Inferring models and structure from biological data: networks and pathways

By

MATTHEW PUTNINS

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Biomedical Engineering and

Computational Biology and Molecular Biophysics

Written under the direction of

Ioannis P. Androulakis

And approved by

_____

_____

_____

_____

_____

New Brunswick, New Jersey

October 2020

ABSTRACT OF THE DISSERTATION

Boolean Networks in Systems Biology

By Matthew Putnins

Dissertation Director:
Ioannis P. Androulakis

Physiological functions are driven by the emergent behaviors of many individual components, whether they are gene, protein, or metabolic interactions. These interactions form biochemical pathways and interaction networks which then lead to more complex cellular or organismal level behaviors that are not knowable from the characteristics of an individual component of that system. Knowing whether a gene is being expressed or knowing the structure of a protein does not necessarily imply the physiological function of either, but within the context of a meaningful biological system, we can infer more complex behaviors. In the enclosed dissertation we present multiple approaches to contextualize biological components into more complex systems. These methods include utilizing Boolean networks to model interactions in a qualitative manner, as well as analyzing expression data in the context of biochemical pathways. We use two distinct approaches for understanding biological systems: We utilize evolutionary algorithms to understand the origin and development of complex systems. This evolutionary framework enables a better understanding of complex network structures as well as evolutionary strategies used in the development of complex biological systems.

We additionally propose a data-driven approach for interrogating gene expression within the context of biochemical pathways. We utilize a novel method for detecting circadian genes and map these genes onto physiologically functional pathways. We utilize this data

to validate methods for constructing a Boolean network to infer the causal relationships which exist within gene pathways. This analysis will improve the applications of high throughput data analysis for the purpose of identifying critical components of complex biological systems.

Acknowledgments

I would like to thank everyone who has supported my efforts through the years and helped me achieve this work. I thank my advisor, Dr. Ioannis Androulakis, and my committee, Dr. David Axelrod, Dr. Li Cai, and Dr. Troy Shinbrot for their assistance in guiding and supporting this work. I thank my wife Alyssa Putnins and my immediate family: Zigi Putnins, Cathy Benson, David Putnins, and Susan Putnins for all the care and patience they've shown me during my journey. I thank my current and former lab mates Alison Acevedo, Rohit Rao, Seul-A Bae, and Jordan Eckhoff for their assistance and camaraderie. I am grateful to the Biomedical Engineering Administrative staff Lawrence Stromberg, Robin Yarborough, Stratos Loukidis, and Linda Johnson for their assistance in navigating my academic life at Rutgers

Table of Contents

List of Tables

List of Illustrations

CHAPTER 1: Background and Motivation

1.1 Systems biology and biological networks

Systems biology represents a philosophy of approaching complex biological problems, often consisting of many different components, from a holistic view. A reductionist view, the opposite philosophical approach, sees a system as being described by the individual behaviors of its components. This includes structural biological approaches, such as designing drug therapies based on protein structure. Deriving pharmaceutical therapies directly from protein structures has proven challenging, however (Van Regenmortel 2001). This limited view of drug development can be overcome by utilizing more systems based methodologies to understand the complex behaviors of biochemical pathways(Regenmortel 2004; Materi and Wishart 2007; Kitano 2002).

Systems biology approaches are often applied to biochemical pathways or genetic and metabolic systems, but may be applicable to more complex population-level dynamics as well(Friedman and Gore 2017; Kitano 2002). The primary focus is on the emergent behaviors of the complex system, rather than the specific properties or behaviors of individual components(Kauffman 2007). In this sense, the behavior of specific components is assumed to not be descriptive of the system as a whole, such as the case of pharmaceuticals designed based on the peptide structure of a protein, and it is more important to understand the relationships and interactions between these components.

The relationships between these components and their emergent properties and behaviors can be studied at various levels of detail. This can be studying discrete time points and qualitative values, such as defining components as active or inactive or using more

refined continuous value models, using a variety of computational techniques which can range from systems of ordinary differential equations to more qualitative, coarse-grained models based on Petri nets, cellular automata, or Boolean networks(Materi and Wishart 2007).

1.2 Boolean Networks

Boolean Networks are simple switching models in which variables are qualitatively defined as being “active” or “inactive”, using Boolean logic to relate individual components. The accuracy of these qualitative descriptors compared to continuous biochemical models has been discussed since Boolean models have been introduced (Kauffman 1969). Kauffman and Glass used continuous models to study biochemical networks based on the activation, inhibition, and decay of molecular species. Within these models, individual species would tend to transition between low and high values and saw these oscillations as a switching behavior that could be described in a binary manner by fitting (Glass and Kauffman 1972). Based on this binary behavior, discrete logical homologs of these models were developed in a way that was representative of the biochemical pathway’s behavior (Glass and Kauffman 1973). Classic, discrete Boolean models have a series of variables that represent nodes N defined by a state X, each with k incoming edges. Although some models restrict k to specific values, they do not necessarily have to be the same for each node of the network. Each variable’s value is defined by a Boolean function (b) as shown in Equation 1. Different models may allow or restrict which Boolean functions are used, depending on what the model will represent. For example, b could be represented by any arbitrary truth table, or it could be limited to

and/or/not gates in order to represent relationships observed in biological regulatory networks(Terfve et al. 2012).

$$X_i(t) = B_i(X_{1i}(t-1), X_{2i}(t-1) \ldots X_{ki}(t-1))$$

Eq. 1

In this case, the state of N_i is determined by the Boolean function B_i which relates all the inputs to the node. In this case, the state X_i can either take on a value of 1 or 0 depending on the output of its relevant Boolean function.

1.3 Identifying critical components of Boolean networks

One of the advantages of Boolean networks is the qualitative nature which allows for flexible structural and functional assessments of the networks. Structural analysis often consists of reducing or removing relatively simple components, such as nodes with a single input and output (Veliz-Cuba 2011). These types of reductions can reduce the amount of data necessary for more detailed models, as well as identify the core functional components of the network. The result of this structural reduction can be used to synthesize a smaller logic network whose function is still reflective of the original complete network(Chudasama et al. 2015). This can allow for complex regulatory pathways to be reduced to a smaller size which is far easier to understand from a qualitative viewpoint and can either be used to produce more detailed networks of the important components or to directly identify potential drug targets.

However, this view of reducing networks through primarily structural means is highly limited, because perturbations to the networks may create unexpected changes in larger, more complex networks. Feedback loops, feedforward loops, and other complex network structures may enhance or diminish small perturbations and may lead to greater or

smaller changes in distant parts of a larger network. Because of this, structural reductions, such as that described above, may hide functionally important network motifs.

Perturbation studies of Boolean networks are often important for identifying drug targets, which is often overlooked in structural based, steady-state reduction described above(Azuaje, Devaux, and Wagner 2010). Even structural reduction methods that are designed to preserve network dynamics may have issues with different network states being reachable based on perturbations, which may have been possible in the unreduced model(Naldi et al. 2009).

1.4 Evolution of biological networks

The complex genetic, regulatory, and metabolic networks which are ubiquitous throughout all domains of life have evolved over time. The evolution of organisms, and their genetic networks, is controlled by a variety of factors which are fundamentally broken into adaptive forces, such as natural selection, and non-adaptive forces, such as random mutation and genetic drift. The interactions between these forces have led to the emergence of complex genetic and protein interaction networks across in all domains of life. Network motifs and small-world network features have been found in many of these pathways (Wuchty, Oltvai, and Barabasi 2003; Dwight Kuo, Banzhaf, and Leier 2006; Alon 2007; Luscombe et al. 2004; Shen-Orr et al. 2002; Van Noort, Snel, and Huynen 2004). These features are found across domains of life, from yeast to bacteria to more complex organisms such as plants and animals (Defoort, Van de Peer, and Vermeirssen 2018; Shen-Orr et al. 2002; Van Noort, Snel, and Huynen 2004).

All forces of evolution, from selective pressure to neutral mutations to genetic drift, influence to formations and differentiation of species(Raeymaekers et al. 2017). Even in the absence of adaptive forces such as natural selection non-adaptive forces of random mutation and genetic drift lead to new species(Lynch 2007; Yu and Miller 2001).

Both natural selection, as well as neutral mutations and network growth, influence the structure and performance of a regulatory network. However, the effect of neutral mutations on network structure depends on the rate that mutations occur, as well as the size or rate of change of the size of the genome of an organism(Bhan, Galas, and Dewey 2002; Van Noort, Snel, and Huynen 2004).

The genome size and mutation rates vary widely between species (Zhang 2003; Gao and Innan 2004; Gu et al. 2002; Drake 1991; Lynch 2010), though there appears to be a relationship between genome size and base-pair mutation rate(Lynch 2010). This base-pair mutation rate is not static, however, and in the presence of lethal stressors it may rapidly change even while the genome size remains relatively static(Swings et al. 2017).

Computational models of evolution often are limited in terms of not including the evolution of functional networks (Van Noort, Snel, and Huynen 2004; Chung et al. 2003; Bhan, Galas, and Dewey 2002; Eisenberg and Levanon 2003) or artificially limit the non-adaptive forces on networks by using fixed mutation rates(Alon 2007; Kashtan and Alon 2005; Wilke et al. 2001) or fixed network size(Knabe, Nehaniv, and Schilstra 2008). Because of this, we propose an evolutionary model that incorporates both adaptive and non-adaptive forces and allows natural and emergent interactions between the two (Chapter 2). This model allows us to differentiate between the adaptive and non-adaptive forces of evolution and allows us to understand how these two distinct forces act in a co-

operative way to drive change within a population. We hypothesize that the non-adaptive forces, such as the mutation rate within a population, are driven by selective pressure much the same way that a phenotype with fitness advantages or disadvantages is driven.

1.5 Genetic pathways and networks in mouse

Evolutionary algorithms and models help us understand the origin of biological networks; however, they need to be compared to the networks and pathways of living organisms to validate the types of network structures we see. We utilize high throughput data microarray expression data to understand how real biological components may interact with each other and interrogate possible relationships between physiological systems. While learning how complex networks evolve over time may provide insight into the evolutionary origins of certain features, expression data allows us to identify those networks which are important in understanding the effects of these emergent interactions in a physiological context. The concept of systems biology includes contextualizing individual components into functionally related groups, something which is lacking in high throughput data studies (García-Campos, Espinal-Enríquez, and Hernández-Lemus). Physiological functions are not produced by individual components but from the emergent behavior of complex interaction networks of genetic, metabolic, and proteomic molecules. These networks are called pathways and are composed of genetic components that are coordinated in a way to produce a single physiological function. The individual genetic components may belong to multiple networks, but contextualizing them within a single pathway can still be a useful tool for understanding micro-array data in a way which allows us to determine how different genes may be related to one another through

both a functional and mechanistic way (García-Campos, Espinal-Enríquez, and Hernández-Lemus ; Jin et al. ; Amadoz et al.).

We make use of whole-genome, whole-organism mouse data (Dyar et al. 2018) within the context of genes and pathways identified within the KEGG database (Kanehisa and Goto 2000) in order to interrogate how different tissues may perform different or related physiological functions and to identify how functional pathways are expressed across different tissue types within an organism. We emphasize identifying genes that have a circadian expression pattern, because of the known link between circadian rhythms, health and disease(Bishehsari et al. 2016; Cutolo and Masi 2005; Doherty 2018; Kaczmarek, Thompson, and Holscher 2017; Lee and Edery 2008). We propose a novel approach to identifying circadian genes (Chapter 3) and use this to identify differentially expressed genes within functional pathways. We further identify and use a set of methods for identifying core pathways components using data from multiple tissues (Chapter 4).

1.6 Outline of the Dissertation

Understanding the context of biological components within a larger system is critical to understanding more complex, emergent properties of biological systems. It is important to understand data within the context of a complex network, but also to be able to interpret potential relationships between components from otherwise decontextualized data. In the following chapters, we propose three unique approaches to understanding complex biological networks. We present a model of the origins and evolution of biological regulatory systems, especially within to context of oscillatory networks (chapter 2). We also approach biological networks from a data-driven approach. Utilizing high throughput omics, we propose a method of detecting oscillatory genes within a large

dataset (chapter 3). Subsequently, we demonstrate methods for utilizing large data sets for the purpose of constructing Boolean models to simulate and study-specific pathways (chapter 4). This is particularly important because omics-level data may not have ideal data for the construction of such data but may be able to lay the framework of causal relationships between components. Our final chapter discusses the implications of larger data sets for the purpose of contextualizing information within the framework of pathways, models, and networks.

CHAPTER 2: An evolutionary model of learning networks

2.1 Background

Regulatory networks and signal transduction pathways are composed of series of biochemical interactions in order to interpret environmental factors and produce a response: When an organism eats, its body needs to produce insulin in order to regulate its blood sugar(Ohneda, Ee, and German 2000). This response is the result of complex networks that have evolved over time, either conserving components that have proven beneficial or mutating aspects which are detrimental or unnecessary as the environment in which the organism exists no longer requires it. Changes to these networks occur incrementally over many generations. Due to a lack of perfect fidelity when DNA is replicated, each generation will have a slightly different genetic makeup than their parents. This change in genetics is caused by random changes, such as point mutations, gene duplication or recombination (Lynch 2007; Knabe, Nehaniv, and Schilstra 2008; Ingram, Stumpf, and Stark 2006; Parter, Kashtan, and Alon 2008; Stelling et al. 2002; Bhan, Galas, and Dewey 2002). These random genetic changes may be neutral, deleterious, or beneficial and are referred to as a non-adaptive force because these changes occur regardless of whether they assist an individual to adapt to their environment. These changes affect the makeup of the regulatory networks within an individual influencing how signals are interpreted or how the network responds to stimuli. Individuals who receive beneficial mutations are more likely to produce offspring, while those with deleterious mutations are less likely to survive and reproduce. This bias in reproduction is referred to as an adaptive force, representing a population's adaption to its environment (Alon 2007; Kashtan and Alon 2005). These two forces,

important for evolution and speciation(Ghalambor et al. 2015; Ghalambor et al. 2007; Raeymaekers et al. 2017), are intrinsically linked: non-adaptive changes create the random variation which can then be selected for through adaptive forces, while adaptive selection may choose individuals with higher or lower mutation rates. Changes in the non-adaptive processes are seen as critical to overcoming stressors (Lavergne, Muenke, and Molofsky 2010; Swings et al. 2017; Komp Lindgren, Karlsson, and Hughes 2003). The flexibility of these non-random processes are also seen as important: bacteria with high mutation rates do not have better rates of survival in response to environmental stressors(Sprouffske et al. 2018), but when wild type bacteria are exposed to environmental stressors their mutation rates will rapidly increase(Komp Lindgren, Karlsson, and Hughes 2003; Lynch 2010; Swings et al. 2017).

Non-adaptive forces, such as the mutation rate, need to be accounted for when studying the effects of evolution. We propose that factors such as mutation rate may not directly affect an individual's fitness (and therefore cannot be directly selected for by environmental pressure), but are indirectly selected for as high mutation rate individuals may be able to overcome stressors and low mutation rate individuals may be less likely to pass on deleterious mutations to their offspring. These factors are not generally included in computational evolution studies of genetic networks.

Models that study the evolution of genetic networks exist in many forms. The simplest form is a model in which the network is evolved without any evolutionary pressure (i.e. no selection or fitness function) and with the network not having any function (Van Noort, Snel, and Huynen 2004; Chung et al. 2003; Bhan, Galas, and Dewey 2002; Eisenberg and Levanon 2003). These models assert a null hypothesis: The structure of

regulatory networks exists because of the rules that create them, not because of the selection of specific functions. These models are proposed as explanations for types of structures found universally within gene networks, such as the scale-free nature of some biological networks(Van Noort, Snel, and Huynen 2004). However, other evidence suggests that biological networks may not be scale-free (Stumpf, Wiuf, and May 2005; Han et al. 2004) and may not have small-world properties(Arita 2004), which indicates that these null-models are not universally applicable.

More advanced models look at functionalized networks, using search algorithms to identify networks that satisfy networks that perform a specific task. Functional network types vary widely and can include electronic logic circuits(Parter, Kashtan, and Alon 2008), neural networks(Kashtan and Alon 2005), and networks that exhibit oscillatory behavior(Knabe, Nehaniv, and Schilstra 2008; Burda et al. 2011). Most of these types of networks can create many topologies which perform the same function, and therefore limitations are set on the search space, which can include limiting the in-degree of nodes(Parter, Kashtan, and Alon 2008), network size (Kashtan and Alon 2005; Knabe, Nehaniv, and Schilstra 2008; Burda et al. 2011). In addition to the types of networks being used, the search algorithms used within these models can vary widely as well. These algorithms are not always evolutionary algorithms, and those which are suffer shortfalls by evolving networks using fixed mutation rates(Alon 2007; Kashtan and Alon 2005; Wilke et al. 2001) or fixed network size(Knabe, Nehaniv, and Schilstra 2008).

Based on our knowledge of biology, these artificial limitations may mask critical evolutionary events. Adaptions to genome size and mutation rates are seen in both plant and bacterial species during stressor events(Lavergne, Muenke, and Molofsky 2010;

Swings et al. 2017). Fixing these rates may miss how a dynamic network size or mutation rate affects the synthesis of a regulatory network or how the conservation of key functional components overrides the otherwise random evolutionary moves.

The primary focus of many of these network models, whether a proposed null model or a functionalized network model, is to identify the origin of regulatory networks motifs, such as feedback or feed-forward loops(Burda et al. 2011; Ingram, Stumpf, and Stark 2006; Knabe, Nehaniv, and Schilstra 2008; Milo et al. 2002). Biological studies cast doubt on the value of motifs. The regulatory network of the yeast *S. cerevisiae*, for example, has over-represented network motifs, but these motifs are limited to recently evolved genes(Meshi, Shlomi, and Ruppin 2007). The study found that these motifs did not correlate with functional enrichment within the network and that they were not evolutionarily conserved. This leads to the conclusion that even if the evolutionary moves which create a regulatory network have some structural bias, selective pressure tends to remove that bias, which underscores the importance of including both the underlying evolutionary mechanics (the non-adaptive steps of evolution) as well as functional selection when studying how biological networks evolve.

We propose a computational evolution model that utilizes Boolean networks to mimic environmental signal transduction, which includes both aspects. We attach mutation rates to individuals, rather than using fixed rates or population-wide-rates allowing for populations to self-select for greater or smaller rates of mutation. Herein we demonstrate the ability of evolution to select for non-adaptive traits, showing that the mutation rate of individuals within a population is tightly linked to the fitness of that population. We further demonstrate the emergence of structurally unique functional networks, indicating

that the structures which are conserved for being functionally important tend to have unique architecture

2.2 Approach

Herein we utilize a novel evolutionary model that captures the adaptive and non-adaptive forces on genetic regulatory networks in changing environmental conditions as seen in Figure 1. This evolution was done using a population of individuals, where each individual has its genetic regulatory network represented by a logic network. To more fully capture the non-adaptive forces, each individual has an additional mutation and duplication/deletion rate, allowing for different members of the population to mutate at different rates, representing mutations of the DNA replicase within an individual.

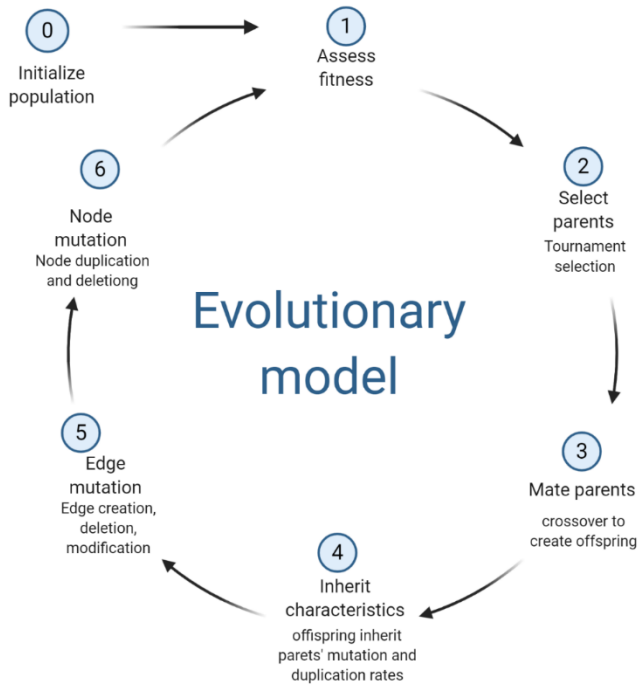

Figure 1: Representation of the evolutionary modell: After the population is initialized, individuals are consistently assessed for fitness before undergoing crossover and mutation. In addition, a duplication/deletion event occurs which allows for the introduction or removal of nodes from the network and then the mutation rate and duplication/deletion rate are modified randomly before the fitness of each individual is assessed again.

Every individual within the population receives an environmental signal, processes this signal within the gene regulatory network, and produces an output signal, as seen in Figure 2. This network is composed of a series of nodes that represent the environmental input, the network's output signal, and the regulatory components within the signaling network. Each edge represents a regulatory relationship between components of the network. Our evolutionary strategy is based on the basic principles of *Genetic Algorithms* (Holland 1975; Goldberg 1989) which encompass computational evolutionary strategies that have been extensively adapted for the evolution of regulatory networks (Noman et al. 2015; Knabe et al. 2010).

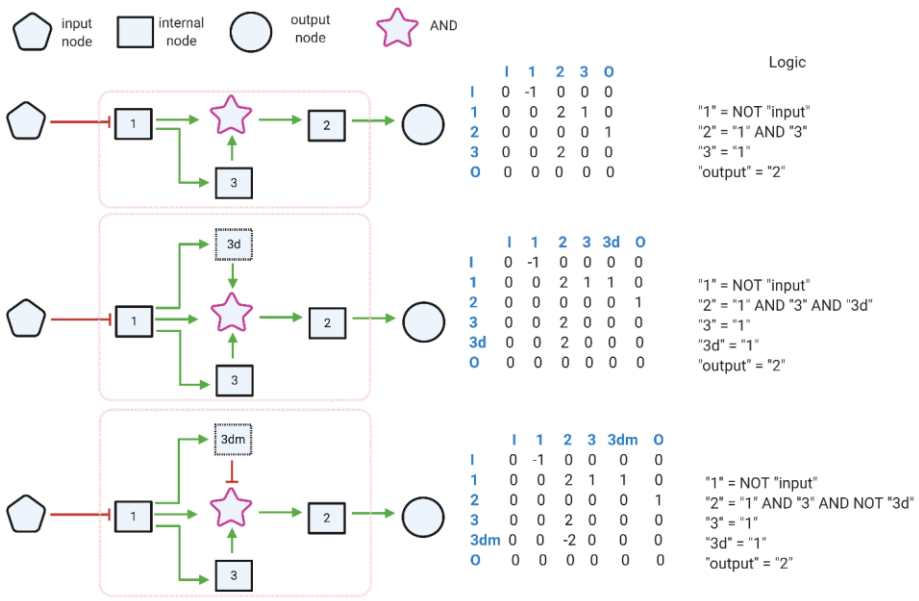

Figure 2: General overview of how an individual network is processed. Each individual receives an environmental signal. This signal is propagated through the signal processing section of the network and produces an output signal. The network output is used to assess the fitness of the individual. This process is repeated for each environmental signal, in our specific case, there are 3 unique environmental signals per network. The sum of the fitness for each input is used as the individual network's fitness. An adjacency matrix expressed the information describing a network. Each row corresponds to a node and each column represents the corresponding connection of that node. The first network describes logic representation. The second network assumed the duplication of node "3" which results in augmenting the logic driving the activity of node "2". The same node's logic is mutated in the third network.

Each individual is made up of the regulatory network itself, a mutation rate, and a duplication/deletion rate. These rates of change are associated with individuals, rather than with the population, which coarsely model the genetic systems which control gene replication during reproduction. The Boolean network is evolved to perform a specific task that determines the fitness of the individual. The mutation rate and duplication rate affect the rate at which components of the network are changed (mutated) or added/removed (duplication/deletion), but do not directly affect the fitness of the individual. By having the mutation rate as an individual trait, we are able to study how selection for functional individuals may affect the selection of higher or lower mutation rate individuals. This enables us to study how adaptive and non-adaptive forces interact. The fitness function of the evolution as based around periodic patterns, which have been used historically as target functions(Knabe, Nehaniv, and Schilstra 2008; Burda et al. 2011) as well as having biological significance(Zhang et al. 2014; Papagiannakopoulos et al. 2016; Baggs et al. 2009). The number of populations used (>100) for each condition was also on par with previous simulations(Kashtan and Alon 2005; Knabe, Nehaniv, and Schilstra 2008). For our specific examples, all networks were evolved to have 3 periods depending on the activity of the input node (i.e., environment) with a regular pattern of $[\tau, 2 * \tau, 4 * \tau]$. For example if τ =3 then the population would have three target outputs of a period 3, period 6, and period 12. This allowed for some control over the complexity of the evolved function in a controlled manner: Boolean networks that produce larger oscillatory periods require more nodes and therefore more Boolean functions. As τ increases, so does the complexity of the network necessary to produce the target outputs.

2.2.1 Boolean network

Each individual regulatory network is modeled using a Boolean network, Figure 2, which consists of an environmental input, an output, and a series of functional genes. The environmental input is represented by a node which has a defined value, this node has no inputs from the rest of the network. The output is used to measure the fitness of the network, and the node representing the output of the network cannot have an output to any other node within the network. This Boolean network is represented by a modified adjacency matrix which determines both the existence of connections between nodes, as well as the logic relationships that determine how multiple inputs into a node are controlled. Each row of the matrix represents a node's outputs and each column represents the node's inputs. For our specific experiments, the first row/column represents the input node and the second row/column represents the output node of the network. If a node has multiple inputs, the logic gate controlling those inputs is determined by the numeric values within the adjacency matrix. Each index on the adjacency matrix can have a value of 0, which represents there is no edge present or a positive or negative integer. The value represents a logical function associated with the connection: All incoming edges with a functional value of 1 are combined using an OR function, while all incoming edges with a functional value of 2 are combined using an AND function. Complex functions, consisting of combinations of AND/OR gates, are then defined by combining multiple functional groups. In the example of Figure 2, the 4^{th} column represents the inputs to Node 4. The presence of 2 inputs each with a value of "2" represents that these inputs will be processed by an AND function.

2.2.2 Selection

Each network within a population is assessed for fitness by comparing the output of the network in response to environmental inputs shown in Figure 2. The fitness is defined as the ability for the output node to produce a specific signal (either constant or oscillatory) for a given environmental input. Each network is given *k* input signals, and each input signal is repeated with *p* permutations of the network, where different functional nodes will be in the active or inactive state at the start of the simulation. Once the network reaches a limit cycle, the period of the output node is measured as a set of actual output periods[O_{a1} ; O_{a2} …O_{ak}] this set represents a 1:1 comparison of environmental stimuli to the network's response. The output oscillation is assessed only on the period of the signal, regardless of the specific pattern. These actual inputs are compared against the set of target periods [O_{t1} ; O_{t2} …O_{tk}]. The fitness of the network is then determined by the difference between the set of actual outputs and the target outputs.

$$Fitness = \sum_{1}^{k}\sum_{1}^{p} O_{ti} - |O_{ai} - O_{ti}|$$

For each environmental input signal, the initial values of each non-input/output node are randomized. The signal is allowed to propagate as before until the network reaches a limit cycle. Once the limit cycle is reached, the period of the output node is recorded. The process of randomizing the values of the network is repeated *p* times, to ensure that the network can process the input signal even when the network is perturbed to different initial values. The fitness is then calculated by summing the distance of the actual signal from the target signal. For our experiments *k* was set to 3 and *p* was set to 32. The value of *k=3* was used so that we could have two constant (nonoscillatory) environmental

signals and one oscillating environmental signal. This allowed for measuring the ability of the networks to create a response for constant and dynamics inputs. The target outputs of the system were set to be different than the input signals in order to remove the possibility of trivial solutions. The value of $p=32$ was selected as that is the number of possible starting positions of 5 boolean variables, which are the starting number of variables in our model (3 internal nodes, 1 input, 1 output). These values were selected because they were sufficient to prevent the evolution of trivial solutions, in this way the input signals are not a zeitgeber which sets the period of the input but an arbitrary environmental condition.

After the fitness of each individual is determined, parental pairs are created with each parental pair creating two offspring. Tournament selection is used to determine both members of each parental pair

2.2.3 Evolutionary moves: Crossover and mutation

After parental pairs are selected, a crossover is used to create two offspring. Crossover is performed by creating a new adjacency matrix, each row of the new matrix is filled in using a row from one of the parental networks. Mutation occurs based on the mutation rate of the parent networks. The mutation rate is the probability that any given index of the adjacency matrix is altered. Each index in the adjacency matrix has a probability of being mutated with a probability equal to the mutation rate of one of its parental networks. This mutation sets the value of the index to either 0, or to a positive or negative value of a functional group. This allows for each edge to be added, deleted, or functionally modified in the same mutation step.

2.2.4 Evolutionary moves: Gene duplication/deletion

After offspring networks were created using crossover and mutation, they undergo duplication/deletion which allows for networks to select their own size. Each individual has a duplication/deletion rate associated with it, which represented the probability that a gene may be duplicated or deleted during the mutation step of the genetic algorithm. For each node, a random number is generated, if it is equal or less than the duplication/deletion rate, the node is selected for a "duplication/deletion" event. There is a 50%probability that the selected node is duplicated, which creates a new node is added to the network, with all incoming and outgoing edges of the node being identical to the duplicated node. Figure 2(Top) shows a network prior to duplication/deletion or mutation, and Figure 2(Middle) shows the network following a duplication event. Duplicated nodes are capable of gaining novel function by mutation, such as the mutation shown in Figure 2(Bottom) where a change from a positive to negative value changes an input into a "NOT" gate signified by a red edge between the nodes. If the selected node was not duplicated, then it is deleted by removing the node and all associated incoming and outgoing edges.

2.2.5 Evolutionary moves: Mutation and duplication/deletion rate

The mutation and duplication/deletion rates are handled in an identical manner. Each new "offspring" will inherit their mutation and duplication/deletion rate from one of the parent networks. These rates are then mutated in a random manner. This allows for the self-selection of mutation and duplication/deletion rates: They do not directly impact fitness and therefore there is no direct selective pressure to select for a specific rate. Rather, there is an indirect pressure to increase the mutation rate if it is beneficial to the

evolutionary process or decrease the mutation rate if it is hindering the evolutionary process as a whole, rather than being directly tied to the fitness of the individual or the time course of the evolution in a determinant manner. Each individual's rate of mutation and duplication/deletion is adjusted by the following formula, where RAN is a random number from 90 to 110 and $Rate_g$ represents the mutation or duplication/deletion rate for the current generation:

$$Rate_{g+1} = Rate_g * (\frac{100}{RAN})$$

This rate is bounded between 0 and 1. This multiplicative random walk was used to represent the genetic basis of gene fidelity during reproduction: if an individual has a lower mutation rate there is a lower likelihood the DNA replication machinery mutates. Because of the bounds and the multiplication used, this produces a random walk which averages close to 0.5. This was used in favor of a more standard additive random walk which would have a constant rate of change in the mutation rate, regardless of what the current mutation rate was.

2.2.6 Identification and structural analysis of core networks

The core network is the set of nodes that are necessary and sufficient for performing the overall function of the network. A series of attacks on the network was used to identify this set of nodes. To do this, a baseline for the network is established. Each Input signal I_i is applied to the input node of the network. The resulting output signal O_i is then taken from the output node.

Two different stuck-at-fault attacks are then applied to each non-input/output node in the network. An attack is then performed by setting a node G_n to a value of either 1 or 0, creating 2 possible attacks per node. This value was then fixed, overriding the existing

logic of the node. Our regular assessment of the network was performed: each potential input was given to the input node, and the resulting output from the output node was measured. If the resulting output does not equal O_i , the attack is then considered lethal, regardless of how close the output may be. If an attack is lethal for any target output, regardless of the input signal, that node is flagged as being part of the core network.

2.2.7 Measure of Population Structural Homogeneity

To assess how homogenous a population of networks is, we developed a measure of "evolutionary distance". This distance score was based on the concept of "edit distance" from graph theory(Gao et al. 2009). We can define the "Evolutionary distance" between two graphs *G* and *G''* similarly to how the graph edit distance can be defined (Serratosa 2019):

$$EED(G,G') = \min_{\forall(e1,e2\ldots ek)\in E(G,G')} \left\{ \underset{(e1,e2\ldots ek)}{CED(G,G')} \right\}$$

Where CED represents the edit cost for each edit path e_i which would convert the graph G to G'. Unlike the graphical moves normally used(Bunke and Allermann 1983), we only allow for "evolutionary" moves, which include node duplication, deletion, and edge mutation. Each of these functions is assigned the same edit cost. Because node duplication and deletion may have different amounts of change in information, we can assume $EED(G,G') \neq EEG(G',G)$ when both graphs have different sizes. To control for this, we measure the EED from the smaller to the larger network only. The edit distance was then calculated by identifying the node-to-node bijection which had the smallest edit distance, where node substitutions could be performed by deleting one node and duplicating another. In the case of size differences, additional nodes were introduced by duplication rather than introducing null nodes. This allows us to determine how related

two networks are together using only the moves allowed under our evolutionary algorithms. A standard graph edit distance may fall short in identifying similarities between networks because of the amount of information that can be added or removed from gene duplication and deletion events.

2.2.8 Simulations and computations

To study the evolution of regulatory networks we performed two general types of simulations: The first simulation modeled the growth of a non-functional network into a functional regulatory system. This was done by using a *naïve* network, which is made up of nodes that have no edges or logical relationships between them. In this way, there is no inherent bias within the network which may influence the structure used or performance of the network over time. The second simulation modeled environmental stressors by changing the target output function of the population partway through its evolution. This forced an already adapted network to re-adapt to a new function. In all cases, the fitness of each individual network was dependent on the period of the output node, regardless of the pattern produced by either the output node itself or any other member of the network. Simulations and calculations were performed on the Amarel high performance computing cluster at Rutgers University. Each population was evolved for a predefined number of generations (3500 for single target evolution, or 3000, 5000, or 8000 generations for two target simulations). These predefined generations were selected based on how well a population could reach its goal. We considered a population to be converged if half or more of its individuals had achieved maximum fitness. For $\tau = 3$ 50% of populations had converged by generation 1000, 62% at generation 3000, and 71% at generation 6000, and 89% at generation 10,000. Based on this, a time limit of 3500 generations was selected

as a non-trivial time limit for our single target experiments. For the most complex single target function used ($\tau = 9$) 35% of populations converged by 3500.

For the switch experiments, the populations were evolved for a set amount of time with a target function of $\tau = 3$, however, during the final 2000 generations, the target function would be changed to $\tau = 6$. For example, if the switch experiment lasted for 5000 generations, then the first 3000 generations would have $\tau = 3$ as a target function and the final 2000 would have $\tau = 6$ as a target function. Approximately 1/3 of simulations reached the original and second target regardless if the switch occurred at generation 1000 (29%), generation 3000 (34%), or generation 6000 (30%). The highest fitness individuals of converged populations would be used for calculating the evolutionary distance within and between populations, while individuals which were maladapted within the population would not be considered for structural analysis.

For each condition, 500 simulations were completed to ensure at least 100 populations converged within the time limit, regardless of how complex the target function was. This provided a sufficient number of populations that achieved the proper target in order to study how networks evolve, especially in non-trivial conditions.

Importantly, the random seeds for each condition are identical. This means that for the switch experiments, where the initial target is $\tau = 3$, at generation 1000 all three switch conditions have identical populations. This is important for comparisons because it means that we can make a direct comparison between the populations that have changed targets and have an environmental disruption and an identical set of populations evolving without an environmental disruption.

2.3 Evolved oscillating systems

2.3.1 Single target evolution

Our first set of experiments were designed to study the evolution of naïve networks toward a single target function. In this case, the naïve networks were initialized with no logic function relating to any of the nodes. During the initial stages of evolution, the naïve networks show a small increase in mutation rate and a rapid increase in network size, on average. The initial mutation rate was set to 0.01, and the initial network size was 5 where there was an input node, output node, and 3 signal transduction nodes. For τ=3, the mutation rate had dropped to $4.97e^{-4}$ by generation 1000, a decrease of 95%, while the average number of nodes in each network increased 5-fold. (Figure 3 top and middle). The population's performance is measured by comparing the oscillatory period of the output node *O* for each of 3 different inputs (Where the input node *I* is set to *I* = [1] *I* = [0] and *I* = [0 1] where it will oscillate between active and inactive at each time point) (Figure 3 bottom). Other values of τ produced networks with slightly different values (larger network sizes and smaller mutation rates) but had the same general characteristics.

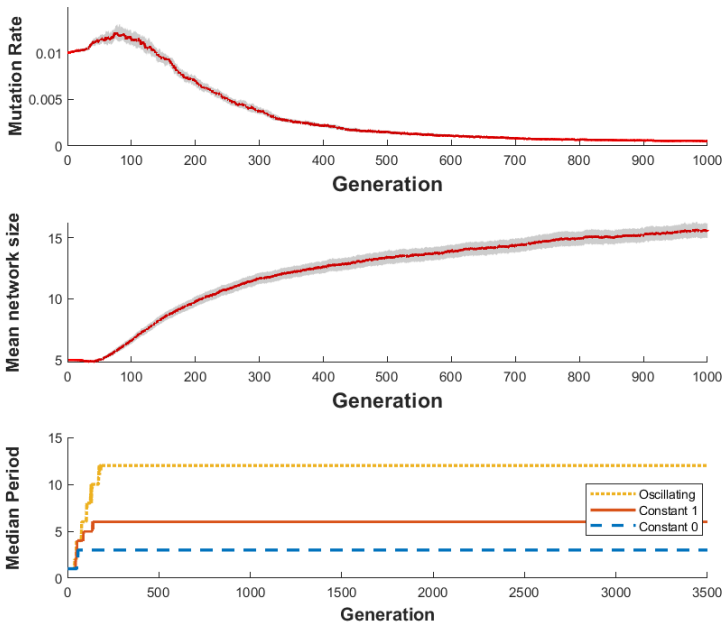

Figure 3: Population-level statistics of the mutation rate, network size, and network output. Top: The median mutation rate over time. The initial mutation rate starts as a fixed value and then is allowed to adapt over time. Initially, the population selected for a higher mutation rate, followed by depression the mutation rate much lower than the initial rate. Middle: The median network size over time for populations with a target function of $\tau = 3$. Initially, there is a high network growth, with most of the network growth being done within the first 500-1000 generations. After this, the median size grows at a slower rate. Both graphs show the phasic nature of these traits: early in the evolutionary history of the population, there is a rapid change in mutation rate and size. As the populations converge to a single function, changes in these traits slow. Bottom: The median period for each of the three inputs. This population has a target of [3 6 12] for inputs of a constant 0, constant 1, and an oscillating signal. Here we can see that the mean value of the outputs approaches the target function within the first several hundred generations.

The tendency of the network size and mutation rate (and hence the number of mutations per generation) to increase initially existed regardless of the target function of the population (τ value), although larger values of τ tended to have larger final network sizes and smaller final mutation rates. We hypothesize that the mutation rate decreased to a lower value in these more complex tasks because these larger networks are more vulnerable to being disrupted by changing the connections between components and further that the lower mutation rate is a response to the larger network size to decrease the number of mutations per generation. The number of mutations that occur is dependent on

the size of the network and the mutation rate, in order to account for variations in these we calculated the average number of mutations per generation. This provides insight into the overall net change of information in the network. Overall, once the target function is found the number of mutations each generation will approach 0, regardless of τ (Figure 4). The peak average number of mutations per generation is 404% when τ=9, 357% when τ=6, and 302% when τ=3 of the starting number of mutations per generation. By generation 1000 these drop 58%(τ=9), 59%(τ=6), and 54% (τ=3) of the starting mutations per generation.

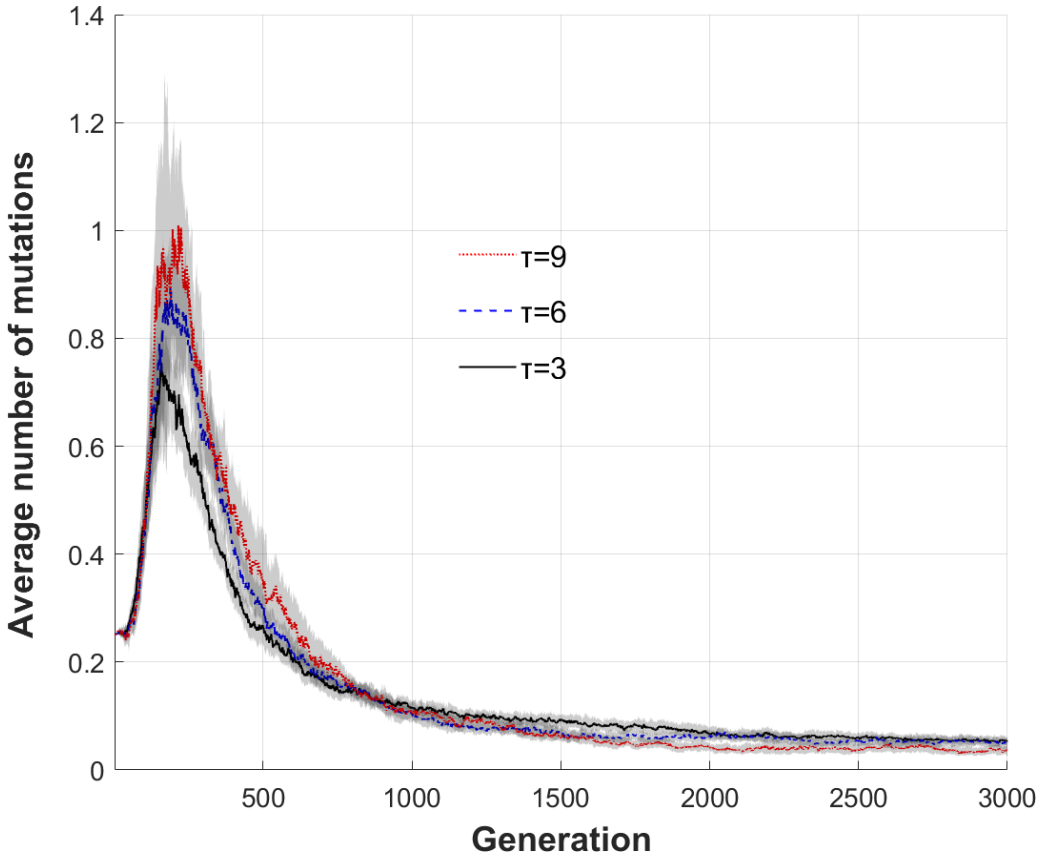

Figure 4: The average number of mutations per generation, which is a function of the network size and the mutation rate, follows the same qualitative evolution regardless of function complexity. A 95% confidence interval is shown in gray around the lines. The more complex the function, the more mutations occur at the maximum, but after the number of mutations peaks there will be a trend to minimize the number

The most striking result from the naïve networks is the suppression of the mutation rate as well as the rapid initial growth of the network size. For $\tau=3$ there is a 22% increase in average per-edge mutation rate. By generation 1000 the average per-edge mutation rate has decreased 95% from its starting value (Figure 3, top). The mutation rate of individual networks has no direct impact on the fitness function, but every population saw a decrease in mutation rate over time.

The overall rate of change within a network is not controlled only by the per-edge mutation rate. The size of the network determines how many edges exist, and therefore larger networks will have a higher number of mutations within a single generation for the same mutation rate. Network size rapidly increases, but there appears to creep upward regardless of the fitness of the population (Figure 3, middle). When we account for network size to determine the overall rate of change of the network, rather than the per-edge mutation rate, we find this decreased rate of change still exists. Regardless of the target period, every population had a decreasing mutation rate over time when there was a single target function. At generation 1000, the average number of mutations per generation was 58%($\tau=9$) 59%($\tau=6$), and 54% ($\tau=3$) (Figure 4).

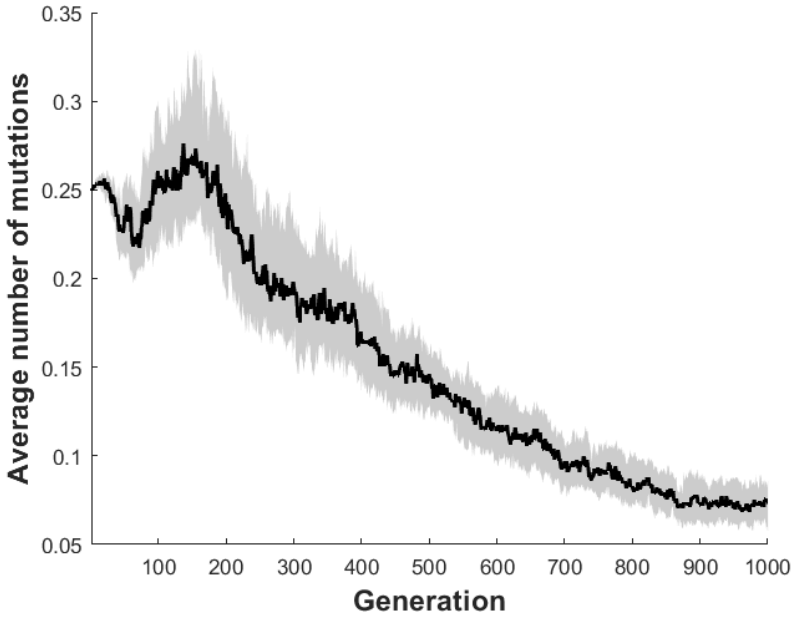

Figure 5: The average number of mutations in the mal-adapted populations over the first thousand generations (black line) and the 95% confidence interval (gray area). The average never increases as high as quickly adapting population, but there is a downward trend in the average number of mutations regardless. If there was no fitness function, the average number of mutations would be expected to increase as the size and mutation rate are both expected to increase.

The early peak mutation rate followed by a decreasing mutation rate over time appears to be independent of whether an individual population finds a maximum fitness network early. This was even true for populations that did not achieve maximum fitness. Mal-adapted populations would still decrease their mutations rates to be protective of the fitness they did gain (Figure 5). These mal-adapted populations had an average decrease of 70% from the start of the simulation through generation 1000, much greater than the decrease of any on-target population. This decreasing mutation rate is a protective mechanism that allows a population to maintain the fitness gains it has made, and so populations which have found second or third best solutions would often see the decrease in mutation rate as well until a better solution was found. The maladapted populations have a larger decrease in the number of mutations compared to the populations which reach their target function within the time limit (70% decrease compared to a 58%

decrease). This may indicate that the maladaption is related to an over the protection of the population where higher mutation rates are overly selected against which slows the ability of the population to improve.

2.3.2 Effect of environmental stressor

The second set of simulations was used to measure how an evolved or evolving network would change in response to a new environmental stressor. These simulations evolved with a target function of $\tau=3$ for 1000, 3000, and 6000 generations before changing the target function to $\tau=6$. At generation 0 the populations were initialized as previously described: a mutation rate of 0.01 and a network size of 5. During the first 1000, 3000, and 6000 generations the simulations proceeded as if they were a single target evolution. After this initial period was complete the fitness function was changed to the new target and populations were given an additional 2000 generations to re-adapt to the new function to standardize the amount of time a population would have to re-adapt, regardless of its starting point. Similar to before, the fitness of individual networks was based entirely on the period of the output node, and not on a specific pattern.

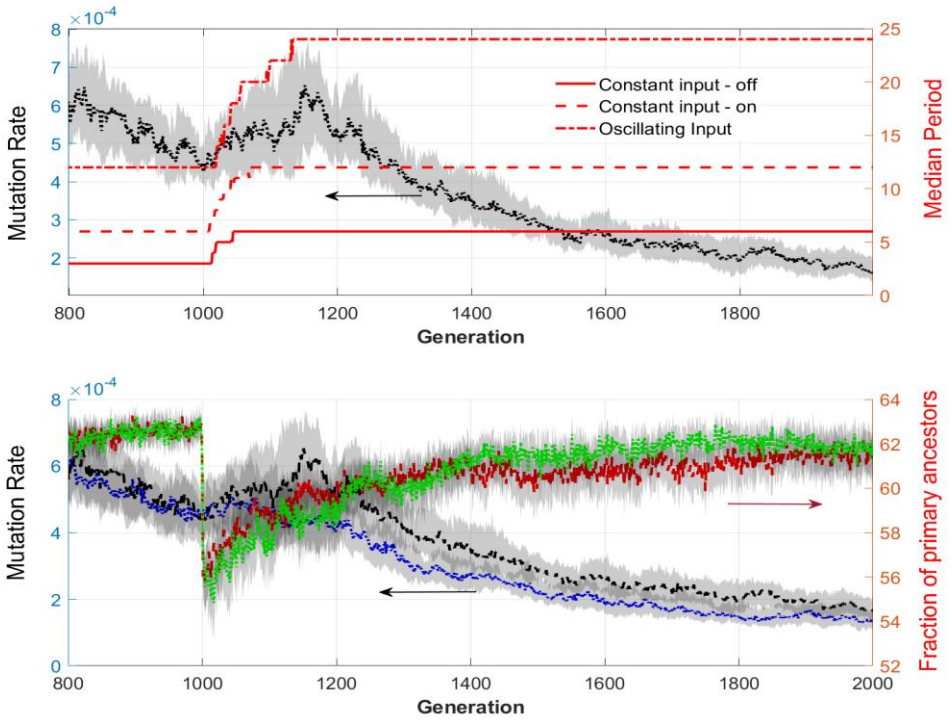

Figure 6: Population-level statistics for the re-adaption of networks after the target function has been changed. Top: The median period of a population with an initial target of τ = 3. At generation 1000 the target is changed to τ = 6. All three input/output functions change within a few hundred generations of the change in function (left). The right axis shows the median mutation rate for the same populations. Here we can see that the mutation rate peaks shortly after the populations have adapted to their new functions. Bottom left axis: The dotted black line shows the median mutation rate of populations which found the new target, compared to populations that did not find the new target within the time limit (blue line). Right axis: The fraction of individuals selected for populations that successfully transition within the time frame (green line) and those which don't (red line) both have an apparent "selection event" where the number of individuals selected for reproduction decreases drastically.

We compared the populations which had evolved to have both the initial target τ as well as the second target τ to populations which only converged to the initial target but failed to adapt not the second target. The major difference found between re-adapted populations and populations which failed to re-adapt within 2000 generations was the recovery of mutation rate (Figure 6, Top). We believe that this is due to a smaller, higher mutation rate subpopulation being selected only after that sub-population has found a higher fitness solution. This is affirmed by the decrease in the number of networks selected for the median number of networks selected for decreased following the switch

to the new function (Figure 6, bottom). The average fraction of individuals selected for the next generation, for the switch occurring at generation 1000, reaches a minimum of 56%. We can see this sub-population selection clearly by studying the change in the number of mutations compared to the average fitness of the population. Figure 7 A shows the evolution of a naïve network, and the fitness sharply increased preceding an increase in the mutation rate. Figure 7 B and C show the functional switch and generation 3000 and 6000, and the fitness recovers sharply at the same time as the change in the number of mutations. The populations do not start to suppress the number of mutations until after the fitness has begun to approach an asymptote, indicating that the population has reached a relatively stable state.

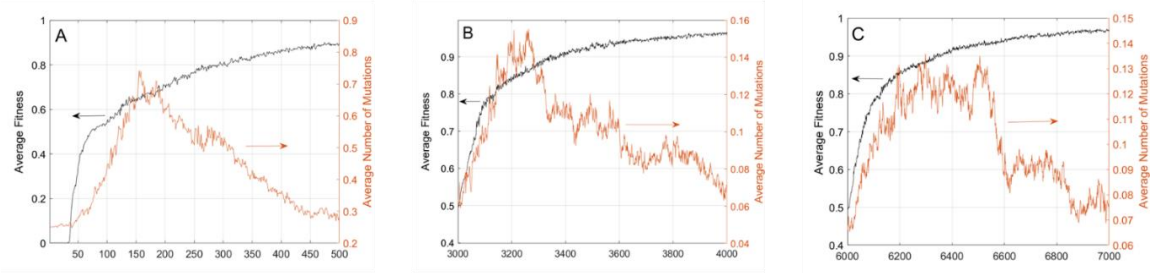

Figure 7: The average population fitness compared to the average number of mutations during naïve evolution (A) and during evolution between two targets at generation 3000 (B) and 6000 (C) in all cases, the slope of the fitness function is steeper initially than the slope of the number of mutations, and the number of mutations does not decrease until after the average fitness has begun to reach an asymptote.

During the simulations in which the target function switches, we do see the per-edge mutation rate, and the number of mutations per generation both increases, like the small peak seen with the naïve networks from the single target simulations. This raises an interesting question of whether the mutation rate increases as a strategy to find a new target function, which would be the obvious conclusion. If this were the case, we would see the mutation rate increase first, and then the fitness of the population increases second. However, our data seems to indicate that the increase in mutation rate is driven

primarily by a selection of a small sub-population of high mutation rate individuals achieving a higher fitness and out-competing the lower mutation rate individuals, as seen by the “selection event” associated with changes in target function, where the average fraction decreases from 63% of networks to about 56% of networks for several generations (Figure 6, Bottom). This causes an indirect adaptive response: The higher mutation rate poses a risk to the population and needs to be decreased, which is why we see peaks and decreases in the mutation rate in our simulations. Because the populations do not know what their maximum fitness is, we would expect this decrease in mutation rate even before all members of the population have achieved maximum fitness.

Further evidence that the change in mutation rate is not being utilized as a population-wide search strategy is how the increase in fitness precedes the increase in the number of mutations per generation (Figure 7). Given that the population-wide mutation rate increases after, rather than before, the fitness of the population increases it is unclear how different factors may affect a population’s ability to re-adapt to a new environmental stressor. Factors such as network size and mutation rate play critical roles in what functions a population may be able to achieve quickly, however, our results demonstrate that no specific feature of the networks affected the ability of a population to adapt to a second target τ: Network size, mutation rate, and structural homogeneity all did not correlate with how quickly a population re-adapted to its new function (Figure 8). Coupled with the fact that the mutation rate tends to peak after the fitness of a population has greatly recovered demonstrates that population-wide changes in adaptive processes proceed non-adaptive changes.

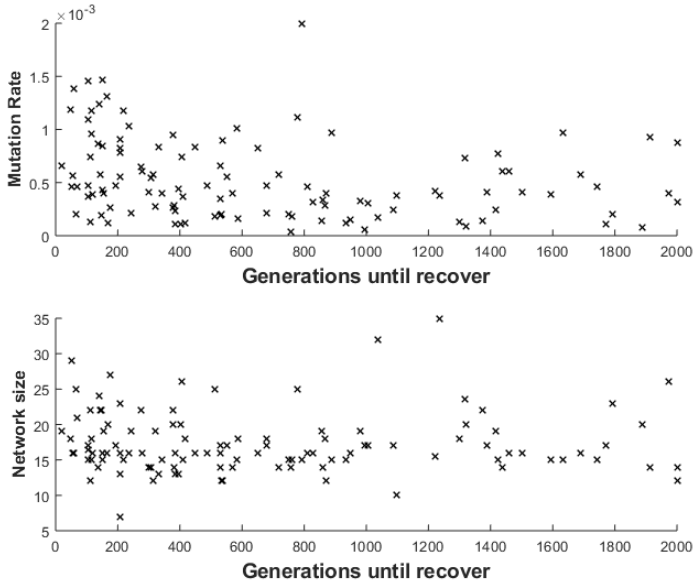

Figure 8: The number of generations it takes for populations to re-adapt do not correlate with either the size or mutation rate of the population as a whole. Top: Scatter plot of the average mutation rate of a population vs the number of generations it took for that population to switch from $\tau = 3$ to $\tau = 6$ Bottom: Scatter plot of the median network size of populations compared to the number of generations it took for a population to switch from $\tau = 3$ to $\tau = 6$. No population-wide statistic is correlated with the number of generations it took for a population to recover, indicating that transitioning between two functions was not caused by a population-wide process but the expansion of a subpopulation.

The mutation rate, and the number of mutations per generation, are a secondary characteristic of each individual: they do not directly affect the fitness of that individual or affect its ability to be selected for the next generation. Rather, these characteristics may affect their offspring: A higher mutation rate may increase the probability of a deleterious mutation, or a network structure may be more robust against mutations. These secondary characteristics, such as mutation rate and network structure, represent the non-adaptive forces of evolution because they do not affect the phenotype of the current generation.

. These two ideas (that population-wide non-adaptive processes do not improve the adaptability of a population, and that mutation rate will rise after the fitness of a population improves rather than before) fit into a broader biological picture as well.

Mutation rate provides no inherent benefit for an organism during stable conditions, but provides an inherent risk: it is possible for a good regulatory network to have a deleterious mutation which results in a loss of function, but it is unlikely for a mutation to be beneficial to an organism if it is already well evolved to survive in a stable ecological niche. Even if there is an environmental stressor or toxin, the probability of mutation making an organism more fit for overcoming the stressor is unlikely.

2.3.3 Evolution of network structure

Rather than study the structure of evolved networks through the use of motifs, we used our evolutionary edit distance metric, which was based on a modified graph edit distance. The evolutionary edit distance calculates the number of moves necessary to transition from one network to another using the moves allowed by our evolutionary algorithm. Understanding that our evolved networks, much like real-world genetic and metabolic interaction networks, have some components which are critical to the function of the network and other components which are not we did not calculate this edit distance for the whole network, but rather the distance was calculated for the core network, which consists of the nodes necessary and sufficient for the function of each network. These cores represent the functional and evolutionarily conserved components of each individual, where mutations or alterations may have a significant impact on the overall fitness. Nodes outside of this core do not influence the output of the network, and any attacks on these nodes will not affect the overall output of the network. Figure 9 top shows the number of cores in a population with a target function of $\tau=3$ at generation 1000. In this case, half of all populations have 90% or more of their individuals

represented by 4 or fewer core networks, and half of all populations had 100% of their individuals represented by 10 or fewer core networks.

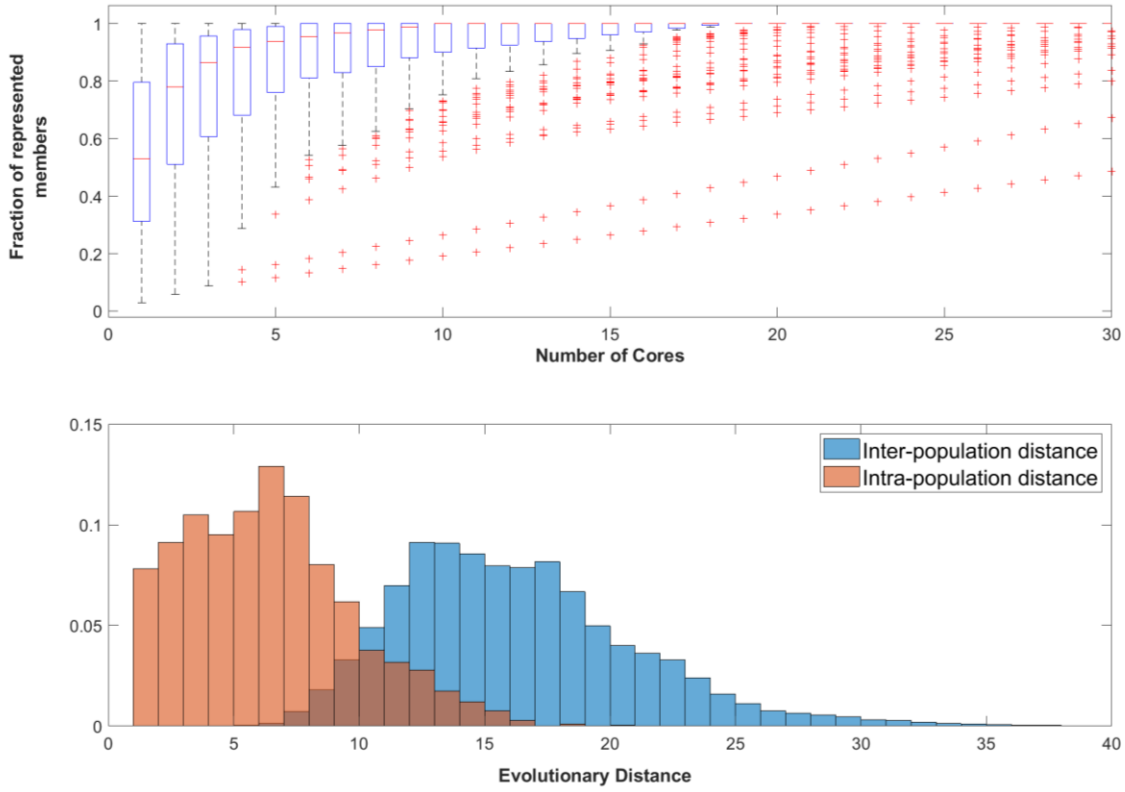

Figure 9: Structure of populations.Top: The distribution of the fraction of individuals whose function is explained by a certain number of cores for $\tau = 3$ at generation 1000. For each box plot, the fraction of individuals within a population who have that number of the most common cores are plotted. At 1 on the x-axis, the number of individuals within a population who share the most common core are plotted, at 5 the fraction of individuals who have the 5 most common cores are plotted. Over half of all populations have 90% or more of their networks represented by only 4 cores. While having a small number of cores is not a unique strategy, it is far more common than having a unique structure for each individual. Bottom: A histogram showing the distribution of evolutionary distances between individuals within the same population (red) and individuals from other populations (blue). We can see here that networks within the same populations are generally highly related, generally requiring 5 or fewer evolutionary moves to convert between the two individuals. In contrast, individuals from different populations have more differences. This implies that while there is a purifying selection within a population for highly related structures, different populations have not converged toward the same structure. The distributions were compared using a Kolmogorov–Smirnov test and determined to be different with a p-value < 0.01.

Once the core networks are identified, we used the evolutionary distance to calculate how similar the core networks were within a population (the intra-population distance). This was compared to the evolutionary distance between the core networks within a popular compared to the core networks of other populations (the inter-population distance). We

used this to measure the evolutionary distance of cores within the same population to cores from other populations and found that populations were more consistent with themselves than they were with other populations with the same target τ (Figure 9 bottom). This data only consists of the distance calculated between different structures, so if there were several copies of the same network within a population, this does not contribute to the intra-population distance. There were zero instances of an identical core structure being found in two different populations, regardless of τ or generation.

Populations from both types of simulations (those with a single target, and those which have a first target and a second target τ) were compared for their diversity. Over time the number of core functions that exist within a population decreased if there was only a single target. Populations that changed target functions maintained their population diversity of network structure, with an average of 53% of all networks sharing the same functional structure (Figure 10 top). However, if the target function remains the same, the number of unique structures drastically decreases, with an average of 67% of all networks sharing the same core structure (Figure 10 bottom).

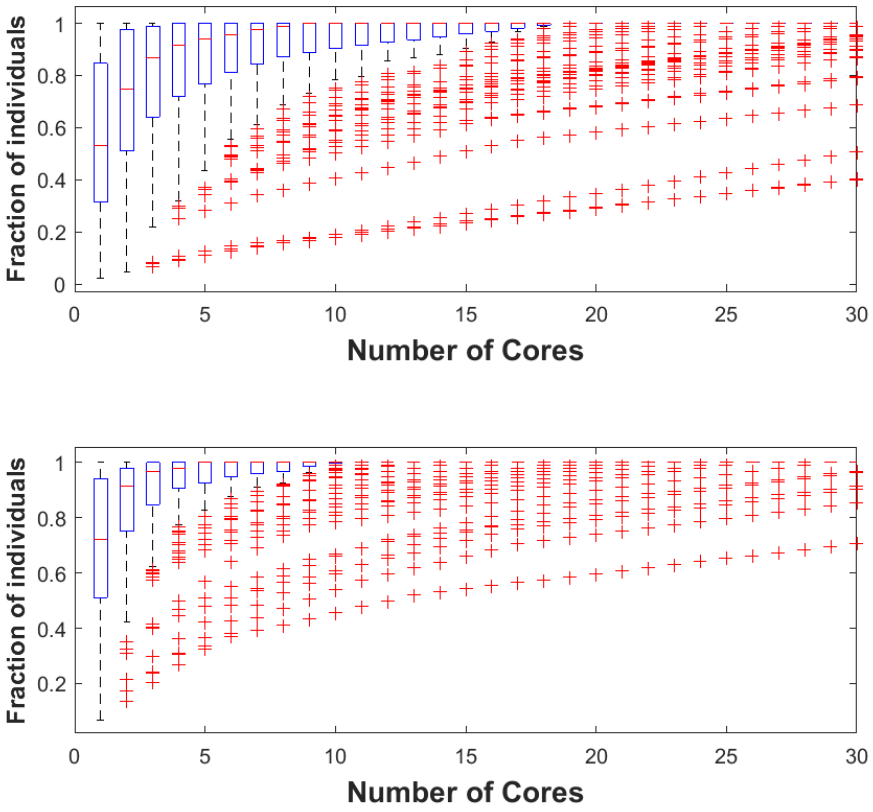

Figure 10:Comparison of population diversity depending on whether an environmental stressor is introduced. Top: The distribution of the fraction of individuals whose function is explained by a certain number of cores for τ = 6 at generation 3000, after the target function was changed from τ = 3 at generation 1000. This is comparable to the distribution in Figure 7 Top. The distribution is very similar, and overall diversity is maintained, with half of all networks having 9 or fewer core functions. Bottom: The distribution of the fraction of individuals whose function is explained by a certain number of cores for τ = 3 at generation 3000. This can be compared to Figure 7 Top, where the overall diversity has decreased. Half of all populations had 5 or fewer unique cores.

Another factor may be related to the diversity of structures within a population. Different structures existed within the same population, as well as different populations finding different sets of solutions. The cause of the inter- and intra-population differences lies with adaptive processes and non-adaptive processes. Each population adapts a small number of strategies that fulfill the target function, and the selection of these strategies varies from population to population. For τ=3 at generations 1000 there was an average of 53% of networks sharing the same identical core structure, and 50% of populations

having 10 or fewer unique core structures (Figure 9 top). These structures appear to be divergent between populations. Within a single population, each core network has a median evolutionary distance of 6, while between populations there is a median evolutionary distance of 15 (Figure 9 bottom). This divergence is strengthened over time as purifying selection and decreasing mutation rates limits the ability of conserved components to easily shift from one strategy to another. We can utilize the switched evolution simulations to see how time affects the diversity of structures in a population. Comparing the structure of the networks before switching at generation 1000 to the structures of networks before switching at generation 3000 allows us to see this purifying selection in action. Over the course of 2000 generations, the percentage of the population which shared the identical functional network rose from 53% to 67% for τ=3. The median number of unique core networks within an average population also decreased from 10 to 5, representing a nearly 50% decrease in the diversity of the populations. This highlights the importance of environmental stressors and competition in maintaining the diversity of a population. For the populations whose target switched from τ=3 to τ=6, there was no significant change in the fraction of networks with a single core network (53%), and 50% of networks had a median of 9 unique structures or fewer (Figure 10).

The decreasing mutation rate creates a freeze, which prevents novel structures from forming. Simultaneously, there is a natural bias of selection leading to a single structure. In this way, the structure and diversity of a population are related to both the adaptive and non-adaptive pressure, which is in turn determined by adaptive pressure as well. However, our data shows that there is no overlap in core networks between populations and that there is a relatively large difference between the structures of individuals within

different populations. This shows that despite the ability for evolutionary conservation to control the network structure, this does not lead to convergent evolution. Due to this, we believe that non-adaptive forces are not responsible for the structure of regulatory networks, but neither is the structure of a regulatory network defined by its function.

2.3.4 Biological implications

Biologically, it appears that some eukaryotes have evolved mechanisms to reduce the randomness of the non-adaptive processes. These mechanisms create different genetic regions, which have different mutation rates, even within the same organism(Chuang and Li 2004). Further work is necessary to understand how these traits interact with a dynamic environment, including different biomes and interactions between different species. Between the evolution and spread of antibiotic-resistant bacteria to changing climate affecting both beneficial and pest species understanding how adaption to new environments may play a critical role in understanding the behavior of these species in the future. We already see the impact of climate change, with certain species thriving due to environmental changes, and many others dying off as changing conditions are not suitable to their survival(Ceballos et al. 2015; De Vos et al. 2015; Bradshaw and Holzapfel 2008; Gienapp et al. 2008; Sirisena and Noordeen 2014). We hope to expand this work to better understand not only how species adapt to stressors, but how multiple species and populations may interact and adapt to one another under stressful conditions.

Our model demonstrates the emergence of several unique, evolutionary events that align with observations from field and laboratory studies. Here we see selection against higher mutation rates at the population level, leading to decreased change over time and overall decreasing diversity within a stable environment. We see a recovery of the population-

wide mutation rate after an external stressor is applied. This indicates that many populations survive environmental stressors by amplifying a subpopulation which has both a higher fitness and mutation rate, rather than a slow, homogenous change to reach a new functional phenotype. The rapid expansion and adoption of new sub-populations seem to agree with the idea of punctuated equilibrium: if there is sufficient disruption to an environment there will be a rapid period of evolution rather than slow adoption of new traits. Although not all populations reach the target function quickly, there seems to be a large initial set of populations that can achieve their function and a long tail of populations that achieve the target despite the decreasing mutation rate.

These results demonstrate that the mutation rate and adaptability of a population are highly dependent on individuals and subpopulations. The presence of competition and environmental changes are essential to maintaining biological diversity, while populations seem to stagnate under constant conditions.

The presence of unique functional components within each population also indicates that the convergent evolution of regulatory structures is unlikely. Even if certain motifs or structures are created more likely through random processes, the selective pressure for functional networks and conservation of those components overrides the random evolutionary steps which take place. Including the interactions between non-adaptive and adaptive forces plays an important role in the development of regulatory networks and should be considered for models that seek to understand the development and evolution of biological networks.

CHAPTER 3: A framework of identifying and contextualizing circadian genes

3.1 Background

Circadian patterns are seen throughout almost all domains of life and are seen as critical to many physiological and metabolic functions (Feillet, Albrecht, and Challet 2006) and are seen as creating important activity cycles which improve survival as an adaptive advantage (Silver et al. 2012; Edery 2000; Getz 2009). Within mammals, these circadian rhythms are entrained to the light/dark cycle and control the physiological functions of the organism which is regulated and primarily driven by a master clock located in the suprachiasmatic nuclei (SCN) (Cardone et al. 2005; Cassone 1990; Feillet, Albrecht, and Challet 2006; Albrecht 2012). This master clock synchronizes peripheral clocks throughout an organism, creating an interconnected network of clocks which coordinates temporal variation(Buijs et al. 2003; Dibner, Schibler, and Albrecht 2010) and coordinates interactions between organism-wide physiological systems (Cutolo and Masi 2005). This integrated, synchronized networks of clocks exert control over many physiological functions necessary for health and wellbeing, and their disruption often results or exacerbates pathological conditions (Bae and Androulakis 2019; Rao and Androulakis 2019; Scherholz, Schlesinger, and Androulakis 2019; Rao, Scherholz, and Androulakis 2018; Bae and Androulakis 2018; Mavroudis et al. 2013). Conversely, targeting these same circadian pathways during the progress of pathological conditions may improve the response of patients suffering from these conditions and may provide novel therapies or improve the efficacy of existing ones(Cunningham et al. 2016; Fang et

al. 2015; Paladino et al. 2010; Nakamura et al. 2016; Zaki et al. 2018; Khaper et al. 2018).

These implications become more interesting as genome and organism-wide data are studied further. A set of core circadian genes and transcription factors have been found to be expressed consistently and coherently across all tissues(Buhr and Takahashi 2013). This core clock interacts with all tissues, and results in broad genetic (Mavroudis et al. 2018; Mure et al. 2018; Zhang et al. 2014) and metabolic (Dyar et al. 2018) circadian patterns throughout organisms. However, many of these studies have found that there are tissue-specific effects: either genes are not consistently expressed across tissues or they're expressed with different patterns in different parts of the body (Mavroudis et al. 2018; Mure et al. 2018; Zhang et al. 2014).

In this work, we establish a computational pipeline for the identification of circadian genes within the context of functional pathways. Without loss of generality, we utilized pathways defined within the KEGG database (Kanehisa and Goto 2000; Aoki and Kanehisa ; Kanehisa et al.). Using these pathways, we seek to contextualize rhythmic patterns within physiologically meaningful pathways. Further, we capitalize on the intensive work in creating these pathways by identifying functional nodes within the defined pathways, which are represented by genetically and functionally related genes. In this way, we can identify not only identical gene behavior across multiple tissues but how similarly functional genes may be expressed in tissues that do not otherwise appear to share gene expression.

3.2 Approach

A novel pipeline for the detection of rhythmic genes expression for the purpose of identifying commonalities between tissues was developed utilizing existing and novel techniques applied to the analysis of gene expression profiles (Figure 11). This pipeline uses statistical tests to detect rhythmic patterns in gene expression and then correlate them in a pairwise manner across multiple tissues, using transcriptomic microarray data from mouse tissue.

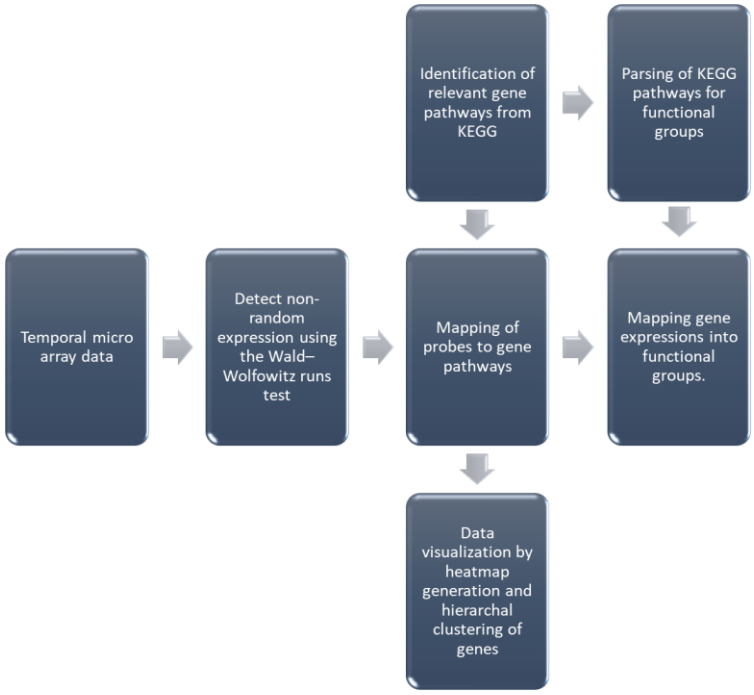

Figure 11:Overview of data processing of the raw microarray data. Data is either visualized and processed as individual genes in the context of pathways, or mapped into functional nodes within those pathways.

3.2.1 Microarray data

Zhang et al. measured timecourse transcriptomic data in 12 organs in mice to characterize circadian gene expression (Zhang et al. 2014). An impressive compendium of protein-coding genes from 6-week-old male C57/BL6 mice was quantified in aorta, adrenal gland, brainstem, brown fat, cerebellum, heart, hypothalamus, kidney, liver, lung, skeletal

muscle, and white fat tissue. The mice were entrained with 12 hours of light, 12 hours of dark for 1 week before being kept in darkness for 48 hours. Three mice were sacrificed every 2 hours during a 48-hour period to collect the tissue specimens at CT0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, and 22. Tissue samples were homogenized, and RNA was extracted as previously described in (Hughes et al. 2009). This darkness was meant to reduce outside noise and was short enough that the mice should retain their entrained circadian rhythms. The RNA from all 3 mice sacrificed at each time point was pooled to average out biological as well as technical variations. The RNA abundances were then quantified using Affymetrix Mouse Gene 1.0 ST arrays.

3.2.2 Data pre-processing

The 48-hour data collected from the mouse tissue was divided into two circadian periods, each 24 hours long. Each circadian profile was then z-scored with respect to the mean and standard deviation of that profile. These z-scored profiles are then used for all further analyses. If either period is found to be differentially expressed from the runs test described below, then that probe is considered to be differentially expressed.

3.2.3 Gene-pathway mapping

Genes identified by the Wald-Wolfowitz runs test as being differentially expressed are mapped into pathways. These pathways represent a group of functionally and mechanistically linked genetic components that are expressed in a coordinated manner to produce a signaling or biochemical function. These pathways are defined in databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG) (Aoki and Kanehisa 2005; Kanehisa and Goto 2000) and Reactome (Fabregat et al. 2018). The present analysis is based on pathways found in KEGG, without loss of generality. Pathways that are disease-

specific (such as cancer pathways) were dropped from the analysis. Certain organ-specific functional pathways (such as digestive specific pathways related to bile production) were dropped from the analysis due to them not being physiologically relevant in all tissue types. We expect the remaining pathway to have genes expressed in multiple pathways, and hope that the comparison of which genes are expressed in a circadian manner or what phases they are expressed in may help relate to different tissue functions. Transcriptomic data from all tissues are mapped onto pathways by converting the Affymetrix probe identifiers within the microarray template are translated into their NCBI Entrez IDs and Gene Symbols using the Affymetrix Mouse Gene 1.0 probe annotation data and converted into KEGG IDs to be sorted into the identified pathways. Affymetrix probe identifiers. This mapping process has previously been described in detail (Acevedo et al. 2019).

3.2.4 Runs test

Rhythmic genes are detected using a one-tailed Wald–Wolfowitz runs test (Wald and Wolfowitz 1940)using binarized transcriptomic data, such that all time points which are equal to or above the mean of the profile are set to a value of “1” and all other time points are set to a value of “0”. The runs test provides a unique method for detecting patterns within gene expression data which is not based on overall expression levels of a gene(Love, Anders, and Huber 2014) or based on curve-fitting or predefined patterns (Straume 2004; Hughes, Hogenesch, and Kornacker 2010). This test is performed by comparing the number of runs in data compared to the expected number of runs to determine if each member of a sequence is drawn from a distribution independent from the member before or after it. For a sequence, *V,* a subsequence, *v,* is defined as a run if

$v_{s+1} = v_{s+2} = \ldots = v_{s+r}$ and $v_s \neq v_{s+1}$ and $v_{s+r} \neq v_{s+r+1}$ while $s+r<N$ where N is the number of elements in V. If each member of the data is drawn independently from the same distribution then the expected number of runs, E, would be:

$$E = \frac{2mn}{m+n} + 1$$

Where m is the number of members of the sequence with a value of "1" and n is the number of members of the sequence with a value of "0". The variance of the expected number of runs would be:

$$\sigma^2 = \frac{(\mu-1)(\mu-2)}{m+n-1}$$

One challenge with implementing the runs test is that the data represents a repeating 24-hour period. If a run begins at the end of a period and continues from the start of the next period, then that run will be split into two runs in our data. To control for this, we will need to time-shift the data sequence V with a length k such that $V_1 \neq V_k$. This will ensure that no runs are interrupted by the arbitrary start and end time of the data collection. The runs test was compared to other methods of detecting circadian genes(Table 1). When tested with a P-value of 0.01 there was a 0.0% false-negative rate. The false-positive rate was 0.9%, which was higher than the other methods described for this purpose. The two primary advantages of the runs test are that each gene is tested independently: there is no reliance on comparing the runs test to a population-wide statistic. The second advantage is that the runs test is wave-form independent. Regardless of the period or waveform, the runs test can detect periodic functions. Due to the data being divided into 24-hour periods, our analysis is useful for detecting circadian rhythms that are close to this period.

If there is significant variation in the periodic signal, the runs test is still capable of detecting those signals but the method of dividing the data into 24 hour periods would be inappropriate. Because of this, we focus our discussion on circadian rhythms.

Table 1: The false positive and negative rates of the Wald Wolfowitz Runs test in detecting circadian genes from synthetic data. Data was produced based on previous literature, comprising of 48 hours of data, with data taken once per hour(Hughes, Hogenesch, and Kornacker 2010). The Wald Wolfowitz test was applied as per the methodology described within this manuscript is compared to the literature values of the false positive and negative rates of each other test compared to similar synthetic data.

	False Positive	False Negative
Wald Wolfowitz Runs Test	0.9%	0.0%
JTK Cycle	0.4%	1.8%
Cosopt	0%	29.6%
Fisher's G	0.2%	30.5%

3.2.5 Identifying synchronous pathway expression across tissues

Because each tissue has a different number of genes that are rhythmically expressed in each pathway, this means that if 100% of genes expressed in one tissue are coherently expressed in a different tissue this could represent a relatively small fraction of the total. Comparisons between the number of genes that are coherently expressed in two tissues to the number of genes expressed rhythmically in those tissues results in a relative distance, which scales poorly for comparing multiple tissues. We require an absolute distance measurement which will be the same regardless of how many genes are expressed in the tissues being compared. We define the absolute distance between two tissues as the difference between the total number of genes within the KEGG database and the number of genes that have coherent expression within both tissues. Agglomerative hierarchal clustering is then performed based on the unweighted average distance between tissues to identify how closely related the expression of a pathway is across multiple tissues.

3.2.6 Functional nodes

Multiple genes within the KEGG database are either orthologs or otherwise functionally related to one another. KEGG provides pathway maps that indicate mechanistic and functional relationships between different components. Many nodes within these pathways contain orthologs (such as AKT1, AKT2, and AKT3), however other nodes contain functionally related components (such as Clock and Npas2). Each pathway file is parsed for functional nodes to identify which nodes are activated by which tissue.

Fisher's Exact Test was used for each pathway to determine if the number of functional nodes was greater than the expected value. In order to calculate this, we determined the number of unique functional nodes listed in the KEGG database as well as the number of functional nodes in which at least 1 gene was expressed for each tissue. The fraction of functional nodes active in a circadian manner within one pathway was compared to the fraction of functional nodes identified in all pathways, with respect to each individual tissue type.

3.3 Identification of non-random genes

Of the 20,310 genes identified through the KEGG database, 13,512 were found to be circadian in at least 1 tissue type. Although 67% of the genes are expressed somewhere in the body in a circadian manner, the majority of these genes are not expressed throughout the body. Only 8 genes were identified to be circadian in all 12 tissue types: Npas2, Fmo2, Nr1d1, Nr1d2, Per3, Dbp, Arnt1, Tspan4. Of these genes, 4 of them belong to the Clock pathway, as defined by KEGG (Npas2, Nr1d1, Per3, Arnt1). Two others, dbp, and Nr1d2, may be part of the core Clock pathway but isn't part of the pathway on KEGG(Yamaguchi et al. 2000; Takahashi 2017).

Table 2:The number of circadian genes identified in each type of tissue. The number of genes in each tissue varies widely, and the fraction of unique. A larger number of gens being expressed in a circadian manner implies that the tissue performs a greater number of functions

Tissue type	Number of circadian genes	Number of unique genes
Brown Fat	2465	508 (21%)
Muscle	1943	462 (24%)
Liver	3484	967 (28%)
Lung	3703	1105 (30%)
Aorta	1603	316 (20%)
Adrenal	1742	327 (19%)
Brainstem	1344	283 (21%)
Cerebellum	1531	397 (26%)
Heart	2229	464 (21%)
Hypothalamus	2186	605 (28%)
Kidney	3041	705 (23%)
White Fat	1519	320 (21%)

The number of circadian genes varied widely between tissues, but many of these genes are not uniquely expressed within a single tissue. Anywhere from 19% to 30% of genes in each tissue are found to be circadian only in that tissue (Table 2). The fraction of unique genes does not strongly correspond to the number of circadian genes expressed. Liver and lung tissue are both highly expressive overall and have a high fraction of their genes uniquely expressed within the tissue. In comparison, cerebellum tissue is the second least expressive tissue, but over a quarter of the genes are unique to the cerebellum.

All identified genes had clear circadian patterns. Figure 12 represents a heatmap of all the genes identified in each of the 12 tissue types used. For each gene, there is a clear active (red) and inactive (green) period. Each tissue has its own circadian patterns as well, whether this is represented by a cascade of activity (such as in lung tissue) or more distinct tissue-wide active and inactive periods (such as identified in kidney tissue). This demonstrates that the Wald-Wolfowitz test used clearly returns circadian genes, without

relying on using population-level statistics or fitting individual traces to previously defined traces.

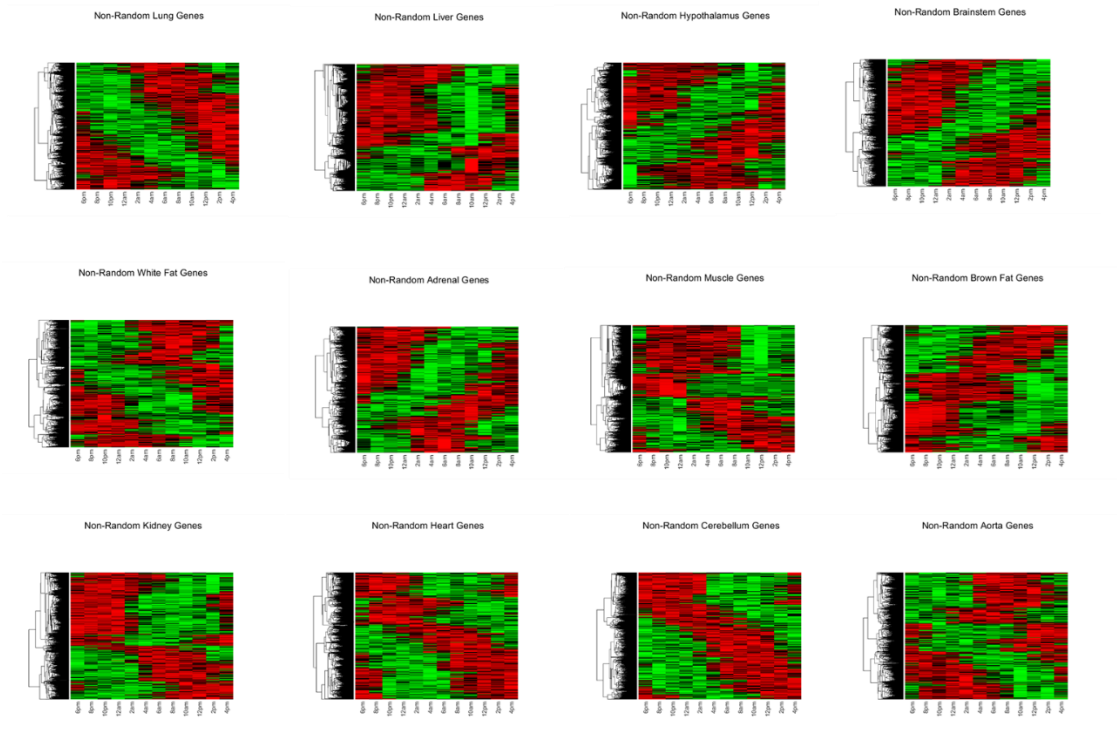

Figure 12: Heatmap of all non-randomly expressed genes in each of the 12 tissue types. All identified genes have clear circadian patterns with clearly identifiable on (red) and off (green) periods. This data is produced in wild type mice.

While most genes are not expressed in many tissues, the majority of the circadian genes expressed in each tissue are not unique to that tissue. One issue with comparing the expression of different tissues is that each tissue has different levels of expression: 391 genes are expressed in both white and brown adipose tissue, which represents over a quarter of the genes in white adipose, but only 16% of the genes in brown adipose. To compensate for this, we analyzed the absolute number of genes being expressed and compared it to the number of genes expressed in the genome as a whole, rather than comparing the shared fraction to the number of genes expressed in either tissue. Figure 13 shows the hierarchal clustering of tissues based on how many genes are expressed in

common, regardless of their function or presence on specific pathways. The three tissues with the most expressed genes (Lung, liver, and kidney) are closely clustered relative to the rest of the tissue types. These tissues also make up the 3 tissues with the largest number of unique genes.

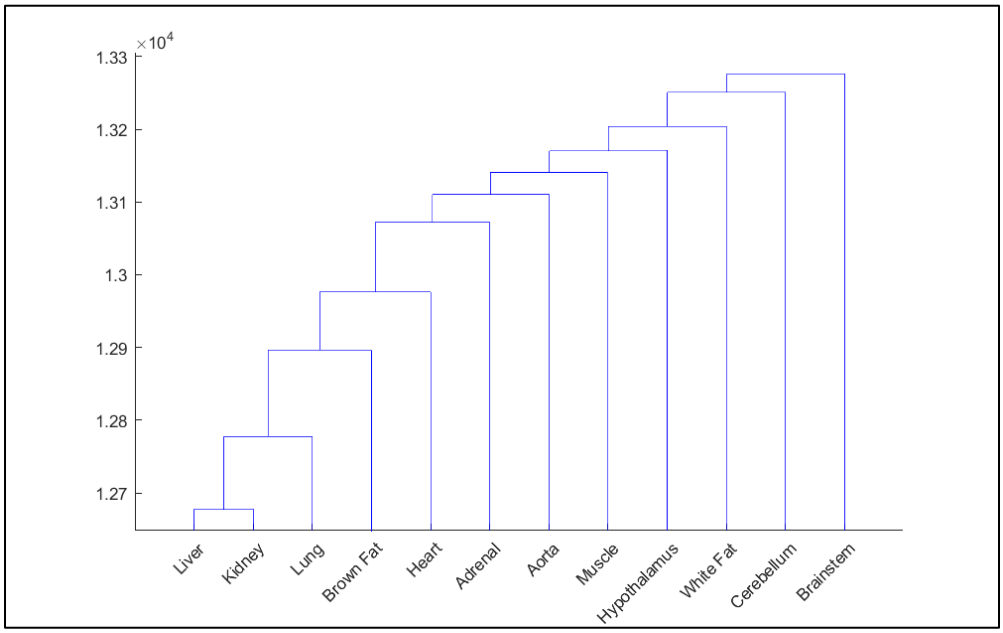

Figure 13: Hierarchal clustering based on the number of genes in common between the various tissue types. The liver and Kidney are the second and third most expressive tissues and have the greatest number of genes in common with one another.

The number of circadian genes each tissue pair have in common corresponds with which tissues are overall highly expressive (Table 3). The overlap between liver, lung, kidney, and brown fat tissue with one another are all high, but the overlap between these tissues and other tissues is high as well because of how expressive they are overall. Muscle and heart tissue has the most overlap outside of those 4 tissues with 392 genes being shared and 1943 and 2229 genes expressed in each tissue. Not all developmentally related

tissues have high overlap: brainstem and cerebellum share 14% of their circadian genes in common.

Table 3: Number of circadian genes expressed in each pair of tissues.

	Muscle	Liver	Lung	Aorta	Adrenal	Brainstem	Cerebellum	Heart	Hypothalamus	Kidney	White Fat
Brown Fat	380	600	637	514	414	250	289	478	314	603	390
Muscle		427	464	276	228	187	196	392	254	418	236
Liver			737	404	508	303	317	549	466	832	376
Lung				444	467	313	361	578	476	728	412
Aorta					290	168	186	337	214	408	279
Adrenal						191	210	336	259	465	247
Brainstem							188	232	266	319	163
Cerebellum								260	254	343	177
Heart									299	530	284
Hypothalamus										437	193
Kidney											336

Genes that are common between different tissues do not necessarily share the same periodic signal. Primarily tissues that are developmentally similar tend to be highly correlated. Tissues which do not appear to be related are not always dissimilar, but those are the tissues where more dissimilarities arise. Figure 14 shows the correlation of shared genes in four tissue types: liver, kidney, brainstem, and cerebellum. A high correlation indicates that the genes have a similar phase, while a negative correlation indicates that the signals are out of phase or even in anti-phase with each other.

While the genes which are expressed in both liver and kidney or both brainstem and cerebellum tend to be highly correlated, the genes which are shared by liver and brainstem tissue do not correlate as well and a larger share of them are close to anticorrelated with each other. This does not seem to correspond to the number of genes shared in common between tissue pairs: hypothalamus and liver tissue have more genes

in common but have a tendency to be less correlated than brainstem and liver tissue. Despite the presence of tissue pairs like hypothalamus and liver tissue, which do not have a positive correlation bias, there are no tissue pairs that tend to have more anti-correlated than correlated behavior for their shared genes (Appendix).

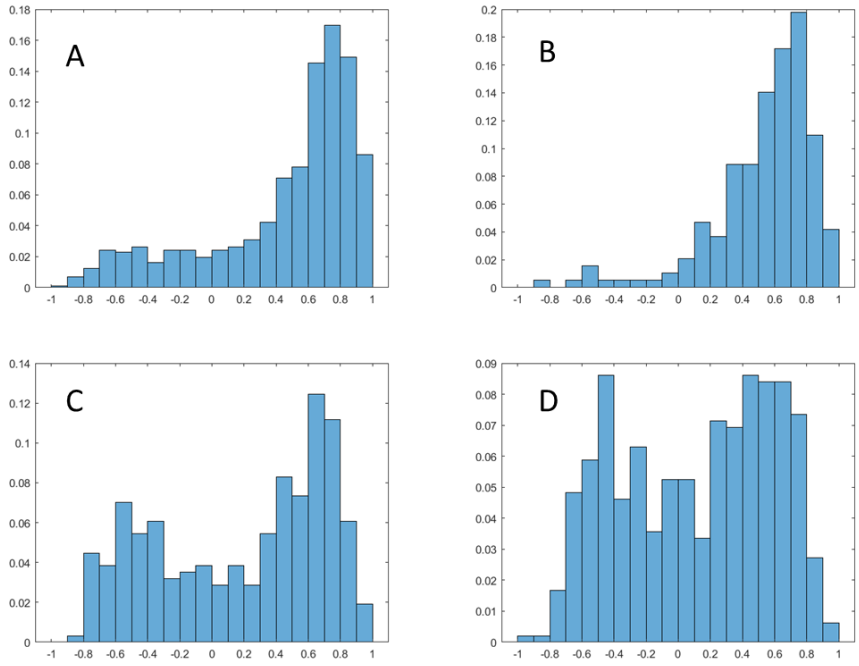

Figure 14: Correlation of shared genes between different tissues. A) Correlation between shared genes in liver and kidney tissues. B) Shared genes between brainstem and cerebellum C) Shared genes between liver and brain stem D) Shared genes in liver and hypothalamus. Even though there may be more genes in common between brainstem and liver tissue, these genes are not well synchronized with each other.

3.4 Pathway level circadian patterns

By placing circadian genes in the context of pathways we can see the time-dependent activity of pathways. Figure 15 shows heat maps of the core circadian rhythm pathway in all 12 tissue types. In brainstem tissue, four of the five genes present in the KEGG pathway are present in all other tissues. We can expect two major types of events within the pathway activity: cascade events where genes are activated consecutively resulting in a similar number of genes being active at all points of the day.

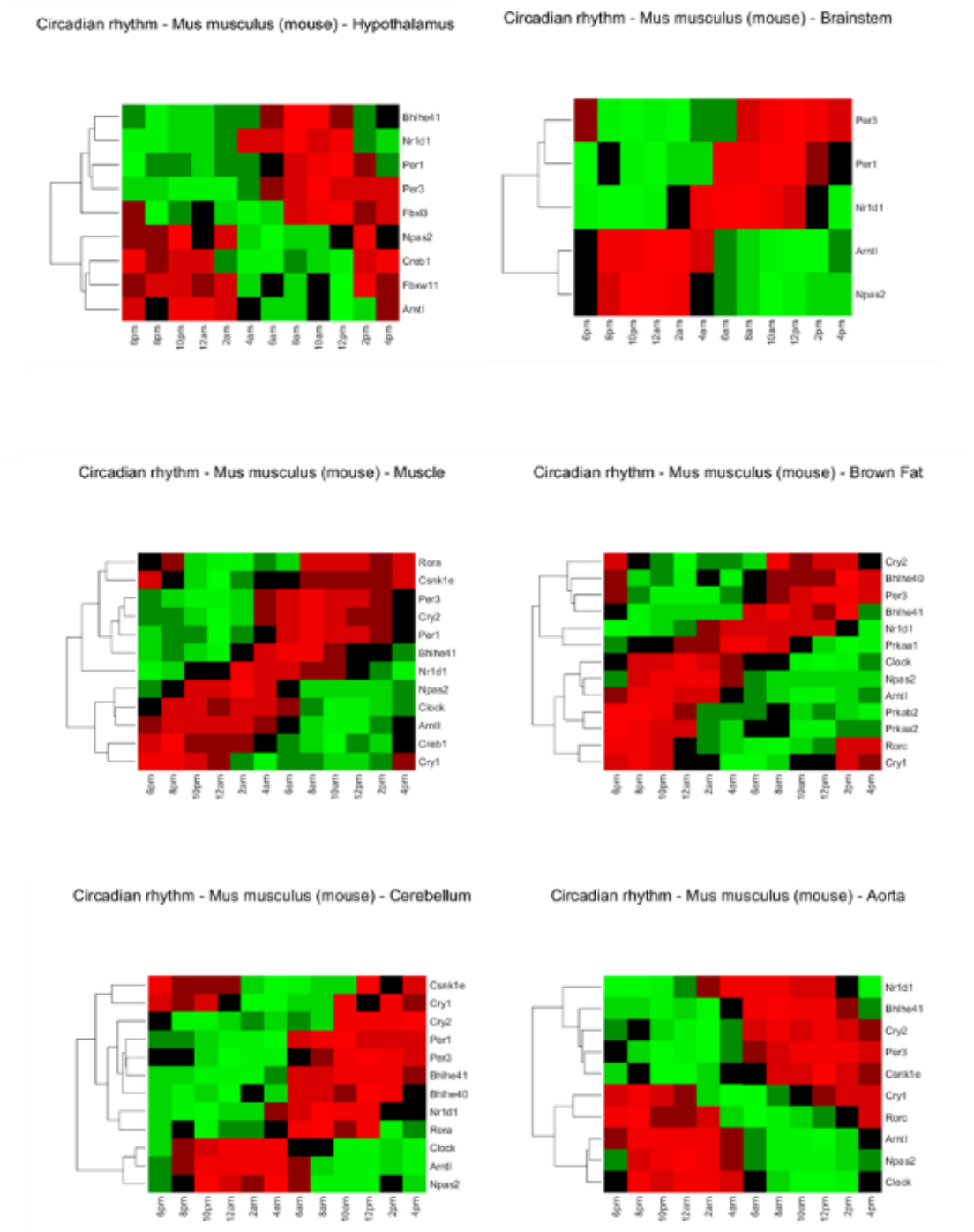

Figure 15:Heat map of circadian genes of the core clock pathways expressed in all 12 tissue types. The number of genes expressed varies broadly, with only 5 genes being expressed significantly in brainstem tissue. Interestingly, 4 of the 5 genes expressed in brainstem tissue are also expressed in a circadian manner in all other tissue types. In each tissue a clear cascade pattern is present.

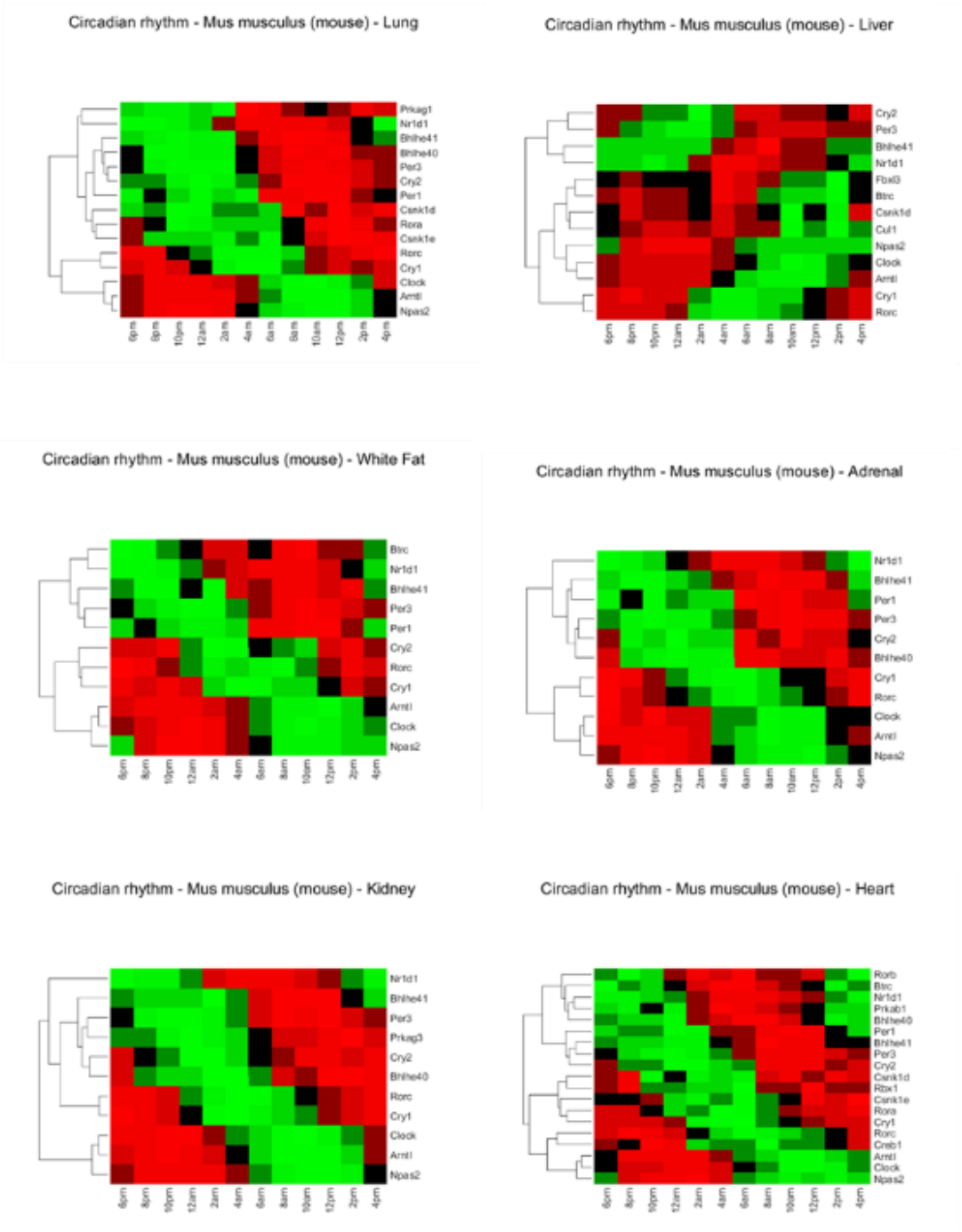

Figure 15 continued

Pathways other than the core circadian rhythm pathway also express largely oscillatory behaviors. These behaviors can be more diverse between tissue types, in terms of the number of genes expressed, how they are expressed, and which genes are expressed. We use VEGF as an example of the diversity commonly found within pathways. Figure 16 shows how the VEGF signaling pathway has a different number of circadian genes expressed in different tissue types, as well as different types of expression patterns in these tissues. These different patterns include qualitative behaviors such as cascade and rush hour type events as well as genes that have different phases in different tissues. Brainstem tissue has only 3 genes expressed in a circadian manner, compared to 16 genes expressed in the heart tissue. Even in highly expressed tissues, there are different qualitative behaviors such as cascades (liver and lung) and rush hour expression (muscle). These qualitative behaviors do not necessarily reflect the overall tissue wide behavior present in Figure 12.

Many of the genes expressed in each tissue are not expressed in all tissue types, which leads to tissue-specific behavior and expression. Even in the case of liver and lung in the VEGF pathway, there are only 2 genes in common despite a similar qualitative behavior: Pik3r1 and Hspb1. However, we find that not all genes expressed in different tissues appear unrelated. We can see that at least one form of AKT (Akt1, Akt2, Akt3) is expressed in 10 of the 12 tissue types, with Akt2 and Akt3 being expressed in liver and lung tissue, respectively.

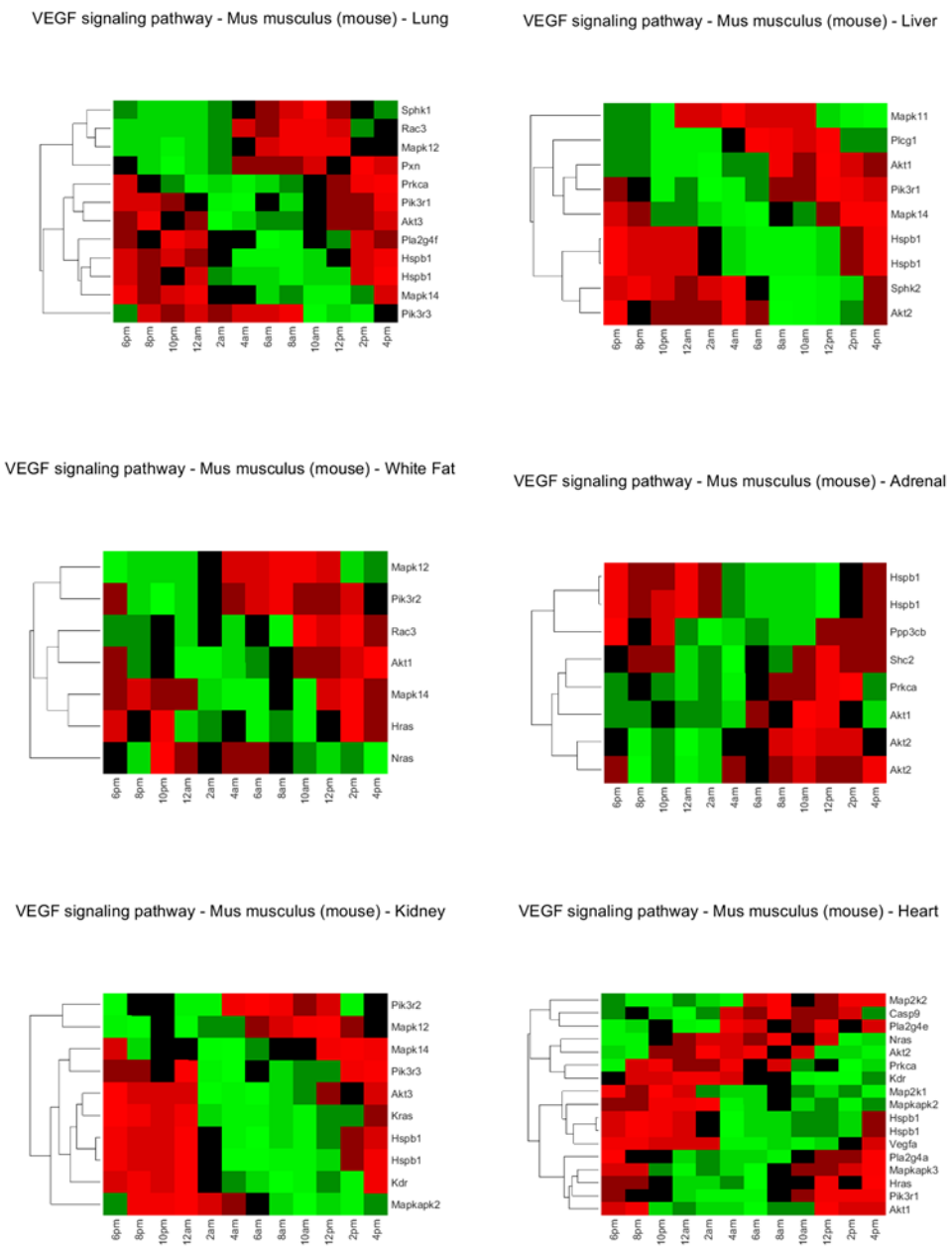

Figure 16: Heat map of circadian genes of the VEGF pathways expressed in all 12 tissue types. Pathways present several qualitative behaviors, including clear, steady cascades (liver and lung) as well as clear rush hour behavior where most of the pathway is active at the same time (muscle).

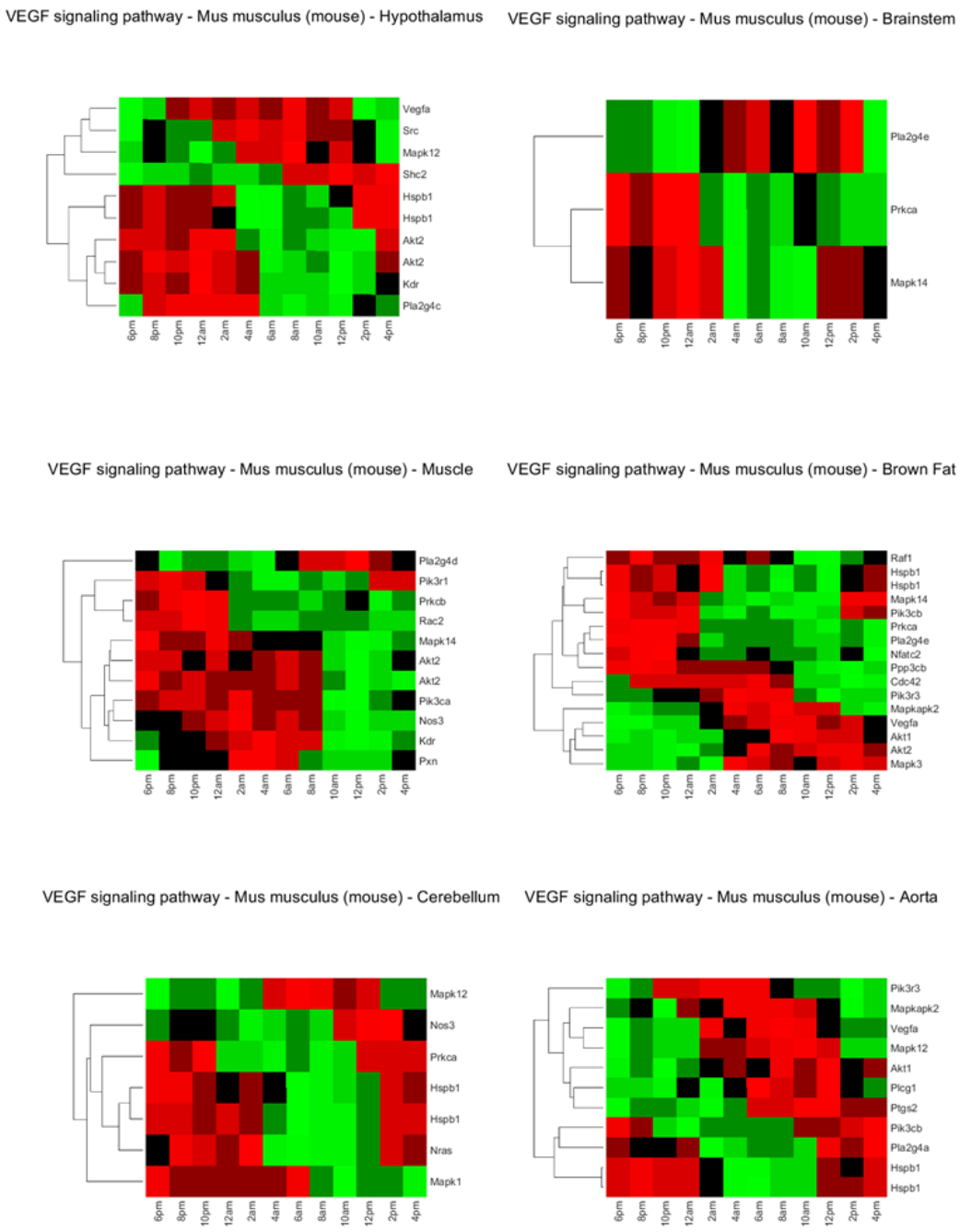

Figure 16 continued

Many genes within a pathway belong to functionally related groups, some of which are made up of orthologs (such as Akt1, Akt2, Akt3), although some functional groups contain genes that are not genetically related to each other (Such as Npas2 and Clock in the circadian pathway).

Figure 17 shows the calcium signaling pathway in both brown fat and lung. The top shows the heat maps of both tissues: There is a clear, circadian cascade in the genes expressed in both tissues. Despite 30 genes being expressed in brown fat, only 20 functional nodes are active.

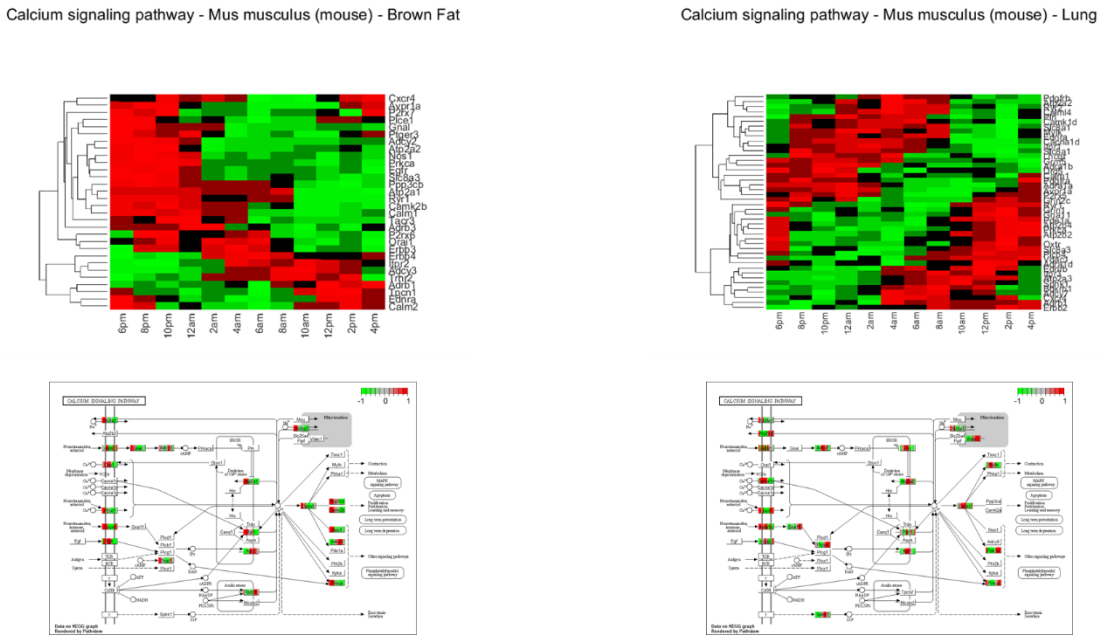

Figure 17:Calcium signaling pathway expressed in both brown fat and lung. Although many genes are different, many of the same "nodes" are occupied within the pathway. Additionally, despite many genes being represented in each tissue, only about 20 functional groups are active in either tissue.

In the case of the calcium signaling pathway, there are only 10 genes in common between brown fat and lung, but there are 13 functional nodes in common. In the previous VEGF example, there are 5 functional nodes in common between the liver and lung compared to

2 genes. Our future work will focus on understanding tissue relationships comparing functional nodes to detect tissue-specific function within different pathways in a more accurate manner. This will include measuring the phase relationships between tissues with similar functional groups. We will also seek to understand the number and function of overlapping functional groups between different tissues to trace relationships more accurately between tissue types.

3.5 Pathway type expression

There are 6 broad categories of pathway categorization for each KEGG pathway. Metabolic, genetic, environmental signaling, cellular regulation, organismal systems, and human disease. Two of these categories (organismal systems and human disease) contain tissue or disease-specific pathways. For that reason, those categories were dropped from the analysis, however, we retained the “endocrine” pathways which fall under organismal systems but were included in this analysis as well because they are less tissue-specific than other organismal pathways, which include subcategories such as digestive, nervous, and circulatory related pathways.

The number of pathways of each type varies widely. There are 91 metabolic pathways, 22 genetic processing pathways, 32 environmental signaling pathways, and 20 cellular process pathways. The endocrine pathways, consisting of 23 pathways. Relatively more endocrine ad environmental signaling pathways were expressed than any other type of pathway. 45% or more of the endocrine pathways are expressed in 10 of the tissue types and 46% or more of all environmental signaling pathways were expressed in 9 tissue types. In the cellular processes, 7 tissues had 40% or more of the pathways expressed. In comparison, only 1 tissue type had more than 20% of the metabolic pathways

significantly expressed (liver), and 2 tissues have 20% or more of the genetic information processing significantly expressed.

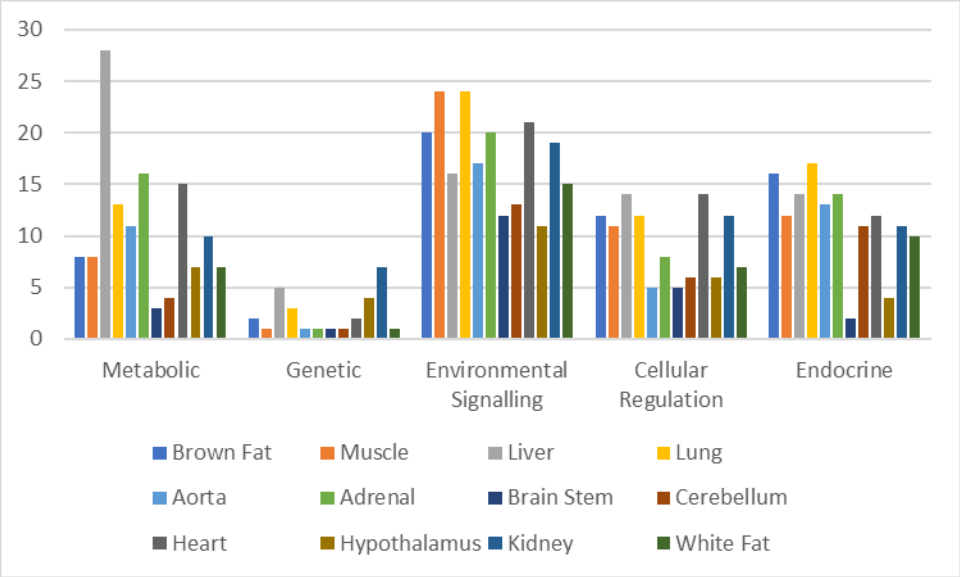

Figure 18:Overview of tissue-level expression of the 4 major categories of KEGG pathways which do not contain tissue-specific or disease-specific pathways.

Many tissues had a few pathways that were uniquely expressed in only that tissue. Most prominently this is found in liver tissue. Within metabolic pathways, liver tissue has far more pathways than any other tissue type. Liver tissue expresses 8 different lipid metabolic pathways, out of 15 pathways defined by KEGG, and expresses 5 different amino acid pathways out of 13 defined by KEGG.

Out of the 12 tissues studied, 8 expressed at least one unique pathway, although only liver tissue had more than 3 pathways uniquely expressed in it:

- Liver tissue significantly expressed 7 pathways that were not found in any other tissue types, 5 of which are metabolic: Autophagy – other, Citrate Cycle, Steroid hormone biosynthesis, Pyrimidine metabolism, Glycine, serine and threonine metabolism, Cysteine and methionine metabolism, DNA replication.

- Muscle tissue expressed 2 pathways not found in other tissues (Both were environmental signaling pathways): Cytokine-cytokine receptor interaction and NF-kappa B signaling pathways.
- Lung tissue was also dominated by signal processing pathways. Lung had 3 unique pathways not expressed in other tissue types: Hedgehog signaling, fructose and mannose metabolism, and RNA polymerase.
- Aorta tissue had 2 unique pathways expressed: Butanoate metabolism and the renin-angiotensin system. The renin-angiotensin system is an endocrine pathway responsible for blood pressure.
- Adrenal tissue uniquely expressed 3 pathways: beta-alanine metabolism, linoleic acid metabolism and ABC transporter pathways.
- Heart tissue has 3 unique pathways expressed: Tyrosine metabolism, other types of O-glycan biosynthesis, and aminoacyl-tRNA biosynthesis.
- Hypothalamus tissue has a single unique pathway: Oxidative phosphorylation.
- Brown fat, brainstem, cerebellum, and white fat tissue had no unique pathways associated with them.

Although there are only a few unique pathways expressed in each tissue, the functionality of many of these pathways is apparent. Liver tissue has an overexpression of metabolic pathways, and aorta tissue has an overexpression of the renin-angiotensin system. Within lung two odd pathways stand out: Hedgehog signaling and fructose metabolism, however both hedgehog signaling pathways and fructose have been identified for healthy lung function(Das, Neogi, and Steinberg 1984; Peng et al. 2015). Future research will focus on how different tissue types express the same functional groups within pathways and

explore the synchronicity of different tissue types, as well as possible tissue-specific expression patterns of functional groups. Because the functional grouping correctly identifies pathways relevant to tissue function, we hypothesize it will help more strongly identify functional relationships between tissue types.

CHAPTER 4: Inference of Boolean models

4.1 Background

Novel drugs or therapies which are developed at the bench-top levels often do not have the desired effect in the clinical setting. This has given way to a great interest in translational medicine: the ability to use basic research to develop clinically relevant treatments and move those treatments into clinical practice(Woolf 2008). Conventional drug development platforms are often focused on targeting single biomarkers or a single metabolic pathway, which is often hampered by the complex interactions between signaling networks, with interactions and cross-talk existing between many metabolic and genetic control pathways(Pietras et al. 1994). Even promising drug candidates run into toxicity or efficacy issues: Across big pharma, biological reasons lead to 43% of Phase II trials being canceled, while issues of safety, efficacy, or bioavailability lead to the cancellation of 50% of Phase I and 61% of phase II trials. (Morgan et al. 2012; Waring et al. 2015).

Metabolic and biochemical pathways often have complex interactions, with deregulation of one pathway often disrupting others. The resulting emergent behaviors of these interconnected networks are difficult to predict from studying the properties of individual genes or proteins. With these emergent properties in mind, one key component of translational medicine is to understand the influence local perturbations may have on the system as a whole(Huber et al. 2013).

Quantitative systems pharmacology (QSP) uses computational models to characterize and simulate these complex biological systems, as well as how these systems behave under

the effects of a disease or drug therapy(Rao et al. 2017; Androulakis 2016, 2015). QSP can help bridge the gap between early research and the clinical setting by providing simulating metabolic pathway deregulations, providing insights into the emergent properties of complex diseases, and predicting novel drug targets (Saez-Rodriguez, MacNamara, and Cook 2015). These models can help researchers develop hypotheses, predict emergency behaviors and can reduce the attrition rate associated with drug development(Leil and Bertz 2014).

Systems of ordinary differential equations (ODEs) are the most common way that QSP models are constructed. Each chemical or biological species (i.e. mRNA, protein, or pharmaceutical) is represented by a variable in the ODEs while the behaviors of these species, such as binding and dissociation rates, chemical kinetics, or transport rates across are defined by the equations which make up the model(Geerts et al. 2013). The mechanism of action for each interaction is taken into account and determines the form of each ODE, although this mechanism of action is often unknown in preliminary research. This problem can be compounded by sparse data in preliminary research, making it difficult to optimize the parameters of these equations.

Because of the amount of data required for ODE models, alternative approaches, and methods for reducing the data needed for a QSP model have been pursued. An alternative approach to ODE models is logic-based models, which can usually be constructed using sparse or qualitative data. These models do not necessarily require knowing the mechanism of reactions, which allows them to be rapidly developed and conceptually easier to interpret than their ODE counterparts. Logic-based network models can be inferred directly from empirical data, developed from a priori knowledge, or use a

combination of the two(Birtwistle, Mager, and Gallo 2013; Zhao, Serpedin, and Dougherty 2006; Chudasama et al. 2015).

Logic models rely on qualitative rules (such as Boolean logic equations) to define the interactions between different components, rather than using ODEs fit to data. The simplest version of a logic model would be a Boolean model, where the variables are limited to either a value of 0 or 1, and the relationship between these variables are described using Boolean logic (AND, OR, NOR). These models create a simplified version of complex networks, when compared to more detailed ODE models, but can provide insights into key regulatory mechanisms or potential drug targets, which allows researchers to gain insights and form hypothesis which can guide research(Le Novere 2015; Chudasama et al. 2015; Lu et al. 2015; Morris et al. 2010). The primary goal of many logic models is to understand potential regulatory mechanisms, identify novel drug targets, or to be used as tools for developing more detailed models, lending themselves to be used earlier in research than ODE models, but they do not provide continuous values of different biological species and may be limited in providing insight about dosing or timing of events.

Boolean models have a long history of being used as tools for simulating complex biological systems. Kauffman first envisioned Boolean network models in 1969 as a way of modeling homeostasis and gene regulation(Kauffman 1969). Boolean models were used to simulate the interactions between multiple components of a system without knowing the mechanistic nature of each component and could be used to study the emergent properties of complex or large networks(Kauffman 1984). Current Boolean models are used in a wide range of fields, with research on advancing the use of logical

modeling coming from all over the globe and organizations such as the CoLoMoTo consortium bringing together research groups with similar interests in the field.

Although Boolean models and ODE models contrast in their level of detail, Boolean models are useful in studying systems holistically while requiring less data to develop than a corresponding ODE model. Boolean models are created by using a series of Boolean algebra equations, similar to how more detailed models are created using a series of ODEs. Boolean algebra allows us to model qualitative information that we may be unable to use to create more detailed models from, by utilizing a number of logical operators (AND, OR, NOT, EXCLUSIVE OR). These logical operators are easily understandable, and potentially inferred from qualitative statements:

1) Glucose is formed in the presence of maltose AND glucosidase
2) Apoptosis occurs if TNF-α OR FasL are present
3) HMG-COA reductase converts HMG-CoA into mevalonic acid as long as Statins are NOT present

Because of the ease and flexibility of these logical statements, it is possible to construct Boolean relationships and models without significant amounts of quantitative data: simply knowing the qualitative relationships between components may be sufficient for creating rudimentary models.

Mathematically, each Boolean function has an output of 1 or 0 (True or false), depending on the input. For example, (A and B) would be equal to 1 if and only if both A and B were equal to 1. In contrast, (A or B) would be equal to 1 if either A or B, or both A and B, were equal to 1.

When these equations are used in a biological context, the values of 0 and 1 are decided in a somewhat arbitrary manner. 0 represents a “low” value of the component, while 1 represents a “high” value of that component.

Boolean networks consist of a series of variables that represent biological components, similar to ODE models. Boolean models utilize discrete time points, which represent events rather than absolute time points, and each variable has a discrete value usually represented as a 0 or 1. A Boolean network consists of *N* nodes, and each node is defined as a variable X_i. Each variable in the network is updated using the Boolean function, B_i:

$$X_i(t) = B_i(X_1(t-1), X_2(t-1) \dots X_n(t-1))$$

Where B_i represents a Boolean function, that defines $X_i\ (t)$ based on the state of the network at the time point ($t-1$). Several methods have been developed for interpreting the Boolean network, each using the same starting definition of a node.

Synchronous Boolean networks are deterministic state functions: The next state of the network depends solely on the current state of the network. the state of the network. Under a synchronous updating scheme, every variable is updated depending on its relevant functions. Because a Boolean network has a finite number of states, and because the next state of the network depends only on the current state of the network it will eventually reach a steady-state behavior. If the network repeats the same path multiple times, it can be described as a limit cycle, and if the network reaches a single point then this can be described as a fixed point. Other updating schemes, such as asynchronous Boolean networks, may exhibit more complex attractor behaviors.

4.2 Approach

In order to construct Boolean models from micro-array data, we utilize a series of statistical tests that determine genes that are differentially expressed in different conditions and cluster genes that have organism-wide similar expression patterns (inter-gene similarities). Clusters that are not coherently expressed across tissues (intra-gene, inter tissue comparison) are removed on the basis that we cannot differentiate between vastly different tissue-specific expression and noise with limited data. Coherent expression requires the genes which are being clustered to correlate with each other, which should indicate they have similar circadian amplitude and phase. Each remaining cluster has its expression profiles averaged, binarized, and is used to infer a Boolean network. This result will construct a core network structure that is widely representative of the gene network across multiple conditions and can be used as a basis for more specific model construction. The result of this pipeline will allow for the direct identification of these relationships, which may help identify where data is too sparse to determine mechanistic relationships (such as where many functionally unrelated genes have similar expression profiles), or may be used to determine which genes are sensitive to attacks. This sensitivity analysis is capable of being used to identify potential drug targets or to reduce the number of variables in the system prior to collecting data for more detailed ODE models.

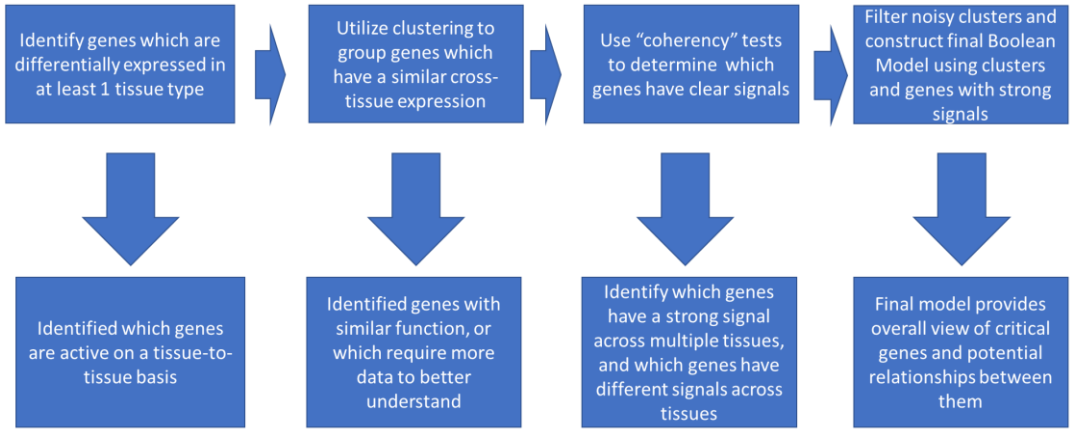

Figure 19: Overview of network construction. Differentially expressed genes are analyzed in the context of multiple conditions to determine which genes are the primary, universal drivers of network activity.

4.2.1 Differential expression

Differential expression is performed in order to determine which genes are differentially expressed in at least 1 condition. This is to ensure that all genes being considered are expressive on some minimal level. This serves primarily as a filtering step, and any genes rejected by differential expression should be rejected by the coherency test as well.

Differential expression was done using the Wald-Wolfowitz Runs test, as described in Chapter 3, with the exception that if a gene was detected in any tissue type it was included in further analysis. The primary advantage of the runs test is that since it is wave-form agnostic and does not depend on variation from a population level average, it can be performed on any number of genes.

4.2.2 Clustering

Genes with highly similar expression profiles will become indistinguishable after binarization. This will lead to an increase in the uncertainty of the final model, and any small differences in the profiles of co-expressed genes may result in networks that show

the possible relationships of one but not both genes. Clustering allows for multiple genes to be grouped together based on their expression profiles, which will reduce the number of variables in the final model and allow for a clearer understanding of the underlying relationships within the network.

Clustering has been used as a method for identifying similar gene expression profiles, as well as specifically used as a method of variable reduction in Boolean model reconstruction(Martin et al. 2007; Bonneau et al. 2006; Guthke et al. 2006; van Someren, Wessels, and Reinders 2000). A *k*-means clustering technique is used to group genes together by their expression profiles. In order to capture similarities across all conditions, the expression vectors for each condition are stacked. For example, if a gene has an expression profile with a length of *T* and has data from 12 different conditions, the gene will use a vector of length 12*T* in the *k*-means clustering.

The resulting clusters are then tested for internal consistency, both between genes within the cluster but between each condition for each gene.

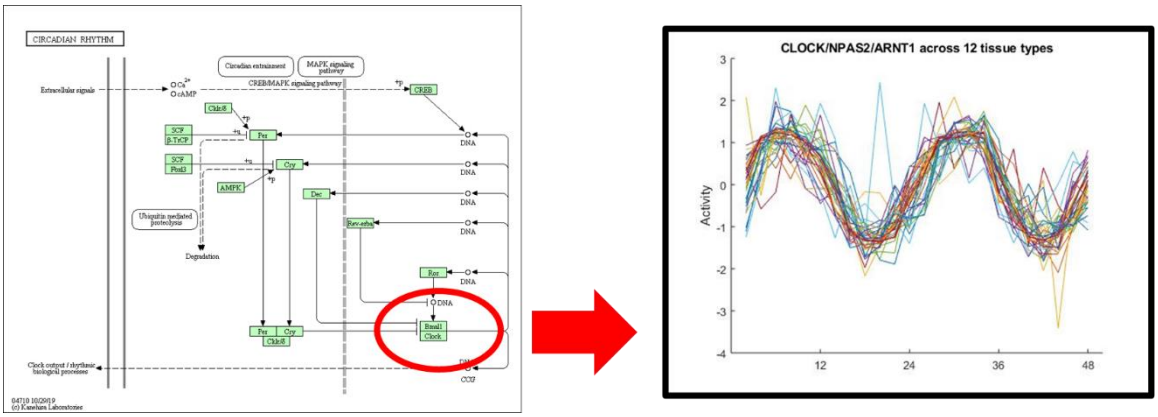

Figure 20:Clustering of multiple genes into a single node. Left shows a functional node, identified in KEGG. Right shows the 48-hour trace of all 3 genes in the node across all 12 tissues. Although some time points do have significant noise, overall, there is significant agreement across all 12 tissues for all 3 genes. Given the limited data available, these 3 genes are indistinguishable from each other.

4.2.3 Coherency

After inter-gene co-expression is used for variable reduction, an intra-cluster (and intra-gene) approach is used. Each expression profile within a cluster is correlated with each other member of that cluster to calculate a distribution of pair-wise Pearson correlation coefficients. This correlation is done not only between genes but between the number of conditions for each gene (for example, CLOCK expression in the liver is correlated to CLOCK expression in the lung). A background set of data is created by scrambling the clusters which were previously identified, and each cluster's distribution is compared to the scrambled dataset using Tukey's Honest Significant Difference test(Tukey 1949). If a cluster is found to be statistically similar to the scrambled data then the variance between conditions may either be due to noise or by local factors, but that cluster is not part of the core network structure of the pathway and is removed from the analysis.

4.2.4 Constructing a Boolean model

The classical Boolean network is a synchronous, deterministic model: All variables are updated at every time point. This means that at each time point it is known what variables are being updated, and at each time point, all variables are updated.

There are two ways this can be modified: a model can be made asynchronous by having different variables updated at each time point. It can also be made non-deterministic, where the variable or variables updated at a time point are not known (i.e. they are selected randomly at each time point).

A deterministic, asynchronous network will have each node updated in a specific, known order(Greil, Drossel, and Sattler 2007). In this way, not every node is updated at the same time, but the model output is the same every time the model is run.

There are multiple non-deterministic, asynchronous update schemes for Boolean networks. The most common are the random asynchronous method, wherein all nodes are updated in random order, and the general asynchronous method where a single node is randomly selected at each time step(Harvey and Bossomaier 1997).

Other modifications to the update scheme can be made to accommodate biological realities. One of the most common is time delays, such as having a node's value depend on the state of its regulating node from more than 1-time step previously(Klemm and Bornholdt 2003).

Boolean networks can be developed using qualitative or quantitative knowledge. Relationships between biological components can be determined through literature-text searches, qualitative data, bioinformatics sources, or high-throughput assays. While literature searches may be labor-intensive, pathway databases such as KEGG or WikiPathways can provide the framework of a model very quickly(Kutmon et al. 2015; Kanehisa and Goto 2000).

Whether it comes from literature or pathway databases, a qualitative understanding of a system can be interpreted as an interaction graph: a simple network of positive and negative connections between variables. The qualitative prior knowledge network (PKN) can then be systemically converted into a Boolean network, either following a simple algorithm or based on the modeler's understanding of specific mechanisms. The most common rules for biological systems assume that all activators are related by OR functions and that any single inhibitor is sufficient to turn a node off (Krumsiek et al. 2010; Flobak et al. 2015). These types of canalizing functions have been used to develop genetic regulatory pathway models, and are based on biological feasibility(Harris et al.

2002). Many researchers may rely on applying these rules to the PKN themselves, rather than relying on algorithms to develop the Boolean network based on the qualitative data(Lu et al. 2015; Chudasama et al. 2015). This may introduce some bias in the development of the network, especially if the relationships between genes appear complex or difficult to reduce to an AND or OR gate.

Quantitative biological data needs to be processed before it can be used to develop a logic model. The first process is called binarization, where the protein or mRNA concentrations are normalized between 0 and 1. The concentration data is broken into two clusters, with the higher-valued cluster having a value of “1” and the lower value having a value of “0”. Methods for clustering data in this manner are well established(MacQueen 1967). The largest drawback of these methods is the loss of continuous data and information and non-deterministic binarization methods producing different results from the same data set (Mircean, Tabus, and Astola 2002). Most Boolean models assume a step function between two variable states represented as a 0 and 1. If a species has more than 2 steady-state values, or the transitions between these two values are not rapid information about how a variable transitions from one state to the other may be lost during binarization.

The second step of data processing involves removing redundant data points. If time course data is used, the kinetics of certain reactions may be slow enough that consecutive time points have the same value after binarization. In a Boolean network, if two consecutive time points are identical then the system has reached a steady-state. Because of this, time points that have the same values after binarization, but do not represent the

steady-state condition of the biological system, need to be removed(Erkkila, Korpelainen, and Yli-Harja 2007).

4.2.5 Binarization

After the gene groups to be used for model construction have been determined, each cluster is averaged. Values below the average are assigned a value of "0" and values above the average are assigned a value of "1", creating a threshold value for each cluster based on its mean(Martin et al. 2007). Various methods for binarization exist for grouping data into "high" and "low" values. Thresholds can be determined in a number of ways, such as by determining an arbitrary fold change in expression levels(Serra et al. 2007) or by observing discontinuities in step sizes between time points(Hopfensitz et al. 2011). We utilized thresholding based on the z-scored data of each profile. Values that were above the mean were assigned a value of "high" and those below the mean were assigned a value of "low" (Figure 21).

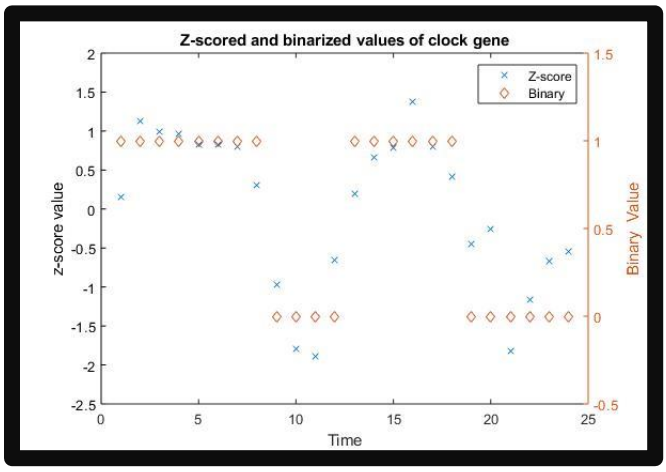

Figure 21: Example of thresholding binarization. The CLOCK gene expression data is binarized based on a threshold determined by the average expression of the gene.

4.2.6 Network inference

The final Boolean model is constructed using the Best-Fit Extension algorithm(Boros, Ibaraki, and Makino 1998; Shmulevich et al. 2001). For the application of this algorithm, redundant time steps are removed before the network inference. A time step is considered redundant if $X(t)=X(t+1)$ for all variable values. In this case, the time step $t+1$ is removed. Methods exist to predict how many bits should change before a transition is considered "significant", however for our purposes, we only removed identical values(Erkkila, Korpelainen, and Yli-Harja 2007).

The Best-Fit Extension was used because of its ability to learn networks even with insufficient or noisy data, while other algorithms (such as REVEAL(Liang, Fuhrman, and Somogyi 1998)) are only able to learn a network if it is error-free. The Best-Fit Extension has similar dynamic and topological accuracy compared to more recent algorithms developed(Ruz, Zúñiga, and Goles 2018; Simak, Yeang, and Lu 2017). These other algorithms can include neural networks, where each edge is given a weight and the weight of each relationship is learned over time(Ruz, Zúñiga, and Goles 2018), or by testing each gene connection individually in order to determine the likelihood of a direct or indirect relationship between genes

For Best Fit, a maximum number of regulators (K) is selected to limit the computation time as well as limit the number of candidate equations which have the same error. For each variable X_I a selection of regulators $X' = (X_1 \ldots X_n)$ where $n \leq K$. A partially defined Boolean function $pdBf(T, F)$ where T is the set of unique states of the set X' at t when the variable $X_i(t+1)$ is True, and F is set of unique states of the set X' at t when the variable $X_i(t+1)$ is False. The error associated with a set X' is the number of inconsistencies within

the pdBf. The error is equal to the sum of the number of sets of values of the selected regulators X' which do not properly define X_I according to the data set available. The algorithm returns the smallest set X' which has the smallest error, discarding larger sets.

Figure 22 shows how the data is sliced into discrete time points. These slices are then used in the Best-Fit-Extension algorithm to infer potential Boolean relationships for each variable. In the case of Npas2, 6 unique Boolean equations are inferred. Several relationships may be qualitatively apparent (such Bhlhe41 being present in every potential function), and the equations may be used in a probabilistic model. However, clustering and inclusion of multiple data sources will reduce this number as discussed later.

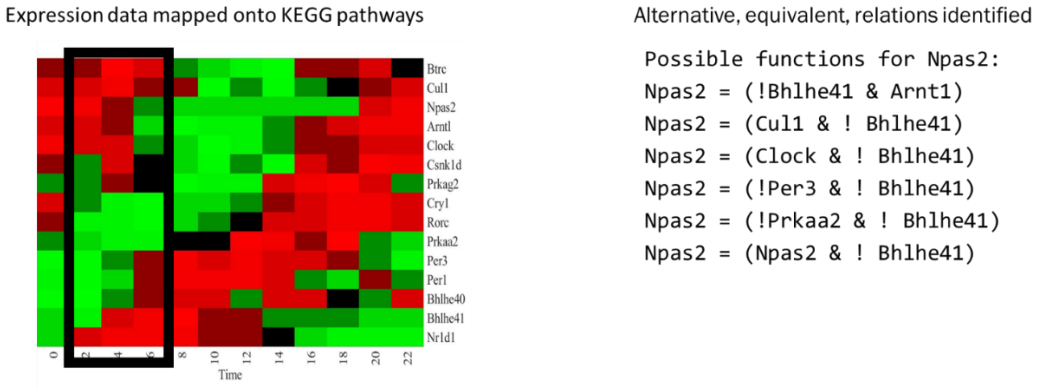

Figure 22: The raw expression data is sliced into transitions which are then used to infer Boolean equations (highlighted on the left). Based on these sliced transitions an exhaustive search finds the "best fit" Boolean equation which requires the fewest possible set of inputs. There may be equally likely possible equations (right), each of which is interchangeable with each other. This does not produce a single model, but rather a set of possible models.

4.2.7 Sensitivity analysis

The identification of components of a Boolean network which are both necessary and sufficient to describe the qualitative behavior of the network, such as oscillations or response to environmental factors, has previously been discussed in the literature. We can consider the subnetwork which is descriptive of the overall behavior of the model as the

"core" network, while nodes outside of this subnetwork can easily be reduced out of the model. Various structural methods to reduce networks without performing sensitivity analysis to the networks have been proposed, but generally are insufficient in preserving components of the network which may be affected by transient perturbations (Shmulevich and Kauffman 2004; Naldi et al. 2009; Saadatpour, Albert, and Reluga 2013; Veliz-Cuba 2011; Chudasama et al. 2015). More traditional sensitivity analysis of Boolean networks generally involves perturbing the network and studying how many nodes are affected by the perturbation or how long it takes for the network to return to its attractor(Shmulevich and Kauffman 2004). One of the issues of these techniques is that not all nodes within a biological network are equally important(del Rio, Koschützki, and Coello 2009).

Similar to the stuck-at-fault mutations in Chapter 2, sensitivity analysis is done by fixing the value of a single node. If this stuck at fault attack affects the value of any other node in the system, then that node is considered removable. After the node is removed, the attacks are performed again. In this way, nodes can be removed sequentially, so that a chain of nodes which have no impact on the network at large can be removed. Unlike Chapter 2, we do not know what nodes are physiologically important to the function of the network and therefore must assess the change in the output of each node within the network rather than only focus on the output node.

4.3 Boolean Network of the Circadian system

It is possible to infer networks from a single data set from a single tissue, but this presents numerous challenges. Figure 23 shows the result of the network inference, representing a single candidate network identified through the network inference.

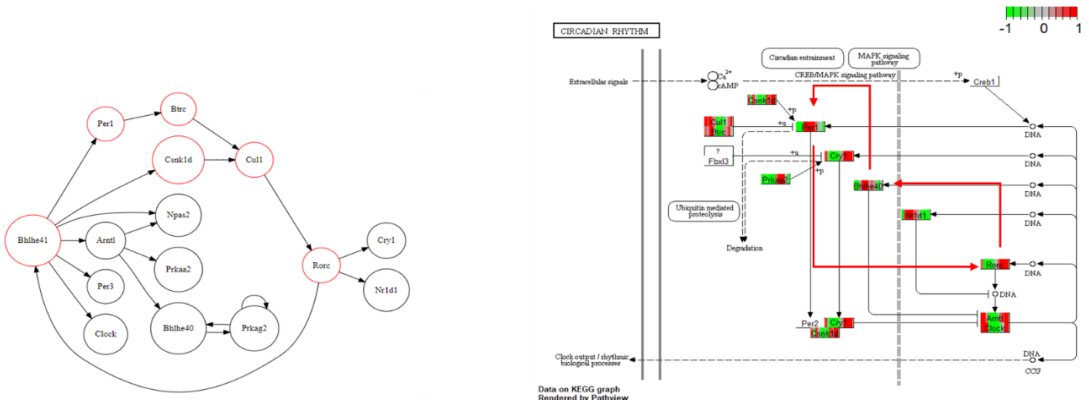

Figure 23: One candidate Boolean network inferred from data from the liver tissue only, demonstrating the discovery of connected feedback loops from the processed data. Although there is a multitude of theoretical networks discovered, the presence of interconnected feedback loops is present in many of them.

Overall, multiple genes had nearly identical expressions, and other molecules were noisy and inconsistent. This led to a total possible 10^16 candidate networks. It is possible to use these candidate networks as a single, probabilistic Boolean network, however, due to the noise present in some of the gene data within a single tissue it becomes unreliable to use all genes with only a single data set. Some of the relationships identified can be confirmed in previous literature (Blhe41 promotes Per1 and Per3), other relationships are either indirectly known (Rorc interacts with Clock which interacts with Blhe41, Rorc does not directly act on Blhe41) or unknown altogether. Interestingly, the candidate networks consistently return nested feedback loops, despite the multiple possible solutions based on the data.

To improve accuracy and reduce noise within individual variables, multiple genes are clustered into gene groups. The data from genes grouped in this way are highly

correlated, although the genes are not necessarily functionally related. However, often these grouped genes do relate to one another.

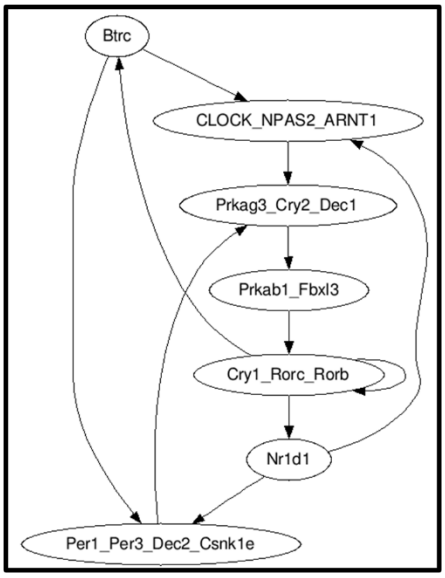

Figure 24: Network inferred from a single tissue, with clustering

By grouping similar genes together, we limit the ability for noise to change 1 or 2 bits within our data, which may upset the inferred model or create uncertainty. After clustering the genes and inferring a network, there are 32 potential candidate networks. However, all 32 candidate networks reduce to the same, identical network seen in Figure 24. Many identified clusters have known, functionally related genes (Clock, Npas2, and Arnt1)

Incorporating multiple tissues require all or most of those data sources to have an expression of the genes being modeled.

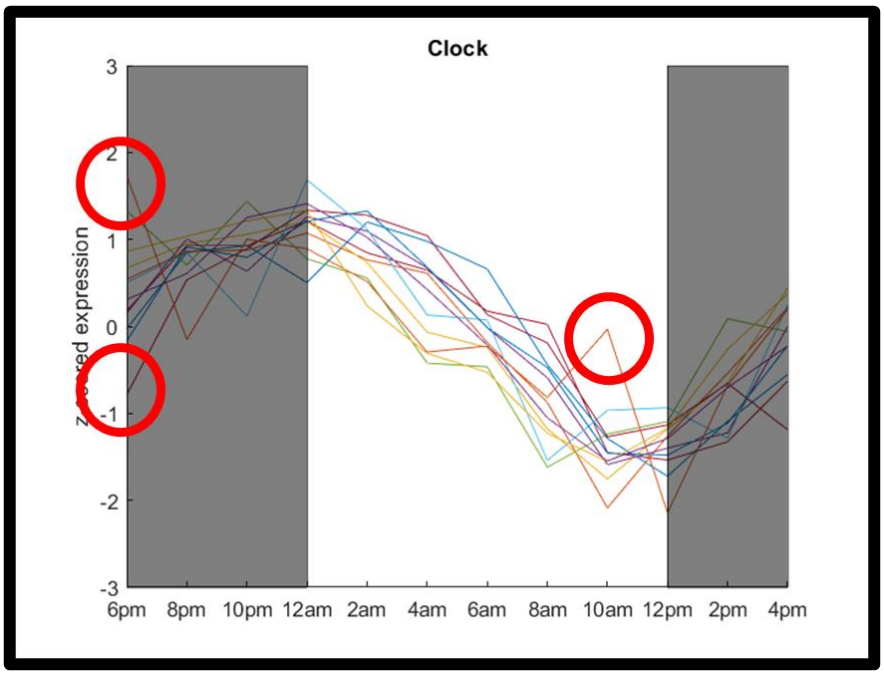

Figure 25:Expression of the clock gene in multiple tissues

Although there is a clear, consistent signal across all 12 tissue types, there are still "hot spots" of noise where a value of 1 may be changed to 0 or vice versa during binarization (Figure 25). As more relevant data sources are averaged together, the probability of noise being an influence on the output model decreases.

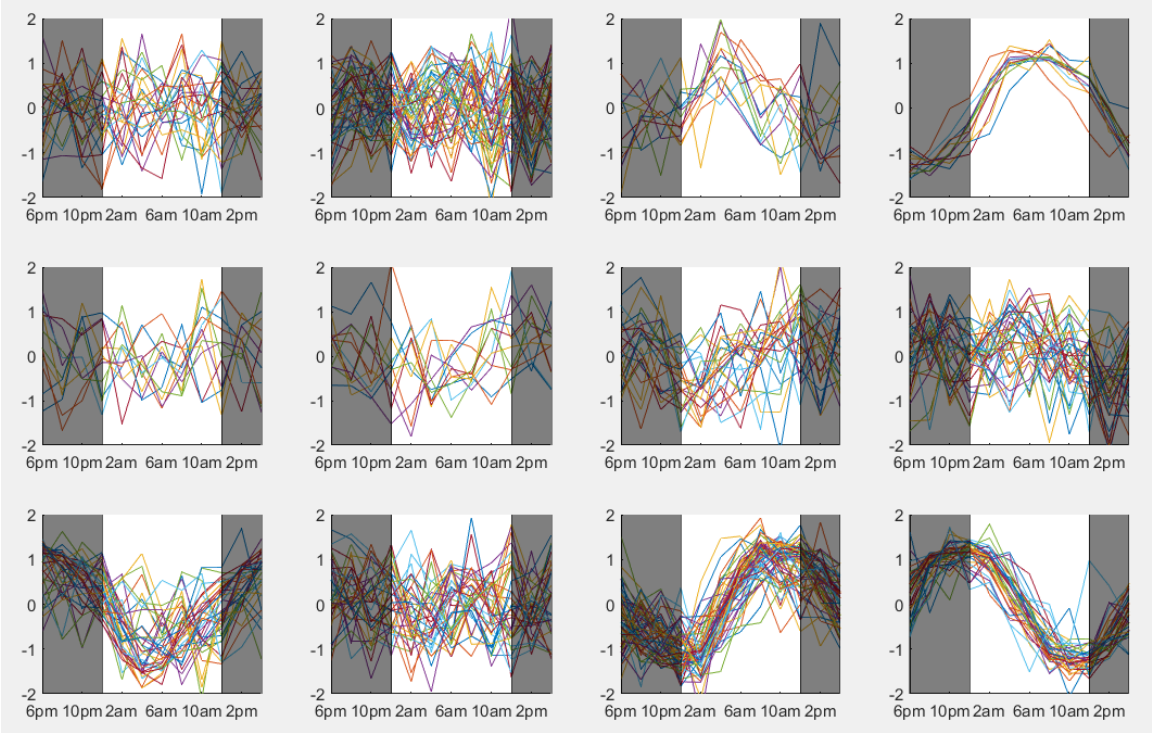

Figure 26: 12 clusters of genes are identified. Only 4 clusters have a positive correlation, indicating that the remaining clusters are either tissue-specific or too noisy to be considered for model development.

Clustering was used to group multiple genes across all 12 tissues. The correlation of the genes in the cluster was then determined, with the distribution of pairwise correlation coefficients of each trace being considered how coherent the cluster is. If a cluster consists of a single gene it will have 12 traces, one from each tissue. If a cluster consists of two genes it will contain 24 traces: one from each tissue per gene grouped together. In this way, if only a single gene is contained in a cluster it may still not be coherent if that gene is expressed inconsistently across the 12 tissues. Our clustering analysis identified 12 potential clusters, however most of these clusters were dominated by noise. Although some individual profiles within each cluster may have been circadian, overall genes within those clusters did not correlate with themselves across tissue types or with the other grouped genes. Of the 12 identified clusters, 4 of them show clear circadian patterns (Figure 26).

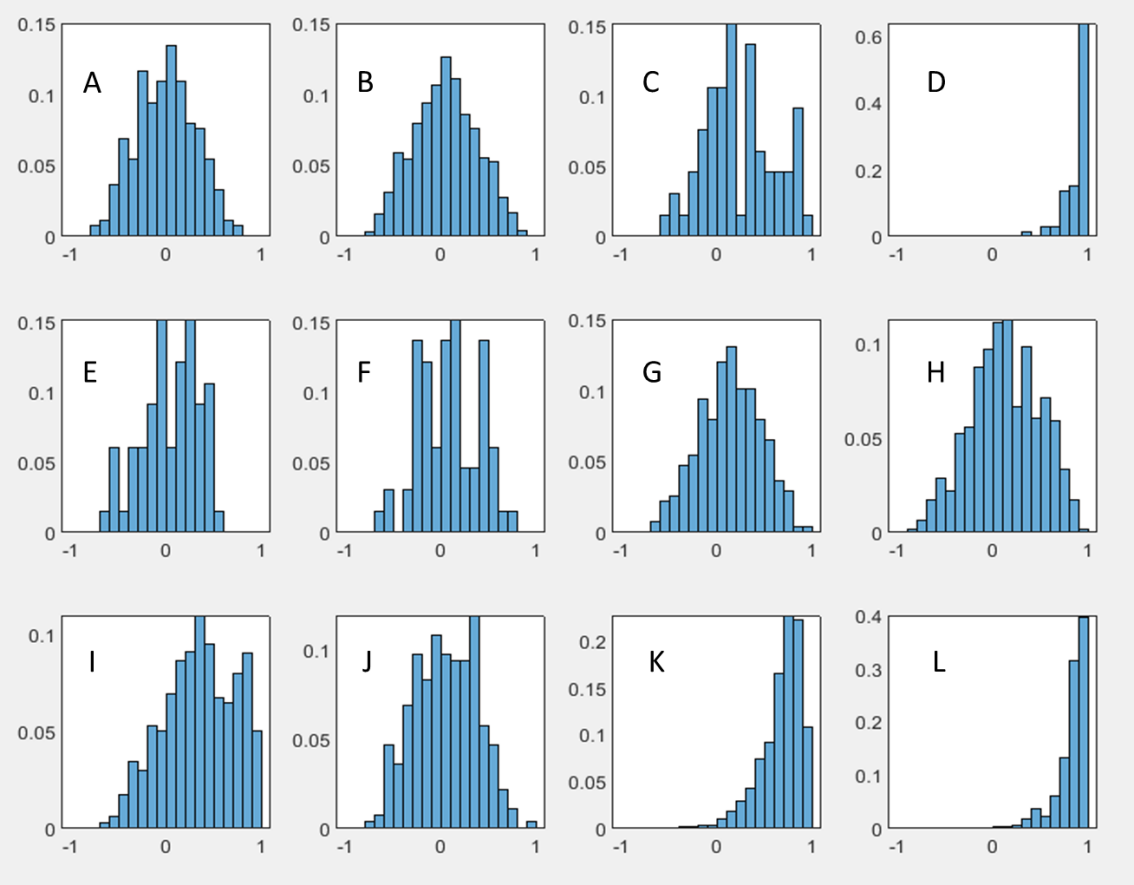

Figure 27: Pairwise correlate of each identified cluster. . Clusters D, I, K, and L are all identified as having a positive pairwise correlation with each identified trace. Cluster C appears qualitatively coordinated, however many traces were found to not correlate with each other.

The pairwise correlations of each cluster tend to show an average correlation of close to 0 (Figure 27). Clusters D, I, K, and L all have distributions of pair-wise correlation coefficients which have a greater median than the other distributions. This median correlation coefficient is compared using a Tukey's Honest Statistical Difference test to detect which distributions have a higher correlation, which corresponds to a higher internal consistency of the cluster identified.

$$C1(t) = C4(t-1) \qquad \text{Equation 1}$$

$$C2(t) = C3(t-1) \qquad \text{Equation 2}$$

$$C3(t) = C2(t-1) \qquad \text{Equation 3}$$

$$C4(t) = C1(t-1) \vee C2(t-1) \vee C3(t-1) \qquad \text{Equation 4}$$

These 4 clusters were then used to construct a final model, which consists of Equations 1-4. These equations form the network identified in Figure 28. One of the important factors concerning the model is its accuracy and proper representation of the system. All interactions between clusters are accurate when compared to the KEGG pathway, although some relationships are missing.

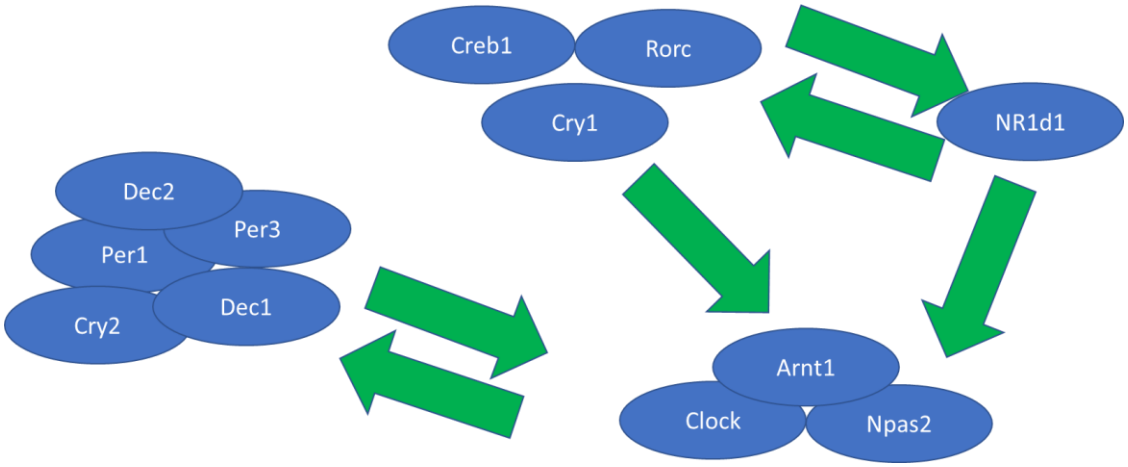

Figure 28:Final inferred Circadian model constructed from multiple tissues. All green arrows represent causal relationships, and all identified relationships are biologically known.

Certain interactions are well established, such as the functional link between Per1 and Per 3, as well as the connection between Per and Cry genes. The identification of multiple, linked feedback loops within the circadian system is notable, and reflective of the known behavior of the circadian rhythm pathway. The relative importance of internal regulation of the clusters (such as any regulation between Dec1/2 and Per1/3) is missing. The

utilization of multiple data sources enabled the elimination of noise, however, given the limited diversity in expression within the pathway this resulted in an accurate, but small, model. Many of the groupings identified are also identified as functional groupings from Chapter 3: Dec1/2, Per1/Per3, Clock/Npas2. Cry2 and Cry1 were identified as being grouped separately, though, and our future work on exploring functional groups will emphasize exploring the phase relationships within groupings, both within tissues and between tissues. Future research for the model building will focus on exploring the use of knock out data sets to improve the detail within the model making process. Knock out experiments would be able to improve the detail within the model by changing the number of state transitions that exist for each node, improving the amount of data available. Knocks out experiments would also allow for mathematically identical genes to be assessed individually: If Per1 was knocked out and there was a change in expression in one gene in the Clock/Arn1/Npas2 cluster, then the function of Per1 may be separated from the function of Per3 and the rest of the cluster. Being able to quickly construct a model in this manner can be utilized in two primary ways. The model can be utilized directly, allowing for the identification of drug targets or combinations of drugs which may lead to novel emergent responses from the system(Lu et al. 2015). This is achieved by perturbing nodes within the network and identifying how the other species in the model react. If a perturbation leads to the activation or inhibition of a gene or protein critical to the progress of a disease then the perturbed node may be a valid drug target and can be explored in more depth. Alternatively, these networks can be used in a larger model building pipeline: being able to determine which species need additional data, or being able to perform sensitivity analysis to reduce the number of species in the model

before constructing more detailed ODE models may help guide the construction of those networks(Chudasama et al. 2015).

CHAPTER 5: Conclusions

Herein we have presented three perspectives to identifying the relational structure of biological networks. Our evolutionary algorithm shows the origin of biological network formation, and how different factors lead to many disparate components forming a coordinated pathway (Chapter 2). Our results have key biological implications in the study of evolution and how selective pressure for phenotypic behavior interacts with the more random mutations and non-adaptive forces which lead to generation to generation changes. Our study on health mouse transcriptomic data reveals a greater understanding of circadian and oscillatory signals within mammals (Chapter 3). There are clear tissue and organ-specific effects, with different genes and pathways being expressed in a circadian manner in different tissue types The presence of highly correlate as well as anti-correlated genes within different tissues indicate that although many genes are not universal, there is still some level of shared expression within tissues. Most interesting is the view of functionally related genes still being expressed in different tissues within the same pathway, even when there appears to be a low overlap of specific genes. We were able to identify pathways within certain tissue types that were strongly tied with the physiological function of each tissue they were found in. This understanding of functional groups was extended into the creation of a model of the pathway, using the circadian rhythm pathway as an example. This highlighted the difficulty of using *in-vivo* data, as there was consistently high noise throughout the data set which brought unique challenges to inferring a Boolean model from the data (Chapter 4). Utilizing multiple data sources, as well as using variable reduction techniques can remove false relationships that were identified when the model was created from a data source. Further research into

using knock out data and potentially creating tissue-specific models will be done to help understand how different pathways are regulated and how circadian patterns result in different behaviors throughout the same organism. Understanding the circadian networks within complex metabolic and signaling pathways can be used to identify potential drug targets or for the purpose of building more detailed models of those systems by establishing a framework and identifying potentially unknown relationships that must be identified.

Acknowledgment of Publications

This dissertation contains significant portions of the following publications:

Putnins, M. & Androulakis, I. P. 2020 ‘Self-selection of evolutionary strategies: adaptive versus non-adaptive forces.’ *Evolutionary Bioinformatics* (submitted)

Putnins, Matthew, and Ioannis P. Androulakis. 2019 ‘Boolean Modeling in Quantitative Systems Pharmacology: Challenges and Opportunities.’ *Critical Reviews™ in Biomedical Engineering* 47:473-488

I, Matthew Putnins, am the original author of this material.

References

Acevedo, Alison, Ana Berthel, Debra DuBois, Richard R Almon, William J Jusko, and Ioannis P Androulakis. 2019. 'Pathway-based analysis of the liver response to intravenous methylprednisolone administration in rats: Acute versus chronic dosing', *Gene regulation and systems biology*, 13: 1177625019840282.

Albrecht, U. 2012. 'Timing to perfection: the biology of central and peripheral circadian clocks', *Neuron*, 74: 246-60.

Alon, U. 2007. 'Network motifs: theory and experimental approaches', *Nature Reviews Genetics*, 8: 450-61.

Amadoz, Alicia, M Hidalgo, Cankut Cubuk, José Carbonell-Caballero, and Joaquín Dopazo. 2018. 'A comparison of mechanistic signaling pathway activity analysis methods', *Briefings in Bioinformatics*, 20: 1655–68.

Androulakis, I. P. 2015. 'Systems engineering meets quantitative systems pharmacology: from low-level targets to engaging the host defenses', *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 7: 101-12.

———. 2016. 'Quantitative Systems Pharmacology: A Framework for Context', *Current Pharmacology Reports*, 2: 152-60.

Aoki, K. F., and M. Kanehisa. 2005. 'Using the KEGG database resource', *Current Protocols in Bioinformatics*, Chapter 1: Unit 1 12.

Arita, Masanori. 2004. 'The metabolic world of *Escherichia coli* is not small', *Proceedings of the National Academy of Sciences*, 101: 1543-47.

Azuaje, Francisco, Yvan Devaux, and Daniel R Wagner. 2010. 'Identification of potential targets in biological signalling systems through network perturbation analysis', *Biosystems*, 100: 55-64.

Bae, Seul-A, and Ioannis P Androulakis. 2018. 'Mathematical Analysis of Circadian Disruption and Metabolic Re-entrainment of Hepatic Gluconeogenesis: The intertwining entraining roles of light and feeding', *American Journal of Physiology-Endocrinology and Metabolism*, 314: E531-E42.

———. 2019. 'Mathematical modeling informs the impact of changes in circadian rhythms and meal patterns on insulin secretion', *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 317: R98-R107.

Baggs, Julie E, Tom S Price, Luciano DiTacchio, Satchidananda Panda, Garret A FitzGerald, and John B Hogenesch. 2009. 'Network features of the mammalian circadian clock', *PLoS biology*, 7: e52.

Bhan, A., D. J. Galas, and T. G. Dewey. 2002. 'A duplication growth model of gene expression networks', *Bioinformatics*, 18: 1486-93.

Birtwistle, M. R., D. E. Mager, and J. M. Gallo. 2013. 'Mechanistic vs. Empirical network models of drug action', *CPT Pharmacometrics Systems Pharmacology*, 2: e72.

Bishehsari, F., F. Levi, F. W. Turek, and A. Keshavarzian. 2016. 'Circadian Rhythms in Gastrointestinal Health and Diseases', *Gastroenterology*, 151: e1-5.

Bonneau, Richard, David J Reiss, Paul Shannon, Marc Facciotti, Leroy Hood, Nitin S Baliga, and Vesteinn Thorsson. 2006. 'The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo', *Genome biology*, 7: R36.

Boros, Endre, Toshihide Ibaraki, and Kazuhisa Makino. 1998. 'Error-free and best-fit extensions of partially defined Boolean functions', *Information and Computation*, 140: 254-83.

Bradshaw, WE, and CM Holzapfel. 2008. 'Genetic response to rapid climate change: it's seasonal timing that matters', *Molecular ecology*, 17: 157-66.

Buhr, E. D., and J. S. Takahashi. 2013. 'Molecular components of the Mammalian circadian clock', *Handbook of Experimental Pharmacology*: 3-27.

Buijs, R. M., C. G. van Eden, V. D. Goncharuk, and A. Kalsbeek. 2003. 'The biological clock tunes the organs of the body: timing by hormones and the autonomic nervous system', *Journal of Endodontics*, 177: 17-26.

Bunke, Horst, and Gudrun Allermann. 1983. 'Inexact graph matching for structural pattern recognition', *Pattern Recognition Letters*, 1: 245-53.

Burda, Z., A. Krzywicki, O. C. Martin, and M. Zagorski. 2011. 'Motifs emerge from function in model gene regulatory networks', *Proceedings of the National Academy of Sciences*, 108: 17263-8.

Cardone, L., J. Hirayama, F. Giordano, T. Tamaru, J. J. Palvimo, and P. Sassone-Corsi. 2005. 'Circadian clock control by SUMOylation of BMAL1', *Science*, 309: 1390 - 94.

Cassone, Vincent M. 1990. 'Effects of melatonin on vertebrate circadian systems', *Trends in Neurosciences*, 13: 457-64.

Ceballos, Gerardo, Paul R Ehrlich, Anthony D Barnosky, Andrés García, Robert M Pringle, and Todd M Palmer. 2015. 'Accelerated modern human–induced species losses: Entering the sixth mass extinction', *Science advances*, 1: e1400253.

Chuang, J. H., and H. Li. 2004. 'Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome', *PLoS biology*, 2: E29.

Chudasama, V. L., M. A. Ovacik, D. R. Abernethy, and D. E. Mager. 2015. 'Logic-Based and Cellular Pharmacodynamic Modeling of Bortezomib Responses in U266 Human Myeloma Cells', *Journal of Pharmacology and Experimental Therapeutics*, 354: 448-58.

Chung, Fan, Linyuan Lu, T Gregory Dewey, and David J Galas. 2003. 'Duplication models for biological networks', *Journal of computational biology*, 10: 677-87.

Cunningham, P. S., S. A. Ahern, L. C. Smith, C. S. da Silva Santos, T. T. Wager, and D. A. Bechtold. 2016. 'Targeting of the circadian clock via CK1delta/epsilon to improve glucose homeostasis in obesity', *Scientific reports*, 6: 29983.

Cutolo, Maurizio, and Alfonse T. Masi. 2005. 'Circadian Rhythms and Arthritis', *Rheumatic Disease Clinics of North America*, 31: 115-29.

Das, DIPAK K, ANITA Neogi, and HARRY Steinberg. 1984. 'Fructose utilization by lung', *Journal of Applied Physiology*, 56: 333-37.

De Vos, J. M., L. N. Joppa, J. L. Gittleman, P. R. Stephens, and S. L. Pimm. 2015. 'Estimating the normal background rate of species extinction', *Conservation Biology*, 29: 452-62.

Defoort, J., Y. Van de Peer, and V. Vermeirssen. 2018. 'Function, dynamics and evolution of network motif modules in integrated gene regulatory networks of worm and plant', *Nucleic acids research*, 46: 6480-503.

del Rio, Gabriel, Dirk Koschützki, and Gerardo Coello. 2009. 'How to identify essential genes from molecular networks?', *BMC systems biology*, 3: 102.

Dibner, C., U. Schibler, and U. Albrecht. 2010. 'The mammalian circadian timing system: organization and coordination of central and peripheral clocks', *Annul Review of Physiology*, 72: 517-49.

Doherty, A. 2018. 'Circadian rhythms and mental health: wearable sensing at scale', *Lancet Psychiatry*, 5: 457-58.

Drake, J. W. 1991. 'A constant rate of spontaneous mutation in DNA-based microbes', *Proceedings of the National Academy of Sciences*, 88: 7160-4.

Dwight Kuo, P., W. Banzhaf, and A. Leier. 2006. 'Network topology and the evolution of dynamics in an artificial genetic regulatory network model created by whole genome duplication and divergence', *Biosystems*, 85: 177-200.

Dyar, Kenneth A., Dominik Lutter, Anna Artati, Nicholas J. Ceglia, Yu Liu, Danny Armenta, Martin Jastroch, Sandra Schneider, Sara de Mateo, Marlene Cervantes, Serena Abbondante, Paola Tognini, Ricardo Orozco-Solis, Kenichiro Kinouchi, Christina Wang, Ronald Swerdloff, Seba Nadeef, Selma Masri, Pierre Magistretti, Valerio Orlando, Emiliana Borrelli, N. Henriette Uhlenhaut, Pierre Baldi, Jerzy Adamski, Matthias H. Tschöp, Kristin Eckel-Mahan, and Paolo Sassone-Corsi. 2018. 'Atlas of Circadian Metabolism Reveals System-wide Coordination and Communication between Clocks', *Cell*, 174: 1571-85.e11.

Edery, I. 2000. 'Circadian rhythms in a nutshell', *Physiol Genomics*, 3: 59-74.

Eisenberg, Eli, and Erez Y Levanon. 2003. 'Preferential attachment in the protein network evolution', *Physical review letters*, 91: 138701.

Erkkila, Timo, Tomi Korpelainen, and Olli Yli-Harja. 2007. "Inference of boolean networks from time series data with realistic characteristics." In *Genomic Signal Processing and Statistics, 2007. GENSIPS 2007. IEEE International Workshop on*, 1-4. IEEE.

Fabregat, A., S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, R. Haw, B. Jassal, F. Korninger, B. May, M. Milacic, C. D. Roca, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, G. Viteri, J. Weiser, G. Wu, L. Stein, H. Hermjakob, and P. D'Eustachio. 2018. 'The Reactome Pathway Knowledgebase', *Nucleic acids research*, 46: D649-D55.

Fang, M., W. R. Guo, Y. Park, H. G. Kang, and H. Zarbl. 2015. 'Enhancement of NAD(+)-dependent SIRT1 deacetylase activity by methylselenocysteine resets the circadian clock in carcinogen-treated mammary epithelial cells', *Oncotarget*, 6: 42879-91.

Feillet, C. A., U. Albrecht, and E. Challet. 2006. '"Feeding time" for the brain: a matter of clocks', *Journal of Physiology - Paris*, 100: 252-60.

Flobak, Åsmund, Anaïs Baudot, Elisabeth Remy, Liv Thommesen, Denis Thieffry, Martin Kuiper, and Astrid Lægreid. 2015. 'Discovery of drug synergies in gastric cancer cells predicted by logical modeling', *PLoS computational biology*, 11: e1004426.

Friedman, Jonathan, and Jeff Gore. 2017. 'Ecological systems biology: The dynamics of interacting populations', *Current Opinion in Systems Biology*, 1: 114-21.

Gao, Li-zhi, and Hideki Innan. 2004. 'Very low gene duplication rate in the yeast genome', *Science*, 306: 1367-70.

Gao, Xinbo, Bing Xiao, Dacheng Tao, and Xuelong Li. 2009. 'A survey of graph edit distance', *Pattern Analysis and Applications*, 13: 113-29.

García-Campos, Miguel A, Jesús Espinal-Enríquez, and Enrique Hernández-Lemus. 2015. 'Pathway analysis: state of the art', *Frontiers in physiology*, 6: 383.

Geerts, Hugo, Athan Spiros, Patrick Roberts, and Robert Carr. 2013. 'Quantitative systems pharmacology as an extension of PK/PD modeling in CNS research and development', *Journal of pharmacokinetics and pharmacodynamics*, 40: 257-65.

Getz, Lowell L. 2009. 'Circadian activity rhythm and potential predation risk of the prairie vole, Microtus ochrogaster', *The Southwestern Naturalist*, 54: 146-50.

Ghalambor, Cameron K, Kim L Hoke, Emily W Ruell, Eva K Fischer, David N Reznick, and Kimberly A Hughes. 2015. 'Non-adaptive plasticity potentiates rapid adaptive evolution of gene expression in nature', *Nature*, 525: 372.

Ghalambor, Cameron K, John K McKay, Scott P Carroll, and David N Reznick. 2007. 'Adaptive versus non-adaptive phenotypic plasticity and the potential for contemporary adaptation in new environments', *Functional ecology*, 21: 394-407.

Gienapp, Phillip, C Teplitsky, JS Alho, JA Mills, and J Merilä. 2008. 'Climate change and evolution: disentangling environmental and genetic responses', *Molecular ecology*, 17: 167-78.

Glass, Leon, and Stuart A Kauffman. 1972. 'Co-operative components, spatial localization and oscillatory cellular dynamics', *Journal of theoretical biology*, 34: 219-37.

———. 1973. 'The logical analysis of continuous, non-linear biochemical control networks', *Journal of theoretical biology*, 39: 103-29.

Goldberg, David E. 1989. *Genetic algorithms in search, optimization, and machine learning* (Addison-Wesley Pub. Co.: Reading, Mass.).

Greil, Florian, Barbara Drossel, and Joost Sattler. 2007. 'Critical Kauffman networks under deterministic asynchronous update', *New Journal of Physics*, 9: 373.

Gu, Zhenglong, Andre Cavalcanti, Feng-Chi Chen, Peter Bouman, and Wen-Hsiung Li. 2002. 'Extent of gene duplication in the genomes of Drosophila, nematode, and yeast', *Molecular biology and evolution*, 19: 256-62.

Guthke, Reinhard, Olaf Kniemeyer, Daniela Albrecht, Axel A Brakhage, and Ulrich Möller. 2006. "Discovery of gene regulatory networks in Aspergillus fumigatus." In *International Workshop on Knowledge Discovery and Emergent Complexity in Bioinformatics*, 22-41. Springer.

Han, Jing-Dong J, Nicolas Bertin, Tong Hao, Debra S Goldberg, Gabriel F Berriz, Lan V Zhang, Denis Dupuy, Albertha JM Walhout, Michael E Cusick, and Frederick P Roth. 2004. 'Evidence for dynamically organized modularity in the yeast protein–protein interaction network', *Nature*, 430: 88-93.

Harris, Stephen E, Bruce K Sawhill, Andrew Wuensche, and Stuart Kauffman. 2002. 'A model of transcriptional regulatory networks based on biases in the observed regulation rules', *Complexity*, 7: 23-40.

Harvey, Inman, and Terry Bossomaier. 1997. "Time out of joint: Attractors in asynchronous random boolean networks." In *Proceedings of the Fourth European Conference on Artificial Life*, 67-75. MIT Press, Cambridge.

Holland, John H. 1975. *Adaptation in natural and artificial systems : an introductory analysis with applications to biology, control, and artificial intelligence* (University of Michigan Press: Ann Arbor).

Hopfensitz, Martin, Christoph Mussel, Christian Wawra, Markus Maucher, Michael Kuhl, Heiko Neumann, and Hans A Kestler. 2011. 'Multiscale binarization of gene expression data for reconstructing Boolean networks', *IEEE/ACM transactions on computational biology and bioinformatics*, 9: 487-98.

Huber, F, J Schnauß, S Rönicke, P Rauch, K Müller, C Fütterer, and J Käs. 2013. 'Emergent complexity of the cytoskeleton: from single filaments to tissue', *Advances in physics*, 62: 1-112.

Hughes, M. E., L. DiTacchio, K. R. Hayes, C. Vollmers, S. Pulivarthy, J. E. Baggs, S. Panda, and J. B. Hogenesch. 2009. 'Harmonics of circadian gene transcription in mammals', *PLoS Genetics*, 5: e1000442.

Hughes, Michael E, John B Hogenesch, and Karl Kornacker. 2010. 'JTK_CYCLE: an efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets', *Journal of biological rhythms*, 25: 372-80.

Ingram, P. J., M. P. Stumpf, and J. Stark. 2006. 'Network motifs: structure does not determine function', *BMC Genomics*, 7: 108.

Jin, Lv, Xiao-Yu Zuo, Wei-Yang Su, Xiao-Lei Zhao, Man-Qiong Yuan, Li-Zhen Han, Xiang Zhao, Ye-Da Chen, and Shao-Qi Rao. 2014. 'Pathway-based analysis tools for complex diseases: a review', *Genomics, proteomics & bioinformatics*, 12: 210-20.

Kaczmarek, J. L., S. V. Thompson, and H. D. Holscher. 2017. 'Complex interactions of circadian rhythms, eating behaviors, and the gastrointestinal microbiota and their potential impact on health', *Nutrition Reviews*, 75: 673-82.

Kanehisa, M., and S. Goto. 2000. 'KEGG: kyoto encyclopedia of genes and genomes', *Nucleic acids research*, 28: 27-30.

Kanehisa, Minoru, Yoko Sato, Miho Furumichi, Kanae Morishima, and Mao Tanabe. 2018. 'New approach for understanding genome variations in KEGG', *Nucleic acids research*, 47: D590-D95.

Kashtan, N., and U. Alon. 2005. 'Spontaneous evolution of modularity and network motifs', *Proceedings of the National Academy of Sciences of the United States of America*, 102: 13773-8.

Kauffman, Stuart. 1969. 'Homeostasis and Differentiation in Random Genetic Control Networks', *Nature*, 224: 177.

———. 2007. 'Beyond reductionism: Reinventing the sacred', *Zygon®*, 42: 903-14.

Kauffman, Stuart A. 1984. 'Emergent properties in random complex automata', *Physica D: Nonlinear Phenomena*, 10: 145-56.

Khaper, N., C. D. C. Bailey, N. R. Ghugre, C. Reitz, Z. Awosanmi, R. Waines, and T. A. Martino. 2018. 'Implications of disturbances in circadian rhythms for cardiovascular health: A new frontier in free radical biology', *Free Radical Biology and Medicine*, 119: 85-92.

Kitano, Hiroaki. 2002. 'Systems Biology: A Brief Overview', *Science*, 295: 1662-64.

Klemm, Konstantin, and Stefan Bornholdt. 2003. 'Robust gene regulation: Deterministic dynamics from asynchronous networks with delay', *arXiv preprint q-bio/0309013*.

Knabe, J. F., C. L. Nehaniv, and M. J. Schilstra. 2008. 'Do motifs reflect evolved function?--No convergent evolution of genetic regulatory network subgraph topologies', *Biosystems*, 94: 68-74.

Knabe, J. F., K. Wegner, C. L. Nehaniv, and M. J. Schilstra. 2010. 'Genetic algorithms and their application to in silico evolution of genetic regulatory networks', *Methods in Molecular Biology*, 673: 297-321.

Komp Lindgren, P., A. Karlsson, and D. Hughes. 2003. 'Mutation Rate and Evolution of Fluoroquinolone Resistance in Escherichia coli Isolates from Patients with Urinary Tract Infections', *Antimicrobial Agents and Chemotherapy*, 47: 3222-32.

Krumsiek, Jan, Sebastian Pölsterl, Dominik M Wittmann, and Fabian J Theis. 2010. 'Odefy-from discrete to continuous models', *BMC bioinformatics*, 11: 233.

Kutmon, Martina, Anders Riutta, Nuno Nunes, Kristina Hanspers, Egon L Willighagen, Anwesha Bohler, Jonathan Mélius, Andra Waagmeester, Sravanthi R Sinha, and Ryan Miller. 2015. 'WikiPathways: capturing the full diversity of pathway knowledge', *Nucleic acids research*, 44: D488-D94.

Lavergne, S., N. J. Muenke, and J. Molofsky. 2010. 'Genome size reduction can trigger rapid phenotypic evolution in invasive plants', *Annals of Botany*, 105: 109-16.

Le Novere, Nicolas. 2015. 'Quantitative and logic modelling of molecular and gene networks', *Nature Reviews Genetics*, 16: 146.

Lee, J. E., and I. Edery. 2008. 'Circadian regulation in the ability of Drosophila to combat pathogenic infections', *Current Biology*, 18: 195-9.

Leil, Tarek A, and Richard Bertz. 2014. 'Quantitative systems pharmacology can reduce attrition and improve productivity in pharmaceutical research and development', *Frontiers in pharmacology*, 5: 247.

Liang, Shoudan, Stefanie Fuhrman, and Roland Somogyi. 1998. 'Reveal, a general reverse engineering algorithm for inference of genetic network architectures', *Pacific Symposium on Biocomputing*, 3: 18-29.

Love, Michael, Simon Anders, and Wolfgang Huber. 2014. 'Differential analysis of count data–the DESeq2 package', *Genome biology*, 15: 10-1186.

Lu, Junyan, Hanlin Zeng, Zhongjie Liang, Limin Chen, Liyi Zhang, Hao Zhang, Hong Liu, Hualiang Jiang, Bairong Shen, and Ming Huang. 2015. 'Network modelling reveals the mechanism underlying colitis-associated colon cancer and identifies novel combinatorial anti-cancer targets', *Scientific reports*, 5: 14739.

Luscombe, N. M., M. M. Babu, H. Yu, M. Snyder, S. A. Teichmann, and M. Gerstein. 2004. 'Genomic analysis of regulatory network dynamics reveals large topological changes', *Nature*, 431: 308-12.

Lynch, M. 2007. 'The evolution of genetic networks by non-adaptive processes', *Nature Reviews Genetics*, 8: 803-13.

———. 2010. 'Evolution of the mutation rate', *Trends in genetics*, 26: 345-52.

MacQueen, James. 1967. "Some methods for classification and analysis of multivariate observations." In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 281-97. Oakland, CA, USA.

Martin, Shawn, Zhaoduo Zhang, Anthony Martino, and Jean-Loup Faulon. 2007. 'Boolean dynamics of genetic regulatory networks inferred from microarray time series data', *Bioinformatics*, 23: 866-74.

Materi, W., and D. S. Wishart. 2007. 'Computational systems biology in drug discovery and development: methods and applications', *Drug discovery today*, 12: 295-303.

Mavroudis, P. D., D. C. DuBois, R. R. Almon, and W. J. Jusko. 2018. 'Daily variation of gene expression in diverse rat tissues', *PLoS One*, 13: e0197258.

Mavroudis, PD, JD Scheff, SE Calvano, and IP Androulakis. 2013. 'Systems biology of circadian-immune interactions', *Journal of innate immunity*, 5: 153-62.

Meshi, Ofer, Tomer Shlomi, and Eytan Ruppin. 2007. 'Evolutionary conservation and over-representation of functionally enriched network patterns in the yeast regulatory network', *BMC systems biology*, 1: 1-1.

Milo, R., S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. 2002. 'Network motifs: simple building blocks of complex networks', *Science*, 298: 824-7.

Mircean, Christian, Ioan Tabus, and Jaakko T Astola. 2002. "Quantization and distance function selecton for discrimination of tumors using gene expression data." In *Functional Monitoring and Drug-Tissue Interaction*, 1-12. International Society for Optics and Photonics.

Morgan, Paul, Piet H Van Der Graaf, John Arrowsmith, Doug E Feltner, Kira S Drummond, Craig D Wegner, and Steve DA Street. 2012. 'Can the flow of medicines be improved? Fundamental pharmacokinetic and pharmacological principles toward improving Phase II survival', *Drug discovery today*, 17: 419-24.

Morris, Melody K, Julio Saez-Rodriguez, Peter K Sorger, and Douglas A Lauffenburger. 2010. 'Logic-based models for the analysis of cell signaling networks', *Biochemistry*, 49: 3216-24.

Mure, L. S., H. D. Le, G. Benegiamo, M. W. Chang, L. Rios, N. Jillani, M. Ngotho, T. Kariuki, O. Dkhissi-Benyahya, H. M. Cooper, and S. Panda. 2018. 'Diurnal transcriptome atlas of a primate across major neural and peripheral tissues', *Science*, 359: eaao0318.

Nakamura, Y., N. Nakano, K. Ishimaru, N. Ando, R. Katoh, K. Suzuki-Inoue, S. Koyanagki, H. Ogawa, K. Okumura, S. Shibata, and A. Nakao. 2016. 'Inhibition of IgE-mediated allergic reactions by pharmacologically targeting the circadian clock', *The Journal of Allergy and Clinical Immunology*, 137: 1226-35.

Naldi, A., E. Remy, D. Thieffry, and C. Chaouiya. 2009. 'A Reduction of Logical Regulatory Graphs Preserving Essential Dynamical Properties', *Computational Methods in Systems Biology, Proceedings*: 266-80.

Noman, N., T. Monjo, P. Moscato, and H. Iba. 2015. 'Evolving robust gene regulatory networks', *PLoS One*, 10: e0116258.

Ohneda, Kinuko, Hooi Ee, and Michael German. 2000. "Regulation of insulin gene transcription." In *Seminars in cell & developmental biology*, 227-33. Elsevier.

Paladino, N., M. J. Leone, S. A. Plano, and D. A. Golombek. 2010. 'Paying the circadian toll: the circadian response to LPS injection is dependent on the Toll-like receptor 4', *Journal Neuroimmunology*, 225: 62-67.

Papagiannakopoulos, Thales, Matthew R Bauer, Shawn M Davidson, Megan Heimann, Lakshmipriya Subbaraj, Arjun Bhutkar, Jordan Bartlebaugh, Matthew G Vander Heiden, and Tyler Jacks. 2016. 'Circadian rhythm disruption promotes lung tumorigenesis', *Cell metabolism*, 24: 324-31.

Parter, M., N. Kashtan, and U. Alon. 2008. 'Facilitated variation: how evolution learns from past environments to generalize to new environments', *PLoS computational biology*, 4: e1000206.

Peng, Tien, David B Frank, Rachel S Kadzik, Michael P Morley, Komal S Rathi, Tao Wang, Su Zhou, Lan Cheng, Min Min Lu, and Edward E Morrisey. 2015.

'Hedgehog actively maintains adult lung quiescence and regulates repair and regeneration', *Nature*, 526: 578-82.

Pietras, RJ, BM Fendly, VR Chazin, MD Pegram, SB Howell, and DJ Slamon. 1994. 'Antibody to HER-2/neu receptor blocks DNA repair after cisplatin in human breast and ovarian cancer cells', *Oncogene*, 9: 1829-38.

Raeymaekers, Joost AM, Anurag Chaturvedi, Pascal I Hablützel, Io Verdonck, Bart Hellemans, Gregory E Maes, Luc Meester, and Filip AM Volckaert. 2017. 'Adaptive and non-adaptive divergence in a common landscape', *Nature communications*, 8: 267.

Rao, R. T., M. L. Scherholz, C. Hartmanshenn, S. A. Bae, and I. P. Androulakis. 2017. 'On the analysis of complex biological supply chains: From Process Systems Engineering to Quantitative Systems Pharmacology', *Computers & Chemical Engineering*, 107: 100-10.

Rao, Rohit, and Ioannis P Androulakis. 2019. 'The physiological significance of the circadian dynamics of the HPA axis: Interplay between circadian rhythms, allostasis and stress resilience', *Hormones and behavior*, 110: 77-89.

Rao, Rohit T, Megerle L Scherholz, and Ioannis P Androulakis. 2018. 'Modeling the influence of chronopharmacological administration of synthetic glucocorticoids on the hypothalamic-pituitary-adrenal axis', *Chronobiology international*, 35: 1619-36.

Regenmortel, Marc H. V. Van. 2004. 'Reductionism and complexity in molecular biology', *EMBO Reports*, 5: 1016-20.

Ruz, Gonzalo A, Ana Zúñiga, and Eric Goles. 2018. 'A Boolean network model of bacterial quorum-sensing systems', *International Journal of Data Mining and Bioinformatics*, 21: 123-44.

Saadatpour, Assieh, Réka Albert, and Timothy C Reluga. 2013. 'A reduction method for Boolean network models proven to conserve attractors', *SIAM Journal on Applied Dynamical Systems*, 12: 1997-2011.

Saez-Rodriguez, Julio, Aidan MacNamara, and Simon Cook. 2015. 'Modeling signaling networks to advance new cancer therapies', *Annual review of biomedical engineering*, 17: 143-63.

Scherholz, Megerle L, Naomi Schlesinger, and Ioannis P Androulakis. 2019. 'Chronopharmacology of glucocorticoids', *Advanced drug delivery reviews*, 151-152: 245-61.

Serra, Roberto, Marco Villani, Alex Graudenzi, and SA Kauffman. 2007. 'Why a simple model of genetic regulatory networks describes the distribution of avalanches in gene expression data', *Journal of theoretical biology*, 246: 449-60.

Serratosa, Francesc. 2019. 'Graph edit distance: restrictions to be a metric', *Pattern Recognition*, 90: 250-56.

Shen-Orr, S. S., R. Milo, S. Mangan, and U. Alon. 2002. 'Network motifs in the transcriptional regulation network of Escherichia coli', *Nature Genetics*, 31: 64-68.

Shmulevich, Ilya, and Stuart A Kauffman. 2004. 'Activities and sensitivities in Boolean network models', *Physical review letters*, 93: 048701.

Shmulevich, Ilya, Olli Yli-Harja, Jaakko Astola, and CG Core. 2001. "Inference of genetic regulatory networks under the best-fit extension paradigm." In

Proceedings of the IEEE—EURASIP Workshop on Nonlinear Signal and Image Processing (NSIP-01), Baltimore, MD, USA, 3-6.

Silver, A. C., A. Arjona, W. E. Walker, and E. Fikrig. 2012. 'The circadian clock controls toll-like receptor 9-mediated innate and adaptive immunity', *Immunity*, 36: 251-61.

Simak, Maria, Chen-Hsiang Yeang, and HH Lu. 2017. 'Exploring candidate biological functions by Boolean Function Networks for Saccharomyces cerevisiae', *PLoS One*, 12: e0185475-e75.

Sirisena, PDNN, and F Noordeen. 2014. 'Evolution of dengue in Sri Lanka—changes in the virus, vector, and climate', *International Journal of Infectious Diseases*, 19: 6-12.

Sprouffske, K., J. Aguilar-Rodriguez, P. Sniegowski, and A. Wagner. 2018. 'High mutation rates limit evolutionary adaptation in Escherichia coli', *PLoS Genetics*, 14: e1007324.

Stelling, J., S. Klamt, K. Bettenbrock, S. Schuster, and E. D. Gilles. 2002. 'Metabolic network structure determines key aspects of functionality and regulation', *Nature*, 420: 190-3.

Straume, Martin. 2004. 'DNA microarray time series analysis: automated statistical assessment of circadian rhythms in gene expression patterning.' in, *Methods in enzymology* (Elsevier), 149-66.

Stumpf, Michael PH, Carsten Wiuf, and Robert M May. 2005. 'Subnets of scale-free networks are not scale-free: sampling properties of networks', *Proceedings of the National Academy of Sciences*, 102: 4221-24.

Swings, T., B. Van den Bergh, S. Wuyts, E. Oeyen, K. Voordeckers, K. J. Verstrepen, M. Fauvart, N. Verstraeten, and J. Michiels. 2017. 'Adaptive tuning of mutation rates allows fast response to lethal stress in *Escherichia coli*', *Elife*, 6: e22939.

Takahashi, Joseph S. 2017. 'Transcriptional architecture of the mammalian circadian clock', *Nature Reviews Genetics*, 18: 164.

Terfve, Camille, Thomas Cokelaer, David Henriques, Aidan MacNamara, Emanuel Goncalves, Melody K Morris, Martijn van Iersel, Douglas A Lauffenburger, and Julio Saez-Rodriguez. 2012. 'CellNOptR: a flexible toolkit to train protein signaling networks to data using multiple logic formalisms', *BMC systems biology*, 6: 133.

Tukey, John W. 1949. 'Comparing individual means in the analysis of variance', *Biometrics*, 5: 99-114.

Van Noort, Vera, Berend Snel, and Martijn A Huynen. 2004. 'The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model', *EMBO Reports*, 5: 280-84.

Van Regenmortel, MHV. 2001. 'Pitfalls of reductionism in the design of peptide-based vaccines', *Vaccine*, 19: 2369-74.

van Someren, Eugene P, Lodewyk FA Wessels, and Marcel JT Reinders. 2000. "Linear modeling of genetic networks from experimental data." In *Intelligent Systems for Molecular Biology*, 355-66.

Veliz-Cuba, A. 2011. 'Reduction of Boolean network models', *Journal of theoretical biology*, 289: 167-72.

Wald, Abraham, and Jacob Wolfowitz. 1940. 'On a test whether two samples are from the same population', *The Annals of Mathematical Statistics*, 11: 147-62.

Waring, Michael J, John Arrowsmith, Andrew R Leach, Paul D Leeson, Sam Mandrell, Robert M Owen, Garry Pairaudeau, William D Pennie, Stephen D Pickett, and Jibo Wang. 2015. 'An analysis of the attrition of drug candidates from four major pharmaceutical companies', *Nature reviews Drug discovery*, 14: 475.

Wilke, C. O., J. L. Wang, C. Ofria, R. E. Lenski, and C. Adami. 2001. 'Evolution of digital organisms at high mutation rates leads to survival of the flattest', *Nature*, 412: 331-3.

Woolf, Steven H. 2008. 'The meaning of translational research and why it matters', *Jama*, 299: 211-13.

Wuchty, S., Z. N. Oltvai, and A. L. Barabasi. 2003. 'Evolutionary conservation of motif constituents in the yeast protein interaction network', *Nature Genetics*, 35: 176-9.

Yamaguchi, Shun, Shigeru Mitsui, Lily Yan, Kazuhiro Yagita, Shigeru Miyake, and Hitoshi Okamura. 2000. 'Role of DBP in the circadian oscillatory mechanism', *Molecular and cellular biology*, 20: 4773-81.

Yu, T., and J. Miller. 2001. 'Neutrality and the evolvability of Boolean function landscape', *Genetic Programming, Proceedings*, 2038: 204-17.

Zaki, N. F. W., D. W. Spence, A. S. BaHammam, S. R. Pandi-Perumal, D. P. Cardinali, and G. M. Brown. 2018. 'Sleep and circadian rhythms in health and disease: a complex interplay', *European Archives of Psychiatry and Clinical Neuroscience*, 269: 365-66.

Zhang, Jianzhi. 2003. 'Evolution by gene duplication: an update', *Trends in ecology & evolution*, 18: 292-98.

Zhang, Ray, Nicholas F Lahens, Heather I Ballance, Michael E Hughes, and John B Hogenesch. 2014. 'A circadian gene expression atlas in mammals: implications for biology and medicine', *Proceedings of the National Academy of Sciences*, 111: 16219-24.

Zhao, Wentao, Erchin Serpedin, and Edward R Dougherty. 2006. 'Inferring gene regulatory networks from time series data using the minimum description length principle', *Bioinformatics*, 22: 2129-35.

Appendix

All pairwise correlations of shared genes

Nearly all genes which are expressed in two or more tissues have similar signaling patterns across each tissue. This can be confirmed by measuring the correlation of shared genes between each tissue pair. If the signal of each shared gene was identical, then we would expect that there would be a single spike at “1”. We find that there is a tendency for the distribution of correlations to be positive, but there are some notable exceptions. Some tissue pairs have a bimodal distribution: a large peak close to 1 and a smaller peak closer to -1. Other tissue pairs have a nearly uniform or normal distribution about 0. These are unique because they indicate that although the tissues have some genes in common, those genes are not expressed in the same manner. This seems to be most common for developmentally distant tissue types (lung and muscle, liver and hypothalamus, liver, and lung).

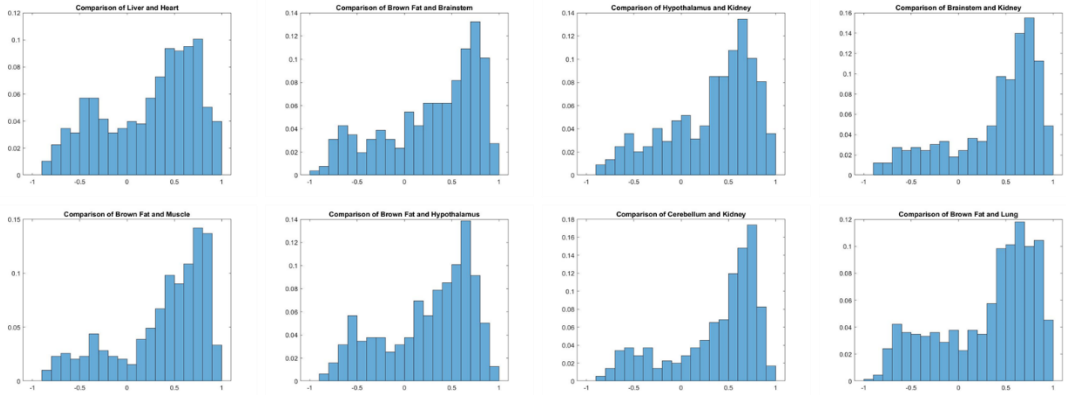

Figure 29: All tissue-to-tissue correlation comparisons of shared genes

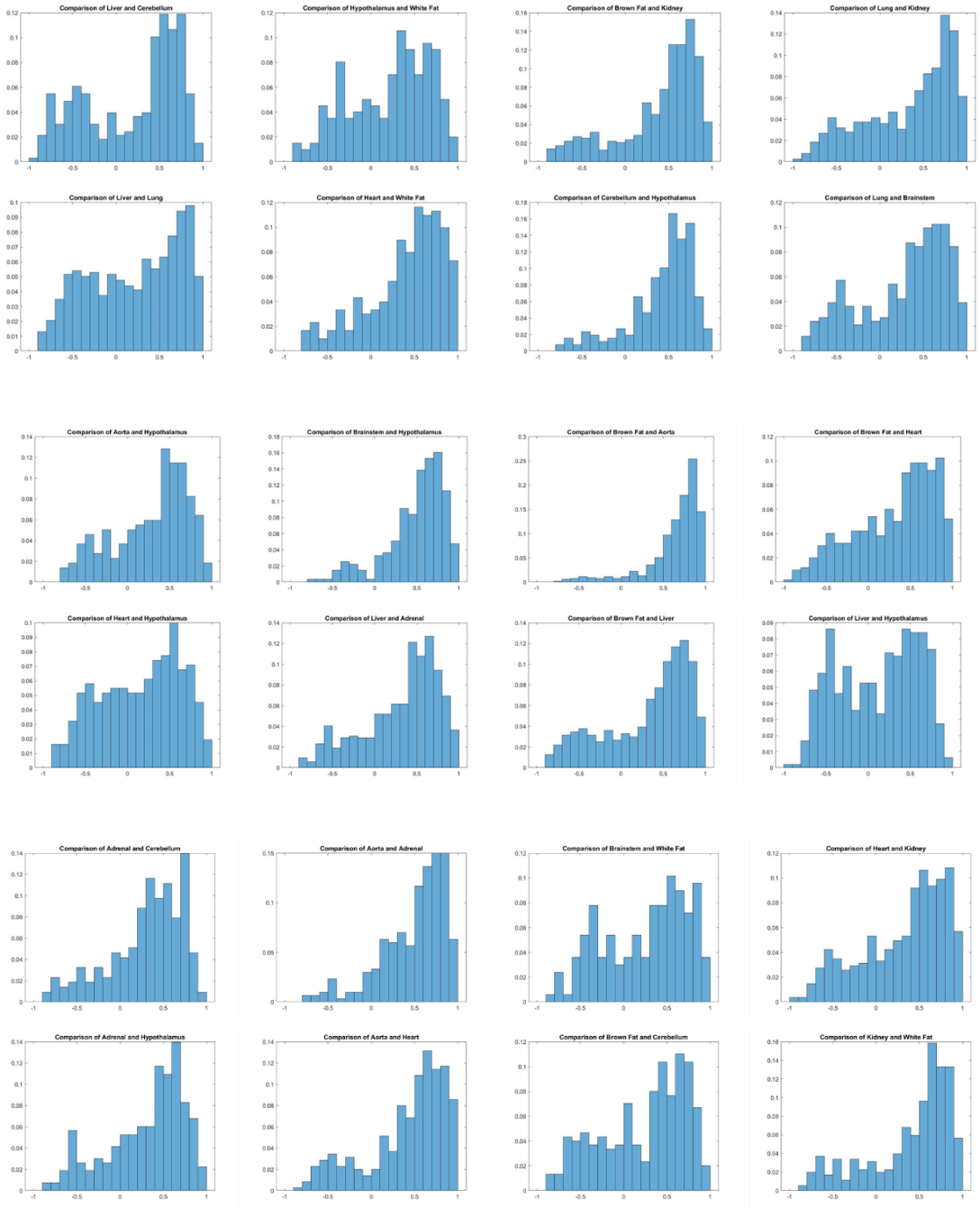

Figure 29 continued

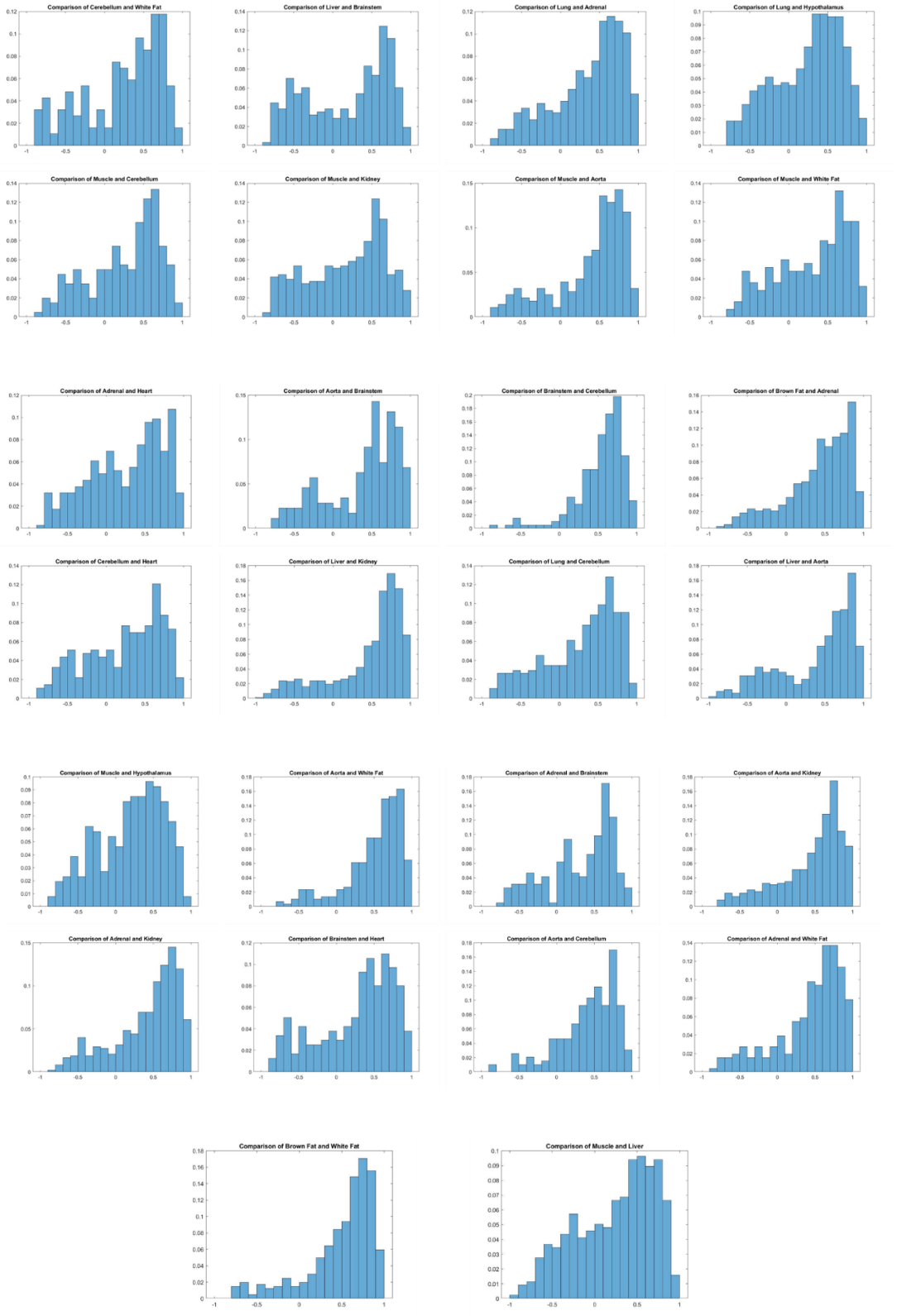

Figure 29 continued