UC Irvine
UC Irvine Electronic Theses and Dissertations

Title
Large Scale Integration, Analysis, and Visualization of Biological Data

Permalink
https://escholarship.org/uc/item/2wp2m4n5

Author
Patel Rajesh, Vishal

Publication Date
2014

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA,
IRVINE

Large Scale Integration, Analysis, and Visualization of Biological Data

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Computer Science

by

Vishal Rajesh Patel

Dissertation Committee:
Professor Pierre Baldi, Chair
Professor Paolo Sassone-Corsi
Associate Professor Xiaohui Xie

2014

© 2014 Vishal Rajesh Patel

DEDICATION

To my mom and dad

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my gratitude to my advisor, Prof. Pierre Baldi, who is not just a great scientist and role-model but has also grown to be a friend. I was extremely fortunate to be learning under his guidance. I thank him and others in the Baldi group who have contributed significantly to the development of Crick, especially Michael Zeller, Jordan Hayes, Nicholas Ceglia, and Yu Liu. I thank Dr. Kenny Daily and Paul Rigor for their work on MotifMap. I thank Peter Sadowski and Dr. Matthew Kayala for all the insightful discussions and advice.

I would like to acknowledge my collaborators for all their scientific contributions. I thank Prof. Paolo Sassone-Corsi, Dr. Kristin Eckel-Mahan, Dr. Selma Masri, and Dr. Lorena Aguilar-Arnal for their wonderful work in Circadian Biology and the extremely fruitful collaboration. I thank Prof. Lan Huang and students from her group. For the pediatric cancer genome project, I thank Dr. Leonard Sender and his colleagues from the Children?s Hospital of Orange County. I thank Prof. Xiaohui Xie for serving on my dissertation committee and all his useful feedback.

I thank my family for their support and encouragement. It wouldn't be wrong to say that I learnt the first concept in science from my mom and the first concept in math from my dad. I thank my parents for their love, care, and support; but I also thank them for all the sacrifices they had to make for me. I thank my brother Biren Patel and sister Jyuthika Patel for always believing in me and encouraging me.

I thank everyone from the MCSB class of 2009. A special thanks to Dr. Chi-Li Chiu and Dr. Jeffrey Suhalim, both brilliant scientists and great to work with, but more important they were my family away from home in Irvine. I would also like to thank all my friends and peers. I specially thank Pavitra Krishnamoorthy, Varsha Natarajan, Naren Vinayak, Sravanthi Sridhar, Natarajan Ramachandran, and Ramnik Kaur for their invaluable support over the last several years.

I would like to thank everyone from The Retail Equation who made my internship a great learning experience and enabled me to positively contribute to several projects. I specially thank Dr. David Speights, my mentor, for several useful discussions, novel ideas, and constant inspiration.

I would like to thank my team at Google who made my internship experience unique and enjoyable. I specially thank Ananth Devulapalli, Samuel Carrijo and Archana Krishnan for their mentoring, constant support, and advice.

I would like to thank the Center for Complex Biological Systems (CCBS) for providing me with the Mathematics, Computational, and Systems Biology Fellowship. I thank the Donald Bren School of Information and Computer Sciences for awarding me with the Dean's Fellowship. My work was also supported by NSF IIS-0513376, NIH LM010235, NIH-NLM T15 LM07443 to Pierre Baldi.

I thank the editors and publishers of Nature Methods, Proceedings of the National Academy

of Sciences, Cell, Cell Metabolism, Trends in Cell Biology, Nature Structural & Molecular Biology, Molecular Metabolism, Molecular & Cellular proteomics, Cellular Signaling, and BMC Bioinformatics for publishing some of my research.

Section 3.2 adapted from Patel *et al.* (2012) and Eckel-Mahan *et al.* (2012).
Section 3.3 adapted from Eckel-Mahan *et al.* (2013).
Section 4.1 adapted from Patel *et al.* (2014).
Section 5.1 adapted from Zeller *et al.* (2014).
Section 5.2 adapted from Aguilar-Arnal *et al.* (2013).

CURRICULUM VITAE

Vishal Rajesh Patel

EDUCATION

Doctor of Philosophy in Computer Science **2014**
University of California - Irvine *Irvine, CA*

Master of Science in Computer Science **2013**
University of California - Irvine *Irvine, CA*

Bachelor of Technology in Industrial Biotechnology **2009**
Anna University *Chennai, India*

RESEARCH EXPERIENCE

Graduate Research Assistant **2009–2014**
University of California, Irvine *Irvine, California*

TEACHING EXPERIENCE

Teaching Assistant **Spring 2012**
ICS 6B, Boolean Algebra and Logic, UC - Irvine *Irvine, CA*

Teaching Assistant **Winter 2012**
ICS 6D, Discrete Mathematics for Computer Science, UC - Irvine *Irvine, CA*

Teaching Assistant **Fall 2012**
CS 175, Projects in Artificial Intelligence, UC - Irvine *Irvine, CA*

Reader **Spring 2011**
ICS 6D, Discrete Mathematics for Computer Science, UC - Irvine *Irvine, CA*

Reader **Winter 2011**
Stats 8, Bio Statistics, UC - Irvine *Irvine, CA*

Reader **Fall 2011**
ICS 171, Introduction to Artificial Intelligence, UC - Irvine *Irvine, CA*

INTERNSHIPS

Software Engineer in Test Intern — **Summer 2013**
Google — *Mountain View, California*

Data Analyst Intern — **Summer 2012**
The Retail Equation — *Irvine, California*

LEADERSHIP

Vice President of Financial Affairs — **2012–2013**
Associated Graduate Students, UC - Irvine — *Irvine, California*

Vice President of Administrative Affairs — **Spring 2012**
Associated Graduate Students, UC - Irvine — *Irvine, California*

President — **2011–2012**
Surabhi, Indian Student Association, UC - Irvine — *Irvine, California*

REFEREED JOURNAL PUBLICATIONS

CircadiOmics: integrating circadian genomics, transcriptomics, proteomics and metabolomics — **2012**
Nature Methods

How pervasive are circadian oscillations? — **2014**
Trends in Cell Biology

Coordination of the transcriptome and metabolome by the circadian clock — **2012**
Proceedings of the National Academy of Sciences

A Genomic Analysis Pipeline and Its Application to Pediatric Cancers — **2014**
IEEE/ACM Transactions on Computational Biology and Bioinformatics

Circadian acetylome reveals regulation of mitochondrial metabolic pathways — **2013**
Proceedings of the National Academy of Sciences

Muscle insulin sensitivity and glucose metabolism are controlled by the intrinsic muscle clock — **2014**
Molecular Metabolism

Leptin Engages a Hypothalamic Neurocircuitry to Permit Survival in the Absence of Insulin — **2013**
Cell Metabolism

MotifMap: integrative genome-wide maps of regulatory motif sites for model species **2011**
BMC Bioinformatics

Development of a novel cross-linking strategy for fast and accurate identification of cross-linked peptides of protein complexes **2011**
Molecular & Cellular Proteomics

Mapping the structural topology of the yeast 19S proteasomal regulatory particle using chemical cross-linking and probabilistic modeling **2012**
Molecular & Cellular Proteomics

Mapping the protein interaction network of the human COP9 signalosome complex using a label-free QTAX strategy **2012**
Molecular & Cellular Proteomics

Combining docking site and phosphosite predictions to find new substrates: Identification of smoothelin-like-2 (SMTNL2) as a c-Jun N-terminal kinase (JNK) substrate **2013**
Cellular Signaling

ABSTRACT OF THE DISSERTATION

Large Scale Integration, Analysis, and Visualization of Biological Data

By

Vishal Rajesh Patel

Doctor of Philosophy in Computer Science

University of California, Irvine, 2014

Professor Pierre Baldi, Chair

Data from decades of life sciences research and literature is being curated and made available for searching and analysis. While considerable work has been done to integrate and re-use this data, what is still lacking is a unifying platform that allows new experimental data to leverage all previously published data effectively.

Crick is an intelligent and scalable platform for data integration, visualization, and searching for meaningful biological hypothesis. It was built to create an effective way to integrate and analyze experimental data in the context of the vast literature of other biologically relevant information. Crick was designed ground-up to solve some of the most challenging problems in biological data such as entity resolution, size, scale, reliability of the data, visualization of high-dimensional information, etc. Crick has been successfully used to identify molecular mechanism regulating circadian metabolism; to understand the complex coupling of circadian oscillating species; to study pediatric cancers; to analyze the dynamic long-range interactions in the genome; and more.

Chapter 1

Big Data in Biology

1.1 Introduction

Humans have been collecting and recording information for a very long time. The oldest proof of written history dates to the 4th millennium BC. We record events, phenomena and every single aspect of our understanding of the world around us. In the age of the internet, we are now experiencing an explosion in the amount of data that is collected and processed. An estimate in 2012 by IBM shows that 2.5 exabytes (2.5×10^{18}) of data is created every single day and this number is increasing. Hence there is a massive amount of data being collected across all domains including but not limited to physics, internet, finance, meteorology, environment and climate studies, politics, marketing and biology.

The domain of life sciences is witnessing a dramatic change in the way we work with data. Data from decades of research is being curated into databases with the ability to be re-used, thereby becoming more accessible. Data from literature is being converted into machine readable formats making it readily available for mining and analysis. However, integration of biological data from diverse databases and various file formats is a growing challenge.

Biological data is a combination of qualitative and quantitative results. There are multiple heterogeneous datasets that can be collected from a system, for example genome sequence, transcriptome, metabolome, proteome, imaging, drug affinity etc. Even when measuring the same biological entity, there are multiple technologies, vendors, experimental procedures and analysis methods which make it harder to compare data across experiments. For instance, one can carry out microarrays or RNA-sequencing to measure the genome-wide expression of transcripts – both of which are offered by several vendors as well as have several methods for normalizing and data processing. With more and more high throughput data being made available, there is a need for a system to integrate new experimental data with all other publicly available data in order to extract biologically meaningful hypotheses. Incompatible exchange of data and the inability to extract results from previous studies have often resulted in whole experiments being repeated. Furthermore, we are also seeing trends of comprehensive data being generated under a specific condition. Like ChIP-seq, transcriptome, and metabolome all measured from the liver of the same animal with identical feeding profile, genetic background, environmental conditions, etc.

1.2 Key Challenges

Biological data has structured, unstructured, semi-structured and multi-structured data. Many of the challenges when dealing with biological data are common to other domains of data science.

1.2.1 Entity resolution or record linkage

Entity resolution or Record Linkage is the process of identifying information in a single data set or across data sets that refers to the same entity (a single molecular species). While

this may seem trivial, it is actually one of the hardest problems in life sciences (Stein, 2003). Many standards have been proposed (Wain *et al.*, 2002; Maltais *et al.*, 2002; Levan *et al.*, 1995) that ameliorate this problem but legacy databases and published literature still contains useful information that is difficult to resolve correctly. In addition, there are subtle difference in the concepts or usage of terms as me move from one domain to another (Stein, 2003).

1.2.2 Size

With recent advances in genomics, proteomics, metabolomics, imaging and other technologies, researchers are generating huge amounts of data. The European Bioinformatics Institute (EBI) which maintains one of the world's largest biological-data repositories, stores ∼20 petabytes (20 x 10^{15} bytes) of data. Of that, genomic data alone accounts for ∼2 petabytes, with the number doubling in under a year. The National Center for Biotechnology Information, or NCBI also stores ∼15 petabases of DNA sequencing data generated per year (Marx, 2013). The Beijing Genome Institute is one of the largest genome sequencing facilities in the world with a capacity of generating 6 TB of data per day. Researchers are already taking on highly ambitious big-data projects like the analysis of the genomes of many cancers; mapping of the human brain; developing better biofuels and other crops; personalized medicine and drug discovery. Hence the size of data generated is only expected to increase over time as we develop better, more sensitive methods for measurement.

1.2.3 Scale

Algorithms, data structures, and analysis methods need to scale with the size of the data, so large collections of data can be efficiently processed. Ideally, scaling should happen without requiring specific high-end hardware. Distributed computing is becoming extremely

common in processing large amounts of biological data. Compute Grids/HPC, Message Passing Interface (MPI) based clusters, and shared-nothing architectures like Hadoop are all being used to effectively process biological data. Each of these methods offer unique scalability and performance characteristics, while they also have their share of advantages and challenges.

1.2.4 Efficient storage and query

Often, the size and scale are so large that it makes the storage prohibitively difficult and expensive. Downloading petabytes of data and analyzing it would be highly inefficient. With cloud based technology, there are ways of minimizing the data movement and allowing queries that are trying to find needles in the haystack. Fortunately, distributed file systems (like Hadoop Distributed File System and GlusterFS) provide a cost efficient way of storing large datasets. In addition, most databases provide standard APIs to access information avoiding downloads, in-house storing, and maintenance of large data stores.

1.2.5 Missing data

Missing data is extremely common in all data sciences and this is especially true of biological data. For instance, in the case of microarrays, there are many sources of missing values such as scratches on the slides, background noise, hybridization failures, spotting conditions, etc. (de Brevern *et al.*, 2004; Scheel *et al.*, 2005). Missing values in experimental data often has a negative impact on downstream analyses of the data.

But in addition, there are several challenges unique to biological data, such as

1.2.6 Obsolete experimental and analysis procedures

As new experimental procedure and technology is developed and adapted, data generated from older methods has become less useful. This is also true of the methods with which we normalize and analyze data. New methods of analysis are published every day. Defining standards for input and output would help create pluggable analysis modules.

1.2.7 Reliability of data

Biological data often contains significant amounts of noise. Similar experimental procedures performed by different researchers produce different results. There are several currently unknown factors that influence and affect the quality and reliability of the data. Statistical tests and performing sufficient replicates are a good way of decreasing the noise. Cost also plays an important role in how many replicates can be performed. Modern technology such as sequencing is not perfect either. For a very accurate picture, the sequencing coverage required would be extremely high. False or incorrectly annotated data is also common within biological databases. This is because of the fact that several of these databases are machine annotated rather than hand curated by experts.

1.2.8 High-dimensional

For any given biological molecule, there are a large set of variables and dimensions that are available or can be measured. For example, in the case of a protein, its sequence, physical properties, chemical properties, structural properties, pathway and interaction features, expression profile etc. can be used in any analysis. One important side-effect of this is overfitting in predictive models and increase in the complexity of algorithms. In addition, it has been shown that the expected distance between the nearest and farthest points in

high dimensional space (Beyer *et al.*, 1999) decreases as the dimensions increase, making clustering ineffective. Principal component analysis (Fodor, 2002) and other methods are often used to reduce the dimensionality.

1.2.9 Complex

Biological data is extremely complex. Let us consider gene expression for example. In the human body, there are over 20,000 genes. Splicing and post-transcriptional modification can result in a much larger number of transcripts. In the human body there are thousands of different cell-types; each of them can have a different gene-expression profile. Furthermore, the gene-expression can oscillate and quickly change with the day-night cycle (Patel *et al.*, 2014) and/or environmental factors – meaning the levels are different depending on what time of the day and where the measurements are performed. In addition, biological systems have a stochastic component that modulates the molecular interactions within the cell (Quackenbush, 2007; Losick and Desplan, 2008). There is a similar degree of complexity in all layers of biological data. In most situations, the average behavior observed by measuring millions of cells is enough to understand the biological systems that we study. However, the conditions under which a healthy cell switches to a disease state would require accounting for the stochasticity (Losick and Desplan, 2008).

1.2.10 Visualization and sharing of results

With the increase in volume and complexity of the data, visualisation has become extremely important. Tables and spreadsheets are useful for sorting and viewing a small set of rows, but unsuitable for unstructured data. Charts and heat maps are often limited to a few different dimensions. Specialized tools are also available for certain sub-domains. For instance, genome browsers (UCSC Genome Browser, Integrative Genomics Viewer, Savant Genome

browser etc.) and circos allow users to visualize regions of the genome and also add custom tracks of experimental data for visualization. Epigenome browser is another example of a custom tool built for the visualization of epigenetic datasets. Networks are also a great way of visualizing the links between the different molecular species, but less ideal when the network is extremely large or when there are too many different types of nodes and edges. There are several software/libraries available for network visualization and layout like Cytoscape, D3js, NAViGaTOR, and Gephi. Amongst these, Cytoscape is the most popular and feature rich option. Thus, effective visualization of biological data often requires a hybrid approach combining several of the above-mentioned methods.

1.2.11 Partial knowledge

Data models become obsolete quickly as new information becomes available. Sometimes the data is not measured and at other times we do not even know what to measure. For instance long-range interactions and the 3D structure of the genome was unknown till a few years ago, and the data model had to account for storing and analyzing this piece of information. Hence it is important to adapt to new data as it becomes available, thereby making the design of the platform that much harder. This also makes maintaining constraints and consistency checks within and across databases difficult.

1.2.12 Too many databases

Biological knowledge is spread across several different databases. While this makes integration and consistency checking harder, it also provides a great way for domain experts and community to make optimal decisions about the data, updating the data, etc. (on Frontiers at the Interface of Computing *et al.*, 2005; Stein, 2003). Databases like GenBank, the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database, UCSC genome

browser, and the DNA Databank of Japan (DDJ) provide well recognized options for storing DNA regions and genomics sequences. Protein databases like UniProt also link back to most of these popular DNA databases. There are databases for specific interactions like BioGRID and STRING for protein-protein interactions; MotifMap and CENTIPEDE for transcriptional regulation edges; KEGG and MetaCyc for enzyme-metabolite reactions etc. Furthermore, there are also several organism specific databases like FlyBase, WormBase, SGD, and Mouse Genome Informatics (MGI). A full list of key databases is provided in Table 2.1. As previously noted, (on Frontiers at the Interface of Computing *et al.*, 2005), the long-term vision would be to have decentralized collection of data while maintaining common standards and practices for design, compatible vocabularies, user interfaces etc. across all databases, allowing easy integration and cross-database queries. This is a non-trivial challenge and intelligent methods of data integration across databases are steps towards solving these challenges.

1.2.13 New information

Last but not the least is dealing with new information. New data is being generated at an astronomical rate and it is not only important to analyze the new data in the context of all previously published data, but also to be able to easily re-analyze older data in the context of the new data. Often inferences would have been missed due to partial knowledge or missing data.

1.3 Summary

In summary, biology contains a large number of different data types, all of which are needed to understand and reliably model the function at the cellular and organism level. Each data

source can only provide a limited view of the underlying molecular mechanisms, and it is the integration of these different lines of evidence under a wide variety of conditions that can help in understanding any biological phenomenon at multiple levels. Integration from multiple sources is always expected to provide more information that any single source of data. In the next chapter, we describe the platform Crick, which provides a scalable and intelligent way for large scale integration of data and analysis of new experimental data.

Chapter 2

Crick

Each data, from microarray to sequencing, from protein interactions to drug binding, comes with its own noise and limitations. It is the combination of diverse lines of evidence that has the power and coverage to solidify inferences, rank hypotheses, and built predictive models in a relevant way. This integration process is not new or unique to biology, of course, and in essence is at the root of IBM's Watson system for the game of Jeopardy (Ferrucci, 2011) and most modern search engines. This integration process is ongoing and raises several computational challenges in its execution.

Crick was designed ground up to solve some of the challenges described in the previous chapter and effectively provide context to any dataset. Broadly, Crick provides three major functions. First, is data integration and the ability to leverage the vast knowledge base of prior information (public databases, published literature etc) to analyze new experimental data. Second, is the ability to search and rank specific molecular hypotheses. Last, is the data visualization which allows easy navigation and manipulation of biological networks with the ability to overlay any specific data of interest.

2.1 Data Integration

Crick extracts information from a variety of biological databases and web servers to build an extensive knowledge base. Table 2.1 provides a list of the main data sources and tools that are integrated into Crick along with a brief description and the corresponding URL.

2.1.1 Representation

A natural way of representing and storing biological data is in the form of a network. Biological networks contain nodes (genes, proteins, sequence of DNA, metabolites, drugs etc) and edges (protein-protein interaction, transcriptional regulation, enzyme-metabolite interaction, drug-protein binding, long-range genomic interactions, etc.). There can be multiple edges between a pair of nodes; edges can be directed or undirected with self-edges. Thus a multi-graph with directed and undirected edges is a suitable form for representing and storing biological data in Crick. Furthermore, annotations about nodes and edges can be stored as key-value attributes to the corresponding nodes and edges. This model of storing information allows one to enumerate all the links in the data, effectively partition and merge the data, and also query specific pieces of the data in the context of the neighborhood in the biological network.

2.1.2 Building

Since the data is represented as a network, the retrieval of information from various sources can be trivially parallelized. Crick builds the network around each node (genes, metabolites, proteins, etc.) separately. Views of the networks centered around a specific node are constructed. All of the relevant information about the nodes and edges in the graph around the node of interest is extracted from databases, file systems, and web services for processing.

Database	Description	URL
TRANSFAC	Databases of transcription factors and transcription factor weight matrices.	http://www.gene-regulation.com/pub/databases.html
JASPAR	Databases of transcription factors and transcription factor weight matrices.	http://jaspar.cgb.ki.se
MotifMap	Genome-wide maps of regulatory binding sites. MotifMap uses transcription factor weight matrices and the Bayesian Branch Length Score to assess evolutionary conservation and identify DNA regulatory elements across entire genomes.	http://motifmap.igb.uci.edu
CENTIPEDE	Predicted transcription factor binding sites.	http://centipede.uchicago.edu/
KEGG	Metabolic pathway information, knowledgebase for metabolites	http://www.kegg.jp
NCBI	Information about mouse mRNA and genes, transcriptional and translational relationships between genes, mRNA and proteins	http://www.ncbi.nlm.nih.gov/nuccore
UCSC Genome Browser	Genomes, alignments, ChIP-seq, and other genomic datasets	http://genome.ucsc.edu
ENCODE	Chip-seq data, DNaseI Hypersensitivity data etc.	http://genome.ucsc.edu/ENCODE
GEO	Gene expression and other genomic datasets	http://www.ncbi.nlm.nih.gov/geo
Circa	Circadian gene expression profiles	http://bioinf.itmat.upenn.edu/circa
UniProtKB	Protein knowledgebase	http://www.uniprot.org
BioGRID	Protein-protein interactions	http://thebiogrid.org
IntAct	Protein-protein interactions	http://www.ebi.ac.uk/intact
Mint	Protein-protein interactions	http://mint.bio.uniroma2.it/mint
STRING	Protein-protein interactions	http://string-db.org/
CORUM	Information about protein complexes	http://mips.helmholtz-muenchen.de/genre/proj/corum
DAVID	Functional enrichment analysis	http://david.abcc.ncifcrf.gov
Phosida	Post-translational modification information	http://www.phosida.de/
PhosphoSitePlus	Protein phosphorylation information	http://www.phosphosite.org/
DrugBank	Drugs and drug targets	http://www.drugbank.ca/
BindingDB	Drugs and drug targets	http://www.bindingdb.org/
PharmGKB	Drugs and drug targets	https://www.pharmgkb.org/

Table 2.1: **Data sources and web-services used by Crick**

Node type	Use
CrickDNARegion	Used to represent any sequence of DNA e.g. a Single Nucleotide Polymorphism
CrickGene	Used to represent genes.
CrickRNA	Used to represent transcripts (mRNA), micro RNAs etc.
CrickProtein	Used to proteins including transcription factors, enzymes, etc.
CrickCompound	Used to represent metabolites, drugs and other chemical compounds found in the biological system.

Table 2.2: **List of node types available in Crick**

These graphs are typically first or second neighbor graphs built around the initial node and finally stored to disk or a database. For instance, a metabolite centric view defined in Figure 2.3 is built around a metabolite and contains all the enzymes that regulate the metabolite, all the metabolites that co-react with it, all transcription factors and protein/drug interactions of the enzymes.

Once the base networks have been built, different annotators and experimental data integrators can be run on top of these networks. This is how Crick can quickly leverage a vast knowledge base of published datasets and literature information. For instance, to analyze a microarray dataset using a microarray annotator, the experimental data can be augmented to the networks with the relevant expression values. In this step the networks can further be pruned for the specific tissue or condition. For example, tissue specific ChIP-seq datasets can be included; genes not expressed in the tissue can be removed hence pruning the graph. At regular intervals all the base networks are recreated and all experimental datasets reintegrated. The process of building Crick networks around *Upp2* is illustrated in Figure 2.1 and sample Crick code (in Python) shown in Figure 2.2.

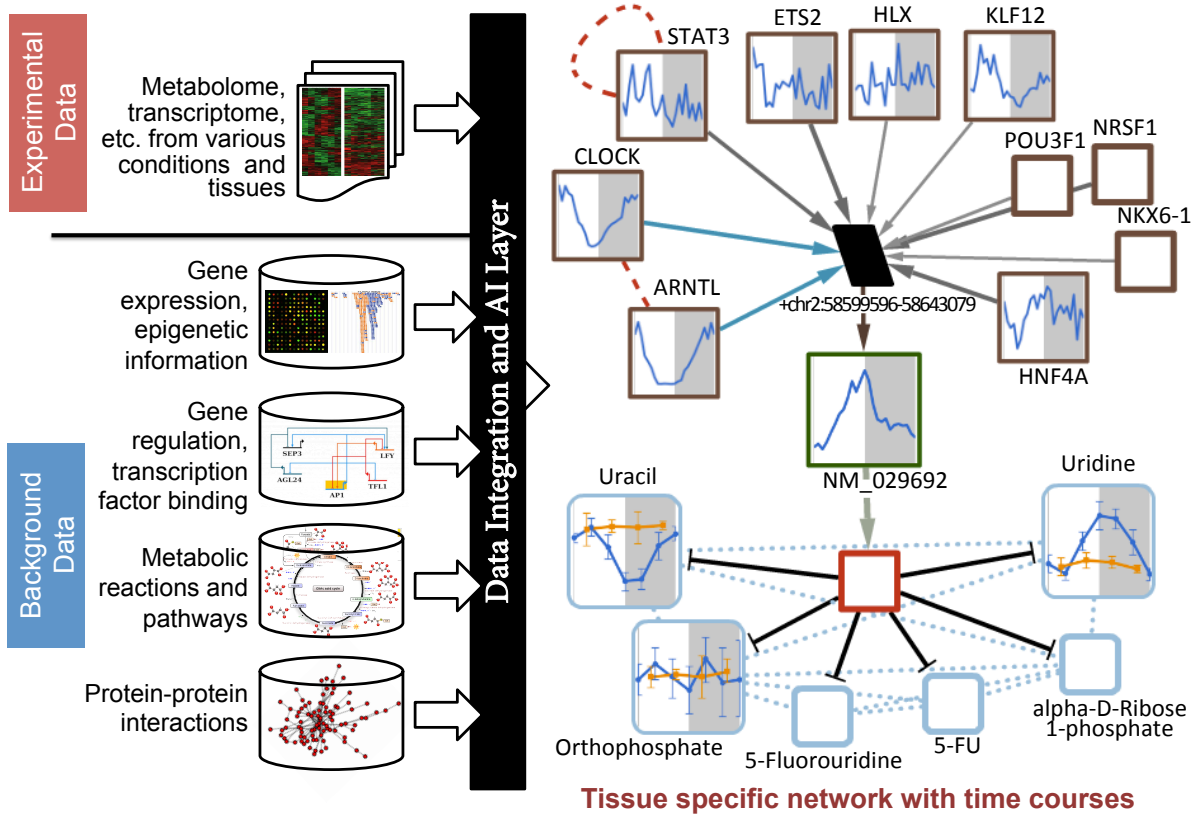


Figure 2.1: **Schematic of data integration.** Starting from the gene *Upp2* (black node labeled with genomic coordinates) Crick builds out the mRNA transcript (green node labeled with RefSeq identifier) and enzyme (red node labeled with UniProt name) along with all relevant nearest neighbors. UPP2 catalyzes a key reaction (black lines) involving uracil and uridine. Regulatory interactions (gray arrows, predicted; cyan arrows, experimentally verified) of transcription factors (brown nodes) that bind to the promoter of *Upp2*, metabolite coreactions (blue dotted lines) and protein-protein interactions (red dashed lines) are also displayed. When available, corresponding time series are shown within the nodes (mRNA data from microarray experiments and metabolite data from mass spectrometry experiments) over the light-dark 24-hour cycle (blue curves, wild type; orange curves, Clock$-/-$).

```
from crick.environment import *
FLAGS.species = const.hg19

cn = CrickNetwork()
cn.add_proteins(["upp2"])

ProteinProteinEdges(cn).build_network()
ProteinDNAEdges(cn).build_network()
EnzymeMetaboliteEdges(cn).build_network()

cn.export_json("upp2.json", networkout=True)
```

Figure 2.2: **Example Python code to build protein-protein interaction, enzyme reaction, and gene regulation for the UPP2 enzyme.**

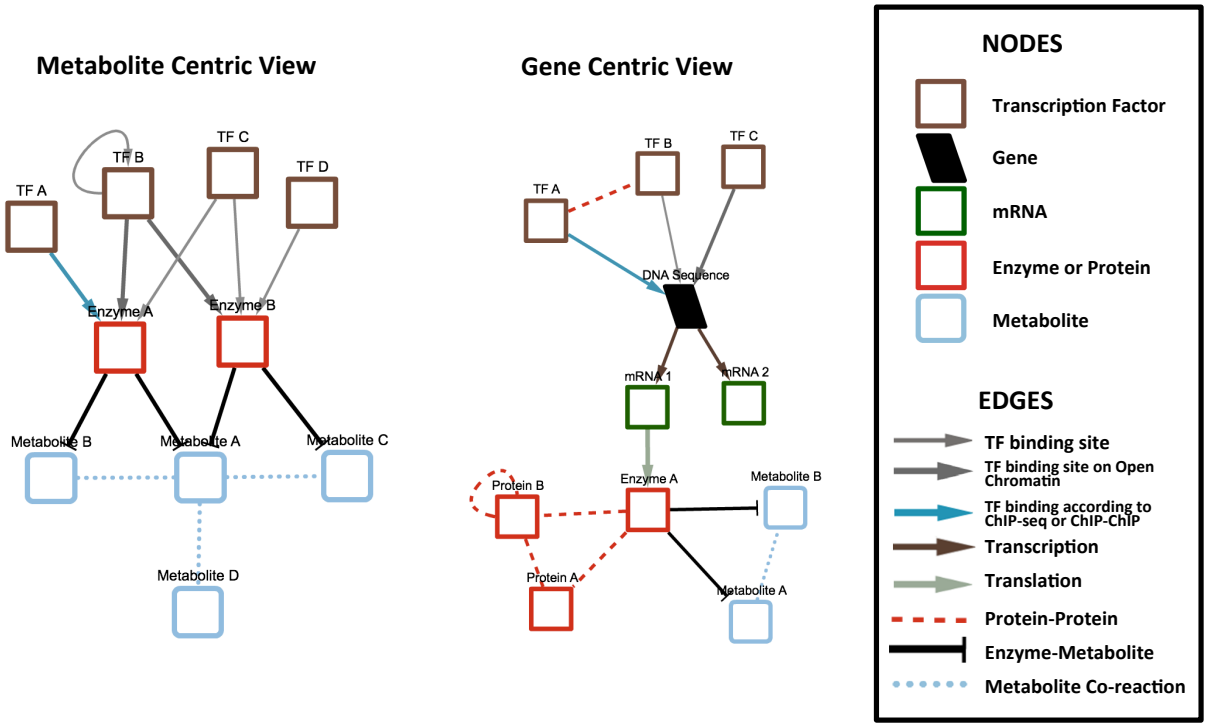


Figure 2.3: **Gene and metabolite centric views of Crick networks.**

2.1.3 Storing

Crick networks can be serialized using Python's pickle format, JSON, or XGMML. JSON and XGMML are more suited for visualization using Cytoscape, while the Python pickle format allows these sub-networks to be quickly loaded, processed, updated, and combined to form new larger networks. The backend database is a scalable and fault tolerant NoSQL datastore. There are several good open-source and proprietary choices for a NoSQL database. Crick uses MongoDB primarily because of the large set of features (counters, regular expression, search, complex indexing, etc.) that are built into the database.

2.2 Analysis and Query

Starting from a small sub-network, it can be expanded by doing a breadth-first or depth-first search and merging with other networks. Additional experimental data to drive the analysis can also be easily overlaid on top of the networks. For instance, after overlaying the expression of all genes from a microarray study, one can calculate the correlation of a node to its connected neighbors. Paths can be correlated and scored from a node of interest to all other nodes in the graph.

Crick also enables searching for specific network-motifs. Networks motifs are a set of nodes and edges that describe a set of predicates that need to be satisfied. For instance, one can describe a network motif with 4 nodes (A, B, C, and D) such that A and B are transcription factors; A and B interact with each other via a protein-protein interaction; C is an enzyme; C is regulated by A and B; C catalyzes a metabolic reaction of D. A motif like this can be easily defined and Crick can enumerate all matching sub-graphs from the network. These network motifs can be extremely complex and have access to all features/data about the nodes and edges. For instance, in the above mentioned motif, one can add constraints - such

as that the gene-expression of A-B-C be correlated; C have a specific gene-ontology; and D participates in a specific metabolic pathway. This also allows estimating graph topology, looking for known molecular feed-forward, feed-back loops etc.

2.3 Data Visualization

The Crick web server uses CytoscapeWeb (Lopes *et al.*, 2010; Shannon *et al.*, 2003) in a HTML5/JavaScript application to display the networks and all node and edge information. The web server allows users to visualize and interactively manipulate any Crick network within the web browser, hence removing the need to download the entire graph and install complex dependencies. This module is implemented using Apache, Python (Flask), and CytoscapeWeb. Data in the form of charts (line charts, pie charts, bar charts etc) or node fill colors (used to represent up and down regulated molecular species) can be used to visualize a specific dataset by drawing them directly on the networks using Cytoscape. While the overlaid graphs are created using Google charts API, any image URL can be shown on the nodes.

During the build process, all nodes, edges, and keywords are indexed using a full-text search library. This allows searching for any Crick network by a combination of one or more keywords.

2.4 Features and Design Choices

2.4.1 Fault tolerance

It is important for a large and complex system like this to be fault tolerant at every level. For example, for most data sources multiple methods of access (web API, database connection, local file system downloads, etc) are defined in Crick. For a given datasource, Crick uses the preferred method for access but in case of failure will automatically extract information from any available channel. This model greatly simplifies the process of downloading data and making sure all parsers are up-to-date. While community driven libraries like BioPython/BioPerl etc. are making this task easier, updating datasets is a huge challenge often forcing researchers to use outdated versions of the data.

The choice of database and file system was also made with fault tolerance in mind. For example MongoDB was run with sharing, replication, and multiple Mongo servers to ensure reliable uptime.

2.4.2 Data sources

Crick contains a large and comprehensive list of data sources, often with multiple sources of databases for the same piece of data. New databases are constantly being added and the design makes it easy to add a new source of node or edge database. Hence Crick is fairly comprehensive and contains a large knowledge base. Information from most databases can be retrieved directly without downloading and processing them in-house.

2.4.3 Performance and scale

Crick uses a compute grid to build networks in parallel and stores the information on network attached file systems or a key-value database. It is important to note that this choice was primarily due to the availability of a large Sun Grid Engine cluster within the group. The network building and information retrieval can be carried out easily by a MapReduce (Dean and Ghemawat, 2008) job that stores the retrieved information to a distributed file system or a distributed key-value database. Crick can reconstruct all of the base networks within a couple of hours – depending on how many compute slots are available, how much data is available locally, how many web requests can be made per second etc. Crick's knowledge base building is embarrassingly parallel and has linear scaling capability.

2.4.4 Auto refresh and cache

Ability to extract data using standard APIs and remote database connections ensures that the networks are always built against the latest version of the data. In addition, APIs often return results in standard format, provide redirection, content-headers etc. Crick also implements a simple persistent-caching mechanism that saves the results of remote requests, decreasing the load on the source web server drastically. This is important since Crick networks are updated and recreated far more often than the data sources.

2.4.5 Experimental data

Crick is designed to ingest experimental data from disparate sources including data analyzed by different techniques or generated using different methods.

2.4.6 Learns IDs over time

Internally Crick assigns a unique identifier for each specific molecular entity. Using cross-reference information, IDs from different databases are connected together and assigned a unique identifier. This means that the Crick assigned ID is transient and resolves to a set of identifiers that can be used based on the database that needs to be queried.

2.4.7 Web application for data visualization

Crick's web based application allows users to view networks without having to install complex dependencies. Crick uses modern NoSQL databases for storing the different objects and networks.

The following chapters describe projects where Crick was successfully used for data analysis, to find novel molecular mechanism, to gather global statistics, etc.

Chapter 3

Understanding the link between metabolism and the circadian clock

3.1 Introduction

Circadian rhythm or the day-night cycle plays a key role in ensuring homeostatic balance with the environment and coordinating many aspects of physiology including the sleep/wake cycle, eating, hormone and neurotransmitter secretion, and even cognitive function (Eckel-Mahan and Sassone-Corsi, 2009; Froy, 2011; Gerstner *et al.*, 2009; Yoo *et al.*, 2004; Takahashi *et al.*, 2008). Disruption of circadian rhythms has been directly linked to health problems ranging from cancer; to insulin resistance, to diabetes, to obesity, and to premature ageing (Takahashi *et al.*, 2008; Antunes *et al.*, 2010; Froy, 2010; Karlsson *et al.*, 2001; Knutsson, 2003; Kohsaka *et al.*, 2007; Lamia *et al.*, 2008; Sharifian *et al.*, 2005; Turek *et al.*, 2005). Research has shown that circadian rhythms are genetically encoded by a molecular clock found in nearly every cell, with a *master clock* located in the suprachiasmatic nucleus (SCN) (Moore and Eichler, 1972; Ralph *et al.*, 1990) of the hypothalamus coordinating and inter-

acting with peripheral clocks throughout the body (Yoo *et al.*, 2004; Takahashi *et al.*, 2008). Central to the molecular rhythmicity of SCN neurons as well as other oscillating cells are transcription factors that drive expression of their own negative regulators (Schibler and Sassone-Corsi, 2002). This attribute of the clock results in a negative transcriptional and translational feedback loop, highly conserved across species, that perpetuates oscillations in gene expression that occur every 24-hrs. In mammals, two bHLH transcription factors, CLOCK and BMAL1 heterodimerize and bind to conserved E-box sequences in target gene promoters, thus driving the rhythmic expression of mammalian Period (*Per1*, *Per2*, and *Per3*) and Cryptochrome (*Cry1* and *Cry2*) genes (Stratmann and Schibler, 2006). PER and CRY proteins form a complex that inhibits subsequent CLOCK:BMAL1-mediated gene expression (Brown *et al.*, 2012; Dibner *et al.*, 2010; Partch *et al.*, 2013). In short, the core of the clock is driven by only a dozen genes (Yan *et al.*, 2008).

Circadian genomic and transcriptomic data sets are accessible, and recently, metabolomics data in the circadian context also have been made available (Eckel-Mahan *et al.*, 2012, 2013; Dyar *et al.*, 2013). Metabolism is a highly dynamic process. It is tightly controlled by the circadian clock, which responds to environmental lighting conditions via the suprachiasmatic nucleus of the brain. Notably, peripheral organs have their own pacemakers, with food serving as a potent zeitgeber for metabolically active tissues (such as liver and muscle). We have explored the circadian metabolome in different tissues and under various physiological conditions, including disparate nutritional and circadian conditions. Thus, the components for building a comprehensive map of circadian networks are thus ready for assembly using Crick. Our initial studies (Eckel-Mahan *et al.*, 2012) had revealed that 60% of liver metabolites show a difference between wild-type and circadian mutant (Clock–/–) mice.

3.2 Liver Metabolome with a Clock Knockout

As an organ involved in glycogen storage, the production of bile acids, and the storage and provider of numerous amino acids and vitamins for the rest of the body, the circadian oscillation of metabolic function within the liver is seminal to understanding circadian physiology. Thus, to determine whether liver metabolites oscillate in a circadian fashion and to determine the extent to which their presence is *Clock*-controlled, liver samples from wild-type and *Clock*-deficient (*Clock*-/-) mice were prepared for GC/MS and LC/MS/MS analysis.

Mass Spectrometry spectra for 538 metabolites were obtained, of which 309 were identified by searching against a standard library (Evans *et al.*, 2009). Metabolites that showed time- or genotype- main effects were grouped based on their major pathway in Figure 3.1. Red fractions of each pie chart in Figure 3.1 depict the percentage of metabolites that change over the circadian cycle. Orange lines depict the relative accumulation of the metabolite in *Clock*-/- livers, while blue lines depict the relative metabolite concentration in WT control livers. A majority of amino acid and xenobiotic metabolites peaked at night (ZT15-ZT21) while peak abundance of nucleotide, carbohydrate, and lipid metabolites occurred at ZT9. Oscillation of amino acid and xenobiotic metabolites was temporally cohesive with the peak hours of energy intake. Of the 538 total metabolites measured, 172 showed a time main effect ($P < 0.05$, ANOVA) and 132 showed a genotype main effect ($P < 0.05$, ANOVA). Of 309 named metabolites, 100 showed a time main effect and 73 showed a genotype main effect. More information about the experiment procedure and results on the analysis can be found in Eckel-Mahan *et al.* (2012).

A

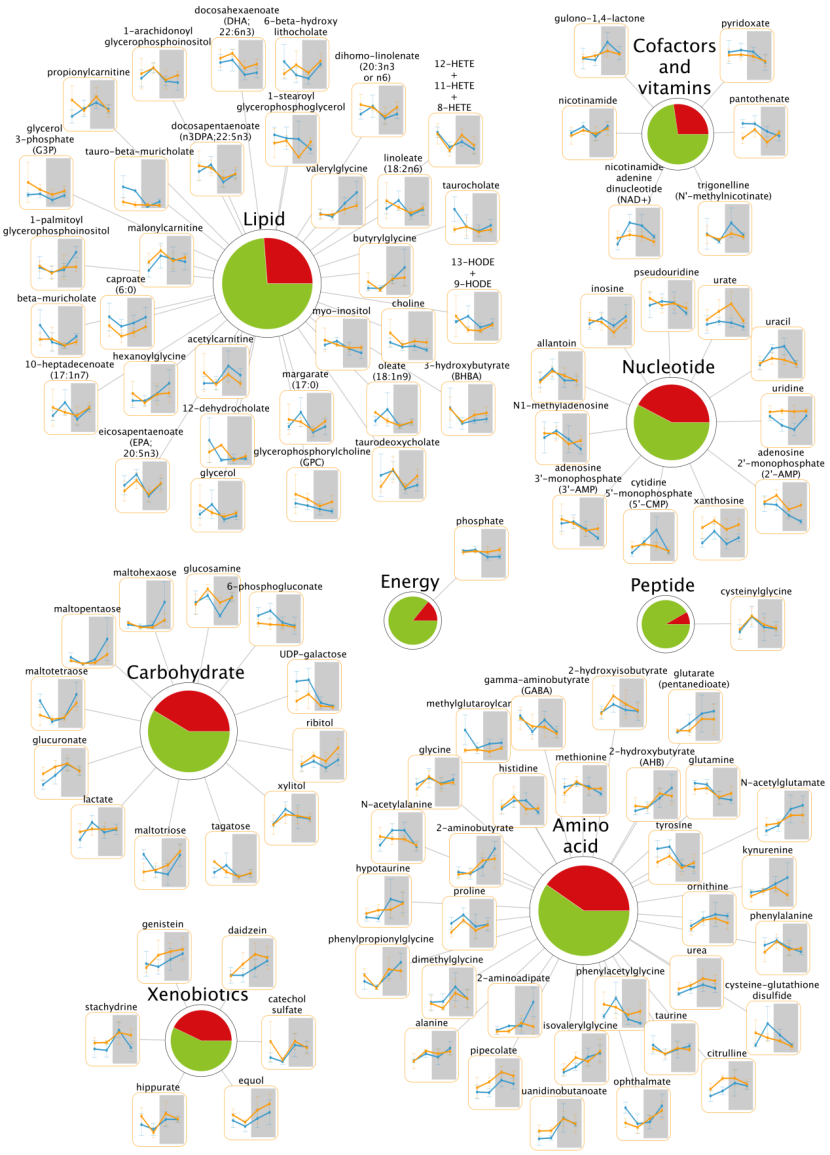


Figure 3.1: **Metabolic pathways of the liver containing diurnally regulated metabolites.** Major metabolic pathways are represented in the liver by numerous metabolites that change in abundance throughout the 24-h cycle (orange lines, Clock–/– metabolites; blue lines, WT metabolites). Metabolites that vary in abundance over time are shown. Pie charts depict the percentage of metabolites that changed over time (red) vs. those that did not (green). N = 5 per genotype per time point.

3.2.1 A Clock-driven metabolome

It has been reported that Clock -deficient mice are rhythmic in the brain due to the presence of the CLOCK paralog, NPAS2 (Debruyne *et al.*, 2006). However, peripheral clocks are arrhythmic in the absence of systemic cues. In the presence of systemic cues (i.e. in vivo), the patterns of liver gene expression are altered in a gene-specific manner (Debruyne *et al.*, 2006; DeBruyne *et al.*, 2007b,a). Clock-deficient mice model a situation in which the local liver circadian clock is impaired. Several food-derived biochemicals were altered in Clock-deficient mice, leading to the hypothesis that a circadian phenotype in energy intake might occur in these mice. Thus, the liver circadian metabolome provided valuable information regarding the metabolic and physiological functions of the circadian clock in vivo.

3.2.2 Constructing a map of the circadian metabolome

To understand this complex metabolic landscape in the backdrop of existing information and the role played by the central clock machinery- we generated comprehensive network of all proteins/enzymes and metabolites using Crick. In addition, experimental datasets listed in Table 3.1 were overlaid on the networks. Finally using Crick, the paths between Clock and all metabolites were enumerated from comprehensive networks; the gene expression of enzymes from Hughes *et al.* (2009) was correlated against the metabolite levels; connected metabolite levels were correlated; and transcription factors regulating enzyme levels were found. One of the top ranking hypothesis that came out of this analysis was the regulation of Uracil levels by the CLOCK transcription factor. This typifies the prominent interfaces found in this study between the circadian liver metabolome and transcriptome and directly demonstrates the interdependence of the metabolome and the hepatic circadian clock.

Specifically, within the uracil-containing pyrimidine metabolic pathway, uridine oscillated in WT livers (P=0.008, ANOVA) while a remarkable absence in oscillation was observed in

Experimental Data	Reference
Liver metabolite levels measured every 6 hours	Eckel-Mahan *et al.* (2012)
Liver transcript levels measured every hour	Hughes *et al.* (2009)
Liver BMAL1 ChIP-seq	Rey *et al.* (2011a)

Table 3.1: **Experimental datasets included in Crick for the analysis of the liver metabolome**

Clock-deficient mice (Fig. 3.2A and C). Uracil showed an oscillation in WT mice that was antiphase to uridine and, like uridine, did not oscillate in the liver of Clock-deficient animals (Fig. 3.2C). Uridine levels were generally elevated in the livers of mutant mice, particularly at ZT9 and ZT15 (ZT9 *Clock*-/- vs. WT= 1.550, $P < 0.001$; ZT15 Clock–/– vs. WT, $P < 0.001$) relative to WT liver concentrations. Values of uracil in liver samples showed a time and genotype main effect (P=0.012 and $P < 0.001$, respectively, two-way ANOVA), and uridine values showed time and genotype main effects (P=0.008 and P=0.001, respectively, two-way ANOVA) as well as a genotype:time interaction (P=0.008). Within the metabolic node, a number of regulatory enzymes, including *Upp2* were implicated (Fig. 3.2A). *Upp2* (uridine phosphorylase 2) is essential for pyrimidine salvage and catalyzes the reversible reaction of uracil and ribose 1-phosphate (or deoxyribose 1-phosphate) into uridine (or deoxyuridine) and orthophosphate (Fig. 3.2B) (Cao and Pizzorno, 2004; Giorgelli *et al.*, 1997). To assess the expression of *Upp2* in vivo, liver *Upp2* mRNA was analyzed in WT and Clock-deficient mice. Correlating with the levels of uracil and uridine in vivo, *Upp2* oscillates in a circadian fashion (P=0.0002, two-way ANOVA, N=5-8 livers per zeitgeber time, per genotype) with a peak at ZT15 ($P < 0.05$, Bonferroni posttest). *Upp2* oscillation was reduced in Clock-deficient mice, and levels of normalized *Upp2* expression were particularly low at ZT15 (Fig. 3.2E) (main effect of genotype, P=0.020, two-way ANOVA). The depression of *Upp2* in Clock-deficient mice correlates with the general increase in their hepatic uridine levels. Based on the enzymatic function of *Upp2*, an increase in uridine levels in the Clock-deficient mice would be expected with a lowering of *Upp2* expression. Conversely, uracil levels were depressed in Clock -/- mice at both ZT9 and ZT15, correlating with the lack of uridine

metabolism. Thus, circadian expression of *Upp2* is cohesive with the observed oscillations in *Upp2*-related metabolites. While *Upp2* mRNA expression showed a strong circadian profile, UPP2 protein also showed an oscillation throughout the circadian cycle (Fig. 3.2F). Using data from MotifMap (Xie *et al.*, 2009; Daily *et al.*, 2011), Crick was able to predict regulatory motifs in the promoter of *Upp2*. Of interest are the E-box elements that Clock and Bmal1 can bind to. Furthermore, Bmal1's binding to the promoter of *Upp2* was confirmed using the ChIP-seq data from Rey *et al.* (2011a) that was included in the network construction. We found the presence of three E-boxes and one canonical E-box between the regions of -143 and -183 of the transcriptional start site. Chromatin immunoprecipitation analysis using an antibody directed against CLOCK revealed that CLOCK binds to this region of the *Upp2* promoter in a circadian fashion, with a peak at ZT15 (Fig. 3.2G). To verify that CLOCK binding followed the expected profile of binding on other target sites in the same livers, CLOCK binding to the *Dbp* promoter was analyzed in the same WT and Clock–/– livers. As expected, CLOCK binding to upstream *Dbp* regulatory elements showed the expected circadian profile, with a peak at ZT9. Thus *Upp2* expression appears to be directly regulated by the circadian clock machinery in a circadian fashion and in a manner that supports the temporal profile of *Upp2* mRNA expression.

3.2.3 CircadiOmics

We also developed CircadiOmics, a computational resource that provides a common repository for circadian cycle related metabolic and transcriptomic data and provides high-resolution biological networks displaying, for instance, metabolites, enzymes, transcription factors and their interactions, and concentration changes over time throughout the circadian cycle. Integration of multiple data sources allows the generation of new testable hypotheses on the interactions of metabolic and circadian processes. The CircadiOmics web server allows users to search for metabolites or genes (Figure 3.3) and view an information-rich, tissue-specific

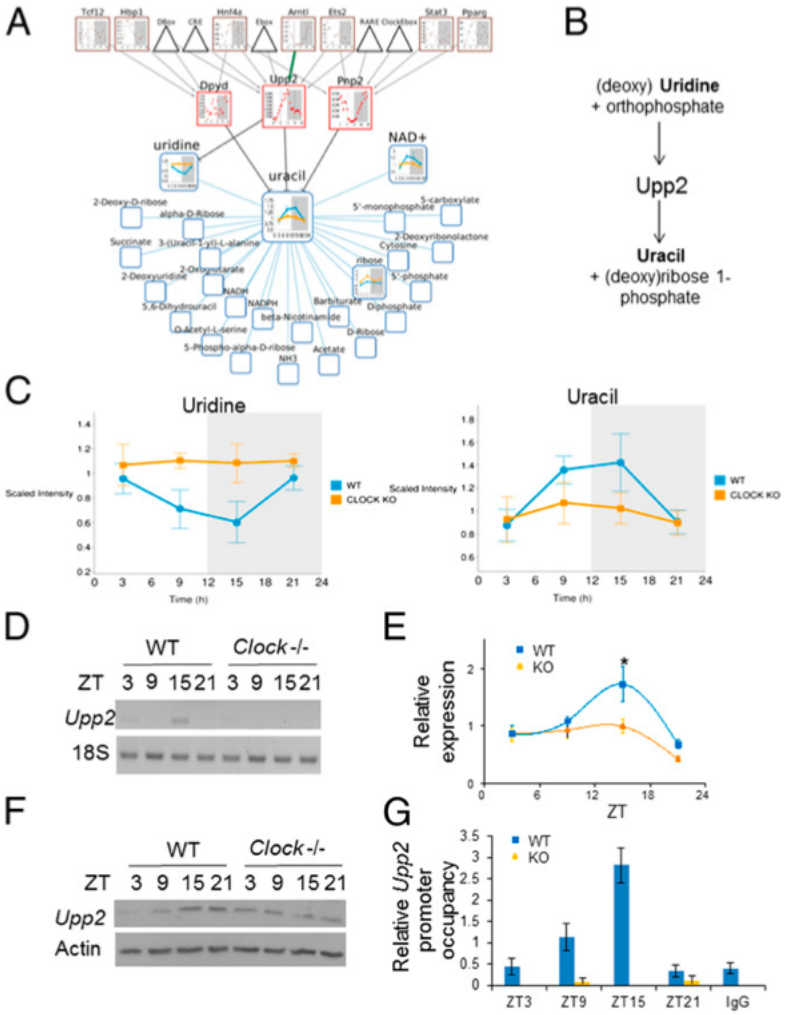
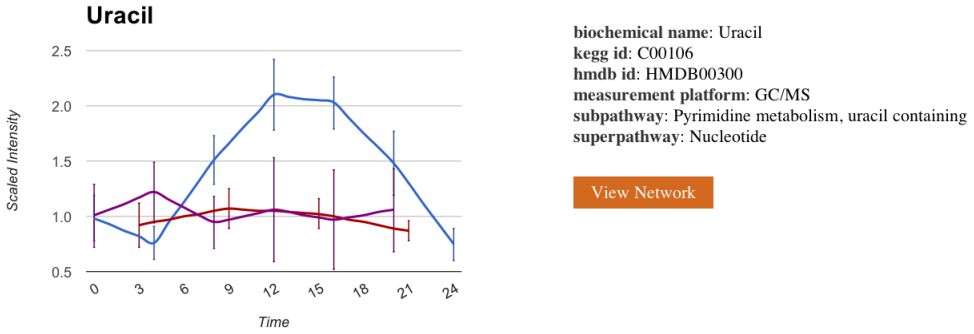


Figure 3.2: **Uracil network built using Crick.** (A) The uracil network predicts an interaction with uridine phosphorylase 2 (UPP2). (B) UPP2 participates in the reversible reaction whereby uridine and orthophosphate are converted to uracil and ribose 1-phosphate. (C) Uridine and uracil oscillate in an antiphase pattern in WT livers but are nonoscillatory in Clock–/– livers. (D and E) Semiquantitative PCR of Upp2 and quantification of Upp2 mRNA in WT (+/+) and *Clock*–/– livers (error SEM). (F) UPP2 and Actin protein expression in WT and *Clock*–/– livers (each band represents five pooled livers) (G) Diurnal binding of CLOCK to the Upp2 promoter. Immunoprecipitation of CLOCK from liver homogenates and quantification of CLOCK-bound target DNA by qPCR normalized to Upp2 input DNA.



Plot?	Tissue/Condition	JTK.p-value	JTK.q-value	Periodicity	Amplitude	Lag
✓	Liver, WT	0	0	24	0.70551	16
✓	Liver, Clock-/-	0.32146	1	23	0.11587	7
✗	Liver, High Fat	0.00002	0.00033	28	0.33676	12
✓	Muscle, WT	1	1	22.18182	0.07048	5.63636
✗	Muscle, Bmal-/-	0.82944	1	24	0.06154	14

Figure 3.3: **Screenshot of CircadiOmics**

biological network from within the web browser. See Figure 3.4 for the gene-centric network of Ass1 overlaid with transcriptomics, proteomics and metabolomics data. CircadiOmics addresses the recently highlighted challenges and opportunities of how to integrate metabolomics with other omics. This resource provides a systems-wide picture taken from a high-resolution lens, allowing researchers to analyze metabolism with a broader perspective and to generate hypotheses on how levels of metabolites may change under various physiological settings. CircadiOmics is available at http://circadiomics.ics.uci.edu/ and allows users to visualize and interactively manipulate these metabolite centric graphs.

3.2.4 The Circadian Clock as a Connector between Metabolome and Transcriptome

Circadian physiology reveals that strong connections between the circadian clock and cellular metabolism exist, but the extent to which this occurs and the nature of these interactions have been largely unappreciated. This study reveals the first integrated map of the circadian metabolome in the liver and provides a detailed depiction of the dynamic interfaces between

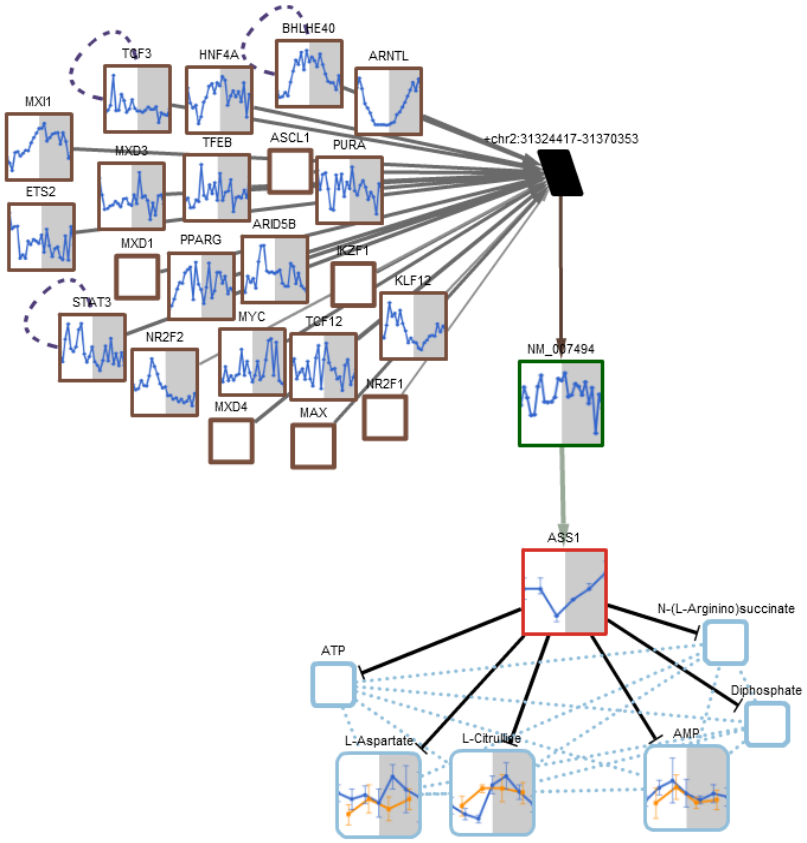


Figure 3.4: **Gene-centric network of *Ass1* overlaid with several experimental datasets**

the metabolome, proteome and transcriptome. The implications are numerous. In addition to revealing scores of oscillating metabolites in the liver, this study suggests that a fraction of the liver metabolome is dependent on local effects of the circadian clock transcriptional machinery. However, as evidenced by metabolites that oscillate in WT and Clock-deficient mouse livers but whose normalized expression differs, it is likely that some metabolic pathways depend more on zeitgebers (such as food or light). In some cases, such as in the production of uracil from uridine, the transcriptional regulation of the enzymes responsible for this reaction is altered in Clock-deficient mice. As analysis by MotifMap identified a potential D-box, E-boxes, and RRE elements in the regulatory region of *Upp2*, it is likely that the aberrant circadian expression is directly related to the disrupted circadian transcriptional network within the Clock-deficient mouse livers. Indeed, the immunoprecipitation of CLOCK from the *Upp2* promoter in WT but not Clock-deficient livers provides direct evidence that this is the case. Furthermore, the decreased and relatively flat expression of *Upp2* in the liver of mutant mice is reflected by the parallel alteration in the levels of uridine and uracil. While CLOCK-mediated gene transcription clearly contributes to metabolic homeostasis in the liver, the altered pattern of food-derived metabolites in the livers of Clock-deficient mice reflects the deviant pattern of food consumption observed in indirect calorimetry experiments. Metabolic homeostasis at the local level then appears to be strongly driven by food acting as a zeitgeber as well as by the classical transcriptional events associated with circadian rhythmicity. The aberrant pattern of food biochemicals (such as essential fatty acids, bile acids, and xenobiotic metabolites) in the Clock-deficient livers strongly suggests that the SCN and hypothalamic cues associated with rhythmic food intake are disrupted in these animals, in a manner that drives early waking, food consumption, and therefore metabolite fluctuations. The molecular events responsible for this advanced circadian food consumption may be related to the causes underlying human diseases such as night eating disorder (NES) where individuals wake during the night to consume food (O'Reardon *et al.*, 2004; Goel *et al.*, 2009; O'Reardon *et al.*, 2005). While circadian disruption occurs in this

situation, the molecular events responsible for this diurnal eating pattern are not yet clear.

3.3 Reprogramming of the Circadian Clock by Nutritional Challenge

The molecular mechanisms by which a high fat diet (HFD) affects the circadian clock are not known. Using high-throughput profiling of the liver metabolome and transcriptome we establish that HFD has multifaceted effects on the clock, including a phase advance of metabolite and transcript oscillations which are maintained on the diet, as well as an abolition of otherwise oscillating transcripts and metabolites. In addition to these disruptive effects, we find a surprising, elaborate induction of newly oscillating transcripts and metabolites. Thus, HFD has pleiotropic effects that lead to a reprogramming of the metabolic and transcriptional liver pathways. These are mediated both by interfering with CLOCK:BMAL1 recruitment to chromatin and by inducing the de novo oscillation of PPARγ-mediated transcriptional control at otherwise non-cyclic genes.

3.3.1 Reprogramming of the circadian metabolome by high-fat diet

To understand how altered nutrients affect circadian metabolism, we explored the effect of HFD in mice by studying the hepatic metabolome, where a large number of metabolites are circadian or clock-controlled (Dallmann et al., 2012; Eckel-Mahan et al., 2012; Kasukawa et al., 2012). After ten weeks on a HFD, mice displayed expected metabolic features. Importantly, the timing and quantity of energy intake was similar between feeding groups. Metabolome profiles were obtained by MS/MS and GC/MS from livers isolated every four

hours throughout the circadian cycle (Evans et al., 2009). A large number of metabolites across several metabolic pathways displayed changes in HFD-fed animals. Of 306 identifiable metabolites, 77% showed a diet effect and 45% showed a time effect (Figure 3.5A and Figures S2). When analyzed for circadian oscillations, 141 metabolites cycled in abundance. Of these, 61 metabolites (43%) oscillated in both feeding conditions (Both), while 42 metabolites (30%) oscillated only in normal chow-fed animals (NC). Importantly, 38 metabolites oscillated only in HFD-fed animals (HF) (Figure 3.5B). Many of the metabolite changes were present at ZT12 and ZT16 (Figure 3.5C) and included numerous nucleotide, amino acid, and xenobiotic metabolites (Figure 3.5D). The metabolite peak profiles differed across several of the metabolic pathways throughout the circadian cycle (Figure 3.5E). Interestingly, the phase and amplitude of remaining oscillatory metabolites also differed. Specifically, metabolites that oscillated in both feeding conditions generally showed a shift in phase when in HFD (Figure 3.5F). Of the phase-shifted metabolites, 28% were delayed in phase, while 72% were phase-advanced in HFD (Figure 3.5F-H). Considering the phase of metabolites that oscillated only in NC or only in HFD, metabolites that oscillated only in HFD tended to peak earlier (Figure 3.5G).

A majority of metabolite oscillations previously shown to be CLOCK-dependent (Eckel-Mahan et al., 2012) are affected by HFD (Figure 3.5I). As seen in our previous experiments, specific metabolic subpathways are circadian. For example, lysine metabolism is highly rhythmic in normal feeding conditions (Eckel-Mahan et al., 2012). In this study, lysine metabolism was highly rhythmic in both feeding conditions. Specifically, glutarate, lysine, 2-aminoadipate, and pipecolate showed oscillatory abundance in both conditions (http://circadiomics.ics.uci.edu/). On the other hand, pyrimidine metabolism displayed rhythmicity only under NC condition. For example, cytidine 5-monophosphate (5-CMP), 2-deoxycytidine, and 2-deoxycytidine 5-monophosphate all lost oscillation in HFD.

Strikingly, HFD completely blocked oscillation of nicotinamide adenine dinucleotide (NAD+)

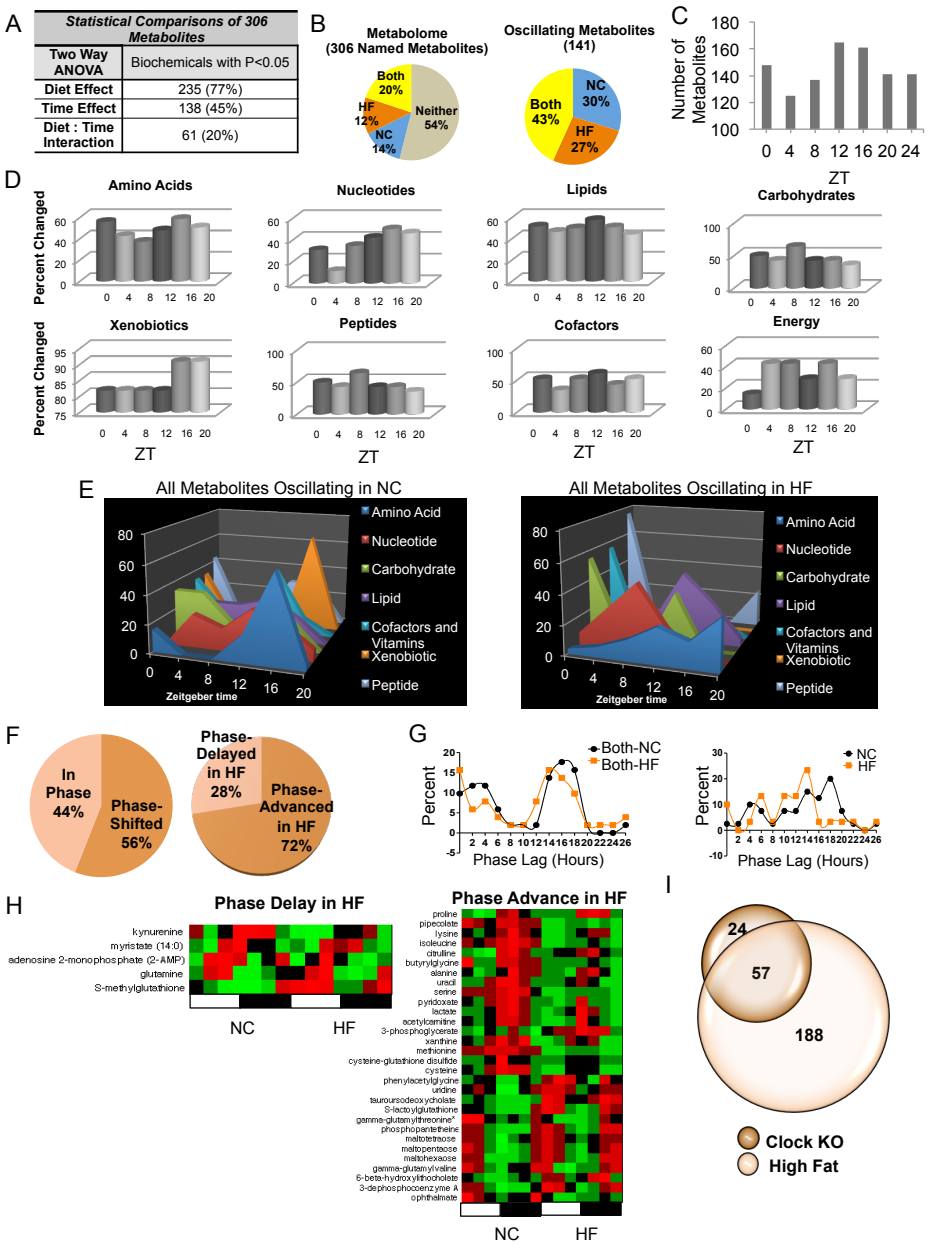


Figure 3.5: **HDF alters the Circadian Profile of the Metabolome.** (A) Number of hepatic metabolites affected by diet or time. (B) The hepatic circadian metabolome consists of metabolites that oscillate in both groups of animals regardless of diet (Both), metabolites that oscillate only in animals fed normal chow (NC), and metabolites that oscillate only in animals fed HFD (HF). $p < 0.05$, JTK_cycle, and $n = 5$ biological replicates. (C) The number of hepatic metabolites altered by the HFD at each zeitgeber time (ZT). (D) Percent of metabolites in a metabolic pathway changing at a specific ZT in HF animals. (E) Metabolic landscapes depict the percent of oscillatory metabolites that peak at a specific ZT for each feeding condition compared to the total number of oscillatory metabolites in that metabolic pathway. (F) Proportion of metabolites that oscillate on both diets that are in phase or phase shifted (left) and the direction of the phase shift (right). (G) Phase graph of metabolites that oscillate in both conditions (left) or only in the NC or HF conditions (right). (H) Heat maps depicting phase-delayed or phase-advanced metabolites in HF livers. (I) Overlap of metabolites that are both CLOCK dependent and sensitive to a HF diet.

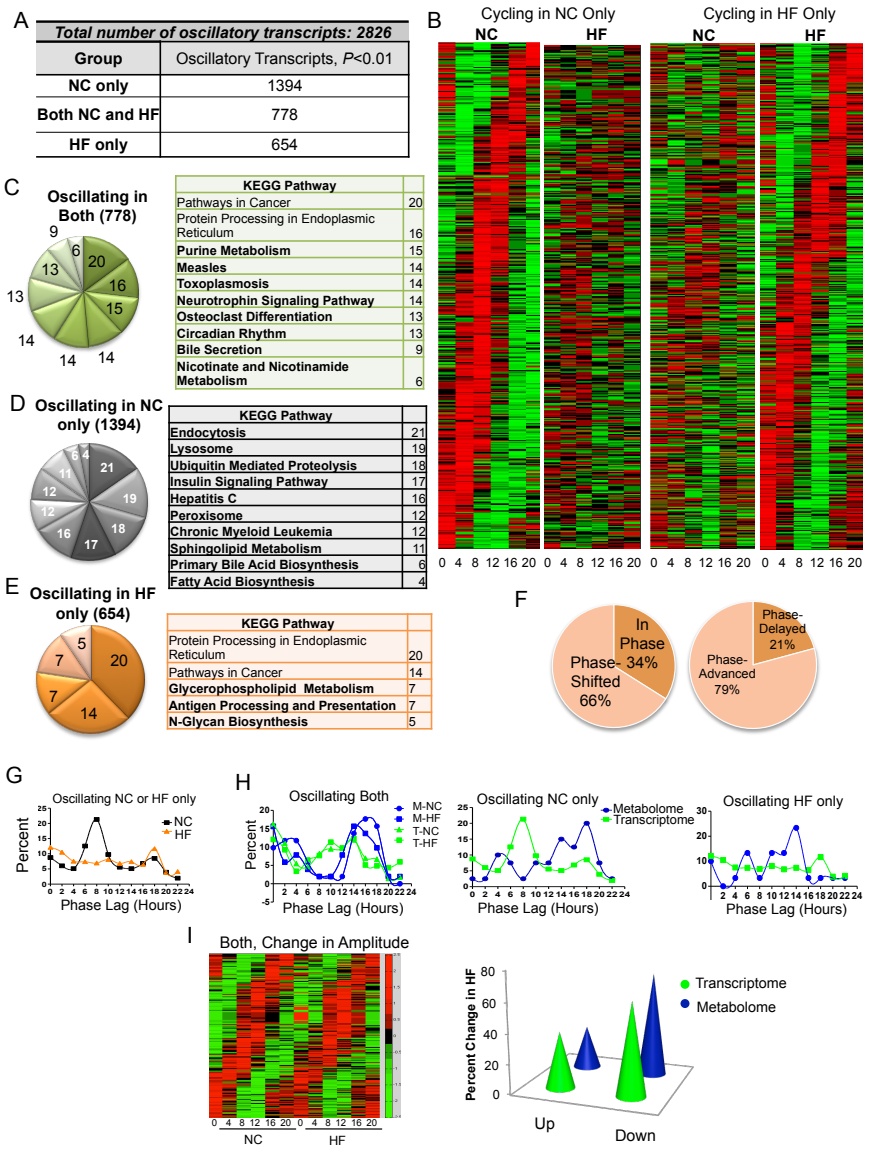


Figure 3.6: **The Circadian Transcriptome Is Reprogrammed by a HFD.** (A) The number of oscillatory transcripts only in NC, only in HF, or in both NC and HF groups ($p < 0.01$, JTK_cycle). (B) Heat maps for NC- and HF-only oscillating transcripts ($p < 0.05$). (C) Gene annotation on oscillating genes with a $p < 0.01$ reveals pathways that are oscillatory in both NC and HF livers (unique pathways in bold font). (D) Pathways in which oscillatory expression is lost by the HF diet. (E) KEGG pathways represented by genes oscillatory only in the HF liver. (F) Proportion of the oscillatory transcriptome shared in both liver sets that is phase shifted (left) and the direction of the phase shift (right). (G) Phase analysis of transcripts that oscillate only in NC or HF. (H) Circadian fluctuations of the metabolome relative to the transcriptome in both (left), NC-only (middle), or HF-only categories (right). (I) Extent of amplitude changes in transcript abundance (heat map and graph) and metabolites (graph) after HF feeding.

(Figure 3.8A). A previous report demonstrated reduced hepatic NAD+ under HFD (Yoshino et al., 2011). Thus, HFD may modulate its negative influence on energy balance by eliminating circadian oscillations in NAD+, rather than inducing a static decrease in total NAD+ content. The lack of circadian NAD+ accumulation under HFD supports the observation that NAD+ is high during fasting (Rodgers et al., 2005) . Animals fed a HFD may never achieve such an energy-depleted state due to the constant and non-oscillatory levels of glucose. The molecular mechanism leading to the impairment in NAD+ oscillation in HFD constitutes a paradigm of clock transcriptional reprogramming through the control of the *Nampt* gene (Figures 3.7B and 3.8E).

A large number of lipid metabolites were affected by HFD. Coenzyme A, a cofactor involved in fatty acid synthesis and beta oxidation, displayed a circadian profile in HFD that was substantially increased in amplitude, as did its precursors phosphopanthetein and 3-dephosphocoenzyme A. Many amino acid metabolites continued to oscillate in both conditions, even though their relative abundance was substantially reduced by the HFD, likely due to increased gluconeogenesis. We conclude that the high fat diet impinges on the circadian metabolome in three possible manners: ablation, phase-advancement, or promotion of oscillation for specific metabolites.

3.3.2 Reprogramming the Circadian Transcriptome

We analyzed the circadian transcriptome using the same liver samples used for the metabolome. In all, 2,799 transcripts oscillated in expression; of these 49.5% (1394) were rhythmic only in the NC condition (Figure 3.6A). An additional 778 were rhythmic in both NC and HF conditions and a surprising 654 were newly oscillating exclusively in HFD (Figure 3.6A and 2B). When analyzed for singular enrichment in metabolic pathways, we found that genes oscillating in both NC and HF showed unique annotations including purine metabolism and

circadian rhythm (Figure 3.6C). The persistence of circadian clock gene oscillation in both NC and HFD validates the notion that circadian oscillation within the core clock genes is highly resistant to perturbation, while clock output genes are more sensitive to food as a zeitgeber (Damiola et al., 2000). Metabolic pathways whose oscillation was uniquely lost in HFD included ubiquitin mediated proteolysis and insulin signaling (Figure 3.6E). The 654 transcripts that gained rhythmicity exclusively in HFD attracted our attention. Only five annotation groups were found, with glycerophospholipid metabolism, antigen processing and presentation, and N-glycan biosynthesis as the uniquely oscillating pathways. Three members of the oligosaccharyltransferase complex (OSTC) were among this newly oscillating group, including a subunit of the Oligosaccharyltransferase Complex Homolog A, Ostc. Importantly, specific N-glycans have been observed to be substantially elevated in the serum of db/db mice and in the serum of human subjects with type 2 diabetes (Itoh et al., 2007). Increased expression of glycan biosynthesis genes has also been observed in serum from humans with type 2 diabetes (Das and Rao, 2007). As 27.6% of all rhythmic genes oscillated in both NC and HF conditions, we analyzed their phase of expression. Of these 778 genes, 34% oscillated in phase while 66% were phase-shifted by HFD (Figure 3.6F). Remarkably, most of the oscillatory transcripts in this category showed a phase profile that was, as for the metabolome, phase-advanced in HFD. Only 21% showed a phase delay while 79% of the shifted transcripts showed a phase advance (Figure 3.6F). Analysis of the phase of transcripts that oscillated only in NC or HFD, revealed starkly different profiles. Specifically, oscillatory transcripts in the NC-only group showed robust peaks between ZT4-ZT12, while the HFD group showed rather an irregular phase pattern (Figure 3.6G). The peak of oscillatory transcripts in the NC-only condition is consistent with when the CLOCK:BMAL1 heterodimer is most active and recruited to circadian contacts (Hatanaka et al., 2010; Kondratov et al., 2003; Rey et al., 2011). When comparing the phase of the transcripts and metabolites that oscillated in both liver sets (Figure 3.6H, left), similar organization was seen, with transcripts and metabolites showing a biphasic pattern and transcript peaks slightly preceding metabo-

lite peaks. Metabolite and transcript oscillations that were lost in HFD (i.e. oscillating in NC only) also showed a temporal organization with transcripts peaking prior to the majority of metabolites (Figure 3.6H, middle). However, coherence was not complete in HFD, with most metabolites peaking at ZT14, and transcripts remaining largely unsynchronized in phase (Figure 3.6H, right). In addition to phase analysis, we studied the oscillation amplitude of both transcripts and metabolites (Figure 3.6I). Of all common oscillators, 62% of the 778 transcripts showed a reduction in amplitude in HFD, while 38% showed an increase. Similarly, 71% of metabolites showed a reduction in amplitude in HFD, while 29% showed an increase. We conclude that the percentage of the circadian metabolome and transcriptome are similarly affected in amplitude by HFD, stressing the coherence between these two groups of oscillators.

3.3.3 Coherence of metabolome and transcriptome

We determined the relationship within metabolic pathways between the transcriptome (of the enzymes) and the metabolome using Crick and the data was included in CircadiOmics (Patel *et al.*, 2012). We classified and grouped metabolite-enzyme edges based on the presence or absence of oscillation as well as additional characteristics of the oscillation, specifically, the phase and amplitude (Figure 3.7A). The most common edge characterization (87 of 384 edges, 23%) revealed that the loss of oscillation for a particular metabolite usually was accompanied by a loss of oscillation for its related transcripts (Figures 3.7A-B). Interestingly, the second most common edge classification involved the loss of oscillatory transcript abundance in HFD but an increase in the amplitude of oscillation in the related metabolite. No phase delay in the transcriptome or metabolome was observed within the top ten edge classification scenarios, suggesting again that a significant effect of HFD is to phase-advance the remaining oscillatory metabolites. Edge classification reinforced the notion that one of the effects of HFD is to reorganize the temporal coherence between the metabolome and transcriptome.

The most common relationships between related transcripts and metabolites involved an opposing state of oscillation in animals in HFD (Figure 3.7A and 3B). A paradigmatic example of a metabolite whose loss of oscillation by HFD is accompanied by a dampened oscillation for its related transcript is NAD+. Circadian NAD+ synthesis depends on the transcriptional control by the clock of *Nampt* gene expression (Nahakata et al. 2009; Ramsey et al. 2009). HFD induces a loss of NAD+ oscillation that parallels a dampening of *Nampt* cyclic transcription (Figure 3.7B; and 3.8E-F). Additional case scenarios include ornithine decarboxylase 1 (Odc1) and ornithine (where a concomitant loss of oscillation occurs in HFD), acyl-CoA synthetase short-chain family member 2 (Acss2) and coenzyme A (where loss of oscillatory transcript in HFD corresponds to an increased metabolite amplitude) and cytochrome P450 monooxygenase (Cyp2a5) and arachidonate (where a phase advance in transcript in HFD corresponds to a lack of oscillation in its related metabolite) (Figure 3.7B).

Importantly, several metabolite and transcript edges within individual pathways mirror each other in HFD-induced gain of oscillation. Remarkable examples are within the amino acid subpathway of cysteine, methionine, S-adenosylmethionine (SAM) and taurine metabolism. Indeed, both SAM and S-adenosylhomocysteine (SAH) showed newly oscillating profiles in HFD (Figure 3.7C). HFD-induced cycling of these metabolites was accompanied by de novo oscillation of several related enzymes, including Ehmt, Trmt2b, Whsc1, Dph5 genes whose products have known or predicted methyltransferase activity. A relevant case is Ahcyl2, the gene encoding the enzyme that catalyzes the reversible conversion of SAH to adenosine and homocysteine, whose oscillation parallels the one of SAH in HFD (Figure 3.7D). Each metabolite and transcript identified in the livers of animals fed NC or HFD were integrated within the computational resource, CircadiOmics (Eckel-Mahan et al., 2012; Patel et al., 2012).

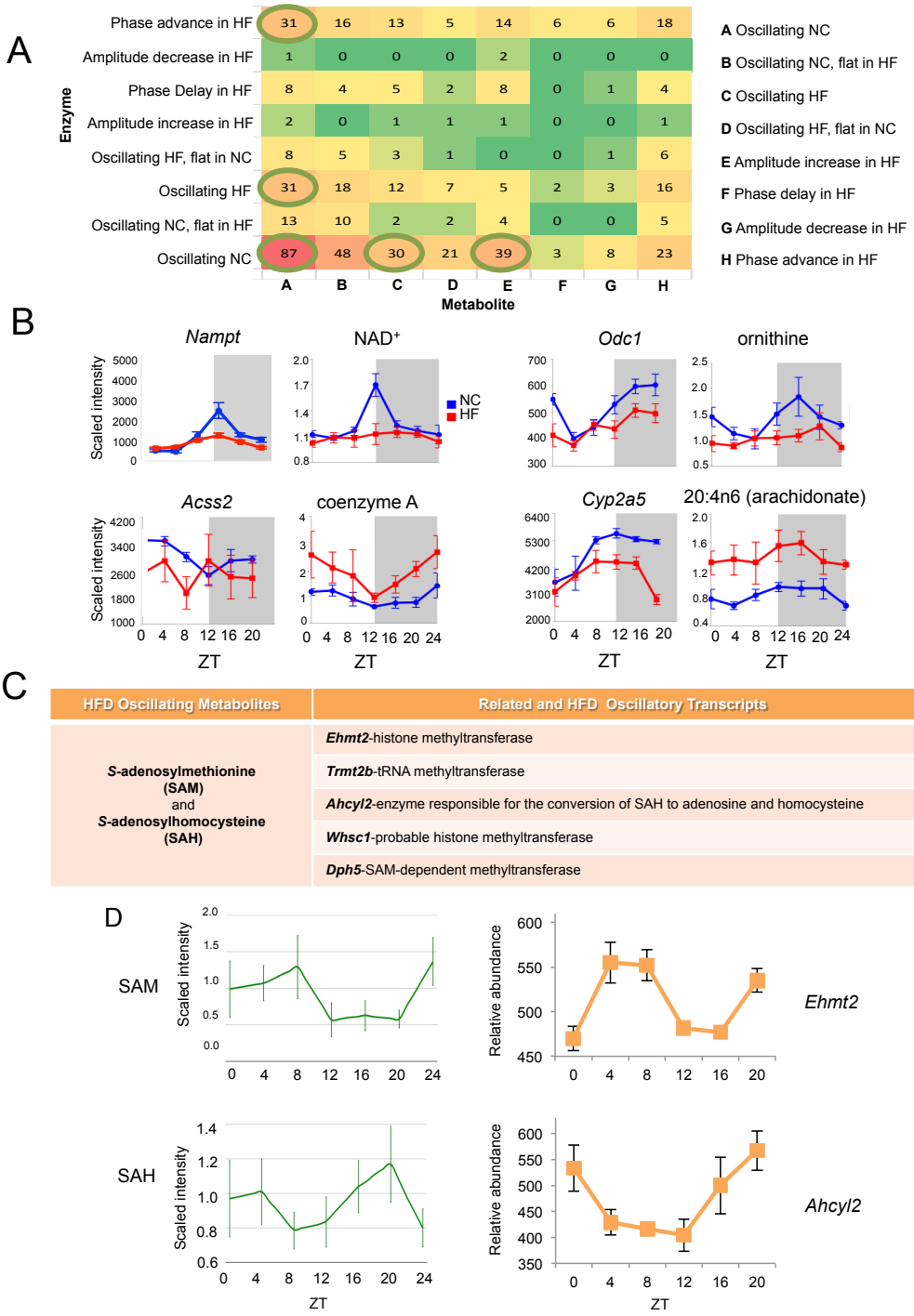


Figure 3.7: **HFD Disrupts Circadian Organization between the Transcriptome and Metabolome**. (A) Heat map showing the relationships between all pairs of metabolites and enzymes in KEGG. (Note: flat is a subset of not, where the maximum abundance does not exceed the minimum by 20%.) Circled are the numbers referring to the five most common relationships. (B) Related enzyme transcripts and metabolites (edges) that follow a particular temporal profile. (C) Metabolites and related transcripts within the SAM node that gain oscillation in HF. (D) Oscillatory abundance of SAM, SAH, and their related enzymes Ehmt2 and Ahcyl2 only in HF. Error bars, SEM.

3.3.4 High-fat diet hinders CLOCK:BMAL1 chromatin recruitment to target Genes

Activation by CLOCK:BMAL1 of target gene promoters has been linked to oscillations of a number of metabolites (Eckel-Mahan et al., 2012; Nakahata et al., 2009; Ramsey et al., 2009). Thus, we investigated the molecular mechanisms by which circadian oscillations are disrupted by HFD. First, we hypothesized that HFD might alter core clock gene expression. Importantly, most of the core circadian genes were rhythmic in the livers of HFD-fed mice (Figure 3.8A), displaying only weak shifts or slightly dampened patterns of oscillation, results which are cohesive with previously published work (Hatori et al., 2012; Kohsaka et al., 2007). Per2 and Bmal1 mRNA showed mild dampening and phase advancement, while Clock expression was unaffected. One case scenario is represented by the gene *Dbp*, whose robust circadian oscillation was phase-advanced in HFD (Figure 3.8A). Since Clock and Bmal1 cyclic transcription is similar in HFD-fed mice (Figure 3.8A), we analyzed protein levels. Importantly, the levels of BMAL1 (Figure 3.8B) and CLOCK proteins were unaltered in livers of HFD-fed animals. Similarly, the phosphorylation profiles of BMAL1 in NC and HFD conditions were similar in different cellular fractions. We next explored whether CLOCK:BMAL1 chromatin recruitment might contribute to the altered pattern of *Dbp* expression by chromatin immunoprecipitation (Ripperger and Schibler, 2006). Remarkably, BMAL1 and CLOCK recruitment was shifted in livers of HFD-fed mice (Figure 3.8C).

Interestingly, transcripts whose oscillation was lost in HFD were in large part peaking between ZT4 and ZT12 (Figure 3.6G), a time period that correlates with prominent CLOCK:BMAL1 recruitment to chromatin targets (Hatanaka et al., 2010; Kondratov et al., 2003; Rey et al., 2011). This parallels the dampening or abrogation of the oscillations of numerous metabolites previously shown to be CLOCK:BMAL1-regulated (Figure 3.8D). A remarkable example is NAD+, whose cyclic levels become flat after HFD, paralleling the profile of *Nampt* transcription (Figure 3.8D-E). Similarly, the oscillations of the metabolites uridine and uracil,

the abundance of which is dependent on the enzymatic activity of CLOCK:BMAL1-driven uridine phosphorylase 2 (*Upp2*) expression (Eckel-Mahan et al., 2012), were depressed in the livers of HFD-fed animals (Figure 3.8D). The amplitude of *Upp2* expression was considerably reduced under HFD (Figure 3.8E). Interestingly, we observed a substantial decrease in CLOCK and BMAL1 circadian occupancy on the *Upp2* and *Nampt* promoters in livers of animals fed a HFD (Figure 3.8F-G). Importantly, oscillation in H3K4me3, a histone modification tightly associated with circadian transcription (Katada and Sassone-Corsi, 2010; Ripperger and Schibler, 2006), significantly decreased at the *Upp2* (Figure 3.8H) and *Nampt* (not shown) promoters in HFD-fed animals. Thus, the profound effect elicited by HFD is caused by either phase-shifted or reduced recruitment of the CLOCK:BMAL1 complex to chromatin at the level of target promoters.

3.3.5 High-fat diet induced reprogramming of the clock by PPARγ

The disruptive effect of HFD on CLOCK:BMAL1 chromatin recruitment (Figure 3.8E-G), does not explain the de novo rhythmicity gained by a large group of genes in HFD (Figure 3.6B). Notably, with the exception of the group of genes whose oscillation is lost under HFD, the group of newly oscillating genes is the largest, over doubling the number of genes that showed phase advancement as a result of the HFD. Transcription factor motif analysis by MotifMap (Xie *et al.*, 2009; Daily *et al.*, 2011) was performed on a regions located 10kb upstream and 3kb downstream of the transcriptional start sites to determine what transcriptional pathways might be most heavily affected by the HFD. Using a Bayesian branch length score of 1 or greater, E-boxes were significantly enriched in genes oscillating only under NC as well as in genes oscillating under both NC and HFD conditions. On the contrary, no enrichment for E-boxes was observed in the group of genes oscillating exclusively in HFD condition. Analysis of the frequency of specific transcription factors binding sites in the promoters of genes whose oscillation was induced by HFD, revealed that HFD promotes the use

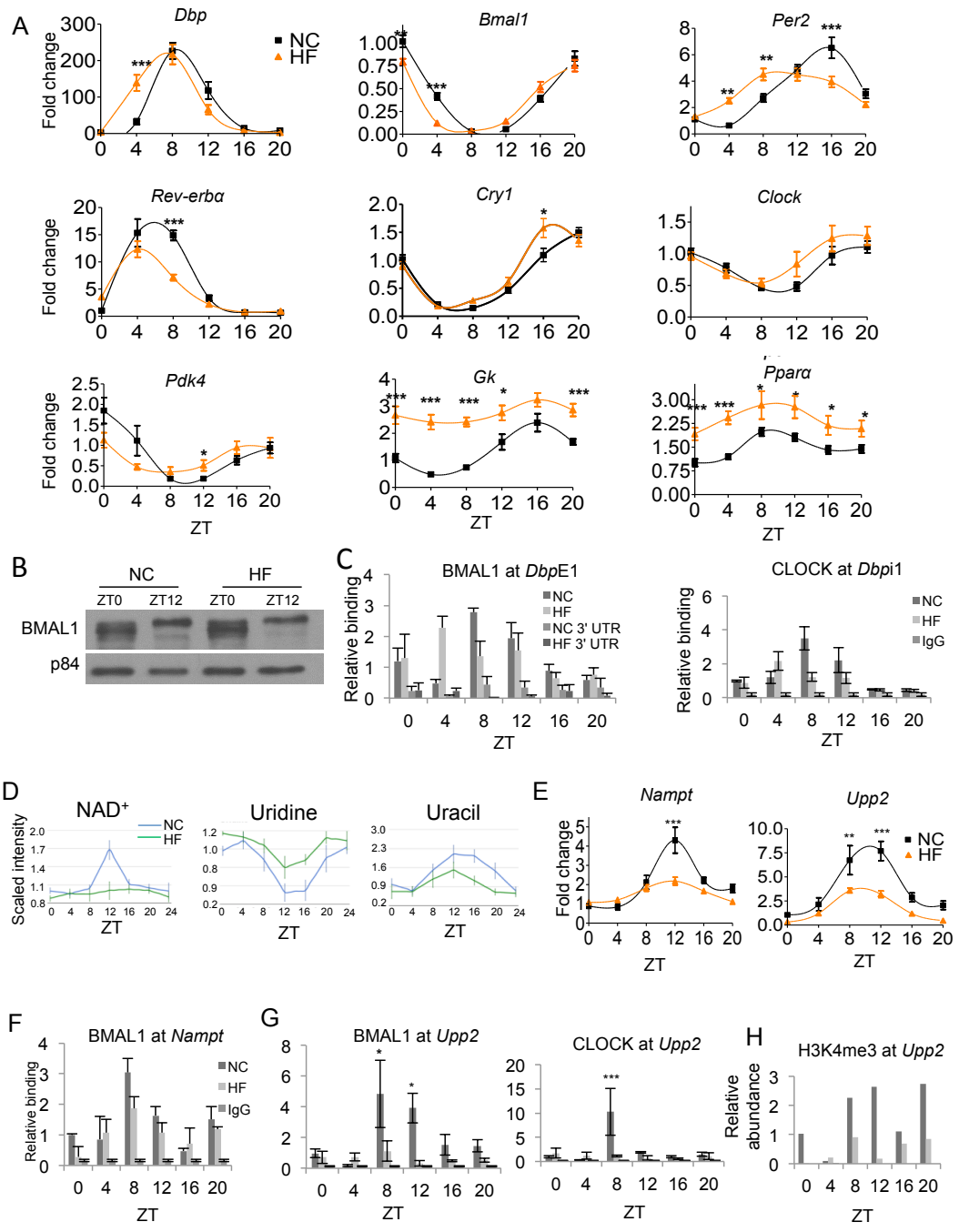


Figure 3.8: **HFD Disrupts Circadian Organization between the Transcriptome and Metabolome**. (A) Heat map showing the relationships between all pairs of metabolites and enzymes in KEGG. (Note: flat is a subset of not, where the maximum abundance does not exceed the minimum by 20%.) Circled are the numbers referring to the five most common relationships. (B) Related enzyme transcripts and metabolites (edges) that follow a particular temporal profile. (C) Metabolites and related transcripts within the SAM node that gain oscillation in HF. (D) Oscillatory abundance of SAM, SAH, and their related enzymes EHMT2 and AHCYL2 only in HF. Error bars, SEM.

of additional transcriptional pathways to reprogram the hepatic transcriptome. One of the most represented transcription factors in the newly oscillating group of genes was PPARγ . Several other transcription factors oscillated in HF only, which included SREBP-1 (Srebf), CREB1 and SRF. PPARγ and SREBP1 were identified as having one or more target sites in 322 and 91 genes, respectively. In line with the idea of increased PPARγ -mediated gene expression under HFD, metabolomics analysis revealed that PPARγ ligands were elevated in livers of HFD-fed animals, specifically 13-HODE, 15-HETE, linolenate and arachidonate (http://circadiomics.ics.uci.edu). PPARγ is a nuclear receptor involved in glucose and lipid metabolism and has been described as a nutrient sensor in metabolic tissues. We found that PPARγ expression was robustly oscillatory in the liver of HFD-fed animals, with a peak at ZT12. While the levels of total PPARγ protein were elevated, but not circadian in HFD-fed mice.

3.4 Summary

In summary, in these studies it was shown that Crick can provide very detailed, high-resolution, and tissue-specific views of the underlying molecular network. Crick can be effectively used to analyze new experimental data to identify novel mechanisms, like the link between *Clock* and Uracil.

Chapter 4

Can most transcripts and metabolites oscillate, and how?

4.1 The Pervasiveness of Circadian Oscillations

It is well known that genetically encoded molecular clocks found in nearly every cell, based on interlocked transcription/translation feedback loops and involving only a dozen genes, play a central role in maintaining these oscillations. However, high-throughput transcriptomic, proteomic, and metabolomic experiments reveal that in a typical tissue ∼10-15% of molecular species oscillate with the day-night cycle and that, beyond the core clock overlap, the oscillating species vary with different cell and tissue types. This suggests that a much larger fraction of all molecular species bear the potential for circadian oscillation. However, recent gene expression experiments reveal that a significant fraction of all transcripts in a cell oscillate in a circadian manner. High-resolution gene expression data from mice liver (Hughes *et al.*, 2009) identified over 3000 transcripts that oscillate with a 24-hour periodicity. In addition, analysis across different tissue types of the same organism reveals little

overlap between the molecular species that oscillate in one tissue versus another beyond the core clock genes. Out of 7000 genes studied, 337 transcripts were found oscillating in SCN compared with 335 in the liver (Panda *et al.*, 2002). However, when the oscillating genes in each tissue were compared, only 28 genes were oscillating in both SCN and liver. Likewise, a low amount of genes was identified as circadian in both muscle and liver (57 genes) (Miller *et al.*, 2007) as well as between heart and liver (Storch *et al.*, 2002). Similar results have also been reported with high-throughput metabolomic studies (Eckel-Mahan *et al.*, 2013, 2012). Furthermore, it has now been established that genetic, environmental, and even diet changes can modify oscillatory patterns by: (1) modifying the phase or amplitude of existing oscillations; (2) suppressing existing oscillations entirely; and (3) giving rise to new oscillations that were not observed in the absence of perturbations (Figure 1A). Indeed, comparing gene expression and metabolite levels in liver tissue from high-fat-fed versus normal-chow-fed mice found over 2,800 transcripts oscillating in expression across both conditions (Eckel-Mahan *et al.*, 2013). Of these, 27.5% (778) did oscillate in both conditions, but often with a change in amplitude and/or phase; 49.5% (1,394) were rhythmic only in the normal-chow-fed condition; and approximately 23% (654) were rhythmic only in the high-fat-fed condition. Similar results were also observed for oscillating metabolites. The results from across tissue and within tissue experiments strongly suggest that a large fraction of all molecular species is capable of oscillating in a circadian manner under some set of conditions.

Researchers have looked at the common denominator (the master clock genes and its interactors), but little has been done to understand the unique and possibly novel oscillations in a tissue or perturbation until recently. In our previous study (Eckel-Mahan *et al.*, 2013) where we compared the transcript and metabolite levels from normal-chow and high-fat diet we noticed a massive reprogramming occurring within the cell. By analyzing not only the transcripts and metabolites that lost their oscillation, but also the transcripts and metabolites that gained a new oscillation as a result of perturbation, we were able to discover compensatory oscillations in important molecular species like SREBP1 (a transcription factor

responsible for lipid synthesis). In this study, we analyze time-resolved transcriptomes and metabolomes from a large group of published and unpublished datasets.

4.2 Results

4.2.1 Comparison of transcriptomes

Analysis across different tissue types and conditions (listed in Table 4.1), reveals surprisingly little overlap between the molecular species that oscillate in one tissue versus another (Figure 4.1) beyond the core clock genes. When the results of circadian experiments conducted over different tissues and conditions (listed in Table 4.1) are aggregated with a stringent oscillatory cutoff ($P < 0.05$), over 13,600 genes are found to oscillate in at least one tissue or condition in mouse. Thus, a significant fraction ($\sim$67%) of the genes in a mammalian genome is capable of generating mRNAs which oscillate in a circadian fashion in at least one tissue or condition. Even comparing transcriptomes from perturbation experiments performed on the same tissue type (liver, see Figure 4.3) show that a major fraction of genes can oscillate under some condition. Thus, tissue-type, environmental, genetic, and even diet perturbations all lead to significant differences in the list of oscillating genes.

4.2.2 Comparison of metabolomes

Similar results are seen when time-dependent metabolite levels from different tissues and conditions (listed in Table 4.1) are aggregated (Figure 4.2). At a P-value less than 0.05, out of the 554 measured metabolites, we find 376 metabolites oscillate in at least one of the conditions. Hence $\sim$68% of the measured metabolites oscillate even with a small set of tissue/conditions explored.

We hypothesize that this may be true even at the protein level, although high-throughput circadian proteomic measurements are not yet widely available. In any case, these results from across tissue and within tissue comparisons strongly suggest that a significant fraction of transcripts and other important molecular species is capable of circadian oscillations in at least one type of tissue or condition. Noise, experimental variability, and statistical significance cutoffs to assess circadian oscillations can account for some of these numbers, but only to a limited extent. For instance, even at a more stringent p-value of 0.01, we still find ∼8,650 genes (Figure 4.4) and ∼300 metabolites (Figure 4.5) that oscillate in at least one experiment.

4.2.3 Effects of perturbation

Genetic, diet, or environmental perturbations may disrupt circadian oscillations and may:

1. Change the amplitude of the oscillations of some of the circadian molecular species;
2. Change the phase of the oscillations of some of the circadian molecular species;
3. Disrupt or even suppress the oscillations of some of the circadian molecular species.

   Indeed, experiments involving genetic knockouts, diet changes, or even simply different mice strains, show these effects (Figure 4.6). For instance, when comparing gene expression in liver tissue from Clock mutant and wild-type mice (Miller *et al.*, 2007), ∼1160 genes show a loss of circadian rhythmicity. However, ∼400 genes oscillate in both conditions, but with a difference in amplitude or phase (Figure 4.6A). Similarly when comparing gene expression and metabolite levels in liver tissue from 10-week high-fat-fed versus normal-chow-fed mice, ∼2200 genes and ∼40 measured metabolites show a loss of circadian rhythmicity, while ∼1520 genes and ∼60 measured metabolites oscillate in both conditions, but with a difference in amplitude or phase (Eckel-Mahan

et al., 2013).

4. Most importantly, these perturbations will also create new circadian oscillations for some of the molecular species. These oscillations are only superficially new in the sense that the potential for oscillatory behavior was already present and in a sense is revealed by the perturbation.

4.2.4 Emergence of new oscillations

When comparing liver samples from *Clock* mutant and wild-type mice, ∼240 genes oscillate in the mutant but not in the wild type. Interestingly, these new oscillations arise in spite of mutating the *Clock* gene. Similarly, when comparing liver samples from mice fed with high-fat chow versus normal-chow, ∼1110 genes and ∼40 measured metabolites oscillate in the high-fat but not in the normal-chow condition (Figure 4.6B). Important transcription factors, enzymes, and metabolites are found among the new oscillating species. For example, SREBF1 a key transcription factor regulating enzymes involved in lipid synthesis shows a new, robust (P = 0.00001), circadian oscillation in the high-fat condition.

Remarkably, novel oscillations are also seen in data collected from the same tissue of wild-type mice, but corresponding to different genetic strains. Specifically, a comparison of liver tissue from C57BL/6J (Eckel-Mahan *et al.*, 2013) and C57/B6 + Black Swiss (Masri *et al.*, 2014) mice strains, uncovers ∼2840 genes that oscillate only in C57BL/6J, ∼890 genes that oscillate only in C57/B6 + Black Swiss, and ∼900 genes that oscillate in both strains (Figure 4.6C). This was true also when comparing liver samples from two animals with a slightly different mixture of C57/B6 and Black Swiss strains (see Liver *Sirt1* WT and Liver *Sirt6* WT in Figure 4.1). Note that this may in part explain also why differences in oscillatory behavior are seen in assays that are presumed to be identical (e.g. wild-type liver tissue), but in reality are not, due to significant differences between wild-type strains (Yalcin *et al.*,

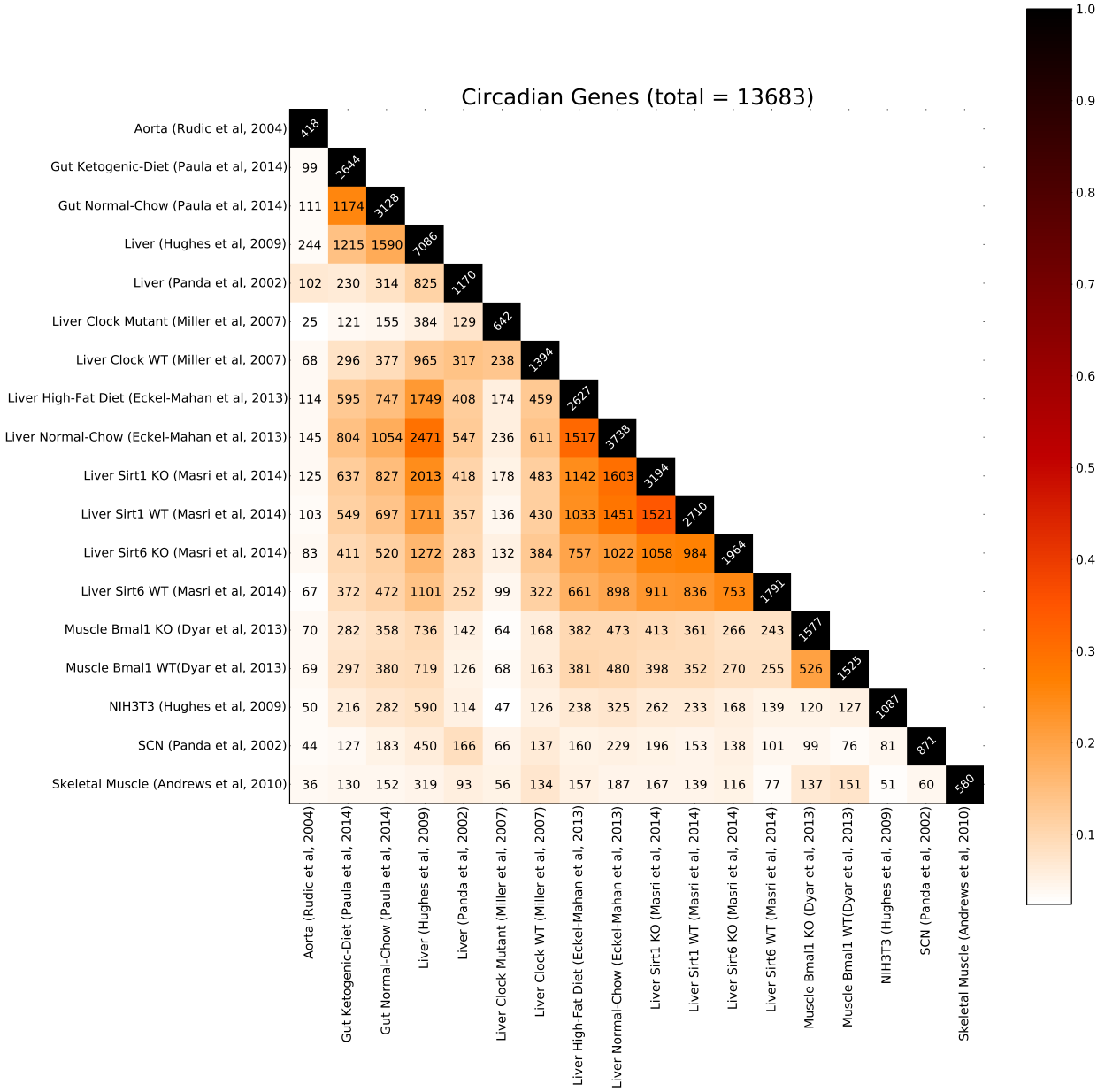


Figure 4.1: **Pairwise comparison matrix across 18 transcriptomic experiments at P $<$ 0.05**. The numbers correspond to the number of oscillating genes/metabolites ($P \leq 0.05$) that are common to both tissues/conditions (i.e. $|A \cap B|$). The color intensity corresponds to the Tanimoto-Jaccard index ($|A \cap B|/|A \cup B|$). In total, there are 13683 (∼67%) genes that oscillate in at least one tissue or condition.

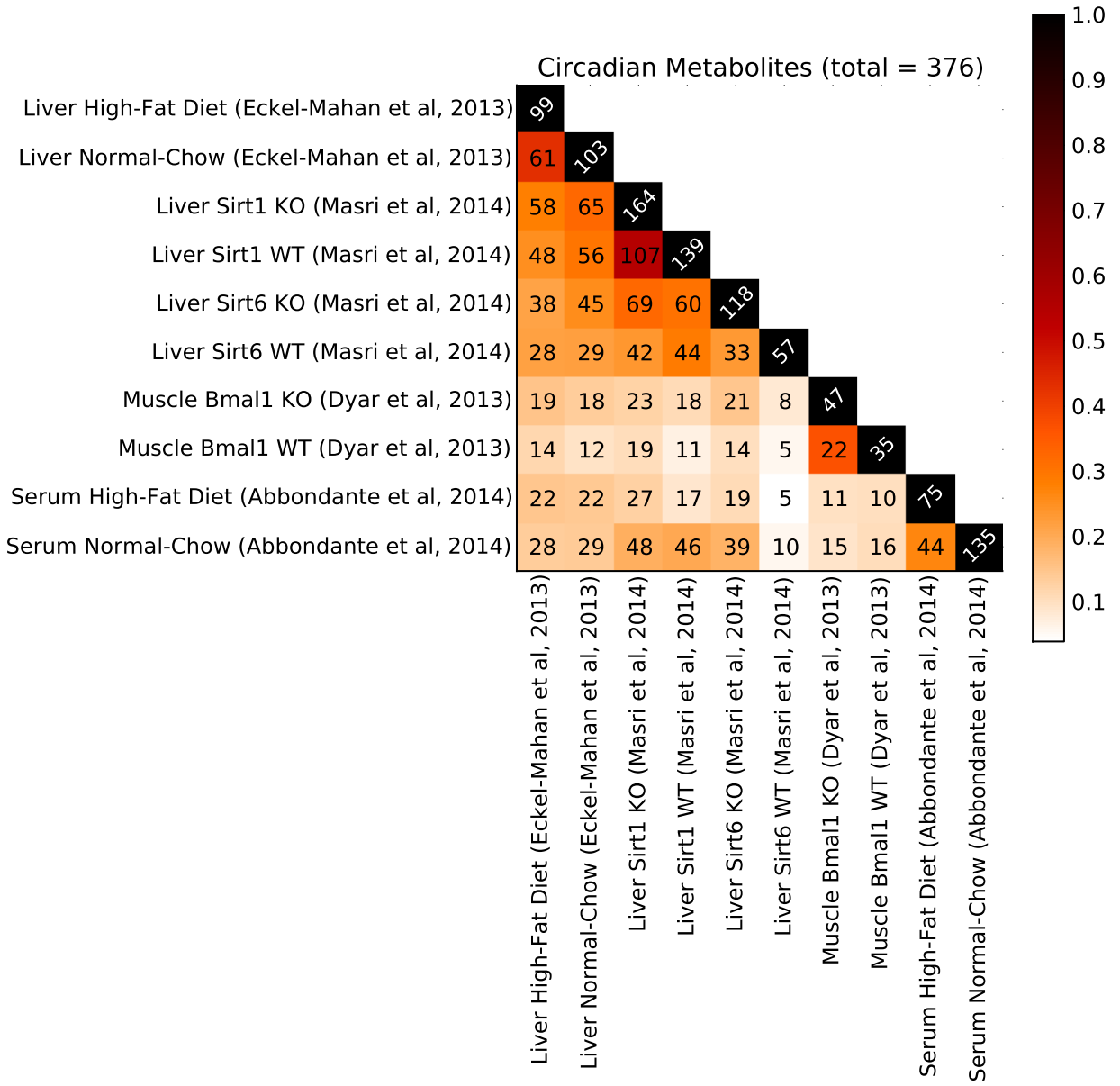


Figure 4.2: **Pairwise comparison matrix across 10 metabolomic experiments at P $<$ 0.05**. The numbers correspond to the number of oscillating genes/metabolites ($P \leq 0.05$) that are common to both tissues/conditions (i.e. $|A \cap B|$). The color intensity corresponds to the Tanimoto-Jaccard index ($|A \cap B|/|A \cup B|$). In total, there are 376 (∼68%) measured metabolites that oscillate in at least one tissue or condition.

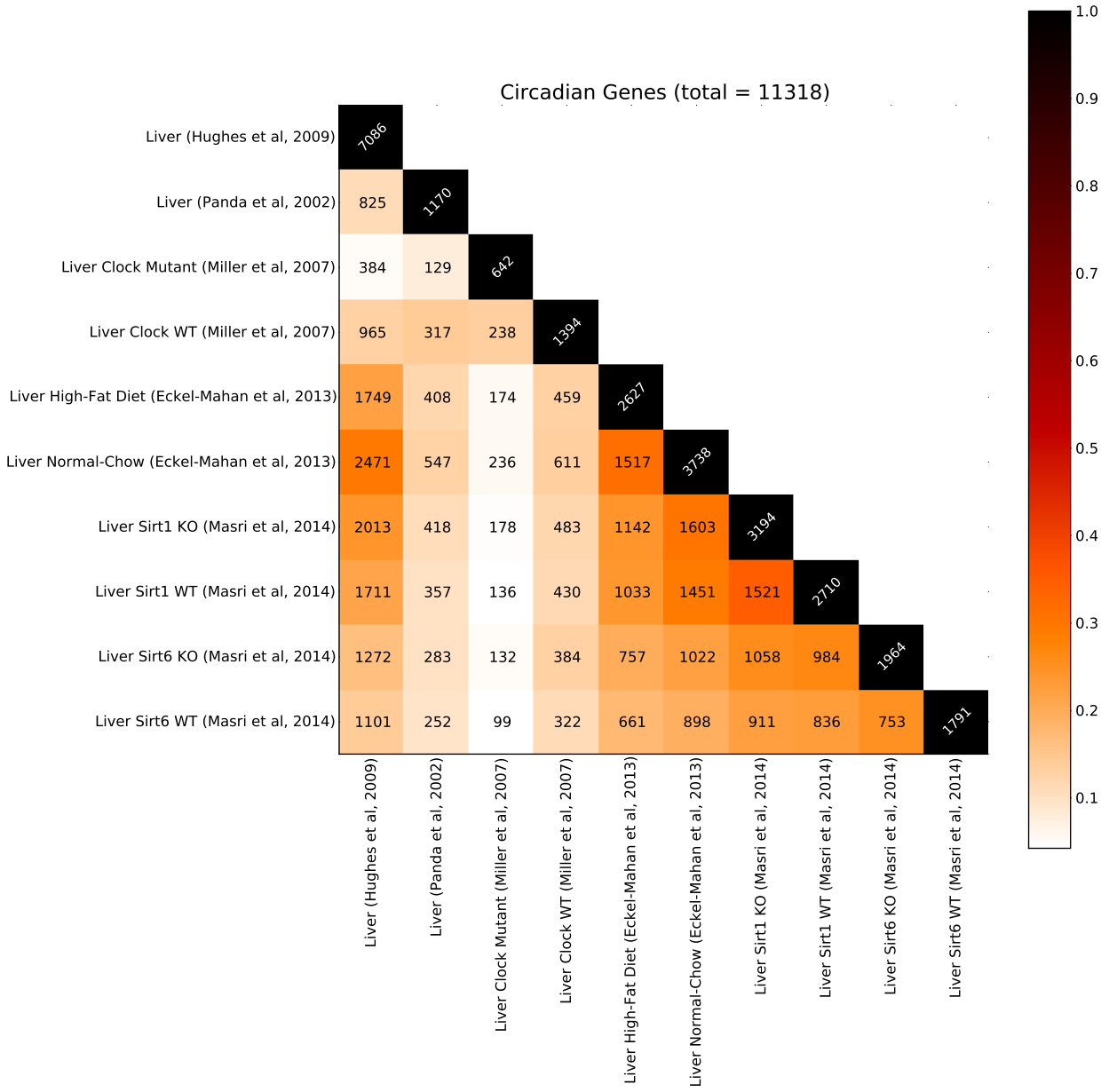


Figure 4.3: **Pairwise comparison matrix across all perturbations experiments from Liver tissue at P $<$ 0.05** The numbers correspond to the number of oscillating genes/metabolites ($P \leq 0.05$) that are common to both perturbations (i.e. $|A \cap B|$). The color intensity corresponds to the Tanimoto-Jaccard index ($|A \cap B|/|A \cup B|$). In total, there are 11318 ($\sim$56%) genes that oscillate in at least one perturbation/condition.

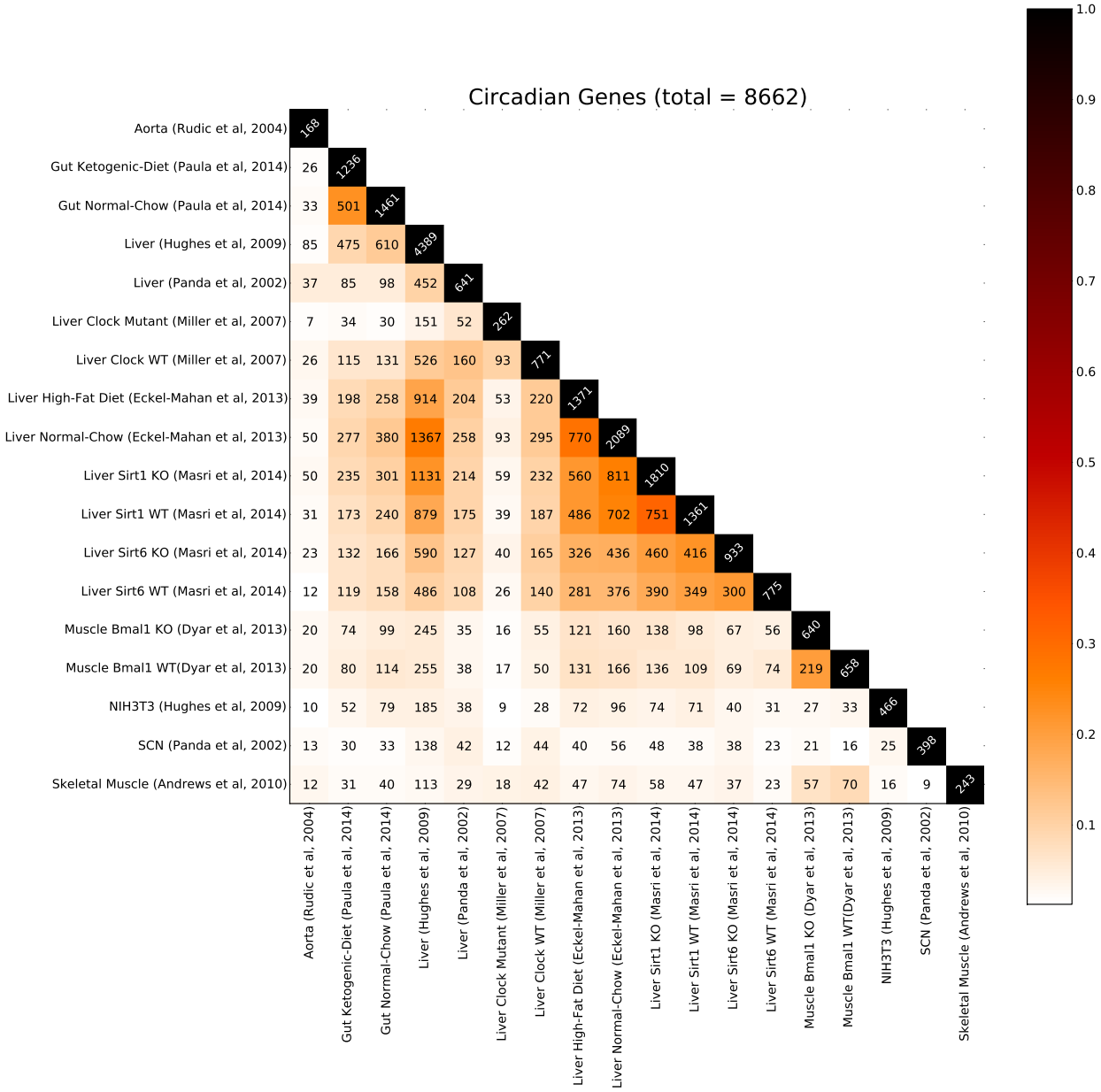


Figure 4.4: **Pairwise comparison matrix across 18 transcriptomic experiments at P < 0.01**. The numbers correspond to the number of oscillating genes/metabolites ($P \leq 0.01$) that are common to both tissues/conditions (i.e. $|A \cap B|$). The color intensity corresponds to the Tanimoto-Jaccard index ($|A \cap B|/|A \cup B|$). In total, there are 8662 (∼43%) genes that oscillate in at least one tissue or condition.

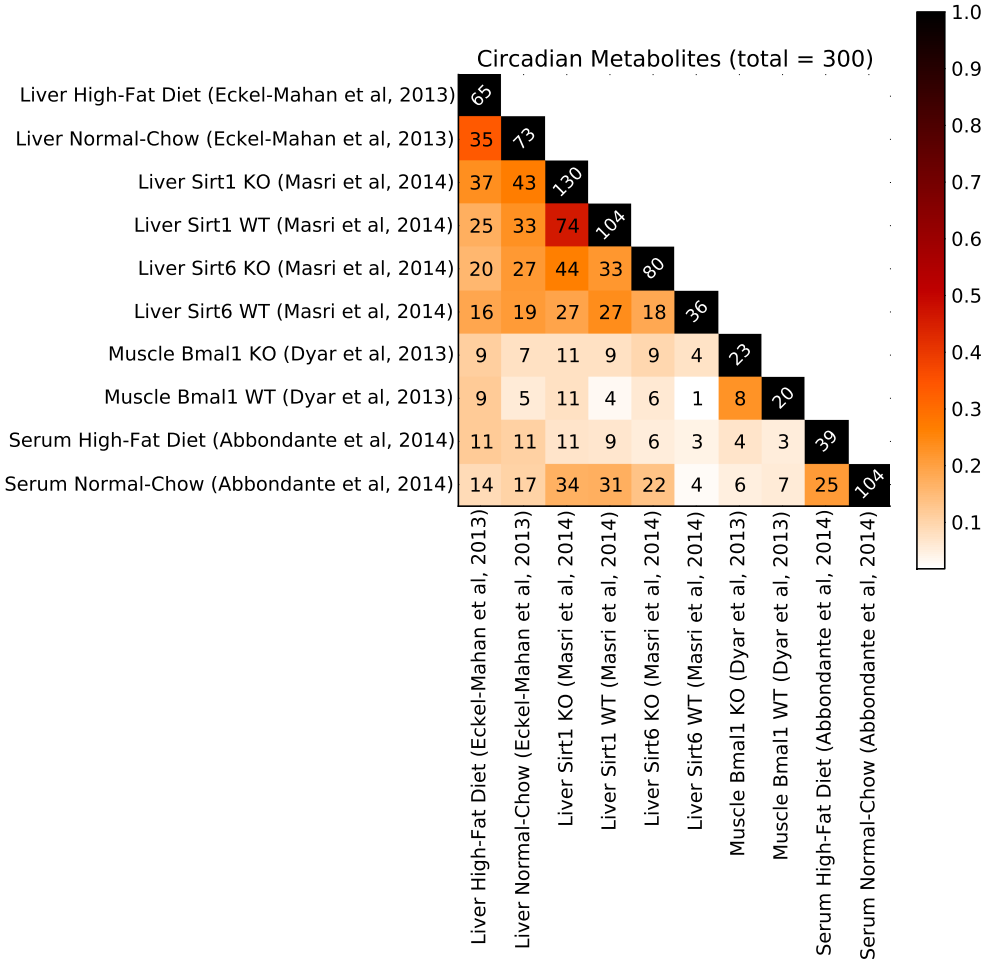


Figure 4.5: **Pairwise comparison matrix across 10 metabolomic experiments at P $<$ 0.01**. The numbers correspond to the number of oscillating genes/metabolites ($P \leq 0.01$) that are common to both tissues/conditions (i.e. $|A \cap B|$). The color intensity corresponds to the Tanimoto-Jaccard index ($|A \cap B|/|A \cup B|$). In total, there are 300 ($\sim$54%) measured metabolites that oscillate in at least one tissue or condition.

2012) and the environments in which they are raised.

4.2.5 Molecular oscillators are directed loops with the same period

In general molecular species in isolation cannot oscillate. What really oscillates are directed loops of interacting molecular species involving various interactions such as regulatory, protein-protein, and enzymatic interactions. In biological networks, there are a large number of such directed loops. For instance, in a network consisting of 21826 genes/proteins with 120988 edges (114493 regulatory edges and 6495 physical protein-protein interactions), we found over 3600 directed loops of size 3 and over 71100 directed loops of size 4. Furthermore, high time resolution circadian data (Hughes *et al.*, 2009) shows that most oscillating genes have a period of approximately 24-hour, with some genes oscillating at harmonic periods of about 12 hours and 8 hours. Very short periods (e.g. periodicity of one hour or less) and periods not commensurate with the day-night cycle (e.g. periodicity of 5 hours) are probably not physiological and thus not observed. The key question then is why do most directed loops exhibit the same 24h periodicity and how does the cell achieve this coupling? The first part of the question can be answered by the fact that the world is drastically different during day and night and since the beginning of life Earth has rotated on its axis over a trillion times ($3.5 \times 10^9 \times 365 = 1.3 \times 10^{12}$), inducing a relentless 24h day-night cycle that has been deeply sculpted in living systems. The second part of the question can be answered by the fact that biological networks are densely connected and several oscillating directed loops share edges between them forming an intricate network of coupled oscillators which can be reprogrammed by modifying the sign of the interactions; creating new interactions; and destroying existing interactions. The coupling in such a network is likely to be non-linear, condition-specific, and heterogeneous.

4.2.6 Role of core clock genes in coupled oscillator network

To understand the factors that contribute to the capability of the cell to reprogram the oscillatory signature, we analyzed the role of the core clock genes in the context of the underlying global molecular network. Using the network with all regulatory edges and protein-protein interactions we calculated the distance of all nodes from *Clock* or *Bmal1* and also the total number of directed loops that *Clock* or *Bmal1* is part of. We found that ∼10% of genes are one hop away and ∼60-70% genes are two hops away from *Clock* or *Bmal1* (see Table 4.2). In addition, there are several directed loops that *Clock* or *Bmal1* is a part of. About ∼10% of genes are connected to *Clock* or *Bmal1* through a directed loop (see Table 4.3) of size 6 or less. Hence in the global molecular network, *Clock* and *Bmal1* are centrally located to modulate circadian rhythms.

4.3 Discussion

Over a trillion day-night cycles have deeply impregnated the 24h periodicity in living systems at all levels: from the molecular to the cellular, to the organism, and beyond. It is interesting to note that the earth is constantly losing angular velocity and rotational energy through a process called tidal acceleration, which leads to a slow lengthening of the day. For instance, about 620 million years ago a day had only about 21.9±0.4 hours. This slow change is helping evolution sculpt the periodicity in all living organisms.

When a complex physical system is significantly perturbed, in general one does not expect to see a large number of new components oscillating at the same frequency, unless these components were already primed and capable of oscillating at this particular frequency, in which case the perturbation simply reveals a preexisting capability. The important question then is why so many genes and metabolites have the potential to oscillate in a circadian

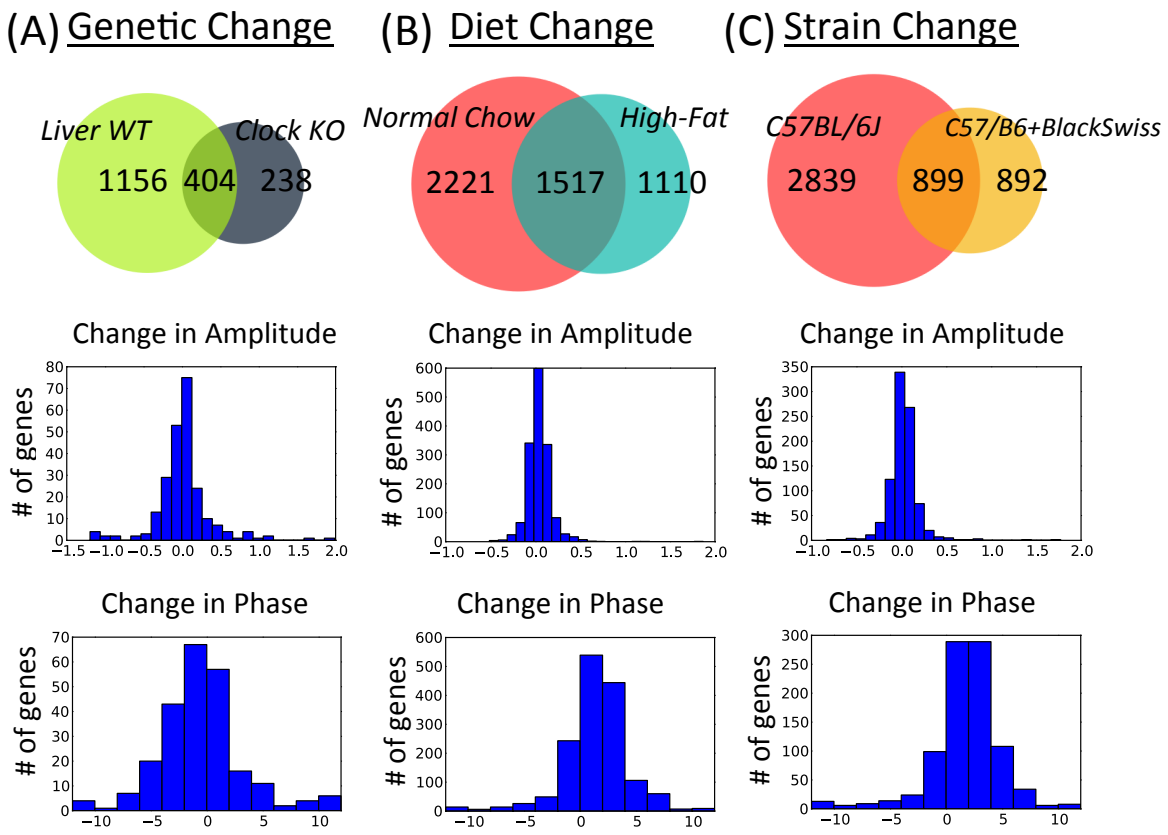


Figure 4.6: **Venn diagrams comparing different pertubations. First row**: Venn diagrams comparing: (A) wild-type and Clock mutant liver gene expression, as an example of genetic perturbation; (B) normal-chow fed and high-fat fed liver gene expression, as an example of environmental perturbation; and (C) C57BL/6J and C57/B6 + Black Swiss liver gene expression, as an example of strain perturbation. In all cases, there is massive reprogramming leading to a large number of new oscillations. **Second row**: Histogram showing the changes in amplitude. **Third row**: Histogram showing the changes in phases (measured in hours).

manner, and how does the cell select which molecular species should oscillate in a given situation?

4.3.1 Mechanism

There are several possible non-exclusive mechanisms by which the cell can create, suppress, or modify interactions between the different species cells in order to rapidly reprogram its oscillatory repertoire. For instance, dynamic changes in the epigenome, like methylation, acetylation, and chromatin remodeling can play a central role in selecting the fraction of oscillating species. In fact recent studies have identified circadian long-range interactions (Aguilar-Arnal *et al.*, 2013) and the role of *Clock* gene as a histone acetyltransferase (Takahashi *et al.*, 2008). More importantly, perhaps, cells create and destroy interaction edges by revealing or hiding transcription factor binding sites. For instance, the core molecular clock is known to bind to a single or pair of E-box sites. E-box sites are short (canonical sequence CACGTG) and frequent in the genome. With a stringent Bayesian Branch Length Score greater than 1, we found over 23800 conserved E-box sites on the genome using MotifMap– several of which are found in the promoter of transcription factors. Using time-resolved ChIP-seq data for *Bmal1*, Rey *et al.* (2011a) identified 2,049 E-box binding sites in mouse liver. Among these, ∼60% (1,319) showed a rhythmic binding and 13% of all *Bmal1* sites had a pair of E-box elements with spacers of 6-7 base pairs. Thus, in a given environment, cells can reveal or hide a fraction of E-box sites thereby controlling which loops are directly, or indirectly, affected by *Clock* and *Bmal1*.

Cells can also rapidly and massively reprogram which species oscillate by acting on densely connected molecular hubs. For example, as previously described *Clock* and *Bmal1* both are densely connected molecular hubs (see Table 4.2). Another example of this is Nicotinamide adenine dinucleotide+, a metabolite that participates in many reactions, plays a central role

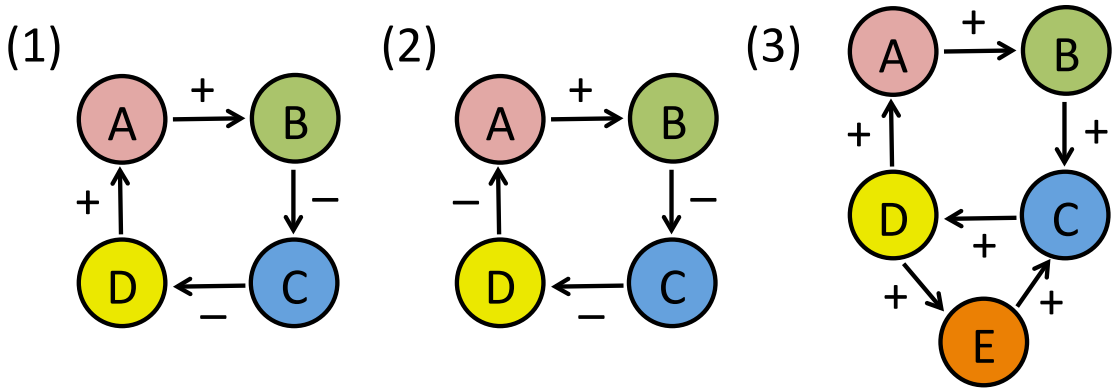


Figure 4.7: **Cycles in biological networks.** (1) A cycle between four molecular species with an even number of negative interactions. Increasing the concentration of A, increases the concentration of B, which decreases the concentration of C, which increases the concentration of D, which further increases the concentration of A (and vice versa if the concentration of A is decreased). Thus in general such a system does not oscillate and will tend to converge to one of several fixed-point attractors. (2) A cycle between four molecular species with an odd number of negative interactions. Increasing the concentration of A, increases the concentration of B, which decreases the concentration of C, which decreases the concentration of D, which then decreases the concentration of A. Thus such a system will tend to oscillate. (3) Example of two interlocked cycles one of size four and another of size three sharing one edge (between C and D) with fixed-point attractors. Changing the sign of the shared interaction creates two oscillatory cycles.

in regulating circadian rhythms (Nakahata *et al.*, 2008; Peek *et al.*, 2013; Ramsey *et al.*, 2009). All these mechanisms could set up cascades among coupled circadian oscillators by which changes in the phases and amplitudes of key oscillators could propagate to neighboring oscillators.

Currently, there is not enough data to try to model these mechanisms on a large-scale in realistic detail. Detailed mathematical models for specific molecular oscillators have been developed (Goldbeter, 1995, 1997). While useful, these are not capable of providing a system view of circadian oscillations, or making system level predictions.

4.3.2 The coupled-oscillator network framework

We propose a novel conceptual framework, whereby as a result of the ancestral and relentless circadian oscillation of the Earth, a large fraction of the molecular species in the cell is capable of oscillations under at least some set of conditions. In a given environment, through epigenetic and other modifications, a cell selects which fraction of molecular species out of its entire repertoire should exhibit circadian oscillations. This framework has important implications for understanding, and may be one day predicting, what happens when significant perturbations are applied to these networks of coupled oscillators.

Arrays of fairly homogeneous coupled oscillators have been studied in physics and other areas (Baldi and Meir, 1990; Strogatz, 2000; Goel and Ermentrout, 2002; Brandt *et al.*, 2006). A fairly general class of models can be written in the form

$$\frac{\partial \theta_i}{\partial t} = \omega_i + \sum_{i}^{n} f(\theta_i, \theta_j) \tag{4.1}$$

where θ_i is the phase of the i-th oscillator, ω_i represents its frequency, and f is the homogeneous coupling function. For instance, in the well-studied Kuramoto model the coupling is given by $f(\theta_i, \theta_j) = sin(\theta_j - \theta_i)$. Other relatively simple models could use Boolean functions or neural networks with Hill-like functions (Baldi and Atiya, 1989; Scheper *et al.*, 1999; Akman *et al.*, 2012) to model the state or concentration of a molecular species as a function of its interacting neighbors. For instance, the concentration y_i of species i could be a non-linear

function of its activation x_i with

$$y_i = f(x_i) = \frac{1}{1 + ce^{-\lambda x_i}} \qquad and \qquad \frac{dx_i}{dt} = \frac{-x_i}{\tau_i} + \sum_j w_{ij} y_j \tag{4.2}$$

here τ_i is the decay time-constant of species i and w_{ij} is a sparse matrix of weights capturing the interactions with neighboring species (Baldi and Atiya, 1989). A general key point is that feedback loops are required for oscillations. Thus in the coupled-oscillator view of molecular networks, the oscillators are not individual molecular species but rather directed loops of interacting species. Under fairly general assumptions, it is known from the theory of dynamical systems that activity in cycles with an even number of negative connections tend to converge to one of several possible stable points, whereas cycle with an odd number of negative connections lead to oscillations (Figure 4.7).

If the sign of an interaction is changed by a perturbation, a corresponding cycle could flip from stable to oscillatory behavior. This effect could extend to multiple cycles sharing the same edge (Figure 4.7) there by propagating the change of phase or amplitude; suppression of existing oscillations; and giving rise to novel oscillations. An additional prediction in this case is that these new oscillations would tend to arise in areas of the network that are adjacent to the previously existing, or newly created oscillators.

4.3.3 Comprehensive networks of circadian molecular species

Significant progress has been made in producing and visualizing comprehensive network maps of which molecular species interact and which molecular species oscillate under particular sets of conditions. CircadiOmics (Patel *et al.*, 2012) produces high-resolution biological networks displaying, for instance, metabolites, enzymes, transcription factors and their interactions,

and concentration changes over time throughout the circadian cycle. These high-resolution maps have been instrumental in predicting and understanding novel molecular mechanism (Eckel-Mahan *et al.*, 2012, 2013). Now, all of the datasets analyzed in this study have been included in CircadiOmics and similar high-resolution networks are made available for all of these experiments. Analysis of the global molecular network shows that there are large number of directed loops capable of circadian oscillations and that *Clock* and *Bmal1* are well situated to modulate circadian rhythms. While these networks are not completely accurate (for reasons such as incomplete information, false positives in predicted regulatory edges, etc.), they still provide a first-order approximation of the underlying real biological networks and are useful to derive global statistical predictions and in the development of coarser but large-scale circadian oscillation models.

4.3.4 Conclusion

In summary, as a result of the world being very different during the night versus the day, over a trillion night-and-day cycles during the course of evolution have deeply sculpted the molecular networks of the cell and made 24h oscillations pervasive. Aggregation of high-throughput transcriptomics and metabolomics experiments across tissue types and conditions reveal that a large fraction of the molecular network of a cell is primed for and potentially capable of oscillating in a circadian manner. Broadly, in a given cell, as a result of many environmental and genetics factors, the cells selects a relatively small fraction of this broad repertoire for circadian oscillations. Thus from a circadian standpoint the cell can be viewed as an intricate network of coupled-oscillators with complex couplings. It is important to note that circadian oscillations are non-rigid and highly plastic, unlike the oscillations in a physical system. This property allows an organism to quickly adapt to perturbations and environmental changes (like a change in diet or time of sleep). Perturbations of all kinds– genetic, epigenetic, and environmental– not only modify existing oscillations, but can also

induce new oscillations through these couplings, revealing the underlying oscillatory fabric that enables flexible cellular reprogramming. The complement of transcripts and other species that oscillate in a circadian manner provides a characteristic signature describing a cell and its physiological condition.

A simple consequence of having so many transcripts oscillating, or with the potential for oscillating, is that one ought to be cautious when drawing differential conclusions from gene expression experiments performed at a single time point. For instance two genes may not be significantly differentially expressed at one time point, but they may be at a different time point, if at least one of the two genes being considered behaves in a circadian manner in the corresponding assay. Ongoing and future work should provide the data to better model and understand circadian coupled-oscillator networks, predict how they respond to perturbations, and used these responses to explain biology and direct therapeutic intervention.

4.4 Methods

4.4.1 Transcriptome analysis

Time-resolved gene expression microarrays from 12 published (Hughes *et al.*, 2009; Miller *et al.*, 2007; Eckel-Mahan *et al.*, 2013; Panda *et al.*, 2002; Andrews *et al.*, 2010) and 6 unpublished experiments (Masri *et al.*, 2014; Tognini *et al.*, 2014) were gathered for analysis. The datasets were downloaded from Circa (Hughes *et al.*, 2009) or GEO (Edgar *et al.*, 2002); or provided by the authors. The transcriptomes along with tissue and perturbation condition (if any) are listed in Table 4.1. To compare gene lists across different microarray platforms/experiments we used DAVID (Huang *et al.*, 2009, 2007) for all gene/transcript ID conversion.

4.4.2 Metabolome analysis

Time-resolved metabolite levels measured using LC/GC chromatography were obtained from 4 published (Eckel-Mahan *et al.*, 2013; Dyar *et al.*, 2013) and 6 unpublished experiments (Masri *et al.*, 2014; Abbondante *et al.*, 2014). The metabolomes along with tissue and perturbation condition (if any) are listed in Table 4.1. The unique compound identifier reported by Metabolon® was used to compare the list of metabolites across the different experiments.

Table 4.1: **List of transcriptomic and metabolomic datasets analyzed**

Tissue/Condition	Strain	Transcript Data	Metabolite Data	Reference
Aorta	unknown	Yes	No	Rudic et al, 2004
Gut Normal-Chow	C57BL/6J	No	Yes	Tognini, Murakami et al, 2014
Gut Ketogenic-Diet	C57BL/6J	No	Yes	Tognini, Murakami et al, 2014
Liver	C57BL/6J	Yes	No	Hughes et al, 2009
Liver	C57BL/6J	Yes	No	Panda et al, 2002
Liver *Clock* WT	C57BL/6J	Yes	No	Miller et al, 2007
Liver *Clock* Mutant	C57BL/6J Clock homozygous mutant	Yes	No	Miller et al, 2007
Liver Normal-Chow	C57BL/6J	Yes	Yes	Eckel-Mahan et al, 2013
Liver High-Fat Diet	C57BL/6J	Yes	Yes	Eckel-Mahan et al, 2013
Liver *Sirt1* WT	Mostly C57/B6 with some Black Swiss	Yes	Yes	Masri et al, 2014
Liver *Sirt1* KO	Mostly C57/B6 with some Black Swiss - *Sirt1* knockout	Yes	Yes	Masri et al, 2014
Liver *Sirt6* WT	Mixed C57/B6 and Black Swiss	Yes	Yes	Masri et al, 2014
Liver *Sirt6* KO	Mixed C57/B6 and Black Swiss - *Sirt6* knockout	Yes	Yes	Masri et al, 2014
Muscle *Bmal1* WT	Cre-negative littermates from cross between C57BL/6 with floxed *Bmal1* and C57BL/6 mouse carrying a Cre recombinase transgene.	Yes	Yes	Dyar et al, 2013
Muscle *Bmal1* KO	Cross between C57BL/6 with floxed *Bmal1* and C57BL/6 mouse carrying a Cre recombinase transgene.	Yes	Yes	Dyar et al, 2013
NIH3T3	C57BL/6J	Yes	No	Hughes et al, 2009
Serum Normal-Chow	C57BL/6J	No	Yes	Abbondante et al, 2014
Serum High-Fat Diet	C57BL/6J	No	Yes	Abbondante et al, 2014
SCN	C57BL/6J	Yes	No	Panda et al, 2002
Skeletal Muscle	C57BL/6J	Yes	No	Andrews et al, 2010

4.4.3 Statistical circadian analysis

JTK_cycle (Hughes *et al.*, 2010) implements a nonparametric statistical test which can be used to determine cycling events. Gene expression and metabolite levels from all experimental datasets were analyzed using JTK_cycle and the corresponding P-values are reported. A gene was considered circadian, if at least one of its transcripts passed the P-value cutoff.

4.4.4 Creation and analysis of circadian networks.

To understand the global structure of the molecular network and the role played by the core clock genes in it we constructed comprehensive networks maps using CircadiOmics. The detailed method to build these networks is described in Patel *et al.* (2012). Briefly, CircadiOmics combines information about all molecular species and their interactions from several databases to provide information-rich and tissue-specific views of the underlying biological network. These networks include metabolic and enzymatic reactions, protein-protein interactions, regulatory edges from MotifMap (Xie *et al.*, 2009; Daily *et al.*, 2011) and some published ChIP experiments, etc. along with the expression profiles from experimental data. Comprehensive first-neighbor graphs were generated in-parallel and finally combined together for network analysis. Networks with all the metabolomes and transcriptomes analyzed in this paper have been included in the CircadiOmics website.

Table 4.2: **Network distance from *Clock/Bmal1***

Distance	# Genes in undirected network	# Genes in directed network with PPI	# Genes in directed network without PPI
1	2300	2293	2286
2	15249	13744	13339
3	758	1750	1035
4	17	437	50
5	1	2	-

We constructed networks with only regulatory edges and protein-protein interactions to estimate the number of loops found in a cell. This network consisted of 21826 genes/proteins, with 120988 edges. There were 114493 regulatory edges and 6495 protein-protein interactions (only physical interactions were considered). The approximate diameter of this graph is 8. Regulatory edges are uni-directional while protein-protein interactions are considered bi-directional.

Table 4.3: **Estimated counts of cycles in the network**

Cycle size	# Cycles with *Clock*	# Cycles with *Bmal1*
2	10	14
3	73	90
4	1007	1097
5	15512	15641
6	260973	253615
7	4570219	4324732

The clock machinery is centrally located and capable of potentially acting on a large fraction of the genome. We estimated the network-distance between *Clock* or *Bmal1* and all other proteins. We computed these on three different networks. First, network where the direction of edges is ignored. Second, network with uni-directional regulatory edges and bi-directional protein-protein interactions. Last, network with only the uni-directional regulatory edges. In all cases, it can be seen in Table 4.2 that ∼10% of genes are one hop away from *Clock*/*Bmal1* and ∼60-70% genes are two hop away.

In general, molecular species do not oscillate in isolation. What oscillates are directed loops also known as cycles. Cycles were counted by enumerating all paths with no repeated nodes which start and end at *Clock*/*Bmal1*. The counts are shown in Table 4.3.

Chapter 5

Crick in Genome Analysis Pipelines

5.1 Using Crick in a cancer genome analysis pipeline

5.1.1 Introduction

At the most fundamental level, cancer is a disease of the DNA, in which changes to the DNA sequence and the molecules that interact with it ultimately lead to uncontrolled cell proliferation. Large-scale cancer sequencing projects, such as the Cancer Genome Atlas (Can, 2013), have already started and produced volumes of data that are already well beyond what can be transferred over the Internet. However, these projects are still at a relatively early stage of development and are fraught with numerous challenges associated with the complexity of the sequencing technology, the lack of standardization, the sheer volume of data, the heterogeneity of cancers, the complexity of cancer biology, and the problem of obtaining proper control samples, to name only a few. Crick is used for the downstream analysis in a computational pipeline for the analysis of high-throughput sequencing cancer data that is currently being applied to pediatric cancer data that is regularly being sequenced,

and further resequenced on recurrence, as a result of a collaboration between the University of California, Irvine (UCI) and the Children Hospital of Orange County (CHOC).

Worldwide, it is estimated that childhood cancer has an incidence of more than 175,000 per year, and a mortality rate of approximately 96,000 per year. In the United States, cancer is the second most common cause of death among children between the ages of 1 and 14 years, exceeded only by accidents, with an incidence of about 12,000 of newly diagnosed cases per year and 1,300 deaths. The most common cancers in children are (childhood) leukemia (34%), brain tumors (23%), and lymphomas (12%). Other, less common childhood cancer types are: Neuroblastoma (7%), Wilms tumor (5%), NonHodgkin lymphoma (4%) , Rhabdomyosarcoma (3%), Retinoblastoma (3%), Osteosarcoma (3%), Ewing sarcoma (1%), Germ cell tumors, Pleuropulmonary blastoma, Hepatoblastoma, and hepatocellular carcinoma. The causes of most childhood cancers are unknown. The CHOC receives on the order of 100 new cases per year, and a project was started in 2012 to sequence the genome from healthy and cancer tissues of a subset of newly diagnosed cases – and therefore with no emphasis on particular tumors or tissue types – together with high-throughput gene expression measurements from cancer cells using RNA-seq.

5.1.2 Analysis pipeline

The goal is to develop an analysis pipeline comprising a combination of in-house and third party software to manage and analyze the raw data produced by these experiments, in a timely manner after they become available, including the identification and ranking of affected genes containing both small and large variants, and their integrative systems biology analyses against the large background of omic, literature, and other data available to us in order to derive inferences of clinical relevance specific to the cancer types of the patients sequenced. This pipeline assists in the analysis of pediatric tumors, as an unbiased

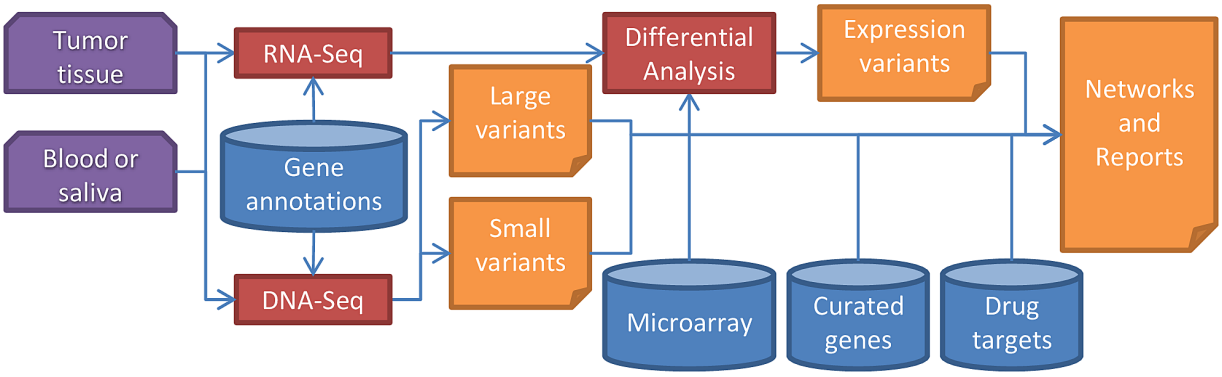


Figure 5.1: **Overview of the genomics analysis pipeline.** Raw sequencing reads are derived from two biological samples per patient and results in a HTML report with ranked genes and pathways.

and automated method for interpreting sequencing results along with identifying potentially therapeutic drugs and their targets. A schematic of the pipeline is shown in Figure 5.1 and more details can be found in Zeller *et al.* (2014).

5.1.3 Crick's network based approach

Relationships between cancer and the sequenced tumor genome are highlighted using Crick's network-based approach that integrates known and predicted protein-protein, protein-TF, and protein-drug interaction data. By using an integrative approach, effects of genetic variations on gene expression are used to provide further evidence of a driver mutations.

It has been shown that cancer cells share in common multiple acquired capabilities that enable the cell to proliferate uncontrollably. These hallmarks of cancer have been highlighted previously (Hanahan and Weinberg, 2000, 2011) and show a wide range of known pathways to be affected across different types of cancer. To visualize the connections between identified genetic and regulatory variations for each patient within known pathways – as well their connections to unaffected proteins – networks are created using Crick and then rendered in a web browser. 478 known pathways were included from KEGG Pathways (Kanehisa *et al.*, 2004) and the NCI Pathway Interaction Database (Schaefer *et al.*, 2009) in order to

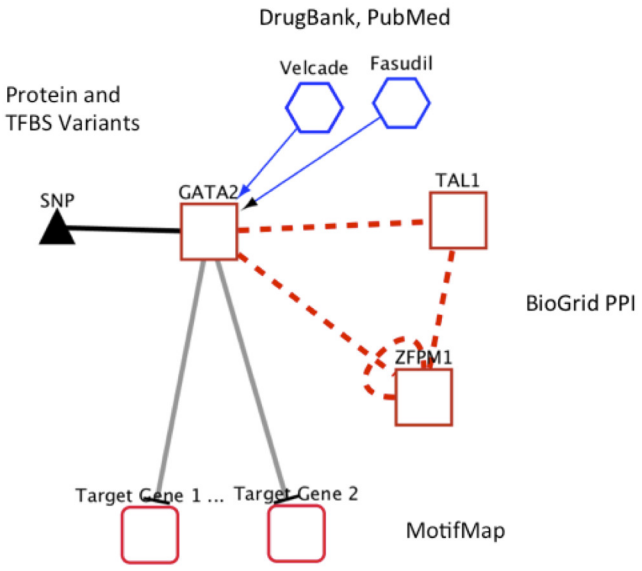


Figure 5.2: **Crick network with drug interactions.** Example of a network with drug, interaction, and transcription databases used to relate transcripts to each other and to potential drugs.

initialize networks with proteins related to a specific pathway. Subsequently, transcription factor (TF)-DNA, TF-TF, protein-protein edges are added to the network based on the publicly available datasets integrated into Crick. Variants on proteins, as well as the proteins identified as differential in the microarray and RNA-seq analyses, are used to highlight portions of the network and help visually interpret the biological role of the mutations. Variant TFBS are visualized by highlighting the edges between transcription factors in a network to the genes which contains a site for that factor within its promoter. Further, drug-protein interactions are added to the network, as described in the next section. Taken together, this network approach assists in investigating potential driver mutations with a focus on identifying potential drug candidates and their targets. An example of such a network is shown in Figure 5.2.

In order to elucidate potential druggable therapeutic targets, we have integrated several publicly accessible databases of drugs for network analyses. We have included well-characterized and predicted drug-effects, binding affinities, and drug-efficacy. These databases include the

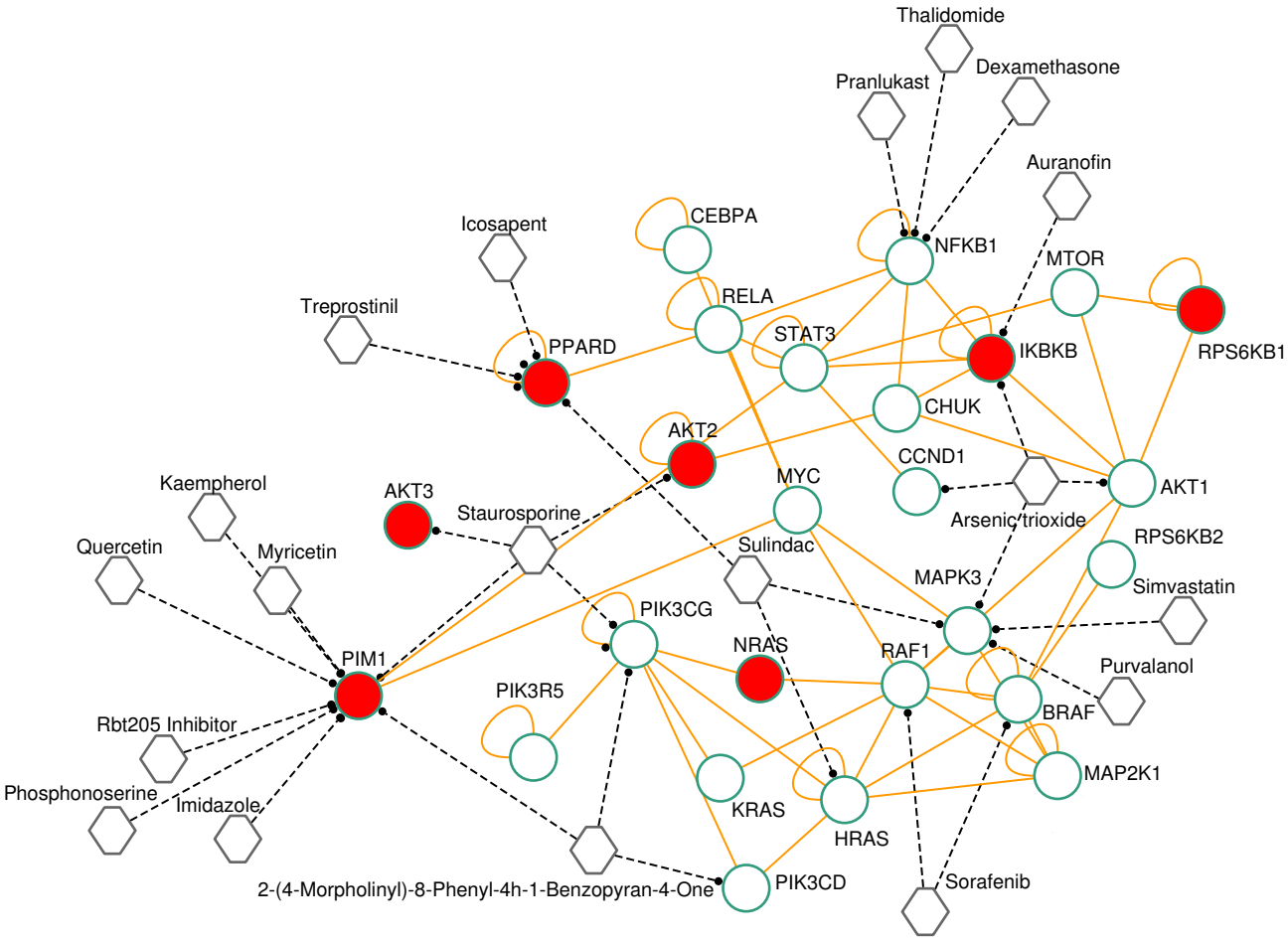


Figure 5.3: **CHOC36 drug-target edges in AML pathway limited to variants.** Circles denote proteins and hexagons denote drugs. Filled circles denote affected proteins, with identified potentially therapeutic drugs circled.

following resources:

- DrugBank (Knox *et al.*, 2011; Wishart *et al.*, 2008, 2006)
- BindingDB (Liu *et al.*, 2007)
- PharmGKB (Whirl-Carrillo *et al.*, 2012)

Each database provides an orthogonal set of annotations from which one can infer potential attenuation of known drug-effect, or perhaps novel drug interaction. Additional drug and drug-target information were also incorporated using semantic web resources for open drug

data. These include Bio2RDF (Belleau *et al.*, 2008; Callahan *et al.*, 2013), Chem2Bio2RDF (Chen *et al.*, 2010), and Linked-Open Drug Data (Samwald *et al.*, 2011). Figure 5.3 shows the corresponding auto-generated drug-target network used in exploring potential therapeutic targets.

Thus, Crick is an essential tool in the downstream analysis of genome-sequencing pipelines.

5.2 Spatial and temporal chromosomal organization driven by the circadian clock

Dynamic transitions in the epigenome have been associated with regulated patterns of nuclear organization. The accumulating evidence that chromatin remodeling is implicated in circadian function prompted us to explore whether the clock may control nuclear architecture. We applied the chromosome conformation capture on chip technology in mouse embryonic fibroblasts (MEFs) to demonstrate the presence of circadian long-range interactions using the clock-controlled *Dbp* gene as bait. The circadian genomic interactions with *Dbp* were highly specific and were absent in MEFs whose clock was disrupted by ablation of the *Bmal1* gene (also called *Arntl*). Using Crick we analyzed these genomic regions of interest and established that the *Dbp* circadian interactome contains a wide variety of genes and clock-related DNA elements. These findings reveal a previously unappreciated circadian and clock-dependent shaping of the nuclear landscape.

5.2.1 Background

Accumulating evidence has shown that chromatin remodeling events are involved in circadian regulation (Aguilar-Arnal and Sassone-Corsi, 2013; Feng and Lazar, 2012). Both genetic and

pharmacological approaches have shown that histone modifiers are implicated in circadian control, which indicates that their coordinated actions are necessary to fine tune a dynamic circadian epigenome. Genome-wide studies comprising mainly chromatin immunoprecipitation sequencing (ChIP-seq) analyses on livers harvested from mice in a circadian fashion have demonstrated that the histone modifications that are introduced by the circadian epigenetic modifiers are indeed rhythmic at many of the circadian gene promoters. These include rhythmic changes in acetylation at histone H3 Lys9 and Lys14 and methylation at histone H3 Lys4 and Lys27 (Koike *et al.*, 2012; Etchegaray *et al.*, 2003; Vollmers *et al.*, 2012), which parallels the rhythmic recruitment of polymerase II (Koike *et al.*, 2012; Le Martelot *et al.*, 2012).

Although these studies provide convincing evidence on the role of chromatin remodeling in circadian function, they do not explore whether nuclear topological organization is influenced by the clock. As the functional compartmentalization of the nuclear interior is being unraveled (Hakim *et al.*, 2010; Rajapakse and Groudine, 2011; Sanyal *et al.*, 2012), the spatial positioning of genes and regulatory elements is becoming increasingly recognized as an important epigenetic regulatory layer (Giles *et al.*, 2010; Edelman and Fraser, 2012; Sexton *et al.*, 2009; Lanctôt *et al.*, 2007). Hence, the three-dimensional folding of chromosomes inside the nucleus has been investigated extensively by fluorescence in situ hybridization (FISH) and chromosome conformation capture (3C) techniques25. The outcomes of these studies have revealed a nonrandom distribution of interphase chromosomes in chromosome territories and in topologically associating domains with common epigenetic marks (Dixon *et al.*, 2012; Lieberman-Aiden *et al.*, 2009; Nora *et al.*, 2012). The positioning configurations of chromosomes and genes diverge between cell types, and they can vary in response to physiological processes such as transcriptional reprogramming, development or disease (Cavalli and Misteli, 2013). In this respect, the circadian clock provides an ideal framework to study the interplay between genome organization and a physiologically dynamic program. Moreover, a potential regulatory role for the circadian clock in coordinating the higher-order

structure of the chromatin has been lacking. In this study we investigated the contribution of the circadian clock to the fine tuning of temporal changes in genome organization. To this end, we explored the circadian genomic interactome of the clock-controlled gene *Dbp* in MEFs. We found that the genomic interactions at the *Dbp* locus change in a way that parallels the circadian cycle progression and the expression of the gene, thereby delineating a *Dbp* circadian interactome. Notably, the *Dbp* circadian interactome was dependent on intact clock machinery, as it was not present in Bmal1-deficient MEFs. We also found that the *Dbp* interactome enclosed other circadian genes and was enriched in functionally related genes.

5.2.2 Circadian long-range genomic interactions

Using 4C (Ohlsson and Göndör, 2007; Simonis *et al.*, 2006), we sought to detect preferential interactions of the *Dbp* gene with other loci in the genome during the circadian cycle. *Dbp* gene was selected because of its robust circadian expression that is dictated by rhythmic CLOCK:BMAL1 binding to E-box DNA elements located on its promoter and coding sequences (Ripperger and Schibler, 2006; Stratmann *et al.*, 2012). More precisely, we designed the bait for the 4C experiment in a region (on chromosome 7) within intron 2 containing two E-boxes (Wuarin and Schibler, 1990). Data was collected from WT MEFs with robust circadian expression and *Bmal1*−/− MEFs with no circadian oscillation. 4C was used to detect the inter-chromosomal (trans) interactions.

To explore interaction frequencies in trans, we applied a running mean procedure to the mouse genome using a window size of 100 kb that was centered at each probe of the microarray. This analysis showed precise distribution and nonrandom interaction patterns for the *Dbp* locus that were coherent with the interaction patterns described previously for other loci in several cell types (Hakim *et al.*, 2011, 2012). The genomic distribution of *Dbp* contacts along the circadian cycle remained largely unaltered and delineated the genomic spatial

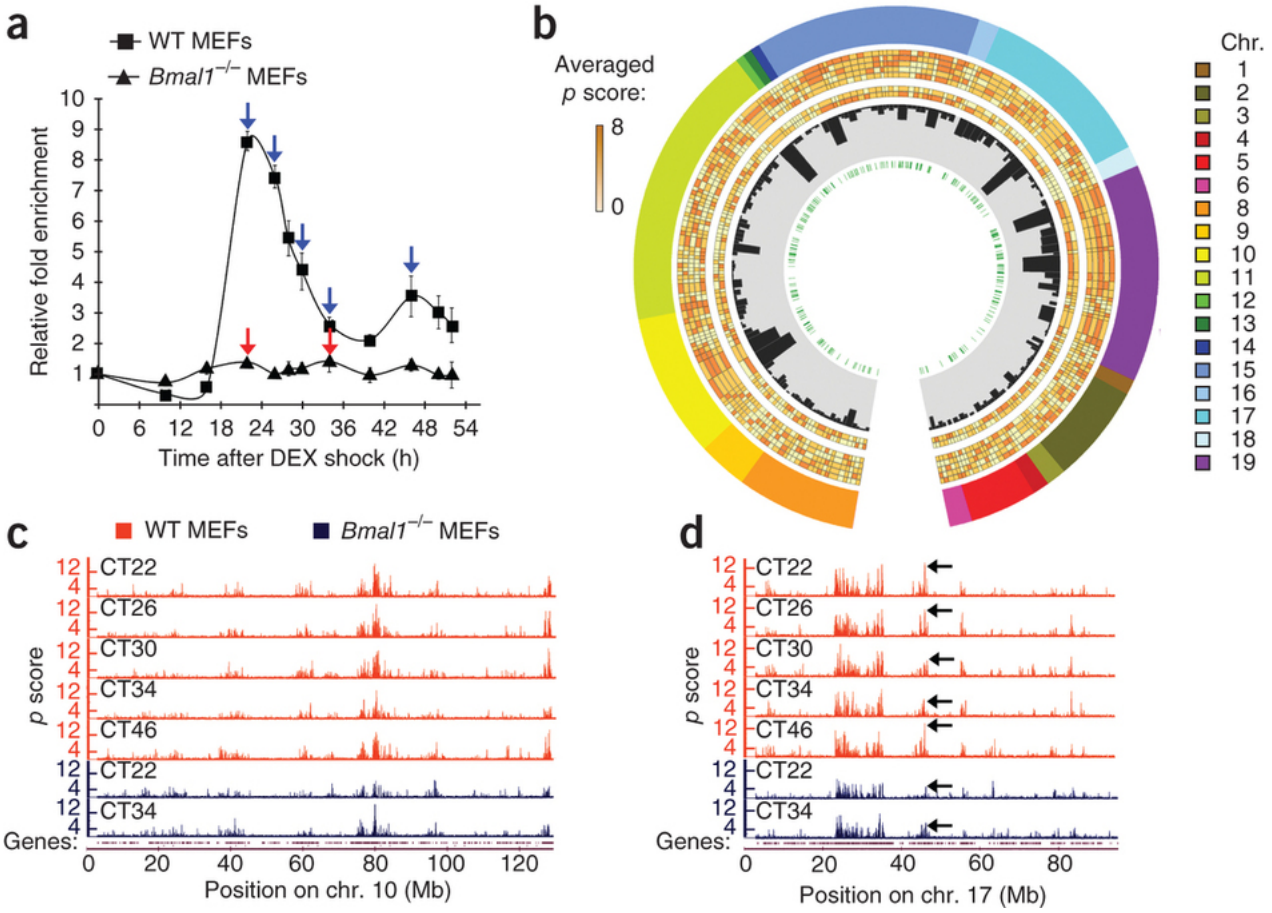


Figure 5.4: **(a) *Dbp* expression profile in WT and *Bmal1*−/− MEFs after synchronization with DEX, as analyzed by quantitative RT-PCR.** The value at time 0 was set to 1. The data were normalized to β-actin (also known as *Actb*) and are represented as the average ± s.e.m. of three independent biological replicates. Blue and red arrows indicate the CT in which WT and *Bmal1*−/− cells, respectively, were harvested for 4C analysis. **(b) Circos plot representing the *Dbp* interactome.** The layers indicate, from the outside to the inside: chromosome, where the chromosome number is indicated as a color code and its length is proportional to the actual length of the interacting regions; averaged p scores for each genomic region shown as a color scale; histogram bars representing the gene content for each region; and E-box element locations. The averaged p scores correspond to each of the 4C experiments, from the outside to the inside: WT CT22, WT CT26, WT CT30, WT CT34, WT CT46, *Bmal1*−/− CT22 and *Bmal1*−/− CT34. **(c,d) Microarray profiles showing the interaction frequencies (p scores from the 4C data) between *Dbp* and mouse chromosomes (chr.) 10 (c) and 17 (d)**. The orange and blue plots represent the data for WT and *Bmal1*−/− MEFs, respectively. The corresponding CT is indicated for each lane. The data sets are highly correlated, but major differences in the interaction frequencies are also apparent (black arrows in d). The genomic positions in mm8 coordinates are indicated on the horizontal axis.

environment of the *Dbp* locus (Figure 5.4b,d).

5.2.3 Regions of interest

We identified 201 genomic regions that contact the *Dbp* gene at any time of the circadian cycle in WT MEFs, and the mean length of these regions was 130 kb (Figure 5.4b). These regions were further analyzed using Crick and all genes and regulatory binding sites were mapped. Whereas some chromosomes, such as chromosomes 1, 3 and 12, interacted only rarely with *Dbp*, others displayed preferential contacts with our bait on different locations (Figure 5.4b, outer layer). For example, on chromosome 11, we found 39 loci that contact *Dbp*. This result indicates close spatial proximity and a high intermingling frequency between territories from chromosomes 7 and 11. Notably, the *Dbp* genomic contacts displayed a four-fold enrichment in gene content over randomized data. We then sought to determine the dynamics of the interaction of *Dbp* contacts during the circadian cycle. Our running mean analysis showed that the genomic locations of *Dbp* contacts remained similar overall along the circadian cycle. Notably, the interaction frequencies of *Dbp* with specific loci varied in degrees according to the locus and CT (Figure 5.4e, with examples indicated by black arrows on the colored genomic plots).

To determine the overall likelihood of interaction of the *Dbp* locus with a given genomic region, we calculated an averaged p score for each of the 201 described contacts at each CT. We calculated the averaged p score by considering the p scores of neighboring probes (Online Methods). We then classified the *Dbp* genomic contacts according to their averaged p score at each CT. Interestingly, specific contacts that followed a cyclic pattern of interaction mirrored *Dbp* circadian gene expression. These genomic regions efficiently contact *Dbp* at CT22, CT26 and CT48, which corresponds to times at which *Dbp* shows the highest expression. These interactions were virtually undetectable at CT34, which is the time of lowest *Dbp* expression

(Figure 5.5a). We identified 29 genomic regions that met these criteria, and these regions comprise the *Dbp* circadian interactome (Figure 5.4b).

5.2.4 Bmal1 is essential for a specific circadian interactome

The *Dbp* contacts in the *Bmal1*−/− MEFs were similar overall to those in WT MEFs and showed a highly correlated profile of peaks and troughs (Figure 5.4c,d). However, there was a markedly reduced contact frequency within the *Dbp* circadian interactome in *Bmal1*−/− MEFs (Figure 5.5a). Notably, a lack of BMAL1 was associated with a loss of circadian oscillation in the interaction of *Dbp* with the 29 selected genomic regions (Figure 5.5a). The profiles of the *Dbp* circadian contacts at both CT22 and CT34 in *Bmal1*−/− MEFs showed low interaction frequencies and were thus highly similar to the profile in WT MEFs at CT34 (Figure 5.5a). These findings indicate that the circadian system contributes to the establishment of a specific subnuclear genomic environment around the *Dbp* gene. To gain insights into the molecular mechanisms that contribute to the specific circadian genomic architecture of the *Dbp* locus, we used MotifMap (Xie *et al.*, 2009; Daily *et al.*, 2011) to identify the transcription-factor binding sites on the promoters of the genes that associate with *Dbp* in a circadian fashion (Figure 5.5b). Among these promoters, we found a 2.5-fold enrichment on promoters containing E-boxes ($P < 0.001$, Fisher exact test). E-boxes are highly conserved DNA elements that bind CLOCK:BMAL1 and are involved in driving circadian gene expression (Ripperger and Schibler, 2006; Hardin, 2004). We speculate that the *Dbp* gene is present within a subnuclear environment that is gradually modified during the circadian cycle and that the circadian molecular machinery is implicated in establishing this pattern.

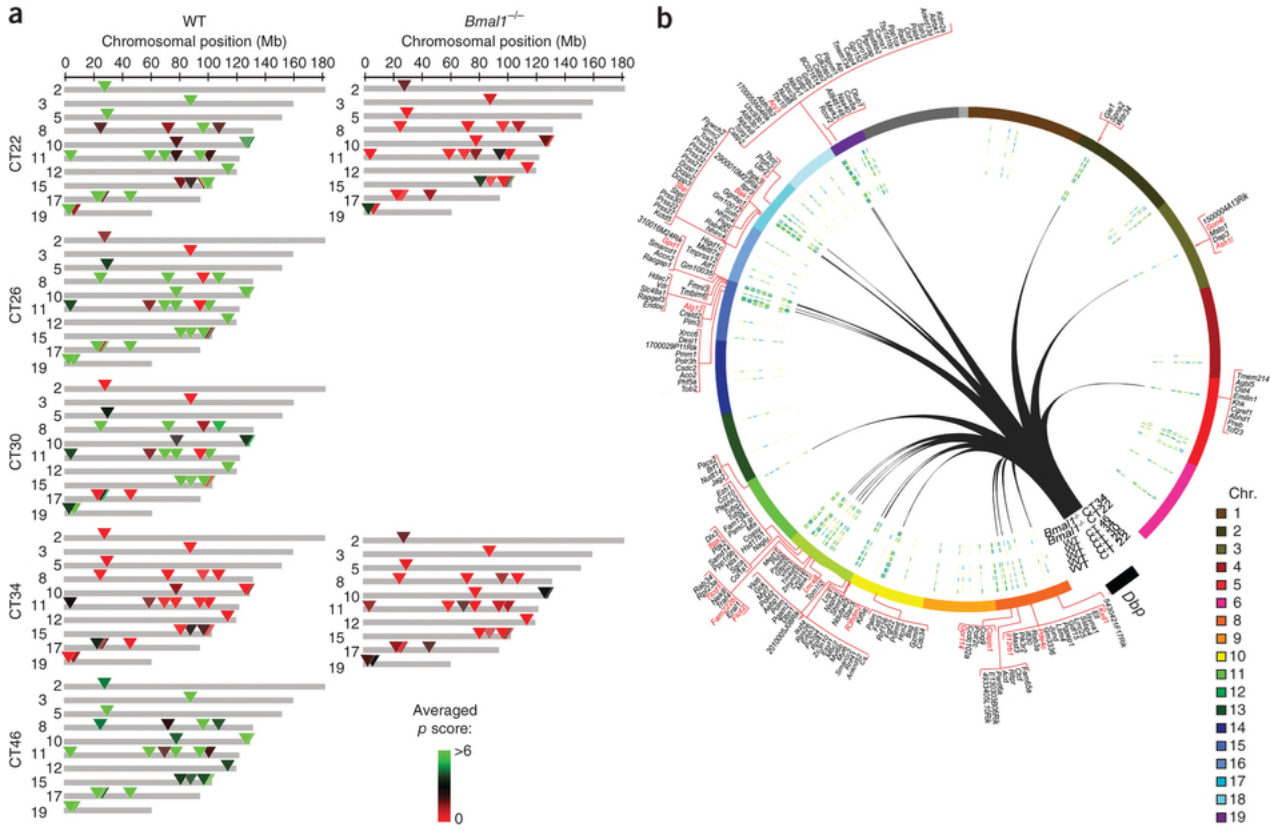


Figure 5.5: **(a) Genomic map of the *Dbp* circadian interactome at the indicated CTs after synchronization with DEX in WT and *Bmal1*−/− MEFs.** Averaged p scores for each region are indicated in a green-red color scale according to the intensity of the interaction, which is proportional to the probe signal (4C over genomic DNA). Colored triangles indicate the positions of the *Dbp* circadian contacts. Areas shown in gray did not show circadian contact. The genomic positions in mm8 coordinates are indicated on the top horizontal axis. Chromosomes that did not present circadian interaction with *Dbp* are not shown here. **(b) Circos plot representing the genome-wide view of *Dbp* circadian interactions (black lines) with the corresponding chromosomes in trans.** The gene content corresponding to each contact region is indicated in the outer layer of the plot. The genes in red are those that presented circadian mRNA accumulation after synchronization with DEX as defined by the gene expression analysis (JTK $P < 0.01$).

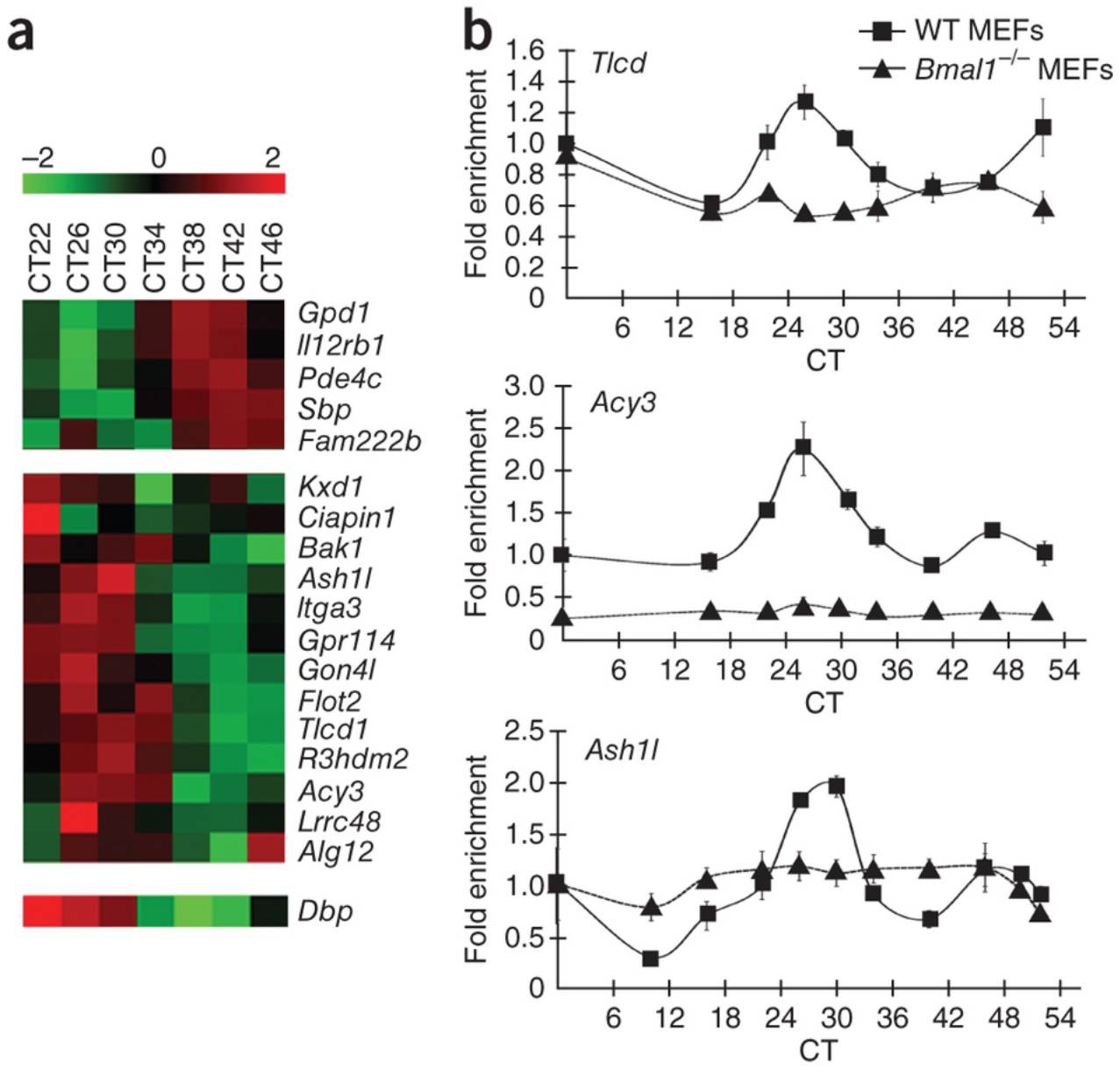


Figure 5.6: **(a) Heat map showing the log2 expression values of the circadian genes in the *Dbp* circadian interactome.** Selected genes were plotted according to their phase. **(b) Quantitative real-time PCR of selected transcripts confirming the microarray data.** Total RNA was collected before synchronization with DEX (CT0) and at 16, 22, 26, 30, 34, 40, 46 and 52 h after DEX induction from WT and *Bmal1*−/− MEFs. Data were normalized to β-actin and are shown as the average s.e.m. of three independent biological replicates.

5.2.5 Features of genes within the *Dbp* circadian interactome

Higher-order genome organization has a major role in efficient responses to stimuli, effective signaling pathways and the coordination of lineage-specific differentiation. We next investigated the time-specific expression of genes within the *Dbp* circadian interactome. More information about the experimental procedure and results from the microarray are included in Aguilar-Arnal *et al.* (2013). By comparing our circadian array data in MEFs with the genes that appear to have circadian contacts, we found 18 genes that cycle in a circadian manner (JTK P < 0.01) (Figure 5.5b, genes in red, and Figure 5.6). These genes can be classified according to their expression phase into two major groups containing 5 and 13 genes (Figure 5.6a). We confirmed the rhythmicity of the selected genes by RT-PCR and found that their mRNA levels oscillated in a circadian manner, further validating our microarray data (Figure 5.6b).

A motif discovery analysis on the promoters of these 18 genes revealed enrichment of recognition motifs for ROR-α and E-box elements, which are known to be centrally implicated in circadian gene expression (Ueda *et al.*, 2005). A comparative analysis with available BMAL1 ChIP-seq databases (Koike *et al.*, 2012; Rey *et al.*, 2011b) indicated that the promoters of many genes located within the *Dbp* circadian interactome bind BMAL1. Moreover, the motif analysis indicated that the promoters of the circadian genes within the *Dbp* interactome may also contain D-boxes. These findings, which are consistent with our microarray analysis, suggest the possibility that the clock machinery itself could be implicated in shaping the nuclear genomic architecture.

The coordinated expression of circadian genes represents a remarkable paradigm of transcriptional control (Doherty and Kay, 2010). Accumulating evidence has recently shaped the notion that distinct changes in chromatin remodeling may be driven by the circadian clock to insure co-regulation of clock-controlled genes. The combined effects exerted by a

variety of chromatin remodelers are assumed to lead to circadian activation and silencing of specific genes. Recent advances in the field of nuclear architecture have suggested that chromosome organization has an active role in many genomic functions. In this context, understanding how the nuclear landscape is modified in a CT-specific manner can lead to insights into the molecular mechanisms that determine circadian rhythms.

5.2.6 Conclusion

Our data reveal that the circadian clock is implicated in shaping temporal and spatial cycles in chromosomal organization. These variations in nuclear organization could provide a genomic frame to assist circadian gene expression of *Dbp*. Our data suggest that the genomic environment of the *Dbp* locus around the circadian cycle remains largely constant. However, several large chromatin domains change their frequency of interaction in trans with *Dbp*, which parallels the progression of the circadian cycle and the transcriptional state of the gene (Figure 5.5). We present genetic evidence that BMAL1 has a critical role in establishing a CT-specific *Dbp* interactome (Figure 5.5).

The outcomes of our comprehensive analysis have direct functional implications. In this respect, we described a spatial clustering of genes and circadian-related DNA elements around the *Dbp* locus. These results point to the existence of subnuclear environments enriched with CLOCK-specific response elements. This finding is in agreement with previous reports showing that spatial congregation of DNase I–hypersensitive sites is observed within DNA sequences that establish trans contacts (Lieberman-Aiden *et al.*, 2009; Hakim *et al.*, 2011). The role of specific transcriptional regulators in establishing the interactome is emerging as a new avenue for understanding the modulation and functions of genome topology (Hakim *et al.*, 2013; Schoenfelder *et al.*, 2010). Here we provide evidence supporting the circadian clock, including the BMAL1 transcription factor, being involved in shaping the nuclear ar-

chitecture during the circadian cycle. This conclusion is further reinforced by our study on BMAL1-deficient cells, in which the changes in the interactome around the circadian cycle were not present. Future studies will be necessary to uncover the precise contribution and the hierarchical organization of all clock regulators in determining circadian genome topology.

5.3 Summary

In summary, Crick forms a important piece in genome analysis pipelines to provide context to experimental data like variants from patient genomes, long-range interactions in the genome, etc.

Bibliography

(2013). The cancer genome atlas homepage.

Abbondante, S., Ceglia, N., Baldi, P., and Sassone-Corsi, P. (2014). *in submission*.

Aguilar-Arnal, L. and Sassone-Corsi, P. (2013). The circadian epigenome: how metabolism talks to chromatin remodeling. *Current opinion in cell biology*, **25**(2), 170–176.

Aguilar-Arnal, L., Hakim, O., Patel, V. R., Baldi, P., Hager, G. L., and Sassone-Corsi, P. (2013). Cycles in spatial and temporal chromosomal organization driven by the circadian clock. *Nature structural & molecular biology*, **20**(10), 1206–1213.

Akman, O. E., Watterson, S., Parton, A., Binns, N., Millar, A. J., and Ghazal, P. (2012). Digital clocks: simple boolean models can quantitatively describe circadian systems. *J R Soc Interface*, **9**(74), 2365–82.

Andrews, J. L., Zhang, X., McCarthy, J. J., McDearmon, E. L., Hornberger, T. A., Russell, B., Campbell, K. S., Arbogast, S., Reid, M. B., Walker, J. R., Hogenesch, J. B., Takahashi, J. S., and Esser, K. A. (2010). Clock and bmal1 regulate myod and are necessary for maintenance of skeletal muscle phenotype and function. *Proc Natl Acad Sci U S A*, **107**(44), 19090–5.

Antunes, L. C., Levandovski, R., Dantas, G., Caumo, W., and Hidalgo, M. P. (2010). Obesity and shift work: chronobiological aspects. *Nutr Res Rev*, **23**(1), 155–68.

Baldi, P. and Atiya, A. (1989). Oscillations and synchronizations in neural networks: An exploration of the labeling hypothesis. *International Journal of Neural Systems*, **01**(02), 103–124.

Baldi, P. and Meir, R. (1990). Computing with arrays of coupled oscillators: an application to preattentive texture discrimination. *Neural Computation*, **2**(4), 458–471.

Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P., and Morissette, J. (2008). Bio2rdf: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, **41**(5), 706 – 716.

Beyer, K. S., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1999). When Is "Nearest Neighbor" Meaningful? In *Proceedings of the 7th International Conference on Database Theory*, ICDT '99, pages 217–235, London, UK, UK. Springer-Verlag.

Brandt, S., Dellen, B., and Wessel, R. (2006). Synchronization from disordered driving forces in arrays of coupled oscillators. *Physical review letters*, **96**(3), 34104.

Brown, S. A., Kowalska, E., and Dallmann, R. (2012). (re)inventing the circadian feedback loop. *Dev Cell*, **22**(3), 477–87.

Callahan, A., Cruz-Toledo, J., and Dumontier, M. (2013). Ontology-based querying with bio2rdf's linked open data. *Journal of Biomedical Semantics*, **4**(Suppl 1), S1.

Cao, D. and Pizzorno, G. (2004). Uridine phosophorylase: an important enzyme in pyrimidine metabolism and fluoropyrimidine activation. *Drugs of today (Barcelona, Spain : 1998)*, **40**(5), 431–443.

Cavalli, G. and Misteli, T. (2013). Functional implications of genome topology. *Nature structural & molecular biology*, **20**(3), 290–299.

Chen, B., Dong, X., Jiao, D., Wang, H., Zhu, Q., Ding, Y., and Wild, D. (2010). Chem2bio2rdf: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC Bioinformatics*, **11**(1), 255.

Daily, K., Patel, V., Rigor, P., Xie, X., and Baldi, P. (2011). MotifMap: integrative genome-wide maps of regulatory motif sites for model species. *BMC Bioinformatics*, **12**(1), 495+.

de Brevern, A., Hazout, S., and Malpertuy, A. (2004). Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering. *BMC Bioinformatics*, **5**(1), 114+.

Dean, J. and Ghemawat, S. (2008). MapReduce: Simplified Data Processing on Large Clusters. *Commun. ACM*, **51**(1), 107–113.

Debruyne, J. P., Noton, E., Lambert, C. M., Maywood, E. S., Weaver, D. R., and Reppert, S. M. (2006). A clock shock: mouse CLOCK is not required for circadian oscillator function. *Neuron*, **50**(3), 465–477.

DeBruyne, J. P., Weaver, D. R., and Reppert, S. M. (2007a). CLOCK and NPAS2 have overlapping roles in the suprachiasmatic circadian clock. *Nature neuroscience*, **10**(5), 543–545.

DeBruyne, J. P., Weaver, D. R., and Reppert, S. M. (2007b). Peripheral circadian oscillators require CLOCK. *Current biology : CB*, **17**(14).

Dibner, C., Schibler, U., and Albrecht, U. (2010). The mammalian circadian timing system: organization and coordination of central and peripheral clocks. *Annu Rev Physiol*, **72**, 517–49.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**(7398), 376–380.

Doherty, C. J. and Kay, S. A. (2010). Circadian control of global gene expression patterns. *Annual review of genetics*, **44**, 419–444.

Dyar, K. A., Ciciliot, S., Wright, L. E., Sjrup Biens, R., Malagoli Tagliazucchi, G., Patel, V. R., Forcato, M., Pea Paz, M. I., Gudiksen, A., Solagna, F., Albiero, M., Moretti, I., Eckel-Mahan, K. L., Baldi, P., Sassone-Corsi, P., Rizzuto, R., Bicciato, S., Pilegaard, H., Blaauw, B., and Schiaffino, S. (2013). Muscle insulin sensitivity and glucose metabolism are controlled by the intrinsic muscle clock. *Molecular Metabolism*.

Eckel-Mahan, K. and Sassone-Corsi, P. (2009). Metabolism control by the circadian clock and vice versa. *Nat Struct Mol Biol*, **16**(5), 462–7.

Eckel-Mahan, K. L., Patel, V. R., Mohney, R. P., Vignola, K. S., Baldi, P., and Sassone-Corsi, P. (2012). Coordination of the transcriptome and metabolome by the circadian clock. *Proc Natl Acad Sci U S A*, **109**(14), 5541–6.

Eckel-Mahan, K. L., Patel, V. R., de Mateo, S., Orozco-Solis, R., Ceglia, N. J., Sahar, S., Dilag-Penilla, S. A., Dyar, K. A., Baldi, P., and Sassone-Corsi, P. (2013). Reprogramming of the circadian clock by nutritional challenge. *Cell*, **155**(7), 1464–1478.

Edelman, L. B. B. and Fraser, P. (2012). Transcription factories: genetic programming in three dimensions. *Current opinion in genetics & development*, **22**(2), 110–114.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Res*, **30**(1), 207–10.

Etchegaray, J.-P. P., Lee, C., Wade, P. A., and Reppert, S. M. (2003). Rhythmic histone acetylation underlies transcription in the mammalian circadian clock. *Nature*, **421**(6919), 177–182.

Evans, A. M., DeHaven, C. D., Barrett, T., Mitchell, M., and Milgram, E. (2009). Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems. *Analytical chemistry*, **81**(16), 6656–6667.

Feng, D. and Lazar, M. A. (2012). Clocks, metabolism, and the epigenome. *Molecular cell*, **47**(2), 158–167.

Ferrucci, D. A. (2011). IBM's Watson/DeepQA. *SIGARCH Comput. Archit. News*, **39**(3).

Fodor, I. (2002). A Survey of Dimension Reduction Techniques.

Froy, O. (2010). Metabolism and circadian rhythms–implications for obesity. *Endocr Rev*, **31**(1), 1–24.

Froy, O. (2011). Circadian rhythms, aging, and life span in mammals. *Physiology (Bethesda)*, **26**(4), 225–35.

Gerstner, J. R., Lyons, L. C., Wright, K. P., J., Loh, D. H., Rawashdeh, O., Eckel-Mahan, K. L., and Roman, G. W. (2009). Cycling behavior and memory formation. *J Neurosci*, **29**(41), 12824–30.

Giles, K. E., Gowher, H., Ghirlando, R., Jin, C., and Felsenfeld, G. (2010). Chromatin boundaries, insulators, and long-range interactions in the nucleus. *Cold Spring Harbor symposia on quantitative biology*, **75**, 79–85.

Giorgelli, F., Bottai, C., Mascia, L., Scolozzi, C., Camici, M., and Ipata, P. L. (1997). Recycling of alpha-D-ribose 1-phosphate for nucleoside interconversion. *Biochimica et biophysica acta*, **1335**(1-2), 6–22.

Goel, N., Stunkard, A. J., Rogers, N. L., Van Dongen, H. P., Allison, K. C., O'Reardon, J. P., Ahima, R. S., Cummings, D. E., Heo, M., and Dinges, D. F. (2009). Circadian rhythm profiles in women with night eating syndrome. *Journal of biological rhythms*, **24**(1), 85–94.

Goel, P. and Ermentrout, B. (2002). Synchrony, stability, and firing patterns in pulse-coupled oscillators. *Physica D: Nonlinear Phenomena*, **163**(3), 191–216.

Goldbeter, A. (1995). A model for circadian oscillations in the drosophila period protein (per). *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **261**(1362), 319–324.

Goldbeter, A. (1997). Biochemical oscillations and cellular rhythms: the molecular bases of periodic and chaotic behaviour.

Hakim, O., Sung, M.-H. H., and Hager, G. L. (2010). 3D shortcuts to gene regulation. *Current opinion in cell biology*, **22**(3), 305–313.

Hakim, O., Sung, M.-H. H., Voss, T. C., Splinter, E., John, S., Sabo, P. J., Thurman, R. E., Stamatoyannopoulos, J. A., de Laat, W., and Hager, G. L. (2011). Diverse gene reprogramming events occur in the same spatial clusters of distal regulatory elements. *Genome research*, **21**(5), 697–706.

Hakim, O., Resch, W., Yamane, A., Klein, I., Kieffer-Kwon, K.-R. R., Jankovic, M., Oliveira, T., Bothmer, A., Voss, T. C., Ansarah-Sobrinho, C., Mathe, E., Liang, G., Cobell, J., Nakahashi, H., Robbiani, D. F., Nussenzweig, A., Hager, G. L., Nussenzweig, M. C., and Casellas, R. (2012). DNA damage defines sites of recurrent chromosomal translocations in B lymphocytes. *Nature*, **484**(7392), 69–74.

Hakim, O., Sung, M.-H. H., Nakayamada, S., Voss, T. C., Baek, S., and Hager, G. L. (2013). Spatial congregation of STAT binding directs selective nuclear architecture during T-cell functional differentiation. *Genome research*, **23**(3), 462–472.

Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, **100**(1), 57–70.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of Cancer: The Next Generation. *Cell*, **144**(5), 646–674.

Hardin, P. E. (2004). Transcription regulation within the circadian clock: the E-box and beyond. *Journal of biological rhythms*, **19**(5), 348–360.

Huang, D. W., Sherman, B. T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M. W., Lane, H. C., and Lempicki, R. A. (2007). David bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic acids research*, **35**(Web Server issue), W169–175.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, **4**(1), 44–57.

Hughes, M. E., DiTacchio, L., Hayes, K. R., Vollmers, C., Pulivarthy, S., Baggs, J. E., Panda, S., and Hogenesch, J. B. (2009). Harmonics of circadian gene transcription in mammals. *PLoS Genet*, **5**(4), e1000442.

Hughes, M. E., Hogenesch, J. B., and Kornacker, K. (2010). Jtk_cycle: an efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. *J Biol Rhythms*, **25**(5), 372–80.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Research*, **32**(suppl 1), D277–D280.

Karlsson, B., Knutsson, A., and Lindahl, B. (2001). Is there an association between shift work and having a metabolic syndrome? results from a population based study of 27,485 people. *Occup Environ Med*, **58**(11), 747–52.

Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, A. C., and Wishart, D. S. (2011). DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.*, **39**(Database issue), D1035–1041.

Knutsson, A. (2003). Health disorders of shift workers. *Occup Med (Lond)*, **53**(2), 103–8.

Kohsaka, A., Laposky, A. D., Ramsey, K. M., Estrada, C., Joshu, C., Kobayashi, Y., Turek, F. W., and Bass, J. (2007). High-fat diet disrupts behavioral and molecular circadian rhythms in mice. *Cell Metab*, **6**(5), 414–21.

Koike, N., Yoo, S.-H. H., Huang, H.-C. C., Kumar, V., Lee, C., Kim, T.-K. K., and Takahashi, J. S. (2012). Transcriptional architecture and chromatin landscape of the core circadian clock in mammals. *Science (New York, N.Y.)*, **338**(6105), 349–354.

Lamia, K. A., Storch, K. F., and Weitz, C. J. (2008). Physiological significance of a peripheral tissue circadian clock. *Proc Natl Acad Sci U S A*, **105**(39), 15172–7.

Lanctôt, C., Cheutin, T., Cremer, M., Cavalli, G., and Cremer, T. (2007). Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nature reviews. Genetics*, **8**(2), 104–115.

Le Martelot, G., Canella, D., Symul, L., Migliavacca, E., Gilardi, F., Liechti, R., Martin, O., Harshman, K., Delorenzi, M., Desvergne, B., Herr, W., Deplancke, B., Schibler, U., Rougemont, J., Guex, N., Hernandez, N., Naef, F., and CycliX Consortium (2012). Genome-wide RNA polymerase II profiles and RNA accumulation reveal kinetics of transcription and associated epigenetic changes during diurnal cycles. *PLoS biology*, **10**(11).

Levan, G., Hedrich, H. J., Remmers, E. F., Serikawa, T., and Yoshida, M. C. (1995). Standardized rat genetic nomenclature. **6**(7), 447–448.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., and Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)*, **326**(5950), 289–293.

Liu, T., Lin, Y., Wen, X., Jorissen, R. N., and Gilson, M. K. (2007). BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.*, **35**(Database issue), 198–201.

Lopes, C. T., Franz, M., Kazi, F., Donaldson, S. L., Morris, Q., and Bader, G. D. (2010). Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, **26**(18), 2347–2348.

Losick, R. and Desplan, C. (2008). Stochasticity and Cell Fate. *Science*, **320**(5872), 65–68.

Maltais, L. J., Blake, J. A., Chu, T., Lutz, C. M., Eppig, J. T., and Jackson, I. (2002). Rules and guidelines for mouse gene, allele, and mutation nomenclature: a condensed version. *Genomics*, **79**(4), 471–474.

Marx, V. (2013). Biology: The big challenges of big data. *Nature*, **498**(7453), 255–260.

Masri, S., Rigor, P., Ceglia, N., Baldi, P., and Sassone-Corsi, P. (2014). *in submission*.

Miller, B. H., McDearmon, E. L., Panda, S., Hayes, K. R., Zhang, J., Andrews, J. L., Antoch, M. P., Walker, J. R., Esser, K. A., Hogenesch, J. B., and Takahashi, J. S. (2007). Circadian and clock-controlled regulation of the mouse transcriptome and cell proliferation. *Proceedings of the National Academy of Sciences*, **104**(9), 3342–3347.

Moore, R. Y. and Eichler, V. B. (1972). Loss of a circadian adrenal corticosterone rhythm following suprachiasmatic lesions in the rat. *Brain Res*, **42**(1), 201–6.

Nakahata, Y., Kaluzova, M., Grimaldi, B., Sahar, S., Hirayama, J., Chen, D., Guarente, L. P., and Sassone-Corsi, P. (2008). The nad+-dependent deacetylase sirt1 modulates clock-mediated chromatin remodeling and circadian control. *Cell*, **134**(2), 329–40.

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N. L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Blüthgen, N., Dekker, J., and Heard, E. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, **485**(7398), 381–385.

Ohlsson, R. and Göndör, A. (2007). The 4C technique: the 'Rosetta stone' for genome biology in 3D? *Current opinion in cell biology*, **19**(3), 321–325.

on Frontiers at the Interface of Computing, C., Biology, and Council, N. R. (2005). *Catalyzing Inquiry at the Interface of Computing and Biology*. National Academies Press, 1 edition.

O'Reardon, J. P., Ringel, B. L., Dinges, D. F., Allison, K. C. C., Rogers, N. L., Martino, N. S., and Stunkard, A. J. (2004). Circadian eating and sleeping patterns in the night eating syndrome. *Obesity research*, **12**(11), 1789–1796.

O'Reardon, J. P., Peshek, A., and Allison, K. C. (2005). Night eating syndrome : diagnosis, epidemiology and management. *CNS drugs*, **19**(12), 997–1008.

Panda, S., Antoch, M. P., Miller, B. H., Su, A. I., Schook, A. B., Straume, M., Schultz, P. G., Kay, S. A., Takahashi, J. S., and Hogenesch, J. B. (2002). Coordinated transcription of key pathways in the mouse by the circadian clock. *Cell*, **109**(3), 307–20.

Partch, C. L., Green, C. B., and Takahashi, J. S. (2013). Molecular architecture of the mammalian circadian clock. *Trends in Cell Biology*.

Patel, V. R., Eckel-Mahan, K., Sassone-Corsi, P., and Baldi, P. (2012). CircadiOmics: integrating circadian genomics, transcriptomics, proteomics and metabolomics. *Nature methods*, **9**(8), 772–773.

Patel, V. R., Eckel-Mahan, K., Sassone-Corsi, P., and Baldi, P. (2014). How pervasive are circadian oscillations? *Trends in Cell Biology*.

Peek, C. B., Affinati, A. H., Ramsey, K. M., Kuo, H. Y., Yu, W., Sena, L. A., Ilkayeva, O., Marcheva, B., Kobayashi, Y., Omura, C., Levine, D. C., Bacsik, D. J., Gius, D., Newgard, C. B., Goetzman, E., Chandel, N. S., Denu, J. M., Mrksich, M., and Bass, J. (2013). Circadian clock nad+ cycle drives mitochondrial oxidative metabolism in mice. *Science*, **342**(6158), 1243417.

Quackenbush, J. (2007). Extracting biology from high-dimensional biological data. *Journal of Experimental Biology*, **210**(9), 1507–1517.

Rajapakse, I. and Groudine, M. (2011). On emerging nuclear order. *The Journal of cell biology*, **192**(5), 711–721.

Ralph, M. R., Foster, R. G., Davis, F. C., and Menaker, M. (1990). Transplanted suprachiasmatic nucleus determines circadian period. *Science*, **247**(4945), 975–8.

Ramsey, K. M., Yoshino, J., Brace, C. S., Abrassart, D., Kobayashi, Y., Marcheva, B., Hong, H.-K., Chong, J. L., Buhr, E. D., Lee, C., Takahashi, J. S., Imai, S.-i., and Bass, J. (2009). Circadian clock feedback cycle through nampt-mediated nad+ biosynthesis. *Science*, **324**(5927), 651–654.

Rey, G., Cesbron, F., Rougemont, J., Reinke, H., Brunner, M., and Naef, F. (2011a). Genome-wide and phase-specific dna-binding rhythms of bmal1 control circadian output functions in mouse liver. *PLoS biology*, **9**(2), e1000595.

Rey, G., Cesbron, F., Rougemont, J., Reinke, H., Brunner, M., and Naef, F. (2011b). Genome-wide and phase-specific DNA-binding rhythms of BMAL1 control circadian output functions in mouse liver. *PLoS biology*, **9**(2), e1000595+.

Ripperger, J. A. and Schibler, U. (2006). Rhythmic CLOCK-BMAL1 binding to multiple E-box motifs drives circadian Dbp transcription and chromatin transitions. *Nature genetics*, **38**(3), 369–374.

Samwald, M., Jentzsch, A., Bouton, C., Kallesoe, C., Willighagen, E., Hajagos, J., Marshall, M., Prud'hommeaux, E., Hassanzadeh, O., Pichler, E., and Stephens, S. (2011). Linked open drug data for pharmaceutical research and development. *Journal of Cheminformatics*, **3**(1), 19.

Sanyal, A., Lajoie, B. R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature*, **489**(7414), 109–113.

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K. H. (2009). PID: the Pathway Interaction Database. *Nucleic acids research*, **37**(Database issue), D674–D679.

Scheel, I., Aldrin, M., Glad, I. K., Sørum, R., Lyng, H., and Frigessi, A. (2005). The influence of missing value imputation on detection of differentially expressed genes from microarray data. *Bioinformatics (Oxford, England)*, **21**(23), 4272–4279.

Scheper, T., Klinkenberg, D., Pennartz, C., and van Pelt, J. (1999). A mathematical model for the intracellular circadian rhythm generator. *J Neurosci*, **19**(1), 40–7.

Schibler, U. and Sassone-Corsi, P. (2002). A web of circadian pacemakers. *Cell*, **111**(7), 919–22.

Schoenfelder, S., Clay, I., and Fraser, P. (2010). The transcriptional interactome: gene expression in 3D. *Current opinion in genetics & development*, **20**(2), 127–133.

Sexton, T., Bantignies, F., and Cavalli, G. (2009). Genomic interactions: chromatin loops and gene meeting points in transcriptional regulation. *Seminars in cell & developmental biology*, **20**(7), 849–855.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, **13**(11), 2498–2504.

Sharifian, A., Farahani, S., Pasalar, P., Gharavi, M., and Aminian, O. (2005). Shift work as an oxidative stressor. *J Circadian Rhythms*, **3**, 15.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature genetics*, **38**(11), 1348–1354.

Stein, L. D. (2003). Integrating biological databases. *Nat Rev Genet*, **4**(5), 337–345.

Storch, K.-F. F., Lipan, O., Leykin, I., Viswanathan, N., Davis, F. C., Wong, W. H., and Weitz, C. J. (2002). Extensive and divergent circadian gene expression in liver and heart. *Nature*, **417**(6884), 78–83.

Stratmann, M. and Schibler, U. (2006). Properties, entrainment, and physiological functions of mammalian peripheral oscillators. *J Biol Rhythms*, **21**(6), 494–506.

Stratmann, M., Suter, D. M. M., Molina, N., Naef, F., and Schibler, U. (2012). Circadian Dbp transcription relies on highly dynamic BMAL1-CLOCK interaction with E boxes and requires the proteasome. *Molecular cell*, **48**(2), 277–287.

Strogatz, S. (2000). From kuramoto to crawford: exploring the onset of synchronization in populations of coupled oscillators. *Physica D: Nonlinear Phenomena*, **143**(1), 1–20.

Takahashi, J. S., Hong, H. K., Ko, C. H., and McDearmon, E. L. (2008). The genetics of mammalian circadian order and disorder: implications for physiology and disease. *Nat Rev Genet*, **9**(10), 764–75.

Tognini, P., Murakami, M., Eckel-Mahan, K., Newman, J., Verdin, E., Liu, Y., Baldi, P., and Sassone-Corsi, P. (2014). *in submission*.

Turek, F. W., Joshu, C., Kohsaka, A., Lin, E., Ivanova, G., McDearmon, E., Laposky, A., Losee-Olson, S., Easton, A., Jensen, D. R., Eckel, R. H., Takahashi, J. S., and Bass, J. (2005). Obesity and metabolic syndrome in circadian clock mutant mice. *Science*, **308**(5724), 1043–5.

Ueda, H. R., Hayashi, S., Chen, W., Sano, M., Machida, M., Shigeyoshi, Y., Iino, M., and Hashimoto, S. (2005). System-level identification of transcriptional circuits underlying mammalian circadian clocks. *Nature genetics*, **37**(2), 187–192.

Vollmers, C., Schmitz, R. J., Nathanson, J., Yeo, G., Ecker, J. R., and Panda, S. (2012). Circadian oscillations of protein-coding and regulatory RNAs in a highly dynamic mammalian liver epigenome. *Cell metabolism*, **16**(6), 833–845.

Wain, H. M., Bruford, E. A., Lovering, R. C., Lush, M. J., Wright, M. W., and Povey, S. (2002). Guidelines for human gene nomenclature. *Genomics*, **79**(4), 464–470.

Whirl-Carrillo, M., McDonagh, E. M., Hebert, J. M., Gong, L., Sangkuhl, K., Thorn, C. F., Altman, R. B., and Klein, T. E. (2012). Pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.*, **92**(4), 414–417.

Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, **34**(Database issue), D668–672.

Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., and Hassanali, M. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, **36**(Database issue), D901–906.

Wuarin, J. and Schibler, U. (1990). Expression of the liver-enriched transcriptional activator protein DBP follows a stringent circadian rhythm. *Cell*, **63**(6), 1257–1266.

Xie, X., Rigor, P., and Baldi, P. (2009). Motifmap: a human genome-wide map of candidate regulatory motif sites. *Bioinformatics*, **25**(2), 167–74.

Yalcin, B., Adams, D. J., Flint, J., and Keane, T. M. (2012). Next-generation sequencing of experimental mouse strains. *Mamm Genome*, **23**(9-10), 490–8.

Yan, J., Wang, H., Liu, Y., and Shao, C. (2008). Analysis of gene regulatory networks in the mammalian circadian rhythm. *PLoS Comput Biol*, **4**(10), e1000193.

Yoo, S. H., Yamazaki, S., Lowrey, P. L., Shimomura, K., Ko, C. H., Buhr, E. D., Siepka, S. M., Hong, H. K., Oh, W. J., Yoo, O. J., Menaker, M., and Takahashi, J. S. (2004). Period2::luciferase real-time reporting of circadian dynamics reveals persistent circadian oscillations in mouse peripheral tissues. *Proc Natl Acad Sci U S A*, **101**(15), 5339–46.

Zeller, M., Magnan, C. N., Patel, V. R., Rigor, P., Sender, L., and Baldi, P. (2014). A Genomic Analysis Pipeline and Its Application to Pediatric Cancers. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. Accepted.