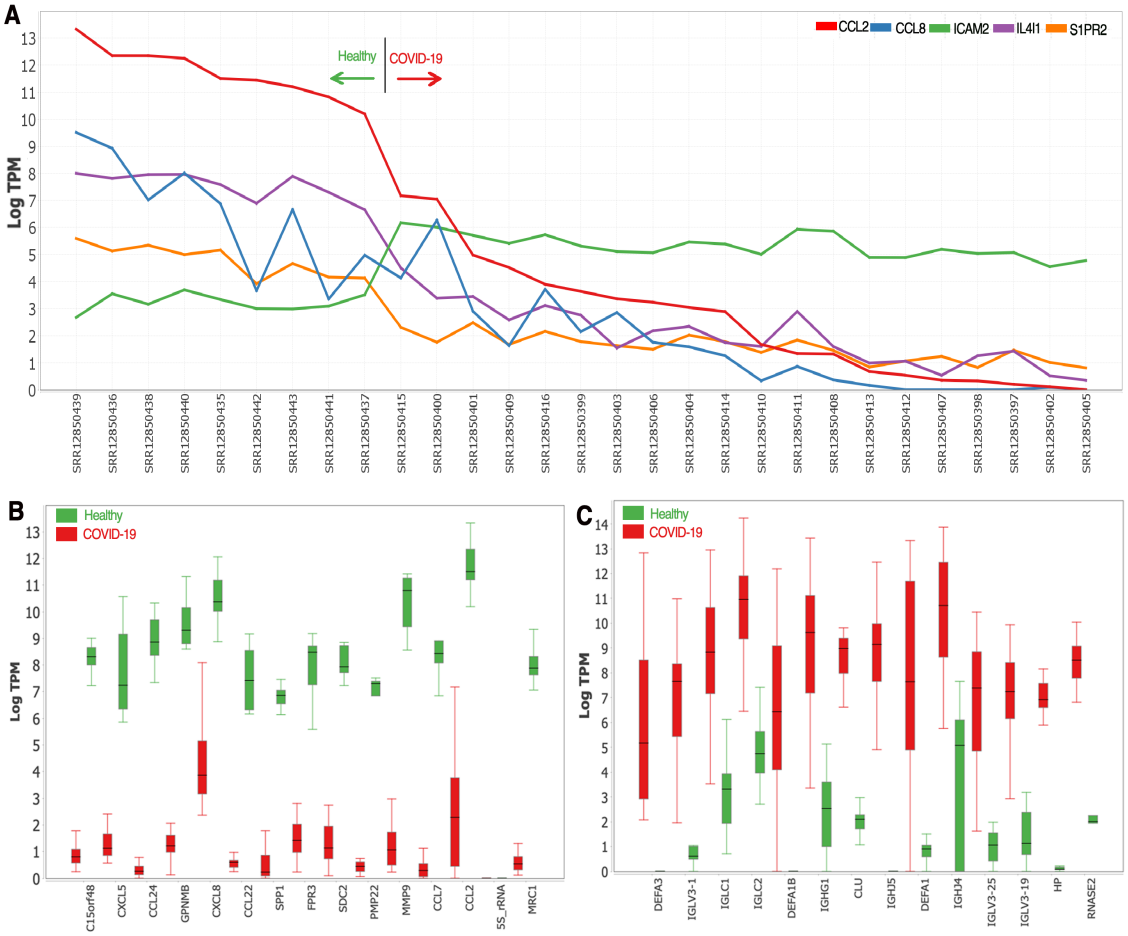

Figure 6.3 Exploratory analyses and visualization using MetaOmGraph (Singh et al., 2020) of RNA-Seq data dervived from monocytes of COVID-19 diseased and healthy individuals. The data were downloaded from NCBI-SRA and processed using `pyrpipe` integrated into the workflow manager, Snakemake. Raw reads from 29 samples of monocytes derived from individuals with COVID-19 (n=20) and healthy (n=9) individuals (SRP287810) were downloaded and processed. Over 60,000 genes are represented in each sample of processed data. For quick preliminary exploration of the data, using MetaOmGraph, we identified genes with more than two-fold change in TPM values with Benjamini-Hochberg adjusted p-value < 0.002 for the non-parametric Mann-Whitney test. **A.** Line chart showing expression pattern of genes non-linearly associated with CCL2 (estimated via Mutual Information (Daub et al., 2004)) in COVID-19 and healthy individuals. **B.** Fourteen genes with highest fold change in healthy v.s COVID-19 diseased individuals. **C.** Fourteen genes with highest fold change in COVID-19 diseased vs healthy individuals.

CHAPTER 7. A PAN-TISSUE PAN-CANCER COMPENDIUM OF HUMAN ORPHAN GENES

Urminder Singh [1,2,3], and Eve Syrkin Wurtele [1,2,3]

[1]Bioinformatics and Computational Biology Program, Iowa State University, Ames, IA 50011.

[2]Center for Metabolic Biology, Iowa State University, Ames, IA 50011.

[3]Department of Genetics Development and Cell Biology, Iowa State University, Ames, IA 50011.

Modified from a manuscript to be published in *Nature Communications*

7.1 Abstract

Orphan genes, defined as genes encoding species-specific proteins, are ubiquitous in evolution. Thousands have been characterized as functionally important. However, their annotation is difficult as their proteins share no homology to proteins in other species. Thus, orphan genes may be under-identified, even in humans. Here, we take a data driven approach to identify novel transcripts in the human genome, leveraging terabytes of RNA-Seq data. We report thousands of novel highly-expressed transcripts that are dynamically and selectively accumulated across tissues and tumors. A majority are inferred as orphan genes. Many are differentially expressed across tumors, genders, and ancestries. These genes may provide promising candidates for novel diagnostics. We validated expression of many using strand-specific and single-cell RNA-seq datasets. Hundreds of these novel genes show evidence of translation. Finally, we report that hundreds of novel orphan genes overlap with deleterious genomic variants and thousands show significant association with overall patient survival. Some of these may play an important role in human disease and physiology and provide targets for therapeutic interventions.

7.2 Introduction

Newly emerged and species-specific, orphan genes provide an organism with a reservoir of genetic elements for evolutionary innovation (Tautz and Domazet-Lošo, 2011; Arendsee et al., 2014; Reinhardt et al.; Van Oss and Carvunis, 2019; Li et al., 2021). Orphan genes code for proteins that bear no detectable homology to proteins of other species. These genes arise continuously in genomes and provide a platform of evolutionary innovation (Arendsee et al., 2014; Van Oss and Carvunis, 2019). Over time, some orphan genes may meet a persistent or intermittent environmental challenge, and thus are selected for and become optimized. These orphan genes will acquire biologically functional roles and continue to evolve under selection (Ruiz-Orera et al., 2015; Arendsee et al., 2014; O'Conner and Li, 2020). However, many of the orphan genes will no longer be useful and consequently, are lost as genes during the course of evolution.

Many theories have been proposed for the evolutionary origins of protein-coding orphan genes: *de novo* from non-genic regions of the genome, introns, novel reading frames within existing genes, or lncRNAs (Van Oss and Carvunis, 2019; Neme and Tautz, 2013; Vakirlis and McLysaght, 2019; Reinhardt et al.; Schmidt et al., 2018; Ruiz-Orera et al., 2015; Vakirlis et al., 2020); rapid divergence of the CDS of an existing gene (Tautz and Domazet-Lošo, 2011; Schmitz et al., 2018, 2020); potentially a hybrid of both mechanisms. A recent paper proposes mitochondrial fostering as a mechanism for orphan gene birth (O'Conner and Li, 2020).

Thousands of orphan genes from diverse species have been functionally-studied (Arendsee et al., 2014; Van Oss and Carvunis, 2019; Li and Wurtele, 2015; McLysaght and Hurst; Singh and Wurtele, 2020); these represent only a tiny fraction of the estimated billions of extant orphan genes (Li et al., 2021).

Orphan genes encode the secreted, paralyzing toxins, including those of tens thousands of species of jellyfish (Arendsee et al., 2014; Ovchinnikova et al.) and parasitic wasps (Dennis et al., 2020); "antifreeze" protein, which have evolved independently in thousands of eukaryotic and prokaryotic species and enable survival in frigid temperatures (Baalsrud et al., 2018); proteins

that remodel internal metabolic and regulatory networks, for example, to mitigate pest and pathogen stress responses (Qi et al., 2019); and proteins needed for gametogenesis (Gubala et al., 2017; Lange et al., 2021).

Because proteins encoded by orphan genes share no homology to other proteins, they are often discarded by conventional gene annotation pipelines (Li et al., 2021). Annotating young orphan genes is challenging even in well-assembled genomes because existing gene annotation pipelines are biased towards canonical gene features and ubiquitous expression (Li et al., 2021; Ruiz-Orera and Albà, 2018).

Young genes have different features compared to *ancient* genes. Proteins of annotated young genes tend to be shorter, a high proportion are monoexonic, the mean high isoelectric point (Pi) is higher, they rarely if ever encode enzymes, and in some species they are GC-rich (Arendsee et al., 2014; Van Oss and Carvunis, 2019; Ruiz-Orera et al., 2015). Thus, models based on *ab initio* prediction methods may not to be suitable to identify younger genes and methods based on experimentally-based direct inference using high-throughput data such as RNA-Seq, Ribo-Seq and proteomics have been used to identify these novel genes (Li et al., 2021; Ruiz-Orera et al., 2015; Ruiz-Orera and Albà, 2018; Blevins et al., 2021; Li et al., 2020; Klasberg et al., 2016). Recently the utility of RNA-Seq data and *ab initio* approaches has been empirically evaluated (Scalzitti et al., 2020; Li et al., 2021).

Using RNA-Seq data to directly identify novel genes is not limited to orphan genes, but extends to identifying yet unannotated ancient protein-coding genes, small ORFs, lncRNAs, alternative isoforms or fusion of annotated genes (Ruiz-Orera and Albà, 2018; Couso and Patraquim, 2017; Li et al., 2021; Pertea et al., 2018; Ruiz-Orera et al., 2014; Blevins et al., 2021; Newtson et al., 2021; Blume et al., 2021; Hon et al., 2017).

One reason a number of transcripts may remain unannotated in genomes is because of their selective expression over tissues under different environments (Ruiz-Orera et al., 2015; Arendsee et al., 2014; Li et al., 2021; Singh and Wurtele, 2020). This is particularly true for young orphan genes. Thus experimental evidence from diverse set of tissues and conditions needs to be

integrated to identify the set of these selectively expressed unannotated genes (Ruiz-Orera et al., 2015; Arendsee et al., 2014; Li et al., 2021).

Further, novel transcripts reported in various studies often do not get incorporated into the community data annotations, (Li et al., 2021; Erady et al., 2021; Martinez et al., 2020; Pertea et al., 2018), and thus are not easily accessible for further study. Because of this, many orphan and other selectively expressed novel genes remain ignored. However, it is important to identify and study the expression patterns of these novel genes.

These novel transcripts can have important functional or prognostic roles in context of human physiology and diseases (Pinskaya et al., 2019; Erady et al., 2021; Pertea et al., 2018; Van Oss and Carvunis, 2019; McLysaght and Hurst; Hon et al., 2017; Lorenzi et al., 2021). Several human-specific orphan genes are known to be highly regulated in multiple cancers and other diseases (McLysaght and Hurst; Samusik et al., 2013; Van Oss and Carvunis, 2019; Suenaga et al., 2014; Ruiz-Orera et al., 2015).

In this study we utilized 27,000 RNA-Seq datasets of non-diseased and cancer tissues from Genotype-Tissue Expression portal (GTEx) (Lonsdale et al., 2013) and The Cancer Genome Atlas (TCGA; `https://www.cancer.gov/tcga`) to extract and characterize a comprehensive compendium of highly-expressed novel transcripts. We developed workflows for uniform and harmonized read alignment, transcript assembly and quantification of these diverse RNA-Seq datasets that reduces study-specific biases in the data (Wang et al., 2018; Lachmann et al., 2018).

We report thousands of highly expressed novel genes, the majority of which are orphans coding for a human specific protein. Thousands of these novel genes show differential expression across multiple tumors, gender and races. Hundreds of the novel genes are significantly associated with overall survival of patients with many tumor types. These novel genes harbour millions of variants, several of which are identified to be pathogenic suggesting their potential role in human diseases. We catalog this comprehensive set of novel genes along with their characteristics, dynamic expression patterns, and translation evidence.

7.3 Results

7.3.1 Identification of highly expressed novel transcripts

We evaluated a total of 26,985 RNA-Seq paired-end samples: 17,283 samples from GTEx comprising 32 tissue types, and 9,702 samples from TCGA comprising 33 tumor types (Supplementary File 1).

Our RNA-Seq processing and meta-assembly pipeline identified a total of 1,384,542 different transcripts with a minimum length of 100 nucleotides (Figure 7.15; Supplementary Figure A1). We filtered out transcripts overlapping with the annotated transcripts and alternatively-spliced transcripts without an ORF, retaining 1,049,303 intronic, intergenic and alternatively-spliced transcripts that potentially code for human specific protein. We merged the transcripts identified with RNA-Seq evidence with the transcripts that are annotated in Gencode, to create a transcriptome assembly of 1,276,527 transcripts from 1,105,283 genes. These include transcripts of a wide range of expression levels, some of which could be the results of non-specific transcription ("noise" (Pertea et al., 2018)).

To retain only those unannotated transcripts with expression levels at least that of an average protein-coding gene, we quantified (Patro et al., 2017; Srivastava et al., 2020) the expression of each of these 1,276,527 transcripts across all 26,985 RNA-Seq samples. We then removed all transcripts with median expression levels lower than the median of all the annotated protein-coding transcripts; 54,794 novel transcripts from 54,748 genes remained (Supplementary Figure A1). We designate these high-expression unannotated transcripts as evidence-based (EB) RNAs. 52,032 contained an Open Reading Frame (ORF) greater than 99 nucleotides (evidence-based protein-coding RNAs, pcEBs)). The other 2,762 were considered as evidence-based non-coding RNAs (ncEBs). In our analyses, we focus on these 54,748 EB annotations, along with all Gencode-annotated coding and non-coding genes.

7.3.2 Thousands of novel EB transcripts are expressed dynamically

We identified the EB transcripts for each tissue type in a tissue-, gender-, and race-specific manner, and merged them together for each tissue. Almost every tissue type highly expresses thousands of novel transcripts. The median expression of all EB transcripts are comparable to that of lncRNAs but lower than annotated protein-coding transcripts (Supplementary File 1).

Among the non-diseased (GTEx) samples, the brain tissue expressed the greatest number of EB transcripts (8,542), while muscle expresses the least (1,234) (Figure 7.1). Testis and brain express high numbers of novel transcripts, as has been noted in earlier studies (Van Oss and Carvunis, 2019; Ruiz-Orera et al., 2015; Pertea et al., 2018; Xie et al., 2012). Brain uniquely expresses 4,037 EB transcripts. A total of 28,189 EB transcripts are highly-expressed in at least one GTEx tissue. Fifty-nine EB transcripts are highly-expressed in each of the 39 GTEx tissues.

TCGA samples taken from bone marrow of individuals with (Acute Myeloid Leukemia (AML) expressed the highest number of EB transcripts (29,631). This number is 18-fold higher than of the next highest TCGA tissue, testis, which expresses (1,616) EB transcripts. Samples from ovary (ovarian serous cystadenocarcinoma) express only 418 EB transcripts (Figure 7.1). Five hundred and fifty five EB transcripts are *uniquely* expressed in Testicular Germ Cell Tumors (testis), and in no other TCGA tissue type; 48 of these EB transcripts are also highly expressed in non-diseased testis.

Fifty-nine novel transcripts are expressed in common in each of the TCGA tissues; 35,212 novel transcripts are expressed in at least one TCGA tissue. A total of 8,607 EB transcripts are expressed in at least one tissue in GTEx *and* in a cancer type in a corresponding TCGA tissue.

7.3.3 The majority of EB transcripts are intronic

Using Mikado compare utility we assessed the locations of all the EB genes with respect to the reference annotation. Out of the 1,384,542 transcripts identified by the meta-assembly pipeline, 709,365 transcripts are from intronic regions of annotated genes. The next highest number of transcripts are from intergenic regions (357,012). After filtering low and spuriously

expressed transcripts, of the total 54,794 expressed EB transcripts, 45,920 are intronic and 7,928 are intergenic. This high number of intronic transcripts might be a result of premature transcripts, intron retention and intronic reads, which can be prevalent even in polyA libraries (Lee et al., 2020; Ruiz-Orera et al., 2015).

We reasoned that those intronic EB transcripts showing high coexpression with annotated transcripts might represent such processing errors. To test this, we examined coexpression patterns in each of the 45,920 intronic EB transcripts with its corresponding annotated transcript. We computed Spearman correlations for the transcript pairs in a tissue specific manner. Dissimilar from the expectation if the EB genes stem from premature transcripts, or intron retention, the vast majority of correlation values for the EB gene-annotated transcript pairs in the 6 tissue/tumors is less than 0.50; the mean value is less than 0.3 for all samples (Figure 7.2; Supplementary File 1). While the mean correlation between any two genes chosen at random is 0.2 (Supplementary Figure A3; Supplementary File 1).

7.3.4 Phylostratigraphy of protein-coding transcripts

In order to identify the human-specific orphan genes, the 52,032 pcEB transcripts and all the protein-coding transcripts annotated in Gencode were classified using the automated R-based software, phylostratr (Arendsee et al., 2019a). This, analysis inferred 43,734 EB transcripts and 1,404 annotated transcripts as human-specific orphans.

To extend the search space to cover as-yet unannotated protein-coding genes (i.e., ORFs not annotated as genes), we applied Liftoff (Shumate and Salzberg, 2021), mapping human EB genes to whole genomes of nine closely related species. 13,921 EB transcripts were mapped to other genomes. The final phylostratal assignments are based on phylostratr (Arendsee et al., 2019b) predictions merged with Liftoff designations (Figure 7.3), yield 29,813 EB orphan transcripts.

7.3.5 Novel genes exhibit differential expression across cancer, gender and race

We identified differentially expressed (DE) genes between TCGA tumor samples and corresponding normal samples from GTEx, controlling for age, gender, and race in the DESeq2 model (Love et al., 2014; Stephens, 2017). We similarly identified genes differentially expressed between gender and race for each tissue or tumor.

Thousands of EB genes show differential expression between tumor and normal samples; figure 7.4 shows the numbers of EB and annotated genes differentially expressed for six tumor types. A higher number of EB genes are downregulated in tumor samples, consistent with more EB genes expressed in GTEx tissues.

Numbers of EB genes that are differentially expressed varies with tissue and tumor type. Around 19,000 EB genes containing an ORF coding for a human-specific protein (EB orphan genes) are downregulated in THCA whereas only 6,953 such genes are downregulated in KIRK. 8,576 EB orphan genes are upregulated in ESCA, but only 849 EB orphan genes are upregulated in THCA. In total, 305 EB orphans are downregulated and 11 are upregulated across all tumor-normal comparisons.

Eight hundred and forty one EB orphan genes are downregulated in males compared to females in BRCA-breast samples. We found 411 EB orphan genes upregulated in males compared to females in KIRC-kidney samples.

We also computed genes differentially expressed among individuals of African, Asian, and European ancestries (as self-reported in the metadata). Maximum number of orphan genes, 670 are upregulated in African populations, compared to European for PRAD-prostate samples. In BRCA-breast samples, 416 genes are upregulated in European as compared to Africans.

We assessed the reproducibility of differential expression results using an independent strand-specific RNA-seq dataset comprising 25 colon and 25 adjacent normal colon tissue (Paredes et al., 2020). Using DESeq2, we computed DE genes between normal colon and adjacent normal colon tissue, controlling for race and age. In this smaller dataset, 113 EB genes are upregulated and 213 are downregulated (more than two fold change and adjusted p-value < 0.05

). We found 66% of the upregulated EB genes are also upregulated in TCGA-COAD and 73% of downregulated EB genes are also deregulated in TCGA-COAD. Compared to TCGA-READ, 53% of the upregulated EB genes are upregulated in TCGA-READ and 82% of the downregulated genes are downregulated in TCGA-READ.

7.3.6 Expression of EB genes in strand-specific RNA-Seq data

To validate the expression of EB transcripts identified from GTEx and TCGA datasets, we used independent RNA-Seq data. We chose strand-specific RNA-Seq data because it allows resolution of ambiguous reads in overlapping genes transcribed from opposite strands and provides more accurate expression counts. This enables us to resolve expression of intronic monoexonic transcripts that may come from the antisense strand (Bussotti et al., 2016; Ruiz-Orera et al., 2015).

We used two human strand-specific RNA-Seq datasets to examine the expression patterns of EB and annotated genes. The first dataset consisted of 8 samples from 4 human tissues, liver, heart, brain, and testis (Ruiz-Orera et al., 2015). A total of 1,690 EB transcripts are expressed with a median of ≥ 1 in this dataset. The second dataset consisted of 50 RNA-Seq samples from colon tumors and adjacent normal tissues (Paredes et al., 2020). A total of 1,803 EB transcripts are expressed with a median of ≥ 1 in this dataset. Eight hundred and ninety EB transcripts are common to both datasets (Supplementary File 7). The majority of the EB transcripts from both these datasets potentially code for a human-specific protein (Figure 7.5 C and D).

11,510 unannotated human EB transcripts have a phylostrata shared with chimpanzee, and are also unannotated in this species. We used eight chimpanzee stranded RNA-Seq samples (Ruiz-Orera et al., 2015) to test the expression of these EB transcripts in chimpanzees, quantifying expression of all annotated chimpanzee RNAs and these 11,510 unannotated transcripts. Five hundred sixty four are expressed these chimp samples (median of ≥ 1) (Figure 7.6; Supplementary File 8).

7.3.7 Novel genes provide cell-specific markers

We hypothesized that many of the EB genes will show cell specific expression patterns. A previous study reported expression of Drosophila orphan genes in specific cell types of the testis using single-cell RNA-Seq (scRNA-Seq) data (Witt et al., 2019). To examine whether the EB genes are regulated in a cell-specific manner, we used three different scRNA-Seq datasets from three different organs: liver (Payen et al., 2021), breast (Bhat-Nakshatri et al., 2021), and testis (Guo et al., 2020). We quantified the EB genes along with all annotated genes using 10x Genomics Cell Ranger 6.0.1 (Zheng et al., 2017). We identified multiple cell-clusters corresponding to different cell types in each dataset by scanpy (Wolf et al., 2018; Traag et al., 2019) (Supplementary Figure A5). Then, we identified marker genes for the clusters as reported by scanpy's *rank_genes_groups* function. EB genes are among the top markers of different cell-type clusters in breast, liver and testis. Figure 7.7 shows selected EB genes that mark particular clusters and have evidence of translation in Ribo-Seq datasets. Because of the cell- and condition-specific expression of many EB genes, we anticipate that some could provide prognostic markers of the various unique cell types.

7.3.8 Novel genes are associated with overall survival

We used Cox regression analysis for each cancer type to identify those differentially expressed genes that are associated with overall patient survival. We controlled for the effect of race and gender in our model and labelled genes significantly associated with survival (adjusted p-value < 0.05).

Thousands of differentially expressed EB genes show a significant association with overall survival in multiple tumors; (Figure 7.8) shows the tumor types with the greatest number of survival-associated EB genes. The higher number of EB genes containing an ORF that encodes a species-specific protein that are associated with favourable prognosis, rather than disfavorable prognosis, might be due in part to the higher number of EB genes that are downregulated in cancer.

7.3.9 EB genes harbour hundreds of millions of variants

To investigate the potential involvement of EB genes in human diseases, we queried multiple databases to identify sequence variants overlapping with these as-yet unannotated regions. First, we queried gnomad (v2.1.1 liftover data) (Karczewski et al., 2020) and found 6,985,430 variants located within the 54,794 EB transcripts. As one approach to quantify the extent to which EB genes might harbour deleterious variants we used the Combined Annotation Dependent Depletion (CADD) scores (Rentzsch et al., 2019). Figure 7.9 shows the distribution of the mean CADD scores for different transcript types. The median of the annotated protein-coding (5.12) and lncRNA (4.21) scores is higher than that of EB genes (3.71). Medians of processed pseudogenes (5.18) and retained introns (6.04) are higher that for annotated protein-coding genes (5.12). A recent study similarly reported higher CADD scores for ORFs present in ncRNAs and pseudogenes (Erady et al., 2021).

In another approach, we investigated the extent to which the Catalogue Of Somatic Mutations In Cancer (COSMIC) overlap with EB genes (Figure 7.10). For this, we focused on COSMIC's "non-coding variants", which are defined as variants that occur within intronic or intergenic regions of the genome (`https://cancer.sanger.ac.uk/cosmic/help/ncv/overview`). EB genes harbour a total of 327,565 variants. The greatest numbers of these variants are located in the EB orphan transcripts. COSMIC data annotates 169,283 of the variants as "pathogenic", as predicted by FATHMM (Shihab et al., 2014). Skin, large intestine and lung contains particularly large numbers of pathogenic variants in EB genes (Figure 7.10).

7.3.10 Translation of novel genes revealed by Ribo-Seq data

To further validate translation of expressed novel genes, we used Ribo-Seq data to identify translating ORFs. We downloaded and processed 289 Ribo-Seq samples over 23 different studies and assessed the translation of all the annotated and EB transcripts in these samples. Using the ribotricer (Choudhary et al., 2020) tool, we identified 943 EB transcripts with evidence of translation (Figure 7.11). As expected, the majority of the protein-coding genes have translation

evidence. Many annotated transcripts labeled as retained intron, nonsense-mediated decay, pseudogenes, and lncRNAs were also detected as translating novel ORFs (Figure 7.11). A caveat in this analysis is that EB genes tend to be sparsely expressed, and we had limited Ribo-Seq samples relative to the nearly 27,000 samples we processed to identify EB genes.

We further investigated the phylostrata and transcript class of the EB genes that have translation evidence. The majority of EB genes with evidence of translation, 646, are human-specific orphan genes. Two intergenic EB genes with translation evidence are from the oldest phylostratum, i.e., cellular organisms (Figure 7.11).

7.3.11 Features of novel and annotated proteins

We compared predicted protein lengths, isoelectric points, GC content, and disorder for the annotated and EB proteins. The novel EB proteins and the annotated proteins show a similar pattern of protein lengths across phylostrata, with the oldest phylostrata having larger proteins (Figure 7.12). The isoelectric point and GC content distributions appear distinctive for annotated and EB proteins. The GC content is lower for the ORFs of the EB proteins as compared to that of the annotated genes (Figure 7.12).

To examine codon usage among EB and annotated genes, we computed Relative Synonymous Codon Usage (RSCU) values for each CDS and applied PCA transformation. A cadre of EB genes form a distinct cluster on the PCA plot (Supplementary Figure A4).

We used the IUPred2A tool (Mészáros et al., 2018) to predict the percentage of disordered residues in annotated and EB proteins. Figure 7.13 shows the distribution of percent disordered residues for the EB and annotated proteins. No significant difference in the distributions of predicted disorder among proteins encoded by human novel EB orphan genes and annotated protein-coding genes is found. However, the alternatively spliced EB genes overlapping annotated genes show high median value for disorder.

7.3.12 Comparison with CHESS annotations

We compared the 54,794 EB transcripts we identified with the CHESS study that identified 116,156 unannotated transcripts from 9,795 of RNA-Seq samples from GTEx (Pertea et al., 2018). Only four EB transcripts are identical to the unannotated transcripts listed in CHESS. 20,404 of the EB transcripts overlap with CHESS transcripts, but are located on the opposite strand. 17,464 additional EB transcripts are "contained" within CHESS transcripts in an intron-compatible manner. 13,630 EB transcripts do not overlap with any CHESS transcript. This low concordance of EB transcripts with CHESS is not surprising. First, the CHESS annotation filtered out the majority of intron-less transcripts. They also discarded any human-specific genes whose predicted proteins were not similar to those other other species, based on a BLAST homology search, thus removing human orphan genes from consideration.

Moreover, our transcriptome meta-assembly pipeline used more datasets, including the TCGA studies. A total of 19,582 of the high-expressed EB are only detected in one of the TCGA samples. Furthermore, our workflow was different in that we used different tools to perform transcriptome meta-assembly. Our filtering criteria for removing low expressed genes considered tissue-, gender-, race-specific expressions.

7.4 Discussion

More than 85% of the human genome is transcribed, however, only 3% of the genome is annotated as protein-coding (Hangauer et al., 2013; Ruiz-Orera et al., 2015; Pertea et al., 2018; Lorenzi et al., 2021). In humans, unannotated transcribed regions have often been often considered transcriptional noise (Pertea et al., 2018; Hangauer et al., 2013; Louro et al., 2009; Kuo et al., 2020) or they are annotated as non-coding RNAs such as lncRNAs (Louro et al., 2009; Hangauer et al., 2013; Pertea et al., 2018; Hon et al., 2017; Raj et al.; Ruiz-Orera et al., 2014; Lorenzi et al., 2021; Kuo et al., 2020). However, the unannotated transcriptome is rich in transcripts that contain ORFs, and a number of studies show that many transcripts assumed to be noise or non-coding have ORFs which could code for novel peptides (Ruiz-Orera et al., 2014;

Raj et al.; Dowling et al., 2020; Van Oss and Carvunis, 2019; Erady et al., 2021; Li et al., 2020, 2021; Dowling et al., 2020; Martinez et al., 2020). By leveraging RNA-Seq data from tens of thousands of samples comprising diverse tumor and non-diseased tissues taken from individuals across ancestries, genders, age and other characteristics, we characterized the dynamic transcriptome of diverse human tissues, annotating EB orphan and and non-orphan transcripts that show high expression across diverse conditions.

In this study, we have chosen a stringent criteria for inclusion of transcribed sequence to be considered as potential novel genes, requiring the median expression of an EB transcript to be equal to or greater than that of the median of all protein-coding transcripts. Therefore, many of the novel transcripts we detected, but we did not include in our analysis are expressed to the level of many transcription factors. However, because we anticipated that many novel transcripts would be highly expressed only under limited conditions, we partitioned our analysis in a tissue-race-gender specific manner. Indeed, we found that computing the median over all conditions would have resulted in low or near-zero median values for many novel transcripts that are highly expressed in specific tissues or tumors. For example, if we had pooled all GTEx samples, before applying filtration for low expression, only 763 EB transcripts would have been considered as highly expressed. Several previous studies, including a large study that identified novel genes from GTEx RNA-Seq data (Pertea et al., 2018), filtered out transcripts based on single exons, similarity search, or presence of repeats. However, because many orphan genes with established functions do have some of these characteristics, and indeed a high proportion of orphan genes are monoexonic, we did not apply this filter.

Proteins with high intrinsic structural disorder are unable to fold into harmful aggregates (Dowling et al., 2020), and orphan gene proteins have been inferred by several studies to be highly disordered (Ruiz-Orera et al., 2014; Basile et al., 2017; Dowling et al., 2020; Van Oss and Carvunis, 2019). However, our analysis of disorder by IUPred2A (Mészáros et al., 2018) indicated there was no significant difference in the distributions of predicted disorder among proteins encoded by human novel EB orphan genes and annotated protein-coding genes. The alternatively

spliced EB genes, which overlap annotated genes, show a slightly higher median value for disorder, consistent with a previous study that reports young ORFs mapping to exons have elevated disorder (Dowling et al., 2020). *De novo* genes with intrinsic structure disorder show positive correlation with their GC content (Van Oss and Carvunis, 2019; Dowling et al., 2020), a trend we also observe (Supplementary Figure A6).

Eighty % of the novel EB genes reside fully within introns of annotated genes. It is challenging to control for gDNA even with abundant DNase treatment (Louro et al., 2009; Bussotti et al., 2016) and contamination due to gDNA or to immature RNAs that contain introns may show up as short spurious monoexonic transcripts after transcriptome assembly. Thus, one interpretation of intronic transcripts is that they are artifacts (Louro et al., 2009). In contrast, an analysis of mouse transcriptomics data that applied multifaceted controls, revealed thousands of intronic monoexonic transcripts (Bussotti et al., 2016), and a recent study of a small set of RNA-Seq samples from six primates found almost twice as many novel intronic transcripts with ORFs as compared to intergenic (Dowling et al., 2020). We evaluated our data in several ways to address these interpretations. Based on the large size of the data we analysed, it is unlikely that contamination due to gDNA would be so consistent and pervasive. The expression of the EB genes is detected across multiple samples, and in stranded and scRNA-Seq datasets. Also, because of our high bar for level of expression for EB genes, the artifacts of contamination would be minimized. Finally, the expression of the vast majority of novel intronic transcripts are not correlated with the corresponding annotated transcripts, as might be expected if the intronic regions were derived from immature RNAs.

Systematic pan-cancer analysis reveals that many of the novel EB genes are differentially expressed and associated with overall patient survival. Thus many of these look promising candidates to for novel diagnostic and therapeutic interventions (Xi et al., 2017; Dvinge and Bradley, 2015).

Using independent Ribo-Seq data, we found evidence of translation for hundreds of EB genes. However, many of the conditions of genetics, development, and disease from which the RNA-Seq

data we worked with were for EB analysis are not represented in the RNA-Seq from which we derived translation evidence; thus, we anticipate that some translated RNAs will not show translation evidence in this dataset. Evidence of transcription, or even translation, does not necessarily imply a cellular function (Ruiz-Orera et al., 2014; Orr et al., 2020). In our study, we show that thousands of EB orphan genes we identified are associated with altered survival in cancer and other conditions. Gene annotated as "ncRNAs" but containing translated ORFs also have been reported to be associated with survival in cancer (Erady et al., 2021).

Many EB genes are highly expressed in cell-specific clusters, implying that some EB genes may provide novel biomarkers for unique cell types; this phenomenon requires further investigation.

In this study, we provide a best-practices automate pipeline for human orphan gene identification. Our approach can be easily extended to any other species (Li et al., 2021). We developed this pipeline keeping in mind reproducibility, scaleability, and ease-of-modification (Singh et al., 2021). All the code used in the analysis is uploaded to GitHub (https://github.com/urmi-21/Human_orphan_genes). All the data generated is made available openly so that this study can serve as a resource for additional study.

7.5 Materials and Methods

7.5.1 RNA-Seq Datasets

RNA-Seq data was from the two most comprehensive human RNA-Seq repositories, the Genotype-Tissue Expression portal (GTEx) (Lonsdale et al., 2013) and The Cancer Genome Atlas (TCGA; https://www.cancer.gov/tcga). All the non-diseased samples from GTEx are polyA+ RNA-seq libraries. The library selection protocol differs for some TCGA cohorts (TCGA-LAML samples used "Alternate sample pipeline" http://gdac.broadinstitute.org/runs/stddata__latest/samples_report/LAML_Notifications.html). Corresponding sample metadata were obtained from the GTEx portal (https://www.gtexportal.org) and GDC portal (https://portal.gdc.cancer.gov/).

Stranded RNA-Seq data for human tissues were downloaded from NCBI-GEO accessions GSE69241 (Ruiz-Orera et al., 2015) and GSE146009 (Paredes et al., 2020). Chimpanzee stranded RNA-Seq data was downloaded from GSE69241 (Ruiz-Orera et al., 2015). Single cell RNA-Seq data for breast, liver and testis tissues were obtained from accessions GSE164898 (Bhat-Nakshatri et al., 2021), SRP285767 (Payen et al., 2021) and SRP214255 (Guo et al., 2020) respectively.

289 samples from 23 studies of RNA-Seq data from NCBI-GEO and RPFdb (Wang et al., 2019) (Supplementary File 5) were downloaded.

7.5.2 Evidence-based orphan gene annotation pipeline

We engineered a scaleable best-practice pipeline for detection of annotated and unannotated genes using the tens of thousands of RNA-Seq samples. The abstract pipeline flowchart is shown in Figure 7.15; the steps are discussed in the following sections.

7.5.2.1 Read alignment and transcriptome assembly

The RNA-Seq raw reads from GTEx and TCGA samples were extracted from the downloaded BAM files using Biobambam (Tischler and Leonard, 2014). Only samples with paired-end layout were considered. The RNA-Seq reads were aligned the human reference genome (version GRCH38 GCA_000001405.15) using STAR (Dobin et al., 2013). Transcripts were assembled using Stringtie (Pertea et al., 2015) with reference GTF annotation as *guide*. We implemented the full RNA-Seq alignment and assemble processing pipeline using pyrpipe (Singh et al., 2021) and ran it on PSC Bridges HPC. The GTEx processing pipeline is available from `github.com/urmi-21/pyrpipe/blob/master/case_studies/GTEx_processing`.

7.5.2.2 Meta-assembly pipeline

From the assembled transcriptomes for each tissue/tumor type, we considered a maximum of the 200 samples that were richest in expression of unannotated genes. We used gffcompare

(Pertea and Pertea, 2020) to compare the assembled transcripts with the reference annotation and select samples that have highest unannotated transcripts.

A consolidated reference annotation was built using a combination of Taco (Niknafs et al., 2017) and Mikado (Venturini et al., 2017) tools. Taco was run to retain transcripts with minimum 100 nt in length with expression based filter turned off. Mikado was tuned to identify transcripts longer than 100 nt. A custom mikado *scoring* file was used to enhance the prediction of novel transcripts. We used orfipy (Singh and Wurtele, 2021) to find all possible ORFs of 99 nt or longer in the transcripts. orfipy results were provided to mikado *serialise* step to annotate the CDS and other features.

First, Mikado and Taco were run independently to merge annotations from the same tissue and tumor type. Then, individual annotations from all non-tumor tissues were merged and the same was done for all tumors. We used gffcompare to compare tumor and non-tumor Taco annotations with corresponding Mikado annotations. Novel transcripts not predicted by Mikado were added to get final GTEx and TCGA annotations. Finally, we used mikado to merge the GTEx, TCGA, and Gencode reference annotations to arrive at single consolidated GTF file (Figure 7.15). This annotation file contained 1,384,542 transcripts.

7.5.2.3 Pre-screening of consolidated assembly

We used Mikado's compare utility to compare our annotation with the Gencode reference annotation. We first removed all the transcripts from our annotation that exactly matched the reference annotation leaving only novel transcripts. We removed fusion products, and all transcripts that overlapped with annotated transcripts except for those marked as alt-spliced. This set contained alt-spliced, intronic and intergenic transcripts.

We selected a subset of alt-spliced transcripts by running phylostratr (choosing a smaller set of reference species for quick analysis) (Arendsee et al., 2019b) to retain only those which were human-specific. Then, we ran cd-hit (Li et al., 2001) on the transcripts (90% similarity over 90% length) and removed redundant transcripts.

For the final compilation of transcripts that we used to quantify gene expression, we added the alt-spliced, intronic, and intergenic transcripts to all annotated transcripts.

7.5.2.4 Quantification pipeline

We used Salmon's selective alignment approach (Patro et al., 2017; Srivastava et al., 2020) for quantification of all the novel EB and annotated transcripts. All human reference genome along with viral sequences were used as decoy. Final transcript level counts were aggregated to gene level counts using a custom python script.

7.5.2.5 Phylostratigraphy

We used phylostratr (Arendsee et al., 2019b) to infer phylostrata of all Gencode annotated proteins and for each EB transcript that contained an ORF of length 99 nt. The annotated proteomes for 241 species sampled over 26 phylostrata were used for phylostratal inference (Supplementary File 4). This run of phylostratr assigned 44 alt-spliced EB genes to older phylostrata. These genes were kept in the analysis.

Additionally, we used Liftoff (Shumate and Salzberg, 2021) to map all EB transcripts to five ape species: chimpanzee, gorilla, macaque, gibbon, orangutan; and to four other animal species: cattle, zebrafish, mouse and rat. 13,921 EB transcripts were mapped to other genomes coding for the homologous proteins. The final phylostrata assignments of EB genes were updated based on phylostratr and Liftoff results.

7.5.3 Filtering novel genes based on level of expression

We retained a total of 54,794 novel transcripts (54,748 genes) filtering out those of low and spurious expression values.

Filtering algorithm:

1. Group samples by tissue sampling site, race, and gender

2. Discard groups with less than 5 samples

3. Within each group, compute median of all transcripts and reduce the data to keep the all the transcripts with median ≥ 1 TPM.

4. Compute the median of the median of all remaining annotated protein-coding transcripts

5. Filter out EB transcripts with median expression less than the median computed in step 4

A less stringent criteria to filter would have been to use annotated long non-coding genes instead of annotated protein-coding genes in step 4.

7.5.4 Intronic transcripts coexpression analysis

To investigate the coexpression patterns of intronic transcripts with the corresponding annotated transcript that contained the intron, we computed pairwise spearman correlations for each intronic and annotated transcript pair. The correlation values were computed using the raw counts in a tissue/tumor specific manner. Calculations were performed in python 3 using pandas.

7.5.5 Differential expression analysis

DESeq2 was used for differential gene expression analysis between TCGA tumor and GTEx normal samples (Love et al., 2014; Stephens, 2017). We controlled for age, gender and race. For gender-specific tissues, only age and race factors were considered. Samples with missing data on age, gender or race were removed from the analysis.

We identified DE genes between tumor and normal samples. Using contrasts (http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#contrasts) we evaluated DE genes between males and females, and among African, Asian and European populations. Differential expression criteria is more than 2 fold change and adjusted p-value < 0.05.

7.5.6 Survival analysis

For each tumor type, the DE genes were analyzed in relation to survival of the individuals. The data for overall survival was obtained from TCGA metadata. Specifically, the columns *days_to_death* and *days_to_last_follow_up* are used.

We used the R package, survival (https://cran.r-project.org/web/packages/survival/index.html), to access the survfit and coxph functions for survival analysis. We used survfit to plot Kaplan-Meier survival curves, and coxph to fit a Cox proportional hazards regression model, while controlling for the effect of age, gender, and race. P values for the Cox regression model were adjusted using the Benjamini-Hochberg method. Genes with adjusted p-value < 0.05 are reported as significantly associated with overall survival.

7.5.7 Processing stranded RNA-Seq

Human stranded RNA-Seq data were obtained from two studies: GSE146009 and GSE69241. The stranded RNA-seq data were run through out salmon quantification pipeline to get the estimated read counts for annotated and EB genes.

Chimpanzee reference genome and transcriptome were obtained from GCA_002880755.3; chimpanzee stranded RNA-Seq data was downloaded from accession GSE69241. To the reference transcriptome, we added 11,510 human EB transcripts that were annotated as being in the chimpanzee genome by Liftoff. We used salmon to quantify the annotated and EB transcripts. The quantification pipeline was implemented using pyrpipe (Singh et al., 2021).

7.5.8 Single-cell RNA-Seq data processing

Single-cell datasets for multiple tissues were downloaded from accessions: SRP301923 (breast), SRP285767 (liver), SRP214255 (testis), All single-cell data used were sequenced using the 10X Genomics' Chromium^{TM} platform. We processed these samples using the 10X Genomics' Cell Ranger version 6.0.1 (Zheng et al., 2017). We created a custom reference transcriptome to be

used with Cell Ranger by running *mkref* command with all the annotated and highly expressed EB genes. Cell Ranger's *count* command was run each RNA-Seq sample with custom reference created with *mkref*.

Downstream analysis were performed using scanpy (Wolf et al., 2018) in a study specific manner. All samples from a single study were concatenated into a single *AnnData* object. Data pre-processing steps included removing cells with too many mitochondrial genes expressed or too many total counts and log normalizing the data.

7.5.9 Ribo-Seq analysis pipeline

We used Ribosome profiling (Ribo-Seq) data from 23 studies and processed a total of 289 samples. Ribo-Seq data were processed on a sample-by-sample basis. Sequencing adapters were removed before aligning the reads to the human reference genome using STAR aligner (Dobin et al., 2013). We used ribotricer (Choudhary et al., 2020) to detect translating ORFs in all the annotated and highly-expressed EB transcripts. ribotricer was run with suggested parameters for human samples.

7.5.10 Variant analysis

To identify variants contained in the annotated and EB genes, We first queried gnomad (v2.1.1 liftover data) (Karczewski et al., 2020). To identify disease causing mutations we using the Catalogue Of Somatic Mutations In Cancer (COSMIC) (Sondka et al., 2018). We used Tabix (Li, 2011) to query the variants from the zipped vcf files.

Variant deleteriousness was examined using Combined Annotation Dependent Depletion (CADD) scores (Rentzsch et al., 2019). Mean CADD score for each transcript was calculated using custom python and bash scripts.

7.5.11 Comparison with CHESS and norfs

We compared with EB annotation with all the novel transcripts in CHESS (v2.2) (Pertea et al., 2018) annotation using the gffcompare tool (Pertea and Pertea, 2020).

7.5.12 Computation of protein disorder and other features

Protein disorder was predicted using IUPred2A (Mészáros et al., 2018) in both short mode. Emboss (Rice et al., 2000) was used to calculate the isoelectric points. Custom python script was written compute CDS length, GC%, Codon Usage (CU) and Relative Synonymous Codon Usage (RSCU).

RepeatMasker (http://www.repeatmasker.org) was used to identify repeats and low complexity regions in all the annotated and EB transcripts. RepeatMasker was run with the *species* parameter set to human.

7.6 Supplementary Data

Supplementary data and files are available from
https://github.com/urmi-21/Human_orphan_genes.

7.7 Acknowledgements

We are grateful to all members of the Wurtele lab and to our COV-IRT colleagues (https://www.cov-irt.org/) for helpful and stimulating discussions. We thank Luca Venturini and other developers of Mikado, for providing support for their tool. We thank Robin Gogerty, Jake Miller, and Levi Baber at Iowa State University for help with administrative and IT services.

This work is funded in part by the National Science Foundation award IOS 1546858, "Orphan Genes: An Untapped Genetic Reservoir of Novel Traits" and by the Center for Metabolic Biology, Iowa State University. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number

ACI-1548562. In particular, it used the Bridges HPC environment through allocations

TG-MCB190098 and TG-MCB200123 awarded from XSEDE and the HPC Consortium.

7.8 References

Arendsee, Z., Li, J., Singh, U., Seetharam, A., Dorman, K. S., and Wurtele, E. S. (2019a). phylostratr: a framework for phylostratigraphy. *Bioinformatics*, 35(19):3617–3627.

Arendsee, Z., Li, J., Singh, U., Seetharam, A., Dorman, K. S., and Wurtele, E. S. (2019b). phylostratr: A framework for phylostratigraphy. *Bioinformatics*, 35(19):3617–3627.

Arendsee, Z. W., Li, L., and Wurtele, E. S. (2014). Coming of age: orphan genes in plants. *Trends in plant science*, 19(11):698–708.

Baalsrud, H. T., Tørresen, O. K., Solbakken, M. H., Salzburger, W., Hanel, R., Jakobsen, K. S., and Jentoft, S. (2018). De novo gene evolution of antifreeze glycoproteins in codfishes revealed by whole genome sequence data. *Molecular biology and evolution*, 35(3):593–606.

Basile, W., Sachenkova, O., Light, S., and Elofsson, A. (2017). High gc content causes orphan proteins to be intrinsically disordered. *PLoS computational biology*, 13(3):e1005375.

Bhat-Nakshatri, P., Gao, H., Sheng, L., McGuire, P. C., Xuei, X., Wan, J., Liu, Y., Althouse, S. K., Colter, A., Sandusky, G., et al. (2021). A single-cell atlas of the healthy breast tissues reveals clinically relevant clusters of breast epithelial cells. *Cell Reports Medicine*, 2(3):100219.

Blevins, W. R., Ruiz-Orera, J., Messeguer, X., Blasco-Moreno, B., Villanueva-Cañas, J. L., Espinar, L., Díez, J., Carey, L. B., and Albà, M. M. (2021). Uncovering de novo gene birth in yeast using deep transcriptomics. *Nature Communications*, 12(1):1–13.

Blume, C., Jackson, C. L., Spalluto, C. M., Legebeke, J., Nazlamova, L., Conforti, F., Perotin, J.-M., Frank, M., Butler, J., Crispin, M., et al. (2021). A novel ace2 isoform is expressed in human respiratory epithelia and is upregulated in response to interferons and rna respiratory virus infection. *Nature Genetics*, 53(2):205–214.

Bussotti, G., Leonardi, T., Clark, M. B., Mercer, T. R., Crawford, J., Malquori, L., Notredame, C., Dinger, M. E., Mattick, J. S., and Enright, A. J. (2016). Improved definition of the mouse transcriptome via targeted rna sequencing. *Genome research*, 26(5):705–716.

Choudhary, S., Li, W., and D. Smith, A. (2020). Accurate detection of short and long active orfs using ribo-seq data. *Bioinformatics*, 36(7):2053–2059.

Couso, J.-P. and Patraquim, P. (2017). Classification and function of small open reading frames. *Nature Reviews Molecular Cell Biology*, 18(9):575.

Dennis, A. B., Ballesteros, G. I., Robin, S., Schrader, L., Bast, J., Berghöfer, J., Beukeboom, L. W., Belghazi, M., Bretaudeau, A., Buellesbach, J., et al. (2020). Functional insights from the gc-poor genomes of two aphid parasitoids, aphidius ervi and lysiphlebus fabarum. *BMC genomics*, 21:1–27.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21.

Dowling, D., Schmitz, J. F., and Bornberg-Bauer, E. (2020). Stochastic gain and loss of novel transcribed open reading frames in the human lineage. *Genome biology and evolution*, 12(11):2183–2195.

Dvinge, H. and Bradley, R. K. (2015). Widespread intron retention diversifies most cancer transcriptomes. *Genome medicine*, 7(1):1–13.

Erady, C., Boxall, A., Puntambekar, S., Jagannathan, N. S., Chauhan, R., Chong, D., Meena, N., Kulkarni, A., Kasabe, B., Bhayankaram, K. P., et al. (2021). Pan-cancer analysis of transcripts encoding novel open-reading frames (norfs) and their potential biological functions. *NPJ genomic medicine*, 6(1):1–17.

Gubala, A. M., Schmitz, J. F., Kearns, M. J., Vinh, T. T., Bornberg-Bauer, E., Wolfner, M. F., and Findlay, G. D. (2017). The goddard and saturn genes are essential for drosophila male fertility and may have arisen de novo. *Molecular biology and evolution*, 34(5):1066–1082.

Guo, J., Nie, X., Giebler, M., Mlcochova, H., Wang, Y., Grow, E. J., Kim, R., Tharmalingam, M., Matilionyte, G., Lindskog, C., et al. (2020). The dynamic transcriptional cell atlas of testis development during human puberty. *Cell stem cell*, 26(2):262–276.

Hangauer, M. J., Vaughn, I. W., and McManus, M. T. (2013). Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet*, 9(6):e1003569.

Hon, C.-C., Ramilowski, J. A., Harshbarger, J., Bertin, N., Rackham, O. J., Gough, J., Denisenko, E., Schmeier, S., Poulsen, T. M., Severin, J., et al. (2017). An atlas of human long non-coding RNAs with accurate 5 ends. *Nature*, 543(7644):199–204.

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443.

Klasberg, S., Bitard-Feildel, T., and Mallet, L. (2016). Computational identification of novel genes: current and future perspectives. *Bioinformatics and Biology insights*, 10:BBI–S39950.

Kuo, R. I., Cheng, Y., Zhang, R., Brown, J. W., Smith, J., Archibald, A. L., and Burt, D. W. (2020). Illuminating the dark side of the human transcriptome with long read transcript sequencing. *BMC genomics*, 21(1):1–22.

Lachmann, A., Torre, D., Keenan, A. B., Jagodnik, K. M., Lee, H. J., Wang, L., Silverstein, M. C., and Ma'ayan, A. (2018). Massive mining of publicly available rna-seq data from human and mouse. *Nature communications*, 9(1):1–10.

Lange, A., Patel, P. H., Heames, B., Damry, A. M., Saenger, T., Jackson, C. J., Findlay, G. D., and Bornberg-Bauer, E. (2021). Structural and functional characterization of a putative de novo gene in drosophila. *Nature communications*, 12(1):1–13.

Lee, S., Zhang, A. Y., Su, S., Ng, A. P., Holik, A. Z., Asselin-Labat, M.-L., Ritchie, M. E., and Law, C. W. (2020). Covering all your bases: incorporating intron signal from rna-seq data. *NAR Genomics and Bioinformatics*, 2(3):lqaa073.

Li, H. (2011). Tabix: fast retrieval of sequence features from generic tab-delimited files. *Bioinformatics*, 27(5):718–719.

Li, J., Singh, U., Arendsee, Z., and Wurtele, E. S. (2020). Landscape of the dark transcriptome revealed through re-mining massive rna-seq data. *bioRxiv*, page 671263.

Li, J., Singh, U., Bhandary, P., Campbell, J., Arendsee, Z., Seetharam, A. S., and Wurtele, E. S. (2021). Foster thy young: Enhanced prediction of orphan genes in assembled genomes. *bioRxiv*, pages 2019–12.

Li, L. and Wurtele, E. S. (2015). The QQS orphan gene of *Arabidopsis* modulates carbon and nitrogen allocation in Soybean. *Plant Biotechnol. J.*, 13(2):177–187.

Li, W., Jaroszewski, L., and Godzik, A. (2001). Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, 17(3):282–283.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580.

Lorenzi, L., Chiu, H.-S., Avila Cobos, F., Gross, S., Volders, P.-J., Cannoodt, R., Nuytens, J., Vanderheyden, K., Anckaert, J., Lefever, S., Tay, A. P., de Bony, E. J., Trypsteen, W., Gysens, F., Vromman, M., Goovaerts, T., Hansen, T. B., Kuersten, S., Nijs, N., Taghon, T., Vermaelen, K., Bracke, K. R., Saeys, Y., De Meyer, T., Deshpande, N. P., Anande, G., Chen, T.-W., Wilkins, M. R., Unnikrishnan, A., De Preter, K., Kjems, J., Koster, J., Schroth, G. P., Vandesompele, J., Sumazin, P., and Mestdagh, P. (2021). The rna atlas expands the catalog of human non-coding rnas. *Nature Biotechnology*.

Louro, R., Smirnova, A. S., and Verjovski-Almeida, S. (2009). Long intronic noncoding rna transcription: expression noise or expression choice? *Genomics*, 93(4):291–298.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550.

Martinez, T. F., Chu, Q., Donaldson, C., Tan, D., Shokhirev, M. N., and Saghatelian, A. (2020). Accurate annotation of human protein-coding small open reading frames. *Nature chemical biology*, 16(4):458–468.

McLysaght, A. and Hurst, L. D. Open questions in the study of de novo genes: what, how and why. 17(9):567.

Mészáros, B., Erdős, G., and Dosztányi, Z. (2018). Iupred2a: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic acids research*, 46(W1):W329–W337.

Neme, R. and Tautz, D. (2013). Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC genomics*, 14(1):117.

Newtson, A., Reyes, H., Devor, E. J., Goodheart, M. J., and Bosquet, J. G. (2021). Identification of novel fusion transcripts in high grade serous ovarian cancer. *International journal of molecular sciences*, 22(9):4791.

Niknafs, Y. S., Pandian, B., Iyer, H. K., Chinnaiyan, A. M., and Iyer, M. K. (2017). Taco produces robust multisample transcriptome assemblies from rna-seq. *Nature methods*, 14(1):68–70.

Orr, M. W., Mao, Y., Storz, G., and Qian, S.-B. (2020). Alternative orfs and small orfs: shedding light on the dark proteome. *Nucleic acids research*, 48(3):1029–1042.

Ovchinnikova, T. V., Balandin, S. V., Aleshina, G. M., Tagaev, A. A., Leonova, Y. F., Krasnodembsky, E. D., Men'shenin, A. V., and Kokryakov, V. N. Aurelin, a novel antimicrobial peptide from jellyfish *Aurelia aurita* with structural features of defensins and channel-blocking toxins. 348(2):514–523.

O'Conner, S. and Li, L. (2020). Mitochondrial fostering: The mitochondrial genome may play a role in plant orphan gene evolution. *Frontiers in Plant Science*, 11:1855.

Paredes, J., Zabaleta, J., Garai, J., Ji, P., Imtiaz, S., Spagnardi, M., Alvarado, J., Li, L., Akadri, M., Barrera, K., et al. (2020). Immune-related gene expression and cytokine secretion is reduced among african american colon cancer patients. *Frontiers in oncology*, 10:1498.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, 14(4):417.

Payen, V. L., Lavergne, A., Sarika, N. A., Colonval, M., Karim, L., Deckers, M., Najimi, M., Coppieters, W., Charloteaux, B., Sokal, E. M., et al. (2021). Single-cell rna sequencing of human liver reveals hepatic stellate cell heterogeneity. *JHEP Reports*, 3(3):100278.

Pertea, G. and Pertea, M. (2020). Gff utilities: Gffread and gffcompare. *F1000Research*, 9.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L. (2015). Stringtie enables improved reconstruction of a transcriptome from rna-seq reads. *Nature biotechnology*, 33(3):290.

Pertea, M., Shumate, A., Pertea, G., Varabyou, A., Breitwieser, F. P., Chang, Y.-C., Madugundu, A. K., Pandey, A., and Salzberg, S. L. (2018). CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome biology*, 19(1):208.

Pinskaya, M., Saci, Z., Gallopin, M., Gabriel, M., Nguyen, H. T., Firlej, V., Descrimes, M., Rapinat, A., Gentien, D., de La Taille, A., et al. (2019). Reference-free transcriptome exploration reveals novel RNAs for prostate cancer diagnosis. *Life science alliance*, 2(6).

Qi, M., Zheng, W., Zhao, X., Hohenstein, J. D., Kandel, Y., O'Conner, S., Wang, Y., Du, C., Nettleton, D., MacIntosh, G. C., et al. (2019). Qqs orphan gene and its interactor nf-yc 4 reduce susceptibility to pathogens and pests. *Plant biotechnology journal*, 17(1):252–263.

Raj, A., Wang, S. H., Shim, H., Harpak, A., Li, Y. I., Engelmann, B., Stephens, M., Gilad, Y., and Pritchard, J. K. Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. 5:e13328.

Reinhardt, J. A., Wanjiru, B. M., Brant, A. T., Saelao, P., Begun, D. J., and Jones, C. D. De novo ORFs in Drosophila are important to organismal fitness and evolved rapidly from previously non-coding sequences. 9(10):e1003860.

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., and Kircher, M. (2019). Cadd: predicting the deleteriousness of variants throughout the human genome. *Nucleic acids research*, 47(D1):D886–D894.

Rice, P., Longden, I., and Bleasby, A. (2000). Emboss: the european molecular biology open software suite.

Ruiz-Orera, J. and Albà, M. M. (2018). Translation of small open reading frames: Roles in regulation and evolutionary innovation. *Trends in Genetics*, 2(5):890.

Ruiz-Orera, J., Hernandez-Rodriguez, J., Chiva, C., Sabidó, E., Kondova, I., Bontrop, R., Marqués-Bonet, T., and Albà, M. M. (2015). Origins of de novo genes in human and chimpanzee. *PLoS Genetics*, 11(12):e1005721.

Ruiz-Orera, J., Messeguer, X., Subirana, J. A., and Alba, M. M. (2014). Long non-coding RNAs as a source of new peptides. *elife*, 3:e03523.

Samusik, N., Krukovskaya, L., Meln, I., Shilov, E., and Kozlov, A. P. (2013). Pbov1 is a human de novo gene with tumor-specific expression that is associated with a positive clinical outcome of cancer. *PLoS One*, 8(2):e56162.

Scalzitti, N., Jeannin-Girardon, A., Collet, P., Poch, O., and Thompson, J. D. (2020). A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms. *BMC genomics*, 21:1–20.

Schmidt, F., List, M., Cukuroglu, E., Köhler, S., Göke, J., and Schulz, M. H. (2018). An ontology-based method for assessing batch effect adjustment approaches in heterogeneous datasets. *Bioinformatics*, 34(17):i908–i916.

Schmitz, J. F., Chain, F. J., and Bornberg-Bauer, E. (2020). Evolution of novel genes in three-spined stickleback populations. *Heredity*, 125(1):50–59.

Schmitz, J. F., Ullrich, K. K., and Bornberg-Bauer, E. (2018). Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. *Nature ecology & evolution*, 2(10):1626–1632.

Shihab, H. A., Gough, J., Mort, M., Cooper, D. N., Day, I. N., and Gaunt, T. R. (2014). Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. *Human genomics*, 8(1):1–6.

Shumate, A. and Salzberg, S. L. (2021). Liftoff: accurate mapping of gene annotations. *Bioinformatics*. btaa1016.

Singh, U., Li, J., Seetharam, A., and Wurtele, E. S. (2021). pyrpipe: a Python package for RNA-Seq workflows. *NAR Genomics and Bioinformatics*, 3(2). lqab049.

Singh, U. and Wurtele, E. S. (2020). Genetic novelty: How new genes are born. *Elife*, 9:e55136.

Singh, U. and Wurtele, E. S. (2021). orfipy: a fast and flexible tool for extracting ORFs. *Bioinformatics*. btab090.

Sondka, Z., Bamford, S., Cole, C. G., Ward, S. A., Dunham, I., and Forbes, S. A. (2018). The cosmic cancer gene census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer*, page 1.

Srivastava, A., Malik, L., Sarkar, H., Zakeri, M., Almodaresi, F., Soneson, C., Love, M. I., Kingsford, C., and Patro, R. (2020). Alignment and mapping methodology influence transcript abundance estimation. *Genome biology*, 21(1):1–29.

Stephens, M. (2017). False discovery rates: a new deal. *Biostatistics*, 18(2):275–294.

Suenaga, Y., Islam, S. R., Alagu, J., Kaneko, Y., Kato, M., Tanaka, Y., Kawana, H., Hossain, S., Matsumoto, D., Yamamoto, M., et al. (2014). Ncym, a cis-antisense gene of mycn, encodes a de novo evolved protein that inhibits gsk3β resulting in the stabilization of mycn in human neuroblastomas. *PLoS Genet*, 10(1):e1003996.

Tautz, D. and Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nature Reviews Genetics*, 12(10):692–702.

Tischler, G. and Leonard, S. (2014). biobambam: tools for read pair collation based algorithms on bam files. *Source Code for Biology and Medicine*, 9(1):13.

Traag, V. A., Waltman, L., and Van Eck, N. J. (2019). From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12.

Vakirlis, N., Carvunis, A.-R., and McLysaght, A. (2020). Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *Elife*, 9:e53500.

Vakirlis, N. and McLysaght, A. (2019). Computational prediction of de novo emerged protein-coding genes. In *Computational Methods in Protein Evolution*, pages 63–81. Springer.

Van Oss, S. B. and Carvunis, A.-R. (2019). De novo gene birth. *PLoS Genetics*, 15(5):e1008160.

Venturini, L., Caim, S., Kaithakottil, G., Mapleson, D. L., and Swarbreck, D. (2017). Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *bioRxiv*, page 216994.

Wang, H., Yang, L., Wang, Y., Chen, L., Li, H., and Xie, Z. (2019). Rpfdb v2. 0: an updated database for genome-wide information of translated mrna generated from ribosome profiling. *Nucleic acids research*, 47(D1):D230–D234.

Wang, Q., Armenia, J., Zhang, C., Penson, A. V., Reznik, E., Zhang, L., Minet, T., Ochoa, A., Gross, B. E., Iacobuzio-Donahue, C. A., et al. (2018). Unifying cancer and normal rna sequencing data from different sources. *Scientific data*, 5:180061.

Witt, E., Benjamin, S., Svetec, N., and Zhao, L. (2019). Testis single-cell rna-seq reveals the dynamics of de novo gene transcription and germline mutational bias in drosophila. *Elife*, 8:e47138.

Wolf, F. A., Angerer, P., and Theis, F. J. (2018). Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):1–5.

Xi, X., Li, T., Huang, Y., Sun, J., Zhu, Y., Yang, Y., and Lu, Z. J. (2017). Rna biomarkers: frontier of precision medicine for cancer. *Non-coding RNA*, 3(1):9.

Xie, C., Zhang, Y. E., Chen, J.-Y., Liu, C.-J., Zhou, W.-Z., Li, Y., Zhang, M., Zhang, R., Wei, L., and Li, C.-Y. (2012). Hominoid-specific de novo protein-coding genes originating from long non-coding rnas. *PLoS Genet*, 8(9):e1002942.

Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1):1–12.

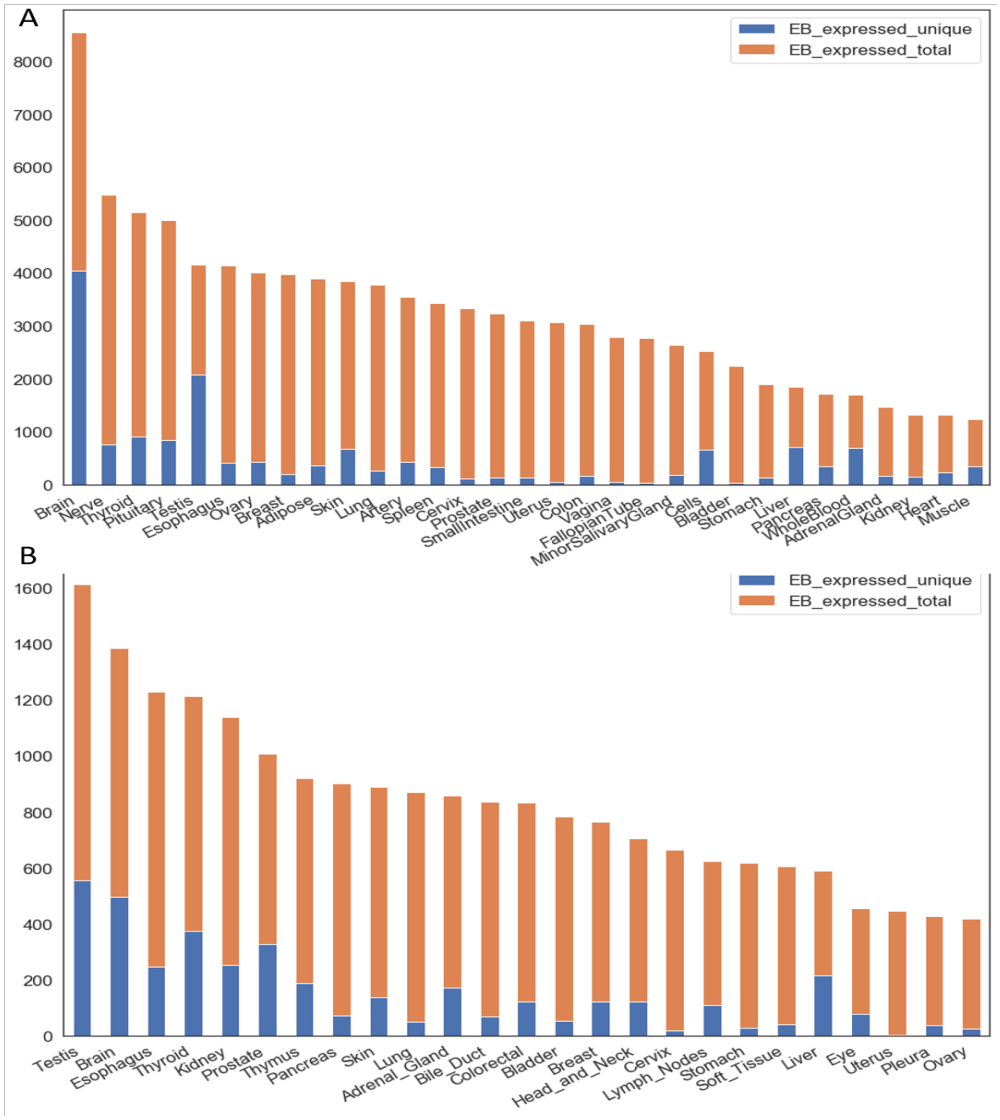

Figure 7.1 Number of highly expressed EB transcripts according to tissue for the GTEx and TCGA cohorts. Barplots represent all transcripts expressed at or above the median of the medians of the annotated protein-coding genes. Transcripts are identified in a tissue-gender-race specific manner and merged by tissue (See Methods). The numbers of transcripts that are uniquely expressed at high levels in only the single tissue or cancer-type are represented by blue bars. **A.** GTEx tissues. **B.** TCGA tissues. TCGA samples from AML, not shown in the plot, expressed the greatest number of EB transcripts, 29,631.

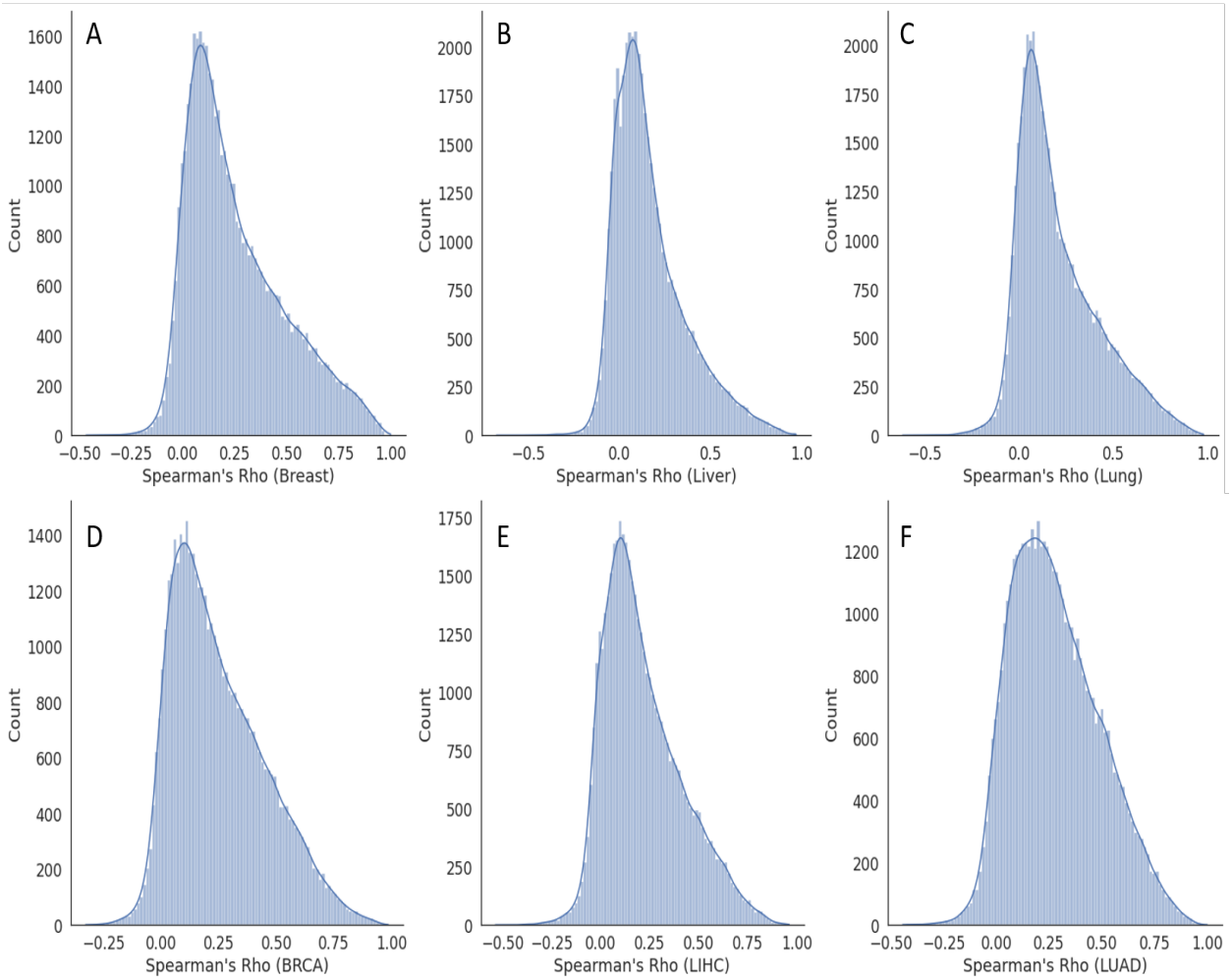

Figure 7.2 Distributions of Spearman's correlation values between each intronic EB transcript and the corresponding annotated transcript. Spearman's correlations are computed for each pair, i.e, intronic EB transcript and corresponding transcript containing the intron, over samples for a tissue or tumor. Distributions of Spearman's correlation values for transcript pairs in: **A.** GTEx breast samples. **B.** GTEx liver samples. **C.** GTEx lung samples. **D.** TCGA BRCA samples. **E.** TCGA LIHC samples. **F.** TCGA LUAD samples.

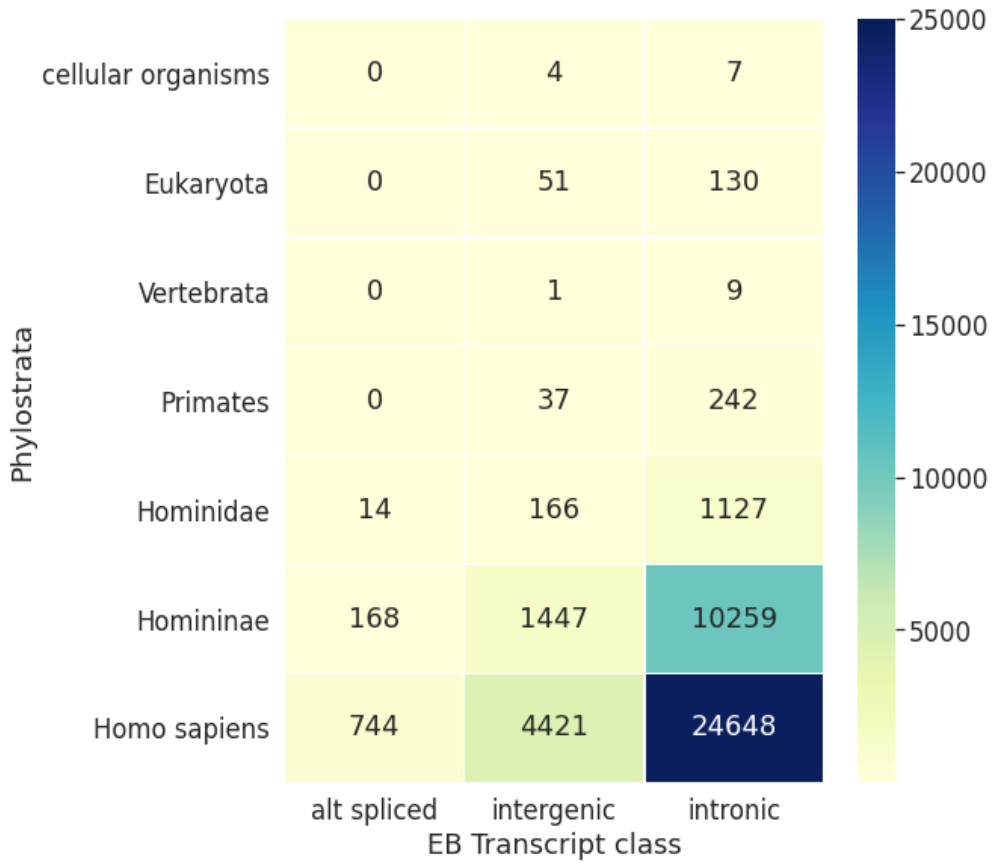

Figure 7.3 Phylostratal assignments for novel transcripts in seven selected phylostrata. Assignments are combined predictions from phylostratr and Liftoff.

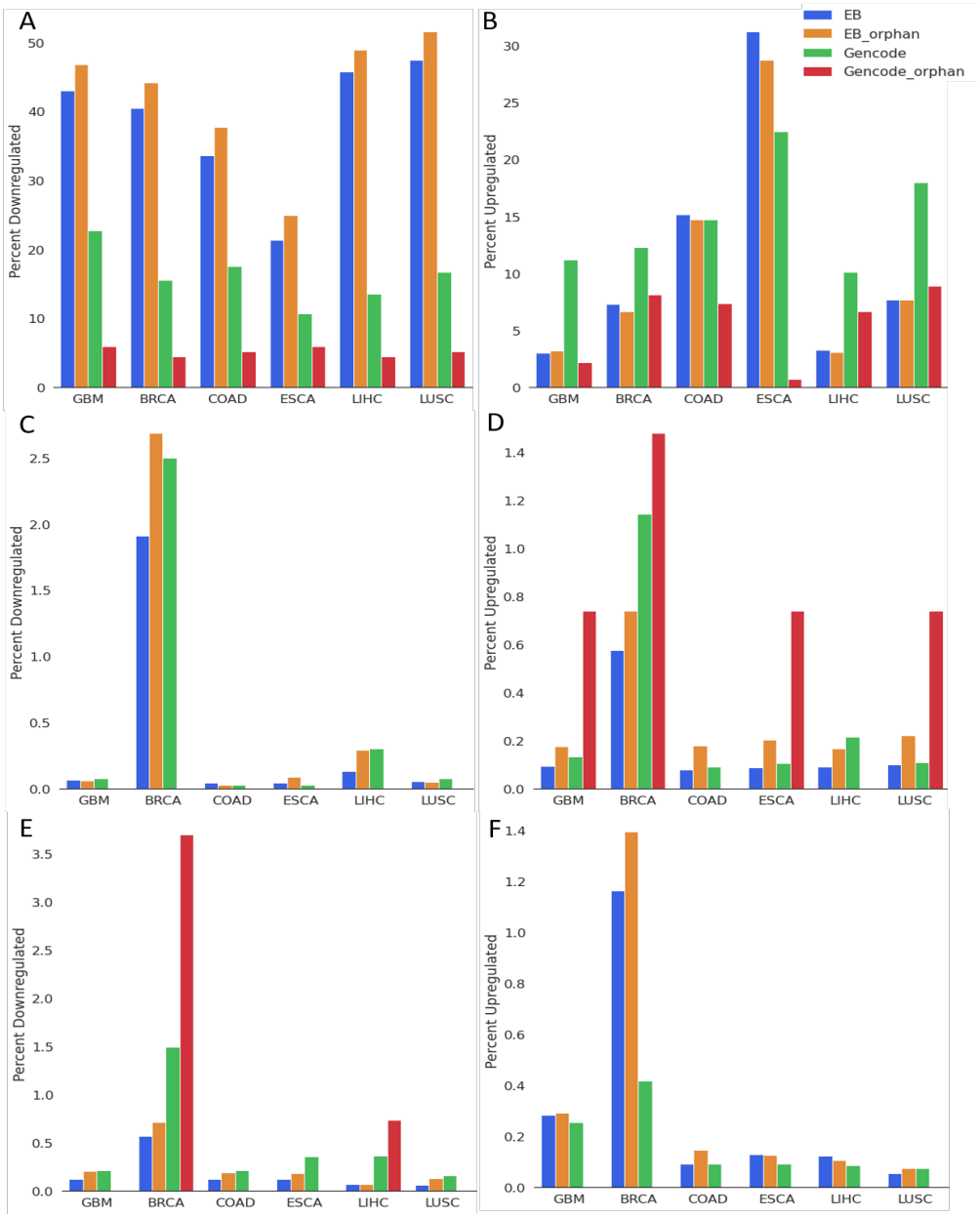

Figure 7.4 Percentage of annotated non-orphan genes, annotated orphan genes, EB non-orphan genes and EB orphan genes that are differentially expressed (DE) in six tumor tissues. **A.** Genes downregulated in tumor compared to normal samples. **B.** Genes upregulated in tumor compared to normal samples. **C.** Genes downregulated in males compared to females. **D.** Genes upregulated in males compared to females. **E.** Genes upregulated in Europeans compared to Africans. **E.** Genes downregulated in Europeans compared to Africans.

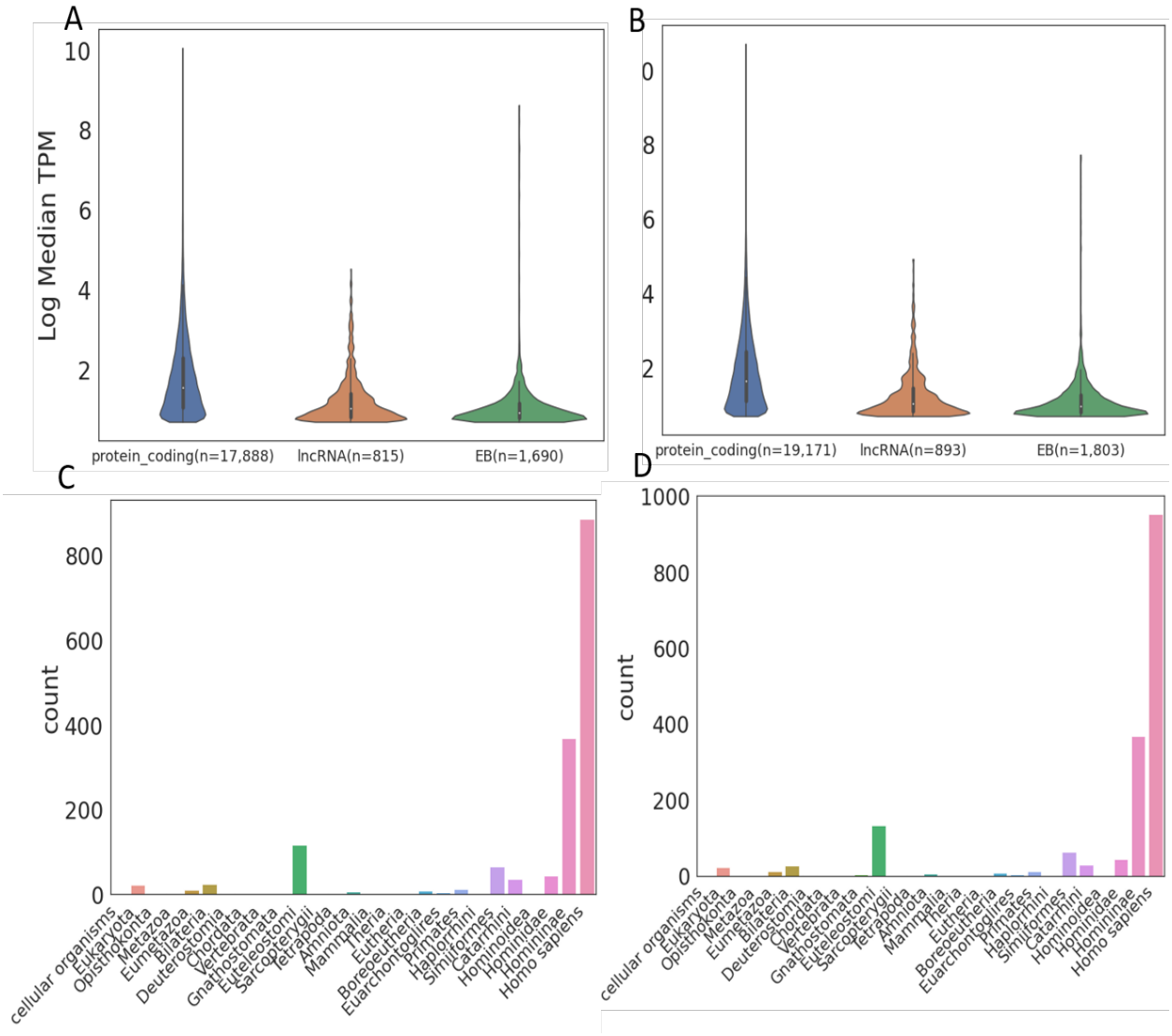

Figure 7.5 Expression of annotated protein-coding and lncRNA, and EB transcripts in strand specific RNA-Seq datasets. Plots show expression in two independent datasets. **A.** Eight RNA-Seq samples from liver, heart, brain, and testis (Ruiz-Orera et al., 2015). A total of 1,690 EB transcripts are expressed with median expression $\geq$ 1 TPM. **B.** Fifty 50 RNA-Seq samples from colon cancer and adjacent normal tissues (Paredes et al., 2020). 1,803 EB transcripts are expressed with median expression $\geq$ 1 TPM. **C.** Frequency of phylostratal assignments for the 1,690 EB transcripts expressed in **A**. **D.** Frequency of phylostratal assignments for the 1,803 EB transcripts expressed in **B**.

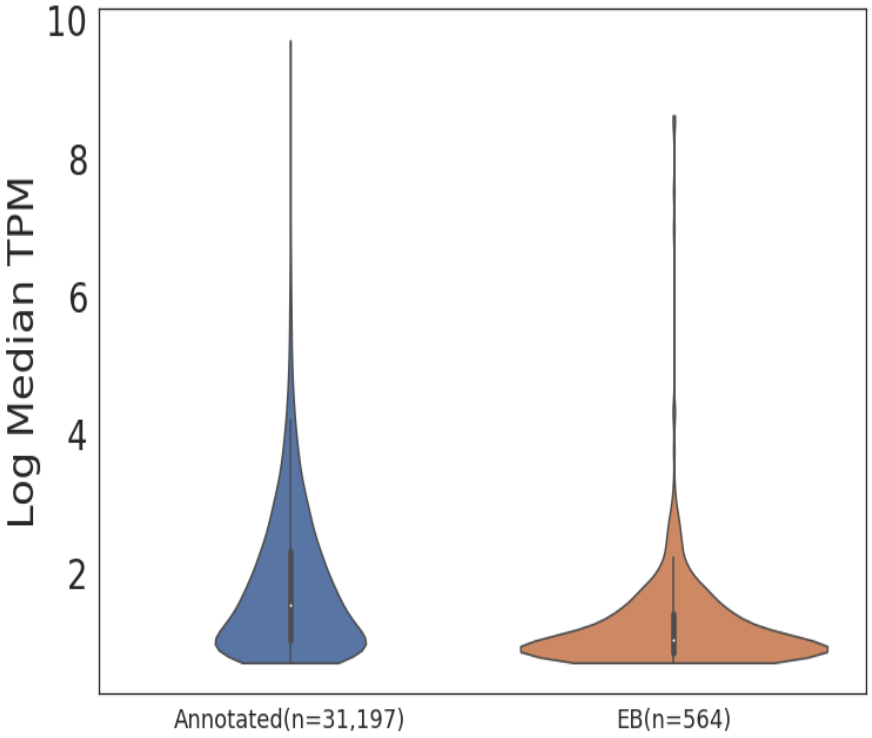

Figure 7.6 Expression distribution of annotated and EB transcripts in chimpanzee strand-specific RNA-Seq data. The data consists of 8 RNA-Seq samples from chimpanzee liver, heart, brain, and testis (Ruiz-Orera et al., 2015). Five hundred and sixty four EB transcripts (5% of the 11,510 unannotated transcripts that we quantified) are expressed with median ≥ 1.

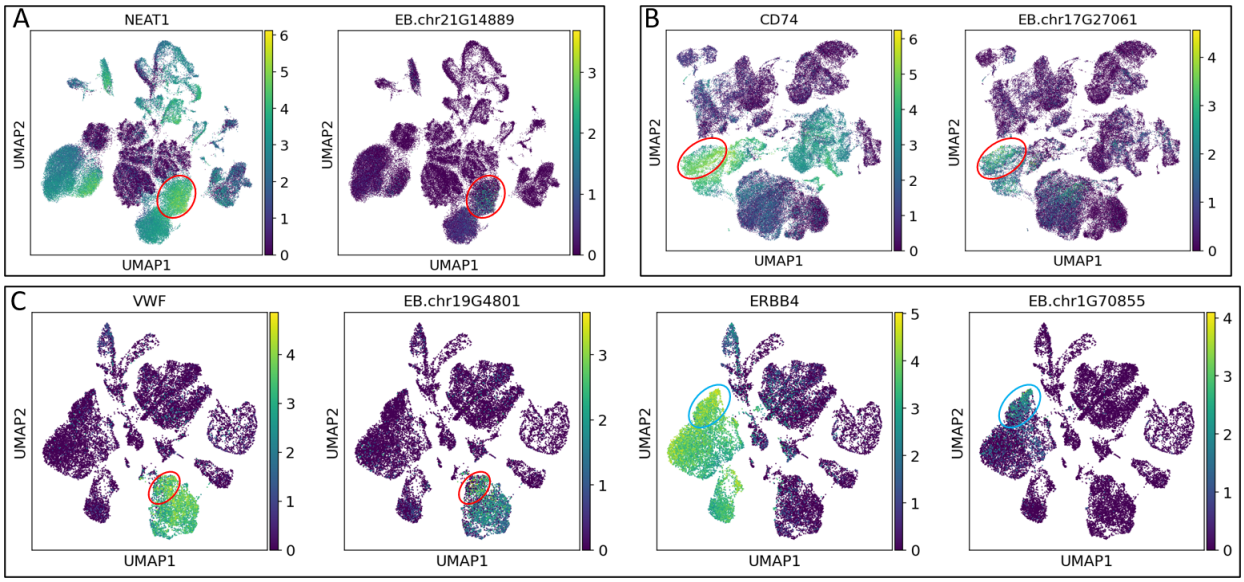

Figure 7.7 Cell-specific expression of novel genes that show evidence of translation using Ribo-Seq data. We examined the expression of novel EB genes in single–cell RNA-Seq datasets. UMAP visualizations. **A.** Liver single-cell RNA-Seq data. We identified 30 clusters of cells (Supplementary Figure A5), using scanpy's leiden function (Wolf et al., 2018; Traag et al., 2019). The novel gene, EB.chr21.G14889 is reported as a marker gene for cluster number 3. The lncRNA nuclear paraspeckle assembly transcript 1 (NEAT1) is the top marker–gene for that cluster. Cluster number 3 is circled in red in the figure. **B.** Breast single-cell RNA-Seq data. Clustering reveals a total of 45 clusters (Supplementary Figure A5). The novel gene, EB.chr17.G27061 is reported as a marker gene for cluster number 11 (circled in red). The CD74 Molecule (CD74) gene is reported as top marker gene for that cluster. **C.** Testis single-cell RNA-Seq data. We identified a total of 31 clusters from the data (Supplementary Figure A5). EB.chr19G4801 and EB.chr1G70855 show cluster-specific expression in two distinct clusters. EB.chr19G4801 is co-localized with marker genes for cluster 8 (circled in red). The Von Willebrand Factor (VWF) gene is one marker for that cluster. VWF is reported as a marker gene for endothelial cells (Guo et al., 2020). EB.chr1G70855 is highly expressed in cluster 2 (circled in blue). Erb-B2 Receptor Tyrosine Kinase 4 (ERBB4) gene also designated by scanpy as a marker for that cluster. Distribution of expression of SOX9 and VIM, designated as markers(Guo et al., 2020), indicates this cluster is a subset of sertoli cells.

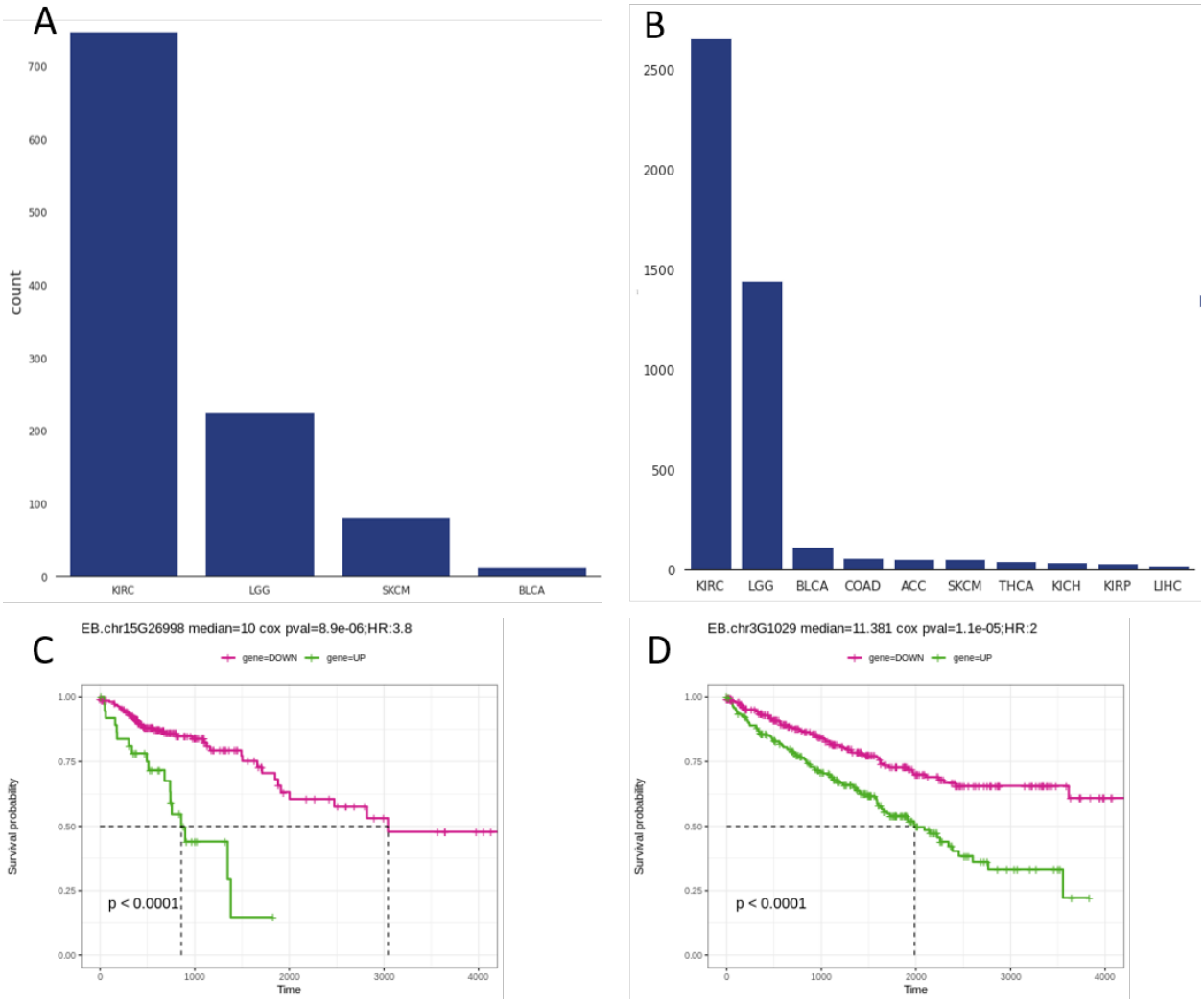

Figure 7.8 Novel genes associated with overall survival in multiple cancer types. **A.** Number of upregulated EB orphan genes associated with unfavourable prognosis. **B.** Number of downregulated EB orphan genes associated with unfavourable prognosis. **C** Kaplan-Meier plot for EB orphan gene EB.chr15G26998 present on chromosome 15. High expression of this gene is associated with poor survival in COAD. **D** Kaplan-Meier plot for EB orphan gene EB.chr3G1029 present on chromosome 3. High expression of this gene is associated with poor survival in KIRC.

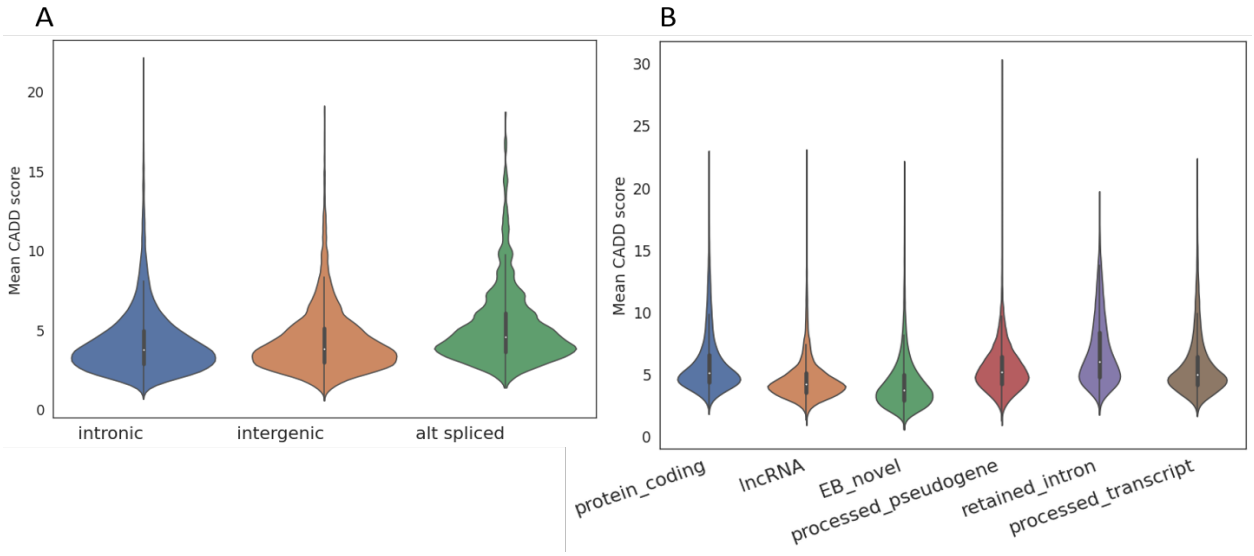

Figure 7.9 Distributions of Combined Annotation Dependent Depletion (CADD) scores (Rentzsch et al., 2019) across annotated and EB transcripts. The CADD score corresponds to the deleteriousness of a variant in the human genome. Violin plots show the medians and distributions of the CADD scores. **A.** Intronic, intergenic and alternatively spliced EB genes. **B.** Annotated transcript types and EB transcripts.

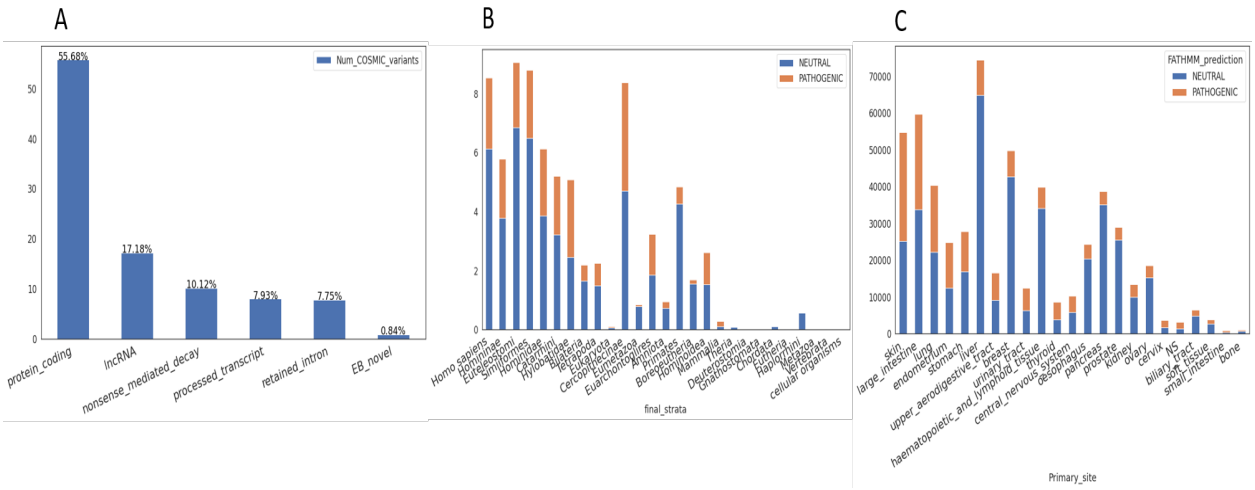

Figure 7.10 Distribution of COSMIC variants and FATHMM predictions for annotated and EB transcripts. We queried the COSMIC non-coding variants (Sondka et al., 2018) that overlap with annotated and EB transcripts. **A, B, C** The numbers of COSMIC variants. **A.** Classes of annotated and EB transcripts. **B.** EB genes grouped by phylostrata and colored by FATHMM predictions of pathogenic variants. Each bar represent number of pathogenic or neutral variants per transcript in the phylostrata. **C.** EB genes grouped by tissue site and colored by FATHMM prediction for pathogenic variants.

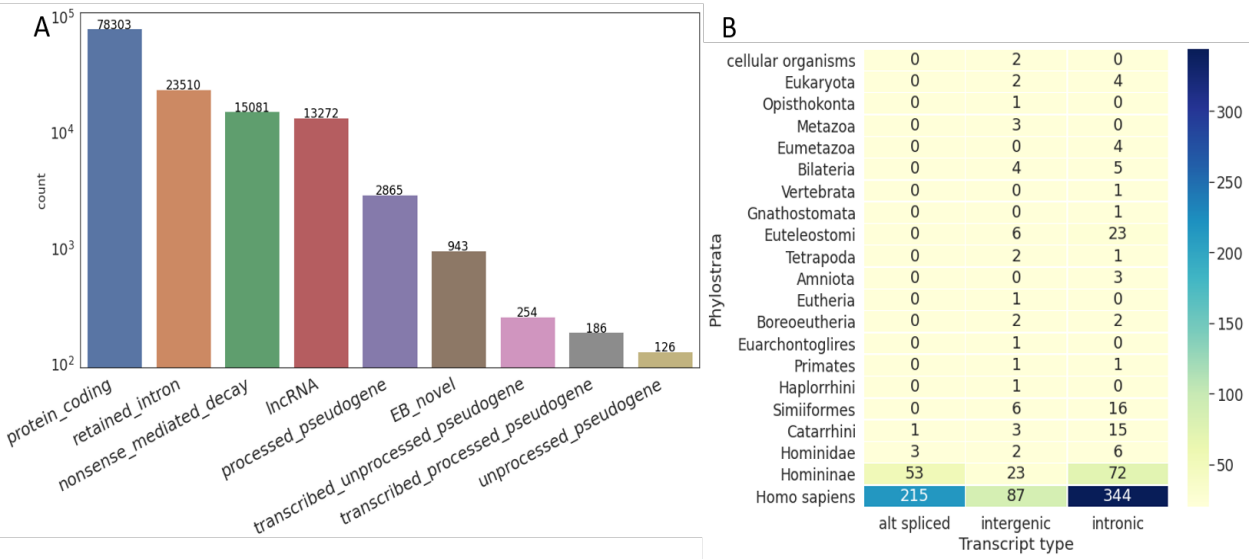

Figure 7.11 Translating ORFs in Ribo-Seq datasets. 289 Ribo-Seq samples in 23 studies were processed and translation of annotated and EB transcripts assessed. **A.** Number of unique translating ORFs detected across various transcript types. **B.** Number of EB transcripts with evidence in Ribo-Seq data, by phylostrata and transcript class.

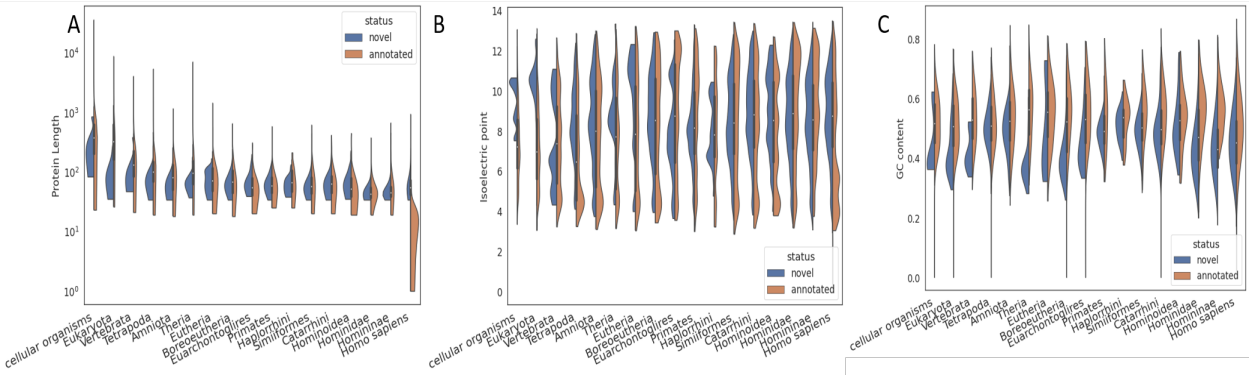

Figure 7.12 Comparison of features across phylostrata for proteins of annotated and EB genes. **A.** Protein length, **B.** Isoelectric point, **C.** GC content.

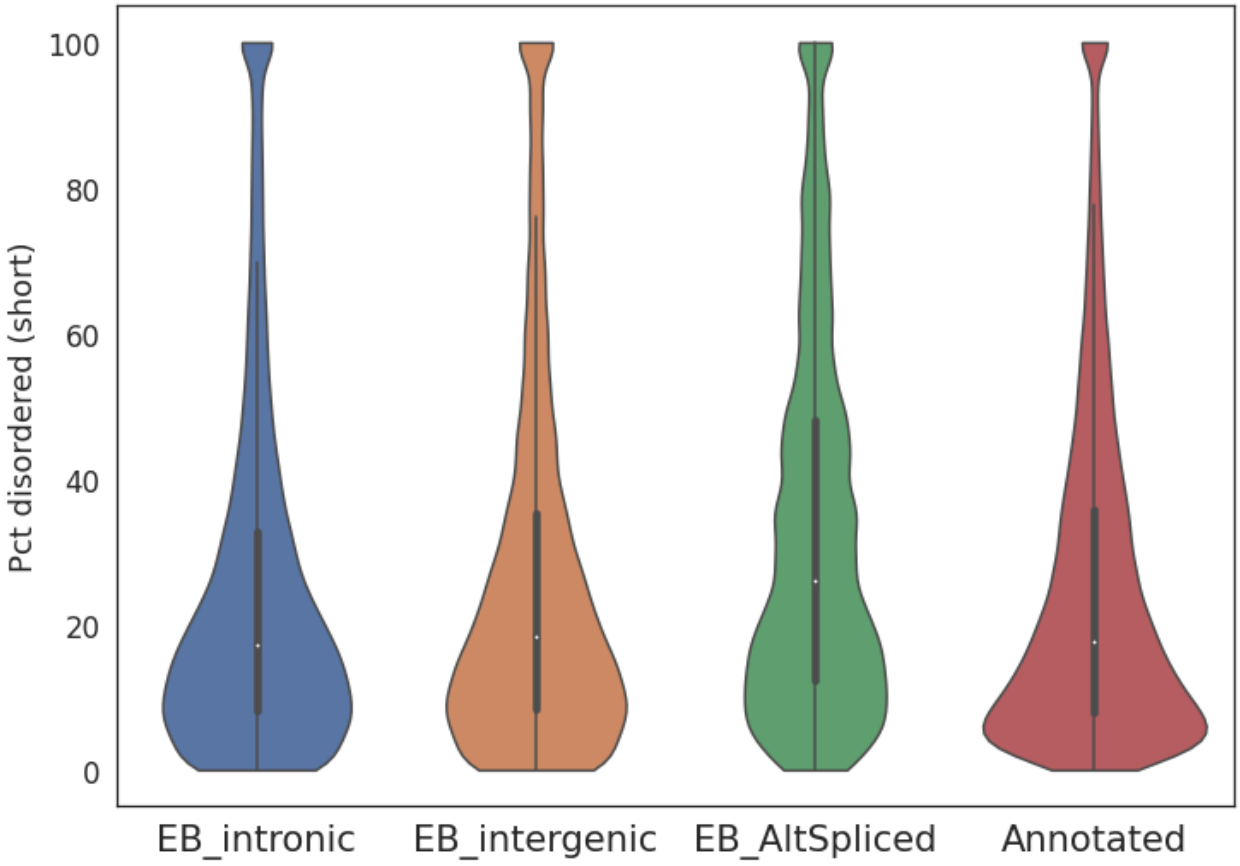

Figure 7.13 Distribution of percent disordered residues predicted for the EB and annotated proteins. Disordered residues within each protein were predicted using IUPred2A's short disorder prediction (Mészáros et al., 2018).

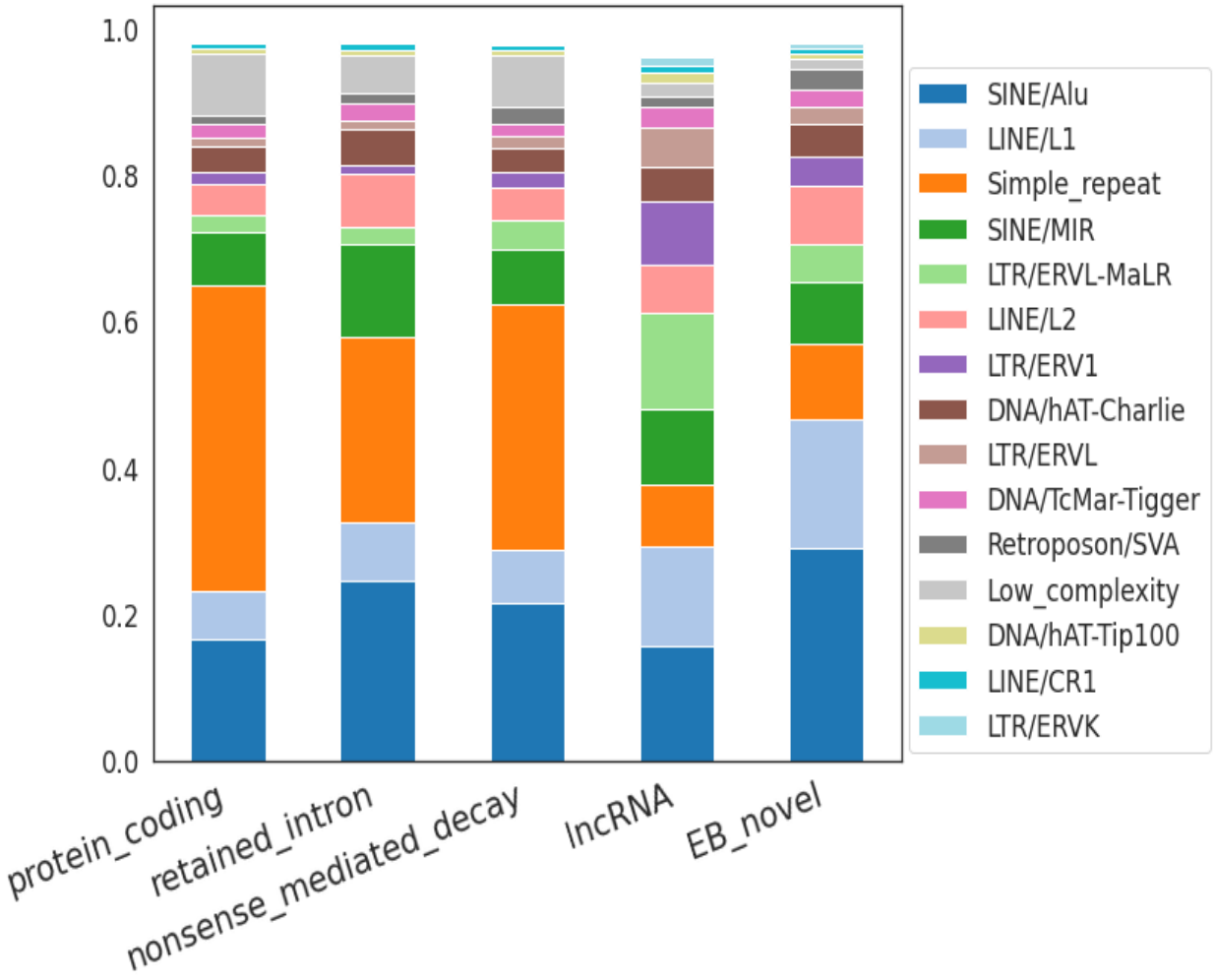

Figure 7.14 EB transcripts have higher proportions of SINE/LINE repeats than annotated protein-coding transcripts. Repeatmasker was used to evaluate the repeats in annotated and EB transcripts. SINE/Alu repeats are most the most prevalent class in EB transcripts, versus "simple repeat" in protein coding annotated genes.

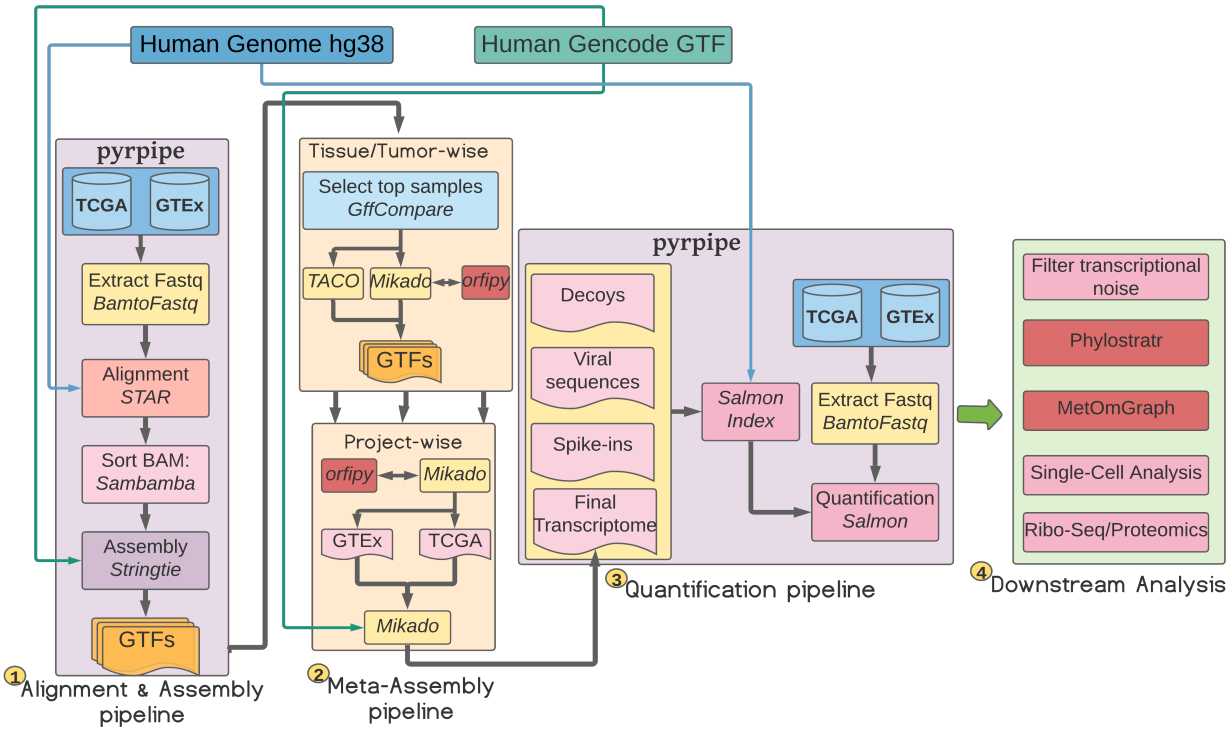

Figure 7.15 Workflow of the study. The alignment and quantification pipelines are implemented in pyrpipe (Singh et al., 2021). For alignment, BAM files were downloaded from GTEx and converted to fastq using biobambam2. STAR (Dobin et al., 2013) was run in 2-pass alignment mode to align reads to the human reference genome (GRCh38 GCA_000001405.15). Transcripts were assembled by Stringtie (Pertea et al., 2015). Individual transcriptomes were consolidated into single transcriptome using our meta-assembly pipeline, consisting of Mikado (Venturini et al., 2017) and Taco (Niknafs et al., 2017). ORFs were identified by orfipy (Singh and Wurtele, 2021). For quantification, a Salmon index was build using human annotated and novel identified transcripts, with whole human genome sequence used as a decoy. BAM files were downloaded from GTEx and TCGA, converted to fastq (using biobambam2) and passed to Salmon's *quant* function for quantification. phylostratr (Arendsee et al., 2019b) was used to infer phylostrata of all Gencode annotated proteins and for each EB transcript.

7.9 Appendix: Supplementary Figures

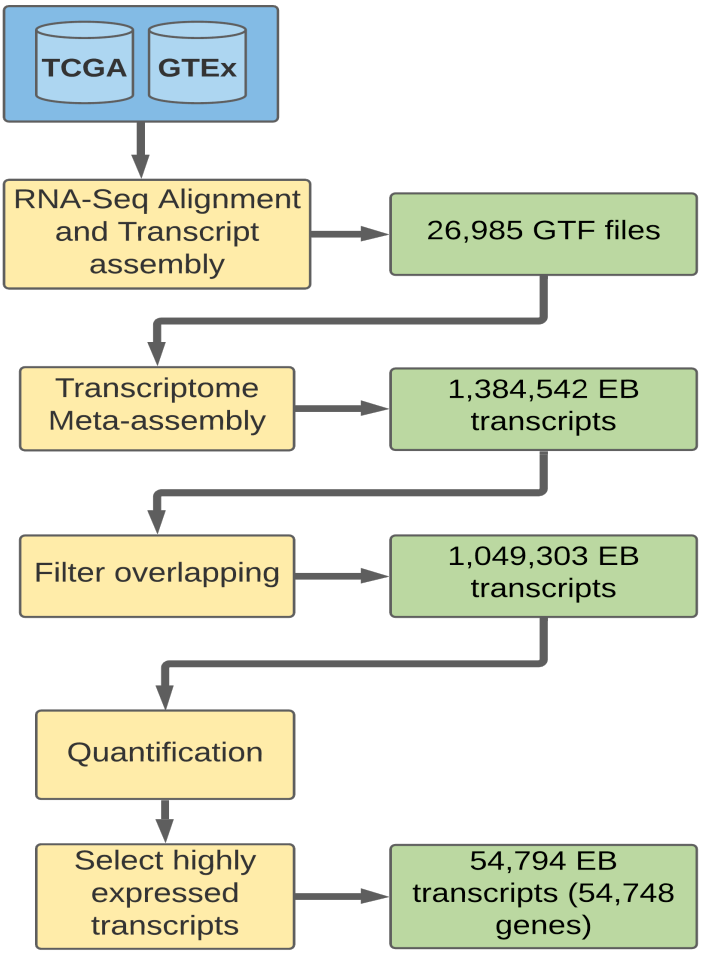

Figure A1 Number of EB transcripts identified at each step of the workflow used in this study.

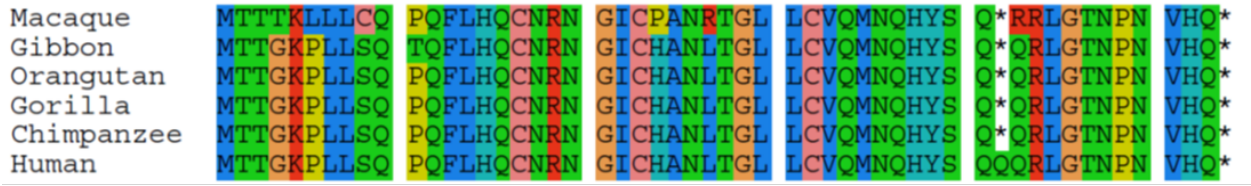

Figure A2 Example of an intergenic EB gene identified in this study, EB.chr6G21600. Liftoff identified mapped the genes to genomes of closely related species where the protein is terminated by an in-frame stop codon. This stop codon is mutated in humans and thus the ORF codes for a longer protein (53 AA).

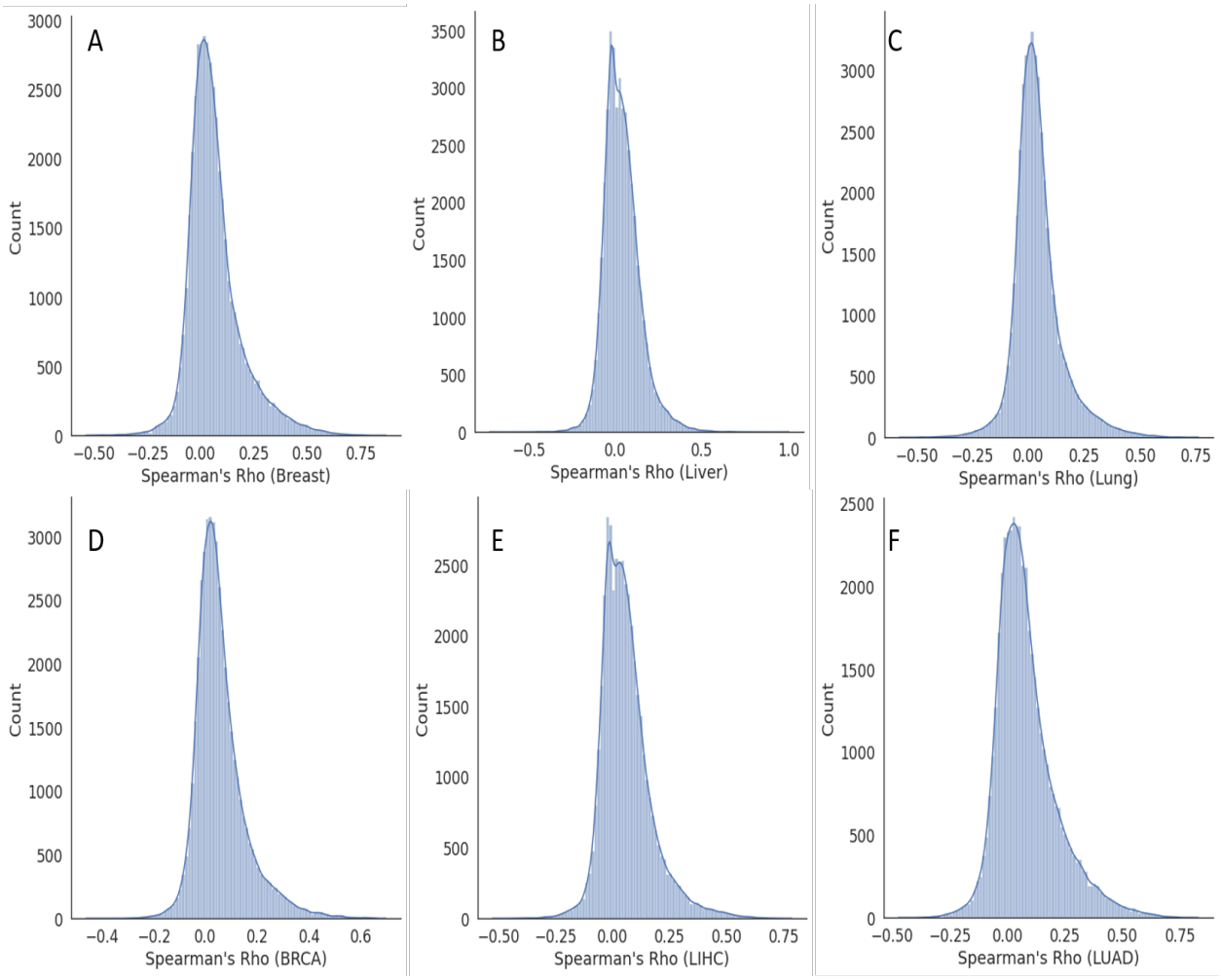

Figure A3 Distributions of Spearman's correlation values between each intronic EB transcript and a randomly chosen annotated transcript. Spearman's correlations are computed for each pair, i.e, intronic EB transcript and corresponding randomly chosen annotated transcript, over samples for a tissue or tumor. Distributions of Spearman's correlation values for transcript pairs in: **A.** GTEx breast samples. **B.** GTEx liver samples. **C.** GTEx lung samples. **D.** TCGA BRCA samples. **E.** TCGA LIHC samples. **F.** TCGA LUAD samples.

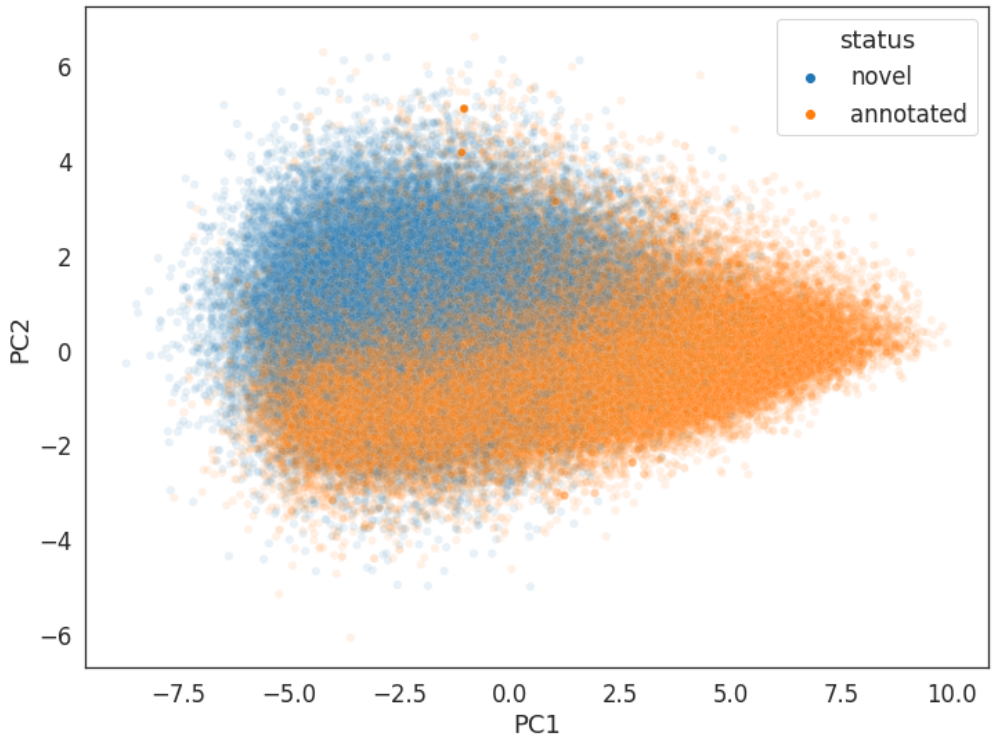

Figure A4 PCA plot using Relative Synonymous Codon Usage (RSCU) values of all pcEBs and annotated protein coding transcripts.

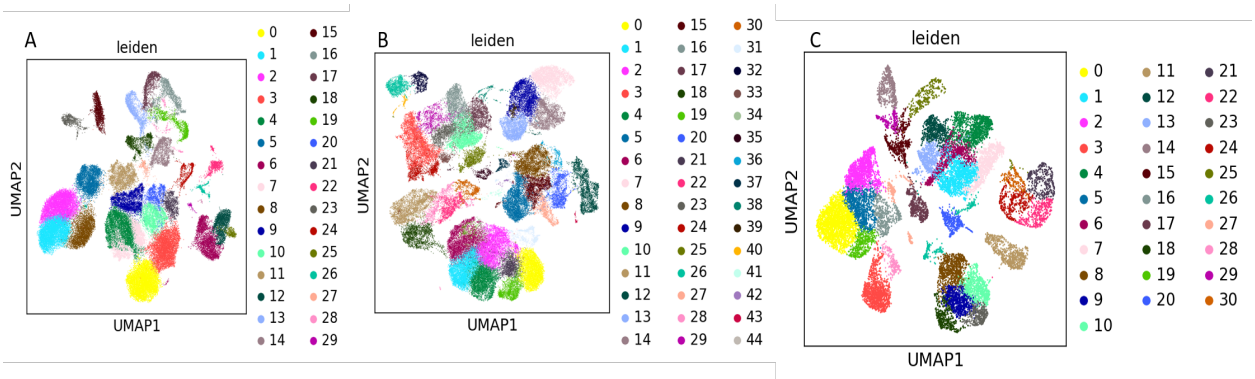

Figure A5 Single-cell RNA-Seq clusters identified in **A.** Liver **B.** Breast and **C.** Testis datasets.

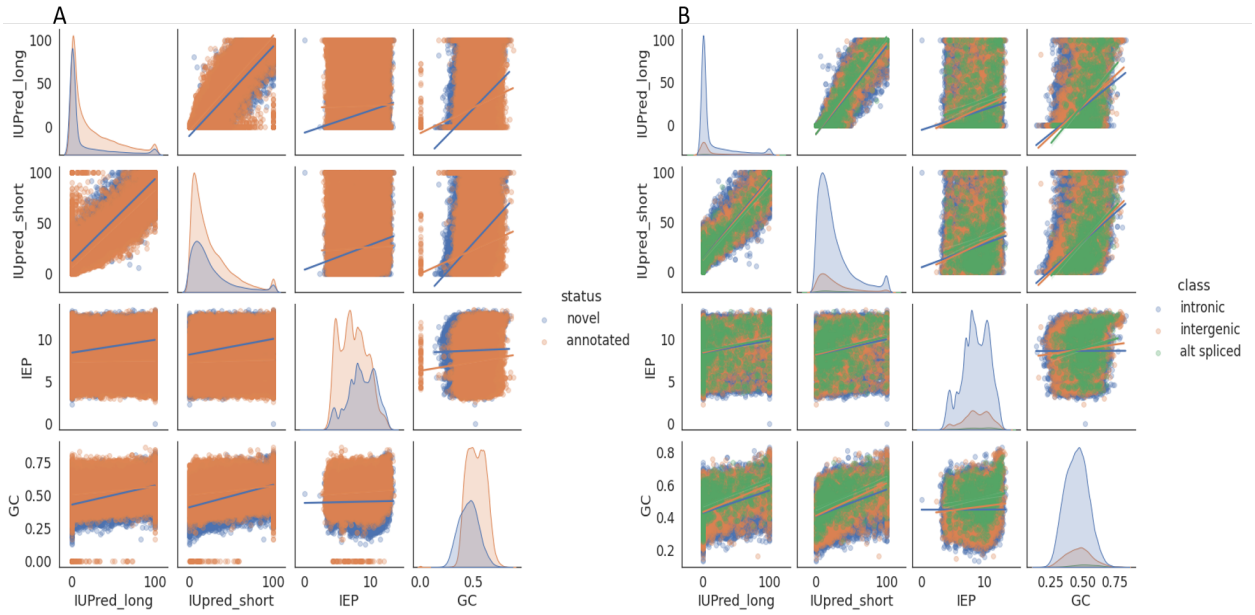

Figure A6 Variation of protein disorder and isoelectric point with GC content. Lines are estimated by fitting a linear regression model. **A.** All EB and annotated proteins. Colors are: orange annotated; blue EB **B.** Only EB proteins. Colors are: orange intergenic EB; blue intronic EB; green alt-spliced EB

CHAPTER 8. GENERAL CONCLUSION

A high quality genome assembly is the first step for reliable gene annotation of any organism. The ab initio gene annotation methods relies on statistical models that are trained on canonical gene features (Li et al., 2021; Scalzitti et al., 2020). Thus the ab initio methods perform poorly for the annotation of genes with non-canonical features, such as orphan genes (Li et al., 2021; Singh and Wurtele, 2021). RNA-Seq evidence based approaches are being used widely to detect the expression of novel genes (Ruiz-Orera et al., 2015; Li et al., 2021; Pertea et al., 2018; Hon et al., 2017; Kuo et al., 2020; Lorenzi et al., 2021; Singh and Wurtele, 2021; Klasberg et al., 2016). Ribo-Seq and proteomics data are helpful to validate translation of novel genes (Ruiz-Orera and Albà, 2018; Raj et al.; Ruiz-Orera et al., 2015; Li et al., 2021; Pertea et al., 2018; Hon et al., 2017).

This dissertation contributes tools for processing, exploration and analysis of big RNA-Seq datasets. Using these tools we provide a pipeline for the annotation of novel orphan genes using RNA-Seq data. First, we described MetaOmGraph, a tool for interactive exploratory analysis of large omics datasets (Singh et al., 2020). MetaOmGraph provides a number of features to explore gene coexpression and differential expression patterns over thousands of samples and infer their functional importance. This approach is particularly helpful for novel orphan genes that lack similarity to existing protein-coding genes. Using MetaOmGraph "domain experts" can easily explore and re-mine existing datasets from various perspectives.

Next, in order to contribute to the ongoing efforts to study the COVID-19 disease, we re-analysed existing RNA-Seq datasets and focused on population specific expression patterns. We utilized MetaOmGraph to for quick exploratory analysis existing human cancer and non-diseased RNA-Seq datasets. We found, a number of genes involved in immune-related functions are differential expressed among African and European populations (Singh et al., 2021a). We were able to validate and replicate the results of MetaOmGraph's exploratory data analysis through

in-depth statistical analysis of the data. Using single-cell RNA-Seq data, we found some of the diferentially expressed genes show cell specific expression patterns. The results of our study have important implications for improving precision treatment of COVID-19 for African Americans.

We also highlighted the unequal representation of minority populations in genomics studies that hinders understanding of diseases in such populations. We suggested that data from genomics studies should have better and open data and metadata sharing models. Since, diseases are a result of complex interactions between one's genetic and environmental factors, we proposed that genomics studies should also record one's socioeconomic, and other environmental factors for researchers to use.

In chapter 5, we presented orfipy, a python tool for fast and flexible annotation of novel coding sequences in large transcriptomics data (Singh and Wurtele, 2021). Compared to existing tools, orfipy provides a number of novel features to find Open Reading Frames (ORFs) in DNA sequences. orfipy is suitable to be used with raw sequencing data (FASTQ format) or with transcriptomics data (FASTA format) from any organism. orfipy scales well with large amount of data and thus is suitable to be used with millions of transcript sequences (Singh and Wurtele, 2021).

Chapter 6 introduced pyrpipe, a python package for reproducible RNA-Seq processing pipelines (Singh et al., 2021b). pyrpipe is based on a novel object-oriented framework that allow for easy implementation of RNA-Seq pipelines, in pure python. To allow for easy reproducibility, pyrpipe provides features like automated logging, easy parameter management and tracking changes to input data and scripts. To scale large workflows pyrpipe is compatible with a number of popular workflow management systems. A dedicated module to access the NCBI-SRA data makes it straightforward to use pyrpipe for the harmonized reanalysis publicly available RNA-Seq data (Singh et al., 2021b).

Finally, chapter 7 makes a significant contribution to the study of human orphan genes. In this study firstly, we engineered a best-practices reproducible and scaleable pipeline for

identification and annotation of orphan genes using RNA-Seq data. This pipeline is implemented in pyrpipe and uses orfipy to accelerate the CDS annotation step.

Leveraging terabytes of RNA-Seq data, we annotated a comprehensive catalog of thousands of novel human genes. Majority of these novel genes are orphan genes, as revealed by phylostratigraphy analysis (Arendsee et al., 2019; Tautz and Domazet-Lošo, 2011). These novel genes were identified from thousands of RNA-Seq samples from multiple diverse non-diseased and tumor tissues. The identified novel genes show strong selective expression patterns based on tissue, tumor, race, and gender.

We investigated the differential expression patterns for these genes and found thousands of novel genes that are differentially expressed across multiple tumors. Cox regression analysis indicated several novel genes are significantly associated with overall patient survival. These genes are promising candidates to explore for therapeutic and diagnostic purposes. We validated the expression of the novel genes using independent stranded RNA-Seq, single-cell RNA-Seq datasets. This analysis confirmed the expression of thousands of novel genes, the majority of which were orphans. A number of novel genes are also expressed in a cell-specific manner. Further investigation is needed to confirm cell-type-specific roles of these orphan genes.

The novel genes overlap with millions of known variants. Using the COSMIC database (Sondka et al., 2018), we found that many of these variants are known to be pathogenic and are spread across a number of tissue sites. This further suggests that some the novel orphan genes might be important players in human diseases.

Future work needs to be done to to experimentally validate the proteins encoded by the novel orphan genes. We plan to query a number of existing proteomics databases and perform proteomics analysis on the most promising candidate orphan genes.

The work presented here contributes to improving the identification and annotation of human orphan and other yet unannotated novel genes. Using thousands of RNA-Seq samples from diverse tissues we compiled a catalog of novel, potentially protein-coding, human orphan genes. As growing evidence suggest, many of these genes might be important in various human diseases.

Thus, this dissertation expands the existing human genome annotation and contributes to the understanding of the cryptic genomic and transcriptomic "dark matter".

The recent release of a complete Telomere-to-Telomere human genome is a promising step towards bridging the gaps in our knowledge of the human genome (Nurk et al., 2021). Coupled with the rapidly improving third-generation long-read sequencing technology, it is now possible to sequence full length transcripts and discover novel isoforms (Clark et al., 2020; Chen et al., 2019; Sun et al., 2021; Sharon et al., 2013; Kuo et al., 2020). With constant improvement in quality and quantity of genomic and transcriptomic datasets, evidence based gene annotation methods will continue to improve and perform better.

In the work presented here, the human reference genome assembly (GRCh38) is used to align the short RNA-Seq reads from multiple tissues of people with diverse ancestries. Although the human reference genome has served its purpose and guided thousands of studies for the exploration of the genetic basis of health and disease, its continuous use a reference genome for diverse populations has several shortcomings (Sherman and Salzberg, 2020; Rosenfeld et al., 2012; Miga and Wang, 2021). The human reference genome assembly is predominantly sequenced from a single individual of European ancestry and thus does not capture the vast genetic variations among different populations (Wong et al., 2020; Nurk et al., 2021). Significant genomic regions present on the genomes of people with other ancestries are missing from the reference assembly (Levy-Sakin et al., 2019; Sherman et al., 2019). In context of RNA-Seq mapping, a recent study (Kaminow et al., 2020) proposed the use of a human pan-genome (Kaminow et al., 2020; Sherman and Salzberg, 2020; Miga and Wang, 2021) that reduces the overall RNA-Seq mapping error by two-fold as compared to using the reference genome. Interestingly, the error rate using the reference genome is reported to be quite small i.e. 0.5%-0.6% (Kaminow et al., 2020).

With the reduced cost of sequencing and the realization that the reference genome is not perfect, researchers have shifted attention to create more population specific reference genomes (Sherman and Salzberg, 2020; Miga and Wang, 2021; Levy-Sakin et al., 2019; Sherman et al., 2019; Sengupta et al., 2021; Consortium et al., 2019; Karczewski et al., 2020). The availability of

these genomes together with high-quality transcriptomics data across diverse conditions will allow investigation into novel population-specific disease-related phenotypic variation (Martin et al., 2014; Daca-Roszak and Zietkiewicz, 2019; Jiang and Assis, 2020; Geoffroy et al., 2020).

8.1 References

Arendsee, Z., Li, J., Singh, U., Seetharam, A., Dorman, K. S., and Wurtele, E. S. (2019). phylostratr: a framework for phylostratigraphy. *Bioinformatics*, 35(19):3617–3627.

Chen, H., Gao, F., He, M., Ding, X. F., Wong, A. M., Sze, S. C., Yu, A. C., Sun, T., Chan, A. W.-H., Wang, X., et al. (2019). Long-read rna sequencing identifies alternative splice variants in hepatocellular carcinoma and tumor-specific isoforms. *Hepatology*, 70(3):1011–1025.

Clark, M. B., Wrzesinski, T., Garcia, A. B., Hall, N. A., Kleinman, J. E., Hyde, T., Weinberger, D. R., Harrison, P. J., Haerty, W., and Tunbridge, E. M. (2020). Long-read sequencing reveals the complex splicing profile of the psychiatric risk gene cacna1c in human brain. *Molecular psychiatry*, 25(1):37–47.

Consortium, G. et al. (2019). The genomeasia 100k project enables genetic discoveries across asia. *Nature*, 576(7785):106.

Daca-Roszak, P. and Zietkiewicz, E. (2019). Transcriptome variation in human populations and its potential application in forensics. *Journal of applied genetics*, 60(3):319–328.

Geoffroy, E., Gregga, I., and Wheeler, H. E. (2020). Population-matched transcriptome prediction increases twas discovery and replication rate. *Iscience*, 23(12):101850.

Hon, C.-C., Ramilowski, J. A., Harshbarger, J., Bertin, N., Rackham, O. J., Gough, J., Denisenko, E., Schmeier, S., Poulsen, T. M., Severin, J., et al. (2017). An atlas of human long non-coding RNAs with accurate 5 ends. *Nature*, 543(7644):199–204.

Jiang, X. and Assis, R. (2020). Population-specific genetic and expression differentiation in europeans. *Genome biology and evolution*, 12(4):358–369.

Kaminow, B., Ballouz, S., Gillis, J., and Dobin, A. (2020). Virtue as the mean: Pan-human consensus genome significantly improves the accuracy of rna-seq analyses. *bioRxiv*.

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443.

Klasberg, S., Bitard-Feildel, T., and Mallet, L. (2016). Computational identification of novel genes: current and future perspectives. *Bioinformatics and Biology insights*, 10:BBI–S39950.

Kuo, R. I., Cheng, Y., Zhang, R., Brown, J. W., Smith, J., Archibald, A. L., and Burt, D. W. (2020). Illuminating the dark side of the human transcriptome with long read transcript sequencing. *BMC genomics*, 21(1):1–22.

Levy-Sakin, M., Pastor, S., Mostovoy, Y., Li, L., Leung, A. K., McCaffrey, J., Young, E., Lam, E. T., Hastie, A. R., Wong, K. H., et al. (2019). Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nature communications*, 10(1):1–14.

Li, J., Singh, U., Bhandary, P., Campbell, J., Arendsee, Z., Seetharam, A. S., and Wurtele, E. S. (2021). Foster thy young: Enhanced prediction of orphan genes in assembled genomes. *bioRxiv*, pages 2019–12.

Lorenzi, L., Chiu, H.-S., Avila Cobos, F., Gross, S., Volders, P.-J., Cannoodt, R., Nuytens, J., Vanderheyden, K., Anckaert, J., Lefever, S., Tay, A. P., de Bony, E. J., Trypsteen, W., Gysens, F., Vromman, M., Goovaerts, T., Hansen, T. B., Kuersten, S., Nijs, N., Taghon, T., Vermaelen, K., Bracke, K. R., Saeys, Y., De Meyer, T., Deshpande, N. P., Anande, G., Chen, T.-W., Wilkins, M. R., Unnikrishnan, A., De Preter, K., Kjems, J., Koster, J., Schroth, G. P., Vandesompele, J., Sumazin, P., and Mestdagh, P. (2021). The rna atlas expands the catalog of human non-coding rnas. *Nature Biotechnology*.

Martin, A. R., Costa, H. A., Lappalainen, T., Henn, B. M., Kidd, J. M., Yee, M.-C., Grubert, F., Cann, H. M., Snyder, M., Montgomery, S. B., et al. (2014). Transcriptome sequencing from diverse human populations reveals differentiated regulatory architecture. *PLoS Genet*, 10(8):e1004549.

Miga, K. H. and Wang, T. (2021). The need for a human pangenome reference sequence. *Annual Review of Genomics and Human Genetics*, 22.

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., et al. (2021). The complete sequence of a human genome. *bioRxiv*.

Pertea, M., Shumate, A., Pertea, G., Varabyou, A., Breitwieser, F. P., Chang, Y.-C., Madugundu, A. K., Pandey, A., and Salzberg, S. L. (2018). CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome biology*, 19(1):208.

Raj, A., Wang, S. H., Shim, H., Harpak, A., Li, Y. I., Engelmann, B., Stephens, M., Gilad, Y., and Pritchard, J. K. Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. 5:e13328.

Rosenfeld, J. A., Mason, C. E., and Smith, T. M. (2012). Limitations of the human reference genome for personalized genomics. *PloS one*, 7(7):e40294.

Ruiz-Orera, J. and Albà, M. M. (2018). Translation of small open reading frames: Roles in regulation and evolutionary innovation. *Trends in Genetics*, 2(5):890.

Ruiz-Orera, J., Hernandez-Rodriguez, J., Chiva, C., Sabidó, E., Kondova, I., Bontrop, R., Marqués-Bonet, T., and Albà, M. M. (2015). Origins of de novo genes in human and chimpanzee. *PLoS Genetics*, 11(12):e1005721.

Scalzitti, N., Jeannin-Girardon, A., Collet, P., Poch, O., and Thompson, J. D. (2020). A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms. *BMC genomics*, 21:1–20.

Sengupta, D., Choudhury, A., Fortes-Lima, C., Aron, S., Whitelaw, G., Bostoen, K., Gunnink, H., Chousou-Polydouri, N., Delius, P., Tollman, S., et al. (2021). Genetic substructure and complex demographic history of south african bantu speakers. *Nature communications*, 12(1):1–13.

Sharon, D., Tilgner, H., Grubert, F., and Snyder, M. (2013). A single-molecule long-read survey of the human transcriptome. *Nature biotechnology*, 31(11):1009–1014.

Sherman, R. M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., Boorgula, M. P., Chavan, S., Vergara, C., Ortega, V. E., et al. (2019). Assembly of a pan-genome from deep sequencing of 910 humans of african descent. *Nature genetics*, 51(1):30–35.

Sherman, R. M. and Salzberg, S. L. (2020). Pan-genomics in the human genome era. *Nature Reviews Genetics*, 21(4):243–254.

Singh, U., Hernandez, K. M., Aronow, B. J., and Wurtele, E. S. (2021a). African americans and european americans exhibit distinct gene expression patterns across tissues and tumors associated with immunologic functions and environmental exposures. *Scientific Reports*, 11(1):1–14.

Singh, U., Hur, M., Dorman, K. S., and Wurtele, E. S. (2020). Metaomgraph: a workbench for interactive exploratory data analysis of large expression datasets. *Nucleic Acids Research*, 48(4):e23–e23. gkz1209.

Singh, U., Li, J., Seetharam, A., and Wurtele, E. S. (2021b). pyrpipe: a Python package for RNA-Seq workflows. *NAR Genomics and Bioinformatics*, 3(2). lqab049.

Singh, U. and Wurtele, E. S. (2021). orfipy: a fast and flexible tool for extracting ORFs. *Bioinformatics*. btab090.

Sondka, Z., Bamford, S., Cole, C. G., Ward, S. A., Dunham, I., and Forbes, S. A. (2018). The cosmic cancer gene census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer*, page 1.

Sun, Y. H., Wang, A., Song, C., Shankar, G., Srivastava, R. K., Au, K. F., and Li, X. Z. (2021). Single-molecule long-read sequencing reveals a conserved intact long rna profile in sperm. *Nature communications*, 12(1):1–12.

Tautz, D. and Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nature Reviews Genetics*, 12(10):692–702.

Wong, K. H., Ma, W., Wei, C.-Y., Yeh, E.-C., Lin, W.-J., Wang, E. H., Su, J.-P., Hsieh, F.-J., Kao, H.-J., Chen, H.-H., et al. (2020). Towards a reference genome that captures global genetic diversity. *Nature communications*, 11(1):1–11.

ProQuest Number: 28413590

INFORMATION TO ALL USERS
The quality and completeness of this reproduction is dependent on the quality
and completeness of the copy made available to ProQuest.

Distributed by ProQuest LLC (2021).
Copyright of the Dissertation is held by the Author unless otherwise noted.

This work may be used in accordance with the terms of the Creative Commons license
or other rights statement, as indicated in the copyright statement or in the metadata
associated with this work. Unless otherwise specified in the copyright statement
or the metadata, all rights are reserved by the copyright holder.

This work is protected against unauthorized copying under Title 17,
United States Code and other applicable copyright laws.

Microform Edition where available © ProQuest LLC. No reproduction or digitization
of the Microform Edition is authorized without permission of ProQuest LLC.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346 USA