DATA AND -OMICS-DRIVEN APPROACHES TO UNDERSTAND THE HEAT STRESS RESPONSE:

THE DEVELOPMENT OF SCALABLE TOOLS AND METHODS TO DRIVE HYPOTHESIS GENERATION

by

Allen Henry Hubbard, Jr.

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Bioinformatics and Systems Biology

Spring 2018

DATA AND -OMICS-DRIVEN APPROACHES TO UNDERSTAND THE HEAT STRESS RESPONSE:

THE DEVELOPMENT OF SCALABLE TOOLS AND METHODS TO DRIVE HYPOTHESIS GENERATION

by

Allen Henry Hubbard, Jr.

Approved: _____

Limin Kung, Jr., Ph.D.
Chair of the Department of Animal and Food Science

Approved: _____

Mark W. Rieger, Ph.D.
Dean of the College of Agriculture and Natural Resources

Approved: _____

Ann L. Ardis, Ph.D.
Senior Vice Provost for Graduate and Professional Education

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Carl J. Schmidt, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Abhyudai Singh, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Eric H. Lyons, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Hagit Shatkay, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Shawn W. Polson, Ph.D.
Member of dissertation committee

ACKNOWLEDGMENTS

I could not have done this work without the support of my lab-mates, friends and family. Special thanks to my advisor, Dr. Carl Schmidt, and IT staff member Gregory Keane for his technical support. Also, great thanks to Heidi Van Every for her support and editorial skills.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This dissertation concerns the broad computational challenges that face labs in the -omics era, in the service of addressing a major agricultural goal – adapting the broiler chicken to heat stress. Its contributions span creation of scalable tools to process raw sequencing reads to statistical methods that integrate multi-omics data and produce novel biological insight. I will present the paradigm for an architecture of powered-by-CyVerse tools, which is leveraged to power the tool fRNAkenseq. CyVerse is a pioneering cyberinfrastructure project to make large scale computing and storage resources accessible to domain scientists and provide a way for tools to share data with one another. fRNAkenseq, a platform for comprehensive analysis for RNA-seq from FastQ to differential expression, relies on CyVerse for cloud-based storage, a grid computing approach, and the ability to access the 30,000 reference genomes curated by the powered-by-CyVerse tool CoGe. fRNAkenseq is among the first of its kind in third party software to leverage CyVerse in such a fashion. To move from data to insight we have developed pipelines and strategies to integrate the complex, tissue rich datasets produced from fRNAkenseq with supplementary metabolomics data. From this data, we generate biological hypotheses and models that extend understanding of the regulation of the heat stress response. In particular, these hypotheses provide context for the co-regulation of sulfur, lipid, and sugar metabolism essential to maintaining homeostasis in the face of heat challenge.

Chapter 1

INTRODUCTION

1.1 Context

Within the next few decades, the impacts of climate change are anticipated to substantially impact poultry livestock yield, partly due to increased frequency of heat waves leading to higher bird mortality and decreased feed efficiency (Rojas-Downing, 2017). Broiler chickens, or lines of chicken raised for meat production, are a mainstay of the global food supply. As in many livestock species, there is an ongoing tradition of intense artificial selection for valuable commercial traits in the modern broiler chicken. These efforts have resulted in a consistent 2-3 percent improvement in the efficiency of broiler meat production per year (McKay, 2009). However, an unintended consequence of this regimen of focused breeding for high-muscle, rapidly growing phenotypes includes an increase in the incidence of skeletal and cardiovascular problems among other serious metabolic issues (Julian, 1998). Some of the negative aspects of the altered physiology resulting from artificial selection are hypothesized to relate to disruption of cellular metabolic systems (Tallentire et al., 2016). Thus, targeted adaptation of broiler chicken to heat stress conditions associated with accelerating trends of climate change will require a stronger understanding of genetic regulation of biochemistry than that which currently exists. Computational approaches are needed to harness the power of increasingly large-scale omics data that will power the rapid breeding for complex traits. This will be an important agricultural objective for genomics and bioinformatics, in the 21st Century.

We have sought to advance this goal by developing novel computational tools and statistical analyses that make it possible for researchers to elucidate regulation of the heat stress response, using modern, high throughput and data intensive techniques. This strategy has required us to develop methods and tools to handle the large datasets associated with our experiments. To do this, we first had to develop the computational infrastructure to handle the transcriptome data associated with large-scale heat stress experiments. This is important, as the resulting volume of data posed a unique informatics challenge both in terms of managing data bottlenecks and identifying causal mechanisms from resulting high dimensional datasets. Such types of problems are not unique to poultry genomics and will intensify as high throughput datasets become increasingly common, for example as advances in genetic engineering encourage combinatorial biology experiments (Zhao *et al*., 2017) and large-scale single-cell sequencing becomes more popular. Anticipating such challenges, this thesis develops concrete computational solutions to challenges associated with data-driven life sciences research, and proposes novel regulatory mechanisms of the heat stress response in the broiler chicken to advance the poultry genomics community. The accomplishments of this thesis include:

- Creation of fRNAkenseq, a user-friendly platform and interface for analysis of transcriptome data that exploits cross-talk with other tools
- Deployment of a statistically sound approach for identifying tissue specific genes, sensitive enough to clarify biology unique to different muscle types (cardiac and skeletal) in fRNAkenseq processed data.

- Development of pipelines that integrate multiple statistical approaches to merge transcriptome and metabolome data, to drive hypothesis generation relating heat stress regulation across the two –omics.
- Linear modeling to identify relationships between metabolites that signify junctions, or forks, between closely related pathways whose regulation shifts under heat stress. Sets of these forks are merged into pathways that provide novel biological insight into sulfur, lipid and sugar metabolism.

1.2 RNA-seq

We have chosen to use transcriptomics to understand the heat stress response in chicken, using high throughput RNA-seq. RNA-seq is a powerful tool for systems biology based projects because it is capable of measuring expression for all transcribed regions of an organism's genome (Nagalakshmi *et al*., 2008). Thus, it provides an excellent lens by which to understand how a treatment or conditions influences a biological system. Because the heat stress response involves many genes and pathways, high throughput methods such as RNA-seq are well suited to

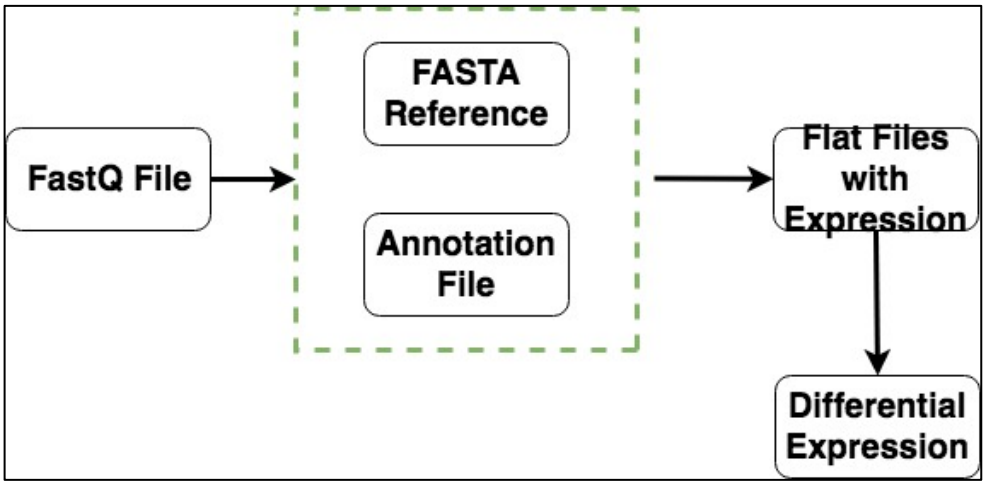

Figure 1: Schematic of workflow for RNA-seq informatics.

uncovering novel regulation that underlies adaptations to this stress. Because changes in gene expression can reflect protein levels, RNA-seq can provide insight into how the regulatory systems of a cell are changing (Maier *et al*., 2009). However, RNA-seq experiments pose unique informatics challenges that require statistical and computational sophistication to effectively process data and gain biological insight (Chu and Corey, 2012).

Typically, the laboratory-based work involved in an RNA-seq experiment and the subsequent sequencing results in the production of a FastQ file. Each FastQ file in our dataset of Illumina short reads contains millions of small fragments (25-50 base pairs) of nucleotide sequences corresponding to fragments of RNA molecules from the transcriptome. Ultimately, each fragment of RNA, or read, in the FastQ file must be mapped to a gene or other genomic feature by aligning the read to its best match in the genome. It is only by determining the appropriate feature each read is associated with

that it becomes possible to quantify expression across genes and other genomic regions.

A number of algorithms exist (Conesa *et al*., 2016) to accomplish these tasks, with many workflows using a series of algorithms to process a single FastQ file. Pipelines typically begin by first using one algorithm to identify the best fit in a reference genome for each read in a FastQ file. This is accomplished during the mapping step of the pipeline. Though there a number of alignment algorithms, most store the alignment information in a file known as a BAM (binary alignment file). The BAM is merely a more efficient, compressed version of a sequence alignment file (SAM) file, which is human readable and stores alignment information and important metrics in a standardized format. Subsequent steps in an RNA-seq informatics pipeline typically use an additional set of algorithms to process alignment files and produce table of either raw read counts for each genomic feature or counts which have been normalized for both the length of the gene and the number of reads in a FastQ file, i.e. Fragments Per Kilobases per Million Reads Mapped (FPKM). The output of this quantification step is often a tab delimited or similarly formatted flat file. These tables provide quantified expression levels for each feature and are frequently used for further statistical analyses. While our research is mostly focused on measuring mRNA expression, read alignments can also detect expression of non-coding RNAs and is not limited to annotated features (Tripathi *et al*., 2017).

After mapping and quantification, downstream statistical analysis attempts to link changes in gene expression to the influence of a treatment. This is often accomplished through the detection of differentially expressed genes. Differentially expressed genes are those whose expression levels differ significantly between control

and treatment conditions. They may influence the levels of enzymes and other proteins critical to controlling tissue-important biology. There are a number of software packages that provide statistical techniques to identify genes whose expression patterns change under treatment. Many of these programs run in the R computing environment and accept as input the number of raw counts produced during read quantification. We have incorporated several of these tools into our pipelines and software that we will use to analyze the heat stress response: edgeR (Robinson *et al*., 2010), DESeq2 (Love *et al*., 2014) and BaySeq (Hardcastle *et al*., 2010).

Each of the packages for differential expression assumes that RNA-seq reads follows a negative binomial distribution (Seyednasrollah *et al.,* 2013). The negative binomial distribution is a modification of the Poisson distribution, intended to take into account the experimental noise that increases the variance of lowly expressed genes, resulting in over-dispersion (Anders and Huber, 2010). However, each tool uses a unique set of assumptions to estimate the parameters of the negative binomial distribution. None has been proven to be universally superior, as the landscape of variation among an experiment will determine which assumptions are most appropriate (Rapaport *et al.,* 2013). Thus, a robust workflow to detect differentially expressed genes should consider the output multiple of multiple prediction tools. Developing the infrastructure and pipelines to rapidly process transcriptome data are but some of the challenges associated with using RNA-seq on a large scale.

1.3 Data Burden to fRNAkenseq Software

In addition to the complexity of developing effective workflows, handling the necessary computational tasks for RNA-seq experiments can be daunting due to the amount of data leveraged at each step. This burden grows with the number of FastQ

samples analyzed. Additionally, many standard workflows require management of files that contain organism specific genome data, against which reads will be aligned (Conesa *et al*, 2016). These, also, can be quite large. For example, the genome sequence against which each read from a FastQ file must be aligned can be one to several GBs, with annotation flat files being somewhat smaller. Genomic reference sequences used for read alignments are stored as Fast-All (FASTA) files that can be procured from large repositories such as the National Center for Biotechnology Information (NCBI) or databases associated with Ensembl, a partnership between the European Bioinformatics Institute and the Wellcome Trust Sanger Institute. The reference FASTA sequences require indexing prior to incorporation in an RNA-seq pipeline, and are often accessed by researchers on as-needed basis using NCBI or Ensembl file transfer protocol (ftp) resources. The data burden of transcriptome analysis, however, is intensified and primarily driven by large sample sizes of FastQ files, as reference files are typically re-used for multiple informatics analyses.

During the course of our studies using RNA-seq to characterize the heat stress response, we have processed over 1,500 RNA-seq library samples. Fully investigating the heat stress response in the chicken, to fulfill the scope of our grant, has required us to manage multiple heat stress trials under various conditions. This experimental design typically produces dozens to hundreds of RNA-seq libraries per trial.

As in many labs, the rate of data production quickly eclipsed our capacity for downstream analysis. For example, a typical ~9 Gigabytes (GB) uncompressed RNA-seq file often requires up to two hours for mapping of reads to a reference genome, merely to generate a binary alignment file (BAM), and an additional 45 to 60 minutes time for quantification of reads in order to provide human interpretable gene level

expression data. This computation time has been much higher in the past, with older-generation algorithms. Relying on the latest generation of mappers has helped reduce this computational time. Nevertheless, we have found the scope of data associated with experiments of this scale can quickly overwhelm in-house resources. While faster algorithms map and quantify reads more rapidly, this often means that issues with managing data burdens emerge more quickly. BAM files from our dataset are up to 12-18 Gigabytes (GBs), and their uncompressed versions, sequence alignment (SAM) files, can be far greater at nearly 50 GBs. The need to compile and organize data on this scale is thus a serious computational challenge. However, this is only the first stage of many informatics challenges that a lab invested in high throughput genomics will face.

Once sequencing files have been processed from raw reads and to produce gene expression tables, researchers must identify genes and the associated biological systems that are altered during the course of treatment. These subsequent challenges, often highly statistical in nature, require robust differential expression analysis as well as means to identify pathways whose genes demonstrate expression changes change under experimental treatments. With each heat stress trial, we had to address these, and other, computational hurdles. They quickly emerged as bottlenecks once we streamlined lab protocol and could produce data quicker than we could process it. Our initial strategy of using simple manual commands to process data in the preliminary stages of our study was rapidly overwhelmed by the burden of data. Thus, we ultimately had to develop novel computational strategies to make our analyses more scalable and effectively manage our datasets.

We initially developed pipelines that could automate steps for mapping and quantification, first using Tophat2 (Trapnell *et al*., 2012) and Cufflinks (Trapnell *et al*., 2012) for these steps, respectively. Tophat and Cufflinks represent the mapping and quantification algorithms, respectively, of the popular Tuxedo suite of RNA-seq analysis tools (Trapnell *et al*., 2012). As technology advanced to more rapidly process larger datasets, we would substitute elements of the Tuxedo suite with algorithms from the next generation of alignment and quantification tools. Hisat2 (Kim *et al*., 2015) replaced TopHat2 as the alignment algorithm in our pipelines, and the quantification algorithm Stringtie (Pertea *et al*., 2015) would replace Cufflinks.

In order to encourage workflow continuity, RNA-seq mapping algorithms have adhered to standards of representing alignment information in the standardized SAM/BAM formats. This convention has encouraged the evolution of our workflows from pipelines to software. While input and output requirements for different generations of algorithms may vary in regards to their preference for SAM/BAM files, the software Samtools (Heng *et al*., 2009) can convert between SAM/BAM files as well as execute file indexing and other file pre-processing as needed. Thus, there are tools such as featureCounts (Li *et al*., 2014) and HTSeq-count (Anders *et al*, 2014) that can process arbitrary SAM/BAM files into tab-delimited tables containing genes and read counts, regardless of the algorithm used upstream to process the FastQ files. For the sake of minimizing computation time our pipelines currently use a python script that parses StringTie output to produce tables of raw counts for downstream differential expression. However, as alignment algorithms continue to develop, these steps can be updated accordingly with suite-specific software, or replaced with an aligner-agnostic program such as featureCounts and HTSeq-count.

Thus, it possible to use software packages in the R computing environment for differential expression analysis pipelines, even as alignment algorithms continue to evolve. Having established workflows for mapping and quantification, we eventually developed separate pipelines to execute differential expression in a robust fashion by executing multiple differential expression algorithms on the same dataset. These workflows evolved from Perl scripts that automated the writing of R scripts for differential expression analyses in DESeq2, edgeR and BaySeq. In order to provide biologists in the lab control over their analysis as we continued to produce volumes of data, we developed a web-base graphical user interface (GUI) using PHP:hypertext processor (PHP) to manage the pipelines. This made it possible such that certain bioinformatics analyses, such as the mapping and quantification of reads in addition to differential expression, could be controlled by biologists with no prior computational experience.

The result of this effort was the first version of the RNA-seq analysis platform fRNAkenseq. fRNAkenseq encompasses differentiated workflows that cover all major steps, from FastQ file to differential expression, of RNA-seq analysis. fRNAkenseq was initially developed to address the informatics needs resulting from the high number of heat stress and other treatment experiments being done by multiple graduate and undergraduate students. In addition to aiding the several manuscripts in publication, analysis of datasets using fRNAkenseq contributed to the papers: "Transcriptomic changes throughout post-hatch development in Gallus gallus pituitary" (Pritchett *et al,* 2016), "RNA-seq:Primary Cells, Cell Lines and Heat Stress. Cytogenetic and Genome Research" (Schmidt, 2015), "Chicken Hepatic Response to Chronic Heat Stress Using Integrated Transcriptome and Metabolome Analysis",

(Jastrebski *et al*., 2017). Several other manuscripts from our large scale RNA-seq datasets are in preparation.

1.4 Integrating Bioinformatics APIs

Ultimately, fRNAkenseq continued to evolve from early in-house pipelines that allowed us to manage initial data bottlenecks, and developed into a complete software platform to keep pace with the informatics bandwidth demanded by our samples. In the course of fRNAkenseq's development, we partnered with collaborators who provided access to computing and storage resources, which became crucial to expanding fRNAkenseq's computational abilities. This was accomplished through leveraging a succession of bioinformatics focused application programming interfaces (APIs).

As computational tools for server based bioinformatics have proliferated, communities of developers have sought to improve the utility of their tools and avoid redundancy of future development, by facilitating data and other resource sharing across software platforms. This is often accomplished by web application programming interfaces (APIs) being leveraged as a way to exchange information between servers in a standardized fashion. In a bioinformatics context, web APIs can be used to handle requests for sequencing files or other, more complex tasks that are necessary for the execution of pipelines. The rapidly evolving class of bioinformatics APIs include those that allow users to access genome files from existing tools or resources and databases, such as the Ensembl REST API (Yates et al., 2015), as well as APIs allowing large batch submission to text mining tools such as The Database for Annotation, Visualization and Integrated Discovery (DAVID) and the Gene Ontology (GO) database, and many similar tools.

However, a number of active software development teams are also producing data sharing and foundational APIs that can power future bioinformatics software development in a comprehensive fashion. Some of these efforts include the Breeding API (BrAPI), (http://docs.brapi.apiary.io/) developed to encourage plant genomics software to leverage cross-talk between the community of diverse plant sequence databases, as well as the job and data-movement management **A** **G**rid **A**nd **V**isualization **E**nvironment (Agave) API suite (Dooley *et al*., 2012). The Agave API allows developers to turn pipelines and workflows into Agave Apps, as well as develop a network of registered storage and execution machines. Agave Apps comprise shell scripts and Javascript Object Notation (JSON) files that register the App with the Agave service. The Agave APIs also include commands to move data across storage and execution machines, on the multiple GB scale commonly associated with bioinformatics datasets. Thus, Agave makes it possible to effectively develop what is known as a "grid computing" scheme for bioinformatics workflows, with different pipelines and data allocated to multiple storage and execution resources. This type of architecture enabled by Agave is critical for fRNAkenseq, as it enables reliable job and data management across machines.

Importantly, Agave enables developers to register arbitrary cloud resources as storage systems to serve as a repository for data processed by Agave Apps. This feature enhances the utility of Agave for bioinformatics developers, who can use external storage resources (cloud, private, or publicly available servers such as the **T**exas **A**dvanced **C**omputing **C**enter (TACC)) to ameliorate the burden of data volume produced by next-generation sequencing workflow. During the course of its development, fRNAkenseq expanded from a web interface that managed RNA-seq

pipelines in the form of shell scripts, to a tool that fully harnessed APIs developed by the Agave team and other software development groups associated with the NSF-sponsored cyberinfrastructure project CyVerse. fRNAkenseq is one of the first tools to incorporate these resources in this fashion to manage large scale next generation sequencing datasets. Ultimately, fRNAkenseq has evolved to use the Agave APIs to fully federate with CyVerse. Only by completely integrating with Agave and other bioinformatics APIs in a novel fashion, could fRNAkenseq enable our lab to keep pace with analyzing the volume of sequencing data we regularly produced, without overwhelming existing resources.

1.5 Using API Driven Development and Downstream Analysis

The first chapter of the thesis will present the paradigm for an architecture of powered-by-CyVerse tools demonstrated by the RNA-seq analysis platform fRNAkenseq. CyVerse is a pioneering cyberinfrasture project to make large scale computing and storage resources accessible to domain scientists and provide a way for tools to share data with one another. CyVerse both provides bioinformatics tools, as well as resources for the community of developers who wish to exploit CyVerse resources to develop their own tools. fRNAkenseq, as a platform for comprehensive analysis for RNA-seq from FastQ to differential expression, relies on CyVerse for cloud-based storage as well as simple grid computing, using the Agave APIs to move data across execution and storage machines. Though currently managed by a development team independent from CyVerse, the Agave APIs make it possible to incorporate CyVerse resources into a robust grid computing structure (using TACC allocations for execution machines, and the CyVerse Data Store for cloud based

storage) as well as gain authentication to shared data with various powered-by-CyVerse tools. Taking advantage of CyVerse's centralized authentication system,

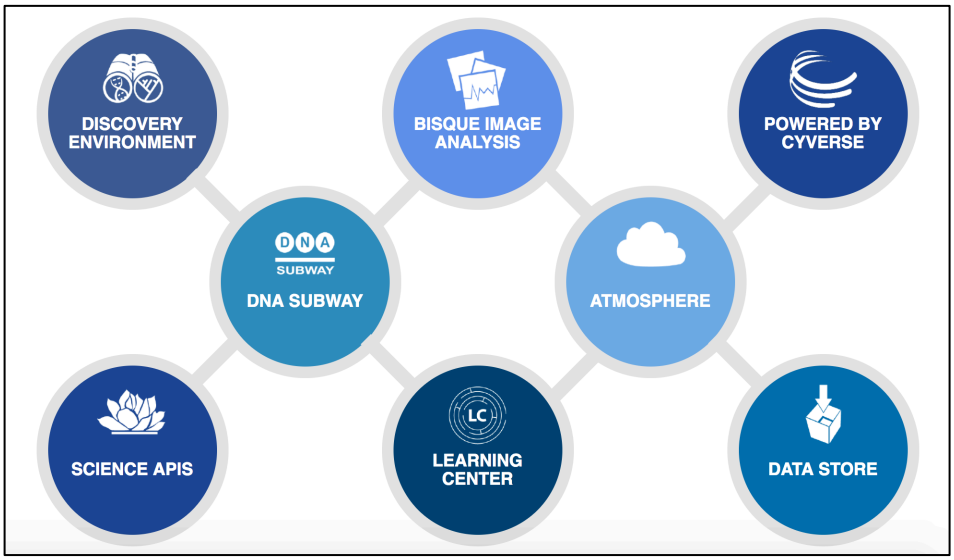

Figure 2: CyVerse offerings include the Science APIs. fRNAkenseq exploits the science APIs to become a powered-by-CyVerse tool in order to communicate with the CyVerse Data Store. fRNAkenseq represents a novel extension of CyVerse resources (CyVerse.org).

fRNAkenseq also has access to over 30,000 reference genomes curated by another powered-by-CyVerse tool, the Comparive Genomics Platform (CoGe). CoGe is an engine for comparative genomics, which leverages a large-scale database to store genomes for syntenic analysis between different organisms (Lyons *et al*., 2008). CoGe also offers a set of **Re**presentational **S**tate **T**ransfer (RESTful) APIs that allow users to access genomes stored in its system, and has expanded to offer various resources and pipelines that visualize and analyze next generation sequencing data.

fRNAkenseq exploits these, making CoGe genomes accessible to fRNAkenseq pipelines.

CoGe also enables users to upload researcher provided genomes, thus making it a vehicle for users to bring draft genomes of non-model organisms into CyVerse. These personally uploaded genomes, owing to authentication with a CyVerse Oauth2 token upon logging into fRNAkenseq, can then be accessed by fRNAkenseq. Though users can now use CoGe genomes for analysis in fRNAkenseq, the initial permission settings of the genome when uploaded to CoGe are maintained. The ability of users to supply their own genome to fRNAkenseq through CoGe is important, as tools for run bioinformatics analyses on organisms that currently lack a high quality draft genome has been identified as an important need by the burgeoning comparative genomics community (Mykles *et al*, 2016). Using CyVerse to facilitate a data sharing relationship with CoGe fulfills a crucial need for the bioinformatics community, by demonstrating how third party tools such as fRNAkenseq can simultaneously leverage multiple CyVerse resources to meet diverse bioinformatics needs. fRNAkenseq's design has allowed it to fully exploit API resources offered by both CyVerse and CoGe to execute important bioinformatics tasks.

fRNAkenseq has been instrumental in managing the data burden associated with our lab's large-scale transcriptome studies, and represents a valuable contribution to the research community. It is accessible to any CyVerse user at the address http://raven.anr.udel.edu. In terms of novelty, fRNAkenseq is among the first of its kind in third party software to leverage CyVerse in such a complete fashion, using the APIs of CyVerse and CyVerse associated tools at a foundational level. Although programs such as SciApps (Lu *et al*., 2016) use Agave APIs to exploit the CyVerse

Data Store and develop workflows as Agave Apps, fRNAkenseq achieves novelty by also easily exchanging data with the powered-by-CyVerse tool CoGe. Other powered-by-CyVerse tools, such as Galaxy, do not use Agave Apps at a foundational level, exploiting CyVerse resources mainly for its cloud data storage. fRNAkenseq's differential expression capabilities are also useful to biologists. The differential expression pipeline of fRNAkenseq runs three differential expression prediction tools (edgeR, DESeq2, and BaySeq), providing a user with output tables with combined

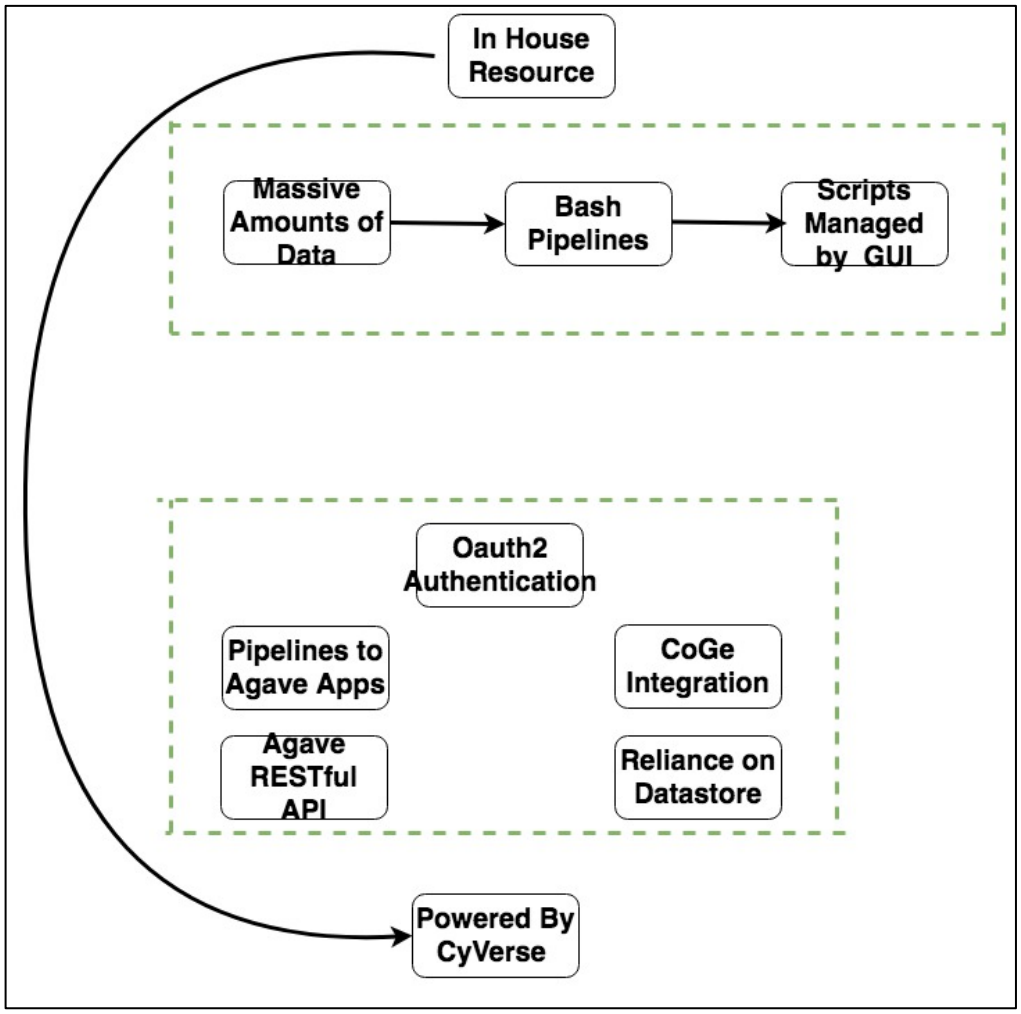

Figure 3: Several features involved in evolution of fRNAkenseq from pipelines, to an in-house resource to a powered by CyVerse tool.

results, separating out enriched genes according to various levels of stringency – significant according to none, one, two or all three enrichment algorithms. The need for pipelines that approach differential expression analysis by using multiple algorithms has motivated the development other software packages, such as RNA-seq

GUI (Russo and Angelini, 2014). However, this program RNA-seq GUI, which is R based, does not encompass mapping and quantification, and must be installed locally.

fRNAkenseq advanced the utility of both CyVerse and CoGe in a novel fashion, allowing CyVerse users the ability to use CoGe genomes on fRNAkenseq RNA-seq pipelines. Owing to the combined features of relying on Agave APIs to manage pipelines, incorporating complex workflows for robust differential expression, as well as leveraging access to CoGe genomes, fRNAkenseq represents a novel bioinformatics tool and demonstrates the potential of powered-by-CyVerse resources. Ultimately, fRNAkenseq and its capacity to effectively accomplish rapid RNA-seq analysis, made it possible for our lab to advance from the preliminary steps of read mapping and quantification to differential expression analysis. However, fRNAkenseq is only a partial solution to the informatics challenges of gaining a systems-biology level understanding of the heat stress response from high throughput data. The processed data must next be mined an integrated into biologically useful models. It is by proposing statistical approaches to analysis downstream of fRNAkenseq data that I have contributed novel biological understanding of the heat stress response in broiler chicken.

1.6 Tools to Proceed From Datasets to Insight

This next phase, moving from data to insight, involves developing additional pipelines and workflows that integrate the complex, tissue rich datasets produced from fRNAkenseq with supplementary metabolomics data. There are critical limitations, for example, of relying on only transcriptome data. Though valuable for the ability to provide expression data for all annotated genes (over 20,000 in the chicken), transcriptome data cannot provide more than circumstantial evidence into biochemical

shifts at the level of metabolites and other compounds. While increasing the dimensionality of datasets, integrating multiple –omics data can improve the biological insight to be gained from high-throughput studies. This gain in dimensionality, resulting from the increase of data points associated with additional data types, can be managed by developing analysis pipelines that more effectively identify biological regulation.

As our laboratory continued to process transcriptome data with the expanded bandwidth of fRNAkenseq, we sought to enhance the biological insight gained from the dataset by supplementing RNA-seq with metabolite data, while also exploiting the diversity of tissues in the dataset to explore organ specific biology. This further analysis required the development of additional pipelines, and the application of statistical algorithms in a novel workflow.

These pipelines begin by leveraging a definition of tissue enrichment. This work will be discussed in detail in Chapter 3. Tissue enrichment involves identifying genes whose expression in a tissue of interest is increased relative to the background samples. Tissue enrichment can be useful to identify modules of genes that control organ-defining physiology. However, finding a strategy of enrichment that is sufficiently stringent as well as sensitive to organ unique biology can be a challenge. One of the first large-scale examples of studies leveraging tissue enrichment to investigate organ-defining biology is the genotype-express (GTEx) pilot study (The GTEx Consortium, 2015). The GTEx protocol identifies enriched genes as those having an expression level that is fivefold greater in the organ of interest compared to background samples. However, we demonstrate that this procedure produces modules of tissue defining genes that often overlap with one another. By leveraging a stringent

z-score test, in which $\frac{\overline{x}(tissue\ interest) - \mu(background)}{\sigma(background)}$ must be greater than five, we demonstrate the ability to identify tissue-defining genes that are unique to each organ, resulting from a large RNA-seq dataset processed by fRNAkenseq. While using such a stringent threshold for tissue-defining genes may eliminate weakly enriched transcripts, focusing on the most robustly expressed genes will be useful for identifying those associated with critical biology. This approach to tissue enrichment is used to explore transcriptome differences between cardiac and skeletal breast muscle.

Understanding the transcriptome aspects of muscle specific types is important because mass of breast muscle has been an important target of artificial selection during the development of the modern broiler chicken (Tallentire *et al*., 2016). In addition to elucidating the heat stress response, clarifying the genetic basis of the resulting physiological changes in muscle has been a major goal of our lab. This research is related to understanding the broiler heat stress response, as we hypothesize that metabolic shifts associated with artificial selection have impacted bird tolerance for heat stress. We shed light on several muscle-related transcription factors whose enrichment patterns differ between breast muscle and cardiac tissues. Understanding the transcriptome underpinnings of muscle type specific physiology provides insight into the developmental biology of the broiler. This is important, as many of these systems critical to muscle development are altered under heat stress. This type of work provides context for them in terms of development.

The large volume of data exploited by fRNAkenseq makes it possible to identify tissue specific biology and emphasize tissue enriched genes for downstream analyses. For example, the identifying of tissue specific genes, providing a form of

feature selection before combining transcriptome and metabolome data into statistical learning pipelines that isolate biomolecules strongly associated with the heat stress response. The term biomolecules is used to refer to both metabolites and genes. Subsequent analysis that will integrate metabolomics and transcriptome data will be from the perspective of tissue defining genes identified through our threshold.

For this step, we leverage an initial step of k-means to separate out compounds by expression patterns, followed by random forest to identify among each cluster those compounds with the strongest ability to classify control and heat stress samples. A final round of principal component analysis (PCA) among these strong biomarkers was able to recapitulate elements of known biological pathways that function under heat stress, as well as propose novel hypotheses relating metabolites and genes not previously connected to these networks. Each algorithm in our pipeline has unique roles in exploiting a different feature of the data. For example, after initial separation by k-means the signature of heterogeneity is exploited by prioritizing compounds with strong classifying power through random forests, and finally by summarizing correlations among compounds by principal components, these pipelines identify potential regulators of heat stress and shed light on systemic metabolic changes. Importantly, this pipeline effectively reduces our dataset to a few compounds representing the key systems involved in the heat stress response: lipid, sugar and antioxidants. The reduction of the set of all possible metabolites and organ specific genes is a critical form of dimension reduction that prioritizes compounds relevant to the heat stress response. This work is described in Chapter 4.

Finally, many of the molecules prioritized by the statistical learning techniques are organized into possible mechanisms that represent concrete biological hypotheses

by building linear models that model metabolite levels in terms of the ratios of other metabolites and identify those that are potentially metabolically related. We use metabolite data to create linear models, as this class of model best fits the data and recapitulates biologically verifiable relationships. The use of ratios to account for behavior of metabolic pathways emerged as a technique to capture the kinetic information represented by relative levels of metabolites at steady state. It has been used in maize (Haries *et al*., 2009) and human studies to produce SNP-metabolite associations (Gieger *et al*., 2008). We extend these methods to identify metabolite-metabolite models that are influenced by the heat stress response. Though there has been progress in using pathway information to integrate SNP and ratio associations (Krumsiek *et al*., 2016), efforts to build complete circuits from metabolomics data have been lacking. It is for this reason that we progress to develop circuits of regulation from these models. The procedure to do this is simple, but effective and provides a form of analysis complementary to our statistical learning pipelines. The interaction terms of several of many linear models containing heat stress responsive metabolites are highly significant, and many have overlapping components. Potential regulatory circuits are built by identifying sets of models of with overlapping components that could represent different regions of a pathway.

Those linear models involving biochemically related compounds, which demonstrate significant interaction terms, could represent the differential direction of metabolic fates in ways that shift between control and heat stress conditions. We call these simple mechanisms “metabolic forks”, in which precursor molecules can be selectively routed to different compounds under regulation. These network motifs are then integrated into larger circuits that provide insight into pathways operate in

coordinated fashion under heat stress, and put in the context of gene expression changes from the transcriptome data. This is critical, because the search time would be computationally expensive, and having narrowed down a set of candidate molecules enables informed selection to build models, dramatically reducing the number of possible relationships that must be evaluated.

The use of linear models to integrate metabolite data into concrete mechanisms is a critical insight, because these ratios capture biochemical information about circuits that would not be detectable otherwise. We will show that linear models capture differential behavior of pathways that sit at the intersection of lipid and sulfur metabolism. It can be seen that certain network relationships would not be identified without the incorporation of metabolites ratios into the models. In particular, the relationship between various sulfur containing species and lipids would not be recognized, without modeling this relationship as a metabolic fork utilizing ratios. The poultry research community has demonstrated active interest in understanding how the levels of such metabolites and their interactions shift under heat stress. In particular, identifying metabolites that influence anti-oxidant production could provide candidates for potentially powerful dietary interventions to improve bird performance (Sahin *et al*., 2013). Importantly, many of the metabolites identified by our analysis are located in pathways adjacent to genes whose expression is significantly changed by heat stress. The putative circuits from these models produce hypotheses that have been validated through literature searches, independently run experiments, or plan on being explored through future feed supplementation studies. This work, detailed in Chapter 4, focuses on samples of liver tissue, as the liver is a metabolic powerhouse for the bird and central to managing sugar and fat metabolism, and influencing

peripheral tissues (Jastrebski *et al*., 2017). This comprehensive approach to explore the heat stress response effectively recapitulates known biology, and also proposes new hypotheses to guide breeding and dietary interventions that could improve bird heat stress tolerance.

1.7 Conclusion and the Way Forward

The work described in this thesis, by producing tools and pipelines for the effective analysis of high throughput sequencing data, has made useful and novel contributions to the life and computational sciences. The powered-by-CyVerse RNA-seq analysis platform, fRNAkenseq, has furthered the capacity of the cyberinfrastructure project CyVerse, by expanding offering of powered-by-CyVerse tools. Additionally, the data sharing capacities across CyVerse resources demonstrated by fRNAkenseq's API driven architecture provides a useful blueprint for developing the next generation of informatics tools, which will need to seamlessly exchange data in order to keep pace with the increasing volume of high throughput datasets. Our lab has used fRNAkenseq to process RNA-seq files on a large scale. These transcriptome files have subsequently been analyzed to produce novel findings regarding tissue-defining biology and regulation of the heat stress response. Additionally, I have developed statistical pipelines that integrate transcriptome files analyzed by fRNAkenseq with supplementary metabolomics data. By doing so, I propose a novel circuit that integrates sulfur, lipid and antioxidant metabolism under heat stress. These mechanisms are corroborated by transcriptome data, and are driving the next phase of experimentation in the lab for additional validation. Thus, my work has developed tools and resources to analyze a dataset investigating an important

problem in biology, generated models and hypotheses to clarify regulation of the heat stress response, and provided candidate mechanisms for future validation.

Chapter 2

A POWERED-BY-CYVERSE TOOL FRNAKENSEQ

2.1 Introduction

RNA-seq is a popular tool to explore biological responses to stimuli. As reagent costs have plummeted, the bottleneck of progressing from experiment to insight has shifted from the biological experiment to data analysis. The challenge to move from data to insight includes managing large-scale data burdens, and development of statistical techniques to identify compounds that drive biological responses. While there has been significant progress in creating web-based tools, these offerings are prone to the limitations that result when system architecture does not anticipate data federation (aggregating data from connected tools into a centralized

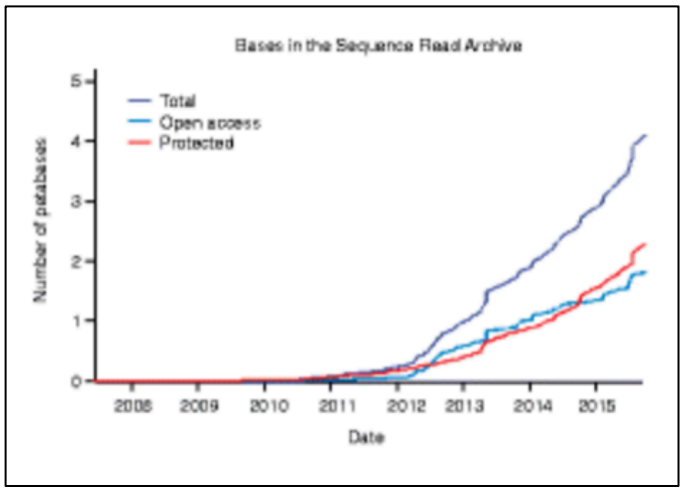

Figure 4: The number of archived data in the short read archive in petabases as function of year. From (Muir *et al.,* 2016).

resource: http://businessintelligence.com/dictionary/data-federation/ , 4/26/2016) or platform crosstalk during the early stages of development. The strain on such tools is likely to grow as data burdens continue to increase. The shortcoming of the current generation of tools can be addressed as a new generation of software platforms evolves that enables shared resources to be exploited through system crosstalk (i.e. the ability to query shared resources, such as databases or sequencing files).

Crosstalk between resources enhances the functionality of individual tools by creating collective abilities not found in any individual platform. For example, a single pipeline could pull input and other data from another tool or platform. Data management could then be distributed across tools and resources. While RNA-seq experiments rely on reference files for alignment, including both a FASTA file and an annotation, the storage of these files is cumbersome. Additionally, many different types of analyses other than RNA-seq have a similar requirement for input genomes. Thus, a system designed around a single database which manages and stores the input data for multiple tools would be a significant improvement in efficiency. This is precisely the relationship enabled by fRNAkenseq's integration with CoGe.

The advantages of such integration will only increase as it becomes unfeasible to have all resources for computational analysis in a single tool. Earlier generations of tools built without consideration of this data-sharing approach suffer from an inability to share reference files and other inputs, and assume the entire load of data burdens that accumulate with job execution. Thus, there is a need for a new a paradigm for development of future tools to meet such challenges. Such tools will exploit the advantages of the Internet to act as a conduit between resources while moving diverse

data types at a genome scale (dozens of GBs). They must also provide the ability to flexibly manage computational resources and job execution.

This pattern of development will provide a useful way to transition from pipelines to software packages that are scalable and relatively lightweight, enabling development to keep pace with the evolving challenges of the genome era. The advantages of job automation and cloud-based data storage enabled by such tools will become more pronounced as genomics technology shifts from the lab to more applied settings, such as the clinic or the farm. Additionally, the tools will be able to adapt as diverse higher throughput technologies, including single cell sequencing and automated data collection in agriculture, become more common.

2.1.1 CyVerse and the Science APIs

CyVerse is a pioneering cyberinfrastructure project developed to make web tools, databases and computing resources accessible to the scientific community (Figure 2), (Merchant *et al*., 2016). Crucially, CyVerse offers access to cloud-based data storage in the form of the CyVerse Data Store. This can serve as an endpoint for processed data as well as hosting for input. CyVerse includes multiple offerings that comprise a suite of tools that biologists can use to process their data, often exploiting Texas Advanced Computing (TACC) resources. Some of these are third party tools, such as fRNAkenseq, that rely on the Data Store for data storage. Others are pipelines that reside in the CyVerse Discovery Environment as lightweight applications for data analysis. The CyVerse Discovery Environment (DE) boasts a graphical user interface (GUI) by which users access a diverse set of applications. It also serves as is a user-friendly introduction to bioinformatics. These applications generally consist of underlying bash script pipelines, accompanied with a JavaScript Object Notation

(JSON) wrapper that allows user input through the GUI in the CyVerse Discovery Environment. Processed data is returned to the Data Store, and inputs can be stored here as well.

Figure 5: The full offering of Agave APIs accessible to developers through the Software **D**evelopment **K**it (SDK). These APIs enable users to interface with CyVerse infrastructure, though default source and endpoint for data resources is customizable.

2.1.2 Agave APIs to Run Sequencing Applications

The Applications in the DE are useful, but are limited by being relatively self-contained. While there are many Apps in the CyVerse Data Store, many of the workflows are not complex enough for the needs of labs with a strong computational component. Additionally, the degree of data sharing between Apps in the DE and third-party tools is minimal. The computing environment of Apps in the DE is also restricted to either a private computing system managed as a Condor cluster, or to resources on TACC machines, as managed by CyVerse developers. Essentially, any of the Apps developed in the DE are confined to running only on the platforms previously allocated for the DE. The lack of control over computing resources is a key limitation for users who wish to develop tools in the DE. This also constrains any

development outside the CyVerse environment that still makes use of its infrastructure.

Thus, an important resource independent of CyVerse but leveraged by some of its tools, such as some Apps found in the Discovery Environment and third party tools, is the Agave (**A** **G**rid **A**nd **V**isualization **E**nvironment) API suite which is accessible to developers through a software development kit (SDK) (Figure 5). Agave provides the means to control job submission to Apps that do not need to reside in the DE, as well as custom management of job execution systems. These job execution platforms can be TACC resources, cloud-based, or private servers. Once an available computing service is registered as an execution system in Agave, Apps can be registered to run on it. App output can be archived to the CyVerse Data Store to avoid costly storage burdens on resources dedicated to job execution. Incorporating Agave Apps into third party tools capitalizes on the advantages of the Data Store while improving opportunities for data sharing between different platforms.

2.1.3 API Integration Blueprint and Utility to Biologists

Through fRNAkenseq's web interface, users query CoGe's database for genomes available for analysis. This includes either 30,000 existing public genomes, or private genomes that they have been granted permission to access. Additionally, users can choose from several manually curated high quality genomes (frnak_approved). Genomes are pulled from CoGe into the CyVerse Data Store using CoGe's get_gff and get_fasta API calls. The files-list query through the Agave web service determines if the genome already exists in the Data Store. CoGe genomes are then processed through indexing, and index files are returned to the user's CyVerse Data Store via another in-app API Agave command: files-upload. This means that

there is a one-time cost of indexing for each genome. This saves significant time during analysis, as index pre-processing can take between 45 minutes and two hours.

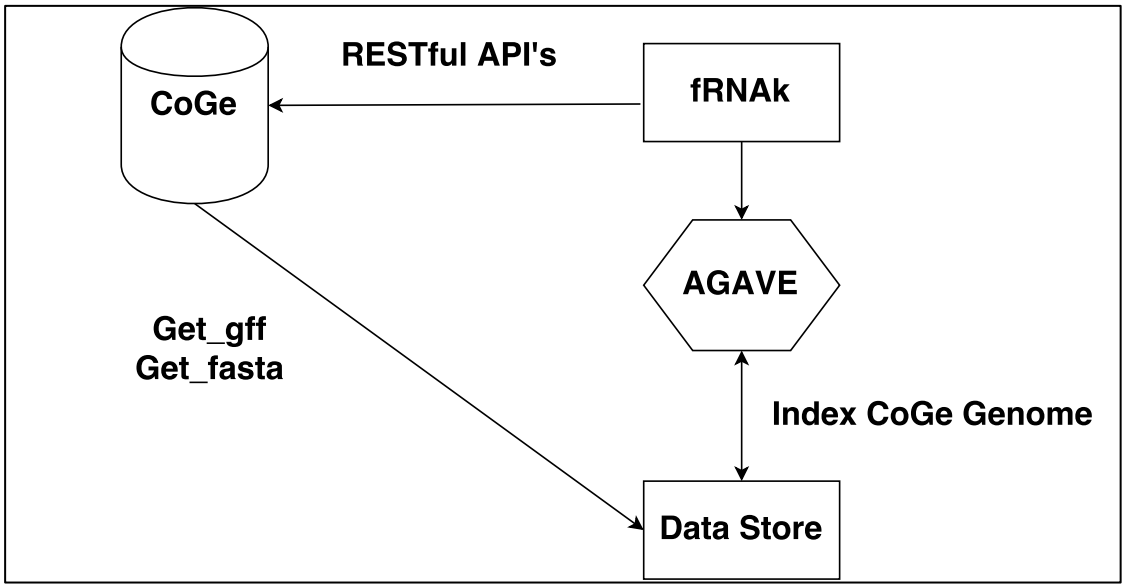

Figure 6: Schema of APIs and their functions for fRNAkenseq. Agave API's move job specific data across machines, while additional other (CoGE) APIs, manage genome files. CoGe RESTful APIs all fRNAkenseq to query CoGe's database.

In total, the fRNAkenseq connection schema between CoGe and TACC resources can be described below (Figure 6). The establishment of this paradigm lays the groundwork for companion tools which are currently in development, such as MInotauR. MInotauR will represent a powered-by-CyVerse design similar to fRNAkenseq, but its pipelines will emphasize analysis of microRNAs.

The ultimate product of fRNAkenseq's innovative backend is a tool that brings together a diverse set of resources to meet the needs of a biology research group with

large scale RNA-seq data. Viewed from the interface, fRNAkenseq provides an accessible portal to algorithms and a workflow that spans mapping and quantification of FastQ files to differential expression using multiple enrichment algorithms. This workflow is optimized for usability by biologists (Figure 7), providing a natural flow from read alignment to differential expression. This design, which is more accessible to a biologist without computational experience, represents an alternative to the loosely-organized Galaxy toolshed.

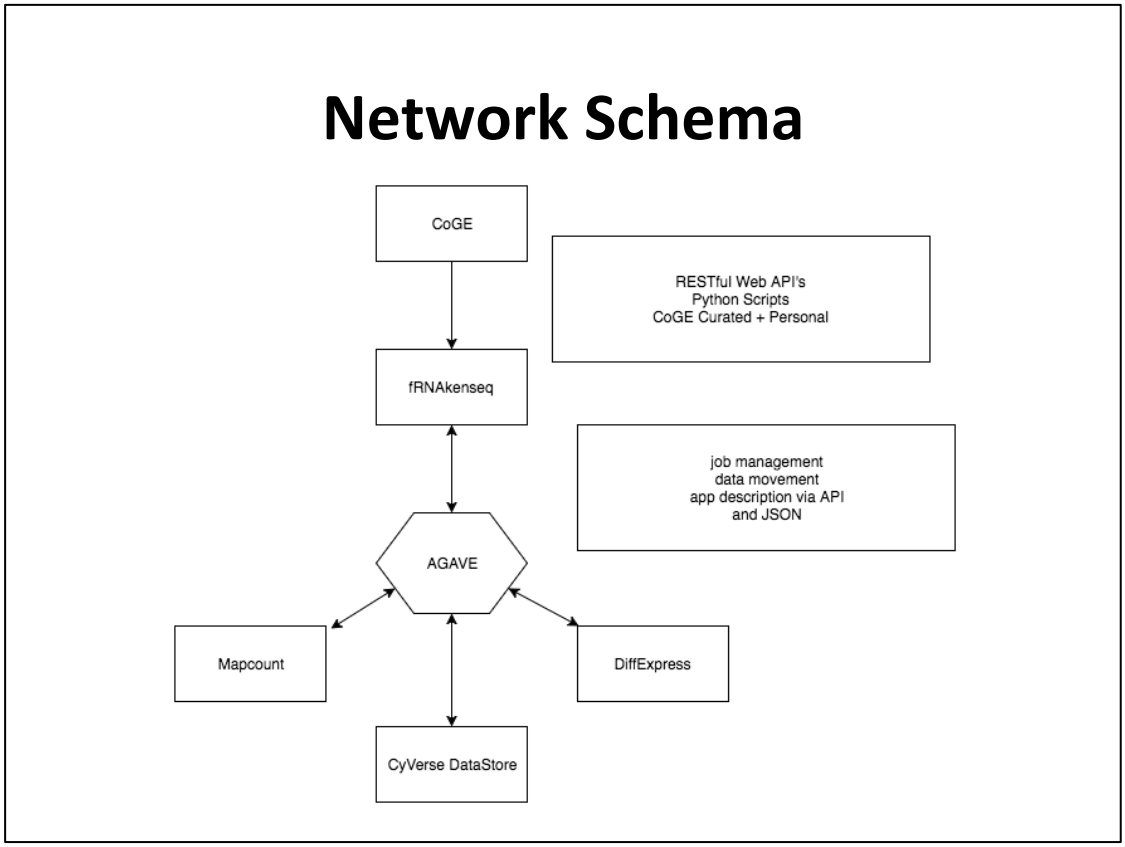

Figure 7: fRNAkneseq backend pipeline components, scripts that manage them, and interface.

2.1.4 fRNAkenseq Components: MapCount

The two main services offered by fRNAkenseq are MapCount and DiffExpress, which span initial mapping/quantification and differential expression analysis, respectively. Output from each of these stages is accessible, at any time, in the user's CyVerse account through the DE. Importantly, DiffExpress leverages three R packages for differential expression producing lists of enriched genes with varying levels of stringency. This allows a biologist to employ a Venn diagram approach to select enriched genes, focusing on the intersection of sets of genes predicted to be differentially expressed by multiple algorithms.

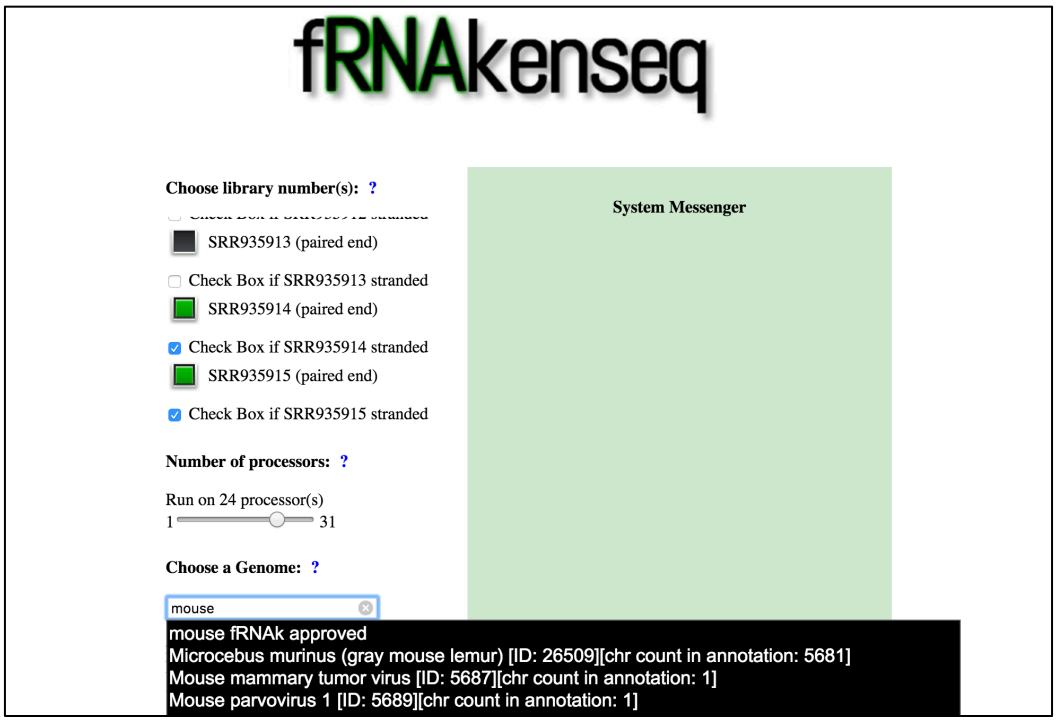

Figure 8: fRNAkenseq MapCount interface showing the selection of a FastQ file and the range of genomes available through CoGe's database

The MapCount stage of fRNAkenseq provides the first steps of an RNA-seq quantification workflow (Figure 8). MapCount offers informatics capacities for both stranded and un-stranded data, as provided in the form of raw read FastQ files. One Agave App processes stranded data, while another App processes un-stranded (Figure 9). Mapping is completed by Hisat-2 (Pertea et al., 2016). Samtools (Li et al., 2009) then converts the mapped output into BAM files (Figure 12). This represents a next-generation iteration of the popular Tuxedo pipeline (Bowtie2/Tophat2, Cufflinks for mapping and quantification, respectively). For further analysis, the standardized BAM files can then be shuttled to various options available in the CyVerse environment (Figure 10).

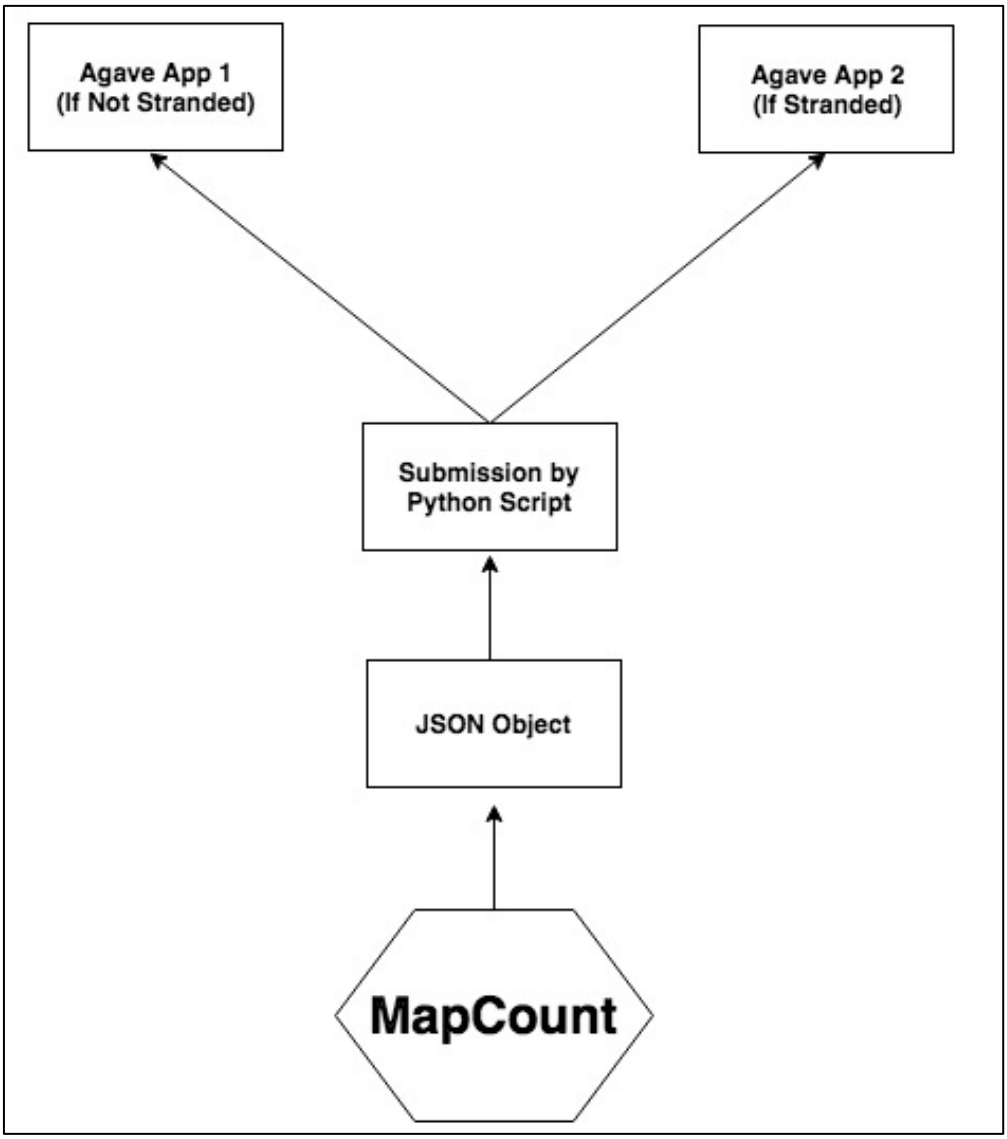

Figure 9: Mapcount workflow and backend integrated with the Agave Apps

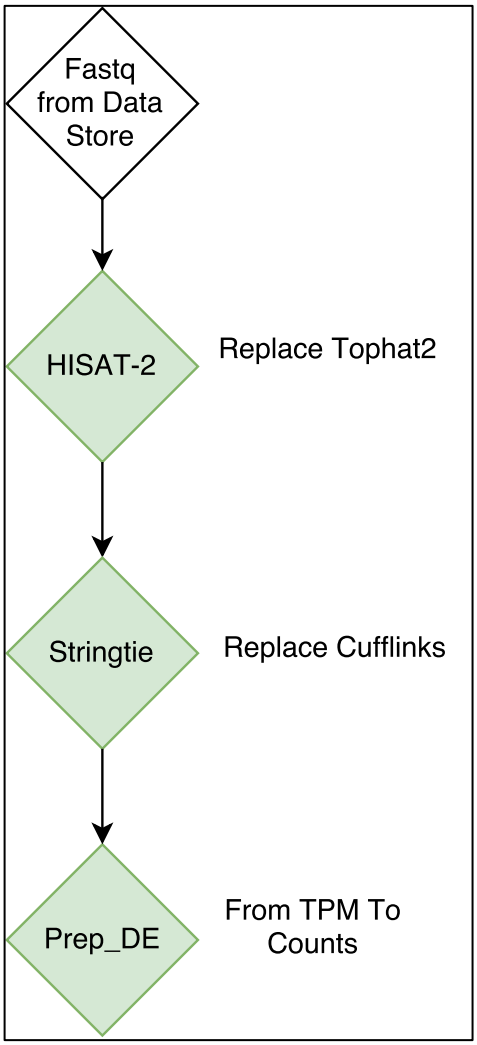

Figure 10: Algorithms of fRNAkenseq MapCount pipeline. Current generation algorithms in the green squares, adjacent to the earlier algorithms they have replaced from the Tuxedo pipeline.

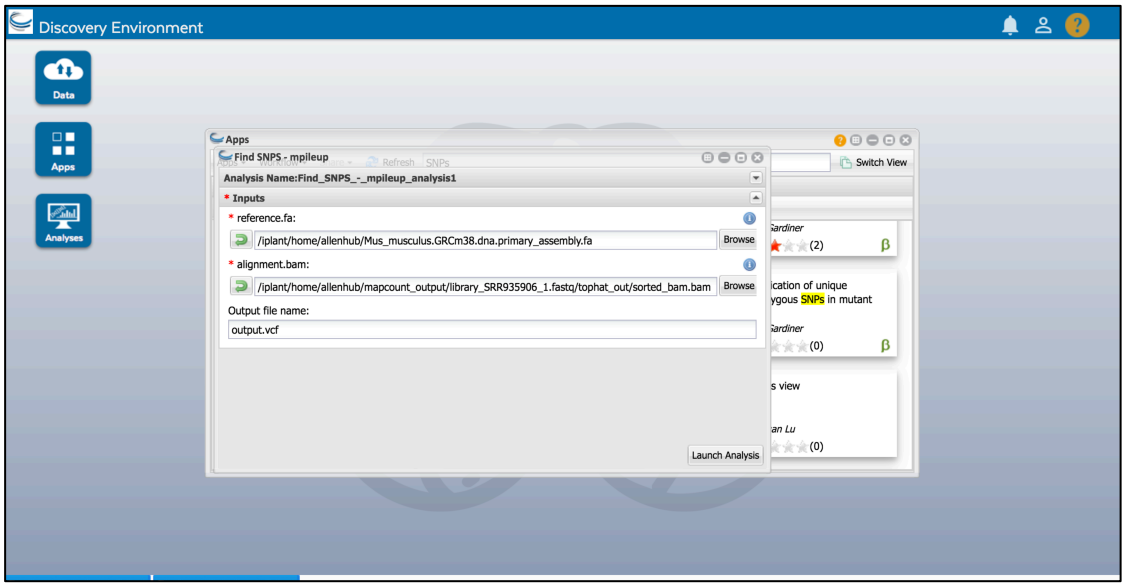

Figure 11: Application of a downstream analysis in CyVerse Discovery Environment executed on a mapped BAM file previously processed by fRNAkenseq

While analyses are run on a separate execution server, all of fRNAkenseq's outputs go directly to the CyVerse data-store for optimal interoperability and crosstalk between fRNAkenseq and other resources in the Discovery Environment. The files are stored in the user's CyVerse Data Store home directory, under mapcount_output, in accordance with the schema in Figure 14.

All mapping and quantification data resides within the mapcount_output directory in the user's CyVerse Discovery Environment account. Relevant files depicted in this schema include the sorted sam, the indexed sam and the sorted bam files, in addition to raw read counts associated with each gene. They may be further analyzed with other CyVerse apps that similarly use the Discovery Environment

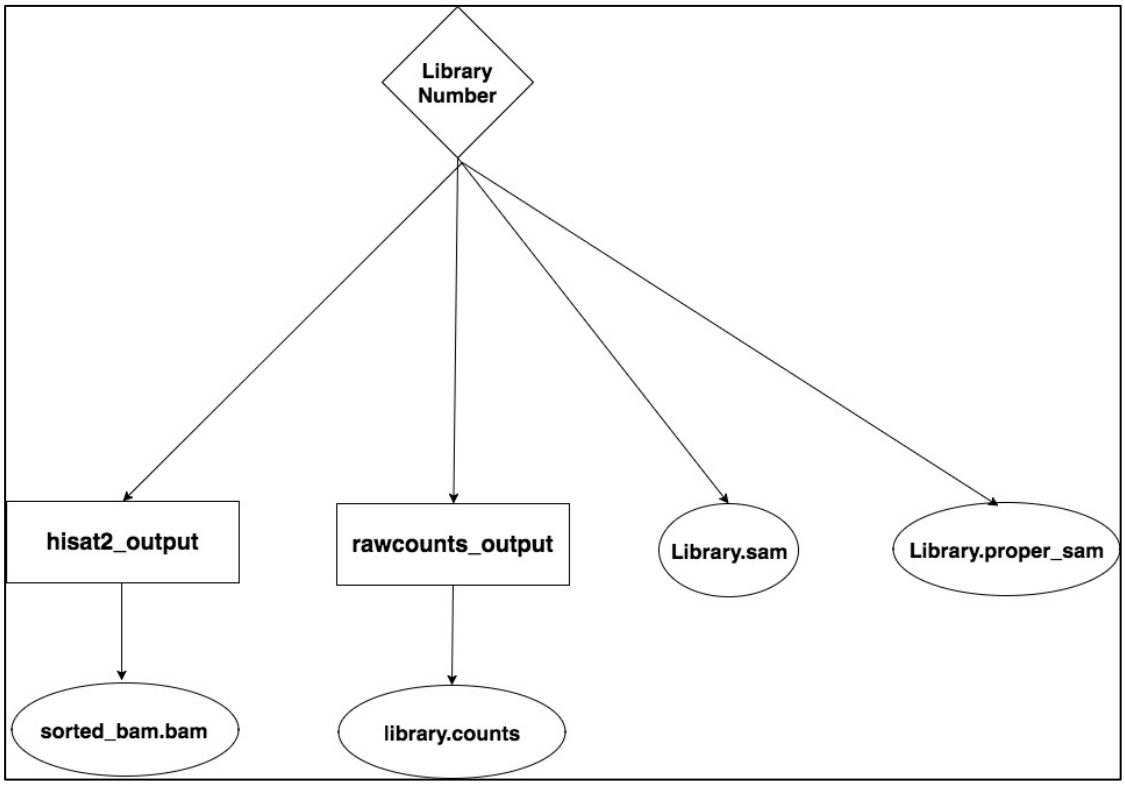

Figure 12: Schema depicting the MapCount directory layout for fRNAkenseq.

cyberinfrastructure. All fRNAkenseq outputs are accessible for further analyses within the CyVerse Discovery Environment, providing a valuable degree of freedom for researchers who would like to extend downstream analyses with customized workflows (i.e. using Apps that process BAM files). Alternatively, a user may progress to differential expression analysis using the second component of fRNAkenseq, DiffExpress.

2.1.5 fRNAkenseq Components II: DiffExpress

DiffExpress is developed for stringent enrichment analysis, using input from the MapCount pipeline (Figure 13). It deploys multiple algorithms in a single run (Figure 14), combining outputs into lists of genes of varying selectivity according to

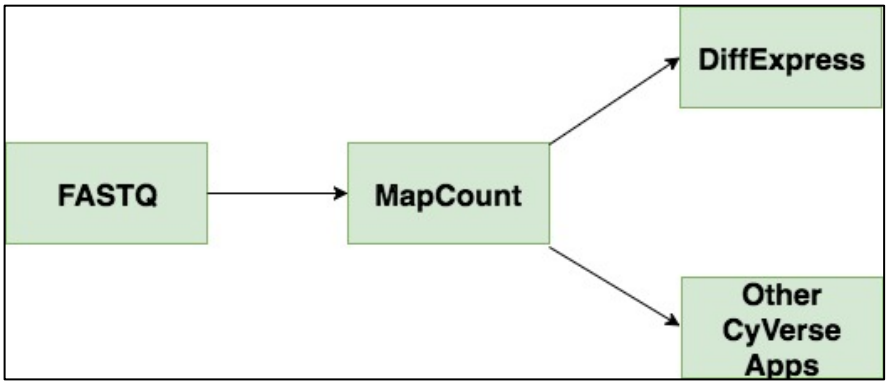

Figure 13: Possible sequence analysis workflow from start to finish with FastQ files using fRNAkenseq and other CyVerse apps

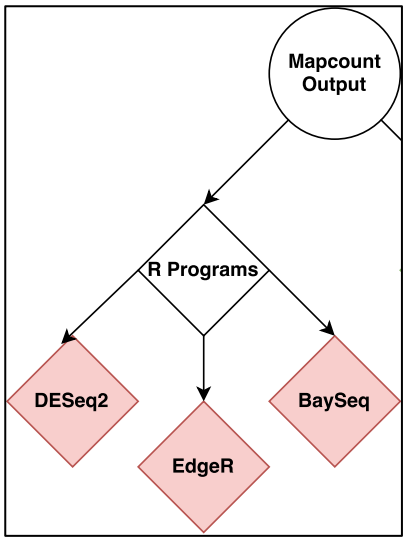

Figure 14: The schema for fRNAkenseq's DiffExpress pipeline representing the different algorithms executed as a single Agave App.

the number of algorithms that have declared them differentially expressed. While the negative binomial null model is incorporated into all featured algorithms, each calculates the significant parameters in a different fashion. Despite many approaches to differential expression, none is found to be universally superior, and the most accurate strategy depends on the profile of the gene and the landscape of the data (Rapaport *et al.,* 2013). To this effect, DESeq2 models variance as a linear

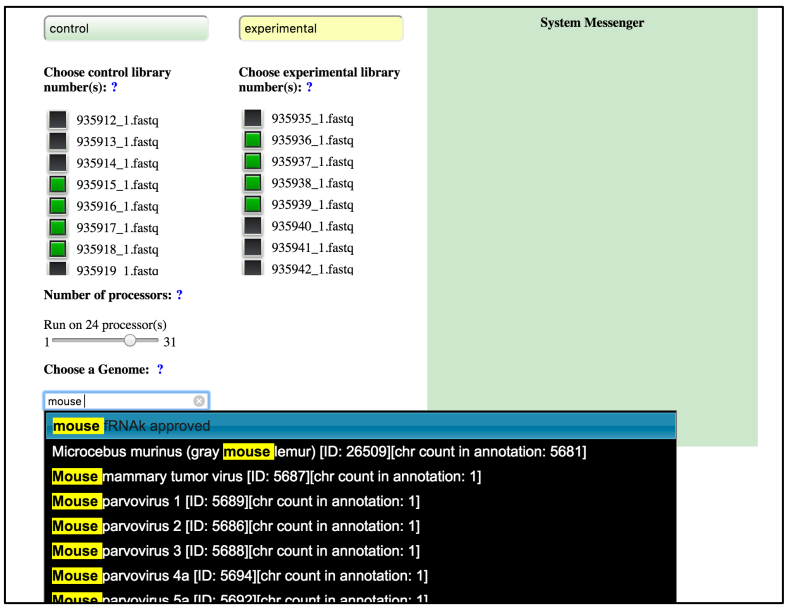

Figure 15: fRNAkenseq's DiffExpress user interface showing the set-up of a sample analysis using libraries previously analyzed by MapCount.

association with the mean of gene expression level (Love, 2014) whereas edgeR uses an empirical Bayes' method to determine most likely variance for a group of genes with a similar expression profile (Robinson *et al.,* 2015). BaySeq relies on a process of empirical sampling of the data to determine posterior probabilities of differential

expression (Hardcastle *et al*., 2010) and return a Bayesian FDR estimate (Seyednasrollah *et al.,* 2013). fRNAkenseq provides output from all three of these differential analysis programs and also applies a Venn-diagram approach to allow users to identify genes that are classified as significantly differentially regulated by more than one statistical approach.

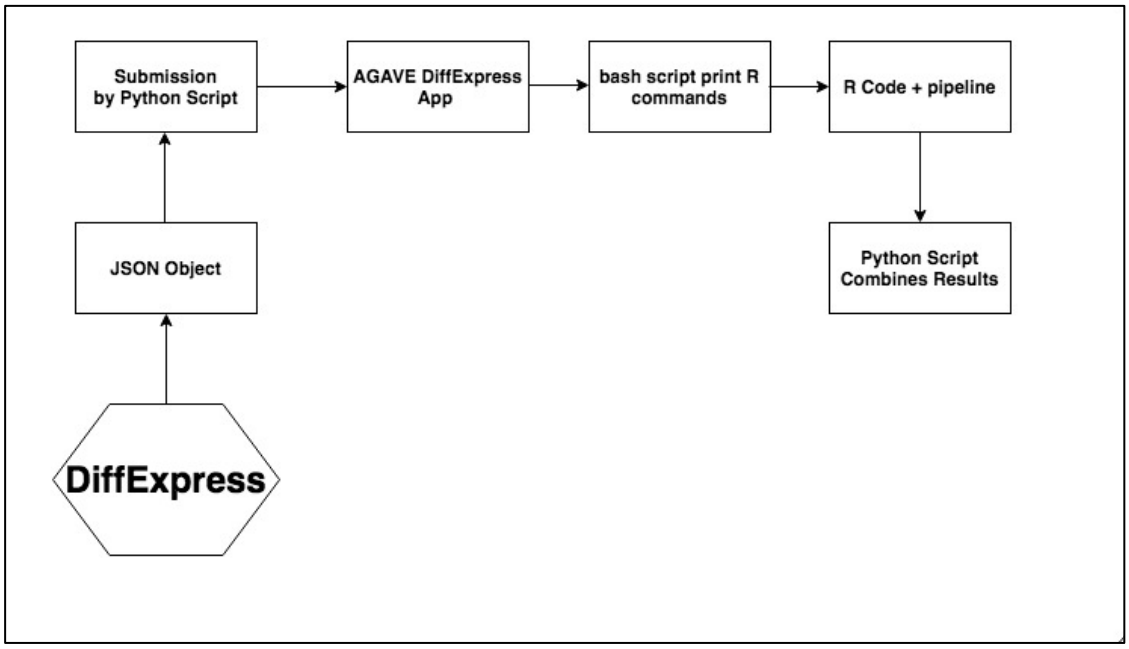

Figure 16: Similar to MapCount, the DiffExpress pipeline has an Agave App at its core. The DiffExpress Agave App, however, has an extra level of complexity in that it creates and executes the code that composes the R pipeline, based on input submitted to DiffExpress, A subsequent Python script identifies genes predicted as enriched by 1-3 different programs.

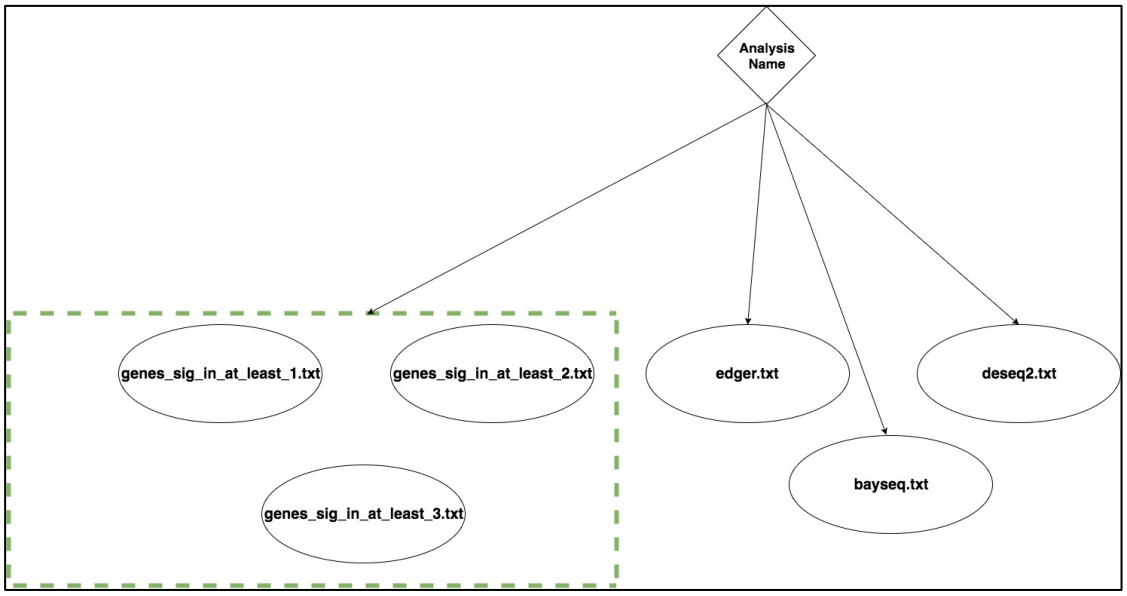

Figure 17: This represents the schema of DiffExpress output within a user's CyVerse directory. Included in this output are the differential expression outputs from three R packages, edgeR, BaySeq and DeSeq2. Files in green rectangle contain genes differentiall expressed according to one, two, or all three of these algorithms.

Each algorithm in DiffExpress (Figure 14) leverages different statistical assumptions to accomplish differential expression analysis. Finally, the output of each of these algorithms is combined into tab-delimited files that integrate differentially expressed genes with various levels of stringency as determined by the number of algorithms that predict them as differentially expressed (Figures 16 + 17). The workflow of the DiffExpress pipeline as an Agave App, though similar to that of MapCount, involves the combination of R and Python scripts (Figure 16).

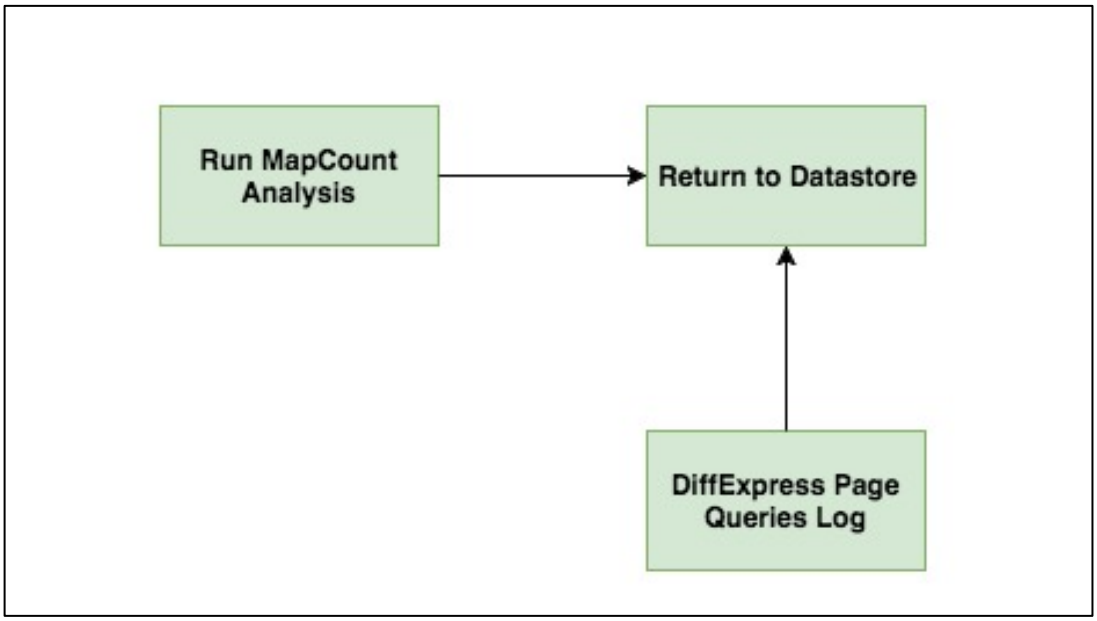

Figure 18: Users have access to data in the Data Store (CyVerse) as well as in CoGe (genome files), through CyVerse's centralized authentication system. Data is moved across different systems according to the task being executed and the API call.

2.1.6 Management of Resources Across Systems

When executing differential expression analysis, it is important to make sure that a user does not try to run DiffExpress on a set of libraries that were analyzed through MapCount using different versions of a species' genomes or accessing the genome of a different species than that used for the MapCount analysis. Implementing this type of safeguard requires first storing the MapCount data analysis information and then having the fRNAkenseq webpage access that data. This is accomplished by executing the schema above, using the fRNAkenseq webpage to execute Agave API calls (Figure 18). This ability to transfer data directly between the webpage and the Data Store through the Agave APIs is also applied during MapCount

when updating the file logs that will store the genome information for subsequent fRNAkenseq DiffExpress runs.

2.2 Discussion: Description of an Agave Apps as a Foundational Unit

By relying on Agave Apps as fundamental units of job execution, fRNAkenseq is able to employ a grid computing strategy (Figures 19 and 20). A grid computing paradigm is one in which different computational tasks are relegated to separate nodes or units. These may be on the same machine or separate ones. This design philosophy

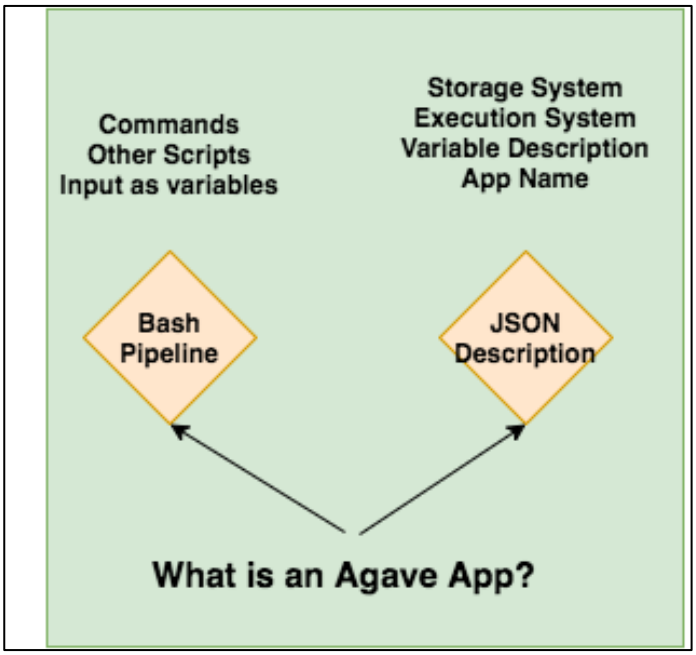

Figure 19: An App posted to the Agave service consists of a shell script pipeline template with its JSON wrapper description. The shell script contains the individual commands for the pipeline. The corresponding JSON wrapper provides execution and storage system information for the app. It also describes the variables that will be passed to the bash template

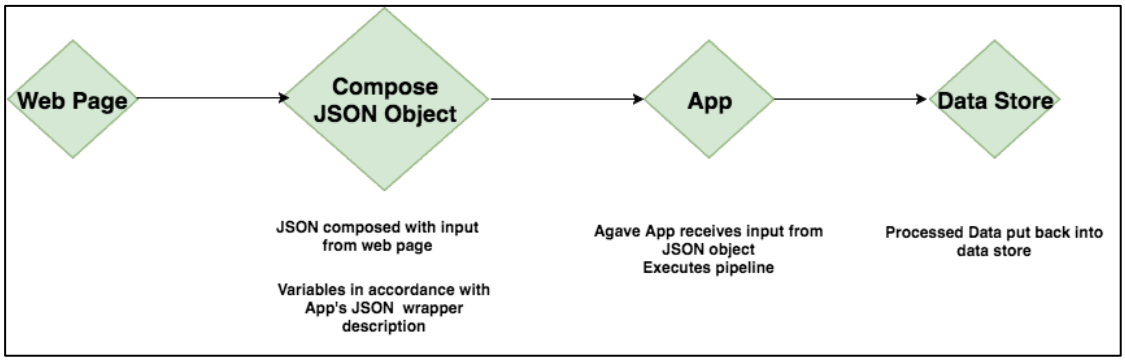

Figure 20: Integration with Agave Apps and a third party web tool such as fRNAkenseq is described. Once an Agave App is registered it can comprise the backend pipeline for a tool. Subsequently, an arbitrary web interface can be developed to compose JSON objects to be submitted to the Agave app. These JSON objects will be consistent with the JSON wrapper of the App that has been posted to the Agave service. When the pipeline executes, the processed data will be returned to the data directory as specified in the App description. The ability to control jobs in this fashion is enabled by the jobs service of the Agave API,

is useful for the development of genome analysis tools because it allows a way to manage the computational loads of job execution and resulting data burdens. One machine, or a set of machines, can be used for job execution, while another system (such as the Data Store) is used for archiving input data and results.

The Agave APIs create a set of diverse utilities that enable software development within the CyVerse computing infrastructure. The Agave Software Development Kit (SDK) makes these offerings available to tool developers in the bioinformatics community (Figure 5). Each service of the Agave API promotes different capabilities spanning from authentication to data transfer. Different subsets of these functions are currently employed to various degrees by the powered-by-CyVerse tools. Some powered-by-CyVerse tools include BioExtract Server (Lusherbourgh et al, 2011), CIPRES (Miller et al., 2010), ClearedLeavesDB (Da et al.,

2014), CoGe (Lyons et al., 2008), iMicrobe, Integrated Breeding Platform (The IBP Breeding Management System Version 3.0.9, 2015) SoyKB (Joshi et al., 2014) and Galaxy (Afgan, 2016). Among the community of powered-by-CyVerse tools, fRNAkenseq is unique through the combination of utilities it offers to biologists and the depth of its reliance on the Agave and other CyVerse APIs as a foundational system. This scheme of integration is complemented by fRNAkenseq's connection with CoGe.

CoGe is a genome resource developed for synteny analysis. Among its features is a database of 30,000 computationally and manually curated genomes from all domains of life. Users may also upload and manage private draft genomes. The ability to customize data in this setting makes the tool an excellent starting point for researchers to begin analyses of non-model organisms for which the only reference files are draft genomes they have produced. CoGe has grown to include bioinformatics pipelines and browser utilities that exploit the database of genomes. Although some of these pipelines, much like fRNAkenseq perform mapping and quantification fRNAkenseq additionally completes differential expression.

CoGe also includes a set of APIs that allows other third party powered-by-CyVerse tools to utilize CoGe's underlying database of genomes. These APIs are independent of the Agave APIs, and fRNAkenseq's use of them is an example of its uniqueness, employing connectivity between third party powered-by-CyVerse tools not found in the Apps in the Discovery Environment. Thus, fRNAkenseq is able to provide novel pipelines to biologists and facilitate a connection between CoGe and the CyVerse DE. This scheme of integration, which incorporates a grid computing

strategy, connection with the data storage, and data sharing with CoGe, separates fRNAkenseq from Galaxy, a similar powered-by-CyVerse tool.

2.2.1 Comparison with Galaxy and CoGe Integration

Galaxy, a software project started in the mid-2000s, is designed to make common bioinformatics tools available to life scientists through web interfaces. fRNAkenseq extends the advantages of web-based tools like Galaxy, and also serves as an example what the next generation of integrated analysis systems can offer. In addition to its immediate utility to biology labs, fRNAkenseq more fully leverages advances in cyberinfrastructure to improve connections between resources through its API-driven design. This development philosophy avoids backend redundancy and improves flexibility for programmers that wish to selectively exploit the utilities of different tools existing on separate platforms.

Similar to the CyVerse DE, Galaxy offers a degree of flexibility in terms of tool development, but also the ability to run different types of instances (i.e. cloud or local). It follows a standard protocol of combining command line tools with wrappers to define Galaxy "Apps" – unrelated to Agave Apps. This paradigm of software development has been effective in meeting the informatics needs for many groups. It also provides an opportunity for researchers to publish their tools as applications within the Galaxy resource. A public instance of Galaxy runs as a powered-by-CyVerse tool in the sense that it exploits TACC resources and utilizes the CyVerse Data Store. However, Galaxy is a self-contained workflow for bioinformatics analyses and does not rely on CyVerse APIs for functions beyond simple access to data. Thus, unlike fRNAkenseq, it does not exist as a third party tool that appropriates the full range of CyVerse and Agave APIs at a truly foundational level. This paradigm

of extended integration demonstrated by fRNAkenseq brings enhanced flexibility for development and data sharing between resources. By relying on Agave for a range of functions from data movement, authentication to job management, fRNAkenseq is able to combine the resources of other powered-by-CyVerse tools in a way that is not possible for a tool like Galaxy, whose substantial pre-existing architecture will limit the scope of integration with CyVerse resources.

Chapter 3

BEYOND DIFFERENTIAL EXPRESSION: TISSUE ENRICHMENT

3.1 Introduction

Once users have accomplished initial file processing through fRNAkenseq, a next step is to relate gene expression data to biological insight. fRNAkenseq makes some early steps in this process (mapping, quantification, and robust differential expression) simple. However, identifying genes that are differentially expressed between conditions is often not enough to elucidate pathways or provide other biologically useful information for tissues of interest in a way that is treatment independent. For example, other methods must be used to identify tissue-specific biology from transcriptome data that has been collected over the course of many large-scale experiments. There is a demand for such techniques, as reliable heuristics to identify tissue-defining genes can provide important forms of feature reduction, partly by reducing organ-specific biology to a relatively small set of genes that are enriched in a tissue of interest. This can be a useful form of pre-processing for subsequent pipelines. Alternatively, the set of genes enriched in a tissue of interest is often biologically informative on its own. Thus, the proliferation of tissue-diverse datasets associated with large-scale experiments represents an opportunity to explore tissue enrichment in its own right. Such data mining produces insights complementary to the original studies for which tissue samples were collected. A subsequent chapter, Chapter 4, will show that it can be useful to understand a biological process, such as heat stress, in the context of enriched genes. This chapter, however, will focus on

using a tissue enrichment threshold to compare genes enriched in breast muscle and those enriched in cardiac tissue. This comparison effectively isolates tissue-specific biology to a high degree of resolution.

3.1.1 Motivation

As researchers regularly accumulate complex data sets with samples from multiple organs, they increasingly want to explore tissue-specific biology. In these types of studies, the analysis shifts from using traditional differential expression methods to leveraging enrichment thresholds that isolate tissue-defining regulation. We develop a method for identifying tissue enrichment, different from the negative binomial model-based differential expression analyses found in DiffExpress (edgeR, BaySeq, DESeq2), and which represents an extension of previous tissue enrichment strategies. This approach is then employed to process large-scale datasets that come out of fRNAkenseq MapCount pipelines, which have processed FastQ files from a diverse set of tissues. This enrichment analysis is useful both independently for comparative studies, and also as an initial step of feature selection for bioinformatics analyses with more advanced pipelines. Our standard of enrichment represents an elaboration and, in many ways, an improvement on methods originally developed by the Genotyping by Expression (GTEx) projects, which were developed to study tissue-specific biology in human organs. The performance of our heuristic on our dataset is compared to that of GTEx methods, after the motivation for our strategy is discussed.

3.1.2 Criteria for Enrichment

Tissue enrichment strategies have diversified to meet the informatics needs of an expanding genomics community. Some of these are characterized in detail through

the review "A Benchmark of Gene Expression Tissue Specificity Metrics" (Kryuchkova-Mostacci and Robinson-Rechavi, 2017). Strategies regarding tissue enrichment include a tissue specificity index, gini-coefficient, and a z-score, $\frac{\overline{x}(tissue\ interest) - \mu(background)}{\sigma(background)}$, approach. Each of these approaches produces a distinct profile of enrichment for different genes across tissues in a tissue-diverse dataset (Kryuchkova-Mostacci and Robinson-Rechavi, 2017). Under the z-score approach, genes whose difference in expression in a tissue of interest compared to background is greater than three standard deviations from the background mean of background tissues are enriched. This z-score approach is commonly used for other types of analyses with continuous values. However, the z-score threshold often requires the assumption of an underlying normal null distribution for data, in which case ninety-five percent of the observations will lie within 2 standard deviations of the mean. The generality of this assumption makes z-score-based metrics a widely applied threshold to sequencing data. Though it tends to correlate poorly with other enrichment strategies, it is the only method that considers the standard deviation of the expression data in enrichment studies (Kryuchkova-Mostacci and Robinson-Rechavi, 2017). We consider this to be an important feature. Thus, for our enrichment analysis, we leverage a modification of the z-score approach, explaining the theoretical motivations for using a cut-off of five standard deviations (5SD) as opposed to three used in the methods described in Kryuchkova-Mostacci and Robinson-Rechavi. We illustrate typical situations in a read alignment dataset that would violate the assumptions of normality required for more relaxed z-score methods, but show how using a cut-off of five standard deviations addresses this problem by remaining applicable for arbitrary distributions. We also demonstrate

empirically that the distribution of reads across samples in our tissue diverse dataset is non-normal (Table 1). We compare this to the ad-hoc threshold for tissue enrichment that is leveraged in GTEx studies

3.1.3 GTEx and Other Enrichment Approaches.

The GTEx project is one of the most well known instances of large-scale tissue enrichment analysis on a large set of samples that resembles our dataset. The GTEx project is an exploration of gene expression across 53 tissues, ultimately focused on understanding the basis for genetic control of tissue specific transcriptomes (The GTEx Consortium, 2015). One step of GTEx computational studies has been to identify tissue enriched genes. The GTEx cutoff for enrichment, which does not consider standard deviation, but instead selects enriched genes according to five-fold increase in means, is effective but ad-hoc. Our strategy, in contrast, incorporates both the mean and standard deviation of expression data, but is distribution independent. Although we previously explored relative tissue expression (RTE) analysis, which relied on the median of tissue expression (Bailer *et al*., 2009), this method produced significant overlap between tissues. We will demonstrate the efficacy of our more stringent method to tease apart tissue-specific muscle biology, and compare to the GTEx approach. In this chapter, we use our enrichment strategy as a lens through which we can understand gene expression differences between cardiac and skeletal muscle. In Chapter 4, we will discuss genes determined to be enriched in liver according to the same methods used in the comparison of skeletal and cardiac tissue. These enriched genes will serve as a foundation for statistical learning pipelines that integrate transcriptome and metabolome data. Tissue enrichment approaches have helped us maximize the insight gained from samples processed by fRNAkenseq.

3.1.4 Improvement Over GTEx

To date, the laboratory's complete dataset comprises over 1500 RNA-seq libraries from diverse tissues. At the time of this analysis, somewhat over half of those samples (800) had been processed. These hundreds of samples, across many different tissue types, provided an opportunity to explore robust enrichment analysis against a large background. To understand fine-tuned variations in regulation underlying different muscle types, one small subset of the total data sample is compared against the background of other tissues at a time. We focus on exploring the difference between cardiac and skeletal muscle using our enrichment definition because these samples represent tissues that possess considerable levels of both unique and overlapping biology. Additionally, increased skeletal muscle growth has been a major target of artificial selection in the broiler chicken over the past several decades (Tallentire *et al*., 2016). Understanding the important and organ specific regulatory genes in this tissue will be useful in defining the baseline biology of breast muscle, improving our understanding of biological systems that may be altered through intense breeding programs.

One of the foundational examples of work with a large tissue-diverse dataset that identifies organ-defining compounds is the GTEx study. Although this dataset emphasized human clinical data, the approach is general enough to be utilized with other tissue-diverse datasets. The GTEx method that was used to identify tissue-enriched genes from the original dataset applied a definition of genes whose mean was fivefold greater in the tissue of interest compared to background tissues. Although this represents an adequate starting heuristic, this approach is bound to over-report lowly expressed genes because it does not consider the difference between the means in the tissue of interest and background tissue in the context of the variance. We

propose a more rigorous and generally stringent definition by setting the threshold such that the z-score, $\frac{\overline{x}(tissue\ interest) - \mu(background)}{\sigma(background)}$ must be greater than five. Owing to Chebyshev's theorem, this cutoff ensures a gene is in at least the 95th percentile of the distribution of expression of the background tissue. This is under the assumption that we have correctly estimated the parameters. These assumptions are justified by the relatively large sample, and by the application of this test as a heuristic to generate hypotheses. Importantly, this approach is an improvement over the ad-hoc boundaries drawn under the GTEx study, being more rigorous from a theoretical standpoint.

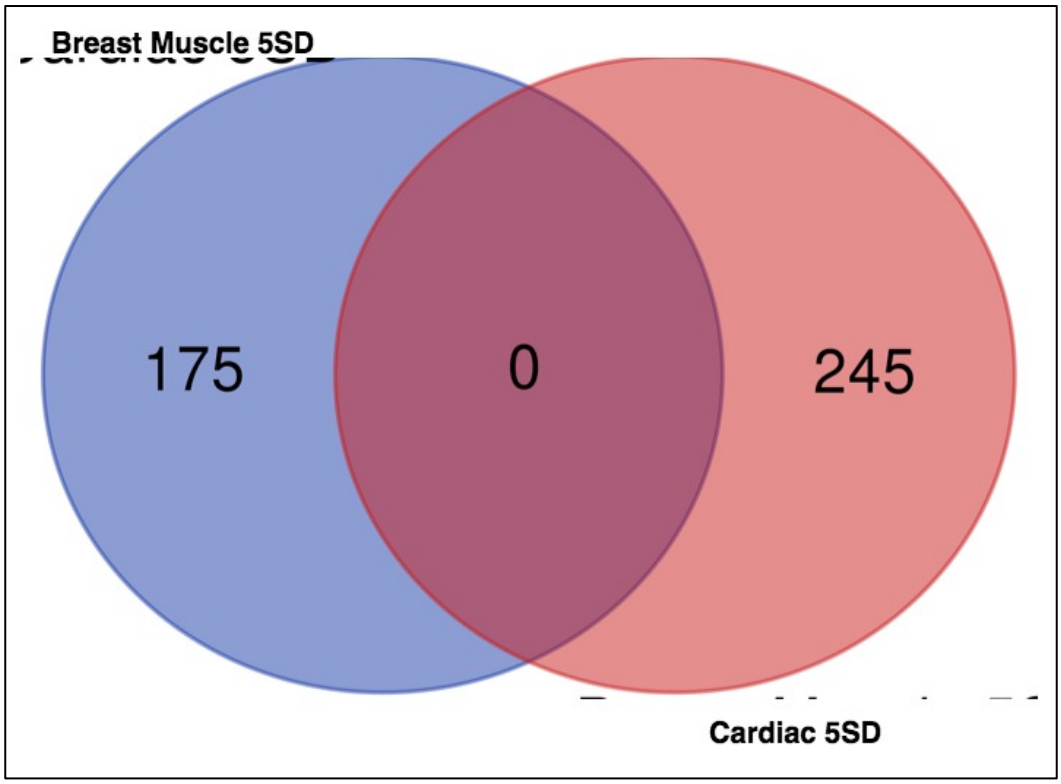

Figure 21: Venn Diagram illustrating specificity of enriched gene lists at the five standard deviation z-score threshold in muscle types.

Our threshold of five standard deviations derives from biological as well as statistical justifications. Five standard deviations can be demonstrated to be sufficiently stringent to isolate tissue specific genes, but not too stringent to exclude relevant biology (Figure 21). The reliance on Chebyshev's theorem, which is common principle discussed in introductory statistic courses, for hypothesis testing can be understood as a more general application of hypothesis testing than what is usually done under assumptions of normality and the empirical rule.

3.1.5 Dealing with Non-normality

Following the assumption of normality, the empirical rule is the heuristic that motivates the calculation of z-score tables used with minimal other assumptions, such as those regarding sample variance that would instead motivate a t-test or chi-square test. Under this approach, it is common to calculate a z-score $\frac{\overline{x}(tissue\ interest) - \mu(background)}{\sigma(background)}$ for a given value to determine the probability it was due to chance, and assume that the null distribution is a normal distribution. Thus, the probability of an observation is determined based off of the distance of that observation, in terms of standard deviations from the mean of that data, assumed to come from a normal distribution centered at μ. Generally, under these sets of assumptions, if a measurement is more than two standard deviations (Figure 22) it is considered significant at a .05 significance level. We initially explored a threshold of two standard deviations from the mean of all tissues to determine if a gene is enriched in a given tissue. This proved an effective first pass at identifying organ-specific biology. However, this threshold resulted in a strong degree of overlap in tissue-enriched genes. This definition of enrichment is thus not sufficiently stringent enough to determine tissue-unique genes among similar organs. Additionally, we have strong

evidence that our data, which concerns gene expression patterns across samples of many different organs, is not normally distributed. For example, these datasets consistently fail the Shapiro-Wilkes test for normality (Table 1).

Tissue	Pass	Fail
Breast Muscle	0	30,161
Liver	0	30,161

Table 1: Genes that have passed or failed test for normality

One of the important motivations for refining the GTEx strategy is the need for stringent identification of enriched genes that can tease apart differences in developmentally related tissues such as cardiac and skeletal tissue. We demonstrate important overlap with the five standard deviation based z-score of our approach and the previous definitions of enrichment given by GTEx, while also illustrating advantages of our approach in identifying tissue specific biology. The need for the stringency of the five standard deviation threshold also has important statistical justification. Out of 30,161 genes, all distributions among background samples fail the Shapiro-Wilk test for normality at the .05 level, causing us to reject the null hypothesis of a normal distribution. Thus, the need to identify enriched genes despite a lack of normality requires a rigorous, non-parametric standard of enrichment.

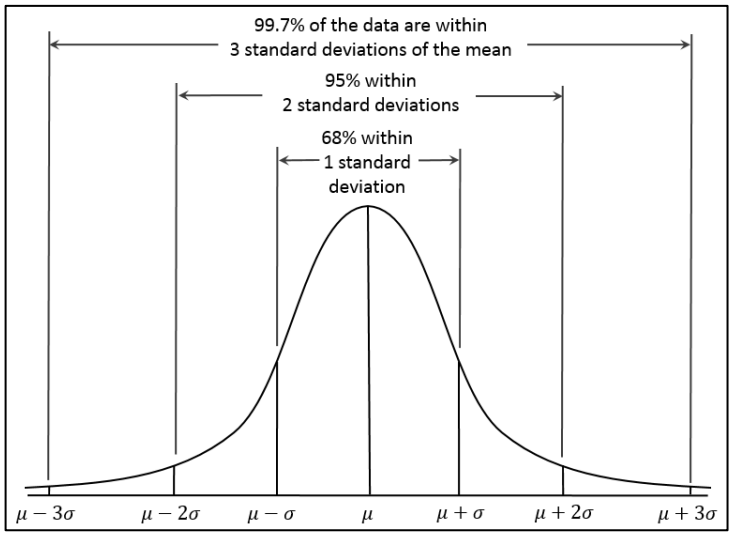

Figure 22: A standard normal distribution and the empirical rule demonstrating the percent of observations that will fall within a given number of standard deviations of the mean.

Even in cases where data does not necessarily appear to be normally distributed, but derives instead from an arbitrary function for the null distribution, it is still possible to determine bounds on the probability that an observation is due to chance by using a z-score. These highly non-normal situations are to be expected in datasets like ours, where the expression of a single gene is compared across libraries of multiple tissue types. Evidence for the lack of normality is shown by the results of the Shapiro-Wilkes test. One situation that could account for this is that gene expression is mostly sparse, i.e. the gene is not expressed (or not expressed consistently) in most tissue of most samples, or unevenly expressed across tissue types. The inevitability of non-normal data distributions is not unique to expression data across multiple tissues; it is a natural consequence of situations involving read alignment in which one sample or region is associated with a disproportionate number

of reads (Figure 23). Understanding consequences and causes of non-normal distributions in RNA-seq data will provide context for different strategies regarding hypothesis testing.

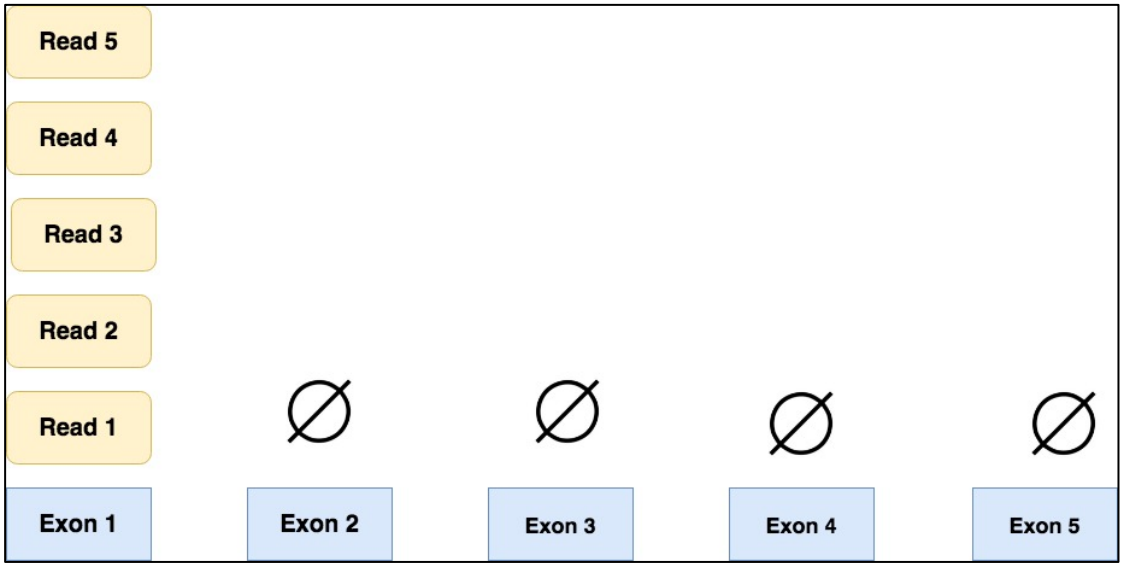

Figure 23: A simple example of non-normality in read distribution across exons of a gene. A similar, though less exaggerated, effect occurs among samples of multiple tissues

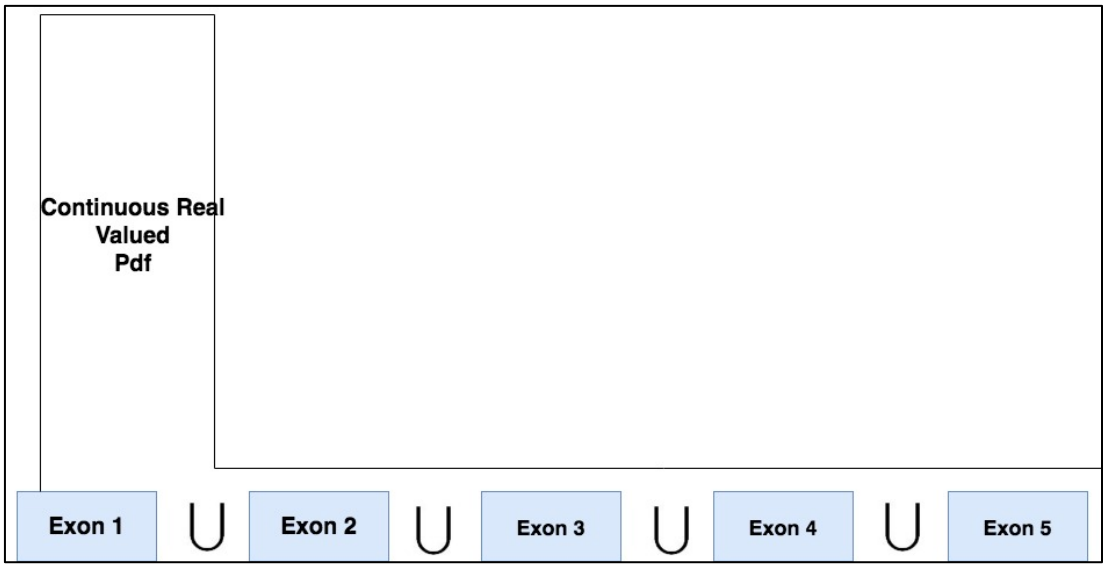

Figure 24: Probability distribution function (PDF) corresponding to the histogram of read alignments per exon in Figure 25.

3.1.6 Models of Read Alignment Associated with Non-normality

Perhaps the simplest example in RNA bioinformatics of a random variable that may not be normally distributed would be the number of RNA reads mapped across exons of a single gene. Due to sequencing chemistry, gene structure, and other influences, it is possible that the probability of reads aligning to one exon versus others is disproportionately high or low. In this situation, the distribution of reads across the various exons would be expected to be non-normal. In the most extreme case, all of the reads could concentrate on a single exon. This is an example of a possible, though extreme type of measurement bias for sequencing data. Examples of exon bias may also result from biologically relevant phenomenon, such as alternative splicing. Such regulation would require disproportionate expression of exons highly utilized across different isoforms and has been identified as an influence on read alignment distributions across exons (Liu et al., 2015). As in the exon-by-exon

consideration for a hypothetical gene, the distribution of reads aligned to a single gene across organs will be driven by both sampling and the influence of biologically meaningful effects. The resulting mostly sparse, uneven or otherwise normality-violating distribution in this case results from distinct tissue expression profiles, as opposed to factors such as sequencing chemistry, gene structure, and similar variables that influence the distribution of reads aligning to distinct exons. The simple toy model of alignment across exons, however, provides an example to illustrate how biologically relevant features of a dataset cause highly non-normal distributions, and how to reason about such situations when hypothesis testing. This model will capture the intuition about how Chebyshev's theorem, extended from Markov's inequality, makes it possible to hypothesis test without assuming a normal distribution, and why such distributions can be important in RNA-seq data.

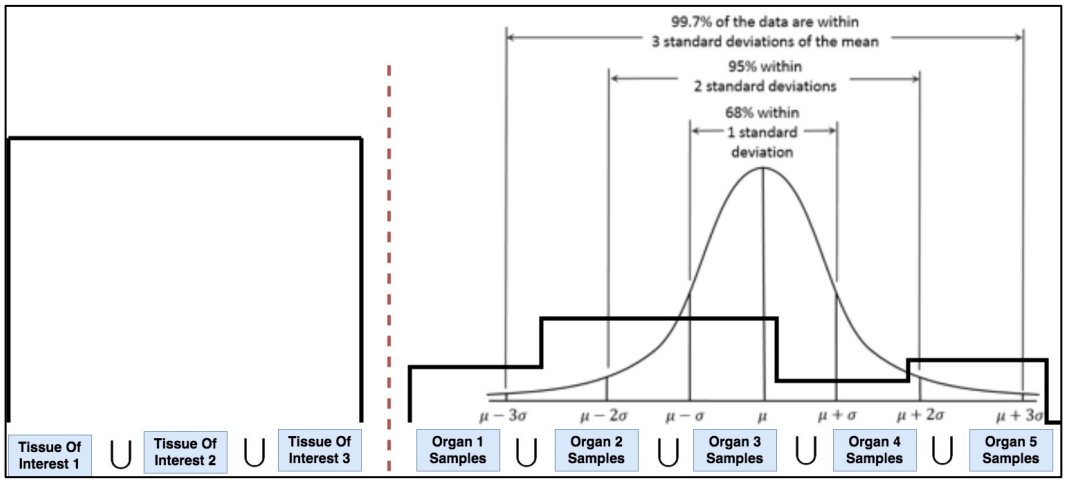

Figure 25: Distribution of reads across tissue of interest, with concentration of many reads in a few samples intuitively creates a violation of a normal distribution.

3.1.7 Using Chebyshev's Theorem, as an extension of Markov's Inequality

The influence of a strong concentration of aligned reads to only a few exons on the probability distribution of reads across all exons has bounds that can be understood through Markov's inequality – and this will lead directly to Chebyshev's inequality. Markov's inequality provides a rigorous description of how a concentration of probability associated with a small region of a function is related to the expectation of that random variable. An example of situation in bioinformatics data in this concept is useful is in understanding the probability that an exon receives a given number of aligned reads. During the normalization for read length in FPKM calculations, the probability of a read aligning to a region of the genome of any given length is assumed to be a constant, determined by the expression level divided by gene length. Figure 23, however, shows that a biologically realistic model that can violate this assumption. The consequences of this type of violation are shown in terms of probability distribution of read alignment across the gene (Figure 24). By taking an average, by definition, one loses considerable information about a distribution and individual samples that compose it. The average number of reads aligned to an exon would incorrectly suggest that each exon receives one read. This is highly inaccurate, however, as in fact one exon has five reads and all others have none.

In the type of situation illustrated in Figure 23, with bias and departure from an easily characterized distribution, the gap between the information expressed by the mean and the impossibility of repeated measurements close to or equal to the mean could skew one's understanding of the distribution. When the mean number of reads per exon is calculated for the gene above (Figure 23), on average each exon receives one read. This is misleading, because while sampling the exons, all except one exon will have no reads aligned. In reality, the probability of a given read aligning to Exon

1 is 1, and for all other exons, the probability of alignment to any other exon is 0. Markov's inequality ensures that in the case of the most extreme bias, or concentration of probability to a specific region, at most the concentration associated with a 1/nth region relative to the mean of the function is n times the mean. In our example, if the average reads across five exons is one, the greatest number of reads that could be contained in a length 1/5 of the gene (or a single exon) is five.

The number of read alignments mapping to the gene may be depicted through a probability distribution over each exon, creating a probability distribution function (PDF). Thus, statements that involve exons can be generalized to more abstract depictions that rely on more general descriptions of the domain under a probability curve. Markov's inequality does so and makes a formal statement about the limits of the expectation of the domain of a function and relates it to the probability distribution that derives from that function.

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

Figure 26: Formal statement of Markov's inequality (Wikipedia).

3.1.8 Read Concentration and Probability

Formally, Markov's Inequality (Figure 26) implies that if all sets, or measurements, of the random variable in our model - which in Figure 25 is the number of aligned reads to exons - are empty except for one, then the highest probability of

selecting the non-empty set will be less than or equal to the mean (one read per exon) divided by the number of total sets/measurements (five exons in the gene). Thus, the expected value of this non-empty set (the one exon with reads) defined as 1/nth the total domain of the function (all exons), is five times the expectation, or mean, of the random variable itself. Thus, the number of reads aligned to one or more non-empty exons can deviate strongly from the number of reads aligned to other exons. Moreover, when picking an exon at random and recording the aligned reads, selecting the exon that contains all five reads is a comparatively rare event. It is only expected to happen twenty percent of the time, as four other exons are have no aligned reads. As the number of exons in the gene increases, but the number of reads aligned stays constant, this would become even more rare. The bias, of course, would increase as well – with a single sample contributing a disproportionately to the dataset.

A similar effect may occur in the distribution of expression across tissue samples for a gene. In this case, most reads will be concentrated in a few samples in which that gene is highly expressed. This will influence the ability to determine enriched genes in a tissue of interest, among a set of background samples. Many of these highly expressing samples may be in the background tissues, which may have some transcriptome similarity to a tissue of interest. It will be important, however, to set bounds on how read concentration in a few samples will influence our knowledge of a null distribution describing gene expression across any type of measurement set.

In the example with the distribution across exons of a gene (Figure 23), the probability of selecting an exon with any reads aligned is a relatively rare event (regardless of how many reads are associated with that exon). With a high number of aligned reads concentrated in a single exon, though, this would not be clear from

averaging across all exons. For example, if one sampled the only exon with reads aligned to it and this exon, for the sake of experiment, was highly over-represented with 300 reads, while the other two exons had no reads, one could incorrectly estimate that on average each exon has 100 read alignments. Thus, intuitively, it can be hard to gain information about a distribution from averages that are biased by unlikely, but highly influential samples. This can be a challenge to hypothesis testing, which depends on determining the probability that a sample belongs to a null distribution, based on parameters inferred from sampling (mean and standard deviation).

Fortunately, the probability of a given sampling that contains these disproportionately influential observations can be quantified. Markov's inequality can be extended further to put bounds on such rare events as producing an unlikely sample, in a more general sense when these events are described in terms of the standard deviation. It is important to note that in Markov's inequality, both a rare observation and its probability are described as expectations of a random variable. Chebyshev's inequality, however, is more general, in that a rare event is described by distance from the mean in terms of standard deviation. As a general way to reason about probability distributions, it does not require normality.

Calculating the probability of an event through Chebyshev's inequality assumes the most relaxed bounds on the null distribution, since no specific distribution is assumed (Figure 27). Doing so results in the most stringent p-value, assuming all parameters are known exactly. Calculations using the empirical rule, which often motivates calculation of a z-score, are well within the bounds of Chebyshev's theorem. For example, assuming a standard normal distribution, 95 percent of observations will fall within two standard deviations of the mean. However,

$$
\begin{aligned}
\operatorname{Pr}(|X-\mu| \geq k\sigma) &= \mathrm{E}\left(I_{|X-\mu|\geq k\sigma}\right) \\
&= \mathrm{E}\left(I_{\left(\frac{X-\mu}{k\sigma}\right)^2 \geq 1}\right) \\
&\leq \mathrm{E}\left(\left(\frac{X-\mu}{k\sigma}\right)^2\right) \\
&= \frac{1}{k^2}\frac{\mathrm{E}((X-\mu)^2)}{\sigma^2} \\
&= \frac{1}{k^2}.
\end{aligned}
$$

Figure 27: Derivation of Chebyshev's inequality as following from Markov's Inequality (Wikipedia).

if the underlying null distribution is not normal, the correct p-value may be much higher than calculated through the empirical rule. Although z-scores are often assumed to follow a standard normal distribution, by the central limit theorem, we will develop a method that is still effective when this may not be the case, owing to the complex pattern of gene expression across tissues that will be skewed towards a few samples. Our tissue threshold for enrichment using a five standard deviation z-score derives simply from the fact that, regardless of the null distribution, 95 percent of the area will be under the curve in accordance with Chebyshev's theorem, regardless of any sampling effects on the z-score distribution. This is quite similar to the less stringent z-score approach mentioned by (Kryuchkova-Mostacci and Robinson-Rechavi, 2017), although under that heuristic only about 91 percent of the distribution is guaranteed to be under the null distribution. Figure 25, for example, depicts a realistic but non-normal distribution for expression across samples from different

tissues. This idealized distribution is quite similar to the example showing gene expression across exons of a single gene. Many genes will be sparse or very lowly expressed in most samples, while highly expressed in organs that share similar biological functions utilizing the gene of interest. The area under the distribution representing expression across tissues will be concentrated over these regions. If the difference of the mean expression of the gene in the samples belonging to the organ of interest is five times the standard deviation of the background dataset, the gene is considered enriched in that tissue. In order to determine enrichment at the .05 level of confidence, this threshold does not require assumption of the null distribution of expression across the general dataset. The most important assumption is that we test the observed mean of expression in the tissue of interest, assuming that we know it exactly (i.e. we do not take into account the variance of expression across samples of the tissue of interest). The justification for this is practical: otherwise, no genes are significant. These are important caveats, but provide a somewhat more rigorous approach than the purely ad-hoc methods of the GTEx study.

3.2 Results

The lists proposed by our enriched method are highly tissue-specific (Figure 21) while those identified by the GTEx analysis yield considerable overlap between breast muscle and cardiac muscle (Figure 28). Additionally, it can be shown that by incorporating the standard deviation of the data into account, we have improved resolution to detect genes that may be expressed at a consistently low to medium level across all background samples, but demonstrate exceptional levels in the tissue of interest. These genes may be associated with processes necessary to homeostasis in all tissues, but extremely important to organ specific biology.

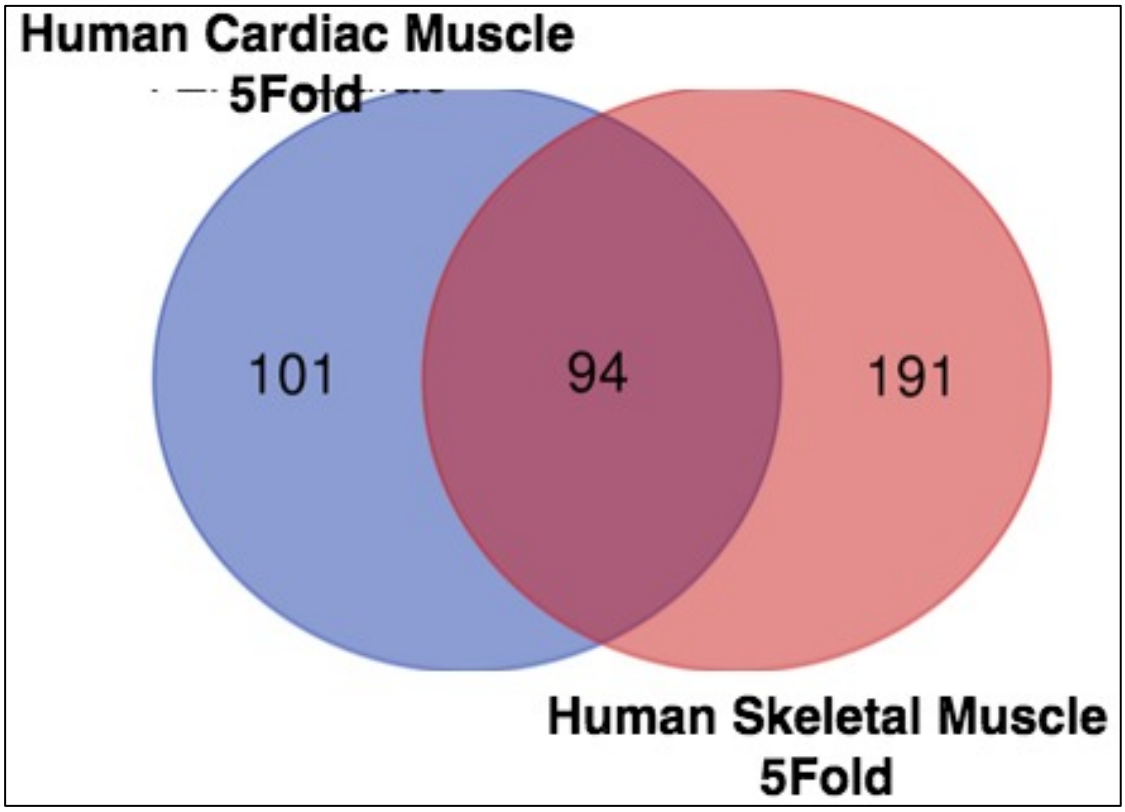

Figure 28: Tissue specificity comparison of GTEx methods (5-Fold higher in tissue of interest) in human, indicating inability of GTEx methods to identify only tissue unique genes.

3.2.1 Comparison with GTEx

For comparison, it is helpful to calculate enriched genes using the GTEx threshold and our five-standard deviation cutoff. These comparisons have a high degree of consensus between them. Overall, the five standard deviations approach

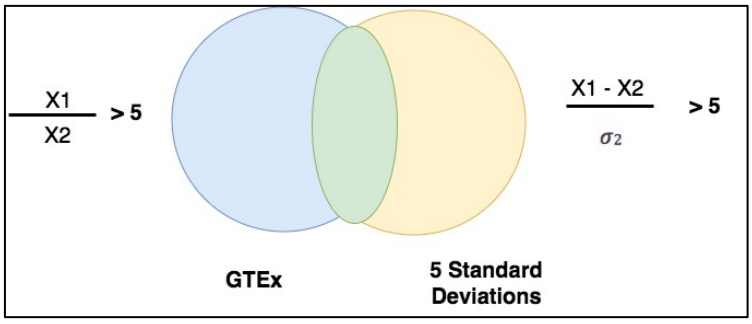

Figure 29: There is considerable overlap between the five standard deviation z-score method and GTEx standard of enrichment applied to our dataset. However, all of the genes identified as enriched according to the five standard deviation z score are unique to muscle type.

is more stringent, in terms of the smaller number of genes identified. It is worth noting that a number of the genes emerge as tissue-enriched in the five standard deviations approach that are not identified by the GTEx threshold. At least some of these genes represent important tissue-specific biology, and are identified by our threshold because they have low standard deviation among our background samples, in addition to having elevated mean expression in the tissue of interest. Our strategy can offer improved sensitivity for these genes. One example of this selection process is the Aconitase 2 (ACO2) gene enriched in cardiac tissue. The mean expression in cardiac tissue is 415 FPKM, while that in the background dataset is 92.37 FPKM. The standard deviation in the background tissue is 57.70 FPKM. In this case, the gene would not be identified as enriched by the GTEx definition. As a mitochondrial-specific gene associated with the tricarboxylic acid (TCA) cycle, ACO2 enrichment in cardiac tissue is consistent with the physiology of the heart, which has characteristically high-energy demands. By focusing on the standard deviation of background tissue, we focus on genes that are lowly, albeit relatively consistently

expressed in background tissue. It is important to mention that, by erring on the side of stringency, our tissue definition will lose sensitivity to genes with a background standard deviation that is relatively large compared to their mean. There are 179 of these such genes which are enriched in breast muscle according to GTEx methods leveraged on our chicken dataset, but not our five standard deviation based z-score (5SD) threshold (Figure 30) and 188 genes enriched in cardiac muscle according to GTEx methods leveraged on our chicken dataset, but not enriched according our z-score (Figure 31). However, by effectively penalizing genes with a relatively large standard deviation relative to their mean in background tissue, we are biasing against lowly expressed and mostly sparse genes. Instead, genes enriched in a given tissue according to our z-score are thus likely to be consistently, but comparatively lowly, expressed in background tissues relative to the tissue of interest. A pattern of consistent expression in all background tissues, but elevated expression in a tissue of interest implies a function in baseline processes which are intensified in the tissue of interest.

Figure 30: Venn Diagram of enriched genes in skeletal muscle according to the five standard deviation based z-score (5SD) threshold compared to calculations using the GTEx definition of 5 five-fold higher expression in tissue of interest.

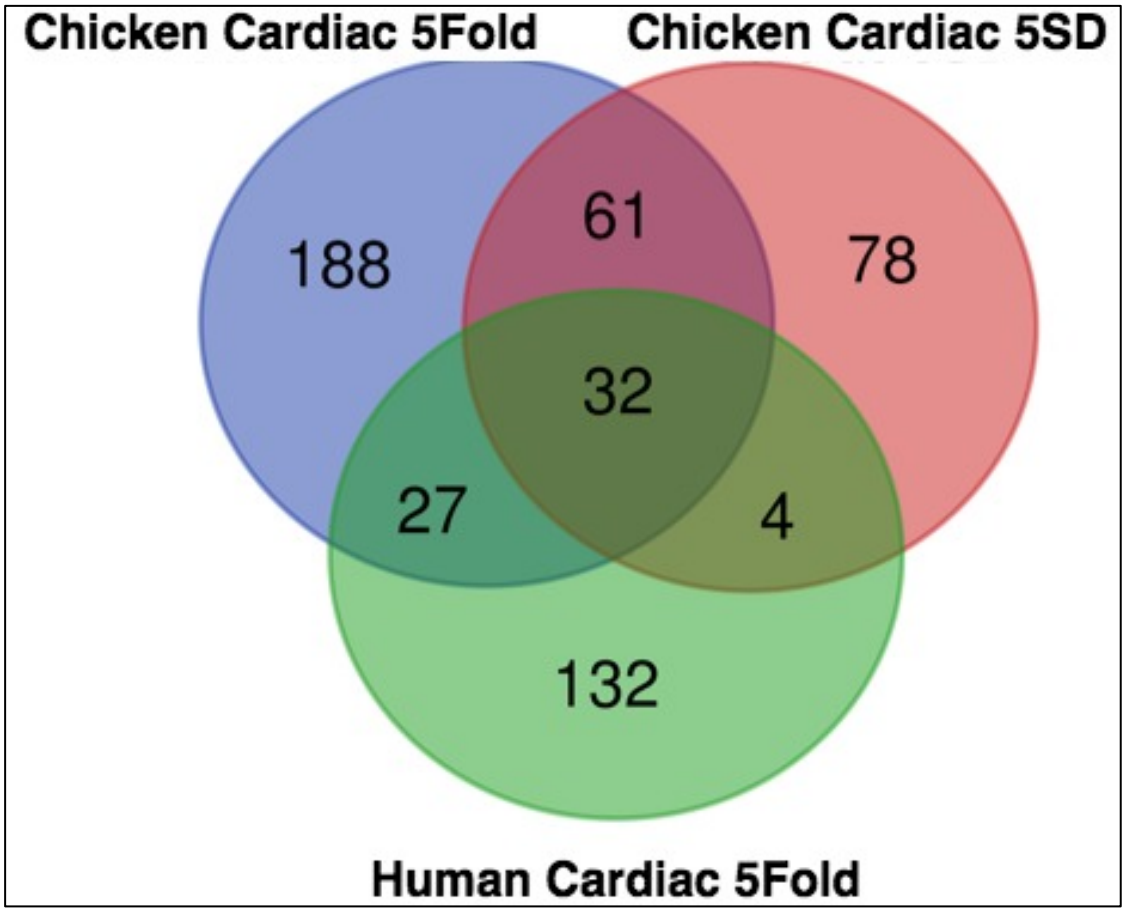

Figure 31: Venn Diagram of enriched genes in Cardiac Tissue according to the five standard deviation based z-score (5SD) threshold compared to calculations with the GTEx definition of 5 five-fold higher expression in tissue of interest.

We leverage these enrichment techniques to compare muscle tissue types in the modern broiler chicken because the effect of artificial selection for breast muscle size has been poorly understood in terms of consequences on cardiac systems (Tickle *et al*., 2014). Elucidating the underpinnings of muscle-type-specific biology is thus important to provide context for aspects of regulation that govern tissue-specific physiology. Enrichment analysis that provides sufficient resolution to separates

muscle-specific genes are required to be effective in understanding metabolic systems that differentiate between different muscle and other tissue types. These tissue-specific modules of enriched genes can provide the foundation for improved understanding of organ physiology. Notably, these are produced by our definition of enrichment, but not by the GTEx threshold.

3.3 Discussion: FAANG and Community Need for Enrichment Strategies

Tissue enrichment approaches can be used to improve biological understanding of individual tissues by isolating genes whose expression may be critically linked to tissue-specific function. This grows increasingly critical as complex datasets representing multiple tissues proliferate. One of the major goals of the post-genome era will be to gain a systems biology level understanding of individual tissues.

As annotations improve and sequencing data accumulates, researchers seek to understand model organisms at the functional genomics level. This level of resolution extends beyond the identification of genes and explores the complicated roles of regulatory elements. Regulatory elements that are the focus of functional genomics include insulators, enhancers, silencers and various types of promoters (Maston *et al*., 2006). A pioneering project that has sought to characterize the functional genomics landscape in humans with tissue enrichment and other approaches is the ENCyclopedia Of DNA Elements (ENCODE) project (The ENCODE Project Consortium, 2012). Complementary to ENCODE is the GTEx consortium, which aims to identify profiles of organ specific biology in health and disease, and whose enrichment protocol we have explored. The success of these initiatives has encouraged similar organizations in various research communities, to which our five standard deviation standard will be useful. One such effort in the animal genomics

community is the Functional Annotation of Animal Genomes Project (FAANG). FAANG is geared towards extending the genomics understanding of livestock species with established reference genomes and a community of committed researchers. Current species of interest include chicken, pig, cattle, and sheep. although the repertoire of species has been expanding (The FAANG Consortium, 2015). FAANG thus presents an opportunity to mine tissue-diverse datasets from a number of species.

Many of the analyzed genes emphasize unexpected levels of organ specificity, thus extending our biological understanding of these tissues. For example, many genes regulating protein degradation emerge as unique to breast muscle. Another class of enzymes encoded by genes that are strongly breast muscle specific are those that regulating splicing activities. The evidence of organ specific degradation and splicing machinery suggests new levels of complexity in transcriptome specificity in breast muscle, derived from the combinatorial arrangements of splicing and ubiquitination machinery. Other pathways, such as the gluconeogenesis/glycolysis cycle, exhibit a complex pattern of enrichment in both organs, with certain modules of these pathways containing genes enriched in cardiac tissue and others in skeletal muscle.

One method to identify important genes and possible protein-protein interactions in tissue-complex datasets involves implementing enrichment thresholds with subsequent network analyses. This provides important functional genomics information, extending approaches of the ENCODE and GTEx studies to the animal genomics community through FAANG. The insights provided by network analysis and enrichment complement one another, identifying specific network nodes that drive tissue specific biology. This provides a more nuanced description of transcriptome profiles shared between organs as well as those that are tissue specific.

3.3.1 Breast Muscle Ubiquitin Profile

Ubiquitination proteins are enzymes that identify other proteins for degradation, by catalyzing isopeptide bonds between a target protein and ubiquitin (Pickart and Ebbins, 2004). Ubiquitination is a highly represented process among the breast muscle specific genes. The ubiquitination process is a multi-step pathway, with each step controlled by different enzymes. The first phase is regulated by one of several ubiquitin-activating proteins, a family known as E1 proteins. An E1 protein activates ubiquitin by reacting with Adenosine Triphosphate (ATP). Activated ubiquitin is next transferred to an ubiquitin-conjugating enzyme (E2) that will interact with ubiquitin ligase (E3) to transfer the ubiquitin to the target protein (Ardley and Robinson, 2005). The diversity of ubiquitination reactions comes from the complexity of different arrangements of E1, E2 and E3. Specificity is most closely regulated by the E3 enzyme, however. There are nine known E1 enzymes, 26 E2 enzymes, and at least 25 E3 enzymes we have identified in chicken. Crucial biology that regulates these processes is still being determined (Lee and Zhou, 2007). Thus, information describing tissue-specific profiles of members of this process is particularly valuable. The number of ubiquitination-related enzymes among breast muscle enriched genes suggests these compounds may encouraging differentiation and maintain organ homeostasis by regulating protein levels through controlled degradation.

The ubiquitin-related genes enriched in breast muscle encode groups of proteins that share interacting partners and motifs. Many of these are Ankyrin repeat and suppressor of cytokine signaling (SOCS)-box containing proteins, for example. These include multiple Ankyrin and SOCS-Box containing proteins (ASB) ASB10, ASB11, ASB14, ASB15, and ASB2. Various ASB proteins are known to interact with Cullin 5 (CUL5) and the Really Interesting New Gene (RING) protein, RING-

Box protein 2 (RBx2) to form ubiquitin E3 complexes (Kohroki *et al*., 2005). These protein complexes play important roles in regulation and development. Importantly, CUL5 is also enriched in breast muscle and is hypothesized to exert an inhibitory influence on cell proliferation by interacting with the SOCS/BC-box/ eloBC/cul5/ RING E3 complex (Petroshki and Deshaies, 2005). The consistent pattern of enrichment of genes encoding ASB proteins and CUL5 is consistent with a regulatory module that controls protein degradation via ubiquitination in a highly specific fashion. Research into hemopoetic stem cells, for example, has revealed that ASB2-alpha causes filamin breakdown through ubiquitination while sparing its other substrates, Janus Kinase Proteins (Lamsoul *et al*., 2012). This may be critical for regulation of tissue defining proteins by controlling their targeted degradation.

Most breast muscle enriched genes associated with ubiquitination encode E3 enzymes. However, an E2 enzyme, ubiquitin conjugating enzyme E2 G1 (UBE2G1), is also enriched. This finding is consistent with the hypothesis that UBE2G1 is specific to muscle protein degradation (RefSeq). Supporting an additional level of tissue specificity derived from ubiquitin regulation, ubiquitin specific peptidase (USP2), is enriched in breast muscle. De-ubiquitination targets of USP2 include mouse double mutant (MDM) genes MDM2, MDM4 and cyclin D1 (RefSeq). This set of ubiquitination related genes enriched in breast tissue suggest protein degradation plays important roles in breast muscle physiology. This is important, as understanding exaggerated muscle characteristics is an important goal for broiler genetics. Other enriched genes, such as transcription factors, have a more explicit link to tissue development and differentiation.

3.3.2 Breast Muscle Transcription Factors

Transcription factors encode a diverse set of proteins that selectively influence the expression of other genes. They are critical to development and differentiation (Spitz and Furlong, 2012). The role of each transcription factor is determined by its structure and functional domain. Due to this complexity, transcription factors are able to influence a diverse set of biological pathways. One domain associated with development-related transcription factors is the homeobox. Homeobox containing proteins are distinguished by a homeobox domain and include several well characterized families such as the homeobox (HOX) and paired box (PAX) families (Holland et al., 2007). There are four groups of HOX genes, the A, B, C, and D families. A number of the HOXA transcripts are enriched in breast muscle, with many of them playing a role in developmental regulation. HOXA7 functions in cell proliferation through a number of possible mechanisms that are still being elucidated (Li et al., 2014). Other HOX proteins play similar roles, with their functions often defined through knockdown experiments. Knockdowns involving HOXA6 and PBX3 interrupt cancer-related proliferation and enhance susceptibility to chemotherapy (Dickson et al., 2013). HOXA3, meanwhile, has been shown to be important to endothelial cell differentiation, with its levels being decreased during differentiation as HOXB3 and HOXA7 increase (Chung et al., 2005).

Other breast muscle enriched transcription factors include non-HOX genes such as the myogenic transcription factors (MYF) MYF5 and MYF6. Myogenic factors are a class of transcription factors regulating muscle growth. MYF6 mutations have been linked to the severe course of Becker muscular dystrophy (Kerst *et al*., 2000). MYF6 is known to encourage muscle regeneration by promoting myoblast amplification. MYF5 mutations are associated with severe deficiencies in myoblast

proliferation (Ustanina *et al*., 2006). Another myogenic gene, MYOG, is critical to skeletal muscle formation, with mutations causing neonatal death (Hasty *et al*., 1993). In addition to explicitly myogenic proteins, nuclear factor of activated proteins (NFATC1) regulates immune function and cellular plasticity (Chen *et al*., 2017). Expression of this gene correlates well with the ubiquitin-related CUL5, suggesting a possible interaction or shared membership in a small network (Figure 30). Immune system transcription factors such as NFATC1 may influence cell proliferation by regulating inflammation and apoptotic pathways. Another immune system related protein that influences transcription is the protein protein kinase C theta (PRKCQ), which activates NF-kB and may link T-Cell activation other transcription factors (RefSeq). Inhibition of PRKCQ expression has been shown to counteract muscle disease in a mouse model of Duchenne's Muscular Dystrophy, by preventing inflammation that impedes muscle regeneration (Marrocco, 2017). Paired-box 7 (PAX7), another transcription factor upregulated in breast muscle, interacts with myogenic factors, with PAX7 and MYOD1 coexpression being associated with activation of quiescent cells (Zammit *et al*., 2006). Single minded family bHLH transcription factor 2, SIM2, another up-regulated transcription factor is associated with neurogenesis (Chrast *et al.,* 1997). Consistent with their functional significance, these transcription factors emerge as hubs in correlation networks (Figure 32). Other forms of transcriptional regulation in skeletal tissue, beyond the direct influence on expression from transcription factors include alternative splicing. A number of alternative splicing genes are also upregulated in breast muscle, along with other genes that may directly influence physiology.

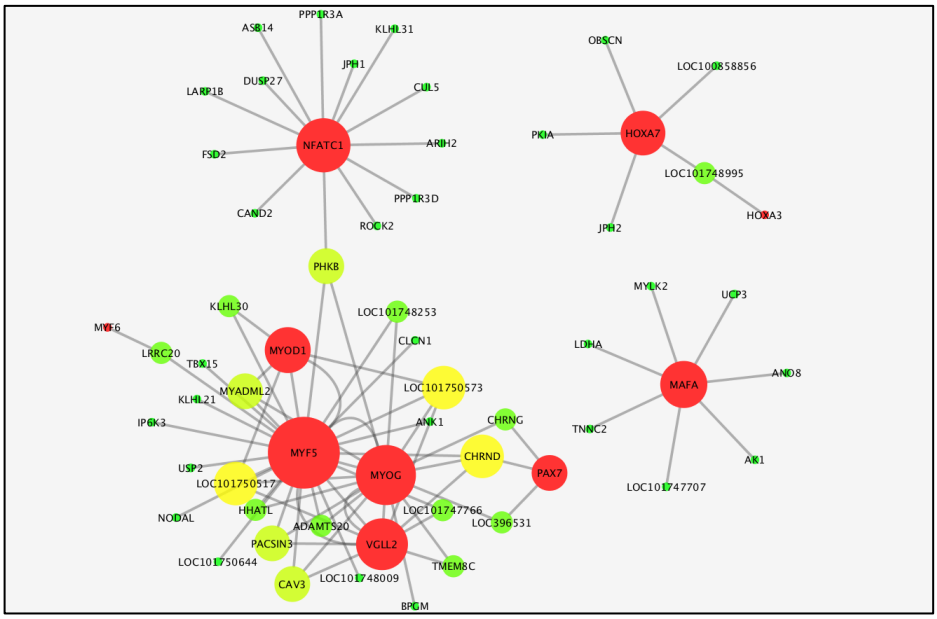

Figure 32: Transcription factors enriched in skeletal muscle according to the 5SD based z-score, breast muscle and which have statistically significant correlations with other enriched genes in the tissue. Transcription factors in red. Node size is reflective of number of interacting partners.

3.3.3 Breast Muscle Spliceosome and Metabolic Physiology

RNA splicing is an important aspect of gene regulation that ensures production of gene isoforms. Several of these are enriched in breast muscle tissue. The breast muscle enriched MYOD1 gene, for example, encodes a protein that regulates cell differentiation by controlling cell cycle arrest. Myogenic Differentiation 1 (MYOD1) expression induces alternative splicing (Ichida *et al.*, 1998) that is essential to myogenesis and is one of two spliceosome related genes enriched in breast muscle. Another gene controlling splicing activity, Breast Carcinoma Amplified 2 (BCAS2), is a major component of the CD5CL/Prp19 complex, which is in turn critical to catalytic activation of the spliceosome (Liu *et al.*, 2016). These may provide regulation at a

scale of isoform-level resolution that is likely critical to the transcriptome profile of breast muscle.

Beyond identifying genes driving tissue-specific morphogenesis and differentiation, this analysis detects enrichment of genes that may be control the transmission of short-term physiological signals. For example, several components of nicotinic acetylcholine receptors, the cholinergic receptor nicotinic (CHRN) genes, are enriched: CHRNA1, CHRNA9, CHRND, and CHRNG. Each of these genes encodes a protein that contributes to the assembly of the multi-component acetylcholine receptor. As a well structure, this receptor is essential to relay neural signals that initiate muscle movements. Consistent enrichment for its various components indicate the relative importance of the receptor to muscle specific physiology. This also shows the importance of the transcriptome in maintaining a physiological homeostasis that allows the bird to quickly respond to stimuli.

Additional genes regulating physiological or metabolic processes include those related to gluconeogenesis and glycolysis. These pathways, which encompass the production and breakdown of sugars, respectively, are critical for extracting energy from sugars. Fructose-bisphosphatase 2 (FBP2) a gene critical to gluconeogenesis, is enriched in breast muscle. FBP2, which encodes an enzyme that hydrolyzes fructose 1,6-bisphosphate (F1,6BP) to fructose-6-phosphate (F6P), is also hypothesized to play an important role in glycogen production as well as mitochondrial health (Pirog *et al*., 2014). Lactate Dehydrogenase A (LDHA), a gene with a prominent role in lactic acid metabolism is also enriched. This is essential to manage the final products of glycolysis under anaerobic conditions. Several enriched genes play roles in steps of glycolysis upstream of LDHA, though LDHA activity is critical enough such that its

inhibition interferes with the enhanced glycolytic stress in cancer cells (Le *et al*., 2010). One such enriched gene upstream of LDHA is glyceraldehyde-3-phosphate dehydrogenase (G3PD), which functions in glycolysis and also functions as an apoptosis-influencing transcription factor (Tarze *et al*., 2007). Another glycolysis-associated gene, phsophoglucomutase (PGM1), regulates an important pathway branch point that interconverts Glycerol-1-Phosphate (G1P) and Glycerol-6-Phosphate (G6P), with the latter being an intermediate to glycolysis and the former serving as a precursor for structural carbohydrates. PGM1 has been shown to be necessary for cell growth under glucose depletion (Bae *et al*., 2014). A final gene associated with sugar metabolism, Bisphosphoglycerate Mutase (BPGM), has until now been associated only with erythrocyte and placental tissues (Pritlove *et al*., 2006). BPGM is involved in the synthesis of 2,3-diphosphoglycerate (2,3-BPG) from the glycolysis intermediate 1,3 biphosphoglycerate (1,3-BPG). 2,3-BPG shifts the equilibrium of hemoglobin towards the deoxygenated state. The enrichment of BPGM in muscle suggests a role for the protein in oxygen transfer in muscle tissue. The diverse set of enriched genes regulating the glycogen/glucose pathways recapitulates known biology, while also suggesting novel relationships that underpin breast muscle specific physiology. Understanding steps of these pathways, which are shared among different muscle types, provides a comparative lens to understand how different types of muscle, i.e. cardiac tissue and breast muscle, differ from one another.

3.3.4 Cardiac Enriched Genes - TCA Cycle Metabolism and Mitochondrial Genes

Cardiac tissue, like skeletal muscle, is enriched for several genes that drive sugar metabolism. However, the functions of these genes are distinct from metabolic

genes enriched in skeletal tissue. These genes enriched in cardiac tissue encode proteins that regulate the intersection of different types of metabolism. Many of these enzymes form multimeric complexes. Pyruvate dehydrogenase A1 (PDHA1), for example, is a subunit of the pyruvate dehydrogenase complex. The pyruvate dehydrogenase complex serves an important function by linking glycolysis to the Tricarboxylic Acid (TCA) cycle through catalyzing the oxidative decarboxylation of pyruvate (Holness et al., 2003). The cardiac muscle enriched gene, pyruvate dehydrogenate kinase 3 (PDK3), meanwhile, inhibits the activity of the pyruvate dehydrogenase complex. This inhibition occurs by phosphorylating the alpha subunit of the pyruvate dehydrogenase complex, which is encoded by PDHA1 (Korotchkina and Patel, 2001). The enzymes encoded by the enriched metabolic genes PDK3 and PDHA1 function in the mitochondria.

Other cardiac enriched genes related to metabolism regulate lipid metabolism. 3-hydroxybutarate dehydrogenase 1 (BDH1), for example, is allosterically activated by phosphatidylcholine (Green *et al*, 1996) and interconverts the products of fatty acid catabolism acetoacetate and (R)-3-hydroxybutyrate. Glycerol-3-phosphate dehydrogenase 1 like (GPD1L), and the homologous glycerol-3-phosphate dehydrogenase 1 GPD1, converts Dihydroxyacetone Phosphate (DHAP) to Glycerol-3-Phosphate (G3P) (Reactome, Ou et al. 2006, Valdivia et al. 2009). The presence of enriched metabolic transcripts related to both sugar and fat metabolism emphasizes the flexibility of energy production in cardiac tissue. This is distinct from metabolic genes enriched in breast muscle tissue, which relate primarily to glycolysis and glycogen synthesis.

At the five standard deviation threshold, cardiac tissue is not enriched for some classes of genes enriched in breast muscle, such as those that regulate alternative splicing and protein degradation. However, just as spliceosome-related genes are enriched only in breast muscle, one class of genes unique to cardiac tissue are those that influence metabolism through controlling mitochondrial metabolism. These genes regulate a number of processes that are localized to the mitochondria, some of which regulate the TCA cycle. The nuclear-encoded gene for the mitochondrial protein 3-Oxoacid CoA Transferase 1 (OXCT1), for example, is the rate-limiting step in ketolysis (Shafquat *et al*., 2013). Ketolysis intersects with many metabolic pathways including fatty acid oxidation, the TCA cycle and gluconeogenesis (Cotter *et al*., 2013). Consistent with the importance of these processes to cardiac tissue, the gene for the mitochondrial protein ACO2 is enriched. ACO2 catalyzes the second stage of the TCA cycle, the inter-conversion of citrate and isocitrate. Other metabolism related genes regulate amino acid breakdown. Acyl-CoA Dehydrogenase, short/branched chain (ACADSB) dehydrogenates acyl-coA derivatives and catabolizes leucine (Andresen *et al*., 2000). Another mitochondrial specific gene enriched in heart that specializes in leucine metabolism is Isovaleryl-CoA dehydrogenase (IVD). The enzyme encoded by IVD functions in the mitochondria as a highly specific Isovaleryl-CoA dehydrogenase (Ikeda and Tanaka, 1983) and is important to valine, leucine and isoleucine catabolism. Perhaps to mitigate the stress of many complex metabolic systems operating simultaneously, cardiac tissue is also enriched for heat shock proteins linked to heart performance. The heat shock protein, heat shock protein family B1 (HSPB1), which is enriched in cardiac tissue, has also been linked to mitochondrial characteristics of heart failure (Marunouchi *et al*., 2013) that may be

related to decreased translocation efficiency into the mitochondria (Marunouchi *et al*., 2014). Another heat shock protein, HSPB7, is enriched in cardiac tissue. Heat shock protein family B7 (HSPB7) is thought to be associated with splicing, a role also shared with HSPB1 (Vos *et al*., 2009). There are no heat shock proteins (HSP) proteins enriched in breast muscle, consistent with the hypothesis that these genes represent biology unique to the enhanced metabolic demands of cardiac tissue. Enrichment for mitochondrial-related genes is unique to cardiac tissue, suggesting that the metabolic contribution of the organelle is more important in cardiac tissue compared to skeletal muscle. This profile is consistent with the importance of fatty acid oxidation to cardiac function.

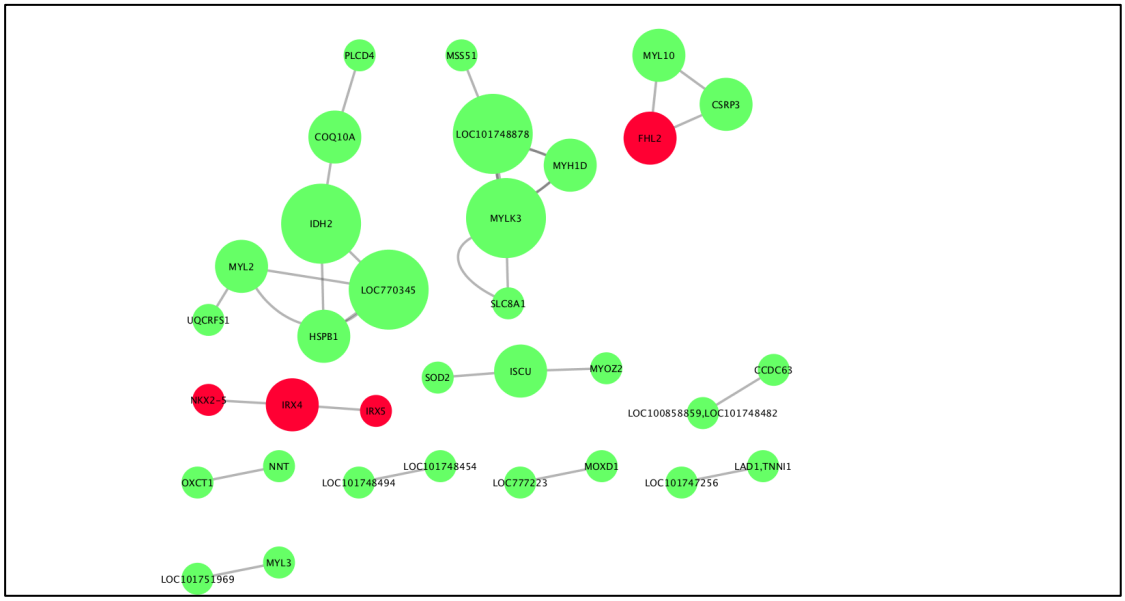

Figure 33: Enriched transcription factors in cardiac muscle, according to the 5SD-based z-score, that have statistically significant correlations with other enriched genes in the tissue. Node size reflective of interacting partners, and transcription factors indicated by red color.

3.3.5 Cardiac Transcription Factors

The number of genes related to transcription factors enriched in cardiac tissue, as in breast muscle, is considerable. Transcription factors enriched at threshold in cardiac tissue include HOXD13, DMRT Like Family B with Proline Rich C-Terminal 1 (DMRTB1), NK2-Homeobox 5 (NKX2-5), PDZ and Lim Domain 1 (PDLIM1), Iroquous Homeobox 4 (IRX4), Iroquous Homeobox 5 (IRX5), Four and a Half LIM Domains (FHL2) and T-Box 20 (TBX20). HOXD13 has been primarily associated with limb development and deformity (Davis and Cappechi, 1996). Other enriched transcription factors are also not canonically associated with cardiac tissue. One gene, DMRTB1, is poorly characterized in cardiac tissue though the closely related paralog, Doublesex and Mab-3-Related Transcription Factor 1 (DMRT1) is critical for sex differentiation (Ottolenghi *et al*., 2002). However, the roles of other enriched transcription factors in cardiac tissue are well established, however. NKX2-5, for example, is critical to heart development and known to be involved in proper differentiation of cardiac tissue (Jay *et al*., 2004). The enriched gene PDLIM1 encodes a protein that possesses two PDZ domains and three LIM domains; this structural complexity gives the protein a variety of functions. One of these is inhibition of the transcription factor NF-kB (Ono *et al*., 2015). Other activity includes interactions with actin types 1 (Kokota *et al*., 2000) and 2. Though not a transcription factor itself, PDLIM1 may play a critical role in cardiac tissue by influencing the behavior of the transcription factor NF-kB (Ono *et al.,* 2015), as well as influencing actin structure (Vallenius *et al*., 2000). The gene for another LIM domain containing protein, FHL2 is enriched in cardiac tissue. FHL2 is associated with cell proliferation, and implicated in a number of cancers (Wang *et al*., 2016). FHL2 is also thought to influence formation of extracellular membranes (RefSeq, 2017). A member of the

TBX family of transcription factors, TBX15, is additionally enriched in cardiac tissue. A member of this family, TBX20, is enriched in breast muscle but not cardiac tissue. TBX20, however, is a member of a conserved network of transcription factors that plays a role in cardiac development across species. Mutations in the gene cause a number of cardiac deformities (Kirk *et al*., 2007). At least one transcription factor, the Iroquois homeobox conataining gene, IRX5, is implicated in physiology, as this gene is essential to controlling the cardiac repolarization gradient (Bruneau *et al*., 2006). Another Iroquois homeobox containing gene, IRX4, is enriched in cardiac tissue. The IRX4 gene influences atrial development, and is associated with cardiac hyperthrophy (Bavrak *et al*., 2008).

The specificity of the gene lists that represent each tissue includes a number of genes associated with lethal mutations. The representative genes associated with both tissues contain members of pathways that regulate muscle structure at a foundational level. This can be further understood by exploring the enriched text mining terms associated with the gene lists from each tissue. It is important to note that fewer cardiac transcription factors have significant correlations with other tissue enriched genes (Figure 33. A similar diagram of breast muscle enriched transcription factors that correlate with other genes enriched in that tissue is shown in Figure 32.

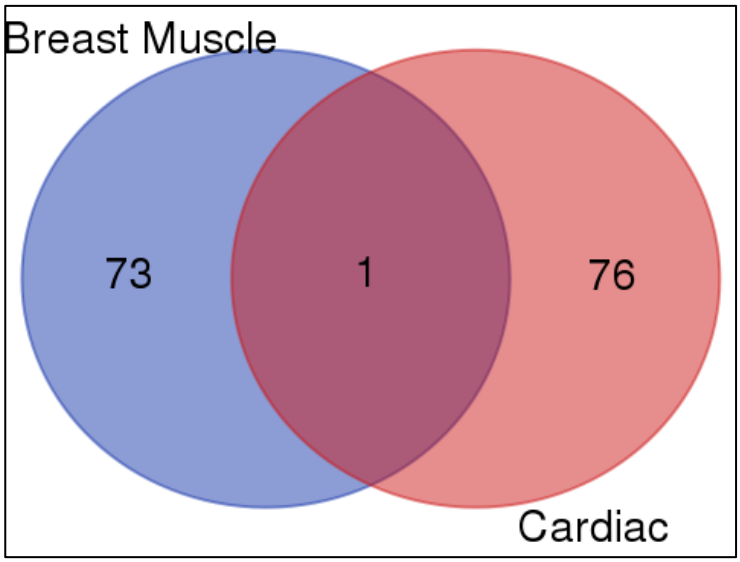

Figure 34: Venn diagram of text mining terms associated with the enriched gene lists, determined by the 5SD based z-score in breast and skeletal muscle tissues.

3.3.6 Text Mining Comparison and Structural Differences

Despite a lack of overlap in the lists of enriched genes in the two organs, at least two broad classes of genes – transcription factors and those that regulate metabolism – seem to play similar roles in cardiac tissue and skeletal muscle. However, when exploring terms enriched in each gene list with the text-mining tool DAVID (Huang et al., 2009), the resulting terms are highly organ specific (Figure 34). This emphasizes the distinction between cardiac and skeletal muscle by considering information about each gene in the previous literature. Using this approach, there is only one term to which a significant number of genes from both tissues apply: sarcolemma. The sarcolemma is a membrane that sheathes striated muscle fiber cells. The cardiac enriched genes relating to this shared term, sarcomere, are Blood Vessel Epicardial Substance (BVES), Popeye Domain Containing 2 (POPDC2),

Tropomodulin 1 (TMOD1). The breast muscle genes mapping to this term are: Caveolin 3 (Cav3), ryanodine receptor 3 (Ryr3) and Sargoclycn Delta (SGCD). BVES (also known as POPDC1) is a cell adhesion protein with a redundant function shared with POPDC2 (Brand et al., 2014). TMOD1 encodes an actin capping protein that binds to the N-terminal of tropomyosin in order to control depolymerization, and thus length of the fibers and shape of the erythrocyte membrane (RefSeq, 2017). Cav3, enriched in breast muscle, is another gene associated with molecular control of muscle structure and its disregulation is associated with muscular dystrophy (Deng et al., 2017) and myasthenia gravis. SGCD plays an important structural role, encoding a member of the sarcoglycan complex, which in turn contributes to the larger dystrophin-glycoprotein complex (RefSeq, 2017). Ryar3 is a ryanodine receptor that functions as a calcium channel (Sorrentino et al., 1994). The different profiles of sarcolemma-related genes between cardiac and skeletal muscle may relate to tissue specific structural organization of fibers. Ryanodine receptors, for example, tend to cluster around T-tubules (Fleischer et al., 1998). T-tubules are extensions of the cell membrane in cardiac and skeletal tissue that begin at the sarcolemma and pass into the interior of the cell and are crucial for the transport of calcium ions necessary for contraction (Hong et al., 2017). BVES and Cav3, two genes enriched in cardiac tissue, are known to co-localize to T-tubules (Alcalay et al., 2013). Ryanodine receptors cluster around T-tubules and their coupling with other proteins are associated with an accelerate EC transmission in skeletal muscle (2ms versus 100 ms), relative to cardiac muscle (Al-Qsairi et al., 2011). Ryr3, which is enriched in breast muscle, controls resting Ca2+ in skeletal muscle (Perez et al., 2005). While the skeletal tissue and cardiac tissue samples demonstrate enrichment of different sarcolemma-related genes

associated with t-tubules, cellular and physiological mechanisms controlling behavior of T-tubules are still being elucidated (Al-Qusairi et al., 2011).

3.3.7 Selective Enrichment of TCA Cycle Genes

While similar to previous studies that develop tissue specific gene expression profiles, such as those leveraged by the GTEx consortium, we have demonstrated a more tissue-specific enrichment threshold. An advantage of this stringency is that resulting gene lists emphasize differences in organelle characteristics between cells of each tissue type. A key insight from this level of resolution is the degree to which mitochondrial proteins drive cardiac physiology, as compared cytoplasmic glycolysis related genes being enriched in breast muscle.

The emphasis on different cellular regions among the enriched genes in each tissue provides important information about how the transcriptome influences physiology. Glucose and glycogen related genes enriched in skeletal muscle tissue, for example, are primarily cytosolic. Cardiac tissue, meanwhile, is enriched for many nuclear encoded mitochondrial proteins that control the TCA cycle. This is consistent with the need for cardiac tissue to have access to ATP (adenosine triphosphate) to fuel the continual muscle contractions associated with heartbeats. Skeletal muscle tissue, meanwhile, depends on a more diverse set of carbohydrate pathways for energy which

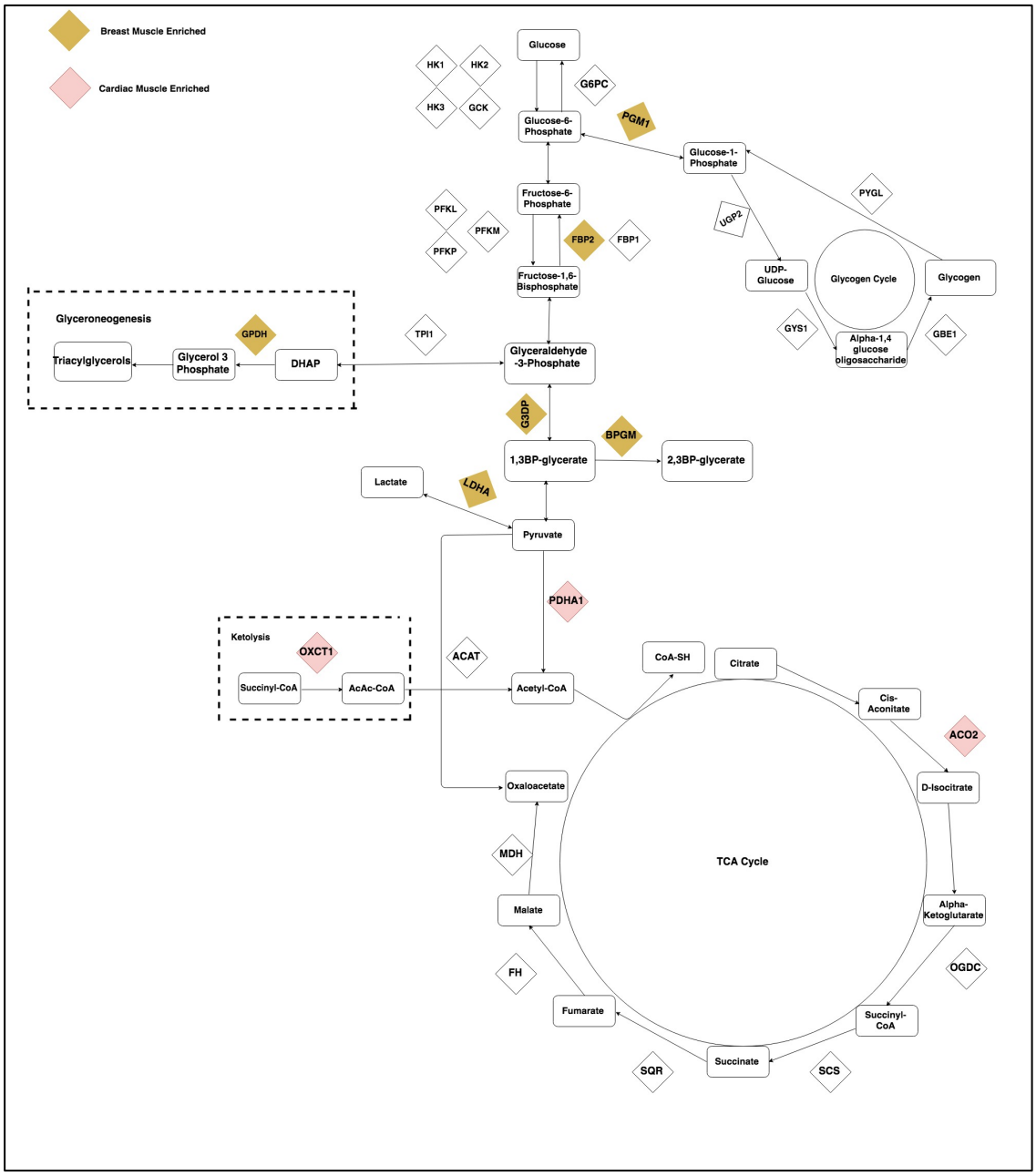

Figure 35: Diagram of TCA cycle and genes related ketone and glycogen metabolism, emphasizing genes that are enriched in breast muscle or cardiac tissue. Enzymes encoded by TCA cycle genes generally function in mitochondria. Glycolysis/gluconeogenesis genes are cytosolic.

must control the careful regulation of glucose and glycogen (Ivy, 1991). The compartmentalization of the metabolic pathways enriched in cardiac tissue and breast muscle can be compared (Figure 35) to understand how these pathways relate to one another despite intracellular separation.

3.3.8 Relationship Between Metabolism and Organelles

This strategy of tissue enrichment provides important insight linking the transcriptome to physiology. However, it also raises many questions about the organelle-specific differences between the two muscle types. The extent to which enrichment for mitochondrial genes represent differences in mitochondrial chemistry between heart and skeletal tissue as opposed to greater mitochondrial number, however, is unclear. Previous studies have shown that while cardiac tissue contains a higher concentration of mitochondria, physiological differences between tissues disappear under normalization for mitochondrial density (Park et al., 2014). Whether or not these differences in enriched genes between muscle types result from increased mitochondrial concentration, they are consistent with different physiological roles for skeletal and cardiac muscle tissue. These functional discrepancies are also emphasized by text mining analysis of enriched gene lists. In fact, only one text mining term is enriched in the lists of tissue-defining genes for both cardiac and breast muscle-defining: sarcolemma. The sarcolemma is a general feature of muscle cells, consistent with its importance in both breast and heart tissue. However, the genes associated with the sarcolemma are different between the two tissues, suggesting the regulatory environments may influence the structure differently in each of the two tissues.

The most significant of these differences involve receptors whose activity regulates signals that trigger muscle contractions, such as the ryanodine receptors. One such receptor, Ryr3, is enriched in breast muscle though expression of this gene is most commonly associated with brain tissue (Zucchi *et al*., 1997). However, Ryr3 is prominently expressed in immature muscle tissue, with its later replacement by Ryr1 during maturation (Smith and Lieber, 2013). Ryanodine receptors control the release of calcium from the sarcoplasmic reticulum in order to control muscle movements (Santulli and Marks, 2015). Ryr3, in particularly, is associated with transient calcium signaling events (Ward and Rodney, 2008). No ryanodine receptors are enriched in cardiac tissue at our threshold, despite cardiac muscle being the primary tissue for expression of Ryr2. Ryr1 is also enriched in breast muscle. These differences perhaps reflect divergent developmental trajectories for cells of each tissue. Several genes associated with a separate set of receptors show similar specificity to breast muscle.

3.3.9 Value of Enrichment Threshold: Comparative Evolution and Feature Subsetting

While tissue enrichment strategies have contributed to better understanding of the human transcriptome across organs through the GTEx and other initiatives, a similar systems level understanding of gene expression across animal species is lacking. This work provides a method for mining tissue-diverse datasets that are increasingly common in animal genomics, and produces novel biological findings. The enrichment of BPGM in the breast muscle samples, for example, represents a potential coupling of glycolysis and aerobic metabolism. BPGM processes a glycolytic intermediate into a compound 2, 3-diphosphoglycerate that makes the deoxygenated state of hemoglobin more favorable. This is critical for the rapid

transfer of oxygen. BPGM is enriched in chicken breast muscle according to both the GTEx five-fold difference in means, as well as the five standard-deviation z-score. This robust enrichment pattern in chicken, but not human, suggests it may be species specific and biologically consequential. For example, this adaptation may be valuable to chicken and other birds, allowing for more effective oxygenation especially to skeletal muscle tissue. This could be an important evolutionary feature related to the increased energy demands of muscles involved with flight and other energy-demanding movements. Importantly, the mean expression of BPGM in breast muscle among broilers and a line of chicken spared intense contemporary breeding pressure, the Illinois line, is not significantly different. This is consistent with BPGM expression being a defining feature of breast muscle across chicken breeds. Another informative, novel finding resulting from applying the 5SD threshold is the enrichment of DMRTB1 in cardiac tissue. DMRTB1, a transcription factor, is canonically linked to a family of sex differentiation factors. The observation that the gene is enriched in cardiac tissue suggests that this family of transcription factors may be far more versatile than previously thought. The biological role of DMRTB1 could be unique to birds, as well, as it is not enriched in the human GTEx data. DMRTB1 is lowly expressed across all background tissues, if it is present in them at all, with a mean FPKM of 0.67.

The fact that these genes, BPGM and DMRTB1, are not identified as enriched in the original human GTEx dataset, but that both pass the GTEx threshold for enrichment (five-fold increase in means) as well as the 5SD z-score, shows the usefulness of using tissue enrichment for comparative biology. Additionally, our 5SD threshold is far more stringent than the GTEx standard, in terms of identifying organ

specific enrichment among a tissue diverse data set. However, we have also shown that it recapitulates established tissue associated biology, while also proposing novel relationships for many enriched genes. Thus, we will use this method of tissue enrichment as a first step of feature selection in downstream pipelines that explore regulation of the heat stress response. We will use this standard of tissue enrichment, for example, to identify liver specific genes before using statistical learning techniques to associate liver enriched genes and metabolite data with the heat stress response. By reducing the set of genes from over 20,000 to several hundred through this standard of tissue enrichment, we have immediately reduced the number of features associated with gene expression. The result is a module of highly tissue-enriched genes. Although our heat stress response studies will focus primarily on the liver, we have demonstrated the general performance of our enrichment strategy by clarifying the transcriptome similarities and differences in two closely related tissues, cardiac and skeletal tissue.

Chapter 4

FROM ORGAN-ENRICHED MODULES TO MECHANISMS

4.1 Introduction: Context for Metabolic Forks

The need to identify core regulatory modules among organs is an important motivation for multi-step informatics pipelines. To develop these, a first step is to subset by organ-enriched genes, for initial feature selection. We have demonstrated the efficacy of using a z-score enrichment strategy on our dataset in Chapter 3. Subsequently, we develop pipelines to analyze metabolomics data in a way that is complementary to and extends the work of tissue enrichment analyses. Modules of genes identified by tissue enrichment are incorporated as a form of initial feature selection in these pipelines. This approach of applying pipelines in an iterative fashion makes it possible to extract the maximum biological meaning from our dataset. We demonstrate this, by producing novel biological findings that enhance understanding of the heat stress response. These chapters will explore the heat stress response from the perspective of liver metabolism. While we have examined muscle-specific difference using tissue enrichment to demonstrate the specificity and usefulness of those techniques, in subsequent chapters we focus on using computational techniques to understand regulation in the liver because it a metabolic powerhouse for the chicken, managing sugar, lipid and antioxidant production (Jastrebski *et al*., 2017)

While enrichment analyses identify tissue-defining genes and DiffExpress detect heat stress responsive genes, those methods only investigate gene expression changes, not how they relate to one another. Such methods also suffer from the

shortcoming that they do not provide direct insight into regulatory changes at the biochemical level, in terms of metabolites. High throughput metabolomics data, however, provides a quantitative way to assess levels of biologically active compounds (Fuhrer and Zamboni, 2015). Though such techniques bring great promise to life science research, there is a need for studies to develop informatics approaches to extract biological insight from large-scale datasets that combine multiple –omics data types (Johnson *et al*., 2014). Subsequently, we develop a strategy to detect regulatory mechanisms at the metabolite level that can then be explored as possible consequences of transcriptome shifts. The transcriptome shifts that underlie changes in metabolites can be contextualized in terms of the modules of tissue-specific genes. Accomplishing this task is multi-tiered, and will require pipelines that use statistical learning techniques to exploit various biologically informative features of the data. This will include modeling ratios of metabolites in terms of other possibly related compounds and precursors, as well as prioritizing compounds whose linear models have significant interaction terms. Candidates for the linear models will be identified from pathways prioritized by an upstream pipeline that uses several statistical learning techniques (k-means, random forest and principal components analysis (PCA)) to identify genes and metabolites strongly associated with the heat stress response. Incorporating ratios of metabolites into linear models will make it possible to detect potential shifts in specific biochemical reactions.

4.1.1 Established Context for Ratios as Extension of Previous Studies

The concept of focusing on the association of ratios of compounds as potential responders to an experimental treatment originates in earlier works integrating metabolomics and genomic data, such as the KORA (Cooperative Research in the

Area of Augsburg) study (Gieger *et al*., 2008). In the KORA study, which was one of the first to integrate high throughput metabolomics with other types of data, the authors focused on relating ratios of compounds to single nucleotide polymorphisms (SNPs) using standard additive models. The ability to combine genomic and metabolomic data represented an innovative paradigm and has improved the biological understanding of SNPs involved in diseases such as diabetes, arthritis and mental illness (Gieger *et al*., 2008). The heuristic of relying on ratios of compounds was motivated by the observation that doing so significantly decreased variance and improved the predictive power of subsequent modeling (Gieger *et al*., 2008). Such an approach also improves the biological interpretation of the quantity that is being measured through the metabolite data.

For example, in the case that the pair comprising the ratio is a substrate-product pair, then one is effectively modeling conversion efficiency between a substrate and product as a function of SNP status. This interpretation is important because it provides a perspective from which SNP data is being related to a phenotype (conversion efficiency of an enzyme) (Gieger *et al*, 2008). While this approach is innovative and useful for leveraging genotype data to explain variation in the chemistry of enzymes, it does not address the needs of identifying concerted metabolic and transcriptome shifts during an experimental treatment (heat stress, in our case). Pursuing this strategy, we developed techniques that would detect deliberate shifts in metabolic regulation in order to relate them to changes in transcriptome data.

Like the early genome-metabolome studies, our approach contextualizes these changes in a mechanistic perspective. We accomplish by relying on small regulatory triplets that relate metabolite levels, instead of a SNP, to the ratio of two other

metabolites. A regulatory triplet is thus a set of three metabolites that may interact with one another through a precursor-product, or similar relationships, such as coupling with the same pathway. Predicted metabolic relationships can be reinforced subsequently by identifying gene expression changes for the enzymes that process each metabolite. Processing metabolic data to produce network skeletons produces the framework of a network in which transcriptome changes can be interpreted. Unlike the genome-metabolome studies, we are modeling continuous metabolomics data. Thus, how these triplets behave under heat stress, as compared to control, can be biologically informative in terms of dynamic regulation of pathways. This is particularly true when these groups showing differential behavior span branches of regulation, as in the special case of regulatory triplets that represent units that we refer to as metabolic forks.

The metabolic fork is defined as a situation in which biological regulation differentially shunts substrates down one path or another. The relative preference of one route is influenced by gene regulation, which must change during heat stress. The ability of our pipelines to detect these situations through metabolome data will be compared with changes in the transcriptome identified by differential expression analysis. This comprehensive approach can identify the genes that may control metabolite relationships, and propose mechanisms for the regulation of precursor flow through a metabolite pathway. Changes to the behavior of metabolic forks could have far reaching effects on overall metabolism. Thus, the ability to detect them is a powerful method to generate hypotheses about novel biology.

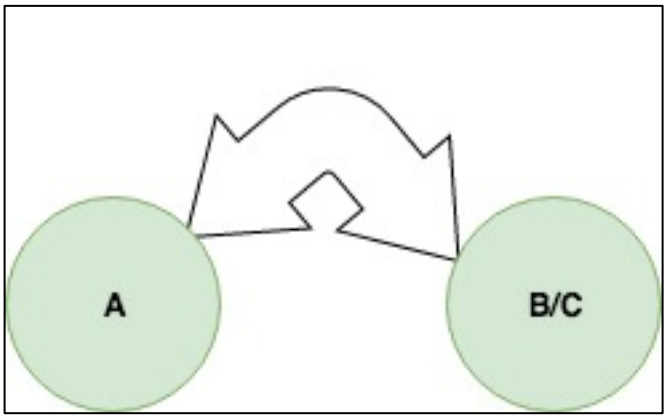

Figure 36: An interpretation of the relationship between a compound, A, and the ratio of two others $\left(\frac{B}{C}\right)$in the case that all three are metabolites. A may be influenced by the ratio of $\left(\frac{B}{C}\right)$ when $\left(\frac{B}{C}\right)$represent fates of precursors for A. Alternatively, A may influence $\left(\frac{B}{C}\right)$ when the two compounds are substrate/product pairs or gene/protein pairs.

In order to detect potential metabolic forks, we calculate the value of correlations for triplets of metabolites of the form A, $\left(\frac{B}{C}\right)$ where A, B and C are compounds from either the metabolome. Triplets that show a change in correlation greater than 1.2, $\left|\mathrm{cor}\left(A,\left(\frac{B}{C}\right)\right)\mathrm{heat} - \mathrm{cor}\left(A,\left(\frac{B}{C}\right)\right)\mathrm{control}\right| > 1.2$, are then used to build linear models and determine, through the p-value of the interaction term, if there is a statistically significant change in the relationship between A and the ratio $\left(\frac{B}{C}\right)$. Using a statistical learning pipeline, in addition to prior knowledge, we will prioritize compounds into a searchable set (about sixty compounds) from which potential metabolic forks can be identified. From these metabolites, we will be able to extract novel relationships through linear models with significant interaction terms.

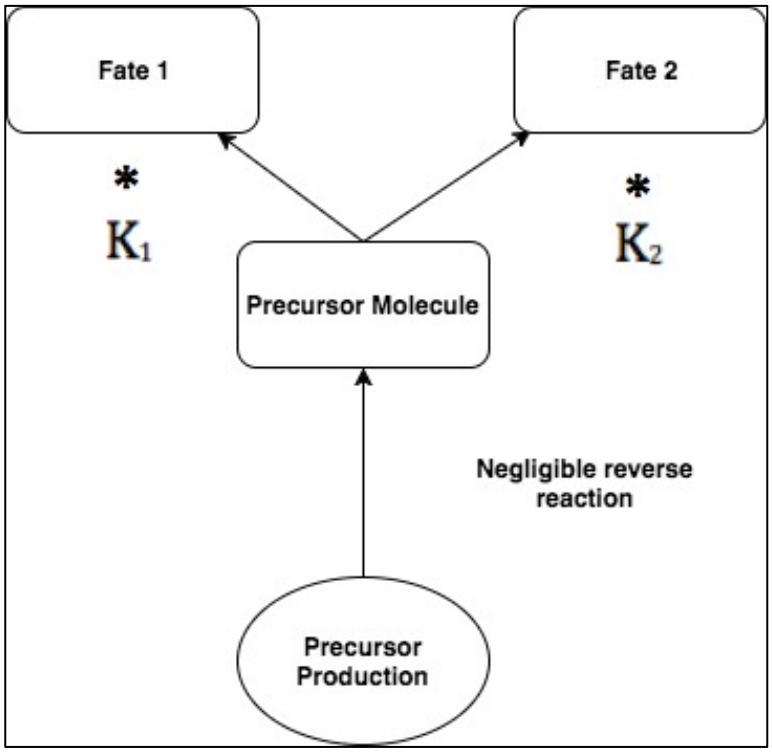

Figure 37: Special case of a metabolic fork, in which compounds are directed to one of two divergent metabolic fates (with reaction back to precursor from either state negligible).

The function $\mathrm{cor}\left(\mathrm{A}, \left(\frac{\mathrm{B}}{\mathrm{C}}\right)\right)$ may detect changes in a potential biochemical relationship that is regulated by a metabolic fork (Figure 37). For example, if $\left(\frac{\mathrm{B}}{\mathrm{C}}\right)$ represents a rate of conversion or levels at steady state, and A is associated with increases or decreases in this ratio, we will consider if it may be driving the production. When B and C represent divergent pathways towards separate compounds, while sharing a common precursor, their ratio may represent the relative probability of a precursor molecule taking one fate over another. The ability to detect these types of changes could be very useful in understanding the physiological consequences of metabolite shifts across multiple pathways. We have hypothesized

that such changes control the heat stress response, and could explain how the metabolic role of nutrients shifts to cause decreases in yield.

It is well known that broiler chickens accumulate muscle mass more slowly under heat stress, thus contributing to lower feed efficiency (Lara and Rostagno, 2013). Enhanced protein catabolism is thought to be a major factor influencing poorer feed efficiency of heat stressed birds (Lara and Rostagno, 2013). However, the metabolic fate of catabolized proteins is uncertain, although it is hypothesized that the liberated amino acids are harnessed to produce energy for the heat stressed bird. Identifying metabolic forks that relate amino acid and sugar production could provide critical evidence clarifying this and other hypotheses.

4.1.2 Interpretation of Metabolite Ratios from a Biochemical Standpoint

Important reactions that produce resources to manage stress may be controlled through changes in gene expression, in order to meet shifting environmental challenges. By influencing levels of the appropriate enzymes, such changes may shift the equilibrium of reactions to favor energy production. A crucial strength in investigating changes in ratios of compounds is that it provides a perspective from which to understand how changes in gene expression may influence equilibrium of metabolite production. This ability to pinpoint specific metabolic changes extends and complements the approach of using statistical learning techniques to identify potential biomarkers for heat stress. We can consider a way in which the ratios model equilibrium changes below, and represent changes in favorability that could regulate the behavior of pathways.

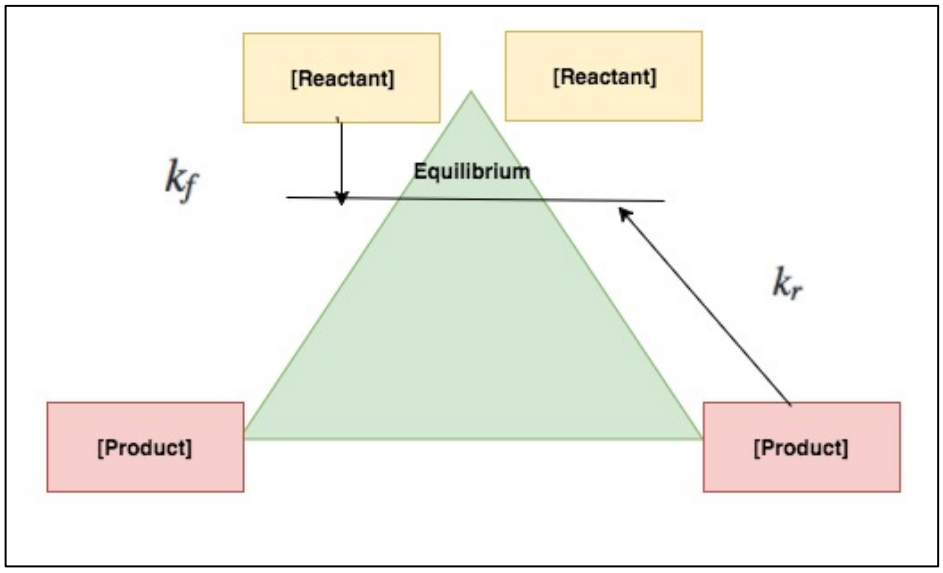

Figure 38: Illustration of an equilibrium point of a reaction, where net movement towards products is countered by backward movement toward reactants. In a biochemical reaction controlled by an enzyme, this equilibrium point may be influenced by gene expression.

At the equilibrium point, the rate of the forward and the reverse reaction are equal. Viewed from another perspective, at equilibrium, the number of molecules switching states between products and reactant is constant. Thus, the probability of a given molecule belonging to either the product state or the reactant state is fixed. When the products are more favored by the reaction, there will be many more product molecules at equilibrium. Conversely, when the reactants are more favored by the reaction, at equilibrium there will be many more reactant molecules. Note that at this equilibrium point, the ratio of products and reactants should be constant (excluding statistical fluctuations) as the same proportion of molecules are in the substrate or product state. One way to control the location of this equilibrium point, in terms of the relative concentrations of reactants and product, would be through enzyme levels.

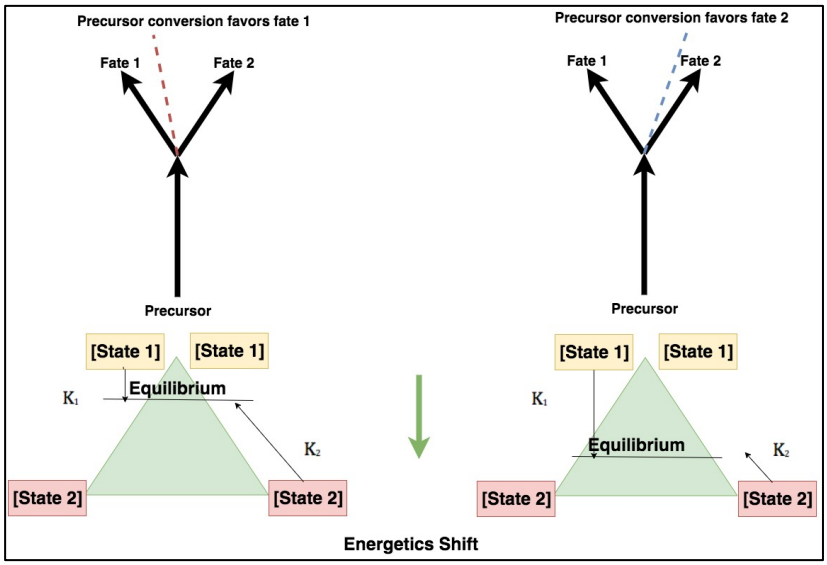

Figure 39: A change to a biochemical reaction in which the forward reaction has become more favorable after regulation of an enzyme, possibly through gene expression changes. The difference between the equilibrium points now results in one state being more energetically favorable than the other, given the current conditions. Depending on the favorability of the subsequent product, a precursor may be more or less likely to be converted into diverging metabolic fates.

The location of a biochemical equilibrium contains information about the relative energetics of one pathway compared to another. Figure 38 illustrates the consequences of a shift in this parameter. The ultimate objective of this statistical approach is to describe changes in equilibrium between reactions as functions of other compounds. Collective shifts across such systems may effectively describe alterations to homeostasis. Alternatively, if the back-reaction towards a precursor is negligible, the equilibrium of the metabolic fork represents relative propensities for one metabolic fate versus another (Figure 39).

4.1.3 Context for Integrating Tissue Enrichment, Statistical Learning Techniques and Linear Models

Identifying metabolic forks highlight differential metabolite regulation under treatment conditions in a way that is consistent with RNA-seq data, so long as changes in levels of relevant enzymes are directly proportionate to transcript levels. Identifying metabolic forks will be computationally intensive, however. Thus, it will be important to use statistical learning techniques to identify compounds that are associated with the heat stress response, from which potential mechanisms can be identified. The resolution of these statistical learning methods is enhanced by sub-setting transcriptome data by tissue-enriched genes. The motivations and insights made possible by this technique have been previously discussed in Chapter 3 of this dissertation. While valuable in itself for tissue enrichment studies, our five standard deviation threshold definition of tissue enrichment will also proves useful in reducing the input set of genes to downstream pipelines. The statistical learning techniques will depend partly on harnessing the signature of heterogeneity of the data at both the transcriptome and metabolome levels, which influences the classifying power of each compound to identify samples from heat stress treated birds.

4.1.4 Identifying Biomolecules Associated with Heat Stress in the Liver

Obtaining biological insight from large-scale transcriptome data is challenging due to biological and technical variance. Careful experimental design can limit unwanted noise. However, when properly harnessed, heterogeneity can be used to detect biological signals that elude traditional enrichment analysis. For example, biological variation relating to a treatment response depends on many variables that are not easily controlled such as allelic or physiological variants. This fact can be informative because many compounds involved in the same process will have similar

patterns of heterogeneity. This can be used to identify relationships between elements of the same pathway, even when their scales of expression and variance differ considerably, by relying on statistical learning strategies. This approach allows the combination of transcriptome and metabolome data to gain a more comprehensive biological understanding of a system. This is particularly helpful in identifying significant features from the large, complex datasets now common in multi-omics studies.

One organ capable of exerting strong influence on both bird growth and thermoregulation, and thus making it an excellent target through which to understand routing of resources, is the liver. This organ has recently proved as a subject for avian studies that leverage multiomics approaches including transcriptomics and metabolomics (Jastrebski et al., 2017). Such work has begun to shed light on differentially regulated genes and metabolites in the tissue under heat stress. However, a systems level understanding in which fluxes in metabolites are related to gene expression, are lacking. This is partly because computational approaches describing the totality of a biological response including gene expression and metabolite production is lacking. We will combine RNA-seq expression and metabolites from the liver to identify genes and compounds that function as biomolecules associated with heat stress. While metabolomics data identifies changes in biologically active compounds, RNA-seq data identifies genes regulate metabolic changes. We offer a geometric interpretation for our statistical procedures, describing how they recapitulate novel biology (Figure 40).

This analysis applies statistical learning approaches on metabolite and gene expression data restricts transcriptome analysis to a core module of liver enriched

genes. These are determined by a definition we propose that proves more stringent than other types of relative expression analysis, as discussed in Chapter 3 of this thesis. Sub-setting in this fashion isolates tissue-enriched genes that reflect unique biology specific to the liver in a tissue diverse dataset, and has been used previously to explore gene expression difference in muscle tissue. The approach of selecting tissue-enriched genes provides a framework to integrate metabolite and transcriptome data by providing an initial step of feature selection. This approach of combining data from different high-throughput technologies makes it possible to identify important features of the high dimensional dataset.

4.2 Methods: Combination of Statistical Learning Techniques

The heat stress response is multi-tiered and involves input from multiple tissues, though we choose to explore it from the perspective of the liver, a metabolic powerhouse. At the cellular level, the heat stress response unfolds across an intricate program of organelle specific changes. Which changes are causal, and which merely correlative with underlying signal or sensing pathways, thus becomes a complex question. However, the variability associated with most basal regulators of the heat stress response should be most closely related to the variation in the heat stress response. By the transitive nature of biological communication, the introduction of noise into the signal diminishes the capacity of downstream molecules, which correlate with, but do not cause the heat stress response, to discriminate between treatment and control samples. From this perspective, the problem of identifying causal molecules from expression profile is well posed as a statistical learning problem that can be addressed through random forests. Random forests can rank candidates on their ability to correctly identify the class of samples as assigned to

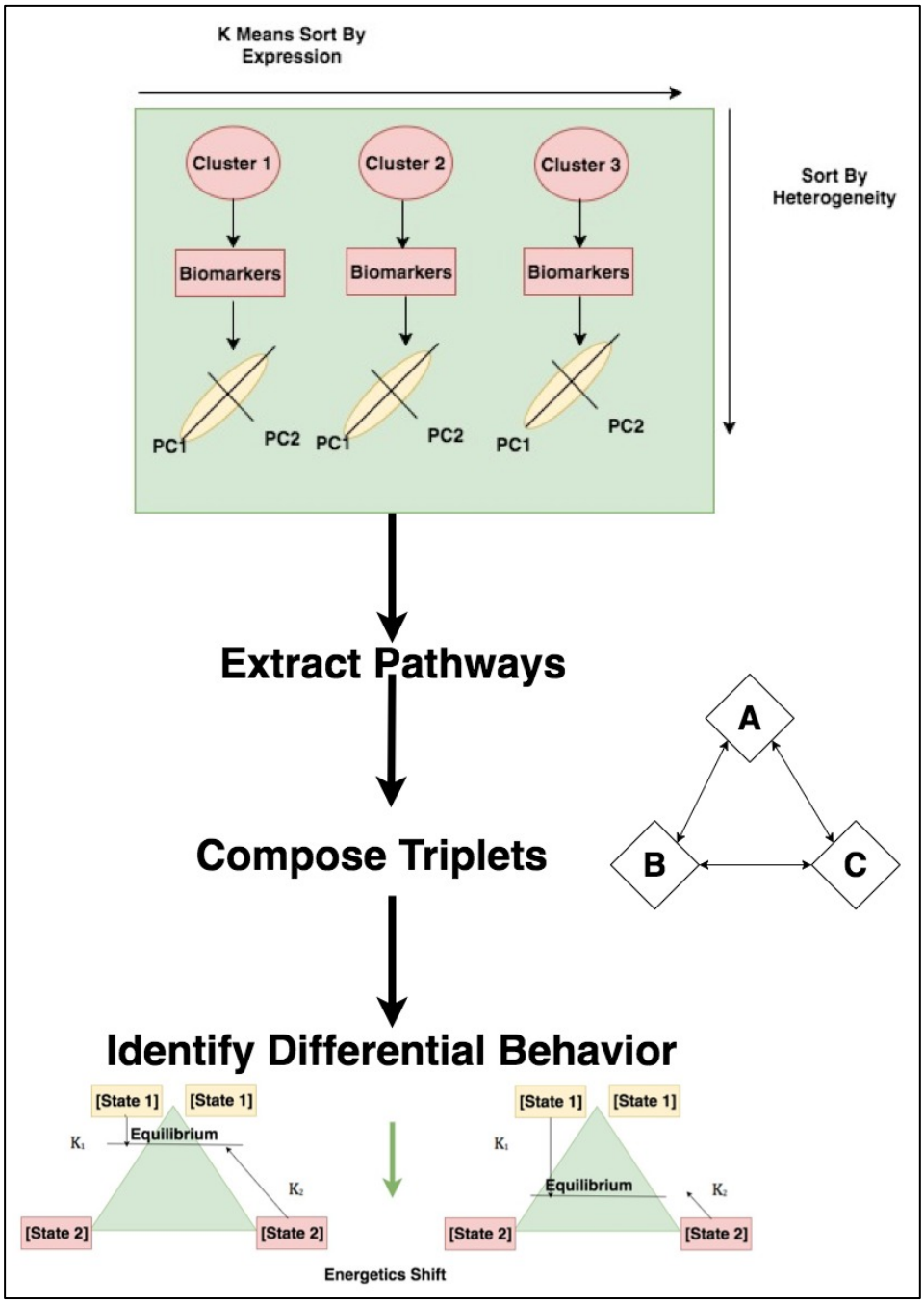

Figure 40: Total pipeline, from data analysis to identifying hypothetical mechanisms.

control or experimental treatment groups. Our approach follows sorting compounds into clusters based on their expression profile, using k-means clustering, prior to application of the random forest algorithm and finally prioritizing these top biomolecules with PCA. Rationale for k-means with k = 3 described in Figures 41A, 41B, and 42. Subsequently, we identify compounds most strongly associated with heat stress among liver enriched genes and metabolites. These compounds can then represent candidates for metabolic forks.

Biomolecules are identified and prioritized to extract pathways from whose elements triplets can be calculated (Figure 40). Triplets showing differential behavior are selected, which detect equilibrium shifts and thus indicate behavior of a metabolic fork.

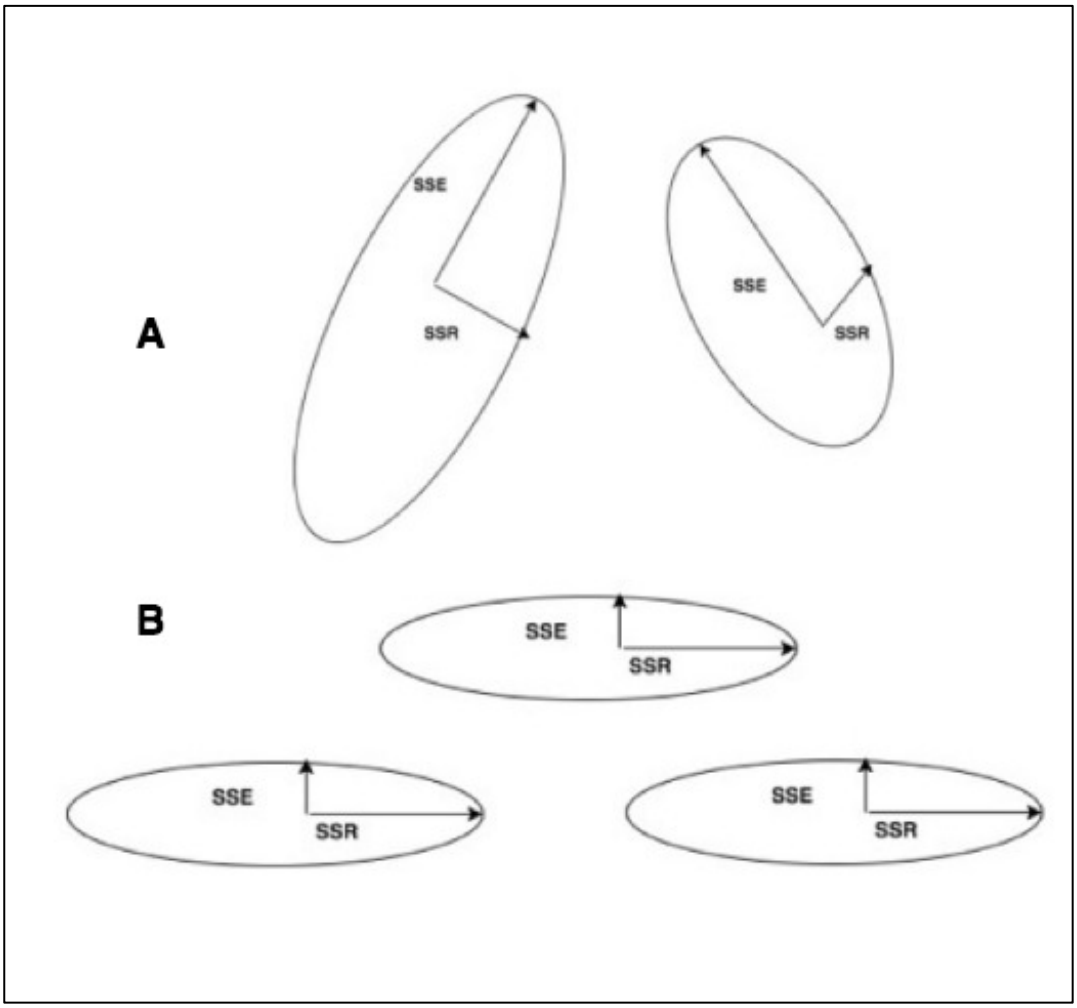

Figure 41 A and 41B: Example of possible models around specific cluster with different k-means selection, illustrating more uniform clustering results with k = 3 (41B) compared to k = 2 (41A).

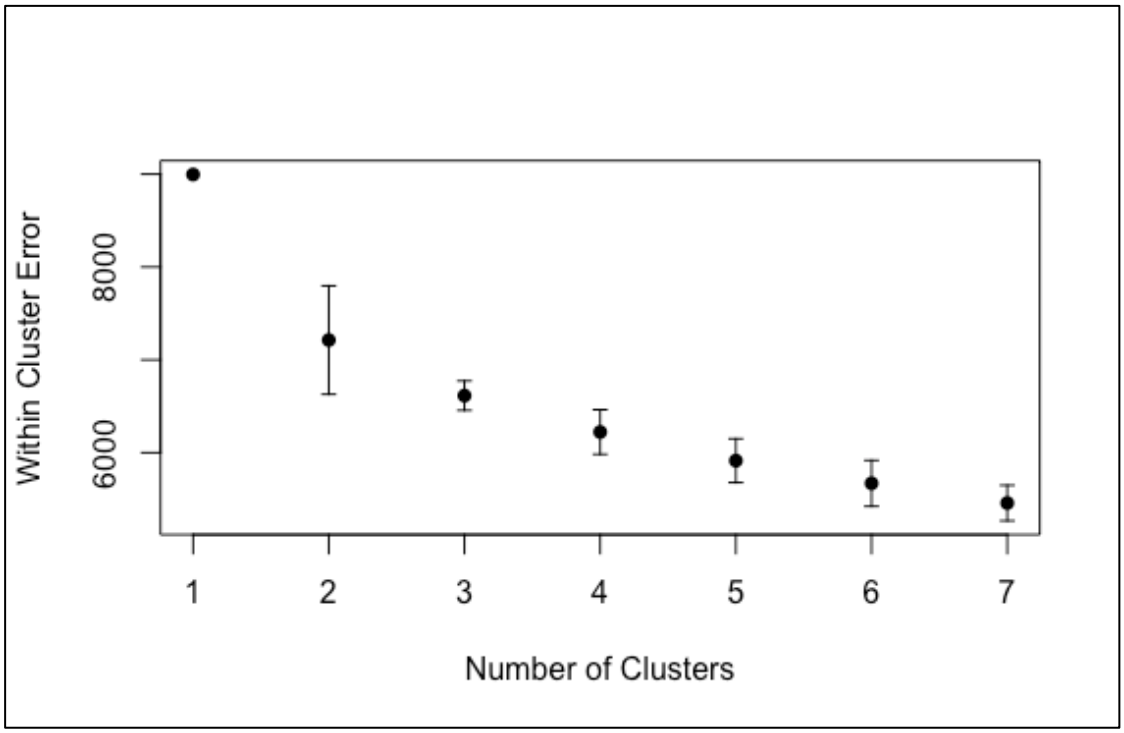

Figure 42: Elbow plot: with K-means = 2, the clusters are somewhat uneven compared to one another. With K = 3, however, we get relatively uniform clusters. The final choice of k = 3 is based on both biological interpretability and statistical properties of each clustering that considers bias-variance tradeoffs.

4.3 Results

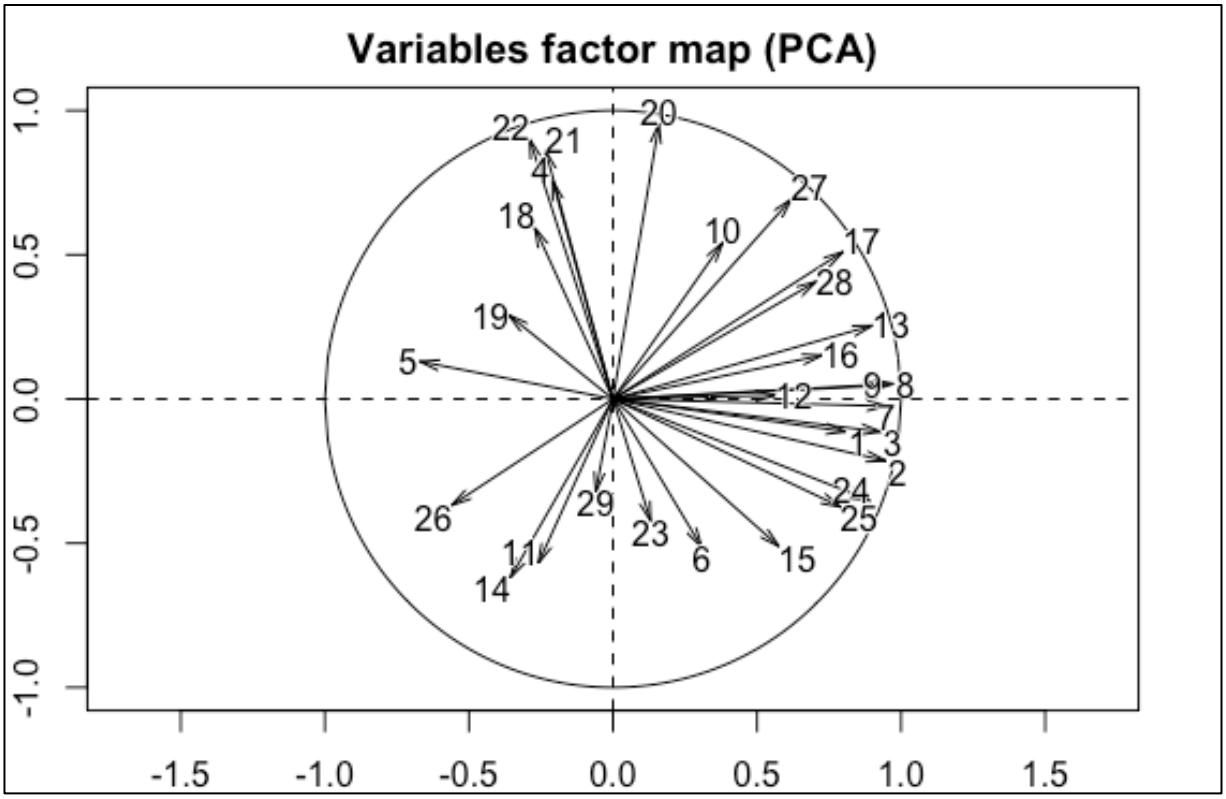

Figure 43: PCA of highly prioritized biomolecules from k-means cluster 1

Table 2: Figure 43 Keys

1. 1_2_dipalmitoyl_GPC__16_0_16_0	**16.** bilirubin__Z_Z
2. 1__6soUnique_enyl_stearoyl	**17.** docosahexaenoate__DHA;_22_6n3
3. 1_arachidonoyl_GPC__20_4n6	**18.** linoleate__18_2n6
4. 1_arachidonoyl_GPE__20_4n6	**19.** margarate__17_0
5. 1_palmitoyl_2_linoleoyl_glycerol__16_0_18_2	**20.** N_acetyltaurine
6. 1_palmitoyl_2_stearoyl_GPC__16_0_18_0	**21**. N_palmitoyltaurine
7. 1_stearoyl_2_arachidonoyl_GPC__18_0_20_4	**22.** N_stearoyltaurine
8. 1_stearoyl_2_arachidonoyl_GPE__18_0_20_4	**23.** Oleoylcarnitine
9. 1_stearoyl_2 arachidonoyl_GPI_18_0_20_4	**24.** sphingomyelin_d18_1_24_1_d18_2_24_0_
10. acetylcarnitine	**25.** sphingomyelin_d18_2_24_1_d18_1_24_2
11. Adipoylcarnitine	**26.** stearoyl_ethanolamide
12. arachidate__20_0	**27.** tartronate__hydroxymalonate
13. arachidonate__20_4n	**28.** taurine
14. beta_guanidinopropanoate	**29**. thiamin_diphosphate
15. betaine_aldehyde	

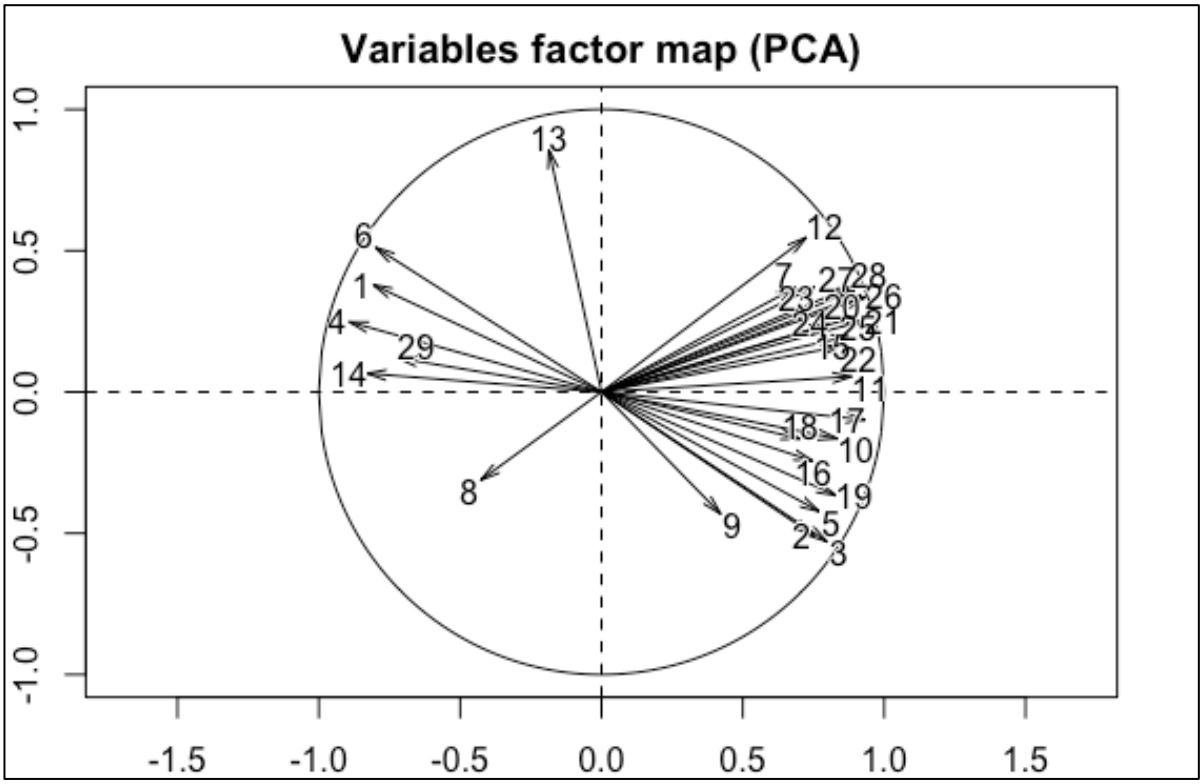

Figure 44: PCA of highly prioritized biomolecules from k-means cluster 2.

Table 3: Figure 44 Keys

1. Gene_C6	**16**. Biopterin
2. Gene_CTSO	**17.** cholesterol
3. Gene_FGG	**18.** Creatinine
4. Gene_HPD	**19.** Dehydroascorbate
5. Gene_ITIH3	**20.** hypotaurine
6. Gene_LIPC	**21.** linoleoylcarnitine*
7. Gene_LOC101748084	**22.** N_formylmethionine
8. Gene_LOC10174882	**23.** Picolinate
9. Gene_LOC417848	**24.** Propionylcarnitine
10. Gene_LOC424748	**25.** sphingomyelin_d18_1_20_0_d16_1_22_0
11. Gene_SLC6A13	**26.**sphingomyelin__d18_1_21_0__d17_1_22_0__d16_1_23_0
12. 1_stearoyl_GPG__18_0	**27.** sphingomyelin__d18_1_22_1__d18_2_22_0__d16_1_24_1
13. 2_hydroxyphenylacetate	**28.** stearoylcarnitine
14. Argininosuccinate	**29.** thiamin__Vitamin_B1
15. behenoyl_sphingomyelin__d18_1_22_0	

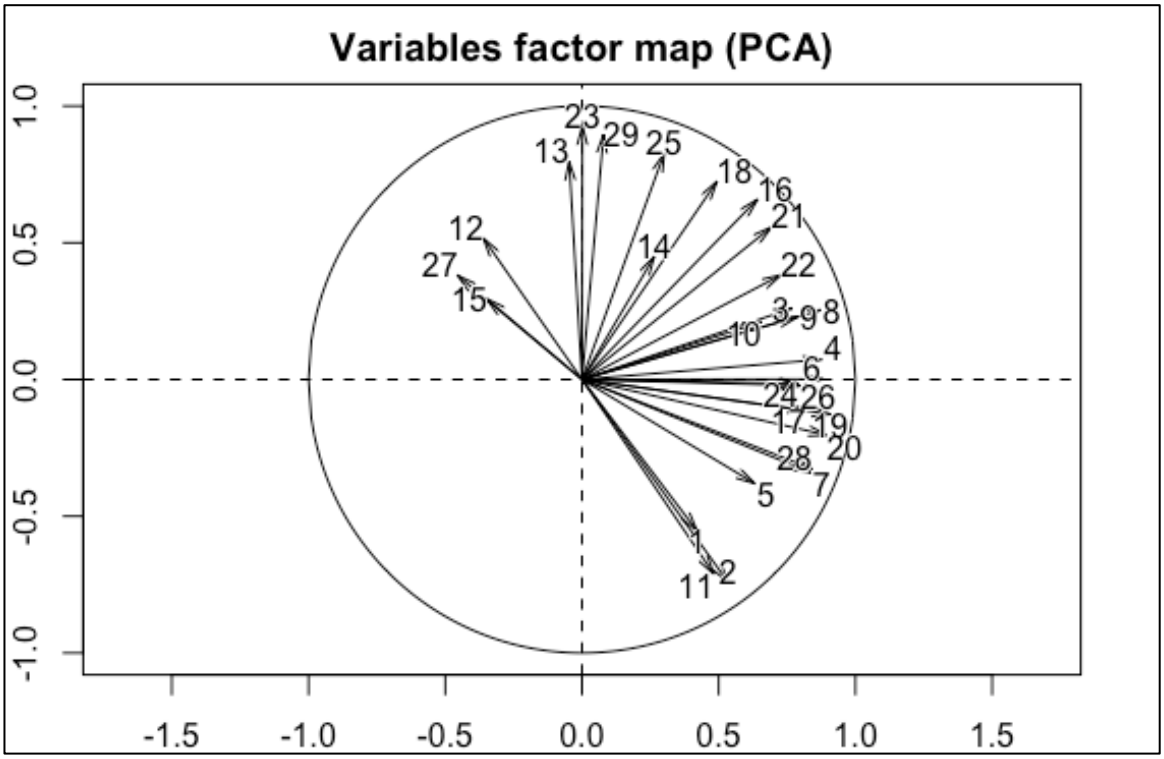

Figure 45: PCA of highly prioritized biomolecules from k-means cluster 3.

Table 4: Figure 45 Keys

1. Gene_NADKD1	**16.** Cysteinylglycine
2. Gene_S100Z	**17.** fructose_6_phosphate
3. 1_palmitoleoyl_3_oleoyl_glycerol__16_1_18_1	**18.** gamma_glutamylcysteine
4. 1_palmitoyl_2_linoleoyl_GPE__16_0_18_2	**19.** glucosamine_6_phosphate
5. 1_palmitoyl_2_linoleoyl_GPS__16_0_18_2	**20.** glucose_6_phosphate
6. 1_palmitoyl_2_oleoyl_GPE__16_0_18_1	**21.** glutathione__reduced__GSH
7. 1_palmitoyl_2_oleoyl_GPI__16_0_18_1	**22.** glycerol_3_phosphate
8. 1_palmitoyl_2_palmitoleoyl_GPC__16_0_16_1	**23.** Glycerophosphoethanolamine
9. 1_palmitoyl_GPE__16_0	**24.** myristoleate__14_1n5
10. 1_stearoyl_2_linoleoyl_GPE__18_0_18_2	**25.** N_acetylglucosaminylasparagine
11. 1_stearoyl_2_linoleoyl_GPI__18_0_18_2	**26.** N6_succinyladenosine
12. 3__dephosphocoenzyme_A	**27.** Phosphopantetheine
13. adenosine	**28.** Pterin
14. adenosine_5__monophosphate__AMP	**29.** UDP_glucuronate
15. coenzyme_A	

4.3.1 Geometric and Biological Consideration of each Statistical Learning Step

A goal of first leveraging k-means analysis is to build more biologically interpretable random forests, with compounds initially separated by expression profile. This reflects the idea that pathways involving essential biological compounds occur across a spectrum of expression profiles. Compounds with different patterns of expression may have distinct biological roles in pathways. Separating out compounds first by this feature prevents compounds from one expression tier crowding out those from another tier when they have similar capacities for classifying samples as control or heat stress. However, the optimal partitioning should produce clusters that are similar in explanatory power. Selecting k = 3 accomplishes this goal by distributing

compounds across clusters that are as similar to one another as possible in terms of their explanatory power (Figures 41A and 41B). This is corroborated by the elbow plot (Figure 42).

Random forest is used to prioritize the genes and metabolites from each k-means cluster that are the strongest classifiers for the heat stress response. Genes or metabolites prioritized by the random forests approach are thus potential biomolecules causally associated with the heat stress regulation. Finally, we organize these strong biomolecules into biologically relevant groupings with PCA. This final step is able to recapitulate elements of major biological pathways – for example grouping together many of the compounds associated with processes such as gluconeogenesis and antioxidant production. This can be seen in figures 43-45. Particularly, fructose-6-phosphate (F6P), glucose-6-phosphate (G6P) and glucosamine-6-phosphate are positioned close to one another on the biplot (numbers 17,19 and 20, respectively in Figure 45) and correlate well with the first principal component, supporting that this principal component represents glucose production. Importantly, this analysis suggests novel biology, by identifying strongly classifying biomolecules in a given k-means cluster that associate with one another through co-linearity. For example, among the PCA biplot of highly prioritized compounds from cluster 1, a number of the taurine related endocannabinoids N_acetyltaurine, N_palmitoyltaurine and N_stearoyltraurine group together (numbers 20,21 and 22, respectively). This is particularly informative, as the biology of these compounds is relatively uncharacterized, but our analysis suggests that they share a similar form of metabolic regulation that is implicated in heat stress. Thus, the additional level of resolution with PCA is useful because although we have already ensured that many of these

biomolecules will be strongly heat stress associated with the prior random forest approach, PCA is able to organize them into groupings that represent potential pathways.

4.4 Discussion:

Our complete pipeline (Figure 40), which combines statistical learning techniques with hypothesis-free modeling of metabolite ratios, is able to propose novel hypotheses while recapitulating significant known biology from the liver metabolome and transcriptome. Significant biology is detected through the statistical learning pipeline alone, prior to calculation of metabolic forks. Relying on classifying power of compounds and PCA effectively identifies changes in compounds with roles across organelles that are increasingly thought to have important functions in the heat stress response.

Much interesting biology, for example, relates to changes in the cell membrane. There are widespread shifts in levels of constituent lipids, for example. The exact mechanisms by which these shifts occur remain unclear, but accumulating evidence suggests these changes in the cell membrane exert important downstream effects on heat stress responsive genes and metabolites. At least some of these may be driven by dietary changes. One such example is the essential fat linoleic acid, which is a precursor to arachidonic acid and emerges as a strong heat stress associated biomolecule and whose detected levels are lower under heat stress. The compound also correlates with two principal components among the heat stress associated biomolecules among its cluster (Appendix, Tables A2 and A3). Downstream arachidonic acid derivatives are similarly decreased, many of which have roles in inflammatory response.

Other biomolecules prioritized through correlation with the same principal component include other lipids, related to signaling and fatty acid oxidation – such as adipoylcarnitine and the taurine related endocannabinoids N-oleoyl taurine and N-Stearoyl taurine (Appendix, Tables A2 and A3). These compounds represent a possible intersection between signaling lipids and sulfur metabolism via coupling with taurine. All of these compounds occur at lower concentrations under heat stress. While the mechanisms of such regulation remain unclear, there is much evidence that suggests lipid changes influence cell state and, potentially, bird metabolism. Lipid changes, in fact, are increasingly recognized as potential regulators of heat stress at a fundamental level (Balogh *et al.,* 2013).

Recent studies have focused on nuances of the heat stress response by revising the model that it is primarily triggered by the presence of unfolded proteins (Hoffman, 2007). For example, lipids in the cell membrane may detect membrane disorder and other physical consequences of heat stress and trigger signal cascades (Balogh *et al*., 2013). The evolutionary value of using a thermo-sensitive organelle such as the cell membrane to refine the heat stress response is the advantage of being able to regulate homeostasis through sensitive adjustments that have meaningful influences on cell fate (Balogh *et al.,* 2013). The inflammatory response may be a significant component in the transition from heat stress to heat stroke.

4.4.1 Heat Stress, Membranes and Lipids

The sophisticated signaling environment created by the cell membrane is comprised of a diverse set of lipids and proteins. Among these is an abundance of sphingolipids that form rafts in the membrane and possess important signaling roles (Simons and Ikonen, 1997). The organization of the cell membrane is intricate and

becomes dynamic under stress response. Important structural changes occur through interactions with membrane proteins, the gating of which possess thermal sensitivity (Torok *et al*., 2014). Additionally, heat causes changes in physical attributes such as diffusion and dimerization rates. Measurements suggest these characteristics change in a predictable fashion during even mild heat stress events (Torok *et al.,* 2014). Thus, the cell membrane is well equipped to sense relative temperature changes.

Not surprisingly, among the compounds prioritized by our pipeline are many lipids. These shifts suggest mixture of changes in compounds with signaling and structural roles. Alterations in lipid content are important in thermal shifts associated with both heat stress and extreme cold. For example, a key adaptation to cold is the increase in membrane fluidity mediated by elevating the fraction of cis-unsaturated fatty-acyl groups in membrane lipids (Vigh *et al.,* 1998). Alternatively, during episodes of heat stress mechanisms to endure temperature shifts focus generally on maintaining the integrity of the cellular processes and such pathways can be causally regulated by changes in cell membrane disorder (Vigh *et al.,* 1998). Regulation of heat shock factors can be influenced by addition of saturated and unsaturated fatty acids, with the former inducing expression and the latter suppressing it (Carratu *et al.,* 1996).

The possibility that the qualities of the cellular membrane make it an ideal substrate in which to store 'memory' or serve as a 'control center' for a physiological response in terms of the composition of density and sensors could be extremely interesting biologically. This could prove extremely important in terms of identifying mechanistic regulators of the general response. Indeed, changes in membrane fluidity induced via alcohols triggers systemic responses paralleling those caused by heat

stress, albeit in the absence of any thermal activation. Such changes include hyperpolarization of the mitochondrial membrane (Balogh *et al.,* 2005). Such experimental work confirms the role of lipids from a regulatory perspective and the influence of the heat stress response across organelles.

Among the cell membrane lipids influenced by heat stress, which are prioritized among their respective clusters, is a number of sphingomyelin species. These are substantially down regulated under heat stress, and emerge as strong classifiers in clusters one and two. This is a potentially significant observation in the context that sphingolipids are up-regulated in the early phases of acute heat stress in studies of yeast (Jenkins *et al.*, 1997). Many of these sphingomyelin species correlate with principal components among their clusters that include the downregulated inflammatory arachidonic acid derivatives (Appendix, Tables A7 and A8). Their general attenuation may be an important aspect of physiological adaptation to the long-term heat stress experienced by the birds, with the pattern of heterogeneity in their levels indicative of bird acclimatization.

4.4.2 Antioxidants and Energy Burden

Heat stress entails a number of challenges that endanger cell function and which must be addressed in order to preserve homeostasis. The management and deployment of protective systems can be quite independent from the initial sensory capacity of the cell membrane. These, for example, can respond to states of cellular stress that could be ongoing in a state of heat stress. Such pathways are essential to the heat stress response, as they relate to management of general consequences of oxidative damage. Several precursors of anti-oxidants, as well as such compounds themselves, are identified as strong classifiers of treatment assignment within each

cluster. These compounds manage the effects of toxic intermediates resulting from increased energy production, mitigating their ability to damage DNA or organelles. Their production may exploit the carbon backbones of amino acids released by catabolized protein.

Given the relationship between oxidation and energy production, some of the classifiers suggest changes in mitochondrial activity. Even slight changes in cell resting state can have dramatic changes on the production of reactive oxygen species and the behavior of the mitochondria (Akbarian, 2016). Molecules associated with mitochondrial performance are computationally recognized as potential biomarkers of the heat stress response. This suggests that mitochondrial conditions are closely related to general heat stress, and that the cell adjusts antioxidant levels accordingly.

At the same time that sugars and other energy-related metabolites show upregulation, an important class of lipids involved in the carnitine shuttle system that transports fatty acids to the mitochondria shows consistent downregulation. These carnitine species (linoleoylcarnitine, stearoylcarnitine, adipoylcarnitine) are identified as strong heat stress associated biomolecules among their clusters and correlate strongly with resulting principal components (Appendix, Tables A1, A2, and A3). Such patterns suggest sweeping downregulation of fatty acid oxidation pathways, as metabolism is increasingly driven by gluconeogenesis. Transcriptome changes support a coordinated shift in lipid and sugar management (Jastrebski *et al*., 2017).

Genes that correlate most highly with the principal components that emerge from the k-means cluster containing gluconeogenesis biomolecules include NAD kinase (NADKD1) and S100 Calcium Binding Protein Z (S100Z), (Appendix, Tables A6 and A7). However, the correlations of these transcripts with the first principal

component are relatively weak compared to the main metabolites associated with gluconeogenesis, i.e. .53 and .41 for S100Z and NADKD1 respectively. Glucosamine-6-phosphate and glucose-6-phosphate have correlations of .89 and .91, for comparison. NADKD1 is a Nicotinamide Adenine Dinucleotide (NAD) kinase responsible for Nicotinamide Adenine Dinucleotide Phosphate (NADP) production, while S100Z is a calcium binding protein. Calcium released during oxidative stress can trigger cell death (Ermak and Davies, 2002). Thus, upregulated S100Z may be important to mitigating apoptosis. The magnitude of correlations for these transcripts with the second principal component summarizing strong classifiers in cluster 3 cluster are relatively stronger, -.7417 and -.5521 for S100Z and NADKD1, respectively, albeit negative. This second principal component correlates strongly with antioxidant-associated compound such as glutathione that would correspond to the increased oxidative stress of gluconeogenesis. NADKD1 may play a role in lipid metabolism, by producing NADP that will be reduced to NADH by the pentose phosphate pathway and thus providing reducing power for lipid production (Pollak *et al*., 2007). Thus, NADKD1 production provides a potential link between gluconeogenesis and lipid production, at the same time lipid oxidation is decreased. The shift away from lipid oxidation is consistent with increases in coenzyme A.

4.4.3 The Metabolic Fork Consistent with Statistical Learning Pipeline

This shift towards gluconeogenesis is supported strongly from a mechanistic standpoint by the metabolic fork relating fat, lipid and sugar production (Figure 46). This triplet, in the context of gene expression data, provides stronger support for

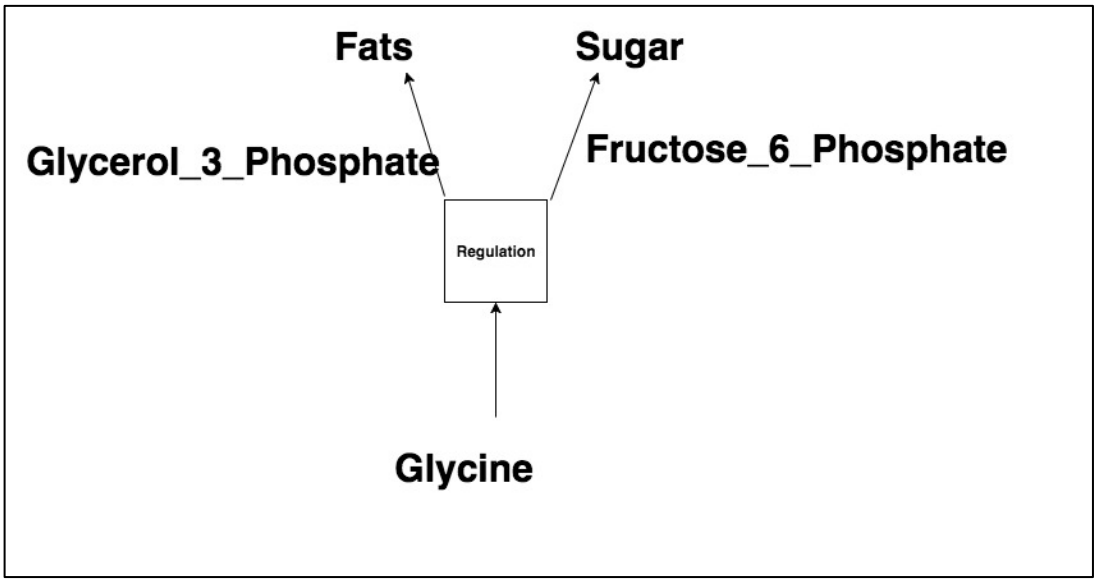

Figure 46: Intersecting pathways captured from a metabolic fork whose linear model shows differential behavior under heat stress.

a causal and directional relationship between these compounds. The three members of the triplet span gluconeogenesis (fructose-6-phosphate), glyceroneogenesis (glycerol-3-phosphate) and amino acid catabolism (glycine). Thus, this metabolic fork provides evidence of on route for the large-scale redirection of carbon resources released from the catabolized glycine. This complements the statistical learning pipeline, which prioritizes biomolecules without determining whether they are causal or merely collinear to biological changes. Pairwise correlations between each node are provided on the edge corresponding edge (Figure 47).

A proposed mechanism for the observed pattern is that catabolized glycine is preferentially shunted towards gluconeogenesis under heat stress, thus contributing to fructose-6-phosphate (F6P) production. This is captured in the linear model (Figure 58). Increasingly fueled by carbon backbones provided by amino acids from

catabolized proteins, gluconeogenesis decouples from glyceroneogenesis under heat stress (Fig 47).

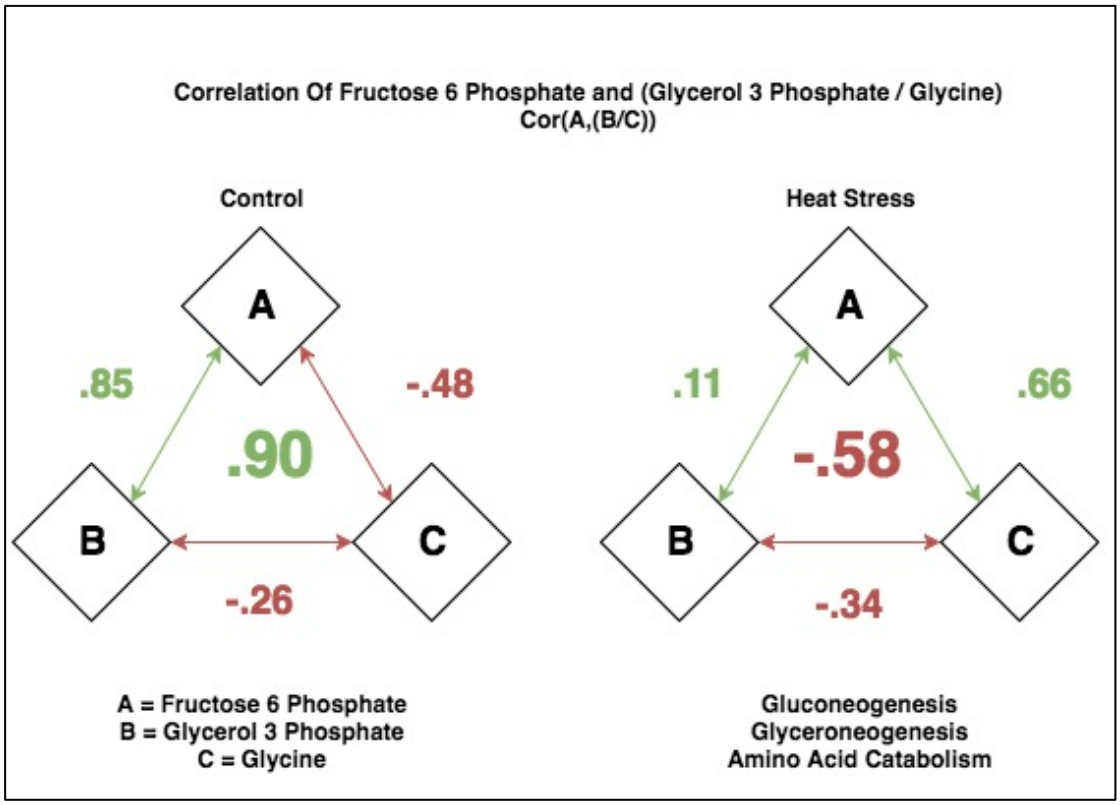

Figure 47: Pairwise correlations of the compounds in the metabolic fork, demonstrating the coupling of glycine with fructose-6-phosphate under heat stress.

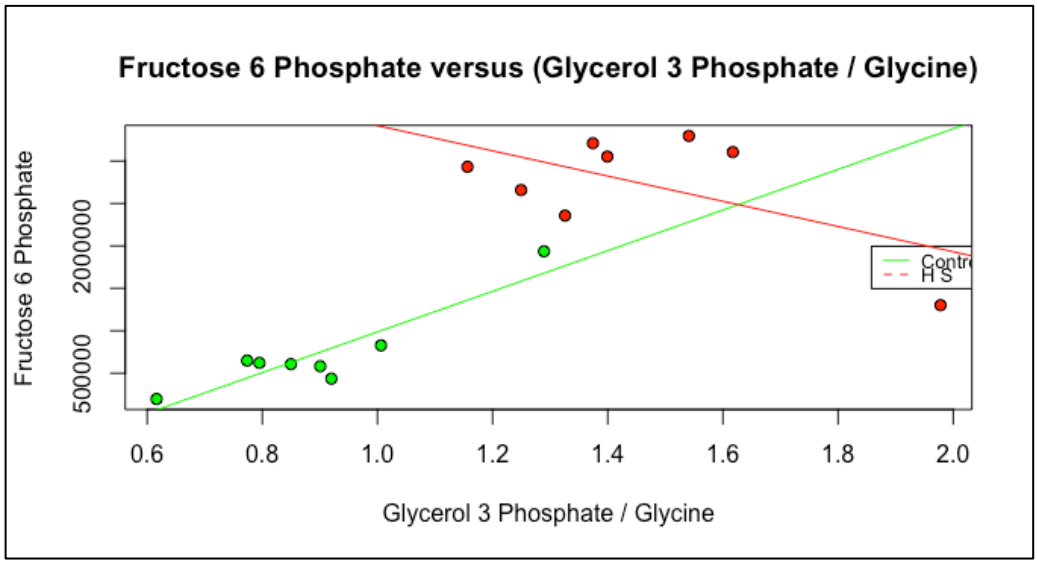

Figure 48: Linear model representing behavior of the triple of fructose-6-phosphate and G3P/Glycine.

The ratio of glycerol-3-phosphate (G3P) to glycine is interpreted as the tendency of catabolized amino acids to become backbones for fats, as opposed to sugars. This is hypothesized to change as a function of increased demand for sugar under heat stress, as given by increases in the sugars glucose-6-phosphate (G6P) and F6P. This is supported by increases in the gene Fructose-Bisphosphatase-2 (FBP2) encoding the rate-limiting gene for gluconeogenesis. Importantly, this computational prediction is independently supported with recent experimental work. According to this mechanism, it would be predicted that glycine supplementation should improve heat stress performance. Recent work has shown that glycine supplementation significantly improves the performance of heat stressed birds according to multiple metrics such as weight gain and feed intake (Awad *et al*, 2017).

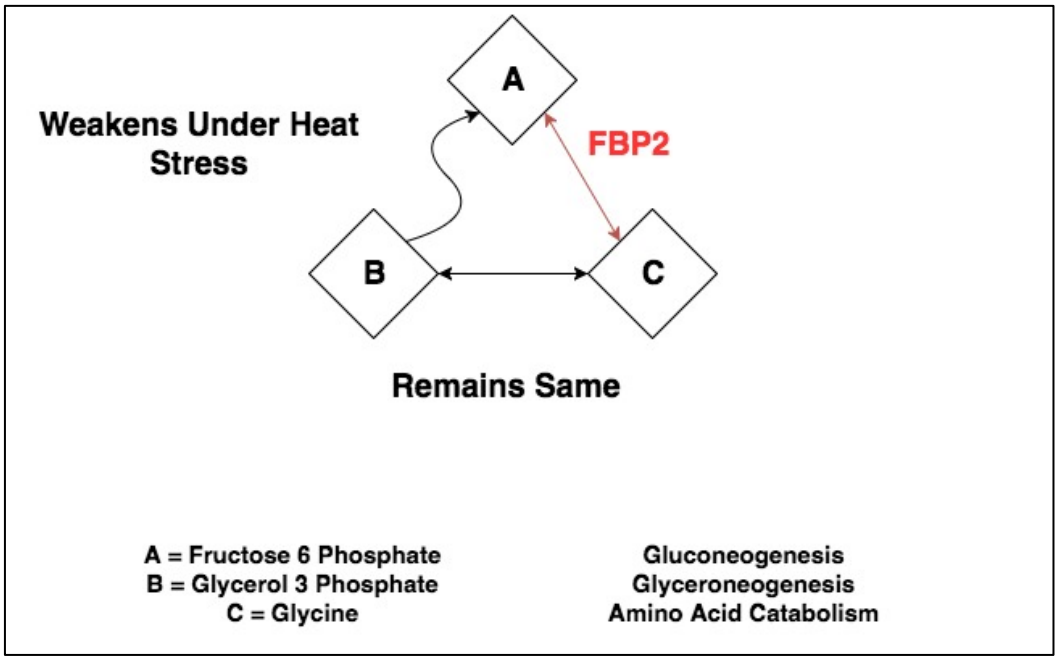

Figure 49: The metabolic fork in context of gene expression data. The coupling between glycine and fructose-6-phosphate is consistent with upregulation of FBP2. Transcriptome upregulation of the gene encoding FBP2 provides evidence for directionality towards F6P.

While this proposed mechanism demonstrates clear directionality, this insight would not be immediately clear from statistical associations alone. The transcriptome data, however, provides clarification of the regulation between the elements of the triplet, by identifying upregulation of the rate-limiting gene for glucoenoegenesis (FBP2). Subsequent work building larger models from metabolic forks will emphasize the utility of evaluating potential mechanisms in the context of transcriptome data.

Chapter 5

TOWARDS CIRCUITRY REGULATING CARBON FLOW UNDER HEAT STRESS

5.1 Introduction

The pipelines in Chapter 4 have produced methods to manage high dimensional datasets and prioritize genes by both classifying power (random forests) and correlation structure (PCA) in heat stress samples. In addition to producing insight on their own, these statistical learning techniques produce candidates for crucial network motifs (the metabolic forks) among prioritized features. The resulting analyses suggest many novel hypotheses about regulation of the heat stress response. Additionally, the signatures of well-known elements of the heat stress response are recapitulated through this analysis. The final chapters of this thesis seek to extend these findings by constructing complete metabolic circuits from many elements identified through the previous statistical analyses. We show that these analyses emphasize regulation that would not be otherwise detectable, and provide testable and novel hypotheses.

Metabolic shifts during heat shift involve a number of signaling cascades (Verghese *et al*., 2012). These include the unfolded protein response and both pro and anti-apoptotic pathways (Fulda *et al.,* 2010). Increasing evidence suggests that biologically active lipids play an important function during heat stress, as signaling agents and maintaining cell membrane integrity (Balogh *et al*., 2013). Establishing the relationship between lipid metabolism and other well-characterized heat responsive

pathways would provide better understanding of the flow of resources to different types of metabolites. This could provide a model describing how small carbon precursors are selectively routed to various fates necessary to sustain signaling, energy production and other processes that must undergo dynamic shifts under heat stress.

Viewing carbon flow as a circuit with dynamics that are affected by gene expression changes produces an effective description and identifies testable biological hypotheses. This perspective describes the mechanisms that manage and create resource pools in the form of biochemically valuable carbon backbones, including cysteine and other catabolized amino acids that are selectively incorporated into various biologically active of molecules. The model that we construct describes the interconnection between production of antioxidants, gluconeogenesis, and production of signaling and structural lipids.

Redirection of carbon backbones occurs at specific points of regulation where molecules are processed into one of two or more available metabolic fates. We have previously introduced this type of regulation in the form of metabolic forks, but have now extended them to build pathway level models. Importantly, this arrangement allows gene expression patterns to implement regulatory logic by selectively directing resources.

When joined, these metabolic forks describe complete circuits of carbon flow. Our computational predictions are consistent with previous literature that explores the putative functions of heat stress responsive molecules and metabolites through such experiments as knockouts. However, by focusing on small but powerful network motifs, we integrate these compounds into a full systems biology circuit model. This

type of comprehensive study has been largely inaccessible prior to the availability of large scale –omics data.

5.1.1 Iterative Linear Models

Extending the previous strategy of using linear models to analyze compounds prioritized by a statistical learning pipeline, we will use similar modeling techniques to evaluate metabolite relationships between compounds representing sulfur and lipid metabolism. These compounds are involved in major processes influenced by heat stress, such as increased antioxidant production and energetically useful lipids.

Detecting metabolic forks that contain compounds from both lipid and sulfur metabolism could improve understanding of how these pathways relate to one another under heat stress. Values for the correlation function of the form $\mathrm{cor}\left(\mathrm{A}, \left(\frac{\mathrm{B}}{\mathrm{C}}\right)\right)$ were calculated where A, B and C represents the levels of metabolites across lipid and sulfur metabolism. In this context, the most biologically informative triplets of the form A, B and C often represent sets of precursors and their resulting metabolic products, or products highly collinear with these compounds. Ratios of compounds are used, as this approach is more sensitive to detecting points of potential regulation for diverging metabolic routes. Triplets whose difference in value for the correlation function was 1.2 or greater between control and experimental conditions, i.e. $\left|\mathrm{cor}\left(\mathrm{A}, \left(\frac{\mathrm{B}}{\mathrm{C}}\right)\right)\mathrm{heat} - \mathrm{cor}\left(\mathrm{A}, \left(\frac{\mathrm{B}}{\mathrm{C}}\right)\right)\mathrm{control}\right| > 1.2$,were selected as representing possible metabolic forks. Linear models were then used to detect differential behavior under heat stress, to identify triplets with significant interaction terms. Per existing methods, all data was log transformed before modeling (Illig *et al*, 2010). The threshold of 1.2 for identifying potentially interesting triplets was a pre-screening method, with the linear models being used to determine a p-value for the difference

between control and experimental conditions. To be considered as a pathway element and incorporated into a circuit, the interaction term must be significant for both models of the form $A \sim \left(\frac{B}{C}\right)$ and $\left(\frac{B}{C}\right) \sim A$. This stringent heuristic is chosen because of ambiguity regarding directionality of the relationship between $\left(\frac{B}{C}\right)$ and A. Though this does enforce an assumption of linearity such an assumption is consistent with the use of correlation, which also measures linear relationships, to identify differential regulation of triplets. Such triplets were subsequently merged with one another to generate pathways.

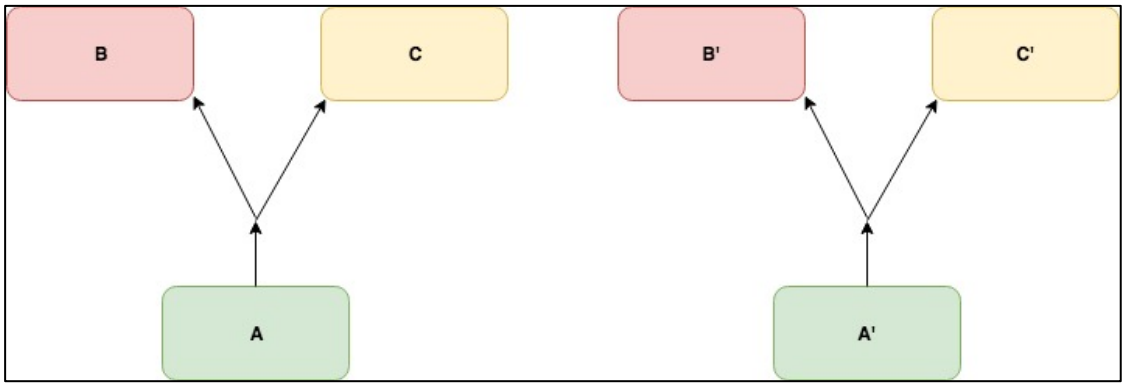

Figure 50: Two triplets, representing distinct potential metabolic forks. Triplets with overlapping elements may be merged, however, to create new biological hypotheses.

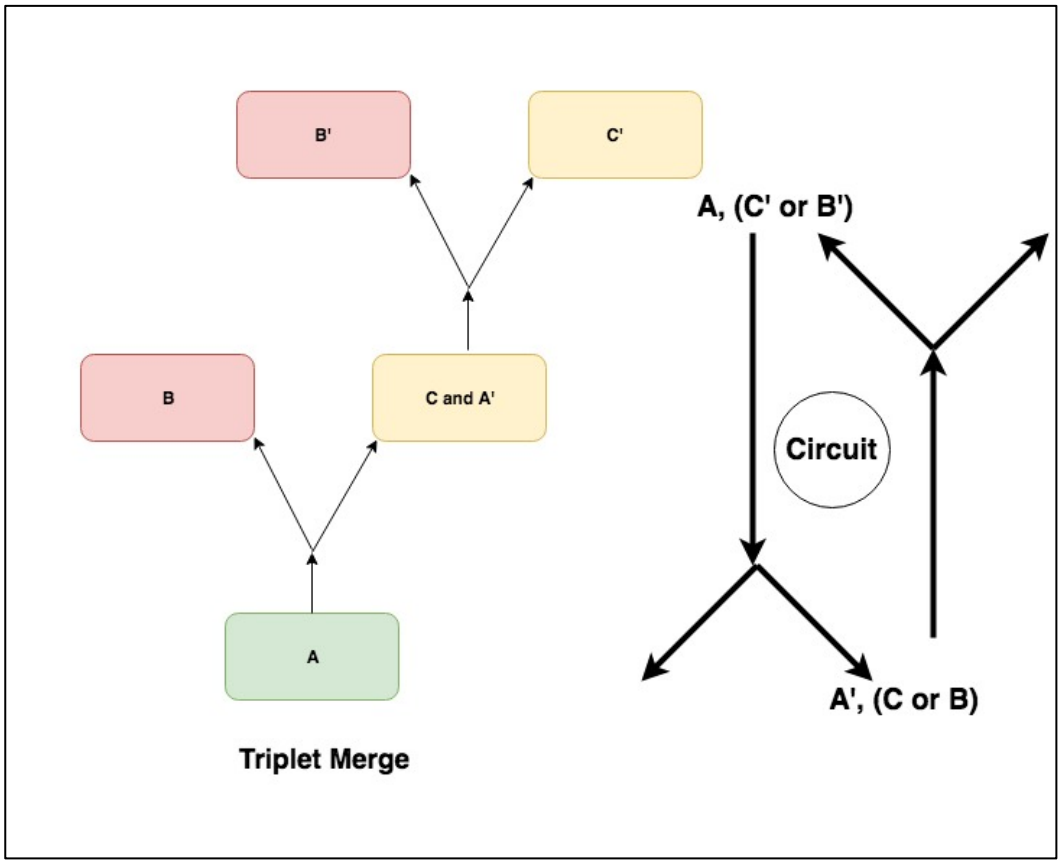

Figure 51: Example of how triplets that pass the differential correlation threshold (1.2) are merged into a circuit, by searching for overlapping components. This is accomplished with an R script that will combine the triplets.

Triplets are merged into a pathway by identifying metabolic forks that share components in common. We focus on a model involving three triplets involving sulfur and lipid regulation, because we hypothesize that the activity associated with these triplets best represents the functioning pathway relating antioxidant and structural and energetic lipids in a novel fashion, and is supported by transcriptome data. Each triplet requires manual inspection, but the network skeletons provide a sound basis for hypothesis generation. Regarding components of some metabolic forks, the

relationships derived from associations are not always causal, nor as direct as would be the case if they were always substrate-product pairs. However, many metabolic forks, in particularly those incorporated into a circuit are still highly biologically informative. Determining the precise nature of the interaction of compounds in a triplet requires manual inspection.

The transcriptome data can be helpful in this regard. For example, hypotheses about the proposed directionality of relationships can be strengthened by gene expression changes. Even when metabolites are not paired directly with precursors in a metabolic fork, one or more compound may be co-linear with a precursor. This precursor may share a carbon source with the compounds in the triplet, linking them biochemically. This type of relationship can be especially informative in a hypothetical circuit that contains precursors for multiple metabolites. Proposed circuits of metabolites exploit the ability of metabolic forks to detect shifts in pathway favorability. Merged triplets thus connect mechanism of regulation into pathways.

A shift in regulation changes the energetic favorability towards one route of the metabolic fork. Importantly, this interpretation with equilibrium shifts representing preferential routing towards separate metabolic fates can model regulatory circuitry. Additionally, we can build a linear model incorporating a gene and two metabolites to describe regulation of this circuit. These triplets that integrate metabolome and transcriptome data are of the form $\left(\mathrm{A}, \left(\frac{\text{gene B}}{\mathrm{C}}\right)\right)$, where A and C are metabolites, but do not have the same interpretation as a potential metabolic fork. However, it can show that the ratio of a gene to metabolite, $\frac{\text{gene B}}{\mathrm{C}}$, correlates differently under heat stress with the metabolite A. This can be informative when metabolites A and C represent distinct biological processes, and gene B links their

metabolism. In this model, the ratio $\frac{\text{gene B}}{\text{C}}$ represents the rate of transcription of gene B, relative to the levels of a metabolite (C) in a pathway coupled to a pathway that produces metabolite B. Candidate genes to be incorporated into triplets can be selected from those identified from a literature search to play a role in circuits created from the metabolome data.

5.2 Results

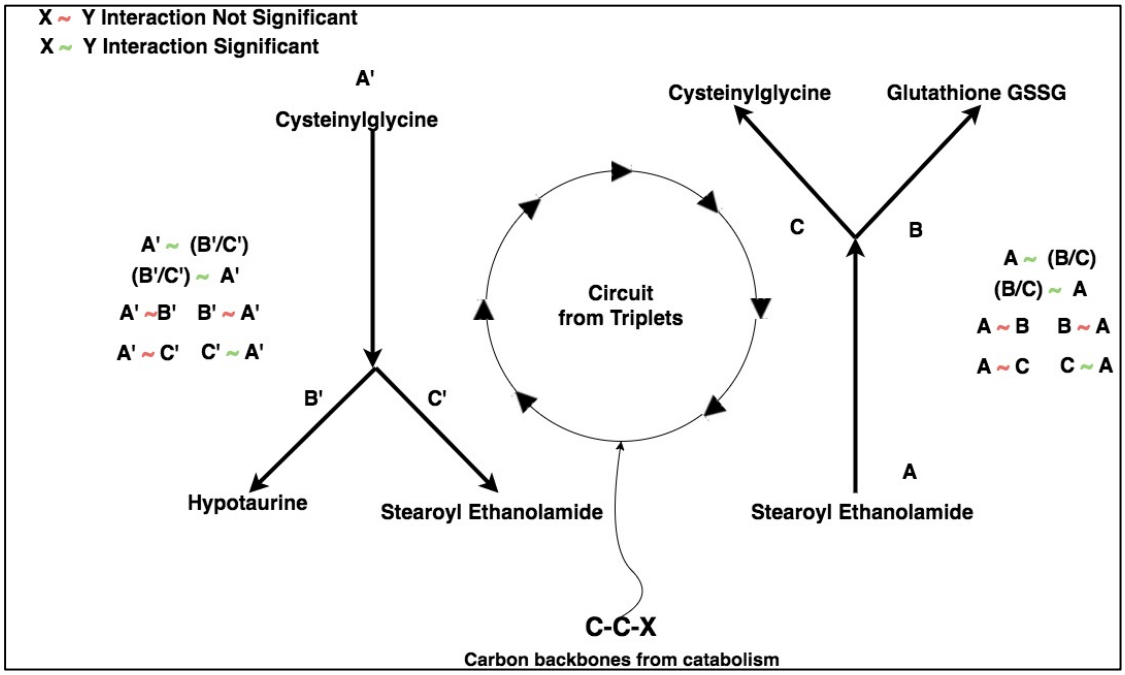

Figure 52: A hypothetical circuit of regulation managing carbon backbones from catabolism. This figure demonstrates that the interaction term of models involving ratios detects relationships that would be missed otherwise.

By merging together forks it is possible to identify small, functional units that we hypothesize represent critical elements of pathways. Merging of metabolic forks

was accomplished by identifying triplets that shared compounds in common (Figure 51). Integrating these isolated units to form controlled regulatory systems identify circuits of carbon and sulfur regulation. Importantly, linear models relying on the ratios of metabolites identify differential behavior not detectable using raw expression measurements alone (Figure 52). These models can then be joined to create larger circuits of regulation. A resulting circuit, relating lipid and sulfur metabolism is subsequently predicted from the data. Such predictions are consistent with previous research relating hypercysteinemia and hyperlipidemia to one another (Herman and Obeid, 2009). This connection has never been elaborated under heat stress, however. The full mechanism relating sulfur, lipid and antioxidant activities to one another is constructed by joining triplets into a network skeleton (Figure 53). In this model, cysteine levels are increased under heat stress (Jastrebski et al., 2017), driving sulfur metabolism coupling lipid and antioxidant production via changes in expression of key regulatory genes detailed below. Redirection of resources to cysteine metabolism is hypothesized to occur at the expense of choline derived signaling and structural lipids, due to changes in gene expression for enzymes related to this process. Differential behavior at each of these forks can be seen in a series of linear models, each of which demonstrate significant interaction terms (Figures 54 A-F and 54 A-F). These branch points are then placed in context of known biology to generate a regulatory model that incorporates both transcriptome and metabolome measurements.

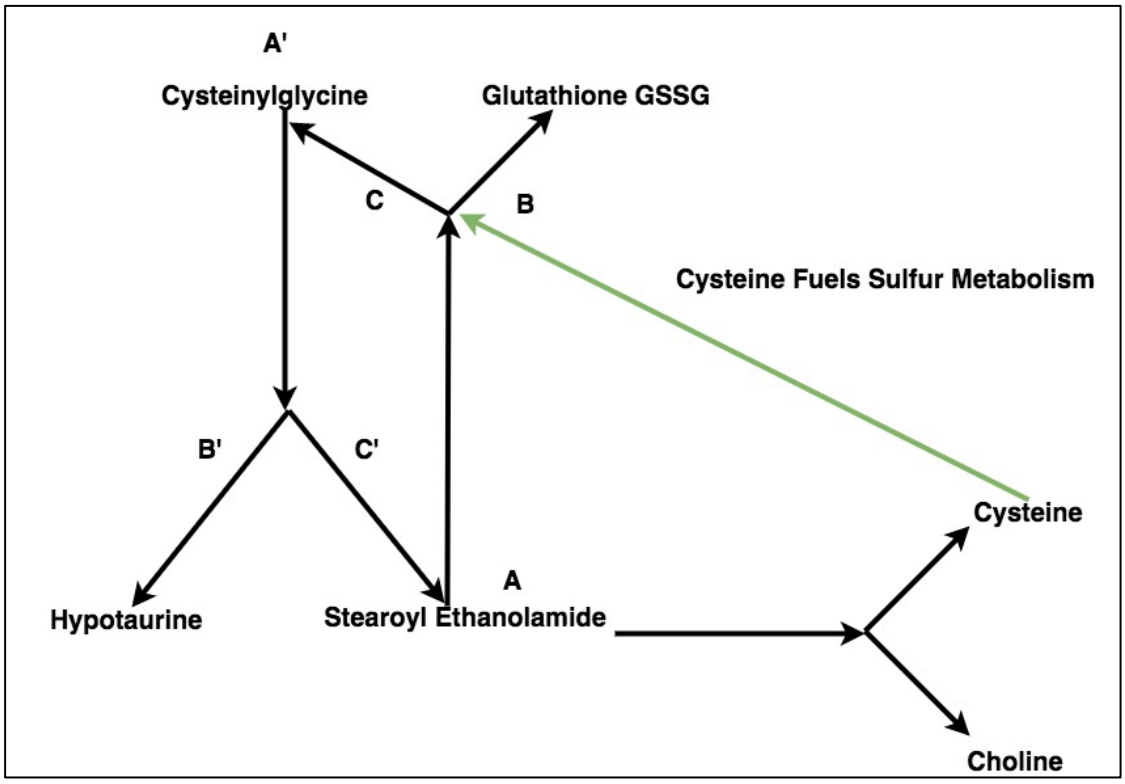

Figure 53: Network skeleton based on merging of triplets. This will provide the hypotheses driving a more complete circuit. Importantly, as lipid production shifts (the triplet with cysteine and choline), cysteine fuels a cycle of antioxidant metabolism represented by the two joined triplets, whose relationship is also summarized in Figure 52. This relationship is indicated by the green arrow.

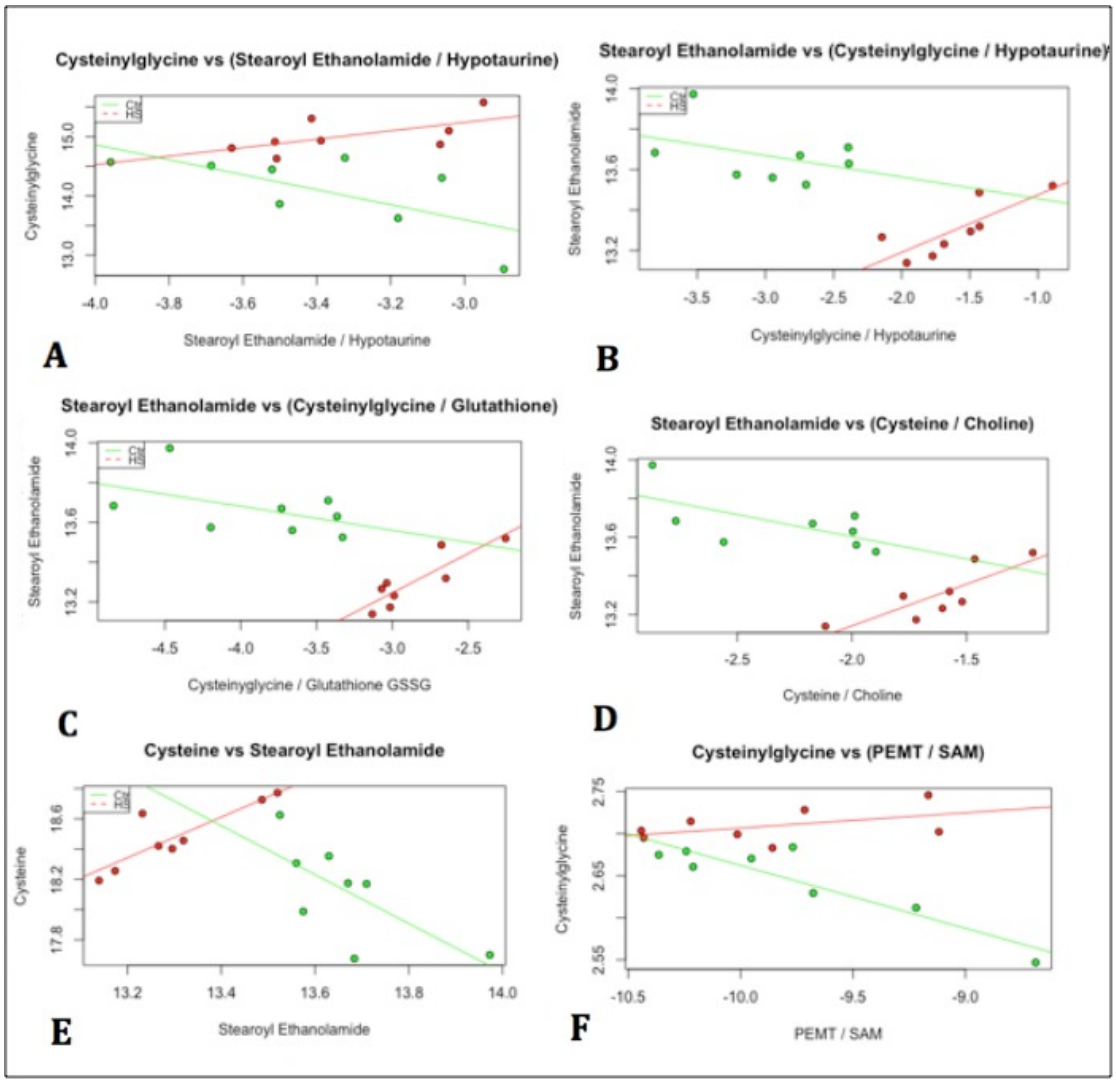

Figure 54: Part 1A-F: Metabolic Forks and Related Models – Levels of A metabolite as a function of ratio $\left(\frac{B}{C}\right)$. Linear models detect differential behavior of the metabolic forks that comprise the circuit. Also shown are the linear models for a triplet involving a gene (54 F) and the general coupling between cysteine and stearoyl ethanolamide (54 E). Figures (58-60) describe each branch-point in detail. All p-values for relevant interaction terms are less than .05.

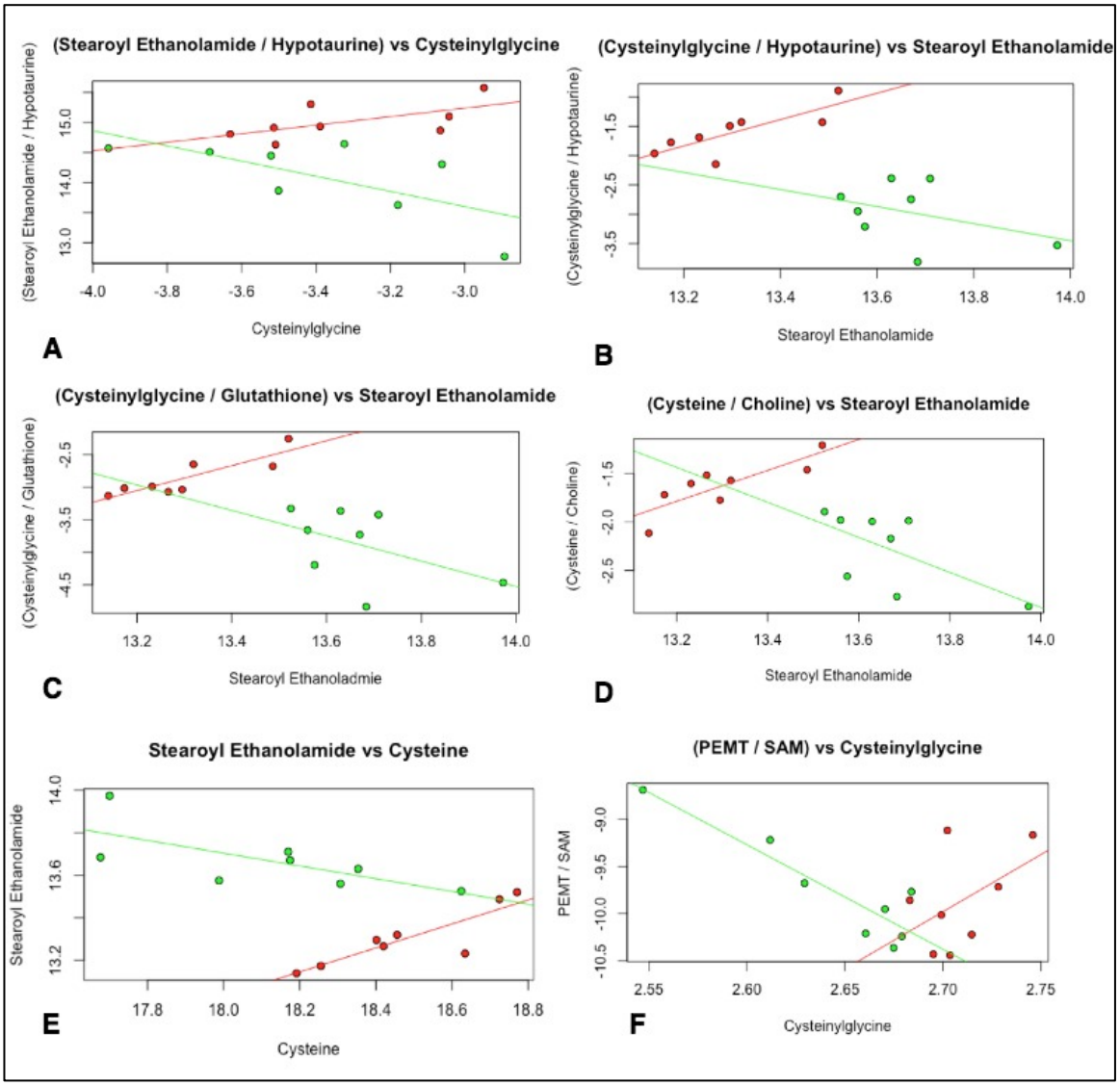

Figure 55: Part 1A-F Metabolic Forks and Related Models in circuit, the ratio $\left(\frac{B}{C}\right)$ as a function of the A metabolite are also shown. Linear models detect differential behavior of the metabolic forks that comprise the circuit. Also shown are the linear models for a triplet involving a gene (54 F) and the general coupling between cysteine and stearoyl ethanolamide (54 E). Figures (60-62) describe each branch-point in detail. All p-values for relevant interaction terms are less than .05.

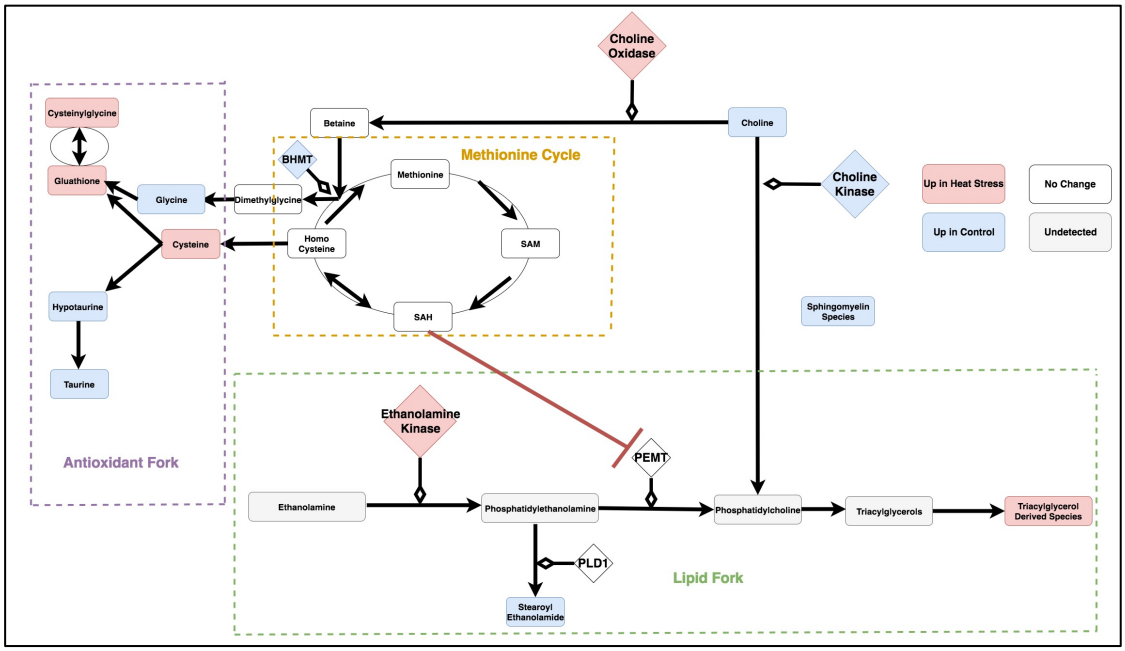

Figure 56: The circuit components as modules summarized by the three categories of antioxidant, lipid and methionine metabolism. SAM: S-Adenosyl-L-methionine, SAH: S-Adenosyl-L-homocysteine, Glutathione GSSG: Glutathione Disulfide, PEMT: Phosphatidylethanolamine N-methyltransferase, BHMT: Betaine--Homocysteine S-Methyltransferase, PLD1: Phospholipase D-1, PEMT: Phosphatidylethanolamine N-methyltransferase.

The proposed metabolic circuit from the network skeleton relates lipid, cysteine and glutathione production to one another (Figure 56). It is derived from the network skeleton proposed (Figure 53) as well as known interactions between lipid and sulfur metabolism, and a resulting literature search to identify genes that regulate these interactions. Though relationships between lipid and sulfur metabolism have been identified (Obeid and Hermann, 2009), the interactions between these pathways under heat stress remains an open area of research. The circuit we propose clarifies this relationship, in terms of many of the compounds associated with pathways

prioritized under the statistical learning pipeline in Chapter 5. This model contains sets of metabolic forks at key regulatory points, operating in concert. One such fork, which is critical to antioxidant production, shifts cysteine metabolism towards glutathione at the expense of taurine (Figure 56). Cysteine, which fuels many sulfur processing pathways, is the only amino acid increased under chronic heat stress. Our work describes mechanisms by which pools of cysteine regulate the long-term heat stress response. This model contextualizes mechanisms predicted by metabolite data with transcriptome data and known biology.

Changes in expression among genes regulating the methionine cycle are critical to the activity of this pathway. Additionally, S-Adenosyl-L-homocysteine (SAH) and Phosphatidylethanolamine N-methyltransferase (PEMT) interactions likely influence ethanolamide metabolism. A similar relationship between methionine metabolism and PEMT features in a putative relationship between hyperlipidemia and hyperhomocysteinemia (Obeid and Hermann, 2009). Changes to sulfur metabolism under heat stress influence lipids in a number of ways beyond the SAH interaction with PEMT.

In our model, choline, the precursor to many fatty acids, is directed away from the production of signaling and structural lipids. Several shifts in gene expression route this resource towards sulfur metabolism. Choline oxidase, the gene encoding the enzyme oxidizing choline to produce betaine, is also up-regulated. Betaine plays an important role in the methionine cycle. Concurrently, transcription of the first enzyme involved in converting choline to phosphatidylcholine, choline kinase, is down-regulated. Betaine levels, however, are unchanged, suggesting redirected choline rescues betaine levels. Further supporting a relationship between sulfur and lipid

metabolism via choline and betaine, Betaine--Homocysteine S-Methyltransferase (BHMT) transcription is down under heat stress. BHMT converts betaine and homocysteine to dimethylglycine and methionine, respectively and mouse knockouts of this gene show highly elevated levels of homocysteine (Strakova *et al*., 2012).

In our model, to prevent depletion of phosphatidylethanolamine-derived lipids such as phosphatidylcholine and stearoyl ethanolamide, phosphoethanolamine kinase is up-regulated. This model predicts that maintaining phosphatidylcholine production, despite dramatically directing resources to antioxidant production may be critical to homeostasis. This pathway can be understood by considering each fork in detail.

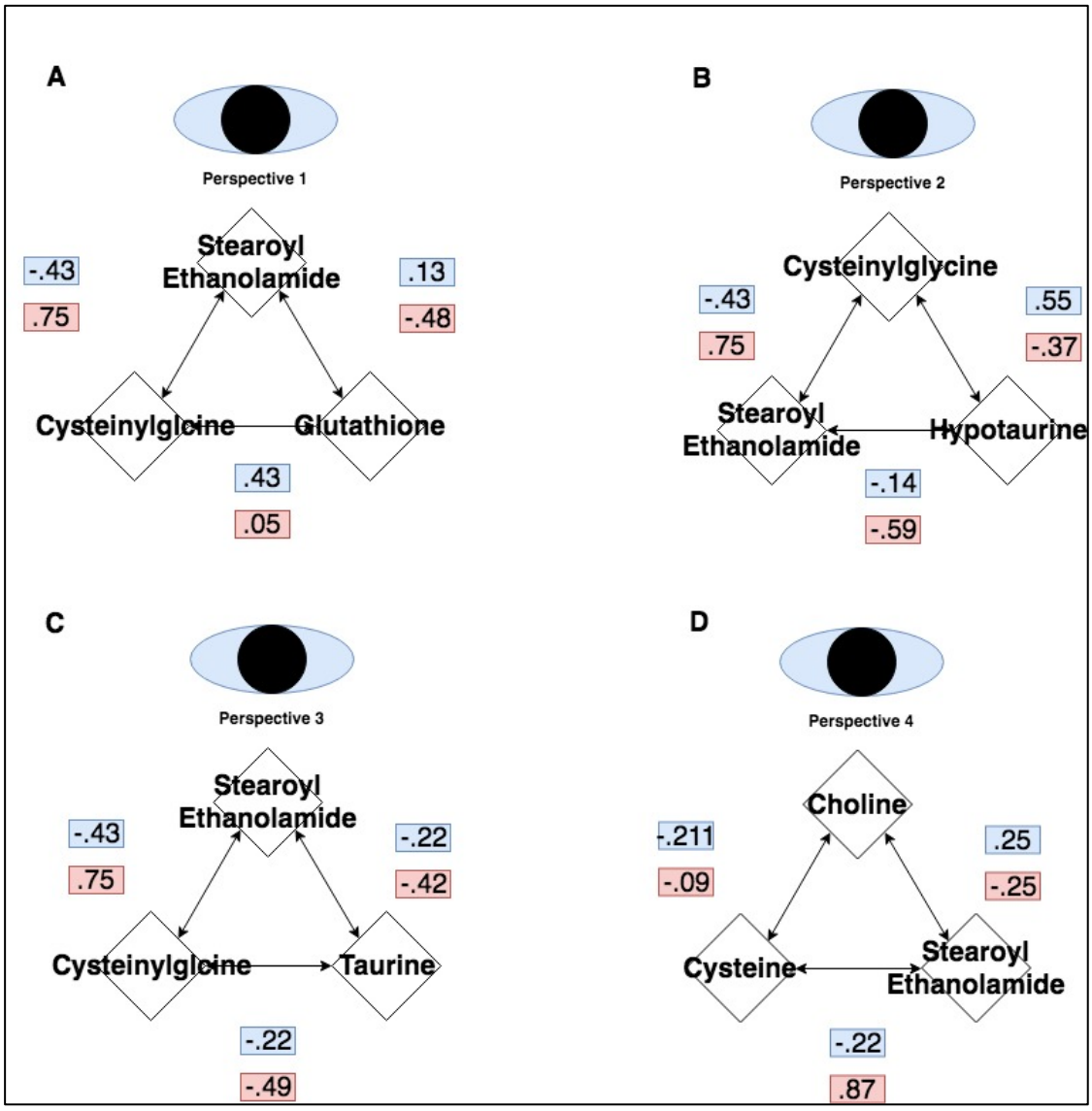

Figure 57: Pairwise correlations for the triplets of metabolites that comprise the linear models describing the circuit of regulation. Each one will be discussed in a regulatory context that notes relevant gene expression changes.

5.3 Discussion

5.3.1 Interpretation of Ratios

Several sets of models describing the proposed lipid, antioxidant, and sulfur circuit demonstrate significant interaction terms, even without using the ratios of metabolites, and relying instead on simple levels of metabolites. Ultimately, the ratio reflects the selective processing of sulfur derived from cysteine to antioxidants under heat stress. We have previously hypothesized that sulfur metabolism transitions de-emphasizes taurine synthesis to enhance antioxidant production during heat stress. This conclusion is bolstered by the pairwise correlations between the sulfur derived amino compounds cysteinylglycine and hypotaurine, representing antioxidant and taurine production, respectively (Figures 54 B and 55 B). Under control conditions, the two share a moderate correlation of .55, as opposed to the weak, but negative correlation of -.37 under heat stress conditions.

Importantly, stearoyl ethanolamide has a positive linear relationship with the ratio of cysteinylglycine and hypotaurine under heat stress conditions. Under control conditions, however, stearoyl ethanolamde has a negative linear relationship with this ratio (Figures 54 and 55 B). This is an important observation, because it suggests that the most biologically relevant feature being modeled through the ratio of cysteinyglycine/hypotaurine is the relative amount of cysteine being processed into antioxidants (glutathione) versus taurine (hypotaurine). Alternatively, the emphasis on antioxidant production as opposed to taurine synthesis may be accomplished through additional, yet to be established mechanisms.

There is no significant change in the expression of the genes regulating the conversion of cysteine to hypotaurine (cysteine sulfinic pathway) or expression of for

the rate-limiting enzyme in glutathione production, glutamate cysteine ligase (GCL). However, levels of important glutathione products (reduced glutathione and cysteinylglycine) are found at higher levels in heat stress than control. Thus, the ratio of cysteinyglycine/hypotaurine is informative, as transcriptome data does not emphasize relative importance of one fate over another for cysteine (glutathione synthesis or taurine production). Gene expression changes are consistent with pooling of homocysteine (decreased expression of BHMT), and the ratio of cysteinylglcine/hypotaurine captures the emphasis of antixodiant production from catabolized amino acids. The biochemistry of each of these mechanisms can be explored in detail (Figures 58 – Figures 61).

5.3.2 Regulation of Individual Forks

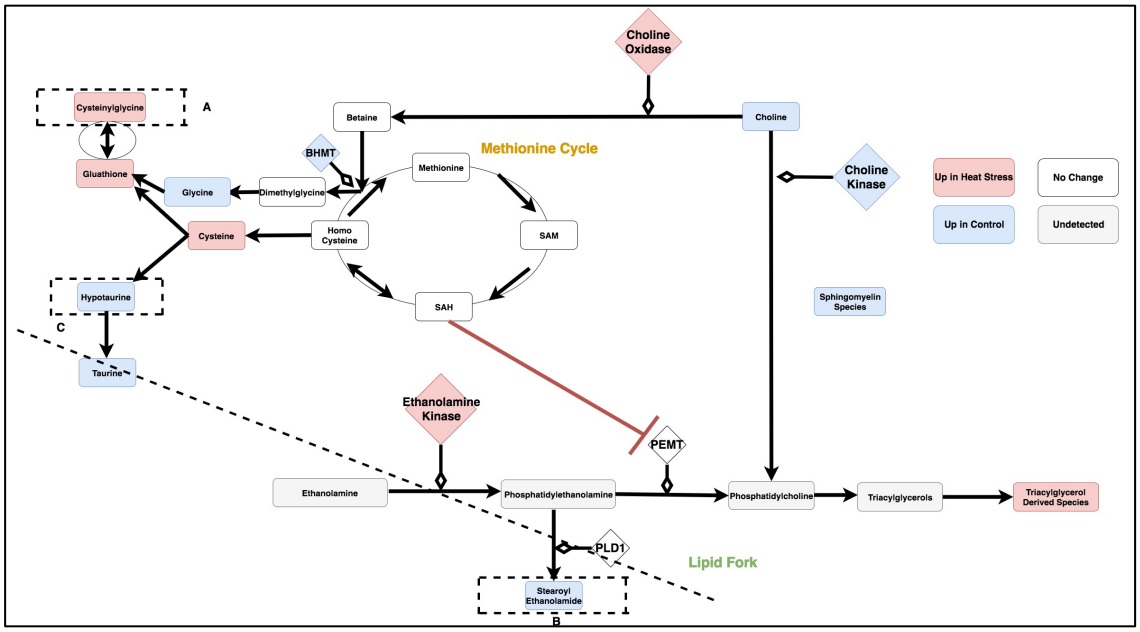

Figure 58: Triplet of cysteinylglycine and (stearoyl ethanolamide / hypotaurine). The compartmentalization of the pathway by regions containing the compounds in the ratio (stearoyl ethanolamide and hypotaurine) is illustrated by the dotted line. For the linear model representing differential behavior of this branch point, see figure 54A. SAM: S-Adenosyl-L-methionine, SAH: S-Adenosyl-L-homocysteine, Glutathione GSSG: Glutathione Disulfide, PEMT: Phosphatidylethanolamine N-methyltransferase, BHMT: Betaine--Homocysteine S-Methyltransferase, PLD1: Phospholipase D-1, PEMT: Phosphatidylethanolamine N-methyltransferase.

Under heat stress, sulfur metabolism favors glutathione at the expense of taurine (Figure 58). (Figuress 54A +55A). This increases the reservoir of anti-oxidants. The proposed circuit also implies changes in lipids, including coupling ethanolamine related compounds, such as phosphatifylethanolamine and stearoyl ethanolamide, to glutathione production through cysteine processing pathways. We

hypothesize this is accomplished through changes in the methionine cycle and choline metabolism. Under heat stress, decreased BHMT transcription preserves cysteine pools, managing the activity of the methionine to S-Adenosyl-L-methionine (SAM)/S-Adenosyl-L-homocysteine (SAH) cycle. A major product of this cycle, SAH, is a potent inhibitor of PEMT. Because higher levels of cysteine would ordinarily fuel methionine metabolism, the balance between cysteine allocation to antioxidants and the methionine cycle may be a critical for lipid production. This is because re-rerouting cysteine to antioxidant production, as opposed to the methionine cycle, would avoid the inhibitory influence of the methionine cycle on lipid production. Models in figures 54F and 55F show that the ratio of PEMT/SAM increases with the glutathione product cysteinylglycine under heat stress. Though not representing a metabolic fork, the ratio of PEMT/SAM is informative because it relates expression of a gene in lipid metabolism (PEMT) to a member of the methionine cycle. As antioxidant production increases, under heat stress conditions, expression of the PEMT lipid-associated gene associated also increases relative to a member of the methionine cycle (SAM). This coupling would further mitigate any interference of the methionine cycle on lipid production.

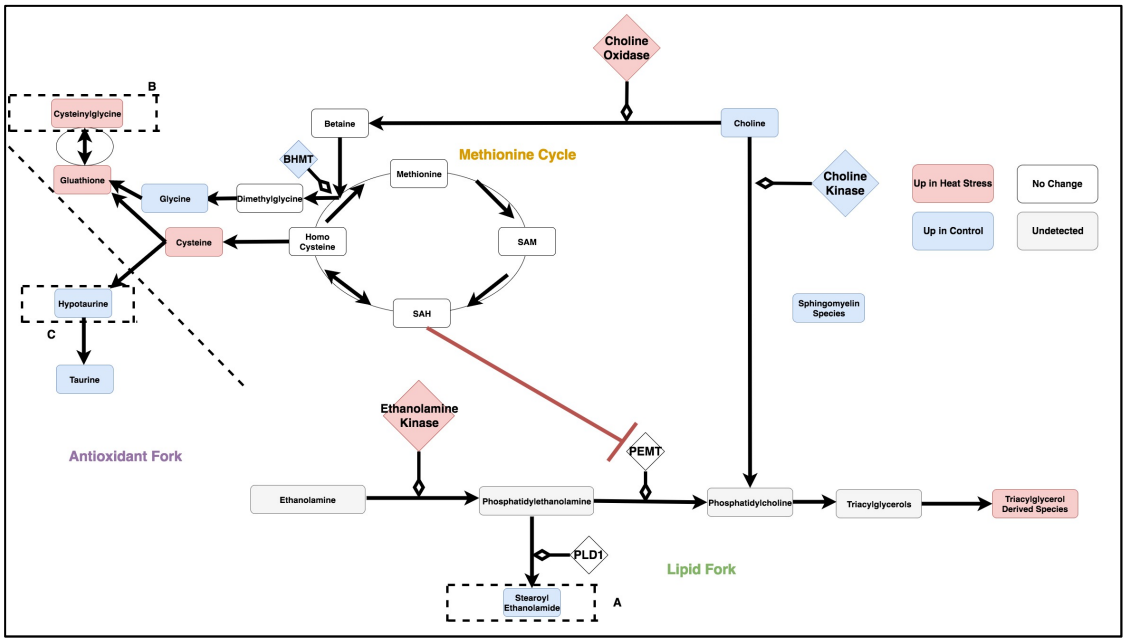

Figure 59: Triplet of stearoyl ethanolamide and (cysteinylglycine / hypotaurine). The compartmentalization of the pathway by regions containing the compounds in the ratio (cysteinylglycine and hypotaurine) is illustrated by the dotted line. For the linear model representing differential behavior of this branch point, see figure 54B. SAM: S-Adenosyl-L-methionine, SAH: S-Adenosyl-L-homocysteine,Glutathione GSSG: Glutathione Disulfide, PEMT: Phosphatidylethanolamine N-methyltransferase, BHMT: Betaine--Homocysteine S-Methyltransferase, PLD1: Phospholipase D-1, PEMT: Phosphatidylethanolamine N-methyltransferase.

Under heat stress conditions**,** stearoyl ethanolamide levels correlate well with ratios of the reduced glutathione derivative, cysteinylglcyine, and hypotaurine (Figures 54B and 55B). This latter quantity represents a metabolic fork underlying sulfur metabolism, which favors glutathione under heat stress (Figure 59). Under control conditions, activation of the sulfur metabolism would inhibit an important component of stearoyl ethanolamide production via SAH-related inhibition of PEMT. This mechanism is countered under heat stress conditions with an increase in the ratio of

PEMT/SAM correlating with rising levels of gamma glutamylcysteine (Figures 54F and 55F). Stearoyl ethanolamide levels and the ratio of the reduced glutathione derivative, cysteinylglycine, and Glutathione GSSG (Glutathione Disulfide) show strong patterns of differential correlation between control and heat stress (Figs 54C + 55C). This is consistent with concerted regulation of several metabolic forks in the underlying circuit of carbon metabolism (Figure 56).

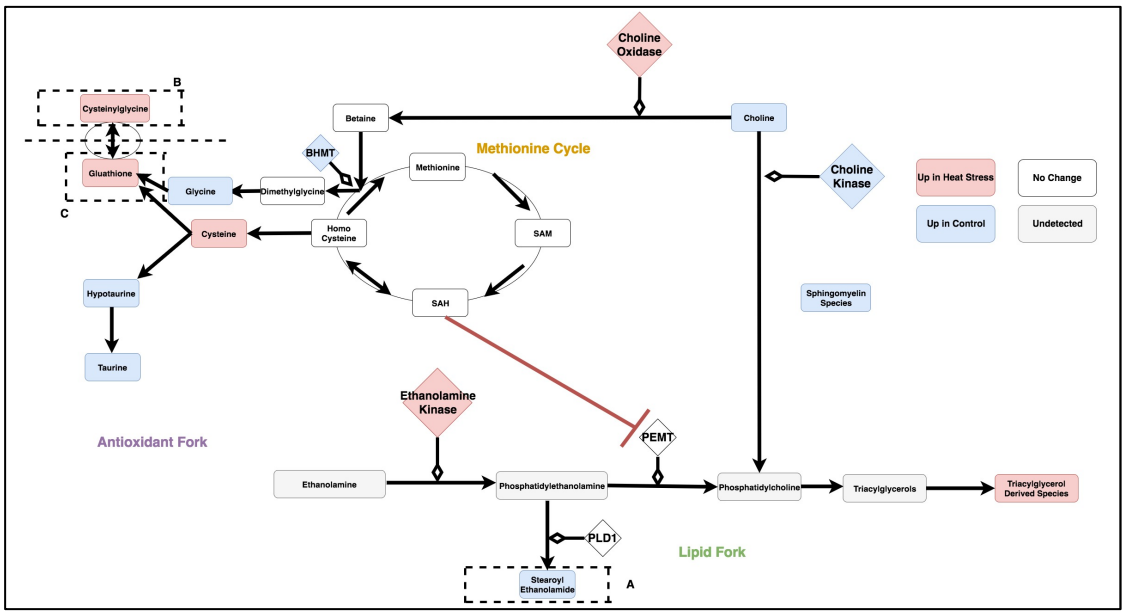

Figure 60: Triplet of stearoyl ethanolamide and (cysteinylglycine / gluathatione). The compartmentalization of the pathway by regions containing the compounds in the ratio (cysteinylglycine and glutathione) is illustrated by the dotted line. For the linear model representing differential behavior of this branch point, see figure 54C. SAM: S-Adenosyl-L-methionine, SAH: S-Adenosyl-L-homocysteine,Glutathione GSSG: Glutathione Disulfide, PEMT: Phosphatidylethanolamine N-methyltransferase, BHMT: Betaine--Homocysteine S-Methyltransferase, PLD1: Phospholipase D-1, PEMT: Phosphatidylethanolamine N-methyltransferase.

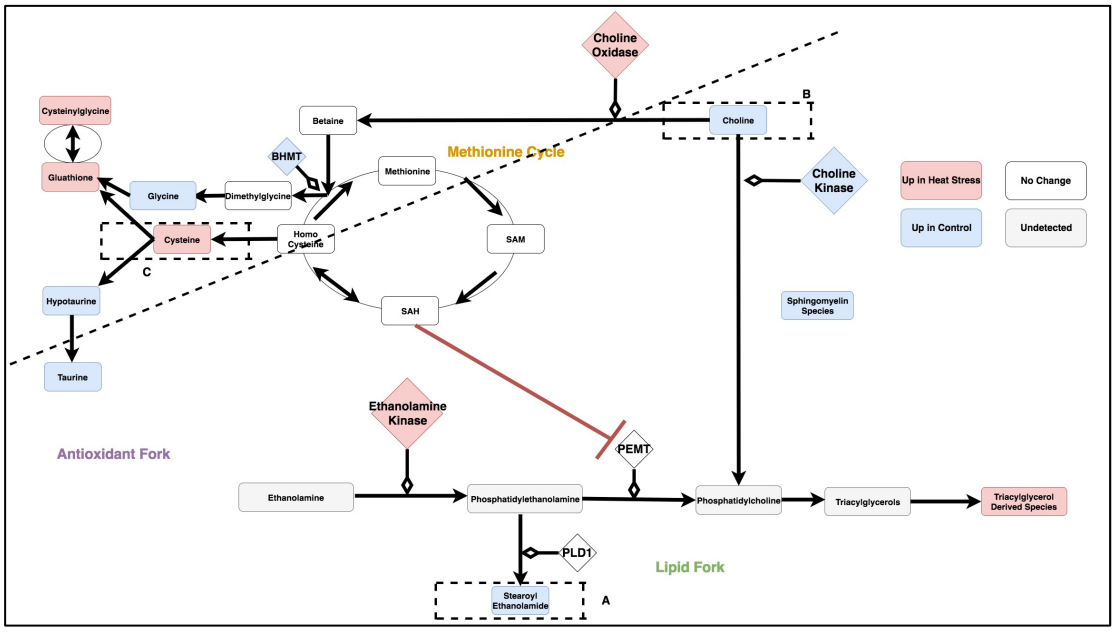

Figure 61: Triplet of stearoyl ethanolamide and (cysteine / choline). The compartmentalization of the pathway by regions containing the compounds in the ratio (choline and cysteine) is illustrated by the dotted line. For the linear model representing differential behavior of this branch point, see figure 54D. SAM: S-Adenosyl-L-methionine, SAH: S-Adenosyl-L-homocysteine,Glutathione GSSG: Glutathione Disulfide, PEMT: Phosphatidylethanolamine N-methyltransferase, BHMT: Betaine--Homocysteine S-Methyltransferase, PLD1: Phospholipase D-1, PEMT: Phosphatidylethanolamine N-methyltransferase.

Stearoyl ethanolamide levels and the ratio of cysteine and choline shows strong patterns of differential correlation between control and heat stress (Figures 54D and 55D). Under the proposed mechanism, as cysteine metabolism is increased during heat stress, choline decreases with its remaining levels critical to maintain betaine (Figure 61).

5.3.3 Relationship between Cysteine and Stearoyl Ethanolamide, Accounted for by Circuit

Stearoyl Ethanolamide and cysteine demonstrate differential relationships between control and heat stress conditions (p-value of interaction term < .05). The physiological roles of stearoyl ethanolamide are not fully established, although it has been shown to have anti-inflammatory properties (Ezzili et al., 2010). This makes the compound similar to many other metabolites involved in the heat stress response. Though stearoyl ethanolamide levels are lower under heat stress, its correlation with cysteine indicates regulatory coupling during heat stress response. This connection would support a metabolic circuit connecting antioxidant and lipid metabolism. Transcriptome data also supports increased utilization of pathways for cysteine production under heat stress in a way that influences lipid production (Figure 54E).

5.3.4 Discussion of Mechanistic Regulation

An important goal of modern genomics is determining the mechanisms that control physiology. Our computational analysis in this chapter provides insight into gene expression changes and associated shifts in metabolite levels that may influence physiology under heat stress. This is analysis provides network context for previous studies that have identified differential gene expression and metabolite levels. Systems biology studies can use multi-omics data to identify elements of regulation, integrating them into concrete networks that generate hypotheses about large-scale regulation. Collectively, these changes in gene expression and metabolic forks identified by this work provide mechanistic context for the differential relationship between stearoyl ethanolamide and cysteine during heat stress. The insights from this study expand the role of carbon of and sulfur flux during the long-term heat stress response.

This work provides a complete model for isolated observations about lipid and sulfur metabolism that have never been integrated into a full circuit. For example, betaine and choline supplementation has variable effects on bird performance with recent studies suggesting it has limited influence on improving broiler performance and cannot overcome the negative influences of heat stress (Kpodo *et al*., 2015). According our extended circuit, the bird is able to effectively maintain betaine levels under heat stress through redirection of choline such that supplementation may be ineffective at shifting network dynamics. We hypothesize gene regulation shunts choline to betaine, preventing the accumulation of resource deficits. Such changes include downregulation of BHMT, upregulation of choline oxidase and downregulation of choline kinase.

The impact of these changes could be dramatic. Modification of choline accounts for 70 percent of phospohatidylcholine synthesis, with the remaining 30 percent derived from PEMT driven methylation of phosphotidylethanolamine (DeLong *et al*., 1999). This latter pathway also has gene expression changes, such as up-regulation of choline kinase. These changes may compensate for altered choline dynamics during long-term heat stress. Thus, the ability of choline to rescue performance from stress may be stress and organism specific. For example, choline supplementation has been shown in clinical studies to improve antioxidant efficiency in cystic fibrosis patients (Innis *et al.,* 2007) despite its efficacy in influencing livestock performance being equivocal (Kpodo *et al*., 2015).

The hypotheses generated by this work propose mechanisms that may underlie associations from GWAS (genome wide association studies). This is important, as previous work on quantifying broiler performance under heat stress, has relied on QTL

(quantitative trait loci) mapping to identify potentially relevant SNP's controlling relevant metrics. One of these resides in the PEMT gene, implicated in sulfur and lipid metabolism, as being associated with body temperature at Day 20 (Van Goor *et al,* 2015). Our proposed circuit includes PEMT as a critical element in a broader network and provides a possible functional role of the previously identified SNP. Building circuits from individual network units provides biological context for statistical observations in a way that relate components from different, but connected, pathways.

Our approach ultimately creates a network integrating compounds whose role in heat stress are well understood and compounds not previously implicated in the heat stress response. This is particularly useful for providing insight into the relatively uncharacterized lipid, stearoyl ethanolamide, that is altered by heat stress. Stearoyl ethanaolamide is as an example of an n-acylamide endocannabinoid. The functions of n-acylamides are best understood in the brain, where they play important roles in signaling and inflammation response (Raboune *et al*., 2014). Thus, changes in ratios of these species can be informative as they may represent preferential routing of carbon resources to lipids that influence inflammation.

The shifts in compounds such as stearoyl ethanolamide are also consistent with a circuit preferentially directing carbon backbones towards gluconeogenesis and triacylglycerol production. Under this complete model, the bird allocates carbon resources to produce signaling molecules as well as to drive antioxidant and energy production pathways. Leveraging computational methods to understand the nuances of carbon and sulfur flow under heat stress provides a significant improvement in understanding the regulation of the response, and generates a number of testable

hypotheses. For example, the role of leveraging glycine released from protein catabolism, in the context our circuit as a route to glutathione production is another mechanism by which glycine supplementation could improve heat stress performance (Awad *et al*., 2017). The importance of glycine in glutathione production may be in addition to the putative role of glycine in energy production discussed in the first metabolic fork in Chapter 4 (Figure 46). These, and other hypotheses, are being incorporated to plan studies in which feed composition is altered with resources thought to be involved in the major circuits. Additionally, we have successfully captured the logic of the carbon flow under heat stress. The transition from simply determining up or down regulation of certain compounds developing a collection of well-characterized mechanisms to be integrated into circuits is a powerful improvement in using systems biology to integrate large- scale multi-omics data.

5.3.5 Future Work and Emphasis on Novelty

Having developed putative circuits regulating the flux of carbon metabolism across multiple pathways, this work has generated a number of hypotheses about how to influence bird growth and performance under heat stress, while using statistical and computational techniques in a novel fashion. These biological insights are being incorporated to plan studies in which feed composition is altered with supplementation of resources involved in the major circuits proposed in this research. Such predictions are made possible by having used computational techniques to understand the regulation of carbon flow under heat stress. The progression from simply determining up or down regulation of certain compounds, which is a common strategy in earlier differential expression studies, to developing a collection of well-characterized

mechanisms to that can be integrated into circuits is a powerful improvement in using systems biology approaches to integrate large- scale multi-omics data.

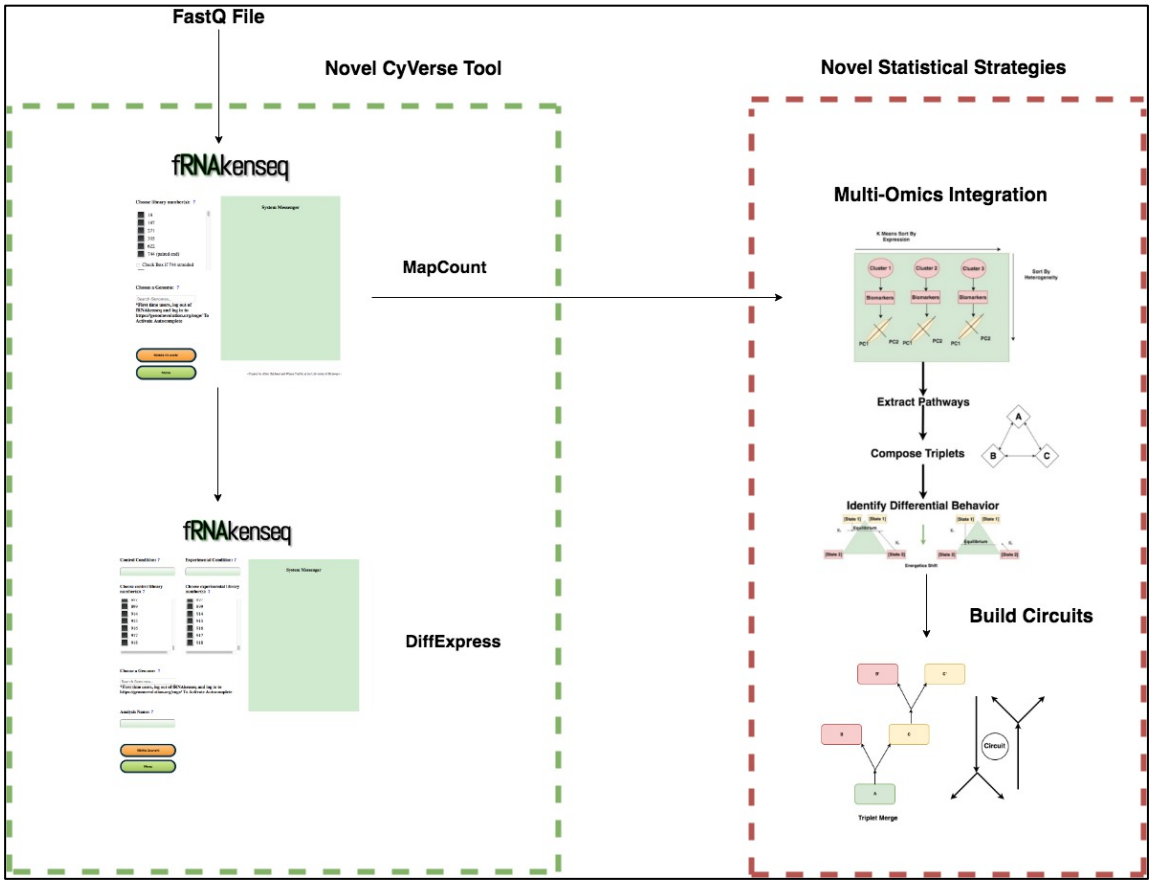

Figure 62: Illustration of the various computational and statistical components of the thesis used to drive biological insight.

Chapter 6

CONCLUSION

This thesis, which represents the material of four manuscripts, describes a natural progression through tools and techniques, from processing raw transcription information into data, through the development of statistical methods that extract biological insight from the resulting high dimensional datasets. Novel aspects of this work are many-fold, encompassing contributions to both biology and bioinformatics. Furthermore, the relevance of this work extends to aid hypothesis generation that provides the foundation for future endeavors in these rapidly-developing fields.

In Chapter 2, we described how work with CyVerse and the construction of the powered-by-CyVerse tool fRNAkenseq demonstrates the value of creating cloud-integrated genomics platforms through utilization of actively developed APIs. fRNAkenseq provides a unique solution to the computation bottleneck of large scale –omics datasets. Its MapCount pipeline provides a method to rapidly process raw sequencing reads into raw or normalized count data. Meanwhile, the DiffExpress pipeline uses multiple differential expression algorithms to effectively deal with false positives when detecting genes whose expression changes under differing conditions. Novel statistical methods developed in this thesis are then utilized to integrate the transcriptome data processed by fRNAkenseq with metabolomics data. The unique infrastructure created by fRNAkenseq facilitates the extraction of maximum biological insight from a complex dataset spanning multiple tissue sources, experimental conditions, and data types. These cohesive analyses uncover novel biology, providing

important detail into exact mechanisms that may control the heat stress response, which can be further tested experimentally.

The steps comprising the post-fRNAkenseq workflow have proven useful both as part of the analysis or as standalone techniques. This was the focus of Chapter 3, in which we proposed using a stringent z-score-based test for tissue enrichment. We demonstrated that this heuristic effectively identifies tissue-specific biology in the large and diverse datasets processed by fRNAkenseq. This can be useful in understanding fine-tuned transcriptome differences between closely related tissues. Many of these insights provided clues into the contribution of gene regulation to tissue defining physiology. This was made possible partly because, due to the stringency of our test, the sets of transcription factors and splicing factors that are enriched are unique to each tissue. It was possible to corroborate the predicted biological roles of these genes that explore consequences of knockouts or deleterious mutations. This provides confidence in the ability of our threshold to identify important biology, and confidence in some of the unexpected predictions that will serve as hypotheses for future research.

We also compared our results with the GTEx method of enrichment, which identifies genes in a tissue with a five-fold difference in means relative to background expression. Our threshold greatly improved tissue specificity because it incorporates standard deviation. The resulting specificity made it possible to decipher subtle biochemical characteristics of each tissue, such as differences in the TCA cycle between cardiac and skeletal muscle. Applying this analysis emphasized that breast muscle tissue is enriched for many genes regulating glycogen and glycolysis metabolism, while cardiac tissue is enriched for nuclear-encoded mitochondrial genes

involved in fatty acid oxidation. The reliance on separate energetic pathways illustrated fundamental biochemical differences between the tissues, which provides context for novel types of regulation suggested by the enrichment analysis. For example, the gene BPGM is enriched in breast muscle and encodes an enzyme that converts a glycolytic intermediate into a metabolite that improves the favorability of de-oxygenated hemoglobin. We have hypothesized that this gene may be an important adaptation in avian skeletal muscle by supporting the oxygenation necessary for flight. This hypothesis is also supported by the fact that BPGM expression is expressed at a similar level in breast muscle between different lines of chicken. In the future, we will further extend this approach to explore gene enrichment patterns in various tissues and how they shift across species. Thus, our enrichment threshold produces a technique useful for comparative genomics.

Notably, our tissue enrichment strategy has also produced candidates for further laboratory experiments. This is important when applying laboratory methods that are generally expensive and relatively low throughput, such as the fluorescent imaging of transcripts. These experiments can be useful to gain spatial information about gene expression, and also to understand the regulation of tissue enriched genes. However, without a hypothesis-guided approach, the efficiency is limited. The stringency of our enrichment method is thus valuable for identifying tissue-specific candidate genes and promoters that can be explored through fluorescent tagging. Additionally, the modules of enriched genes provided a context for GWA studies, in which SNPs that fall in the region of enriched genes may be prioritized. Finally, modules of enriched genes can be integrated into downstream pipelines, by serving as an initial form of feature selection that reduces the transcriptome to tissue important

genes. The heat stress response can therefore be understood from the perspective of such tissue-enriched modules. From this, a pipeline was developed that leverages multiple statistical learning techniques to integrate the expression of liver tissue defining genes and metabolite levels, and prioritize those associated with heat stress regulation. We focus on the liver for this analysis, because it is a source of lipid, antioxidant and sugar production. Levels of these compounds have far-reaching physiological consequences for the performance of the bird. Thus, understanding those most closely associated with the heat stress response could produce candidates for interventions that improve bird performance, such as dietary changes or improved genetic selection.

Chapter 4 describes the culmination of several analytical methods into a comprehensive statistical learning pipeline for integrating metabolomic and transcriptomic data and generating novel hypotheses. We demonstrate how this approach effectively prioritizes heat stress biomolecules by their expression levels (k-means), classifying power (random-forests), and their correlations with one another (PCA), and how these methods recapitulate heat stress responsive pathways. Importantly, this pipeline provides a basis for integrating arbitrarily complex sets of continuous data to identify compounds strongly associated with the heat stress response. This is a decidedly useful advantage enabled through the ability of each step of the pipeline to exploit a different feature of the data. For example, the k-means step first separates out compounds into distinct clusters, random forest then prioritizes biomolecules that best classify heat stress samples, and PCA lastly organizes these biomolecules into highly correlated groups. Thus, the most important features relevant to understanding a biological response are classifying power of each

biomolecule and correlation with potentially related biomolecules. This makes the pipeline amenable to identifying biomolecules across -omics associated with a generic treatment regimen, such as a disease state. In the future, we will use this pipeline to integrate complementary proteomics data to provide a more in-depth systems level understanding of the heat stress response.

The total workflow that incorporates identifying genes enriched in liver tissue and leveraging these three statistical learning techniques also proposes novel hypotheses that can be tested. These are enhanced by building linear models from the prioritized biomolecules that that take as input three metabolites - A, B, and C – to link metabolite A to the ratio of B and C, $\frac{B}{C}$, into a regulatory triplet. Some of these models relate precursors and products, which we refer to as metabolic forks. One of the metabolic forks exhibiting differential behavior between control and heat stress conditions involves the amino acid glycine, the sugar F6P, and the fat precursor G3P. Under heat stress conditions, glycine and F6P are more closely coupled at the expense of the correlation between F6P and G3P under control conditions. Therefore, we propose the following hypothesis:

- Under heat stress conditions, carbon backbones from glycine released from catabolism are directed towards sugar production (F6P), with much greater preference relative to lipid production (G3P). This change is mediated by increased expression of FBP2.

One way to test this hypothesis proposed by our statistical learning pipeline and metabolic fork is to determine if glycine supplementation improves bird performance under heat stress. Recent studies have, in fact, shown this to be the case (Awad *et al.*, 2018). Another method to verify this hypothesis regarding the fate of

catabolized glycine would involve radiolabeling feed-supplemented glycine. We are interested in exploring this and other forms of validation regarding glycine and glucose production. Encouraged by the ability of the linear models associated with metabolic forks to produce biologically informative hypotheses, we also sought to apply this method to some of the other pathways prioritized by the statistical learning pipeline, such as a lipid and sulfur metabolism. Calculating metabolic forks that could explain the regulation of these pathways would make it possible to propose targeted hypotheses about the relationship between lipid and sulfur metabolism.

To extend this analysis, we constructed network skeletons out of metabolic forks. By identifying relationships in a network context, it became possible to construct pathway models over which transcriptome data could be overlaid. The proposal of this complete and novel pathway which connects sulfur and lipid metabolism under heat stress allows us to relate individual mechanisms to a complete system. Moreover, the metabolic forks of this model can be shown to represent targeted hypotheses about sulfur and lipid regulation which have been validated individually in other contexts, yet never related to the heat stress response. A list of several such hypotheses and their experimental validation as mechanisms is provided below:

- **Hypothesis**: The methionine cycle influences lipid production under heat stress

  Literature Evidence: SAH, a product of the methionine cycle, inhibits methyl transferases such as PEMT involved in lipid production (Obeid and Hermann, 2009).

- **Hypothesis**: Lipid and sulfur metabolism are related via the relationship between choline and cysteine.

  Literature Evidence: Choline can improve hyperhomocysteniemia in patients with cystic fibrosis (Innis *et al*., 2007)

  Experimental Evidence: BHMT is down-regulated, choline oxidase is up-regulated under heat stress, indicating a shift in choline to betaine
- **Hypothesis**: Betaine regulation increases cysteine levels for antioxidant production

  Literature Evidence: Mouse knockouts of this gene show highly elevated levels of homocysteine (Strakova *et al*., 2012)

  Experimental Evidence: Betaine levels are reduced under heat stress. Further supporting a relationship between sulfur and lipid metabolism via choline and betaine, Betaine--Homocysteine S-Methyltransferase (BHMT) transcription is downregulated under heat stress. BHMT converts betaine and homocysteine to dimethylglycine and methionine, respectively and mouse knockouts of this gene show highly elevated levels of homocysteine (Strakova *et al*., 2012).

This final chapter of the thesis represents one of its major meaningful biological contributions. By applying this method of integrating metabolic forks to create networks, a consistent model of regulation is generated that describes how shifts in lipids and sulfur metabolism can be coordinated during heat stress. This constructs a solid foundation that facilitates hypotheses generation and prioritization. Importantly, the network skeleton that results provides a framework in which differentially expressed genes can be placed. Ultimately, we are able to couple these genes to mechanisms of carbon regulation, via cysteine and choline, in a way that further

relates antioxidant and biologically active lipids through the re-routing of precursors due to changes in gene expression. Not only are these insights novel contributions to the understanding of heat stress response in avian genomics, they also are highly applicable in the clinical context, as the relationship between sulfur and lipids is still an active area of research (Obeid and Hermann, 2009).

This thesis presents a complete approach to analyzing diverse and complicated data. It begins by solving computational problems associated with large-scale biological experiments exploring the heat stress response, then proposes statistical approaches to identify tissue enriched genes, and subsequently integrates these into pipelines to identify biomolecules across the transcriptome and metabolome that are important to the heat stress response. Finally, pathways associated with these biomolecules are used to develop candidates for metabolic forks and generate networks that prioritize hypotheses to drive future experimentation. This adaptable, integrated approach can be applied in many other contexts involving large-scale diverse datasets, where deciphering specific biological insight is a persisting challenge.

REFERENCES

Afgan, E., Baker, D., van den Beek, M., Blankenberg., D., Rouvier, D., Cech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., Gruning, B., Guerler, A., Hillman-Jackson, J., Von Kuster, G., Rasche, E., Soranzo, N., Turage, N., Taylor, J., Nekrutenko, A., Goecks, J. 2016. The Galaxy Platform for Accessible, Reproducible And Collaborative Biomedical Analyses: 2016. Nucleic Acids Research. 44(W1): W3-W10. DOI:10.1093/nar/gkw343. PMID: 27137889.

Akbarian, A., Michiels, J., Degroote, J., Majdeddin, M., Golian, A., De Smet, S. Association between heat stress and oxidative stress in poultry; mitochondrial dysfunction and dietary interventions with phytochemicals. Journal of Animal Science and Biotechnology. 2016. 7: 37 DOI: 10.1186/s40104-016-0097-5. PMCID: PMC4924307.

Anders, S., Pyl, P.T., Huber, W. HtSeq-a Python framework to work with high-throughput sequencing data. 2015. 31(2):166-169. DOI: 10.1093/bioinformatics/btu638. PMCID: PMC4287950.

Andresen, B.S., Christensen, E., Corydon, T.J., Bross, P., Pilgaard, B., Wanders,R.J., Ruiter, J.P., Simonsen, H., Winter, V.,Knudsen, I., Schroeder, L.D., Gregersen, N., Skovby, F. Isolated 2-methylbutyrylglycinuria caused by short/branched-chain acyl-CoA dehydrogenase deficiency: identification for a new enzyme defect, resolution of its molecular basis, and evidence for distinct acyl-coA dehydrogenase in isoleucine and valine metabolism. American Journal of Human Genetics. 67(5):1095-103. DOI: 10.1086/303105. PMID: 11013134.

Ardley, H., Robinson, P.A. E3 Ubiquitin Ligases. Biochemistry. 2005. 41:15-30. DOI: 10.1042/bse0410015. PMID: 16250895

Arribas, J., Gimenez, E., Marcos, R., Velazquez, A. Novel antiapoptotic effect of TBX15 reduces apoptosis in cancer cells. 2015. Apoptosis. 20(10):1338-46. DOI: 10.1007/s10495-015-1155-8. PMID 26216026.

Awad, E.A., Idrus,Z., Soleimani Farjam,A., Bello, A.U., Jahromi, M.F. Growth performance, duodenal morphology and the caecal microbial population in female broiler chickens fed glycine-fortified low protein diets under heat stress conditions. 2018. British Poultry Science. DOI: 10.1080/00071668.2018.1440377.

Balogh, G., Horvath, I., Nagy, E., Hoyk, Z., Benko, S., Bensaude, O., Vigh, L. "The hyperfluidization of mammalian cell membranes acts as a signal to initiate the heat shock protein response". 2005. FEBS J. 272(23):6077-86. DOI: https://doi.org/10.1016/j.febslet.2013.05.016. PubMed PMID: 16302971.

Balogh, G., Peter, M., Glatz, A., Gombos, I., Torok, Z., Horvath, I., Harwood, J.L., Vigh, L. Key Role of Lipids in Heat Stress Management. 2013. 587(13):1970-80. DOI: https://doi.org/10.1016/j.febslet.2013.05.016. PubMed PMID: 23684645.

Bavrak, F., Komorcu-Bayrak, E., Mutlu, B., Kahveci, G., Erginel-Unultuna, N. Genetic Analysis of the Irx4 gene in hypertrophic cardiomyopathy. 2008. 36(2):90-5. PMID: 18497553.

Biernacki, M., Skrzydlewska, E. Metabolism of Endocannabinoids. 2016. Postepy Hig Med Dosw (Online). 70(0):830-43. PubMed PMID: 27516570.

Bommelje, C.C., Weeda, V.B., Huang, G., Shah, K., Bains, S., Buss, E., Shaha, M., Gonen, M., Ghossein, R., Ramanathan, S.Y., Singh B. Oncogenic Function of SCCRO5/DCUN1D5 requires its Neddylation E3 Activity and Nuclear Localization. Clin Cancer Res. 2014. 20(2):372-81. DOI: 10.1158/1078-0432.CCR-13-1252. PMID: 24192928

Brand, T., Simrick, S.L., Poon, K.L., Schindler, R.F.R. The cAMP-binding Popdc proteins have a redundant function in the heart. Biochem Soc Trans. 2014. 42(2):295-301. DOI: 10.1042/BST20130264. PMID 2462234.

Carratu, L., Franceschelli, S., Pardini, C., Kobayashis, G., Horvath, I., Vigh, L., Maresca. Biochemistry Membrane Lipid Pertrubation Modifies the Set Point of the Temperature of Heat Shock Response in Yeast (A9-desaturase/membrane physical state). 1996. Proc Natl Acad Sci U S A. 93(9): 3870–3875. PubMed PMID: 39451.

Chatzakos, V., Slatis, K., Djureinovic, T., Helleday, T. Hunt, M.C. N-acyl Taurines are anti-proliferative in prostate cancer cells. 2012. Lipids. 47(4):355-61. DOI: 10.1007/s11745-011-3639-9. PMID:22160494.

Chen, N.M., Nesse, A., Dyck, M.:., Steuber, B., Koenih, A.O., Lubeseder-Martellato, C., Forster, T., Boheneberge, H., Kitz, J., Reuter-Jessen, K., Griesmann, H., Gaedcke, J., Grade, M., Zhang, J.S., Tsai, W.C., Siveke, J.P., Schildhaus, H.U., Strobel, P., Johnsen, S.A., Ellenrieder, V., Hessmann, E. Context-Dependent Epigenetic Regulation of Nuclear Factor of Activated T Cells 1 in Pancreatic Plasticity. 2017. Gastroenterology. 152(6):1507-1520.e15. DOI: 10.1053/j.gastro.2017.01.043. PMID: 28188746

Chrast, R., Scott, H.S., Kudoh, J., Rossier, C., Minoschima, S., Wang,Y., Shimizu, N., Antonarakis, S.E. Cloning of two human homologs of the drosophila single-minded gene SIM1 on chromosome 6q and SIM2 on 21q within the Down Syndrome chromosomal region. 1997. Genome Res. 7(6):615-24. PMID: 9199934.

Chu, Y., Corey, D.R. RNA Sequencing: Platform Selection, Experimental Design and Data Interpretation. 2012. Nucleic Acid Ther. DOI: 10.1089/nat.2012.0367. PMCID: PMC3426205.

Chung, N., Jee, B.K., Chae, S.W., Jeon, Y.W., Lee, K.H. 2009. HOX gene analysis of endothelial cell differentiation in human bone marrow-defined mesenchymal stem cells. 2009. Mol Biol Rep. 36(2):227-35. DOI: 10.1007/s11033-007-9171-6. PMID: 17972163.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., M.W. Szczesniak, Gaffney, D.J., Elo, L.L., Zhang, X., Mortazavi. A Survey of Best Practices for RNA-seq data Analysis. 2016. Genome Biol. DOI: 10.1186/s13059-016-0881-8. PMCID: PMC4728800.

Corbin, K.D. Zeisel, S.H. Choline Metabolism Provides Novel Insights into Non-alcoholic Fatty Liver Disease and its Progression. 2012. Curr Opin Gastroenterol. 28(2): 159-165. DOI: 10.1097/MOG.0b013e32834e7b4b. PubMed PMID: 22134222.

Costantini, D.L., Arruda, E.P., Agarwi, P., Kim, K.H., Zhu, Y., Zhu, W., Lebel, M., Cheng, C.W., Park, C.Y., Pierce, S.A., Guerchicoff, A., Pollevick, G.D., Chan, T.Y., Kabir, M.G., Cheng, S.H., Husain, M., Antzelevitch, C., Srivastava, D., Gross, G.J., Hui, C.C., Backy, P.H., Bruneau, B.G. The Homeodomain Transcripion factor Irx5 Establishes the Mouse Cardiac Ventricular Repolarization Gradient. 2005. Cell. 123(2):347-58. DOI:10.1016/j.cell.2005.08.004. PMID:16239150.

Das, A., Buksch, A., Price, C.A., Weitz, J.S. ClearedLeavesDB: and online database of cleared plat leaf images. 2014. Plant Methods. PMID: 24678985. DOI: 10.1186/1746-4811-10-8.

Davis A.P., Capecchi M.R. A mutational analysis of the 5' HoxD genes: dissection of genetic interactions during limb development in the mouse. 1996. Development. 122 (4): 1175–85. DOI:10.1007/s11033-007-9171-6. PMID 8620844.

Delong, J., Sen, Y.J., Thomas, .M.J., Cui, Z. Molecular Distinction of Phosphatidylcholine Synthesis between the CDP-Choline Pathway and Phosphatidylethanolamine Methylation Pathway. 1999. The Journal of Biological Chemistry. 274(42):29683-8. DOI:10.1074/jbc.274.42.29683. PubMed PMID:10514439.

Deng, Y.F., Huang, Y.Y., Lu, W.S., Huang, Y.H., Xian, J., Wei, H.Q., Huang, Q. The Caveolin-3 P104L mutation of LGMD-1C leads to disordered glucose metabolism in muscle cells. 2017. Biochem Biophys Res Commun. 486(2):218-223. DOI: 10.1016/j.bbrc.2017.02.072. PMID: 28232187.

Dickson, G.J., Liberante, F.G., Kettyle, L.M., O'Hagan, K.A., Finnegan, D.P.J., Bullinger, L., Geerts, D., McMullin, M.F., Lappin, T.R.J., Mills, K.I., Thompson, A. *HOXA/PBX3* knockdown impairs growth and sensitizes cytogenetically normal acute myeloid leukemia cells to chemotherapy. 2013. Haematologica. 98(8): 1216–1225. DOI: 10.3324/haematol.2012.079012. PMID:23539541.

Diks, S. H., Bink, R. J., van de Water, S., Joore, J., van Rooijen, C., Verbeek, F. J., den Hertog, J., Peppelenbosch, M. P., Zivkovic, D. The novel gene asb11: a regulator of the size of the progenitor compartment. 2006. J. Cell Biol. 174: 581-592. DOI: 10.1083/jcb.200601081. PMCID: PMC2064263.

Dooley, R., Vaughn, M., Stanzione, D., Terry, S., Skidmore, E. "Software-as-a-Service: The iPlant Foundation API", 5th IEEE Workshop on Many-Task Computing on Grids and Supercomputers (MTAGS). IEEE, 2012.

The ENCODE Project Consortium. An Integrated Encyclopedia of DNA Elements in the Human Genome. 2012. Nature. 489(7414):57-74. DOI: 10.1038/nature11247.

Ermak, G., Davies, K.J. Calcium and Oxidative Stress: From Cell Signaling to Cell Death. 2002. Mol Immunol. 38(10):713-21. PubMed PMID: 11841831.

Ezzili, C., Otrubova, K., Boger, D.L. Fatty Acid Amide Signaling Molecules. Bioorg Med Chem Lett. 2010. 20(20):5959-5968. DOI: 10.1016/j.bmcl.2010.08.048. PMID:20817522.

Fergus, J.E., Wu, Y., Smith, K., Charles, P., Powers, K., Wang, H., Patterson, C. ASB4 is a hydroxylation substrate of FIH and promotes vascular differentiation via an oxygen-dependent mechanism. 2007. Mol Cell Biol. DOI:10.1128/MCB.00511-07. PMID: 17636018.

Fleischer, S., Ogunbunmi, E.M., Dixon, M.C., Fleer, E.A. Localization of Ca2+ Release Channels with Ryanodine in Junctional Terminal Cisternae of Sarcoplasmic Reticulum of Fast Skeletal Muscle. 1985. Proc Natl Acad Sci U S A. 82(21):7256-9.

Fuhrer, T., Zamboni, N. High-throughput Discovery Metabolomics. 2015. Current Opinions Biotechnology. 31:73-78. DOI: 10.1016/j.copbio.2014.08.006. PMID: 25197792.

Fulda, S., Gorman, A.M., Hori, O., Samali, A. Cellular Stress Responses: Cell Survival and Cell Death. 2010. International Journal of Cell Biology. ID 214074.

Gieger, C., Geistlinger, L., Altmaier, E., Hrabe de Angelis., Kronenberg, F., Meitinger, T., Mewes, H.W., Wichmann, H.E., Weinberger, K.M., Adamski, J., Illig, T., Suhre, K. Genetics Meets Metabolomics: A Genome-Wide Association Study of Metabolite Profiles in Human Serum. 2008. PLoS Genet. 4(11):e1000282. DOI:10.1371/journal.pgen.1000282. PubMed PMID: 19043545.

Green D, Marks AR, Fleischer S, McIntyre JO. Wild type and mutant human heart (R)-3-hydroxybutyrate dehydrogenase expressed in insect cells. 1996. Biochemistry. 35(25):8158-65. DOI:10.1021/bi952807n. PMID:8679568

Griffith, O..L., Pepin, F., Enache, O.M., Collisson, E.A., Spellman, P.T. A Robust prognostic signature for hormone-positive node-negative breast cancer. 2013. Genome Med. 11:5(10):92. DOI: 10.1186/gm496.

The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* 08 May 2015. Vol. 348, Issue 6235, pp 648-660. DOI: 10.1126/science.1262110

Gutkowska, J., Jankowski, M., Antunes-Rodrigues, J. The Role of Oxytocin in Cardiovascular Regulation. 2014. J Neuroendocrinol. DOI: 10.1111/j.1365-2826.2011.02235.x. PMID:21981277.

Harjes, C.E., Bai, L., Eun-Ha, K., Yang, X., Skinner, D.J., Fu, Z., Mitchell, S., Li, Q., Fernandez, M.G.S., Zaharieval, R.B., Fu, Y., Palacios, N., Li, J., DellaPenna, D., Brutnell, T., Buckler, E.S., Warburton, M.L., Rocheford, T. Rare genetic variation at Zea mays crtRB1 increases β-carotene in maize grain. 2010. Nature Genetics. 42(4):322-7. DOI: 10.1038/ng.551.

Heng, L., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. The Sequence Alignment/Map Format and SAMtools. 2009. Bioinformatics. 25(16):2078-2079.

Hoffman, N.R. The Plasma Membrane as First Responder to Heat Stress. 2009. Plant Cell. 21(9):2544. DOI: 10.1105/tpc.109.210912. PubMed PMID: PMC2768915.

Holness, M.J., Sugden, M.C. Regulation of pyruvate dehydrogenase complex activity by reversible phosphorylation. 2003. Biochem Soc Trans. 35(25):8158-65. DOI:10.1021/bi952807n PMID: 8679568.

Hong, T.T., Shaw, R.M. Cardiac T-Tubule Microanatomy and Function. 2017. Physiological Reviews. 97(1):227-252. DOI: 10.1152/physrev.00037.2015. PMID 27881552.

Huang, D.W., Sherman, B.T., Lempicki, R.A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. 2009. Nucleic Acids Res. 37(1):1-13.

Human Metabolome Database. Record Name: 8,11 14-Eisotrienoic Acid. URL: http://hmdb.ca/metabolites/HMDB0925

The IBP Breeding Management System Version 3.0.9. (December 2015) The Integrated Breeding Platform. https://www.integratedbreeding.net/breeding-management-system.

Illig, T., Gieger, C., Zhai,H., Romisch-Margl, W., Wang-Sattler, R., Prehn, C., Altmaier, E., Kastenmuller, G., Kato, B.S., Mewes, H.W., Meitinger, T., de Angelis, M.H., Kronenberg, F., Soranzo, N., Wichmann, H.E., Spector, T.D. A genome-wide perspective of genetic variation in human metabolism. Nat Genet. 2010. 42(2):137-41. DOI: 10.1038/ng.507. PubMed PMID: 20037589.

Innis, S.M., Davidson, G.F., Melynk, S., James, S.J. Choline-related supplements improve abnormal plasma methionine-homocysteine metabolites and glutathione status in children with cystic fibrosis. 2007 American Society for Clinical Nutrition. 85(3):702-8. PMID:17344490.

Ichida, M., Endo, H., Ikeda, U., Matsuda, C., Ueno, E., Shimada, K., Kagawa, Y. MyoD is Indispensable for muscle-specific alternative splicing in mouse mitochondrial ATP synthase gamma-subunit pre-mRNA. 1998. J Biol Chem. 273(14):8492-501.

Ikeda Y, Tanaka K (1983). "Purification and characterization of isovaleryl coenzyme A dehydrogenase from rat liver mitochondria". J. Biol. Chem. 258 (2): 1077–85. PMID: 6401713.

Ippolito, D., Lewis, J.A.., Chenggang, Y., Leon, L.R., Stallings, J. Alteration in Circulating Metabolites during and after heat stress in the conscious rat: potential biomarkers of exposure and organ-specific injury. 2014. BMC Physiology. DOI: 10.1186/s12899-014-0014-0.

Ishikura, K., Miyazaki, T., Song-Gyu, R., Endo, S., Nakamura, Y., Matsuzaka, T., Miyakawa, S., Ohmori, H. Effect of Taurine Supplementation of the Alterations in Amino Aid Content in Skeletal Muscle with Exercise in Rat. 2011. J Sports Sci Med. 10(2):306-314. PMCID: PMC3761861.

Iwasa K., Furukawa, Y., Yoshikawa, H., Yamada, M. Caveolin-3 is aberrantly expressed in skeletal muscle cells in myasthenia gravis. 2016. J Neuroimmunol. PMID: 27863830

Jastrebski, S.F., Lamont, S.J., Schmidt, C.J. Chicken hepatic response to chronic heat stress using integrated transcriptome and metabolome analysis. 2017. PloS One. 12(7):e0181900. DOI: 10.1371/journal.pone.0181900. PMID: 28759571.

Jenkins, G.M., Richards, A., Wahl, T., Mao, C., Obeid, L., Hannun, Y. Involvement of yeast sphingolipids in the heat stress resonse of Saccharomyces cerevisae. 1997. 272(51):32566-72. PubMed PMID:9405471.

Joshi, T., Fitzpatrick, M.R., Chen, S., Liu, Y., Zhang, H., Endacott, R.Z., Gaudiello, E.C., Stacey, G., Nguyen, H.T., Zu, D. “Soybean Knowledge Base (SoyKB): a Web Resource for Integration of Soybean Translational Genomics and Molecular Breeding.” Nucl Acids Res. (1 January 2014) 42(D1): D1245-1252. Doi:10.1093/par/gkt905

Julian, R.J. Rapid Growth Problems: ascites and skeletal deformities in broilers. 1998. Poultry Science. 77(12):1773-80. PMID: 9872578. DOI: 10.1093/ps/77.12.1773

Kamura T., Hara T., Matsumoto M., Ishida N., Okumura F., Hatakeyama S., Yoshida M., Nakayama K., Nakayama K.Cytoplasmic ubiquitin ligase KPC regulates proteolysis of p27(Kip1) at G1 phase. Nat. Cell Biol. 6:1229-1235(2004)

Katz, A.M. Regulation of Cardiac Muscle Contractility. 1967. J Gen Physiol. 50(6): 185–196. PMCID: PMC2225748

Kerst, B., Mennerich, D., Schuelke, M., Stottenburg-Didinger, G., von Moers, A., Gossrau, R., van Landeghem, F.K., Speer, A., Braun, T., Hubner, C. Heterozygous myogenic factor 6 mutation associated with myopathy and severe course of Becker muscular dystrophy. Neuromuscul Disord, 2000 Dec. PMID 11053684

Kile, B.T., Schulman, B.A., Alexander, W.S., Nicola, N.A., Martin, H.M., Hilton, D.J. The SOCS box: a tale of destruction and degradation. Trends Biochem Sci. 2002 May;27(5):235-41.

Kim, D., Langmead, B., Salzber, S. HISAT: a Fast Spliced Aligner with Low Memory Requirements. 2015. Nature Methods 12:357-360. Doi: 10.1038/nmeth.3317.

Kirk, E.P., Sunde, M., Costa, M.C., Rankin, S.A., Wolstein, O., Castro, M.L., Butler, T.L., Hyun, C., Guo, G., Otway, R., Mackay, J.P., Wadd, L.B., Cole, A.D., Haywar, C., Keogh, A., Macdonald, P., Griffiths, L., Fatkin, D., Schooler, G.F., Zorn, A.M., Feneley, M.P., Winlaw, D.S., Harvey, R.P. Mutations in Cardiac T-Box Factor Gene *TBX20* Are Associated with Diverse Cardiac Pathologies, Including Defects of Septation and Valvulogenesis and Cardiomyopathy. Am J Hum Genet. 2007 Aug; 81(2): 280–291. Published online 2007 Jun 15. doi: 10.1086/519530 PMCID: PMC1950799

Kohroki, J., Nishiyama, T., Nakamura, T., Masuho, Y. ASB proteins interact with Cullin5 and RBx2 to form E3 ubiquitin ligase complexes. FEBS Lett. Dec 19; 579(30):6796-802. Epub 2005 Nov 28.

Kohsaka S, Shukla N, Ameur N, Ito T, Ng CK, Wang L, Lim D, et al. A recurrent neomorphic mutation in MYOD1 defines a clinically aggressive subset of embryonal rhabdomyosarcoma associated with PI3K–AKT pathway mutations. Nat Genet 2014 May 4 [Epub ahead of print].

Korotchkina, L. G., Patel, M. S. Site specificity of four pyruvate dehydrogenase kinase isoenzymes toward the three phosphorylation sites of human pyruvate dehydrogenase. J. Biol. Chem. 276: 37223-37229, 2001.

Kotaka, M., Kostin, S., Ngai, S., Chan, K., Lau, Y., Lee, S. M. Y., Li, H., Ng, E. K. O., Schaper, J., Tsui, S. K. W., Fung, K., Lee, C., Waye, M. M. Y. Interaction of hCLIM1, an Enigma family protein, with alpha-actinin 2. J. Cell. Biochem. 78: 558-565, 2000.

Kouassi Kpodo, M. O. Smith, and R. Beckford. Response of heat-stressed broilers to dietary choline and betaine I: Performance and carcass characteristics. 2015. Poultry Science.

Kovats, R.S., Shakoor, H. Heat Stress and Public Health: A Critical Review. Annual Review of Public Healt. Vol. 29:41-55. DOI:https://doi.org/10.1146/annurev.publhealth.29.020907.090843

Krumsiek, J., Stuckler, F., Suhre, K., Gieger, C., Spector, T.D. Network-based metabolite ratios for an improved functional characterization of genome-wide association study results. bioRxiv. 2016.

Kuwahara, K., Barrientos, T., Pipes, G.C., Li, S., Olson, E.N. Muscle-specific signaling mechanism that links actin dynamics to serum response factor. Mol Cell Biol. 2005 Apr;25(8):3173-81.

Hasty P, Bradley A, Morris JH, Edmondson DG, Venuti JM, Olson EN, Klein WH (August 1993). "Muscle deficiency and neonatal death in mice with a targeted mutation in the myogenin gene". Nature. 364 (6437): 501–6. PMID 8393145. DOI:10.1038/364501a0.

Ikeda Y, Tanaka K. Purification and characterization of isovaleryl coenzyme A dehydrogenase from rat liver mitochondria. 1983. J. Biol. Chem. 258 (2): 1077–85. PMID: 6401713.

Iwasa K., Furukawa, Y., Yoshikawa, H., Yamada, M. *et al*. Caveolin-3 is aberrantly expressed in skeletal muscle cells in myasthenia gravis. 2016. J Neuroimmunol. 301:30-34 Doi: 10.1016/j.jneuroim.2016.10.011 PMID: 27863830

Ivy, J.L. Muscle Glycogen Synthesis Before and After Exercise. 1991. Sports Med. 11(1):6-19. PMID: 2011684

Jonhson, C.H., Ivanisevic, J., Benton, P.H., Siuzdak, G. Bioinformatics: the next frontier of metabolomics. 2015. Anal Chem. 87(1):147-156. DOI: 10.1021/ac5040693. PMCID: PMC4287838.

Kamura T., Hara T., Matsumoto M., Ishida N., Okumura F., Hatakeyama S., Yoshida M., Nakayama K., Nakayama K. 2004. Cytoplasmic ubiquitin ligase KPC regulates proteolysis of p27(Kip1) at G1 phase. Nat. Cell Biol. 6:1229-1235 DOI: 10.1038/ncb1194. PMID:15531880.

Kerst, B., Mennerich, D., Schuelke, M., Stottenburg-Didinger, G., von Moers, A., Gossrau, R., van Landeghem, F.K., Speer, A., Braun, T., Hubner, C. 2000. Heterozygous myogenic factor 6 mutation associated with myopathy and severe course of Becker muscular dystrophy. Neuromuscul Disord. 10(8):572-7. PMID: 11053684.

Kile, B.T., Schulman, B.A., Alexander, W.S., Nicola, N.A., Martin, H.M., Hilton, D.J. The SOCS box: a tale of destruction and degradation. 2002. Trends Biochem Sci. 2002. 27(5):235-41. PMID:12076535.

Kirk, E.P., Sunde, M., Costa, M.C., Rankin, S.A., Wolstein, O., Castro, M.L., Butler, T.L., Hyun, C., Guo, G., Otway, R., Mackay, J.P., Wadd, L.B., Cole, A.D., Haywar, C., Keogh, A., Macdonald, P., Griffiths, L., Fatkin, D., Schooler, G.F., Zorn, A.M., Feneley, M.P., Winlaw, D.S., Harvey, R.P. 2007. Mutations in Cardiac T-Box Factor Gene *TBX20* Are Associated with Diverse Cardiac Pathologies, Including Defects of Septation and Valvulogenesis and Cardiomyopathy. Am J Hum Genet. 81(2): 280–291. DOI: 10.1086/519530. PMCID: PMC1950799.

Kohroki, J., Nishiyama, T., Nakamura, T., Masuho, Y. ASB proteins interact with Cullin5 and RBx2 to form E3 ubiquitin ligase complexes. 2005. FEBS Lett. 579(30):6796-802. PMID:16325183. DOI:10.1016/j.febslet.2005.11.016

Kohsaka S, Shukla N, Ameur N, Ito T, Ng CK, Wang L, Lim D. A recurrent neomorphic mutation in MYOD1 defines a clinically aggressive subset of embryonal rhabdomyosarcoma associated with PI3K–AKT pathway mutations. 2014 Nat Genet. DOI: 10.1038/ng.2969. PMID:24793135

Korotchkina, L. G., Patel, M. S. Site specificity of four pyruvate dehydrogenase kinase isoenzymes toward the three phosphorylation sites of human pyruvate dehydrogenase. 2001. J. Biol. Chem. 276: 37223-37229. DOI:10.1074/jbc.M103069200. PMID:11486000.

Kotaka, M., Kostin, S., Ngai, S., Chan, K., Lau, Y., Lee, S. M. Y., Li, H., Ng, E. K. O., Schaper, J., Tsui, S. K. W., Fung, K., Lee, C., Waye, M. M. Y. Interaction of hCLIM1, an Enigma family protein, with alpha-actinin 2000. J. Cell. Biochem. 78: 558-565. PMID: 10861853.

Kovats, R.S., Shakoor, H. Heat Stress and Public Health: A Critical Review. Annual Review of Public Health. 29:41-55. DOI: https://doi.org/10.1146/annurev.publhealth.29.020907.090843. PMID:18031221.

Kuwahara, K., Barrientos, T., Pipes, G.C., Li, S., Olson, E.N. Muscle-specific signaling mechanism that links actin dynamics to serum response factor. 2005. Mol Cell Biol. 25(8):3173-81. DOI: 10.1128/MCB.25.8.3173-3181.2005 PMID:15798203.

Lamsoul, I., Erar, M., van der Ven, P.F.M., Lutz, P.G. F. Filamins but Not Janus Kinases Are Substrates of the ASB2α Cullin-Ring E3 Ubiquitin Ligase in Hematopoietic Cells. 2012. PLoS One.7(8): e43798. DOI: 10.1371/journal.pone.0043798. PMID:22916308.

Langmead, B., Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. 2012. 9(4):357-9. Nature Methods. DOI: 10.1038/nmeth.1923. PMID: 22388286.

Lara, L.J., Rostagno, M.H. Impact of Heat Stress on Poultry Production. Animals (Basel). 2013 3(2): 356-369. DOI: 10.3390/ani3020356. PMCID: PMC4494392.

Lee, J. T., Wheeler, T. C., Li, L., Chin, L.-S. Ubiquitination of alpha-synuclein by Siah-1 promotes alpha-synuclein aggregation and apoptotic cell death. 2008. Hum. Molec. Genet. 17: 906-917. DOI:10.1093/hmg/ddm363. PMID:18065497.

Lee, J., Zhou, P., DCAFs, the Missing Link of the CUL4-DDB1 Ubiquitin Ligase. 2007. Molecular Cell. 26(6):775-80. DOI:10.1016/j.molcel.2007.06.001. PMID:17588513.

Lei, Y., Henderson, B.R., Emmanuel, C., Harnett, P.R., DeFazio, A. Inhibition of ANKRD1 sensitizes human ovarian cancer cells to endoplasmic reticulum stress-induced apoptosis. 2015. Oncogene. 34(4):485-95. DOI:10.1038/onc.2013.566. PMID:24531715.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. The Sequence Alignment/Map format and SAMtools. 2009. Bioinformatics. DOI: 10.1093/bioinformatics/btp352. 25(16):2078-2079 PMCID: PMC2723002.

Li, Y.R., Li, J., Zhao, S.D., Bradfield, J.P., Mentch, F.D., MAggadottir, S.M., Hou, C., Abrams, D.J., Chang, D., Gao, F., Wei, Z., Connolly, J.J.,Cardinale, C.J., Bakay, M., Glessner, J.T., Li, D., Kao, C., Thomas, K.A., Qiu, H., Chivacci, H.M., Kim, C.E., Wang, F., Snyder, J., Richie, M.D. Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases. 2015. Nature Medicine. 2015. 21(9):1018-21. DOI:10.1938/nm.3933.

Li, Y., Yang, X.H., Fang, S.J., Qin, C.F., Sun,R.L.,Liu, Z.Y., Jian, B.Y., Wu, X., Li, G. HOXA7 stimulates human hepatocellular carcinoma proliferation through cyclin E1/CDK2. 2015. Oncol Rep. 33(2):990-6. DOI:10.3892/or.2014.3668. PMID:25501982.

Liao, Y., Smyth, G.K., Shi, W. featureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features. 2014. 30(7):923-930. DOI: 10.1093/bioinformatics/btt656.

Liu, X., Zhang, L., Chen, S. Modeling Exon-Specific Bias Distribution Improves the Analysis of RNA-Seq Data. 2015. PLOS one. DOI: https://doi.org/10.1371/journal.pone.0140032.

Lu, Z., Wang, L., van Buren, P., Ware, D. SciApps.org – a federated platform powered by CyVerse. 2016. International Conference on Intelligent Systems for Molecular Biology

Lusherbourgh, C.M., Jennewein, D.M., Brendel, V.P. "The BioExtract Server: a web-based bioinformatics workflow platform" Nucleic Acids Res (2011). 39 (suppl_2): W528-532.

Lyons, E., Pedersen, B., Kane, J., Alam, M., Ming, R., Tang, H., Wang, X., Bowers, J., Paterson, A., Lisch, D. (2008) Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar and grape: CoGe with rosids. Plant Phys 148:1772–1781.

Mabb, A.M., Ehlers, M.D. Ubiquitination in Postsynaptic Function and Plasticity. Annu Rev Cell Dev Biol. 2010. 26: 179-210. DOI: 10.1146/annurev-cellbio-100109-104129. PMID:20604708.

Maerki, S., Olma, M. H., Staubli, T., Steigemann, P., Gerlich, D. W., Quadroni, M., Sumara, I., Peter, M. 2009. The Cul3-KLHL21 E3 ubiquitin ligase targets Aurora B to midzone microtubules in anaphase and is required for cytokinesis. J. Cell Biol. 187: 791-800. PMID:19995937 DOI: 10.1083/jcb.200906117.

Maier, T., Guell, M., Serrano, L.. Correlation of mRNA and protein in complex biological samples. 2009. 583(24):3966-3973. DOI: https://doi.org/10.1016/j.febslet.2009.10.036.

Mangan, S., and Alon, A. Structure and Function of the Feed-Forward Loop Network Motif, PNAS, (2003), 100, 11980-11986.

Marrocco, V., Fiore, P., Benedetti, A., Pisu, S., Rizzuto, E., Musaro, A., Madaro, L, Lozanoska-Oscher, B., Bouche, M. Pharmacological Inhibition of PKCθ Counteracts Muscle Disease in a Mouse Model of Duchenne Muscular Dystrophy. 2017. 16:150-161. doi: 10.1016/j.ebiom.2017.01.001. PMID:28089792.

Marunouchi, T., Abe, Y., Murata, M., Inomata, S., Sanbe, A., Takagi, N, Tanonaka, K. 2013. Changes in small heat shock proteins HSPB1, HSPB5 and HSPB8 in mitochondria of the failing heart following myocardial infarction in rats. Biol Pharm Bull. 36(4):529-39. PMID:23546289.

Maston, G.A., Evans, S.K., Green, M.R. 2006. Transcriptional Regulatory Elements in the Human Genome. Annu Rev Genomics Hum Genet. 7:29-59. DOI:10.1146/annurev.genom.7.080505.115623 PMID:16719718.

Mazella, J., Beraud-Dufour, S., Devader, C., Massa, F., Coppola, T. Neurotensin and its receptors in the control of glucose homeostasis. 2012. Front Endocrinol (Lausanne). 3(143). DOI: 10.3389/fendo.2012.00143. PMCID: PMC3515879.

McCay, J.C. Biology of Breeding Poultry. 2009. ISBN 9781845933753. DOI: 10.1079/9781845933753.0003.

McDaneld, T.G., Hannon, K., Moody, D.E. Ankyrin repeat and SOCS box protein 15 regulates protein synthesis in skeletal muscle. 2006 Am J Physiol Regul Integr Comp Physiol. 290(6):R1672-82. DOI:10.1152/ajpregu.00239.2005 PMID:16424087.

Merchant, Nirav, et al., "The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences," PLOS Biology (2016), doi: 10.1371/journal.pbio.1002342.

Miller, M.A., Pfeiffer, W., Schwartz, T. Creating the CIPRES science gateway for inference of large phylogenetic trees. 2010. Gateway Computing Environments Workshop (GCE). New Orleans, LA, 2010, pp. 1-8. DOI: 10.1109/GCE.2010.5676129URL.

Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D.J., Salichos, L., Zhang, J., Weinstock, G.M., Weinstock, Isaacs, F., Rozowsky, J., Gerstein, M. "The Real Cost Of Sequencing: Scaling Computation to Keep Pace with Data Generation." *Genome Biology* 2016, 17:53. 23 March 2016.

Mykles, D., Burnett, K., Durica, D., Blake, J., Mccarthy, F., Schmidt, C., Stillman, J. Resources and Recommendations for Using Transcriptomics to Address Grand Challenges in Comparative Biology. Integrative and Comparative Biology. DOI: 10.1093/icb/icw083.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., Snyder, M. The transcriptional landsape of the yeast genome defined by RNA Sequencing. 2008. Science. 320(5881):1344-1349. DOI: 10.1126/science.1158441.

Neubauer G, King A, Rappsilber J, Calvio C, Watson M, Ajuh P, Sleeman J, Lamond A, Mann, M. 1998 "Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex". Nat Genet. 20 (1): 46–50. DOI:10.1038/1700. PMID:9731529.

Obeid, R., Herrman, W. Homocysteine and lipids: S-Adenosyl Methionine as a key intermediate. 2009. FEBS Lett. 583:1215-1225. DOI: https://doi.org/10.1016/j.febslet.2009.03.038.

Ono, R., Kaisho, T., Tanaka, T. PDLIM1 Inhibits NF-kB-mediated Inflammatory Signaling by Sequestering the p65 subunit of NF-kB in the cytoplasm. 2015. 5:18327. Scientific Reports. DOI:10.1038/srep18327.

Ottolenghi, C., Fellous, M., Barbieri, M., McElreavey, K. Novel Paralogy Relations Among Human Chromosomes Support a Link Between the Phylogeny of Doublesex-related genes and the Evolution of sex Determination. 2002. 79(3):333-43. *Genomics*. DOI:10.1006/geno.2002.6711 PMID: 11863363.

Owen, J.B., Butterfield, O. Measurement of Oxidized /Reduced Glutathione Ratio. Methods Mol Biol. 2010:648-269. Doi:1007/978-1-60761-756-3_1B

Pantoja-Melendez, C.A., Miranda-Duarte, A., Roque-Ramirez, B., Zenteno, J.C. Epidemiological and Molecular Characterization of a Mexican Population Isolate with High Prevalence of Limb-Girdle Muscular Dystrophy Type 2A Due to a Novel Calpain-3 Mutation. 2017. PLoS One. https://doi.org/10.1371/journal.pone.0170280 PMID: 28103310.

Park, S.Y., Gifford, J.R., Andtbacka, R.H.I., Trinity, J.D., Hyngstrom, J.R., Garten, R.S., Diakos, N.A., Ives, S.J., Dela, F., Larsen, S., Drakos, S., Richardson. 2014. Cardiac, skeletal, and smooth muscle mitochondrial respiration: are all mitochondria created equal. Am J Physiol Heart Circ Physiol. 307(3): H346–H352. doi: 10.1152/ajpheart.00227.2014. PMID: 24906913.

Pasutto F., Keller K.E., Weisschuh N., Sticht H., Samples J.R., Yang Y.F., Zenkel M., Schlotzer-Schrehardt U., Mardin C.Y., Frezzotti P., Edmunds B., Kramer P.L., Gramer E., Reis A., Acott T.S., Wirtz M.K. 2012. Variants in ASB10 are associated with open-angle glaucoma. Hum Mol Gen. 21(6):1336-49. DOI: 10.1093/hmg/ddr572 PMID:22156576.

Pertea, M., Kim, D., Leek, J.T., Salzberg. Transcript-level Expression Analysis of RNA-seq Experiments with HISAT, StringTie and Ballgown. 2016. Nature Protocols. DOI:10.1038/nprot.2016.095.

Pertea, M., Pertea, G.M., Antonescu, C.M., T.C., Chang, Mendell, J.T., Salzberg, S. StringTie Enables Improved Reconstruction of a Transcriptome from RNA-seq Reads. 2015. 33:290-295. Nature Biotchnology. DOI:10.1038/nbt.3122.

Perez, C.F., Lopez, J.R., Allen, P.D. Expression Levels of RyR1 and RyR3 Control Resting Free Ca2+ in Skeletal Muscle. Am J Physiol Cell Physiol. 2005. 288(3):C640-9.

Petroski, M.D., Deshaies, R.J. 2005. Function and Regulation of cullin-RING E3 Ligases. Nature Reviews Molecular Cell Biology. 6(1):9-21. PMID 15688063

Pickart, C.M., Eddins, M.J. Ubiquitin: structures, functions, mechanisms. Volume 1695: Issues 1-3. 55-72. 2004. Biochimica et Biophysica Acta (BBA) – Molecular Cell Research. 1695(1-3):55-72. DOI:10.1016/j.bbamcr.2004.09.019. PMID:15571809.

Pietras, R.J., Nemere, I., Szego, C.M. "Steroid hormone receptors in target cell membranes." Endocrine. 2001. Apr;14(3):417-27.

Pirog, M., Gizak, A., Rakus, D. Changes in quaternary structure of muscle fructose-1,6-bisphosphatase regulate affinity of the enzyme to mitochondria. 2014. Int J Biochem Cell Biol. 48:55-99. Doi:10.1016/j.biocel.2013.12.015. PMID: 24412565.

Pollak, N., Dolle, C., Ziegler, M. The Power to Reduce: Pyridine Nucleotides – small Molecules with a Multitude of Functions. 2007. Biochem J. 402(2):205-18. DOI:10.1042/BJ20061638. PubMed PMID: 17295611.

Pritchett, E.M., Schmidt, C.J., Lamont, S.J. Transcriptomic Changes Throughout Post-Hatch Development in Gallus gallus Pituitary. 2016. J of Molecular Endocrinology. DOI: 10.1530/JME-16-0186.

Pritlove, D.C., Gu, M., Boyd, C.A., Randeva, H.S., Vatish, M. 2006. Novel Placental Expression of 2,3-bisphosphogylceart mutase. Placenta. 27(8): 924-7. DOI:10.1016/j.placenta.2005.08.010. PMID: 1624616.

Qu, Q., Mao, Y., Xiao, G., Fei, X., Wang, J., Zhang, Y., Liu, J., Cheng, G., Chen, X., Wang, J., Shen K. 2015. USP2 promotes cell migration and invasion in triple negative breast cancer cell lines. Tumour Biol. 36(7):5415-23. DOI: 10.1007/s13277-015-3207-7. PMID: 25687182.

Raboune, S., Stuart, J., Leishman, E., Takacs, S.M., Rhodes, B., Basnet, A., Jameyfield, Mchugh, D., Widlanski, T., Bradshaw, H.B. "Novel endogenous *N*-acyl amides activate TRPV1-4 receptors, BV-2 microglia, and are regulated in brain in an acute model of inflammation". Front. Cell. Neurosci., 01 August 2014 |https://doi.org/10.3389/fncel.2014.00195

Raz, V., Buijze, H., Verwey, N., Anvar, S.Y., Aartsma-Rus, A., van der Maarel, S.M. 2014. A novel feed-forward loop between ARIH2 E3-ligase and PABPN1 regulates aging-associated muscle degeneration. Am J Pathol. Apr. DOI: 10.1016/j.ajpath.2013.12.011 PMID: 24486325.

Rezvani, K., Teng, Y., Pan, Y., Dani, J.A., Lindstrom, J., Garca Gras, E.A., McIntosh, J.M., De Biasi, M. 2009. UBXD4, a UBX containing protein, regulates the cell surface number and the stability of α3-containing nicotinic acetylcholine receptors. Journal of Neuroscience. 29(21):6883-6896. DOI: https://doi.org/10.1523/JNEUROSCI.4723-08.2009.

Ripps, H., Shen, W. "Review: Taurine: A 'very essential' amino acid" Molecular Vision. 2012; 18: 2673-2686.

Riuzzi, F., Sorci, G., Sagheddu, R., Sidoni, A., Alaggio, R., Ninfo, V., Donato, R. 2014. RAGE signaling deficiency in rhabdomyosarcoma cells causes upregulation of PAX7 and uncontrolled proliferation. J Cell Sci. 127(8):1699-711. DOI: 10.1242/jcs.136259.

Roden, J.C., King, B.W., Trout, D., Mortazavi, A., Wold, B.J., Hart, C.E. Mining Gene Expression Data by Interpreting Principal Components. 2006. BMC Bioinformatics. 7:194. DOI:https://doi.org/10.1186/1741-2105-7-194.

Russo, F., Angelini, C. RNASeqGUI: a GUI for analyzing RNA-seq data. 2014. Bioinformatics 30(17):2514-2516. Doi: 10.1093/bioinformatics/btu308

Saghatelian, A.,McKinney, M.K.,Bandell, M. A FAAH-regulated class of N-acyl taurines that activates TRP ion channels. 2006. Biochemistry. 45(30):9007-15. DOI:10.1021/bi0608008. PMID:16866345.

Sahin, K., C. Orhan, M. O. Smith, and N. Sahin. Molecular targets of dietary phytochemicals for alleviation of heat stress in poultry. 2013. World's Poultry Science Journal, 69: 113-123.

Samad, A.F., Suliman, B.A., Basha, S.H., Manivasagam, T., Essa, M.M A Comprehensive In Silico Analysis on the Structural and Functional Impact of SNPs in the Congenital Heart Defects Associated with NKX2-5 Gene-A Molecular Dynamic Simulation Approach. 2016. PLoS One. 11(5):e0153999. DOI: 10.1371/journal.pone.0153999.

Samant, R.S., Clarke, P.A., Workman, P. 2014. E3 ubiquitin ligase Cullin-5 modulates multiple molecular and cellular responses to heat shock protein 90 inhibition in human cancer cells. Proc Natl Acad Sci U S A. 111(18):6834-9. DOI: 10.1073/pnas.1322412111. PMID: 24760825.

Santulli, G., Marks, A. Essential Roles of Intracellular Calcium Release Channels in Muscle, Brain, Metabolism and Aging. 2015. Current Molecular Pharmacology. 8(2):206-222. DOI: 10.2174/1874467208666150507105105. PMID:25966694.

Schmidt, Carl. (2015). RNA-seq:Primary Cells, Cell Lines and Heat Stress. Cytogenetic and Genome Research. 145. 145. 10.1159/00430927.

Sewduth, R.N., Jaspard-Vinassa, B., Peghaire, C., Guillabert, A., Franzi, N., Larrieu-Lahargue, F., Moreau, C., Fruttiger, M., Dufourcq, P., Couffinhal, T., Duplaa, C. The Ubiquitin ligase PDZRN3 is required for vascular morphogenesis through Wnt/planar cell polarity signaling. 2014. Nature Communications. 5: 4832. DOI:10.1038/ncomms5832 PMID:25198863.

Shafqat, N., Kavanagh, K.L., Sass, J.O., Christensen, E., Fukao, T., Lee, W.H., Oppermann, U., Yue, W.W. 2013. A Structural Mapping of Mutations Causing Succinyl-CoA:3-ketoacid CoA transferase (SCO) deficiency. Journal of Inherited Metabolic Disease. 36(6):983-7. DOI:10.1007/s10545-013-9589-z. PMID: 23420214.

Shao, W., Zumer, K., Fujinaga, K., Peterlin, B.M. 2016. FBXO3 Protein Promotes Ubiquitylation and Transcription Activity of AIRE (Autoimmune Regulator). 2016. J Biol Chem. 291(34):17953-63. DOI: 10.1074/jbc.M116.724401. PMID: 27365398

Simons, K., Ikonen, E. Functional Rafts in Cell Membranes. 1997. Nature. 387(6633):569-72. DOI: 10.1038/42408. PubMed PMID:9177342.

Smith, L.R., Meyer, G., Lieber, R.L. 2013. Systems analysis of biological networks in skeletal muscle function. Wiley Interdiscip Rev Syst Biol Med. 5(1):55-71. DOI: 10.1002/wsbm.1197. PMID: 23188744.

Sorrentino, V., Giannini, G., Malzac, P.,Mattei, M.G. Localization of a novel ryanodine receptor gene (RYR3) to human chromosome 15q14-q15 by in situ hybridization. Genomics. 18(1):163-5. DOI:10.1006/geno.1993.1446. PMID:8276408.

Spitz, F., Furlong, E.M. Transcription factors: from enhance binding to developmental control. Nature Reviews Genetics 13, 613-626. DOI:10.1038/nrg3207. PMID:22868264.

Strakova, J., Gupta, S., Kruger, W.D., Dilger, R.N., Tryon, K., Li, L., Garrow, T.A. Inhibition of betaine-homocysteine S-methyltransferase in rats causes hyperhomocysteinemia and reduces liver cystathione B-synthase activity and methylation activity. 2011. Nutr Res. 31(7):563-571 DOI: 10.1016/j.nutres.2011.06.004. PMCID: PMC315641.

Tallentire, C.W., Leinonen., I., Kyriazakis, I. Breeding for Efficieincy in the Broiler Chicken: A Review. 2016. 36:66. Agronomy for Sustainable Development.

Tarze, A., Deniaud A., Le Bras, M., Maillier, E., Molle, D., Larochette, N., Zamzami, N., Jan, G., Kroemer, G., Brenner, C. 2007. GAPDH, a novel regulator of the pro-apoptotic mitochondrial membrane permeabilization. Oncogene. 26(18): 2606–20. DOI:10.1038/sj.onc.1210074. PMID:17072346.

Tickle, P.G., Paxton, H., Rankin, J.W., Hutchinson, J.R., Codd, J.R. Anatomical and Biochemical Traits of Broiler Chickens across Ontogeny. Part I. Anatomy of the Muscoloskeletal Respiratory Apparatus and Changes in Organ Size. 2014. Peer J. doi: 10.7717/peerj.432. PMCID:PMC4103091.

Torok, Z., Crul, T., Maresca, B., Schutz, G.J., Viana, F., Dindia, L., Piotto, S., Brameshuber, M., Balogh, G., Peter, M., Porta, A., Trapani,A., Gombos, I., Glatz, A., Gungor, B., Peksel, B., Vigh, L., Csoboz, Z., Horvath, I., Vijahan, M., Hooper, P., Harwood, J.L., Vigh, L. Plasma membranes as heat stress sensors: From Lipid controll molecular switches to Therapeutic Applications. 2014. Biochim Biophys Acta. 1838(6):1594-618. DOI: 10.1016/j.bbamem.2013.12.015. PubMed PMID:24374314.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H.R., Salzberg, S.L., Rinn, J.L., Pachter, L. Differential Gene and Transcript Expression Analysis of RNA-seq Experiments with TopHat and Cufflinks. Nature Protocols. 2012. 7:562-578. Doi: 10.1038/nprot.2012.016.

Tripathi, R., Chakraborty, P., Varadwaj, P.K. Unraveling long non-coding RNAs through analysis of high-throughput RNA-sequencing data. 2017. Non-Coding RNA Research. DOI: https://doi.org/10.1016/j.ncma.2017.06.003.

Ustanina, S., Carvajal, J., Rigby, P., Braun, T. 2007. The myogenic factor Myf5 supports efficient skeletal muscle regeneration by enabling transient myoblast amplification. Stem Cells. DOI:10.1634/stemcells.2006-0736. PMID:17495111.

Valvona, C.J., Fillmore, H.L., Nunn, P.B., Pilkington, G.J. 2015. The Regulation and Function of Lactate Dehydrogenase A: Therapeutic Potential in Brain Tumor. Brain Pathol. 26(1):3-17. DOI: 10.1111/bpa.12299. PMID:26269128.

Van Der Veen, J.N., Lingrell, S.,Vance, D.E. The Membrane Lipid Phosphatidylcholine Is an Unexpected Source of Triacylglycerol in the Liver. 2012. Journal of Biological Chemistry. DOI: 10.1074/jbc.M112.381723. PMID: 22610093.

Van der Zwaag, B., Burbach, J.P., Scharfe, C., Oefner, PJ., Brunner, H.G., Padberg, G.W., Van Bokhoven, H. Identifying new candidate genes for hereditary facial paresis on chromosome 3q21-q22 by RNA in situ hybridization in mouse. 2005. Genomics. 86(1):55-67. DOI:10.1016/j.ygeno.2005.03.007. PMID 15953540.

Van Goor, A., Bolek, K.J., Ashwell, C.M., Persia, M.E., Rothschild, M.F., Schmidt, C.J., Lamont, S.J. Identification of Quantitative Trait Loci For Body Temperature, Body Weight, Breast Yield, and Digestibility in an Advanced Intercross Line of Chickens Under Heat Stress. 2015. Genetics Selection Evolution. 47:96. DOI: https://doi.org/10.1186/s12711-015-0176-7.

Verghese,J., Abrams, J., Wang, Y., Morano, K.A. Biology of the Heat Shock Response and Protein Chaperones: Budding Yeast (*Saccharomyces cerevisiae)* as a Model System. 2012. DOI: 10.1128/MMBR.05018-11 Microbiol. Mol. Biol. Rev. 76(2):115-1581.

Vigh, L., Horvath, I., Maresca, B., Harwood, J.L. Can the stress protein response be controlled by 'membrane-lipid' Therapy. 2007. Trends in Biochemical Sciences. 32(8):357-363. DOI: 10.1016/j.tibs.2007.06.009. PMID:17629486.

Vigh, L., Maresca, B., Harwood, J.L. Does the Membrane's Physical State Control the Expression of Heat Shock and other Genes? 1998. Trends Biochem Sci. 23(10):369-74. PubMed PMID:9810221.

Vos, M.J., Kanon, B., Kampinga, H.H. HSPB7 is a SC35 Speckle Resident Small HeatShock Protein. 2009. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research. 1793(8):1343-1353. DOI: https://doi.org/10.1016/j.bbamcr.2009.05.005.

Wang, Q., Wang, Z., Tian, X., Tang, R., Xu, X.Four and a half LIM domains 2 contributes to the development of human tongue squamous cell carcinoma. J Mol Histol. 2016 Apr;47(2):105-16. doi: 10.1007/s10735-016-9654-7. Epub 2016 Jan 12.

Wang, X., Lin, H., Gu, Y. "Multiple Roles of dihomo-y-linoleic acid against proliferation diseases" Lipids Health Dis. 2012:11:25. Doi:10.1186/1476-511X-11-25.

Ward, C.W., Rodney, G.G. Does a lack of RyR3 make mammalian skeletal muscle EC coupling a 'spark-less' affair. 2008. J Physiol. 586(Pt 2): 313-314. DOI: 10.113/jphysiol.2007.148643. PMCID: PMC2375586.

Watanabe, T. K., Kawai, A., Fujiwara, T., Maekawa, H., Hirai, Y., Nakamura, Y., Takahashi, E. Molecular cloning of UBE2G, encoding a human skeletal muscle-specific ubiquitin-conjugating enzyme homologous to UBC7 of C. elegans. 1996. Cytogenet. Cell Genet. 74: 146-148, 1996. PMID:8893823. DOI:10.1159/000134403.

Watson, C.M., Crinnion, L.A., Murphy, H., Newbould, M., Harrison, S.M., Lascelles, C., Antanaviciute, A., Carr, I.M., Sheridan, E, Bonthron, D.T., Smith, A. Deficiency of the myogenic factor MyoD causes a perinatally lethal fetal akinesia. 2016. J Med Genet. 53(4):264-9. DOI: 10.1136/jmedgenet-2015-103620.

Wells, A., Kopp, N., Xioxiao, X., O'Brien, D.R., Yang, W., Nehorai, A., Adair-Kirk, T.L., Kopan, R., Dougherty, J.D. 2015. The anatomical distribution of genetic associations. Nucleic Acids Research. 43:22.

Yates, A., Beal, K., Keenan, S., McLaren, W., Pignatelli, M., Richie, G.R.S.,Ruffier, M., Taylor, K., Vullo, A., Flicek, P. The Ensembl REST API: Ensembl Data for Any Language. 2015. Bioinformatics. 31(1):143-145. DOI: https://doi.org/10.1093/bioinformatics/btu613.

Yu, W., Li, Y., Zhou, X., Deng, Y., Wang, Z., Yuan, W., Li, D., Zhu, C., Zhao, X., X., M., Huang, W., Luo, N., Yan, Y., Ocorr, K., Bodmer, R., Wang, Y., Wu, X. 2008. A novel human BTB-kelch protein KLHL31, strongly expressed in muscle and heart, inhibits transcriptional activities of TRE and SRE. Mol Cells. PMID: 18719355

Yuan, L., Liu, S., Bai, X., Gao, Y., Liu, G., Wang, X., Liu, D., Li,T., Hao, A., Zhen, W., Oxytocin Inhibits lipopolysaccharide-induced inflammation in microglial cells and attenuates microglial activation in lipopolysaccharide-treat mice. 2016. Journal of Neuroinflammation. 13:77.

Zhao, D., Shen, J.P., Sasik, R., Ideker, T., Mali, P, Lab, I. Combinatorial Crispr-Cas9 Knockout Screen. 2017. Nature Protocol Exchange. DOI:10.1038/protex.2017.063

Zucchi, R., Ronca-Testoni, S., The Sarcoplasmic Reticulum Ca^{2+} Channel/Ryanodine Reception: Modulation by Endogenous Effectors, Drugs, and Disease States. Pharmacological Reviews. 49(1):1-51. PMID:9085308.

I

PCA TABLES

Table A1: Significant correlations for the top 30 biomarkers in cluster 1 with PC1

Compound	Correlation with PC	P Value of Correlation
1_stearoyl_2_arachidonoyl_GPE__18_0_20_4_	0.973411108	8.17E-19
1_stearoyl_2_arachidonoyl_GPC__18_0_20_4_	0.949414773	4.18E-15
1__6soUnique_enyl_stearoyl_	0.947359337	7.07E-15
1_stearoyl_2_arachidonoyl_GPI__18_0_20_4_	0.942057468	2.50E-14
1_arachidonoyl_GPC__20_4n6_	0.929897571	3.05E-13
arachidonate__20_4n6_	0.899789752	3.16E-11
sphingomyelin__d18_1_24_1__d18_2_24_0_	0.891821542	8.45E-11
1_2_dipalmitoyl_GPC__16_0_16_0_	0.808014535	1.16E-07
docosahexaenoate__DHA;_22_6n3_	0.798469517	2.10E-07
sphingomyelin__d18_2_24_1__d18_1_24_2_	0.789957155	3.47E-07
bilirubin__Z_Z_	0.725230729	8.56E-06
taurine	0.705899307	1.89E-05
tartronate__hydroxymalonate_	0.61912611	0.000342551
betaine_aldehyde	0.576008864	0.001075968
arachidate__20_0_	0.563029256	0.001473619
acetylcarnitine	0.380762315	0.041573998
stearoyl_ethanolamide	-0.559317412	0.00160854
1_palmitoyl_2_linoleoyl_glycerol__16_0_18__	-0.670628916	6.86E-05

Table A2: Significant correlations for the top 30 biomarkers in cluster 1 with PC2.

Compound	Correlation with PC	P Value of Correlation
N_acetyltaurine	0.954472146	1.04E-15
N_stearoyltaurine	0.901295841	2.60E-11
N_palmitoyltaurine	0.8608777	2.09E-09
1_arachidonoyl_GPE__20_4n6_	0.756796066	2.03E-06
tartronate__hydroxymalonate_	0.692836171	3.11E-05
linoleate__18_2n6_	0.593372295	0.000691859
acetylcarnitine	0.541858196	0.002396802
docosahexaenoate__DHA;_22_6n3_	0.510264184	0.004683614
taurine	0.407260431	0.028325624
stearoyl_ethanolamide	-0.368251095	0.04934983
sphingomyelin_d18_2_24_1_d18_1_242	-0.37425016	0.045489132
Oleoylcarnitine	-0.42348255	0.022071949
betaine_aldehyde	-0.510989226	0.00461539
1_palmitoyl_2_stearoyl_GPC_16_0_18_0	-0.513650003	0.004372239
adipoylcarnitine	-0.569611132	0.001258411
beta_guanidinopropanoate	-0.618823499	0.000345515

Table A3: Significant correlations for the top 30 biomarkers.

Compound	Correlation with PC	P Value of Correlation
linoleate__18_2n6_	0.737139185	5.09E-06
adipoylcarnitine	0.720754209	1.03E-05
margarate__17_0_	0.715830593	1.27E-05
stearoyl_ethanolamide	0.667060782	7.75E-05
1_palmitoyl_2_linoleoyl_glycerol__16_0_18_2_	0.629414484	0.000254203
beta_guanidinopropanoate	0.618738902	0.000346348
betaine_aldehyde	0.608525643	0.000460849
1_palmitoyl_2_stearoyl_GPC__16_0_18_0_	0.592667967	0.000704707
1_arachidonoyl_GPE__20_4n6_	0.513716944	0.004366265
acetylcarnitine	0.509219436	0.004783433
oleoylcarnitine	0.454866424	0.013168719

Table A4: Significant correlations for the top 30 biomarkers in cluster 2 with PC1

Compound	Correlation with PC	P Value of Correlation
sphingomyelin_1*	0.935398433	1.04E-13
cholesterol	0.932829799	1.74E-13
linoleoylcarnitine	0.931565767	2.22E-13
hypotaurine	0.916576738	2.94E-12
sphingomyelin_2*	0.908244464	1.01E-11
sphingomyelin_3*	0.896807074	4.61E-11
Gene_SLC6A13	0.885360942	1.78E-10
stearoylcarnitine	0.882086144	2.55E-10
behenoyl_sphingomyelin__d18_1_22_0_	0.881483363	2.72E-10
N_formylmethionine	0.844331641	8.58E-09
Gene_LOC424748	0.83462041	1.83E-08
dehydroascorbate	0.829149409	2.74E-08
propionylcarnitine	0.803345422	1.55E-07
Gene_FGG	0.797405686	2.24E-07
Gene_ITIH3	0.770180345	1.03E-06
picolinate	0.755204507	2.19E-06
Gene_CTSO	0.750255405	2.78E-06
biopterin	0.748358136	3.04E-06
1_stearoyl_GPG__18_0_	0.724755853	8.74E-06
creatinine	0.702867662	2.12E-05
Gene_LOC101748084	0.688118298	3.70E-05
Gene_LOC417848	0.421609557	0.022730276
Gene_LOC101748827	-0.426372363	0.021086958
thiamin__Vitamin_B1_	-0.720611773	1.04E-05
Gene_LIPC	-0.799094603	2.02E-07
Gene_C6	-0.80811097	1.15E-07
argininosuccinate	-0.829406255	2.69E-08
Gene_HPD	-0.892524003	7.77E-11

1*: d18_1_21_0__d17_1_22_0_d16_1_23_0_
2*: d18_1_20_0__d16_1_22_0_
3*: d18_1_22_1__d18_2_22_0__d16_1_24_1_

Table A5: Significant correlations for the top 30 biomarkers in cluster 2 with PC2.

Compound	Correlation with PC	P Value of Correlation
2_hydroxyphenylacetate	0.858128947	2.67E-09
1_stearoyl_GPG__18_0_	0.547061713	0.002132962
Gene_LIPC	0.511422315	0.004575044
Gene_C6	0.379715473	0.042184541
picolinate	0.372804189	0.046396969
stearoylcarnitine	0.372439684	0.046628074
Gene_LOC101748084	0.370803858	0.047676496
Gene_ITIH3	-0.422671653	0.022355021
Gene_LOC417848	-0.431339147	0.019478493
Gene_CTSO	-0.507217221	0.004979808
Gene_FGG	-0.52952873	0.003136762

Table A6: Significant correlations for the top 30 biomarkers in cluster 2 with PC3.

Compound	Correlation with PC	P Value of Correlation
Gene_LOC101748827	0.764749525	1.36E-06
Gene_LOC417848	0.697169244	2.64E-05
creatinine	-0.370667834	0.04776451

Table A7: Significant correlations for the top 30 biomarkers in cluster 3 with PC1.

Compound	Correlation with PC	P Value of Correlation
glucosamine_6_phosphate	0.908333066	9.99E-12
glucose_6_phosphate	0.896905228	4.55E-11
1_palmitoyl_2_oleoyl_GPE__16_0_18_1_	0.88275415	2.37E-10
1_palmitoyl_2_linoleoyl_GPE__16_0_18_2_	0.876784526	4.47E-10
1_palmitoyl_2_palmitoleoyl_GPC__16_0_16_1	0.872140482	7.15E-10
pterin	0.842162231	1.02E-08
1_palmitoyl_2_oleoyl_GPI__16_0_18_1_	0.836085477	1.64E-08
fructose_6_phosphate	0.819623864	5.37E-08
N6_succinyladenosine	0.799891815	1.92E-07
myristoleate__14_1n5_	0.79219482	3.05E-07
1_palmitoyl_GPE__16_0_	0.789460668	3.57E-07
1_palmitoleoyl_3_oleoyl_glycerol_16_1_18_1_	0.769924732	1.04E-06
glycerol_3_phosphate	0.727885235	7.64E-06
glutathione__reduced__GSH_	0.693131484	3.07E-05
1_stearoyl_2_linoleoyl_GPE__18_0_18_2_	0.655701187	0.000112805
cysteinylglycine	0.642033649	0.00017373
1_palmitoyl_2_linoleoyl_GPS__16_0_18_2_	0.631028195	0.000242351
Gene_S100Z	0.53484916	0.002796371
gamma_glutamylcysteine	0.493156219	0.006562248
1_stearoyl_2_linoleoyl_GPI__18_0_18_2_	0.48161764	0.008161479
Gene_NADKD1	0.419911845	0.023340809
phosphopantetheine	-0.45438759	0.013277576

Table A8: Significant correlations for the top 30 biomarkers in cluster 3 with PC2.

Compound	Correlation With PC	P Value of Correlation
glycerophosphoethanolamine	0.926562169	5.59E-13
UDP_glucuronate	0.896634062	4.71E-11
N_acetylglucosaminylasparagine	0.826979934	3.20E-08
adenosine	0.797166015	2.27E-07
gamma_glutamylcysteine	0.723775342	9.11E-06
cysteinylglycine	0.658985686	0.000101361
glutathione__reduced__GSH_	0.557487932	0.001678892
3__dephosphocoenzyme_A	0.516101315	0.004157998
adenosine_5__monophosphate__AMP_	0.448790444	0.014606585
glycerol_3_phosphate	0.382077553	0.04081694
phosphopantetheine	0.380002726	0.042016301
1_palmitoyl_2_linoleoyl_GPS__16_0_18_2_	-0.379979627	0.04202981
Gene_NADKD1	-0.55217634	0.001898469
1_stearoyl_2_linoleoyl_GPI__18_0_18_2_	-0.710759687	1.56E-05
Gene_S100Z	-0.741713038	4.14E-06

Table A9: Significant correlations for the top 30 biomarkers in cluster 3 with PC3.

Compound	Correlation with PC	P Value of Correlation
coenzyme_A	0.816408535	6.68E-08
3__dephosphocoenzyme_A	0.758400863	1.87E-06
phosphopantetheine	0.661404476	9.36E-05
Gene_NADKD1	0.580619473	0.000959209
1_stearoyl_2_linoleoyl_GPE_18_0_18_2	0.567883041	0.001312083
adenosine	0.415577597	0.024960561