Decoding heterogeneous big data in an integrative way

XIA ZHANG
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd
Part of the Bioinformatics Commons, and the Genetics Commons

Recommended Citation
ZHANG, XIA, "Decoding heterogeneous big data in an integrative way" (2013). Graduate Theses and Dissertations. 13630.
https://lib.dr.iastate.edu/etd/13630

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Decoding heterogeneous big data in an integrative way

by

Xia Zhang

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Genetics

Program of Study Committee:
Jeanne M. Serb, Major Professor
M. Heather West Greenlee
Dan Nettleton
Jeffrey M. Trimarchi
Maura McGrail

Iowa State University
Ames, Iowa
2013

Copyright © Xia Zhang, 2013. All rights reserved.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>NOMENCLATURE</td>
<td>vii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>viii</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>ix</td>
</tr>
<tr>
<td>CHAPTER 1. GENERAL INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Structure of the thesis</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Background</td>
<td>2</td>
</tr>
<tr>
<td>1.2.1 Revolutionary biotechnologies that shaped post-genomic era</td>
<td>3</td>
</tr>
<tr>
<td>1.2.2 Technology-driven opportunities and challenges</td>
<td>6</td>
</tr>
<tr>
<td>1.2.3 Big and heterogeneous data and their analyses</td>
<td>8</td>
</tr>
<tr>
<td>1.2.4 Systematic approaches</td>
<td>11</td>
</tr>
<tr>
<td>1.3 General statement of problems</td>
<td>13</td>
</tr>
<tr>
<td>1.4 Rationale and specific objectives</td>
<td>15</td>
</tr>
<tr>
<td>1.5 Literature Review</td>
<td>19</td>
</tr>
<tr>
<td>1.5.1 Cell fate determination and differentiation of mouse retina</td>
<td>19</td>
</tr>
<tr>
<td>1.5.2 Prioritization of candidates and integrations of networks</td>
<td>22</td>
</tr>
<tr>
<td>1.5.3 Integration across genomes</td>
<td>30</td>
</tr>
<tr>
<td>CHAPTER 2. MOUSE RETINAL DEVELOPMENT: A DARK HORSE MODEL FOR SYSTEMS BIOLOGY RESEARCH</td>
<td>34</td>
</tr>
<tr>
<td>2.1 Abstract</td>
<td>34</td>
</tr>
<tr>
<td>2.2 Introduction</td>
<td>35</td>
</tr>
<tr>
<td>2.3 Results and Discussion</td>
<td>37</td>
</tr>
<tr>
<td>2.4 Summary</td>
<td>51</td>
</tr>
<tr>
<td>2.5 Authors’ Contribution</td>
<td>53</td>
</tr>
<tr>
<td>CHAPTER 3. EnRICH: EXTRACTION AND RANKING USING INTEGRATION AND CRITERIA HEURISTICS</td>
<td>63</td>
</tr>
<tr>
<td>3.1 Abstract</td>
<td>63</td>
</tr>
<tr>
<td>3.2 Background</td>
<td>65</td>
</tr>
<tr>
<td>3.3 Implementation</td>
<td>67</td>
</tr>
<tr>
<td>3.4 Result and discussion</td>
<td>72</td>
</tr>
<tr>
<td>3.5 Conclusion</td>
<td>75</td>
</tr>
<tr>
<td>3.6 Authors’ Contribution</td>
<td>76</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure 2.1 The retinal cell types in the adult mouse retina ............................................................................................. 54
Figure 2.2 Time course of cell genesis in the developing mouse retina ............................................................................................. 55
Figure 2.3 A network of genes essential for Müller glia development ............................................................................................. 56
Figure 2.4 A network of genes essential for ganglion cell development ............................................................................................. 57
Figure 2.5 A network of genes essential for bipolar cell development ............................................................................................. 58
Figure 2.6 A network of genes essential for amacrine cell development ............................................................................................. 59
Figure 2.7 A network of genes essential for horizontal cell development ............................................................................................. 60
Figure 2.8 A network of genes essential for rod and cone photoreceptor cell development ............................................................................................. 61
Figure 3.1 EnRICH graphical user interface .................................................................... 77
Figure 3.2 EnRICH visualization window ....................................................................... 78
Figure 3.3 Case study workflow ......................................................................................... 79
Figure 4.1 Matching between mouse and human ........................................................ 94
Figure 4.2 The matchings of four organism contrasts and their evolutionary relationships ............................................................................................. 95
Figure 1 (Appendix C) Bipartite graph ............................................................................ 141
Figure 2 (Appendix C) Maximum cardinality bipartite matching ............................... 142
Figure 3 (Appendix C) Maximum weighted bipartite matching ................................. 143
Figure 4 (Appendix C) The trade-off between cardinality and weight ........ 145
Figure 5 (Appendix C) Bipartite matchings with the same cardinality ........ 146
LIST OF TABLES

Table 2.1 Pairwise correlation coefficients between genes of the photoreceptor-specific seed-network ................................. 62
Table 4.1 The identifications of putative orthologs of four methods .................................................................................. 96
## NOMENCLATURE

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BACs</td>
<td>Bacterial artificial chromosomes</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>cDNA</td>
<td>Complementary DNA</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>ncRNA</td>
<td>Non-coding RNA</td>
</tr>
<tr>
<td>ESTs</td>
<td>Expressed sequence tags</td>
</tr>
<tr>
<td>SNP</td>
<td>Single nucleotide polymorphism</td>
</tr>
<tr>
<td>NGS</td>
<td>Next generation sequencing</td>
</tr>
<tr>
<td>ChIP-chip</td>
<td>Chromatin immunoprecipitation with chip (microarray)</td>
</tr>
<tr>
<td>RNA-Seq</td>
<td>RNA sequencing</td>
</tr>
<tr>
<td>ChIP-Seq</td>
<td>Chromatin immunoprecipitation with sequencing</td>
</tr>
<tr>
<td>GO</td>
<td>Gene ontology</td>
</tr>
<tr>
<td>KEGG</td>
<td>Kyoto Encyclopedia of Genes and Genomics</td>
</tr>
<tr>
<td>CNV</td>
<td>Copy number variation</td>
</tr>
<tr>
<td>MS</td>
<td>Mass spectrometry</td>
</tr>
<tr>
<td>SAGE</td>
<td>Serial analysis of gene expression</td>
</tr>
<tr>
<td>CAGE</td>
<td>Cap analysis of gene expression</td>
</tr>
<tr>
<td>MPSS</td>
<td>Massively parallel signature sequencing</td>
</tr>
<tr>
<td>coIP/MAS</td>
<td>Co-immunoprecipitation coupled with mass spectrometry</td>
</tr>
<tr>
<td>LUMIER</td>
<td>Luminescence-based mammalian interactome mapping</td>
</tr>
<tr>
<td>NMR</td>
<td>Nuclear magnetic resona</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

I would like to thank my two major professors Jeanne M. Serb, M. Heather West Greenlee for being supportive, understanding, patient and positive, and other committee members, Dan Nettleton, Jeffrey M. Trimarchi and Maura McGrail, for their guidance and support throughout the course of this research. I would also like to thank Vasant Honavar and Julie A. Dickerson who served on my committee till preliminary exam.

In addition, I would like to offer my appreciation to the coordinator of IG (Interdepartmental Genetics) program Linda Wild. She is very nice and helpful to students and always wants to do the best for her jobs, and also the department faculty and staff. The colleagues I worked with during my research assistantship and teaching assistantship gave me sincere help and with them, my time at Iowa State University was a wonderful experience.

Finally, I really want to thank my family and friends for their love and care.
ABSTRACT

Biotechnologies in post-genomic era, especially those that generate data in high-throughput, bring opportunities and challenges that are never faced before. And one of them is how to decode big heterogeneous data for clues that are useful for biological questions. With the exponential growth of a variety of data, comes with more and more applications of systematic approaches that investigate biological questions in an integrative way. Systematic approaches inherently require integration of heterogeneous information, which is urgently calling for a lot more efforts.

In this thesis, the effort is mainly devoted to the development of methods and tools that help to integrate big heterogeneous information. In Chapter 2, we employed a heuristic strategy to summarize/integrate genes that are essential for the determination of mouse retinal cells in the format of network. These networks with experimental evidence could be rediscovered in the analysis of high-throughput data set and thus would be useful in the leverage of high-throughput data. In Chapter 3, we described EnRICH, a tool that we developed to help qualitatively integrate heterogeneous intra-organism information. We also introduced how EnRICH could be applied to the construction of a composite network from different sources, and demonstrated how we used EnRICH to successfully prioritize retinal disease genes. Following the work of Chapter 3 (intra-organism information integration), in Chapter 4 we stepped to the development of method and tool that can help deal with inter-organism information integration. The method we proposed is able to match genes in a one-to-one fashion between any two genomes.

In summary, this thesis contributes to integrative analysis of big heterogeneous data by its work on the integration of intra- and inter-organism information
CHAPTER 1
GENERAL INTRODUCTION

1.1 The structure of this thesis

Background (Chapter 1): I describe the big picture that this thesis relates to and provide essential background knowledge.

General statement of problems (Chapter 1): I summarize general problems that we can try to solve within the scope of this thesis.

Rationale and specific objectives (Chapter 1): Based on the general statement of problems, I provide a rationale for and identify the specific objectives addressed in this thesis.

Literature Review (Chapter 1): I review concepts and the body of literature that is essential to the specific objectives addressed in this thesis.

Chapter 2: Mouse Retinal Development: a Dark Horse Model for Systems Biology Research, Bioinformatics and Biology Insights 2011: 5 99-113. I justify the use of the mouse retina as a powerful model system to explore systematic and integrative approaches. I then summarize gene networks employed in the developing mouse retina. Last, I describe how these networks were re-discovered using high-throughput data.

Chapter 3: EnRICH: Extraction and Ranking using Integration and Criteria Heuristics, BMC Systems Biology 2013, 7:4. I describe the software we developed to filter, as well as integrate, both list and network data and present a case study in which this software was successfully used to identify retinal disease genes.
Chapter 4: Plugging into the tree of life: genome-wide homolog identification between model and non-model organisms. I introduce the genome-wide approach we proposed to match homologous genes/proteins in a one-to-one fashion between organisms, and present results on its performance.

Chapter 5: I summarize the contribution of my thesis work to the scientific community.

1.2 Background

In 1990, the National Institutes of Health (NIH) and the Department of Energy (DOE) initiated the Human Genome Project (HGP)[1]. The human genome was estimated to have about 3 billion base pairs and with the technologies at the time, it was broken into smaller pieces with a length ranging from 150,000 to 200,000 base pairs [1]. These smaller pieces were ligated into vectors of BACs so they could be inserted into bacteria and copied by bacterial DNA replication to make BAC clones. With BAC clones, human DNA was prepared in quantities large enough for sequencing. Before sequencing, each BAC clone was mapped to chromosomes to determine the precise location of the DNA sequence and its relationship to DNA sequences in other BAC clones [1]. Then, shotgun sequencing was applied to the DNA sequence in each BAC clone. With this approach known as ‘hierarchical shotgun’ [2], it was expected to take 15 years to finish the HGP. The private company Celera Genomics led by Craig Venter, a previous NIH scientist, claimed it would complete human genome sequencing using much less time and cost than the publicly funded team [3, 4]. The Celera Genomics approach known as ‘whole genome shotgun’ used similar sequencing technologies, but differed from ‘hierarchical shotgun’ in the strategies to break
up sequences and then put them back together [5]. ‘Hierarchical shotgun’ broke up the whole genome into large fragments of DNA of known positions and then sequenced each fragment of DNA by shotgun sequencing. This strategy requires the construction of a whole-genome map of these DNA fragments. ‘Whole genome shotgun’ shredded the whole genome into small pieces that are directly sequenced and then all the small pieces were put back together based on sequence overlaps. Celera Genomics spurred the HGP team to change its strategy and sped up the completion of HGP. In June 2000, the head of public HGP, Francis Collins, and that of Celera Genomics, Craig Venter, jointly announced the completion of human genome project, paving the way for post-genomic era.

1.2.1 Revolutionary biotechnologies that shaped post-genomic era

In addition to human genome project, Craig Venter also was involved in developing a gene discovery and tagging strategy known as Expressed Sequence Tags or ESTs[6] and pioneered its use in gene discovery [7, 8]. The technological advantage of this approach was that only a distinguishable fraction of cDNA was sequenced. So an EST is a sub-sequence of one cloned cDNA. Since cDNA is complimentary to mRNA, ESTs also represent portions of expressed genes and can be used to identify gene transcripts. Since 1991 when the term EST was coined[6], EST data experienced an exponential growth. Likely because ESTs have a broad application in gene discovery, genome annotation, gene structure identification, and SNP characterization [9]. ESTs even exerted an important effect in the design of cDNA microarrays [10], the most significant technology in gene expression profiling before Next Generation Sequencing or NGS technologies became popular.
Microarray technology [11] evolved from Southern blotting [12], a method routinely used to detect a specific DNA sequence in DNA samples. The specific DNA sequence is a so-called ‘probe’, and sequences in DNA samples that can be hybridized with the ‘probe’ are target sequences. The core biological principle that the microarray utilizes is the same as Southern blotting. That is, the hybridization between two complementary DNA strands. But in microarray, instead of a single probe, there could be a large number of probes. In this way many tests can be simultaneously carried out to detect probes in target sequences.

However, microarray technology is not simply an extension of Southern blotting by adding fluorescent technology [13]. Microarray technology requires solutions to at least three challenges. The first challenge is to put a synthesized probe on a solid chip. The second challenge is to generate detectable signals that can be quantitatively read to indicate the hybridization between target sequences and probes. Last, the target-probe signal must be separated from noise as cross hybridization may produce confounding signals. For the first challenge, Stephen Fodor and colleagues developed the techniques to synthesize oligonucleotides on a solid matrix by combinatorial chemistry synthesis [14]. They also solved the array-reading problem by adoption of fluorescent labeling techniques and confocal laser scanning, alleviating issues with the second challenge. For the third challenge, David Lockhart designed a single mismatch in oligonucleotide to eliminate confounding signals [15]. In 1993, Stephen Fodor co-founded Affymetrix [16], a company that produces a large variety of oligonucleotide microarrays. Based on the technology developed for DNA microarrays, other types of microarrays such as the protein microarray [17] and chemical compound microarray [18, 19] also emerged. Important technology such as chromatin immunoprecipitation (ChIP)ChIP-chip [20, 21] became possible with
microarray. Chromatin immunoprecipitation (ChIP) is a technique developed to investigate interactions between proteins and *in vivo* DNA, but when combined with microarray technology, these investigations can be genome-wide. That is, for a DNA-binding protein of interest, its binding sites across the whole genome could be detected at the same time.

As microarray technology became widely adopted in profiling gene expressions, next generation sequencing (NGS) technologies [22], driven by a strong demand in low cost sequencing, were also exponentially advanced [22, 23]. After the HGP was completed, several model organisms such as fly, mouse, rat, chicken, dog, etc. were also sequenced [24]. But the sequencing technology that generated these genomes was still Sanger-based capillary sequencing, the same as that worked in HGP. The Sanger-dominant sequencing landscape shifted to a NGS one, as multiple commercial sequencing instruments became available since 2005. These instruments all share a common mechanism of data generation that is radically different from Sanger-based capillary sequencing, though each of them has its distinct specifics regarding the method of library amplification, sequencing, detection and post-incorporation etc [25]. *In sample preparation*, BAC clones and DNA isolations are replaced by ligations of platform-specific adaptors to ends of DNA fragments and amplifications on solid surface. *In instrument use*, sequencing reactions are designed as a series of automatic repeating steps instead of semi-automated implementation of Sanger chemistry, making these instruments able to generate a much higher throughput of sequences per run at a lower cost. The nice qualities of NGS technologies have led them to not only replace Sanger sequencing, but also to occupy what used to belong to DNA microarray. Looking back, EST data contains sequences of gene transcripts and helped to
design the platform of DNA microarray. DNA microarray profiles expressions of gene transcripts. Now, with RNA-Seq [26], both the sequences and expressions of gene transcripts can be investigated. The aforementioned ChIP-chip is also evolving into ChIP-Seq [27].

Though many technologies contribute to the status quo in post-genomic era, technologies described in the paragraphs above composed the theme of the story, and continue to play important roles as this revolution grows.

1.2.2 Technology-driven opportunities and challenges

In 2001, Francis Collins and Craig Venter published their reports on the human genome [2, 5]. Since then, more model organisms have been sequenced. The availability of multiple genomes of model organisms led to the emergence of new fields, such as comparative genomics [28], which employs comparative analyses between two or more genomes. Studies on evolutionary relationships, gene identification and regulatory element identification embody the boom of comparative genomics.

In parallel to comparative genomics is functional genomics. The early studies using microarrays made scientists realize that changes in gene expression in isolation are difficult to interpret without sufficient knowledge of these genes [29, 30]. To make sense of gene expression, an approach widely adopted by researchers was to identify putatively relevant genes by comparing gene expression between two phenotypes. Microarray analysis became a powerful way to identify candidate genes for a phenotype of interest, accelerating the process to connect genotypes and phenotypes [31]. With microarray-related technology such as ChIP-chip, scientists can generate interactions between DNA-binding proteins and DNA in vivo in a high-throughput manner, escalating studies on
functional elements in the genome. Functional genomics emerged [32, 33] and developed as a field to study properties and functions of genes and gene products by integrative utilization of data generated by high-throughput technologies.

As NGS rapidly develops, it becomes more and more affordable to sequence new genomes and profile transcriptomes, deepening and widening the biological questions that scientists can investigate. With the enormous opportunities brought by high-throughput technologies, come a variety of challenges that scientists never faced before. The first challenge is high-throughput data analysis requires knowledge of statistics, computer science and biology. So it is important to optimize data analysis methods and to develop user-friendly tools for biologists to easily organize the data and conduct the analysis[34, 35]. The second challenge lies in the heterogeneous sources of data [36, 37]. Diverse data sources represent different lines of evidence, and leveraging all data is absolutely the ideal situation. However, this requires methods and tools that biologists can use to do an integrative analysis on different types of data in order to address questions that are better answered by such analysis. The third challenge is, as high-throughput data increases exponentially, the old systems of storing, retrieving and analyzing data are becoming inefficient in, and even incapable of, providing what scientists need [38-40]. To tackle the three challenges, an interdisciplinary effort should be devoted to generate methods and tools that can cope with and interpret big, heterogeneous biological data as well as develop information systems that effectively handle the data flow.
1.2.3 Big and heterogeneous data and their analyses

Overall, large, heterogeneous data can be grouped into several big categories. The first category is nucleic acid sequence data, which mainly consists of genomic DNA sequences and several subgroups of RNA sequences (e.g. small RNA and ribosome RNA). The second category is gene expression data at the RNA-level that measures expressions of gene transcripts. To date, this category contains mostly high-throughput expression data such as microarray and RNA-Seq, and these data are very helpful to effectively infer gene activities in a spatially defined and time-resolved way. The third category is protein data that includes protein-level expression, protein sequences and even secondary and three-dimensional structures of proteins. In addition to protein-level expression, protein structures are also important indicators/predictors of their biological functions, helping to translate static genetic information to dynamic carriers of cellular activities. The fourth category is interaction data such as protein-protein interactions, protein-DNA/RNA interactions and genetic interactions. Protein-protein interactions and protein-DNA/RNA interactions are physical interactions detected by techniques such as yeast two-hybrid [41] and ChIP-chip/ChIP-Seq [42], while genetic interactions, for example, detected by double mutant animals, reveal functional relations. The fifth category is metabolic data including metabolites that are intermediates and products of metabolism and biochemical processes between them. Metabolic data [43, 44] captures the cellular physiology that RNA expression and protein data alone cannot, and thus complements the big picture from genes to gene expressions to cellular activities. The sixth category is annotation data that comes from various sources such as standard biological annotation terms such as GO [45] and KEGG [46], and scientific literature or a more summarized version of them such as
Gene Wiki [47, 48]. Taken together, genomics data draws a map of a biological system. Transcriptomics and proteomics data illuminate intricate parts of this system while interaction data shows its underlying dynamics. Metabolic data can be seen as biochemical results that are generated from as well as interact with the intricacies of a biological system. They are very important to understanding how genetic information prints the integrative physiology of a biological system. Annotation data is contained in standard repositories that document what biologists learn about each aspect of this biological system and can be very helpful when leveraging from the known to study the unknown. In paragraphs below, data that reveal the inner side of a biological system, that is, genomics, transcriptomics and proteomics data, will be discussed with regards to its data analysis.

**Sequence data** now are mostly from genome sequencing/transcriptome profiling short reads generated by an NGS platform. The major analysis is to assemble short reads into a genome or gene transcripts. The bioinformatics community has created a variety of algorithms and tools to deal with genome/transcriptome assembly [49-52]. Sequence assembly and alignment programs are the most important tools to analyze NGS data. Most sequence assembly programs adopt a/the De Bruijn graph as the data structure and assemble short reads to long contigs and then to scaffolds by resolving the De Bruijn graph [53, 54]. Assembly programs also cater to either genome or transcriptome assembly based on different requirements genomes and transcriptomes impose on them. For example, a genome has many more repetitive regions and thus requires a huge and complicated De Bruijn graph to resolve, while transcripts are relatively small and utilize simple De Bruijn graphs. To identify structural changes such as a deletion, insertion, SNP, and CNV or to annotate genes/transcripts, aligners are must-run programs. Alignment
programs tuned to short reads employed FM-index on Burrow Wheeler transform, and those tuned to longer reads alignment use Hash-index and seed-to-extend strategy [55]. Complementing assembly and alignment programs, some tools [56] such as PASA [57] and Blast2GO [58, 59] focus more on gene structure prediction and gene function annotation. Together these tools compose a source pool from which scientists can customize their own analysis pipelines targeting specific purposes.

Gene transcript expression data has two sources: microarray and RNA-Seq. A typical microarray analysis pipeline includes pre-processing such as background correction, within array or across array normalization, and statistical analysis such as linear fitting and multiple testing to identify differentially expressed genes [30, 60, 61]. A variety of statistical methods and tools were developed for microarray data, as the microarray became a dominant method to profile gene expression [62]. R Bioconductor packages such as ‘affy’ and ‘limma’ became mainstream tools in scientific studies using microarrays [63], because these packages provide a full range of functions on microarray analysis and can be used in combination with other R packages to make the analysis more adaptable to specific needs. RNA-Seq analysis shares the statistical philosophy of microarray analysis on the identification of differentially expressed genes even it requires statistical models different from what microarray analysis requires. However it can also have its own agenda in specific analyses and involves another spectrum of tools [64]. For example, from RNA-Seq data structural variations of transcripts can be detected while microarray does not provide such information. In RNA-Seq data, short reads must be aligned to assembled transcripts or reference genome and then summarized into expression counts and then normalized into
expression value. In contrast, raw expression values in microarrays are usually the data handed to scientists and to get expression values only normalization is involved.

In addition to gene transcript expression data, **protein abundance data** is also an important expression data subcategory. Due to alternative splicing, post-translational modification and protein degradation regulation, RNA-level expression is related to, but does not always correlate to protein-level expression [65, 66]. Proteomic studies that survey protein abundance provide a direct measure of protein-level expression. There are two major ways of proteome analysis [67], the top-down and the bottom-up. The top-down means characterizing intact protein by mass spectrometry (MS) without proteolysis. Due to its technical challenges, the top-down is rarely used. The widely used bottom-up approach consists of protein fragmentation, fractionation and MS analysis [67]. In quantitative proteomics, peptide identification, peptide quantification and assembling from peptide to protein involve a wide range of methods, software and databases that are beyond the scope of work in this thesis. I referred to the review written by Matthiesen R [68] if the reader is interested in this topic.

1.2.4 Systematic approaches

The term “system” was rooted from general systems theory [69] and was defined as “an entity that maintains its existence through the mutual interaction of its parts.” Although systems biology has specific definitions from various fields, the consensus view of systems biology is a biological system with emergent properties that do not arise from an individual part or the linear assembly of them [70, 71]. In this thesis, systems biology is taken as a research methodology[72] to explore the systems behavior of a biological entity at
potential levels that cannot be studied by a reductionist methodology. Specific approaches that adopt this methodology are referred as ‘systematic approaches’ here.

Limitations of systems biology as a methodology include, but are not limited to, three factors. First, incomplete data makes model construction difficult and biased. Second, when heterogeneous data are incorporated, each dataset will have its own error and bias. Thus, it is challenging to make them comparable, complicating the error estimate and pre-experiment evaluation of the model. Third, model validation by experiments often is not affordable. To overcome these limitations, a lot of more effort is still needed for data integration, model construction and model validation.

Systems biology inherently requires a model to represent the components of a system so how the mutual activities of these components finally produce system behavior can be studied. In recent years, networks (graphs) have been the most prevailing data structure to construct disease models in systems biology studies [73-75]. This is because the most challenging diseases, like cancer and diabetes, usually arise from multiple factors and their complicated interactions [75-78]. These interactions can propagate the effect of a singular molecular perturbation to other parts of the system, eventually changing the system state [75, 78, 79]. Network/graph models in which nodes represent components and edges represent interactions is the most direct representation of the dynamics within a system. So network analysis becomes a very important part in systematic approaches.

As we mentioned in previous sections, the revolutionary biotechnologies generate tons of heterogeneous data awaiting integrative analyses. And such analyses are necessary or even routine when systematic approaches are applied. In section below, I will identify
how this thesis can contribute to the integrative analysis and thus help to pave the way for prosperous applications of systematic approaches.

1.3 General statement of problems

High-throughput technologies in the post-genomic era such as microarray, ChIP-chip, NGS, RNA-Seq, ChIP-Seq (see 1.2.1) have been driving the exponential growth of biological data (see 1.2.3). The enormous data reservoir has been filled by heterogeneous sources: genomics, transcriptomics, proteomics, interaction data, metabolic data and annotation data (see 1.2.3). Powerful technologies and abundant data provide biologists unprecedented opportunities to ask questions that can only be answered with a systematic and integrative view. For example, how do cells of different types arise from a homogeneous cell pool during development of an organism? To answer these questions, connections must be made between genes to gene expression and regulation/interactions to determine how the components work together as a system. With unprecedented opportunities come tremendous scientific and technological challenges such as developing methods and tools to analyze data, creating information systems to effectively handle data flow, and systematically integrating information to interpret data (see 1.2.2). There are many angles to view these challenges and one of them is the biologist’s view. Biologists would like to leverage heterogeneous big data for clues to answer their questions, yet methods and tools that can help them to implement this approach are relatively scarce. This problem could be read into three pieces in a divide-and-conquer philosophy. First, is it necessary to have a model system (see 1.2.4) from which biologists can formulate their questions, collect data, and conduct data analysis. Second, biologist-friendly methods and
tools must be developed to leverage heterogeneous big data (see 1.2.3). Third, methods and tools developed in one model system must be extended to other biological systems.

All three points mentioned above are so general that exhaustive explorations of all of them are far beyond the scope of this thesis. However, a heuristic strategy can still be employed to approach this problem with full consideration to all three aspects. In other words, for this thesis it is feasible to focus on one model system, use it to ask appropriate questions, and more importantly, develop one or two information-leveraging methods or tools that work for, yet are not bound by, the investigation on this specific model system. So what kind of leveraging strategy is broad and intuitive enough to generate methods and tools that have a wide application? For leveraging heterogeneous big data, integration is a must since different sources of information must be comprehensively utilized. Prioritization is also necessary to identify a focus in clues from a sea of data for further in-depth research. Methods and tools that help to integrate and prioritize heterogeneous big data would alleviate the need of leveraging heterogeneous big data and thus contribute to coping with the challenges imposed by high-throughput technologies and exponential growth of dataset size.

In summary, this thesis will try to solve at least two problems:

1. Conduct a heuristic exploration on a model system that is appropriate for asking questions that would be better answered by systematic and integrative approaches.

2. Develop methods/tools that can integrate and prioritize large, heterogeneous datasets to facilitate a systems biology investigation.
1.4 Specific rationale and objectives

Based on the general statement of problems, I provide a rationale to target three specific objectives as shown below.

Rationale 1

We know cell fate determination in the developing vertebrate retina is dependent upon a number of gene interactions (see 1.5.1). We also know that the developing vertebrate retina is an ideal model system to study cell fate determination and differentiation (see 1.5.1). First, it is a multicellular tissue whose development is well characterized and within which the sequence of cell genesis is well documented. Second, the development and cell genesis of this tissue is largely conserved among vertebrates. Third, this tissue is highly accessible and is very amenable to in vivo hypothesis testing. The mouse retina is such an ideal model system that has long been studied. With the developing mouse retina, activities of gene networks that underlie the cell fate determination and differentiation would take place in well-characterized cells with known birthdates and known locations within the tissue. Also, the role of hypothesized gene candidates and network interactions in cell fate determination and differentiation can be readily assessed. Additionally, knowledge learned on this model system can be leveraged to study other vertebrate retinas. So the developing mouse retina is an excellent model that systematic and integrative approaches can be applied to.

Retinogenesis is a developmental process during which different types of retinal cells arise from a pool of retina progenitor cells to form a functional retina. Many studies on this process have generated a reservoir of prior knowledge on essential genes to direct this process. These known essential genes mostly were discovered in small-scale gene-
knockout studies and thus only cover tiny patches of the whole map [80, 81]. At the same time, studies employing high-throughput technologies are generating more and more data awaiting further exploration [82] (Only two examples are cited here [83, 84] due to the large volume of data). So utilizing prior knowledge to extract relevant information from ever-increasing amount of data is urgently needed. To utilize prior knowledge, it is important to consider it in an in-depth developmental context as well as to summarize it in a data form that can be readily used to guide further investigation. This is because it does not make any sense to explore the effect of genes on cellular fate choice without specifying developmental context such as when (developmental period) and where (location within tissue) the gene is expressed. More importantly, in a certain developmental context, how these genes work together to exert the overall effect must be summarized in a simplified form that even researchers outside this specific area can easily understand.

**Objective 1**

Based on the above rationale, the first objective of this thesis is to summarize prior knowledge on cell fate determination and differentiation of mouse retinal cells in a data form appropriate for systematic approaches and leverage the prior knowledge to heuristically search high-throughput data.

**Rationale 2**

For questions such as what is the molecular basis of cellular determination and differentiation in developing retina, identifying key genes and their interactions within a certain developmental context is the first and foremost job. While clues from small-scale studies may be summarized to help identify key genes and their interactions, the
overwhelming amount of data that comes from high-throughput screening platforms is far beyond the scope of a manual search. Biologists urgently need tools that can assist them in integrating information to extract the most useful results. For example, when faced with a large number of gene candidates generated from microarray or RNA-Seq, biologists would want to narrow down a long list of genes to identify the most promising candidates for more in-depth investigation. In addition to the prioritization of gene candidates, integrating interactions among genes to construct a composite network is necessary to gain a bird’s eye view of known knowledge and grasp essential clues. Putative gene relationships can come from sources such as physical interactions identified by ChIP-chip/ChIP-Seq and yeast two hybrid, genetic interactions identified by genetic array, genetic knock-out and knock-in experiments, and coexpressed interactions identified by expression profiling. With these different sources of putative gene relationships, how could we integrate them for an even more reliable prediction on gene relationships for our biological questions? To approach this in a quantitative way would be difficult since these sources are across heterogeneous platforms and a quantitative measure that can be associated with a biological interpretation is tough to define. To approach this problem in a qualitative way might be a good choice, yet a qualitative integration method and tool is absent.

**Objective 2**

From the rationale, I concluded it was necessary to develop a tool that helps prioritize candidate genes and their relationships by qualitative integration of data from heterogeneous sources for systematic investigation and it would be the second objective of this thesis.
Rationale 3

The second objective aims to generate a tool that can qualitatively integrate heterogeneous information for prioritization of gene candidates and construction of networks of interest. When heterogeneous data are integrated to answer a biological question, it is more likely the data used are from the same organism. However, in some scenarios data from multiple organisms may be used. For example, comparing two or more organisms to detect conserved or divergent parts of the network of interest, or using closely related species as prior knowledge to leverage queries of the unknown in the organism of interest. Unlike intra-organism integration, inter-organism integration is quite complicated due to the fact that genes or proteins must be first matched or translated across organisms. An intuitive strategy is to use DNA or protein sequence homology as a bridge between organisms. However, a biological truth is one gene in one organism may have multiple homologous counterparts in another organism, and even worse two genes may share the same homologous counterparts. This is a major hurdle for application of network information since the across-genomes projections are based on multiple interacting genes instead of one individual gene. That gene correspondence between two genomes that is not one-to-one can dramatically affect the network topology, resulting in noise that may eclipse homologous relationships. So matching two genomes at the genome-wide scale and in a one gene to one gene fashion is highly desirable and a method and tool to solve this problem would largely contribute to integrating information across organisms.
Objective 3

Based on the above rationale, the third objective of this thesis is to develop a method/tool that can quickly match across genomes in a one gene to one gene fashion at the genome-scale.

1.5 Literature Review

1.5.1 Cell fate determination and differentiation of mouse retina

Multicellular organisms, as the name implies, have multiple cells as well as multiple cell types. For example, a human is a multicellular organism since it is composed of multiple cell types (skin cell, bone cell etc.). During the development of a multicellular organism, cells have to become specialized. For example, multiple cell types arise from a zygote (the initial cell formed when the egg and the sperm joined). In developmental biology, determination means cells choose a particular fate, while differentiation follows determination to elaborate a cell fate-specific developmental program. To know how a cell is determined to a particular cell fate is a fundamental question in developmental biology and has huge potential in biomedical application. For example, pluripotent cells such as embryonic stem cells and induced pluripotent stem (iPS) cells, are unspecialized cells that can divide and renew themselves for long periods and give rise to specialized cell types. The underlying logic in stem cell therapy is to derive specialized cells that are desired for a certain kind of disease [85]. So knowing how a particular cell fate is determined will largely contribute to stem cell therapy in at least two ways: one is to induce a desired cell fate, and the other is to identify a marker profile that characterize and confirm the identity of a particular cell type. Previous studies showed that it is molecular regulation, especially
transcriptional regulation that integrates extracellular and intracellular signals to program cell fate [86]. So understanding cell fate determination requires identification of cell fate regulators and investigation of their underlying mechanisms.

Several principles are inherently applied to the study of cell fate determination. First, the developmental stages need to be characterized for this model organism in order to know where and when the cell type of interest will appear. Second, manipulation of cells of interest and molecules of interest are allowed for this model organism in order to conduct experiments. Current methodology to study cell fate determination can still be divided into two categories even though they both share principles mentioned above. The first category of approaches usually employs the knock down, knock out and over-expression of several candidate genes in a cell population of interest either in vivo or in vitro to see whether these candidate genes are regulators of this cell type. Studies falling into this category either use small-scale screen to investigate the roles of a few candidate genes in the proliferation, death, morphology and function of a cell population of interest, or use high-throughput technology to profile general expression patterns of genes in a cell or cell population of interest. The second category of approaches emerged in recent years and adopts a systems biology strategy that utilizes various types of data to construct, test and optimize a model by experimental, mathematical and computational methods. These models [87-89] tried to capture the essential dynamic of interaction of multiple genes that play a role in cell fate determination.

The first and the second categories are complementary to each other and neither of them is dispensable. The first category lays down the foundation of, and will continue to exert important effect in, cell fate determination studies. But the lack of a systematic view
of the first category leaves a gap between various experimental data and a multi-dimension capture or modeling of a real situation, which drives the emergence of the second category. However, the second category is still in an exploratory stage in several ways: the first is the lack of data on a specific model system. Systems biology takes advantage of a variety of information. However, without the focus on a specific model system, the information is diluted due to the dispersion of them among different model systems. Second, since cell fate determination may not only involve cell intrinsic programming but also cell extrinsic signals [90], the context is worth careful consideration: a context of diffused systems like the immune system or a context of cultured tissue system is different to that of a complex tissue. Third, specific data integration and model construction in systems biology approach are still in the exploratory stage [91-93].

The vertebrate retina is a sheet structure composed of several neuronal layers, lying in the back of eye. Superficially it kind of functions like a film in the old-styled camera even it is actually more than that (which is beyond the scope of this thesis): the strikes of light on retina initiate a cascade of signals representing the visual information; and then the visual information received in the retina is relayed to the visual center of the brain [94]. The vertebrate retina consists of five well-defined layers. They are outer nuclear layer (ONL), outer plexiform layer (OPL), inner nuclear layer (INL), inner plexiform layer (IPL) and ganglion cell layer (GCL). These layers consist of seven cell types: rod photoreceptor, cone photoreceptor, bipolar cell, ganglion cell, amacrine cell, horizontal cell and Muller glial cell [80]. The seven cell types that comprise the retina are derived from a common pool of retinal progenitor cells [95]. The sequence of retinal cell genesis is highly conserved among vertebrates,[96-100] following a general progression of retinal ganglion cells (RGCs),
horizontal cells (HCs), cone photoreceptors followed by amacrine cells (ACs), and subsequently bipolar cells (BCs), rod photoreceptors and Müller glial cells (MCs).

The developing vertebrate retina is an excellent model to study cell fate determination. First, development of the retina is well characterized [101-103] and the sequence of cell genesis and differentiation is well documented. Thus, gene activities that underlie the fate determination in a particular retinal cell type will take place in known cells with known birthdates and known locations within the tissue. Second, the retina is highly accessible and is very amenable to in vivo hypothesis testing [104], thus the role of hypothesized gene candidates and their interactions in cell fate can be readily assessed to some extent. Third, with relatively limited well-defined cell types, retina is a complex tissue and thus cell fate determination can be examined within the context of a complex tissue. Fourth, retina in several vertebrates such as mouse, frog, fish and chick are extensively studied and the conservation of cell types and cell genesis among vertebrates offers a great opportunity to consider cell fate determination in the perspective of evolution. The nice properties mentioned above makes the developing vertebrate retina a model system that systematic approaches can be applied to.

1.5.2 Prioritization of candidates and integrations of networks

Prioritization of candidate genes

High-throughput technology is the method of scientific experimentation that allows the researcher to quickly conduct thousands and even millions of tests. Microarray is one of the most prominent high-throughput technologies, with either single array or two-color array to detect RNA expression levels of genes in a biological sample at the whole genome scale. Most microarray experiments are conducted to either obtain a differentially
expressed gene list between two conditions or profile the coexpression of genes across conditions. RNA-Seq [105] utilizes next generation sequencing technologies to sequence cDNA to get RNA content information. Other methods like SAGE [106, 107], CAGE [108, 109] and MPSS [110] also profile the RNA expression by sequencing. All these technologies have general goals similar to microarray analyses: to obtain differentially expressed genes among conditions of interest or coexpression gene clusters across conditions. Gene expression data generated by these high-throughput technologies are stored in publicly available databases like ArrayExpress and Gene Expression Omnibus. To identify differentially expressed genes, a procedure which includes normalizing array data [111], fitting a model to test parameters [112, 113] and controlling the false discovery rate [114] is already well established. A couple of tools [115, 116] to analyze microarray data are freely available on line, and statistical methods for RNA-seq are also emerging as next generation sequencing is increasingly used [22, 117]. While hundreds and even thousands of candidate genes now can be generated from a single high-throughput experiment, they are unlikely to get experimentally validated due to time and cost constraints. Thus it is imperative to identify the most potentially promising candidates through prioritization and ranking. Various ways [118-126] and tools [127-131] have been proposed to address this need and they are generally grouped into three categories.

The underlying principle of the first and most significant category is to find genes that are most relevant to a particular phenotype, for example, a disease. This principle evolved from the utilization of GO and KEGG enrichment [132-135] to an integrating methodology that utilizes heterogeneous data sources. These methods rank candidate genes by their similarities to the training data in data sources including published
literature, functional annotation ontology, signaling pathway, sequence, expression and interactions. Training data can be a set of seed genes involved in a certain biological process or the keyword describing this biological process. For many more specifics about these methods and tools (which are not very related to this thesis), we refer to the review by Tranchevent et al. [136]. While the similarity-based methods successfully utilize heterogeneous data sources and prior knowledge to rank candidates, they are also limited or biased by the training data and thus not sensitive to candidate genes that are from less well annotated biological processes or involved in cross talk between different pathways. Additionally, computational tools that implement these methods automatically query data source from pre-determined public databases and thus cannot be used in a case-specific way.

The second category is the statistical ranking of candidate genes. This approach essentially seeks to separate systematic effect from random effect. That is, to make the top genes most enriched in true positive. In a series of studies [120, 121, 137-139] that compared the performances of different gene ranking methods such as WAD (weighted average difference)[140], AD (average)[140], fold change (FC), rank products (RP)[141], moderated t statistic (modT)[113], significance analysis of microarrays (samT)[142], shrinkage t statistic (shrinkT)[143], and intensity based moderated t statistic (ibmT)[144], found that the fold change-based gene ranking methods work better in terms of reproducibility for differentially expressed genes. Particularly, the WAD, RP or ibmT statistical rankings had higher levels of sensitivity and specificity than other methods.

In addition to high-throughput technologies, candidate genes can be obtained by many other ways, including generated from analyses like differentially correlated
genes[145], summarized from literature, or as query results from databases or small-scale experiments. Accordingly, methods to prioritize candidate genes are not limited to the major categories reviewed above. For example, Oti et al. employed a cross-species comparison strategy from the evolutionary angle and prioritized disease genes using gene coexpression conservation across distantly related species [119]. Thus, it is impossible to cover all kinds of candidate genes and prioritization methods here. But the point is, biologists currently are faced with a large number of candidate genes from heterogeneous sources and lists of prioritized candidate genes generated by various prioritization methods. Biologists greatly need tools that can help to integrate lists of candidate genes.

**Integration of networks**

Many biological questions require investigation of gene relationships, which require the generation of interaction data. There are generally three groups of interaction data. The first group is direct, physical interaction of two entities within a biological system. The most prominent interaction types include protein–protein interaction, protein–DNA interaction, and metabolic interaction between metabolites. Popular approaches to detect protein–protein interactions are yeast two hybrid, coIP/MAS and LUMIER [146]. DNA-protein interactions can be assayed from small-scale experiments like Gel mobility shift, DNase Footprint and other footprinting methods, to high-throughput experiments such as ChIP-chip and ChIP-Seq. NMR spectroscopic and MAS based techniques are used to profile metabolic interactions [147]. Interactions in the second group are usually inferred from various experimental data. They may foretell physical interactions between genes or reveal their shared biological roles. For example, positive and negative interactions from genetics experiments showed the direct or indirect regulatory effects among genes; co-expression
interactions reveal that some genes have similar expression pattern and thus more likely work together for the same task. Although interactions at this level are not physical, they have experimental evidence to support potential physical interactions. Interactions in the third group are predicted by computational/statistical model instead of experimental data. For example, the putative regulatory sequence motifs shared among genes may suggest these genes are co-regulated. To properly interpret interactions at this level, the data used and the underlying assumptions of models should be carefully inspected. As discussed above, there is a huge amount of interaction data generated from heterogeneous sources. To make interaction data easily accessible, public databases such as MINT[148], BioGRID[149], IntAct[150] and etc. have been developed to store and query these data. In addition to these well-known databases, a lot of other databases that are more specific for certain purposes also store interaction data.

Network construction and integration are sometimes interwoven together. For example, one may need to integrate interactions from heterogeneous sources to construct a network; one may also wish to integrate several well-constructed networks to compare reproducibility. So the two terms are used here interchangeably, with network construction emphasizing the way to put interaction data into a network, and network integration the way to deal with heterogeneous sources. As mentioned above, there are three levels of interaction data. Only interactions of the first level are generated by experiments. Thus, in the most common scenario, a network has to be derived from both experiments and other data sources. A lot of effort has been devoted to methods and tools of network construction and integration.
Current strategies mainly fall into four groups. The first group includes those that utilize correlation coefficients, Euclidean distances, and information theoretic scores such as the mutual information to describe the dependencies between genes (nodes). Take a coexpression network as an example. Correlation coefficients of genes across several conditions are calculated and then genes are connected if their correlation coefficients pass a statistically reasonable cutoff value. Methods in this group have the advantage of simplicity and low computational costs and thus make the inference of large-scale network easy to handle. But the drawbacks are twofold: 1) the network is determined by an arbitrary threshold. And 2) the network is static yet gene activities are quite dynamic across conditions being used. The second group uses differential equations to describe gene activities as a function of other genes and non-gene factors. By solving differential equations, the association of each gene with other genes and non-gene factors can be revealed. This group requires the specification of function and some certain constraints for function parameters. The third group is Boolean networks in which nodes can be “on” and “off” and edges are represented by Boolean functions. To construct a Boolean network, continuous gene expression signals have to be transformed into binary data (“on” and “off”) and Boolean functions be inferred by various algorithms. While Boolean networks have inherent limitation since two states cannot adequately describe gene expression, they are dynamic, easy to interpret and very suitable for gene regulatory relations. The fourth group is Bayesian networks in which genes (nodes) are random variables associated with probability distributions and the relations among genes (edges) are described as conditional probabilities. Bayesian networks are very flexible at
integrating different types of data and prior knowledge. They can also be either static or
dynamic up to the given data.

As network construction strategies infer more and more networks from ever
increasing number of data sets, network integration has become imperative, because
researchers are faced with heterogeneous networks. Pioneering efforts have been devoted
to the network integration from several aspects. The first is data exchange: incompatible
format hampers the sharing of the interaction data stored across databases. BioPAX [158],
a computable language that can represent various pathway and interaction data,
established a standard syntax for this kind of information, largely facilitating the data
exchange. Although several databases of interactions almost adopt BioPAX, text files are
still the most popular ones among biologists. This is because, on one hand, there is still a
huge amount of interaction data not in BioPAX, such as interactions stored on personal
webpages, databases oriented for specific purposes, or data generated from independent
experiments. On the other hand, biologists who do not have programming skills or are
unwilling to do unnecessary format-converting work would likely adopt the easily-
perceived text file. With more effort devoted to tools such as Sig2BioPAX, a command line
JAVA program which converts text file to BioPAX format [159], text files can easily merge
into the data exchange flow. The second is data collection: cPath [160] and Pathway
Commons [161] can query and collect interactions by source (database) and interaction
type from a list of databases. While inter-database query and collection are almost
automated for several significant databases, manual collection and editing are still
unavoidable due to two reasons. The first reason inter-database mass-collection only
covers a certain list of databases. Second, compared with collection within a specific
database, mass data collection loses some information (such as experimental system and etc.), which may be useful to biologists. The third aspect of network integration is integration methods and tools: CABIN [162] enables the integration of interactions from multiple sources of evidence by assigning each source a weight based on the confidence in the evidence. GraphWeb [163] is a web server that allows the combination of multiple networks into a global network by specifying the values in edge and node settings. CABIN and GraphWeb assume that all edges in a source network have the same confidence in evidence and assign all the edges in the network the same score or weight. GeneMANIA [164, 165] allows more weighting strategies of interaction sources including query dependent weighting, GO-based weighting and equal weighting. GeneMANIA does excellent job when it comes to merging heterogeneous sources of data, user interface, and speed, but like CABIN and GraphWeb, some detailed information are lost due to the assumption that a source network is a homogeneous network. For example, all physical interactions are taken as the same in ignorance of different experimental evidence codes.

Network visualization is almost imperative in network integration due to several reasons. First, compared with other data structures, networks (graphs) rely more on image representation to convey information. Second, visualization is expected to give viewers the insight that may not be possible without it. Third, interactive visualization helps users assemble the network. For these three reasons, we can see visualization is quite data-oriented and study-specific. In addition, network visualization is coupled with network integration to some extent. For example, if the purpose is to merge large-size networks, visualization should be able to display a large-size network in an excellent layout that helps users to see the global features; if the purpose is to map across large-size networks to get a
single network with higher confidence, then in addition to the global layout, visualization should be able to distinguish the overlapping interactions. Generally, network size, interaction type and the needed analysis are the factors that fit the features of network visualization. A couple of network visualization tools are freely available now. For example, Cytoscape [166-168] is an influential network visualization platform, and its plug-ins render Cytoscape the ability to do various visual analyses of networks. VisANT [169-171] is a network visualization and analysis platform leaning towards the metagraph feature and gene set enrichment analysis. GenMAPP [172-174] focuses more on pathway analysis of high-throughput expression data. For other network and pathway visualization tools like Osprey [175], CellDesigner [176], BioLayout [177, 178], ProViz [179] and so on, each has its own set of build-in features. If the reader is interested in these tools, we refer to the review by Gehlenborg et al. [180]

1.5.3 Integration data across organisms

Integration of knowledge or data across organisms is widely applied in functional studies and evolutionary studies, or a mix of both. In functional studies, data integration is driven by the motive to leverage as much information as possible to study the function of one single gene or a set of genes. To name a few, functions of gene copies in several organisms are used to infer gene function in another organism [181]. Functional relationships of a set of genes are predicted by integrating high-throughput data from multiple organisms [182]. In evolutionary studies, it is necessary to incorporate knowledge from multiple organisms to study the evolutionary relationships at genome-level or gene-level [183]. For example, copies of the same gene from multiple organisms are used to construct a phylogenetic tree that reveals evolutionary relationships between genes. In
some studies, evolutionary and functional insights are equally important. For example, orthologous gene sequences are utilized to identify common allelic sequences that have similar effect on phenotypic variation, helping develop functional marker across several plant species[184].

Leaving aside different emphases of studies, the underlying logic of all examples listed above is the same. That is, DNA sequences encode RNA sequences and then proteins that are responsible for functions, so conserved DNA sequences encode common features of different organisms and likewise divergent DNA sequences reveal functional differences among them. So DNA /RNA /protein sequence homology is the key to any application that are across organisms. This means sequence alignment is very important for data integration across organisms. Many sequence alignment algorithms and software [55, 185-192] have been developed to cater to a variety of needs such as pairwise sequence alignment, multiple sequence alignment, motif finding, short read sequence alignment and genomic analysis to identify splice site junction, introns and ncRNAs. For the need of integrating data across organisms, pairwise sequence alignment and multiple sequence alignment are most important since they are the first and foremost step to identify homologous relationships which serve like bridges between organisms.

Even the underlying principle utilized in integration of knowledge across organisms is the same, specific strategies, methods and tools hinge on study emphases. In situations where function is the concern, sequence homology itself is more like a tool to leverage heterogeneous data of one organism in order to annotate that of another. For example, in tool like Blast2GO, to annotate the unknown genes, GO terms of its homologous genes are used and a Blast search of query sequence against NCBI database is how homology is used.
Similarly, a tool called IMP [193] constructed functional association networks from pathway annotations as well as GO terms of multiple model organisms to predict the function of query gene in any organism. More data sources [182, 194] such as microarray data and protein-protein interaction data are added into this category of methods and tools. In addition to functional prediction, data of multiple species are also used to prioritize candidate genes [195, 196] and construct biological networks [197]. It is worth noting here homologous relationships can be further divided into homologs, orthologs and paralogs [198]. Homolog is a general term that refers to genes that share significant sequence homology. Orthologs are homologs that descend from common ancestral sequence that are then separated by speciation event, while paralogs are homologs that stem from gene duplications within the genome. Since paralogous genes may or may not share more sequence similarity than orthologous genes, it is hard to tell them from each other by pure sequence homology. The term ortholog is often mis-used or loosely used in non-evolutionary studies to actually mean homologs that share very high sequence similarity. But to really name orthologous relationships, the construction of phylogenetic tree is necessary. So unlike that in functional studies, the way homology is used becomes quite important since it must adapt to different purposes such as phylogenetic analysis, detection and characterization of duplication and recombination events, estimation of selection pressures and population genetic inferences. On the contrary to heterogeneous data used in functional studies, evolutionary studies mostly use sequence data.

In this chapter, I summarized two general problems from the overview of background. To approach the two general problems within the scope of this thesis, three
specific objectives were identified and rationales provided. Chapter 2, 3 and 4 describe how the three specific objectives were achieved.
CHAPTER 2. Mouse Retinal Development: 
a Dark Horse Model for Systems Biology Research

Modified from a paper to be published in
Bioinformatics and Biology Insights

Xia Zhang¹,³, M Heather West Greenlee²,³,⁴,§, Jeanne M Serb¹,³

Abstract

The developing retina is an excellent model to study cellular fate determination and differentiation in the context of a complex tissue. Over the last decade, many basic principles and key genes that underlie these processes have been experimentally identified. In this review, we construct network models to summarize known gene interactions that underlie determination and fundamentally affect differentiation of each retinal cell type. These networks can act as a scaffold to assemble subsequent discoveries. In addition these summary networks provide a rational segue to systems biology approaches necessary to understand the many events leading to appropriate cellular determination and differentiation in the developing retina and other complex tissues.

¹Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa, USA
²Department of Biomedical Sciences, Iowa State University, Ames, Iowa, USA
³Interdepartmental Genetics Program, Iowa State University, Ames, Iowa, USA
⁴Bioinformatics and Computational Biology Program, Iowa State University, Ames, Iowa, USA
§Corresponding author: M. Heather West Greenlee, Department of Biomedical Sciences, 2008 Veterinary Medicine, Iowa State University, Ames, IA 50010. Phone: (515) 294-9251. Email: mheather@iastate.edu
Introduction

Multicellular organisms are made of tissues with multiple specialized cell types. Understanding the determination and differentiation of heterogeneous cell types within the context of complex tissues is fundamental to many areas of biology. This knowledge will have widespread application in treatment of developmental disorders and disease states such as cancer and will be critical for successful bioengineering and transplantation of tissue types to replace damaged or degenerate structures. The determination and differentiation of a given cell within a tissue is the culmination of the expression of many gene products and their subsequent intra- and intercellular signaling events. To address the challenge of understanding cell fate determination and differentiation we must adopt a broad systems biology approach to adequately take into account the activities of large numbers of genes and signaling pathways.

One emerging systems-based strategy to analyze and integrate large datasets is to generate network models, in which genes or proteins are represented by nodes and their relationships by edges in the graph (network). However, most large expression datasets are too sparse to infer high statistical confidence gene relationships which are based on estimate of a covariance matrix [199]. In addition, the networks generated de novo are often large, and do not facilitate prioritization of candidate genes and gene relationships for hypothesis based validation. To address this problem, we have previously described a heuristic approach that uses a seed network to summarize prior knowledge of a small part of the gene network involved in cellular development [200, 201]. The seed network can then be used to query large datasets in order to identify additional molecules with putative relationships to seed genes. These candidate molecules can then be used to expand the
network and are the basis for generating testable hypotheses to validate their functional role.

Cell fate determination and differentiation in the vertebrate retina provides many opportunities to generate and utilize systems-based tools and approaches to understand development of cells within complex tissues. First, development of the retina is well-characterized [101-103] and the sequence of cell genesis and differentiation is well-documented and largely conserved among vertebrates [96-100]. Thus, activity of gene networks that underlie the fate determination and differentiation in a particular retinal cell type will take place in known cells with known birthdates and known locations within the tissue. Second, the retina is highly accessible and is very amenable to in vivo hypothesis testing [104], thus the role of hypothesized gene candidates and network interactions in cell fate determination and differentiation can be readily assessed. Third, we can build on the foundational system-based approaches developed through the study of single cell organisms like yeast [202], diffuse systems like the immune system [203], or cultured tissue systems [87], and extend these methods to examine the development of more complex tissues that comprise living organisms.

Here we review what is presently known about the genetic networks that underlie cell fate determination and differentiation in the developing retina. The developing retina is an extensively reviewed [80, 81, 95, 204, 205] system regarding cell fate determination during retinogenesis, but a summary of literature-curated gene networks underlying differentiation of each retinal cell type has not been previously presented. In order to demonstrate its potential as a model to study determination and differentiation of multiple cell types within the context of a complex tissue, we have assembled seed networks to
summarize what is known about the genes and their relationships that underlie cell fate
determination and largely influence the differentiation of each of the basic retinal cell
types.

Results

Retinal Cell types

The mature mouse retina is composed of seven basic cell types, six neuronal and
one glial (Figure 2.1). While this review focuses on only the differentiation of the basic cell
types, many retinal cells can be further subdivided morphologically, biochemically and
functionally [206-216]. Photoreceptors (rods and cones) reside in the outer nuclear layer
(ONL) and are responsible for phototransduction and necessary for vision [217].
Photoreceptors synapse with bipolar cells, neurons that reside in the inner nuclear layer
(INL). Bipolar cells relay visual stimulus to retinal ganglion cells in the ganglion cell layer
either directly or indirectly via amacrine cells, which also reside in the INL. Other cells
present in the INL are horizontal cells, which mediate lateral interactions between
photoreceptors and Müller glia that play a critical role in retinal homeostasis [218]. Axons
of the retinal ganglion cells project into the visual centers in the brain, thereby relaying the
visual information detected by the retina. While appropriate processing of visual stimuli
requires the function of all retinal cell types, most blinding retinal diseases are the result of
the degeneration of photoreceptors or ganglion cells [219, 220]. Interestingly, the seven
cell types that comprise the retina are derived from a common pool of retinal progenitor
cells [95]. Thus, the developing retina provides a relatively simple, yet elegant system to
study the generation and maturation of a complex tissue. We know that the cell fate
decisions made by retinal progenitor cells are governed by an intrinsic genetic program
that determines their response to extrinsic cues from their environment [80, 95]. The sequence of retinal cell genesis is highly conserved in vertebrates [97, 100, 221-224], following a general progression of retinal ganglion cells (RGCs), horizontal cells (HCs), cone photoreceptors followed by amacrine cells (ACs), and subsequently bipolar cells (BCs), rod photoreceptors and Müller glial cells (MCs) (Figure 2.2). Based on this general progression of birth order, retinal cell types can be divided into cohorts of early-born cells which include ganglion cells and cone photoreceptors, and late-born cells which include rod photoreceptors, bipolar cells and Müller glia [97].

Gene Families that Underlie the Specification of Retinal Cell Types

There are a number of genes that are well known to act in the specification of and/or largely influence the differentiation of retinal cells. They compose a regulatory network that can integrate extrinsic information through signaling pathways like Notch, as well as implement intrinsic programming via transcription factors, many of which can be grouped into the basic helix loop helix (bHLH) gene family and the homeobox gene family.

The family of basic helix loop helix (bHLH) genes is characterized by an α helix-loop-α helix structural motif. The bHLH genes Mash1, Math3, NeuroD, Math5 and Ngn2 regulate each other to specify neuronal types in developing retina [205, 225], while Ptf1a [226, 227], Bhlhb4 [228] and Bhlhb5[229] have roles in the development of more specific retinal cell types or subtypes. Other family members such as Hes1 and Hes5 work as effectors of Notch signaling. These bHLH genes interact with several other genes including Pax, CVC, POU, Lim, Sox and Dlx. The Pax gene subfamily has critical roles in embryogenesis [230] and Pax6 functions as an early regulatory gene in the development of eye [231]. In contrast, the CVC homeodomain subfamily members Vsx1 and Chx10 have more specific roles in
retinogenesis across vertebrate species like mouse [232, 233], chicken [234] and fish [235, 236]. The POU homeodomain subfamily members have a variety of functions related to neural development [237], and genes Brn3b, Brn3c and Brn3a are important in the development of retinal ganglion cells. The LIM homeodomain gene subfamily are involved in neural patterning [238] with Isl1 and Lim1 playing crucial roles in retinal development. The Sox subfamily genes [239] and Dlx genes [240] are indispensable in many aspects of development including neurogenesis, and Sox2 [241], Sox4 [242], Sox8 [243], Sox9 [243], Sox11 [242] are implicated in retinogenesis. Like the bHLH genes, some homeobox genes like Vsx1 [244, 245], Barhl2 [246] and Irx5 [247] appear to specify retinal cell subtypes. Together these genes (see Appendix A. Supplementary Information) work in concert to specify cell fate in the developing retina.

Using a scaffold of these gene family members, we developed a seed network to summarize key gene relationships that govern the development of each of the retinal cell types in mouse retina. These seed networks are based on published studies that have demonstrated a role for the seed genes in the determination and differentiation of retinal cell types via either loss of function experiments [226, 248-255] gain of function experiments [241, 256] or transcriptional regulation experiments [243, 257]. Genes involved in the specification of multiple retinal cell types (see Appendix A. Supplementary Information) are not always included, in a given seed network due to the lack of strong evidence they interact with other essential genes in the seed network specifying a particular cell type. These seed networks can be used in two complimentary ways: 1) to design database queries to identify additional key molecules for cell-specific development, 2) to assemble a comprehensive summary of known gene relationships and identify key
decision points in cell-specific specification that may be important regulatory targets for future application.

**Müller Glial Cells**

The gene relationships that underlie Müller glia determination and differentiation are summarized in the seed network in Figure 2.3. Müller glia are the only glial cells to arise from the retinal progenitor cell population. Thus, the factors which influence the progenitor cell choice between gliogenesis and neurogenesis are critical for the creation of these cells. Previous work has demonstrated that Notch signaling plays a major role in the choice between neural and glial cell fate [258]. Notch is a transmembrane receptor that functions at the cell surface to both receive extracellular signals and to regulate gene expression in the nucleus. Notch signaling is widely used to control developmental processes in many animal species [259]. In the developing retina, the Notch pathway is implicated in the control of progenitor cell proliferation and apoptosis, as well as the multipotency of progenitor cells [260]. In addition to its role in maintaining the undifferentiated and proliferative state of retinal progenitor cells (RPCs), Notch also seems to regulate the neuronal versus glial cell fate choice by inhibiting the photoreceptor cell fate in mouse retina [261, 262].

As essential effectors of Notch signaling [263], bHLH genes Hes1 and Hes5 have partly overlapping but distinct roles in Müller cell determination and differentiation. Both Hes1 and Hes5 are thought to repress expression of neuronal bHLH genes [205]. However, their specific target genes appear to be different, since Hes1 maintains the progenitors and inhibits both neuronal and glial differentiation, whereas Hes5 cooperatively regulates maintenance of progenitors but promotes the glial cell fate [248, 249]. Specifically, Hes1 is
known to inhibit the proneuronal gene Mash1 [264] and thus promotes glial cell
determination. Consistent with their different effects, both Hes1 and Hes5 are expressed in
undifferentiated cells while Hes5 is also expressed in differentiating Müller glial cells.

The homeobox gene, Rax, promotes the glial cell fate choice, potentially via
activation of promoters of Notch1 and Hes1 [258]. The SRY box genes Sox8 and Sox9 have
also been implicated in the specification of Müller glial cells [243, 265], though neither of
them alone is sufficient to induce Müller glial cell differentiation. Notch signaling regulates
Sox8 and Sox9 transcription, though it does not appear to be through its activation of Hes1
and Hes5 [243].

Retinal Ganglion Cells

The gene relationships that underlie retinal ganglion cell determination and
differentiation are summarized in the seed network in Figure 2.4. The bHLH gene Math5
plays a critical role in retinal ganglion cell (RGC) development. The targeted deletion of
Math5 results in the loss of more than 80% of RGCs [251], and a cell fate shift to other
retinal cell types [251, 266, 267]. It seems that Math5 underlies RGC differentiation in two
ways. First, Math5 is very important to activate a downstream transcriptional network that
controls ganglion cell differentiation and development [268, 269]. Second, Math5
suppresses other bHLH proneuronal genes such as Math3, NeuroD and Ngn2 that are
involved in the adoption of other retinal cell fates [267, 269]. The available evidence
suggests that Math5 is directly regulated by Pax6 [270, 271]. Downstream of Math5, Brn3b
and Isl-1 are known to play critical roles in ganglion cell differentiation [251, 272, 273].
Brn3b, a POU subfamily gene, while not required for the initial commitment of RGC fate, is
essential for early retinal ganglion cell differentiation [250, 274]. Homozygous disruption
of Brn3b leads to a selective loss of 70% RGCs [250] suggesting not all RGC differentiation is dependent on Brn3b [275]. Consistent with this, it is hypothesized that Brn3b regulates genes important for formation of RGC axons and axon path-finding [274, 276]. In addition to loss of Brn3b, deletion of the Lim family gene Isl-1 also causes a marked reduction in the number of ganglion cells [252]. Recent studies indicate that both Isl-1 and Brn3b regulate genes such as Eomes and Shh [272, 273]. Eomes is a T-box transcription factor, now known as a direct target of Brn3b and required for RGCs and optic nerve development [277]. Other Brn3b-related genes are also found to contribute to ganglion cell development. For example, the zinc finger protein Wt1, acts upstream of Brn3b, and plays a role in the development of RGCs [278, 279]. Barhl2 functions downstream of Brn3b to regulate the maturation and survival of RGCs [246]. Math5 and Brn3b are essential for ganglion cell determination. In addition here are other Brn3b dependent-genes [257], Math5-dependent genes [269], and genes identified in RGC single cell expression studies [280]. However, the relationships of these genes to the network described here are not yet understood and were not included in our seed network.

Bipolar Cells

Compared to other retinal cell types, data supporting the relationships among genes essential for bipolar cells specification and differentiation are relatively sparse; however, the genes with key regulatory roles in bipolar cells determination and differentiation are summarized in Figure 2.5. The bHLH gene Mash 1 plays a pivotal role in bipolar cell differentiation. In both rat and mouse, the onset of Mash1 expression most closely correlates with the appearance of bipolar cells and Müller glia [281, 282]. In Mash1 -/- retinal explants, the differentiation of all late born retinal cells (bipolar cells, rod
photoreceptors and Müller glia) was delayed, and the number of the mature bipolar cells was significantly reduced, though the number of vimentin-positive cells (likely Müller glial cells) was increased [283]. Additionally, Mash1 is expressed by a subset (10-30%, depending on age) of the total proliferating progenitor cells, providing a molecular maker of heterogeneity among retinal progenitor cells (RPCs) [282]. Together, this evidence suggests that Mash1 plays a role in the commitment and/or differentiation of late born retinal cells, particularly bipolar cells.

Mash1 and Math3 are co-expressed in various regions of CNS suggesting these genes may have some functional redundancy [254]. Interestingly, the *Xenopus* homolog of Math3, Ath3, was shown to directly convert non-neuronal or undifferentiated cells to a neural fate [284], though the phenotype of Math3 (-/-) mice suggests Math3 is not essential for neuronal commitment [254]. However, in Math3 (-/-)-Mash1 (-/-) mice, in regions where the two genes are normally co-expressed, neuronal fate is blocked at the neural precursor stage and cells that normally differentiate into neurons adopted the glial fate. The retinas in these animals lack bipolar cells and have a significantly increased population of Müller glia [254]. It has been shown that Math3 and Mash1 are expressed by differentiating bipolar cells in the retina [282, 284]. However, misexpression of Mash1 or Math3 does not promote bipolar cell generation, rather it inhibits Müller gliogenesis[285]. Taken together these studies suggest that Mash1, with the cooperation of Math3, prevents gliogenesis in the developing retina and contributes significantly, but not entirely, to the specification of the bipolar cell fate.

The expression of the homeobox gene Chx10 is also integral to bipolar cell fate. Chx10 is restricted to the inner nuclear layer (INL) in the mature retina, though in the
developing mouse eye, the Chx10 transcript is confined to the anterior optic vesicle and all neuroblasts of the optic cup [286]. Loss of Chx10 results in reduced proliferation of retinal progenitors and a specific absence of differentiated bipolar cells [255]. Misexpression of Chx10 induces generation of inner nuclear layer cells [285], while misexpression of Mash1 or Math3 together with Chx10 increases the number of mature bipolar cells while decreasing the mature Müller glial cell number [285]. Thus, it is proposed that Chx10 confers the specific inner nuclear layer identity to retinal neurons while bHLH genes such as Mash1 and Math3 subsequently specify the bipolar cell fate [285]. In addition, Chx10 promotes bipolar cell fate determination by inhibiting photoreceptor specification, presumably by acting downstream of Otx2 or other Otx genes [287]. Otx2 subcellular localization is hypothesized to play a role in the rod versus bipolar cell fate choice [288]. In the retina of a postnatal, bipolar-cell-specific-Otx2 conditional knockout mouse the expression of mature bipolar cell markers is significantly down-regulated [289], demonstrating its importance in bipolar cell differentiation.

**Amacrine Cells**

The gene relationships that underlie amacrine cell determination and differentiation are summarized in the seed network in Figure 2.6. For amacrine cell specification, the bHLH gene Math3 cooperates with another bHLH gene, NeuroD, and amacrine cells are completely missing in Math3-NeuroD double mutant retinas. The cells in the double knockout retinas that fail to differentiate into amacrine cells adopt both ganglion and Müller glial cell fates. However, while these genes are necessary for amacrine cell fate determination, they are not sufficient; misexpression of either Math3 or NeuroD alone cannot induce amacrine cell genesis [256].
In the Pax6-knockout mouse retina, the retinal progenitor cells become totally restricted to an amacrine cell fate [270]. While misexpression of Pax6, Math3 or NeuroD alone does not induce amacrine genesis, the misexpression of a combination of bHLH genes Math3 or NeuroD with homeobox genes Pax6 or Six3 (the transcription of which is independent of Pax6 [290]) does promote amacrine cell genesis [256]. Furthermore, misexpression of Pax6 with only Math3 results in the production of amacrine cells and horizontal cells, while the combination of Pax6 and NeuroD predominantly increases only the number of amacrine cells, suggesting that when expressed with Pax6, NeuroD is more specific for amacrine cell differentiation than Math3 [256]. The homeobox gene, Sox2, is expressed in a subset of amacrine cells and misexpression of Sox2 results in a dramatic increase of amacrine cells in INL. Experimental evidence indicates that Sox2 transcriptionally induces Pax6 and may also induce NeuroD [241]. Taking all these data into account, it appears that Sox2 functions upstream of Pax6 and NeuroD to affect/promote amacrine cell fate.

The expression of the forkhead gene family member Foxn4 in mouse retina correlates closely with the birth date of amacrine cells and misexpression of Foxn4 promotes amacrine cell genesis [253]. Further, Foxn4-null mice exhibit a significant decrease in amacrine cells and a complete loss of horizontal cells [253]. The effect of Foxn4 on amacrine cell differentiation may be via activation upstream of NeuroD and Math3 signaling, since in Foxn4-/- retinas there is a marked downregulation of NeuroD and Math3 with no observable alteration in Math5, Ngn2, Chx10 or Pax6 expression [253]. Downstream of Foxn4 is Ptf1a[291] Lineage tracing reveals that Ptf1a expression in the developing mouse retina marks the horizontal and amacrine cell precursors [227].
Ptf1a affects the differentiation of a small population of amacrine cells and the entire population of horizontal cells. While Foxn4 may influence amacrine cell differentiation via NeuroD and Math3, Ptf1a does not appear to work in this way as in the Ptf1a-null retina expression of the two genes was unaffected [226, 227].

Horizontal Cells

The gene relationships that underlie horizontal cell determination and differentiation are summarized in the seed network in Figure 2.7. It appears that amacrine (Fig. 6) and horizontal (Fig. 7) cell fates are linked as they share several key regulatory genes including Foxn4, Ptf1a, Math3, and Pax6 (Figs. 6 and 7). As previously mentioned, misexpression of Pax6 with Math3 results in an increase of both horizontal cells and amacrine cells, though the effect on horizontal cell genesis is greater (14% increase) than the effect on amacrine cell genesis (7% increase) [256]. At the same time, deletion of Foxn4 results in complete loss of horizontal cells, presumably via the downregulation of Math3 [253].

Prox1, the Prospero-related homeobox 1, is also important for horizontal cell differentiation. Prox1 is expressed in, and is required for efficient cell cycle exit for, early RPCs (but not in late RPCs) [292]. Prox1-null retinas exhibit a complete loss of horizontal cells and the misexpression of Prox1 results in the production of horizontal cells [292, 293]. Considering the fact that there is a lack of Prox1 expression in Foxn4-null retina and a downregulation of Prox1 in Ptf1a-null retina [226, 253], Prox1 seems to promote horizontal cell fate by acting downstream of the Foxn4-Ptf1a axis. Downstream of Foxn4-Ptf1a-Prox1 is another essential gene, Lim1 [291]. Lim1 is required for specific
morphogenesis of horizontal cells in chick retina [294]. In mouse retina, Lim1 is essential to instruct the differentiation and migration of horizontal cells to the correct laminar position [295, 296].

**Cone and Rod Photoreceptors**

Both cones and rods employ phototransduction, a process that captures and converts photons of light to an electrical signal; however, each cell type expresses a particular visual protein (opsin) to absorb a specific portion of the light spectrum. In mice, cones express either a S-opsin (short wavelength sensitive) or a M-opsin (middle wavelength sensitive) while rods express rhodopsin. Interestingly, both rod and cone photoreceptors share several key genes essential for cell fate specification and differentiation. Thus, the relationships of genes underlying the differentiation of cones and rods are shown together in a single network (Figure 2.8).

NeuroD is the only bHLH gene known to be essential for photoreceptor differentiation. NeuroD is expressed in developing photoreceptors and is maintained in a subset of mature photoreceptors in the adult mouse retina [297, 298]. In the NeuroD-null retina, the number of rods is reduced, while the number of the bipolar cells is increased in a dose-dependent fashion [297]. Misexpression of NeuroD not only blocks gliogenesis, but also favors rod photoreceptor differentiation while reducing bipolar cell differentiation [297]. NeuroD is also necessary for sustained expression of TRβ2, an essential gene for cone photoreceptor development [298].

Photoreceptor cell types are generated by common activity of genes like Crx (Cone rod homeobox), Nrl (neural retina leucine zipper), and Nr2e3 where the expression of one gene promotes rods and suppress cones. For example, Crx is expressed early in the
developing retina, and is predominantly expressed in photoreceptors in mature retina [299]. Crx transactivates the Rhodopsin promoter and acts synergistically with Nrl to drive rhodopsin expression in rods [299]. Crx also activates cone opsins [300, 301]. Two genes are known to suppress Crx function Ataxin-7 [302, 303] and BAF [304], both contribute to photoreceptor degenerative disease. Otx2, a member of Otx homeobox gene family, transactivates Crx [305] and misexpression of Otx2 directs retinal progenitor cells towards photoreceptor fate but Crx does not [305].

Nrl is a basic motif –leucine zipper transcription factor preferentially expressed in rod photoreceptors [306, 307], which positively regulates rhodopsin [308, 309]. In the Nrl-/− mouse retina, cone-like photoreceptor cells are clearly different from WT rods and cones, revealing a functional transformation from rods to S-cones [310, 311]. From these results, it is inferred that Nrl modulates rod-specific genes as well as inhibits S-cone differentiation through the activation of Nr2e3 [311, 312]. Nr2e3 expression is restricted to photoreceptor cells. It is a ligand-dependent transcription factor that requires itself for the repression of its own transcription [313, 314]. Mutation of Nr2e3 causes enhanced S cone syndrome (ESCS) [315], a retinal degenerative disease in humans that results in an abundance of short-wavelength sensitiely cones (S cones) at the expense of rod photoreceptors [316]. It is hypothesized that when photoreceptors are first generated the defective Nr2e3 cannot prevent a ‘default’ shift of rod progenitors to an S-cone fate, producing a large number of S-cones and an absence of rods [317]. This is supported by the fact that Nr2e3 acts as a repressor of cone-specific genes in rods [318], and directly interacts with Crx to enhance rhodopsin and repress cone opsins [319].
In addition to upstream genes including Otx2, Crx, Nrl and Nr2e3 as well as photoreceptor-specific genes like rhodopsin, S-opsin and M-opsin, retinoid receptors are indispensable for appropriate photoreceptor differentiation. Retinoid receptors belong to a steroid receptor superfamily of proteins that serve as ligand-dependent transcription factors. Retinoic acid (RA) plays its role in transcription through retinoic acid receptors (RARs) and retinoid X receptors (RXRs). 9-cis RA binds to and transactivates both RXRs and RARs [320]. In addition, 9-cis RA directs progenitor cells to the rod cell fate through activation of members of the steroid/thyroid superfamily of receptors [321]. Another effector of this family, thyroid hormone (TH), is found to induce progenitor cells to differentiate into cones in embryonic rat retinal cultures [322]. Many effects of TH are mediated by TH receptors (TRs) [323]. The most important TR in retina development is TRβ2. TRβ2 is expressed in the outer nuclear layer of the embryonic retina [324, 325]. The mouse retina has an opposing S-cone (greater expression ventrally) and M- cone (expressed more dorsally) distribution. Deletion of TRβ2 in mice causes the selective loss of M-cones and a concomitant increase in S-opsin immunoreactive cones, disturbing the gradient of an opposing S- (ventral) and M- cone (dorsal) distribution [326]. TH is also required to inhibit S-opsin and activate M-opsin expression [327]. Other studies confirm that thyroid hormone action is required for normal cone opsin expression during mouse retinal development [328, 329]. RXRγ cooperates with TRβ2 to suppress S-opsin in all immature cones and in dorsal cones of the mature retina though it is not necessary for M-opsin regulation [330]. Finally, RXRα acts in synergy with Crx to activate many cone-specific genes [331].
Identification of experimentally-determined gene relationships in a high throughput gene expression dataset

The gene relationships in the seed networks described above are supported by experimental evidence and thus have been validated in the narrow sense by identifying direct or indirect interactions between two genes under particular experimental conditions. The next step to identify the 'system' of genes that work together to influence cell-specific determination and differentiation will require the use of large gene expression datasets and potentially additional dataset types such as protein-protein interaction datasets, ChIP-Chip datasets, datasets from animals with specific mutations, etc. We have previously demonstrated the successful application of literature-derived seed-networks to query high-throughput gene expression datasets [200, 201]. One motivation for this review article was to assemble the available experimental evidence in a way that it might be readily applied to future studies of other cell types, and perhaps to even guide the experimental design processes that underlie the generation of new datasets.

An implicit assumption when using large gene expression datasets is that legitimate gene relationships will be discoverable by identifying a correlation of expression between them. An important question, then is, are known experimentally-determined gene relationships identifiable in large gene expression datasets as high correlation coefficients? We used the seed-network that describes photoreceptor differentiation (Figure 2.8) to address this question.

Using previously published data collected from developing rod photoreceptors isolated from the retina at E16, P0, P2, P6 and P10 [332], we calculated the correlation coefficients between all pairs of genes (edges) present in the seed network (Table 2.1). In
the photoreceptor seed-network, there were 13 genes and 17 edges (relationships) between them. Two genes (BAF and 9-cis-RA) were not present in the dataset, which left 15 edges to identify. Seven of the 15 edges were recognized as high correlation coefficients (>0.85) and an additional three of the 15 edges were supported with weaker correlation coefficients (>0.45).

Thus, two-thirds of the seed-network relationships are present in the dataset and nearly half of the seed-network relationships are strongly correlated. Encouragingly, our result suggests that a significant number of legitimate gene relationships can be discovered using gene expression data. Previously, we have used seed networks to discover new candidate genes by focusing on genes that were correlated with multiple seed-network genes [200, 201]. Ultimately, it appears that it will require a combination of datasets and approaches to describe the entire gene network that underlies cell fate determination and differentiation.

Summary

The seed networks presented here can be the basis for queries of high throughput datasets to identify larger, more comprehensive networks that participate in cellular fate specification and differentiation in the developing mouse retina. In addition to summarizing prior knowledge of these processes, seed networks can also be the basis for comparative studies between tissue types within a species or between diverged organisms in order to identify genetic pathways that are conserved through development and evolution [201, 333-335]. While a more generalized gene-by-gene comparative approach has been effective in identifying orthologs that may play a role in a complex process or a disease state in different organisms [336-339], it is the conservation of not only the gene,
but of its relationships to other genes in a network, that dramatically increases the likelihood that the gene, in fact, functions in similar way. Being able to include relacional data is one advantage of the seed network approach over more generalized comparative studies. The effectiveness of a cross-species seed network approach has been demonstrated elsewhere [201, 334].

These seed networks were constructed to help demonstrate the potential of the developing vertebrate retina as a model system for the development and evaluation of systems based approaches. In addition to its characteristic organization and developmental time course, there is a significant amount of high throughput data that has been collected from the developing retina [340-344], and single cells from the developing retina [280, 332, 345, 346]. Because of its characteristic organization during development, candidate molecules that are generated using systems based approaches can be rapidly, albeit cursorily, evaluated based on in situ spatial and temporal expression [201, 347]. Finally, due to its accessibility, candidates can be functionally evaluated in developing retinas using in vivo electroporation to either drive overexpression or knockdown expression of candidate molecules [348].

Networks and network representation of processes have an important role in the implementation of systems based approaches and the analysis of large datasets and complex processes. Demonstrating the ability of these seed networks to effectively focus the generation of hypotheses from high throughput data sets would significantly advance the discoveries that depend upon this type of data. In addition, we have also demonstrated that seed networks are an effective way to do comparative analysis of retinal development and use knowledge of one model system to drive discovery in another [201]. The use of
seed networks to identify conserved networks that act in similar ways (as opposed to conserved genes) will be tremendously useful in the extrapolation of discovery in one model system to another. Thus, development of systems based approaches to investigate cell fate determination in the developing mouse retina will not only lead to important discoveries in the developing retina, but strategies that can be broadly generalized to address many biological questions.

**Authors’ Contribution**

XZ, JMS and MHWG conceived the idea and drafted this manuscript. XZ reviewed the literature and conducted data analysis.
**Figure 2.1.** The retinal cell types in the adult mouse retina. The adult mouse retina is comprised of three cellular layers separated by two synaptic layers. Rod and cone photoreceptors reside in the outer nuclear layer (ONL), and form synaptic contacts in the outer plexiform layer (OPL) with horizontal cells and bipolar cells, both of which reside in the inner nuclear layer (INL). In addition, amacrine cells and the cell bodies of Müller glia are found in the INL. Synaptic contacts between bipolar cells, amacrine cells and ganglion cells are present in the inner plexiform layer (IPL) and ganglion cells reside in the innermost cellular layer, the ganglion cell layer.
Figure 2.2. Time course of cell genesis in the developing mouse retina. Retinal cell types are listed on the Y-axis, developmental time on the X-axis. Birth of the animal is indicated as 0, embryonic development is left of 0, postnatal development to the right. The approximate time course of cell genesis is indicated by the bar adjacent each cell type. This figure is based on the work reported by Young [97].
Figure 2.3. A network of genes essential for Müller glia development. Edges in this graph are based on evidence that Rax promotes notch1 and Hes1 transcription [258]. Notch signaling positively regulates expression of hes1, hes5 [263], sox8 and sox9 [243, 265], and Hes1 suppresses the proneuronal gene Mash1 [264, 349]. Blue edges between genes indicate activation, while red edges indicate repression.
Figure 2.4. A network of genes essential for ganglion cell development. Edges in this graph are based on evidence that Pax6 activates Math5 expression [350] and Math5 suppresses Math3 and NeuroD to promote ganglion cell fate [267, 269]. In addition, Math5 promotes Brn3b and Islet1 expression [251, 272, 273], which in turn positively regulate genes like Eomes [277], Shh [351] and Barhl2 [246]. Brn3b is also activated by Wt1 [279]. Blue edges between genes indicate activation, while red edges indicate repression.
**Figure 2.5.** A network of genes essential for bipolar cell development. The edges in this graph are based on evidence that Otx2 may affect the competence of progenitor cells to adopt a bipolar vs. rod photoreceptor cell fate [288], that Chx10 is hypothesized to work downstream of Otx2 to promote bipolar cell fate [287], and that Chx10, together with Mash1 and Math3, specify bipolar cell fate [254, 285]. Dotted edges indicate indirect or poorly characterized gene relationships.
Figure 2.6. A network of genes essential for amacrine cell development. The edges in this graph are based on evidence that Sox2 activates Pax6 and NeuroD to promote amacrine cell fate [241, 270], that Pax6 and Six3, with the cooperation of Math3 and NeuroD, specify amacrine cell fate [256], and that Foxn4 positively regulates Ptf1a, Math3 and NeuroD expression [226, 227, 253]. Blue edges between genes indicate activation while dotted edges indicate indirect or poorly characterized relationships between genes.
**Figure 2.7.** A network of genes essential for horizontal cell development. The edges in this graph are based on evidence that Foxn4 positively regulates Math3 and Ptf1a expression [253], that coexpression of Pax6 and Math3 promotes horizontal cell fate [256], and that Ptf1a positively regulates Prox1 expression [226, 253, 292] which in turn affects Lim1 expression [291].
**Figure 2.8.** A network of genes essential for rod and cone photoreceptor cell development. The edges in this graph are based on evidence that Otx2 activates Crx [305] while Ataxin-7 [303] and BAF [304] repress Crx transactivation. Crx and Nrl synergically activate the rod specific pigment rhodopsin [299], while Crx promotes expression of M and S cone-specific opsins [300, 301]. Nr2e3, which is activated by Nrl represses expression of both S- and M-opsin [312]. NeuroD is necessary for sustained expression of TRβ2 [298], which inhibits S-opsin and activates M-opsin expression [327]. RXRα [331] promotes cone-specific gene expression while RXRγ [330] represses S-opsin expression. Blue edges between genes indicate activation, while red edges indicate repression.
Table 2.1. Pairwise correlation coefficients between genes of the photoreceptor-specific seed-network. Pearson correlation coefficients were calculated based on the developmental gene expression in rod photoreceptors isolated from retina at ages E16, P2, P6 and P10 [332]. Two genes, BAF and 9-cis RA were not present in the expression dataset and therefore no correlation coefficient could be calculated (NO DATA). The seed-network is shown in Figure 2.8.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Gene</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAF</td>
<td>crx</td>
<td>NO DATA</td>
</tr>
<tr>
<td>ataxin7</td>
<td>crx</td>
<td>0.658655867</td>
</tr>
<tr>
<td>crx</td>
<td>rhodopsin</td>
<td>0.596816525</td>
</tr>
<tr>
<td>crx</td>
<td>s-opsin</td>
<td>-0.287648519</td>
</tr>
<tr>
<td>crx</td>
<td>m-opsin</td>
<td>0.072625073</td>
</tr>
<tr>
<td>nrl</td>
<td>nr2e3</td>
<td>0.995021406</td>
</tr>
<tr>
<td>nrl</td>
<td>rhodopsin</td>
<td>0.910738221</td>
</tr>
<tr>
<td>nr2e3</td>
<td>rhodopsin</td>
<td>0.867910201</td>
</tr>
<tr>
<td>nr2e3</td>
<td>s-opsin</td>
<td>-0.983763422</td>
</tr>
<tr>
<td>neurod</td>
<td>trb2</td>
<td>-0.27275455</td>
</tr>
<tr>
<td>neurod</td>
<td>rhodopsin</td>
<td>0.097272022</td>
</tr>
<tr>
<td>rxrg</td>
<td>s-opsin</td>
<td>0.940893433</td>
</tr>
<tr>
<td>trb2</td>
<td>s-opsin</td>
<td>-0.919529756</td>
</tr>
<tr>
<td>rxra</td>
<td>s-opsin</td>
<td>-0.59339804</td>
</tr>
<tr>
<td>rxra</td>
<td>m-opsin</td>
<td>0.455414641</td>
</tr>
<tr>
<td>9-cis RA</td>
<td>rhodopsin</td>
<td>NO DATA</td>
</tr>
</tbody>
</table>
CHAPTER 3. EnRICH: Extraction and Ranking using Integration and Criteria Heuristics

Modified from a paper to be published in
BMC Systems Biology

Xia Zhang¹, ³, M Heather West Greenlee², ³, ⁴, §, Jeanne M Serb¹, ³

Abstract

Background

High throughput screening technologies enable biologists to generate candidate genes at a rate that, due to time and cost constraints, cannot be studied by experimental approaches in the laboratory. Thus, it has become increasingly important to prioritize candidate genes for experiments. To accomplish this, researchers need to apply selection requirements based on their knowledge, which necessitates qualitative integration of heterogeneous data sources and filtration using multiple criteria. A similar approach can also be applied to putative candidate gene relationships. While automation can assist in this routine and imperative procedure, flexibility of data sources and criteria must not be sacrificed. A tool that can optimize the trade-off between automation and flexibility to simultaneously filter and qualitatively integrate data is needed to prioritize candidate genes and generate composite networks from heterogeneous data sources.

Results

We developed the java application, EnRICH (Extraction and Ranking using Integration and Criteria Heuristics), in order to alleviate this need. Here we present a case study in which we used EnRICH to integrate and filter multiple candidate gene lists in order to identify potential retinal disease genes. As a result of this procedure, a candidate pool of several hundred genes was narrowed down to five candidate genes, of which four are
confirmed retinal disease genes and one is associated with a retinal disease state.

Conclusions

We developed a platform-independent tool that is able to qualitatively integrate multiple heterogeneous datasets and use different selection criteria to filter each of them, provided the datasets are tables that have distinct identifiers (required) and attributes (optional). With the flexibility to specify data sources and filtering criteria, EnRICH automatically prioritizes candidate genes or gene relationships for biologists based on their specific requirements. Here, we also demonstrate that this tool can be effectively and easily used to apply highly specific user-defined criteria and can efficiently identify high quality candidate genes from relatively sparse datasets.

1Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa, USA
2Department of Biomedical Sciences, Iowa State University, Ames, Iowa, USA
3Interdepartmental Genetics Program, Iowa State University, Ames, Iowa, USA
4Bioinformatics and Computational Biology Program, Iowa State University, Ames, Iowa, USA
§Corresponding author: M. Heather West Greenlee, Department of Biomedical Sciences, 2008 Veterinary Medicine, Iowa State University, Ames, IA 50010. Phone: (515) 294-9251. Email: mheather@iastate.edu
**Background**

Hundreds to thousands of candidate genes, or genes of interest, can now be generated from a single experiment utilizing high throughput screening technologies. However, the number of candidate genes that can be experimentally studied in-depth is often constrained by time and cost. Therefore, prioritization of candidate genes is a critical step in the experimental process. Approaches to identify ‘the most promising’ candidates are becoming increasingly more sophisticated. For example, when microarray studies were initially reported, ‘the most promising’ candidates were often the most differentially expressed and could be obtained by a simple ranking of candidates based on fold change. As more data has become available, biologists have begun to look for ways [122, 352-354] to use multiple data sources to increase the accuracy of candidate gene prioritization. Some tools have already been developed to address this need [127-131, 136, 355]. These tools prioritize candidates by their similarity to genes already known to be important for a particular biological process (e.g., genes known to regulate cell cycle in yeast). Multiple data sources including published literature, gene sequence, functional annotation, etc. can be considered when comparing the similarity of candidates to ‘known genes’. These tools [127-131, 136, 355] have made important progress towards the problem of candidate prioritization. However, these tools use data queried from predetermined sources, such as public databases, and include embedded criteria. Thus, these software packages have limited utility.

Biologists, with expertise in a given area, generally already have a list of criteria that could be applied to identify high quality candidates. Likely, for a given set of experiments and resulting datasets, the best candidates may satisfy one set of criteria in one dataset and
a separate set of criteria in another dataset. Currently, there is no tool that allows simultaneous consideration of heterogeneous datasets to identify candidates that satisfy multiple criteria. This problem does not only relate to candidate genes, but also to putative relationships between genes in networks.

Putative gene relationships can be inferred from many heterogeneous sources (e.g., physical interactions, genetic interactions, expression correlation and interactions predicted by computational models). While each of these relationships from a given dataset should be interpreted differently (and subject to very different criteria), the ability to easily hypothesize gene relationships based on their meeting appropriate criteria in multiple datasets is an attractive prospect. This task not only calls for an automated filtering and integration tool, but also demands great flexibility of data sources and the ability to set filtering criteria. Finally, for proper interpretation, visualization of the resulting network must facilitate inspection by 1) retaining the original data sources of each putative relationship and 2) providing a mechanism to easily manage the size of the displayed network. While some tools have been developed to generate composite networks from multiple data sources (e.g., the Cytoscape [356] plugin CABIN [162], GraphWeb [163] and GeneMania [164]), they do not fully address the problems stated above. For example, CABIN supports only one filter for a single source network and thus multiple criteria cannot be applied. GraphWeb [163] does not support filtering by user-defined criteria and interactive network visualization. GeneMania [164] helps to predict the function of a set of input genes by utilizing functional association data to generate a functional relevant network, but does not address integration of user-determined data and filtration with user-defined criteria.
We identified the need for a tool that is able to: 1) filter individual datasets using appropriate criteria and then integrate them to prioritize candidates that meet the criteria in multiple datasets; 2) allow users to define the most appropriate datasets and filtering criteria; and 3) provide an interactive visualization to facilitate the generation of an integrated network with a manageable size and connectedness. To address the open demand of filtering and qualitative integration of heterogeneous datasets, we have developed a stand-alone, portable and flexible java application with its own user-interactive visualization. EnRICH (Extraction and Ranking using Integration and Criteria Heuristics) will assist biologists in prioritization of genes and gene relationships from heterogeneous-source data.

**Implementation**

EnRICH was implemented in Java (SE 6 JDK). EnRICH visualization was written in Processing (http://processing.org/), an open-source programming language to create images, animation and interactions. The separation of non-visual and visual modules of EnRICH lays a flexible foundation for future development and provides the user easy access to both the text and visual output results.

**Design**

The objectives of EnRICH are firstly to provide a tool for integration of multiple or heterogeneous data sets to prioritize candidate molecules that fulfill user-defined criteria, and secondly to make the integration process flexible and simple for biologists who have little programming skill. Our aim-oriented design principles are 1) user-defined data sources and criteria, 2) simplicity which allows straight-forward application of user-
defined criteria to filter user-defined datasets, and 3) platform independence.

The overall architecture of EnRICH is reflected in its workflow-like graphical user interface (GUI) (Figure 3.1). The first component (numbered as step 1 in the GUI) accepts a single file or a directory of files as input data and lists all files that can be selected for analysis. The second component (numbered as step 2 in the GUI) allows the user to display the selected file as a table and edit the table. The third component (numbered as step 3 in the GUI) enables the user to specify filtering criteria for each attribute of the selected file. The fourth component (numbered as step 4 in the GUI) displays all uploaded files for the user to customize an integration pool. It also provides the user running options on whether to apply filters that are already specified in step 3. The fifth component is a dialog window, which appears when the integration run is finished, and gives the user the option to save or visualize the result. For network data, EnRICH has an additional visualization component where the user can do an interactive visual analysis of the integrated network.

Input data

The current version of EnRICH accepts two types of data: list and network. A list is a set of elements that could be genes, proteins, etc., which have their own unique identification code or name. List data can come from a large variety of sources. For example, a list of genes can be differentially expressed genes (DEGs) from the analysis of a microarray experiment, genes identified by genome-wide association mapping, or genes retrieved from a database query. Each list member may have one or more attributes. For example, each gene in a list of DEGs has its own significance value, functional annotation, etc. For EnRICH, list data is represented as a named matrix that is composed of one column
of elements and zero to multiple attribute columns. Attributes can either be value attributes that will be taken as mathematical values or label attributes treated as tags.

A network is a set of nodes that are interconnected by edges representing particular relationships between nodes. Like list data, network data can originate from heterogeneous sources including yeast two-hybrid experiments, computational or statistical inferences, literature summaries or database queries. Although there are several standard languages or formats for network representation, we assume that biologists may not be familiar with those standards. Thus, EnRICH applies a popular node-pair/edge list format as the input format for network data, where an edge is denoted by the pair of nodes it connects. In the matrix format, network edges are represented by two columns of node names. Like list data, network edges may have values and label attributes. Accordingly, a network is a named matrix consisting of two columns of nodes and zero to multiple attribute columns. EnRICH allows blank fields in the attribute column when data are missing.

Running mode

EnRICH runs in two modes: undefined (without filters) and defined (using specific criteria to filter attributes). The undefined mode simply ignores the attributes of networks or lists. Each list or network is considered as a source, and all sources will be merged together. The defined mode simultaneously considers integration of networks or lists as well as user-defined criteria (which filters out elements that do not meet the criteria) over each network or list. For both types of running modes, candidates (edges of a network or elements of a list) are ranked by their reoccurrence across all sources after integration. The filtering process is completely user-defined. Because the filter is totally attribute-based, the
user sets filters most appropriate for their biological question, which may include a combination of filters for each attribute, and even multiple filters for multiple attributes. For example, two of the comparison operators (<, <=, >, >=, ==) applied at the same time can be used to set a cutoff range for value attributes or several tags can be used (with an OR operator between them) when the user wants to select multiple label values (e.g. two annotations) for one attribute. When there are multiple attributes, multiple filters (with an AND operator between them) can be applied simultaneously.

Text Output

EnRICH saves output results as a tab-delimited text file. In the output text file, the user can see what files were integrated, which filters were applied to each file, and the result. For list data, the result is a table, which consists of three columns: the label of an element, its reoccurrence across all lists, and names of source-lists. For network data, the result includes four tables: node statistics, edge statistics, nodes, and edges. The node degree reveals topological importance of the node, so the table of node statistics contains two columns, one column is the node degree (the number of connections a single node has) and the other is the number of nodes that are greater than or equal to (>=) this node degree. For the table of edge statistics, one column is edge reoccurrence (the number of times a single edge is recovered across all datasets) and the other is the number of edges that have an edge reoccurrence that are greater than or equal to (>=) this edge reoccurrence. The table of nodes and the table of edges are quite similar. Each has a column of nodes/edges, their reoccurrence, and source-networks. The only difference between the node and edge tables is that a node is represented by the node label and an edge is denoted
as two node labels. The table of edges is a tab-delimited data table composed of several columns such as node label name, edge reoccurrence and source. Therefore, if desired, the user can directly copy or import them into another network visualization tool such as Cytoscape [166, 356].

**Visualization**

EnRICH enables an interactive visual analysis of the integrated network without depending on a third-party visualization software. EnRICH network visualization consists of two components for user interaction: the integrated network and the plot of network statistics (Figure 3.2). In the integrated network, an undirected edge is drawn as a blue line while a directed edge is drawn as a pink line with a pink arrow to indicate the direction of the interaction (e.g. transcriptional regulation). A blue line with a pink arrow is used to denote merged undirected and directed edges. All edges and nodes can be repositioned, without changing connections, by clicking and then dragging the item on the screen. In addition, the user can click to show or dissipate node labels and edge sources at the node- and edge-specific level, instead of the whole network level. The plot component has two plots: 1) the number of nodes vs. node degree plot and 2) the number of edges vs. edge-reoccurrence plot. The number of nodes and the number of edges are two aspects of the network size, while the node degree reveals topological importance of node. Edge reoccurrence is the number of times the edge is recovered in different data sources, which implies the reliability of an edge. In conjunction, the two plots are used to balance the visualization of network size and quality. All data points in the two plots are clickable to redraw the integrated network at the selected level of node degree or edge reoccurrence.
This interactive plot gives the user an easily visible comparison of node degree, edge reoccurrence and network size, and allows the user to simultaneously visualize the network at corresponding levels. EnRICH also allows the exportation of the network as an image file in TIFF format, which is widely supported.

**Result and Discussion**

**Application of EnRICH**

Retinal disease genes are genes that, when knocked out or mutated, cause retinal degeneration (https://sph.uth.tmc.edu/retnet/disease.htm). The identification of retinal disease genes is a major goal of retinal degenerative disease research, and as part of the effort, there have been a significant number of experiments that describe transcriptional changes during normal retinal development [280, 332, 343, 344, 357-359]. Here, we present a case study in which we use EnRICH to integrate multiple gene lists to identify potential retinal disease genes.

*Nrl* [309, 311, 312, 360] is a retinal disease gene that is associated with the retinal degenerative disease enhanced s-cone syndrome [316]. When *Nrl* is mutated, the resulting phenotype is an abundance of s-cone photoreceptors at the expense of rod photoreceptor differentiation [310, 311], leading to the eventual death of all photoreceptors. During normal development, *Nrl* influences the cone versus the rod cell fate decision by activating rod-specific genes, including the genes *Rho* and *Nr2e3* [361]. *Rho* [362-364] is a rod-specific gene, the mutation of which leads to rod photoreceptor cell death and retinal degeneration. *Nr2e3* [318, 319] is also essential during retinal development, as it promotes the expression of rod-specific genes (including *Rho*) and represses the expression of cone-specific genes in rods. The mutation of *Nr2e3* also causes enhanced s-cone syndrome [314].
Based on the known regulatory relationships between these three disease genes and their importance for normal photoreceptor development, we rationalize that the behavior of these genes would make good criteria to identify additional retinal disease genes.

Using these assumptions, we defined the following criteria to identify retinal disease genes: 1) candidates must be highly co-expressed with Nrl, Nr2e3 and Rho during rod photoreceptor development of wild-type mice; and 2) candidates must be disregulated when Nrl is knocked out (as Nr2e3 and Rho are). With these criteria in mind, we decided to use a microarray dataset [332] (GSE4051), which profiles gene expression in isolated rod photoreceptors at multiple developmental stages (E16, P2, P6, P10, 4-weeks) in both Nrl-knockout and wild-type mice. In these microarrays, we confirmed that Nr2e3 and Rho are highly co-expressed with Nrl in wildtype while in Nrl-mutant no statistical evidence supports that they are still co-expressed.

According to the corresponding workflow (Figure 3.3), we prepared, and subsequently integrated, three types of gene lists which are: Type 1) Genes that are co-expressed with Nrl, Nr2e3 and Rho in developing wild type rod photoreceptors; Type 2) Genes that are co-expressed with Nrl, Nr2e3 and Rho in developing photoreceptors isolated from Nrl-mutant retinas; and Type 3) Differentially expressed genes (DEGs) at each age when comparing gene expression in wild-type rod photoreceptors to Nrl-knockout rods. Each list contained attributes that were used to apply criteria filters (i.e. pairwise correlations for type 1 and 2, age at which expression was up or down regulated for type 3). To carry out the workflow, we first specified filtering criteria for each list. This is a key element of EnRICH, where users can simultaneously query multiple datasets to generate an ‘integrated result’. For this experiment, eight list datasets were integrated. The filtering
criterion for six of the lists was an absolute value of the correlation coefficient greater than 0.9, while the filtering criterion for the two differentially expressed gene lists was the developmental time points P6 and P10 (for criteria on each single list, see Appendix B. Additional file 1). Candidates that satisfied these filtering criteria in eight lists were identified as the highest priority candidates. All the lists were prepared from standard analyses of the dataset GSE 4051 (calculation of co-expression coefficients within a genotype and differentially expressed genes between genotypes).

The execution of our workflow generated five candidate genes (see Appendix B. Additional file 2) from an initial pool of 272 unique differentially expressed genes (see Appendix B. Additional file 3). Based on a literature/database search, four of our five candidate genes (pde6b, gnb1, guca1a and cgna1) are confirmed retinal disease genes [365-374], and the fifth gene (kcne2) has been shown to be up regulated during a neuroinflamatory response in the retinas of diabetic rats [375], making it a reasonable candidate for a disease gene as well. Thus, in our example analysis to identify disease genes, 80% of our candidates are known disease genes, while the remaining candidate has a demonstrated tie to the diseased retina, and is perhaps a high quality candidate. Using a Fisher test we also concluded that retinal disease genes are significantly overrepresented in the genes prioritized by EnRICH, compared with genes not prioritized by EnRICH (see Appendix B. Additional file 3). Our case study demonstrates that a well-conceived data integration and criteria-based filtration, as implemented in EnRICH, can effectively identify a limited number of high quality candidate genes for careful hypothesis-based investigation. Conversely, if the number of candidates returned is too small, slight adjustments in the filtration criteria may be easily made to generate a larger, while still
Conclusions

EnRICH is a free java application which can qualitatively integrate results from large, heterogeneous data sources while simultaneously applying filters to each of them. It allows the user to define data sources, and to integrate them as well as specify multiple sorting criteria specific to each data source. It provides interactive network visualization tool for the user to identify an integrated network with a desirable balance between network size and quality. With EnRICH, biologists have an automated yet flexible integration tool to carry out their data analysis and effectively prioritize candidate genes for further investigation.

Availability and requirements

Project name: EnRICH

Project home page: http://xiazhang.public.iastate.edu/

Operating system(s): platform-independent

Programming language: Java

Other requirements: Java 1.4.2 or higher

License: GNU General Public License

Any restrictions to use by non-academics: NO

Additional files

Additional file 1: Gene lists and their filtration criteria prior to integration. This file includes a supplementary table that displays names and descriptions of gene lists for integration in case study and some further explanation on the sources of gene lists.
Additional file 2: Gene candidates resulting from the EnRICH filtration and prioritization analyses. This file is the supplementary table of the five gene candidates from the prioritization by using EnRICH in case study.

Additional file 3: Description of data processing. This file includes detail description of data pre-processing for case study, and analysis of the significance of case study result.

Authors’ contributions

XZ, JMS and MHWG conceived and designed this software, and drafted this manuscript. XZ coded this software and conducted case study presented in this manuscript. All authors read and approved the final manuscript.
Figure 3.1. EnRICH graphical user interface. There are four major components in the user interface, which are numbered as 1, 2, 3 and 4. Component 1: upload input data; Component 2: browse or edit the selected file; Component 3: specify filtering criteria of the selected file; Component 4: select files to define integration pool.
**Figure 3.2.** EnRICH visualization window. Left: integrated network from synthesizing data. Circles represent nodes, while circle size represents node degree (the number of connections one node has). Lines represent edges of a network and line stroke represents the amount of edge reoccurrence (the number of times one edge is recovered across data sources). Undirected edges are represented by blue lines, while directed edges are pink lines with pink arrows. The merged edge of undirected and directed data is denoted by blue line with a pink arrow. Right: top panel is the statistical plot of node degree vs. the number of nodes. Bottom panel is the statistical plot of edge reoccurrence vs. the number of edges. In the software, all data points in the two plots are clickable to update visualization of the integrated network.
Figure 3.3. Case study workflow. A. Rationale illustrates our current knowledge of Nrl, Nr2e3 and Rho and their behaviors in the dataset we analyzed. B. Workflow displays the steps that we must go through in order to apply our criteria to identify candidate genes. These steps are executed by iRank to obtain candidate genes. B. Workflow is the implementation of the investigation process based on the rationale (A).
CHAPTER 4. Plugging into the tree of life: genome-wide homolog identification between model and non-model organisms

Modified from a manuscript in preparation

Xia Zhang¹, ³, M Heather West Greenlee², ³, ⁴, Jeanne M Serb¹, ³, §

Abstract

In the recent past, genomic or transcriptomic studies were limited to a small number of model organisms that had complete genome sequences available. Rapidly-advancing next generation sequencing technologies have created numerous opportunities to investigate previously understudied organisms. And applications of these technologies have been generating new datasets in an exponential way. To quickly annotate new datasets becomes a bottleneck of data analysis, hurdling further investigations to some extent. This exposes the need to identify putative orthologs between model and non-model organisms in a fast and genome-wide fashion. We propose a fast, across-genome matching method that alleviates the need for screening all homologous gene relationships between two genomes for one-to-one gene matches as putative orthologs. Our results demonstrated this method performed better or as good as other putative orthologs-identifying methods. Our results also showed the relationship between the global homology and the number of putative orthologs between two genomes fits a nonlinear growth model. Furthermore, parameters of this nonlinear growth model are tightly related to evolutionary relationships. Our results suggest exploiting this relationship may also be a prosperous approach to predict evolutionary relationships between known and unknown organisms. Using the tool package (AGEM) that was created to implement the method we proposed, investigators
would be able to easily generate a gene dictionary to translate between any two genomes for further analysis.

1Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa, USA
2Department of Biomedical Sciences, 2008 Veterinary Medicine, Iowa State University, Ames, IA 50010, USA
3Interdepartmental Genetics Program, Iowa State University, Ames, Iowa, USA
4Bioinformatics and Computational Biology Program, Iowa State University, Ames, Iowa, USA
§Corresponding author: Jeanne M Serb, Corresponding author email: serb@iastate.edu

Introduction

Recent technological advances in high-throughput sequencing have allowed for the genomic or transcriptomic profiling of nearly any organism [22, 25, 376, 377]. This technological advance is particularly important for the development of a more comprehensive [holistic] view of biological systems which was previously limited to model organisms but is expanding to include non-traditional models [378, 379]. However, the next challenge/bottleneck for research on these lesser known organisms lies in gene annotation. One approach to this problem is to identify similar or homologous DNA regions between traditional models where gene function has been experimentally demonstrated, with similar, putatively homologous, DNA regions in non-model organisms [56, 380]. This task is not trivial, because each query sequence of a non-traditional model requires 1) the unknown sequence (query) to be aligned against a target genome or database of traditional models and 2) appropriate filtering of alignment results (hits). Once putative homologs have been identified for the unknown sequence, the homolog match can then be leveraged
to infer gene function. In this approach, it is extremely important to effectively exploit everything that is known about the identified homologs. In addition to GO terms [58, 59], gene interactions between the identified homologs are one important informational source to infer function. However, the application of interactions requires projections between two genomes that are based on interacting genes, instead of a set of individual genes. In other words, across-genome gene/protein correspondence that is not one-to-one can dramatically affect the network topology, resulting in noise that may eclipse homologous relationships. Thus, it would be very helpful to effectively obtain cross-organismal matching in which each query protein/gene sequence of one genome has its own unique homologous counterpart in another genome. In addition, this matching must be genome-wide to avoid creating a 'moving target' where the one-to-one correspondence changes with the subset of genes. While non-model systems must utilize sequence homology for gene annotation, model organisms also heavily rely on sequence homology to compare and integrate data cross species to infer evolutionary relationships[183], characterize biological functions [381], and evaluate transferability of research results in clinical trials [382, 383], etc. In situation such as comparison of networks between two organisms[384], matching across genomes to get a one-to-one gene/protein match is highly desired. In all, this post-genomic era exposes the need of model and non-model organism to creatively tap sequence homology to plug into the tree of life.

When it comes to utilizing sequence homology for genome/gene comparisons, there are at least three essential steps. The first step is to align query sequences against an appropriate set of target sequences. Thus, it is necessary that an aligning algorithm or tool is selected based on the research objective. For example, investigating phylogenetic
relationships requires a sequence alignment of multiple homologs across different organisms, while a (reciprocal) blast against closely related genomes is more appropriate when trying to identify functionally similar sequences. **The second** step is to prune the sequence alignment results, in order to reduce the amount of noise introduced into subsequent analyses. For example, blasting a sequence against a genome to identify putative homologs probably will return a set of hits that will include false positives even if the e-value cut-off is extremely low. This could arise when the query and subject share a high sequence identity over a very short sequence length. **The third** step is to apply the pruned alignment hits to downstream analyses, which in our case is requires obtaining a one-to-one gene/protein match.

Assuming the aforementioned first and second steps are carefully implemented, modeling the genome matching as maximum **bipartite matching** (See Box 1.)[385] provides a good solution. Maximum bipartite matching reduces the complicated relationships between two parties to a matching that only contains one-to-one correspondence, while maximizing some sort of measure associated with the matching. Actually, one of the typical maximum bipartite matching algorithms, the maximum weighted bipartite matching, was successfully applied to problems such as the identification of putative orthologs [386] and the evaluation of genome architecture [387]. Here we propose the maximum weighted bipartite matching at fixed cardinality as a new way to approach genome matching. This algorithm matches putatively homologous gene sequences between two genomes by simultaneously maximizing their sequence homology (**maximum weighted**) (See Box 1.) and maximizing the number of matched genes (**maximum cardinality**) (See Box 1.).
Moreover, it provides an insight on how the evolutionary relationship between genomes affects the matching dynamics.

In order to test this proposed method, we executed this algorithm on four genome comparisons: mouse vs rat; mouse vs human; mouse vs chicken; and mouse vs zebrafish. Our results demonstrate that our method performed better or at least as good as relevant work in matching homologous genes/proteins. Further, our results demonstrated the relationship between the genome-wide homology and the number of matched homologous genes/proteins. From the results, we also found an interesting feature that may be used to predict evolutionary relationships of genomes.

During the implementation of our studies, we created a python tool called AGEM. It is freely available online at https://github.com/versaille/AGEM and provides an alternative for investigators whose analyses require or would benefit from one to one genome-wide gene matching between organisms.

**Results**

**Maximum weighted bipartite matching at fixed cardinality**

Bipartite matching models the elements of the two entities that need to be matched as a bipartite graph and then searches a matching on this bipartite graph (See Box 1.). We modeled the relationship between two different classes of objects (e.g., genomes) at the resolution of gene/protein as a bipartite graph (See Box 1.). Under this model, genes/proteins of two genomes would be the vertices $V$ of a bipartite graph $G$, and would be in either $X$ or $Y$ partitions of $V$ depending on which genome they belong to. Putative homologous relationships of genes/proteins between the two genomes would be represented by the edges $E$ between vertices in $X$ and vertices in $Y$, and quantitative scores
that measure the homology extent would be the weights $W$ associated with these edges.

We then applied the algorithm of maximum weighted bipartite matching at fixed-cardinality to this model (For more details on this method, see Technical Supplement). This algorithm searches the maximum weighted bipartite matching (See Box 1.) for any potential cardinality (see Box 1) from minimum to maximum. That is, for a specific number of matched elements (or a specific cardinality), there could be multiple matchings (See Box 1.) and only the matching (see Box 1.) that maximizes the sum of weights of edges is chosen. Maximum cardinality bipartite matching (referred as ‘maximum cardinality’ below, see Box 1.) and maximum weighted bipartite matching (referred as ‘maximum weighted’ below, see Box 1.) are the two major types of bipartite matching. Both are produced during the execution of this algorithm. Maximum cardinality bipartite matching maximizes the number of homologous gene/protein pairs between two genomes. Maximum weighted bipartite matching maximizes the global sequence homology measure between homologous gene/protein pairs.

To estimate the performance of this algorithm, we used the number of matched homologous gene pairs between two genomes as an indicator. Our results are as good as or better than that of relevant work, as shown in Table 4.1. We implemented this algorithm on data from four organismal comparisons: mouse to rat, mouse to human, mouse to chicken, and mouse to zebrafish. We compared the results to relevant work that are shown as ‘EGM’ (Encapsulated gene by gene match) [386] and ‘HomoloGene’ [388]. ‘EGM’ is an ortholog-identifying method that employs maximum weighted bipartite matching to match across the same gene family among different organisms. ‘HomoloGene’ is a NBCI database that detects and stores the orthologous and paralogous gene groups for several eukaryotic
model organisms. We used the number of homologs identified by ‘EGM’ and ‘HomologGene’ as a benchmark to which we compared the performance of our algorithm. For ‘mouse-rat’ and ‘mouse-human’ comparisons, both ‘maximum cardinality’ and ‘maximum weighted’ beat ‘EGM’ and ‘HomoloGene’. For the ‘mouse-zebrafish’ comparison, ‘EGM’ generates a larger number than ‘maximum cardinality’ and ‘maximum weighted’. We attributed this to a less stringent filtering criterion on alignment hits used by ‘EGM’. As we reduced the strictness our filtering criteria (still more strict than what is applied by ‘EGM’), the number of matched homologs increases from 10729 to 12310 for ‘maximum cardinality’ and from 10705 to 12282 for ‘maximum weighted’ (See Table 4.1, ‘reciprocal & stringent’ and ‘reciprocal & less stringent’ columns). For both ‘mouse-chicken’ and ‘mouse-zebrafish’, ‘HomoloGene’ seems to win over all other methods listed in Table 4.1. However, ‘HomoloGene’ builds homologous gene groups for multiple organisms. That is, the number given by ‘HomoloGene’ is the number of genes that can be matched to all other eukaryotic groups that it stores instead of only mouse, meaning some genes out of the 13,149 for chicken and of 14,183 for zebrafish might not even be matched to mouse.

Our results also show how pre-processing steps influence the performance of the algorithm. For all four organismal contrasts, the results of maximum cardinality and of maximum weighted bipartite matching are further categorized by different pre-processing strategies into three sub-columns that are labeled as ‘unilateral & stringent’, ‘reciprocal & stringent’ and ‘reciprocal & less stringent’. ‘Stringent’ and ‘less stringent’ indicate the filtering criteria of sequence alignment results, or the cut-off levels (see Table 4.1 legend for more details) that define the homologous genes (see Method/Sequence alignment and homologs). ‘Reciprocal’ and ‘unilateral’ denote how identified homologous relationships
were parsed for the implementation of this algorithm (see Table 4.1 legend for more details). For each organismal contrast, there is quite a difference between ‘stringent’ and ‘less stringent’, meaning the criteria that define the homologous genes have a significant impact on the final result. This is understandable, since the less stringent the criteria are, the result will include more putatively homologous relationships between two genomes. The reason that ‘reciprocal’ and ‘unilateral’ does not differ much are very likely due to our stringent filtering criteria of sequence alignment results, since with strict filtering criteria, homologous genes are more prone to be reciprocal instead of unilateral.

**Cardinality vs. Global weight**

To display the dynamics between the number of matched genes/proteins and the global sequence homology between two genomes, we graphed cardinality against maximum global weight for every matching generated by the execution of algorithm as previously described. As shown in Figure 4.1A, there is a correlated relationship between cardinality and global weight which fits a nonlinear growth model. That is, as cardinality increases, global weight increases too, but not at a linear rate. Instead, global weight increases at a decreasing rate. This means the global sequence homology goes up at a declining rate as the number of matched homologous genes/proteins rises. As either of them reaches its own maximum, cardinality and global weight converge with each other. The trade-off between maximum cardinality and maximum global weight only happens within a tiny range near to maximum global weight (Figure 4.1B). Results from other organismal contrasts like mouse vs. rat, mouse vs. chicken and mouse vs. zebrafish show similar patterns. This demonstrates that the maximum number of matched homologous genes/proteins to some extent means the maximum global sequence homology.
Knowing how cardinality and global weight relate to each other is important to increase efficiency of matching. The algorithm we described and executed here can simultaneously generate cardinality and global weight. However, provided the relationship of cardinality and global weight prevails, it would more efficient to choose to compute either maximum cardinality bipartite matching or maximum weighted bipartite matching based on the specific need or objective of the project, because the two types of matching require very different computational intensities. The Hopcraft-Karp algorithm for maximum cardinality bipartite matching has a running time of $O(|E|\sqrt{|V|})$ (here $|E|$ is the number of edges and $|V|$ is number of vertices). And the Hungarian algorithm for maximum weighted needs a running time of $O(|V|^2 \log |V| + |V||E|)$. This means it would be much faster to compute maximum cardinality bipartite matching than maximum weighted bipartite matching as the number of vertices $|V|$ increases. Since global weight and cardinality converge with each other as either of them reaches maximum, maximum cardinality, instead of maximum weighted bipartite matching would be an appropriate selection to compute. This would dramatically speed the computation time, especially when the number of genes included in the dataset is very large.

Genome matching and evolutionary relationship

From figure 4.2A, we can see for all organismal contrasts, the general relationship between cardinality and global weight resembles a nonlinear growth pattern. What differs among organism contrasts are the height, length and curviness. Comparing these differences in the context of their evolutionary relationships (Fig. 2B), we found that when two organisms more recently shared a common ancestor, the higher the curve of the
contrast between them is, and the faster the global weight increases as cardinality increases. Considering the height of the curve represents the global weight between two organisms, it means the global weight may indicate the global homology between two organisms. This does not only happen to the absolute height of the curve, but the height of each point on the curve. That is, for each cardinality value (or value on x axis), the global weight (corresponding value on y axis) is higher for more closely related organisms. Translating cardinality into the number of matched homologous gene pairs, this also means that, as the number of homologous gene pairs being matched increases, the global homology increases faster for evolutionarily similar organisms. We regarded the curve as a nonlinear growth model and quantitatively describe it by function below:

\[ Y = c - ae^{-bx} \]

In this function, \( Y \) represents global weight and \( X \) represents cardinality. The function parameter \( b \) implies the speed of growth, \( c \) describes the full growing size and \( a \) determines the starting size. In figure 4.2A, \( a \) equals to 0 for all four curves while \( b \) and \( c \) are bigger for evolutionarily closer organisms. This model suggests its parameters are able to indicate evolutionary relationships or vice versa.

**Discussion**

We modeled the matching between any two genomes in a one-to-one fashion as a bipartite matching problem and then proposed maximum weighted bipartite matching at fixed cardinality as a solution to approach this problem. We also demonstrated how the pre-processing steps such as aligning sequences and pruning sequence alignment results influence the performance of our method. Additionally, we built a python tool AGEM that
can implement the matching algorithm for other investigators to match datasets across genomes that are not included in this study.

The relationship between global weight and cardinality fits a nonlinear growth model. What’s more, the closer two organisms are evolutionarily, the bigger the parameters \( b \) and \( c \) of this nonlinear growth model is. Thus, the parameters for a given matching here can indicate similarity between organisms and may be a good predictor of evolutionary relationships. This feature has the prospect of being applied to predict evolutionary relationships for known and unknown genomes. Such prediction is based on the global effect of the whole genome, instead of only single gene or a couple of genes. Traditionally, to delineate evolutionary relationships between organisms, evolutionary biologists first do multiple sequence alignments on protein sequences of the same conserved gene among different organisms and then construct a phylogenetic tree based on the alignment result. This process inherently limits the data being used from one gene to at most a few genes, limiting perspective. Increasing the data being used from a couple of genes to the whole genome may produce more accurate estimate of evolutionary relationships. However, two important questions highlight the need for further comparisons using more genomes. First, does the aforementioned feature (model parameter vs. evolutionary relationship) hold true for other genomes? Second, even it holds true, what is the resolution between closely related organisms? That is, using this feature we can easily resolve mouse to human from mouse to rat, but can human to chimpanzee from human to rhesus monkey be resolved as easily? Answers to the two questions will help determine at what relatedness this approach is appropriate.
Our method provides a fast solution to identify putative orthologs at the genome-wide scale. However, like EGM, this tool is not appropriate for the definitive identification of orthologs. To label ortholog/paralog in a strict evolutionary sense, a phylogenetic tree that includes out-group should be constructed, this process is effective at the gene-level but not yet the genome-level. So instead of labeling ortholog/paralog, we used stringent cutoff on alignment length and percent identity to remove artifacts and control the homogeneity of homologs being matched. Our method quickly screens homologous relationships between two genomes and generates a genome-wide batch of putative orthologs, providing a search base for more in-depth study.

The utility of our approach can be illustrated in two different examples. First, by generating a genome-wide batch of putative orthologs, evolutionary biologists can gain a global outlook to better target a subset of genes for further analysis and then apply traditional method of constructing phylogenetic tree. Second, comparison of networks between organisms requires genome-wide one-to-one gene matching.

The method we proposed here, together with the tool we created to help implement the method, will help to address the needs of those working on genomes that are not currently within the public database.

**Methods**

**Sequence alignment and homologs**

Protein sequences of mouse (*Mus musculus*), rat (*Rattus norvegicus*), human (*Homo sapiens*), chicken (*Gallus gallus*) and zebrafish (*Danio rerio*) were downloaded from the National Center for Biotechnology Information (NCBI) genome database ([ftp://ftp.ncbi.nih.gov.genomes/](ftp://ftp.ncbi.nih.gov.genomes/)). We used the NCBI standalone BLAST (ncbi-blast-
2.2.27+) to run the sequence alignment

(ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/). We made five blast databases, one for the protein sequences of each organism. Then we did a reciprocal blastp between mouse and mouse, mouse and rat, mouse and human, mouse and chicken, and mouse and zebrafish to identify putative homologous hits. We set the cut-off e-value (=expectation value) at 1e-30 which is particularly strict. To further prune the alignment hits, we filtered them by alignment length, percentage of identical matches, and bit score

[http://www.ncbi.nlm.nih.gov/books/NBK21097/] and used as a homology score below. Two levels of filtering criteria were used to define the homologous relationship. For ‘stringent’, only hits that have an alignment length of 200 amino acids (aa) or greater, a percentage of identical matches of 50 or greater than that, and a bit score that equals to or greater than 200 aa were regarded as homologs and kept for further analyses. For ‘less stringent’, the cutoffs are alignment length being 120 aa, percent identity being 50 and bit score being 150.

**Analysis pipe**

Each pair of homologs has a homology score (=bit score) after alignment. We used this homology score as the weight of an edge between a pair of homologs in two ways. For ‘reciprocal’ way, only those pairs of homologs that have each other in their hits are kept and the weight is the average of their reciprocal homology score (computed by the script combine_reciprocal.py in our python tool AGEM). For ‘unilateral’ way, any pair of homologs that is either reciprocal or unilateral is kept. If being reciprocal, the weight is the average of their reciprocal homology score. If being ‘unilateral’, the weight is the half of the only homolog score (computed by the script combine.py in our python tool AGEM). After that,
we searched the maximum cardinality or the potentially maximized number of matched homologs between the aforementioned four organism contrasts (computed by the script hopcraft_karp.py in our python tool AGEM). Also we searched maximum weighted bipartite matching at fixed cardinality that includes from minimum to maximum (computed by the script weighted_diagnostics.py in our python tool AGEM)

Software availability

The free python tool we created can be found and downloaded at the open source site [http://www.github.com/versaille/AGEM](http://www.github.com/versaille/AGEM).

Authors’ Contribution

XZ, JMS and MHWG conceived the idea and drafted the manuscript. XZ proposed the specific method, coded the python package AGEM and conducted data analysis.
**Figure 4.1.** Matching between mouse and human. Both A and B are based on the matching between mouse and human with the preprocessing strategies being ‘unilateral & stringent’. A plots cardinality against global weight of the maximum weighted bipartite matching at this specific cardinality. B shows some of the specific data points of the plot A, especially the data points that are around the maximum of cardinality and of global weight. Entries in B shaded in grey are the maximum of cardinality 18628 and the maximum of global weight 19482865.
**Figure 4.2.** The matchings of four organism contrasts and their evolutionary relationships. A is a plot that displays the relationship between cardinality and global weight of the maximum weighted bipartite matching at this specific cardinality of four organism contrasts which are mouse and rat (black), mouse and human (blue), mouse and chicken (green) and mouse and zebrafish (red). What were used to plot A are results with the preprocessing strategies being ‘reciporcal & stringent’. B indicates the evolutionary relationships between mouse and the other four organisms that include rat, human, chicken and zebrafish in colors correspond to plot A.
Table 4.1. The identifications of putative orthologs of four methods. This table lists the number of matched homologs that came from four methods/sources for four organism contrasts. The four methods include two typical bipartite matching methods that are maximum cardinality and maximum weighted, one published method referred as EGM that employed maximum bipartite matching in one of its steps, and the NCBI database HomoloGene that stores homologs of several well established eukaryotic genomes. The four contrasts are between organisms that are increasingly evolutionarily divergent from each other. They are mouse and rat, mouse and human, mouse and chicken, and mouse and zebrafish. ‘Stringent’ and ‘less stringent’ indicate different filtering criteria on sequence alignment results (blastp hits), and ‘uni’ (unilateral) and ‘re’ (reciprocal) denote the ways the filtered alignment results are handled before they are passed to the listed methods. ‘Reciprocal’ means only reciprocal homologous relationships were passed to the implementation of this algorithm while ‘unilateral ’ means both reciprocal and unilateral homologous relationships were. ‘Stringent’ means satisfying these cutoffs (evalue<=1e-30, alignment length>=200aa, percent identity>=50%, bit score>=200), ‘less stringent’ (evalue<=1e-30, alignment length>=120aa, percent identity>=50%, bit score>=150).

<table>
<thead>
<tr>
<th>Organism Contrast</th>
<th>The number of matched homologs</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Maximum cardinality</td>
<td>Uni Stringent</td>
<td>Re Stringent</td>
<td>Re Less stringent</td>
<td>Maximum weighted</td>
</tr>
<tr>
<td>Mouse-Rat</td>
<td>18443</td>
<td>18360</td>
<td>20855</td>
<td>18402</td>
<td>18322</td>
</tr>
<tr>
<td>Mouse-Human</td>
<td>18528</td>
<td>18370</td>
<td>20529</td>
<td>18482</td>
<td>18339</td>
</tr>
<tr>
<td>Mouse-Chicken</td>
<td>10773</td>
<td>10604</td>
<td>11898</td>
<td>10757</td>
<td>10593</td>
</tr>
<tr>
<td>Mouse-Zebrafish</td>
<td>11095</td>
<td>10729</td>
<td>12310</td>
<td>11048</td>
<td>10705</td>
</tr>
</tbody>
</table>
Box 1. Bipartite Matching Glossary

**Graph**: a graph is a set of interconnected objects also referred as vertices, being represented by $G(V, E)$ in which $V$ represent the set of vertices and $E$ denotes the set of edges that describe the interconnected relationships among vertices. A graph is sometimes called a network and vertices are sometimes referred as nodes.

**Bipartite graph**: a bipartite graph $G(V(X, Y), E)$ is a graph has its vertices $V$ being divided into two disjoint sets $X$ and $Y$, and edges $E$ being drawn between vertices in $X$ and in $Y$.

**Matching**: given a graph $G(V, E)$, a matching $M$ is a subset of edges of this graph that do not share common nodes.

**Bipartite matching**: given a bipartite graph $G(V(X, Y), E)$, a bipartite matching $M(M \subseteq E)$ is a subset of $E$ that contains only non-adjacent edges (or, no two edges share a vertex).

**Cardinality**: given a set, cardinality is a measure that tells the number of elements in this set. For example, a set has 10 elements, so its cardinality is 10.

**Maximum cardinality bipartite matching**: given a bipartite graph $G(V(X, Y), E)$, maximum cardinality bipartite matching is the matching $M(M \subseteq E)$ that maximizes cardinality, or the number of matched pairs, between the two partitions $X$ and $Y$.

**Maximum weighted bipartite matching**: given a bipartite graph $G(V(X, Y), E)$ and the weights $W(E)$ associated with the edges $E$, maximum weighted bipartite matching is the matching $M(M \subseteq E)$ that maximizes $W(M)$, the sum of weights of edges in $M$.
CHAPTER 5

SUMMARY

This thesis heuristically explored ways to decipher heterogeneous big data for useful clues that help to answer questions in an integrative way. Chapter 1 gives an overview of the big background, a brief description of rationale and specific objectives and a literature review that is essential to the specific objectives.

Chapter 2 describes review work on prior knowledge to show developing retina is an excellent model to study cellular fate determination and differentiation in the context of a multicellular tissue, shows network models we constructed to summarize known gene interactions that underlie determination of each retinal cell type and displays how we leveraged high-throughput data to rediscover the network supported by experimental evidence. These networks provide a rational segue to systems biology approaches necessary to understand the many events leading to appropriate cellular determination and differentiation in the developing retina and other complex tissues. The ability of these networks to effectively focus the generation of hypotheses from high-throughput data sets would significantly advance the discoveries that depend on this type of data. It will be tremendously useful in the extrapolation of discovery in one model system to another if the network models are used to identify conserved networks act in similar ways. So investigation in more organisms is worth future effort.

Chapter 3 describes tool development work on EnRICH that is able to qualitatively integrate multiple heterogeneous datasets and use different selections to filter each of them, provided the datasets are tables that have distinct identifiers
and attributes. A case study is presented to show how EnRICH is used to integrate and filter multiple candidate gene lists to identify potential retinal disease genes. As a result of the case study, a candidate pool of several hundred genes was narrowed down to five candidate genes, of which four are confirmed retinal disease genes.

With EnRICH, biologists have an automated yet flexible integration tool to construct and analyze composite network and effectively prioritize candidate genes for further investigation. For the applications of EnRICH in network data, I name a few cases here. The first case is to extract interactions of interest out of huge amount of such information by filtering their attributes. For example, a mouse protein-protein network can be really huge for a biologist who is interested in mouse retinal development to look for valuable information. Fortunately proteins in such a network are more likely to be functionally annotated. So the biologist can relay the function information of proteins to their interactions and then use EnRICH to get interactions of interest (e.g. only interactions that are composed by proteins who are known to work for retinal development are extracted). The second case is to compare networks. For example, with a neural development network and a retinal development network, the biologist may want to know what are conserved or differentiated between the two networks. By using EnRICH to integrate the two networks, the biologist would easily discern the conserved from the differentiated, or vice versa. The third case is to integrate across interaction data from different platforms. For example, to find out how genes may work with each other, the biologist may wish to leverage heterogeneous information such as protein-protein
interactions, genetic interactions, computationally predicted interactions for a more reliable hypothesis. With EnRICH, this can be easily done within minutes.

Even though Current version of EnRICH can deal with heterogeneous intra-organism information. The results of prioritization and integration may get improved with the addition of inter-organism information. This requires the addition of the function that is able to convert/map genes between organisms (research on conversion method led to Chapter 4).

Chapter 4 describes a method we proposed to address the need of across-genome matching in a one gene to one gene fashion. In this method, matching between any two genomes is modeled as a maximum bipartite matching problem and the algorithm of maximum weighted bipartite matching of specific cardinality is used to solve this problem. To implement this proposed method we created a python tool package called AGEM. The result from implementing this method on four pairs of organisms indicates this method is better than other methods in terms of the number of putative orthologs.

Our method and tool AGEM enables biologists to generate dictionary to translate genes in a one-to-one fashion between any two genomes and thus has a broad use in integrating data across organisms, especially in leveraging model organisms to annotate non-model organisms. For example, studies on how genes interact with each other in the development of non-model organism cotton are very desirable due to the social-economic value of cotton. However, the lack of interaction data of cotton hinders such efforts. With AGEM, the biologist can map
sequences of cotton to *Arabidopsis* genes and thus leverage the interaction data of *Arabidopsis* to hypothesize gene interactions of cotton.

By studying on the performance of this method, the relationship between global homology and the number of putative orthologs was discovered to follow a nonlinear growth model. And parameters of the nonlinear growth model are relate to closeness of evolutionary relationships and thus could be used to predict unknown evolutionary relationships among organisms. Implementation of this method on more organisms can help determine at what relatedness the feature that is promising to predict evolutionary relationships is appropriate.

Chapter 5 summarizes the results of previous chapters and points out where the future research should focus on.
REFERENCES

1. **Human Genome Project**


3. Golden F, Lemonick MD: The race is over. The great genome quest is officially a tie, thanks to a round of pizza diplomacy. Yet lead researcher Craig Venter still draws few cheers from his colleagues. *Time* 2000, **156**(1):18-23.


12. Maskos U, Southern EM: Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization


16. **Affymetrix** [www.affymetrix.com](http://www.affymetrix.com)


64. Garber M, Grabherr MG, Guttman M, Trapnell C: **Computational methods for transcriptome annotation and quantification using RNA-seq.** *Nat Methods, 8*(6):469-477.


82. **Gene Expression Omnibus** [http://www.ncbi.nlm.nih.gov/geo/]


134. Tanabe M, Kanehisa M: **Using the KEGG database resource.** *Curr Protoc Bioinformatics, Chapter 1*:Unit1 12.


198. Jensen RA: **Orthologs and paralogs - we need to get it right.** *Genome Biol* 2001, **2**(8):INTERACTIONS1002.


200. Hecker LA, Alcon TC, Honavar VG, Greenlee MH: **Using a seed-network to query multiple large-scale gene expression datasets from the developing retina in order to identify and prioritize experimental targets.** *Bioinform Biol Insights* 2008, **2**:401-412.


220. Parameswaran S, Balasubramanian S, Babai N, Qiu F, Eudy JD, Thoreson WB, Ahmad I: Induced pluripotent stem cells generate both retinal ganglion cells and photoreceptors: therapeutic implications in degenerative changes in glaucoma and age-related macular degeneration. Stem Cells, 28(4):695-703.


290. Oliver G, Mailhos A, Wehr R, Copeland NG, Jenkins NA, Gruss P: **Six3, a murine homologue of the sine oculis gene, demarcates the most anterior border of the developing neural plate and is expressed during eye development.** *Development* 1995, **121**(12):4045-4055.


312. Oh EC, Cheng H, Hao H, Jia L, Khan NW, Swaroop A: *Rod differentiation factor NRL activates the expression of nuclear receptor NR2E3 to*


376. Gibbons JG, Janson EM, Hittinger CT, Johnston M, Abbot P, Rokas A: Benchmarking next-generation transcriptome sequencing for


398. R [http://www.r-project.org/]


400. **limma**  

401. **qvalue**  

402. **doBy** [http://cran.r-project.org/web/packages/doBy/index.html]


**APPENDIX A**

**SUPPLEMENTARY INFORMATION (CHAPTER 2)**

**Supplementary Table 1.** A summary of genes regarded as important for retinal cell determination and differentiation in this study. Individual citations experimentally describe the developmental effect of a gene on the specification or differentiation of a specific cell type.

<table>
<thead>
<tr>
<th>GENE</th>
<th>RETINAL CELL TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Muller</td>
</tr>
<tr>
<td>Bm3b</td>
<td></td>
</tr>
<tr>
<td>Bm3a</td>
<td></td>
</tr>
<tr>
<td>Bm3c</td>
<td></td>
</tr>
<tr>
<td>Barhl2</td>
<td>[246]</td>
</tr>
<tr>
<td>Chx10</td>
<td></td>
</tr>
<tr>
<td>Crx</td>
<td></td>
</tr>
<tr>
<td>Dlx1</td>
<td>[394, 395]</td>
</tr>
<tr>
<td>Dlx2</td>
<td>[394, 395]</td>
</tr>
<tr>
<td>Irx5</td>
<td>[247]</td>
</tr>
<tr>
<td>Lim1</td>
<td></td>
</tr>
<tr>
<td>Islet1</td>
<td>[272, 273]</td>
</tr>
<tr>
<td>Otx2</td>
<td>[288, 289]</td>
</tr>
<tr>
<td>Pax6</td>
<td>[270, 271]</td>
</tr>
<tr>
<td>Prox1</td>
<td></td>
</tr>
<tr>
<td>Rax</td>
<td></td>
</tr>
<tr>
<td>RPF1</td>
<td>[396]</td>
</tr>
<tr>
<td>Sox2</td>
<td></td>
</tr>
<tr>
<td>Sox8</td>
<td>[243]</td>
</tr>
<tr>
<td>Sox9</td>
<td>[243, 265]</td>
</tr>
<tr>
<td>Bhlhb4</td>
<td></td>
</tr>
<tr>
<td>Bhlhb5</td>
<td></td>
</tr>
<tr>
<td>Hes1</td>
<td>[249, 258, 263]</td>
</tr>
<tr>
<td>Hes5</td>
<td>[248, 263]</td>
</tr>
<tr>
<td>Math5</td>
<td>[251, 266-269, 271]</td>
</tr>
<tr>
<td>Mash1</td>
<td></td>
</tr>
<tr>
<td>Gene</td>
<td>Reference(s)</td>
</tr>
<tr>
<td>--------</td>
<td>--------------</td>
</tr>
<tr>
<td>Math3</td>
<td>254, 256</td>
</tr>
<tr>
<td>NeuroD</td>
<td>256</td>
</tr>
<tr>
<td>Ptf1a</td>
<td>226, 227</td>
</tr>
<tr>
<td>Ataxin7</td>
<td>302, 303</td>
</tr>
<tr>
<td>BAF</td>
<td>304</td>
</tr>
<tr>
<td>Eomes</td>
<td>277</td>
</tr>
<tr>
<td>Nrl</td>
<td>311</td>
</tr>
<tr>
<td>Nr2e3</td>
<td>315, 318, 319</td>
</tr>
<tr>
<td>RXRα</td>
<td>331</td>
</tr>
<tr>
<td>RXRγ</td>
<td>330</td>
</tr>
<tr>
<td>Shh</td>
<td>351, 397</td>
</tr>
<tr>
<td>TRβ2</td>
<td>298, 326</td>
</tr>
<tr>
<td>9-cis RA</td>
<td>320, 321</td>
</tr>
</tbody>
</table>
**APPENDIX B**

**ADDITIONAL FILES (CHAPTER 3)**

**Additional file 1:** Gene lists and their filtration criteria prior to integration. This file includes a supplementary table that displays names and descriptions of gene lists for integration in case study and some further explanation on the sources of gene lists.

**Supplementary Table 1. Gene Lists for integration and criteria on them for filtration**

<table>
<thead>
<tr>
<th>Gene List</th>
<th>Gene List Description</th>
<th>Attribute</th>
<th>Filter on attribute</th>
<th>Potential candidate genes are expected to be</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wt_Nrl.txt</td>
<td>Genes coexpressed with Nrl across all five time points (E16, P2, P6, P10, 4-WEEK) in wild type</td>
<td>Spearman correlation coefficient</td>
<td>&gt;</td>
<td>0.9</td>
</tr>
<tr>
<td>Wt_Nr2e3.txt</td>
<td>Genes coexpressed with Nr2e3 across all five time points (E16, P2, P6, P10, 4-WEEK) in wild type</td>
<td>Spearman correlation coefficient</td>
<td>&gt;</td>
<td>0.9</td>
</tr>
<tr>
<td>Wt_Rho.txt</td>
<td>Genes coexpressed with Rho across all five time points (E16, P2, P6, P10, 4-WEEK) in Nrl-mutant</td>
<td>Spearman correlation coefficient</td>
<td>&gt;</td>
<td>0.9</td>
</tr>
<tr>
<td>Nrl_Nrl.txt</td>
<td>Genes coexpressed with Nrl across all five time points (E16, P2, P6, P10, 4-WEEK) in Nrl-mutant</td>
<td>Spearman correlation coefficient</td>
<td>&gt;</td>
<td>0.9</td>
</tr>
<tr>
<td>Nrl_Nr2e3.txt</td>
<td>Genes coexpressed with Nr2e3 across all five time points (E16, P2, P6, P10, 4-WEEK) in Nrl-mutant</td>
<td>Spearman correlation coefficient</td>
<td>&gt;</td>
<td>0.9</td>
</tr>
<tr>
<td>Nrl_Rho.txt</td>
<td>Genes coexpressed with Rho across all five time points (E16, P2, P6, P10, 4-WEEK) in Nrl-mutant</td>
<td>Spearman correlation coefficient</td>
<td>&gt;</td>
<td>0.9</td>
</tr>
<tr>
<td>Deg_Up.txt</td>
<td>Up-regulated genes (Nrl-mutant VS. Wild type)</td>
<td>Time points at which genes are</td>
<td>P6, P10</td>
<td>Present</td>
</tr>
</tbody>
</table>
Note: The GEO dataset GSE4051 [332] is used to prepare differentially expressed gene lists and coexpressed gene lists. Differentially expressed genes between wild type and Nrl-mutant at four developmental time points (E16, P2, P6 and P10) are obtained with the cut-off q-value controlled at 0.05 (For details, see Additional file 3). Coexpressed genes of Nrl/Nr2e3/Rho are obtained by calculating Spearman correlation coefficients between Nrl/Nr2e3/Rho and other genes across the five time points (developmental stages) from wild type and Nrl-mutant. R [398] and R packages (affy [399], limma [400], qvalue [401] and doBy[402]) are used to do the computation. All lists in the above table can be downloaded from http://xiazhang.public.iastate.edu/demo.html
**Additional file 2:** Gene candidates resulting from the EnRICH filtration and prioritization analyses. This file is the supplementary table of the five gene candidates from the prioritization by using EnRICH in case study.

<table>
<thead>
<tr>
<th>Gene symbol</th>
<th>Full name</th>
<th>Previously known retinal disease genes</th>
<th>NCBI Links</th>
<th>Wikigenes Links</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>pde6b</strong></td>
<td>phosphodiesterase 6B, cGMP, rod receptor, beta polypeptide</td>
<td>yes</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td><strong>cnga1</strong></td>
<td>cyclic nucleotide gated channel alpha 1</td>
<td>yes</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td><strong>gnb1</strong></td>
<td>guanine nucleotide binding protein (G protein), beta 1</td>
<td>yes</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td><strong>kcne2</strong></td>
<td>potassium voltage-gated channel, Isk-related subfamily, gene 2</td>
<td>no</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td><strong>guca1a</strong></td>
<td>guanylate cyclase activator 1a (retina)</td>
<td>yes</td>
<td>5</td>
<td>10</td>
</tr>
</tbody>
</table>

6: http://www.wikigenes.org/e/gene/e/18587.html  
7: http://www.wikigenes.org/e/gene/e/12788.html  
8: http://www.wikigenes.org/e/gene/e/14688.html  
9: http://www.wikigenes.org/e/gene/e/246133.html
http://www.wikigenes.org/e/gene/e/14913.html

Note:
Red: genes coexpressed with nrl, nr2e3, rho in wildtype and are not coexpressed with these genes in nrl-mutant, but are down regulated at p6, p10 (like rho and nr2e3) in nrl-mutant
Green: genes coexpressed with nrl, nr2e3, rho in wildtype and are not coexpressed with these genes in nrl-mutant, but are up regulated at p6, p10 (like rho and nr2e3) in nrl-mutant
**Additional file 3:** Description of data processing. This file includes detail description of data pre-processing for case study, and analysis of the significance of case study result.

**The Differentially Expressed Genes (DEGs) of GSE 4051**

GSE 4051 [332] is a dataset of Affymetrix arrays. We downloaded the dataset GSE4051_RAW.tar from NCBI database Gene Expression Omnibus (GEO) [82]. We used R Packages to analyze this dataset. The package ‘affy’ [399] is used to read and pre-process the raw data. The package ‘limma’ [400] is used to fit the linear model (see below), and compute empirical Bayes moderate t-statistics. In order to control multiple testing FDR (False Discovery Rate) at 5%, we also used the package ‘qvalue’ [401] to compute q-values and set the cut-off as 0.05.

**Linear model**

GSE 4051 [332] profiles gene expression in isolated rod photoreceptors at five developmental stages (E16, P2, P6, P10 and 4-weeks) from wild type and Nrl-knockout. We are interested in what genes respond differently between wild type and Nrl-knockout at each of the four developmental stages (E16, P2, P6 and P10). So our linear model takes the mathematical form:

\[ Y_{ijk} = \mu + g_i + t_j + gt_{ij} + s_k + e_{ijk} \]

Where the responsible variable \( Y_{ijk} \) represents signal intensity for genotype \( i \), time point \( j \) and biological replication \( k \). The parameter \( \mu \) denotes the average effect, \( g_i \) is the fixed effect of genotype, \( t_j \) is the fixed effect of developmental
stage, $g t_{ij}$ indicates the interaction between genotype $i$ and developmental stage $j$, $s_k$ represents the fixed effect of replicate $k$, and $e_{ijk}$ is the random error.

**The Significance of the Case Study Results**

We found 272 unique differentially expressed genes from the analysis of GSE 4051 [332]. The exclusion of three genes ($Nrl$, $Nr2e3$ and $Rho$ are bait genes) results in a candidate pool of 269 genes. By using EnRICH to prioritize this candidate pool, we obtained five genes with the highest priority, and four of them are confirmed retinal disease genes. According to our search for Retinal Disease Genes (see below), only one previously known retinal disease gene ($Rbp3$) was not re-discovered by EnRICH. The Fisher’s exact test (see below) also indicates retinal disease genes are significantly enriched in the fived prioritized candidate genes.

**Search for Retinal Disease Genes**

We used NCBI database Genetics Home Reference [403] to search for the list of documented genes that contribute to the genetic conditions of Retinitis Pigmentosa (Retinitis pigmentosa is a retinal disease caused by abnormalities of the photoreceptor rods and cones or the retinal pigment epithelium). Genes prioritized by EnRICH were compared to this list to determine if they were known retinal disease genes.

**Fisher’s exact test**

We performed a Fisher’s exact test to check for enrichment of retinal disease
genes in the high priority genes identified using EnRICH.

<table>
<thead>
<tr>
<th>Identified as high priority by EnRICH?</th>
<th>Known as retinal disease genes?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes 4_d 1_e 5_b</td>
</tr>
<tr>
<td>No</td>
<td>1_f 263_g 264_c</td>
</tr>
<tr>
<td></td>
<td>5_h 264_k 269_a</td>
</tr>
</tbody>
</table>

a The total number of genes in the candidate pool
b The total number of genes selected by EnRICH
c The total number of genes not selected by EnRICH
d The number of genes that are selected by EnRICH as well as known as retinal disease genes
e The number of genes that are selected by EnRICH yet not known as retinal disease genes
f The number of genes that are not selected by EnRICH yet known as retinal disease genes
g The number of genes that are neither selected by EnRICH nor known as retinal disease genes
h The total number of genes that are known as retinal genes in the candidate pool
k The total number of genes that are not known as retinal genes in the candidate pool

Test result computed by R:
p-value = 1.168e-07
Alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
3.806447e+01 4.503600e+15

The p-value for this test is extremely small, from which we can reject the null hypothesis that retinal disease genes are equally represented between genes selected by EnRICH and genes not selected by EnRICH. Alternatively, the test result favors the hypothesis that retinal disease genes are overrepresented in genes selected by EnRICH.
APPENDIX C
TECHNICAL SUPPLEMENT (CHAPTER 4)

Part 1: Definition of the glossary

**Graph**: a graph is a set of interconnected objects also referred as vertices, being represented by \( G (V, E) \) in which \( V \) represents the set of vertices and \( E \) denotes the set of edges that describe the interconnected relationships among vertices. A graph is sometimes called a network and vertices are sometimes referred as nodes.

**Bipartite graph**: as shown in Figure 1, a bipartite graph \( G (V (X, Y), E) \) is a graph has its vertices \( V \) being divided into two disjoint sets \( X \) and \( Y \), and edges \( E \) being drawn between vertices in \( X \) and in \( Y \).

**Matching**: given a graph \( G (V, E) \), a matching \( M \) is a subset of edges of this graph that do not share common nodes.

**Bipartite matching**: given a bipartite graph \( G (V (X, Y), E) \), a bipartite matching \( M (M \subseteq E) \) is a subset of \( E \) that contains only non-adjacent edges (or, no two edges share a vertex)

**Cardinality**: given a set, cardinality is a measure that tells the number of elements in this set. For example, a set has 10 elements, so its cardinality is 10.

**Maximum cardinality bipartite matching**: as shown in Figure 2, given a bipartite graph \( G (V (X, Y), E) \), maximum cardinality bipartite matching is the matching \( M (M \subseteq E) \) that maximizes cardinality, or the number of matched pairs, between the two partitions \( X \) and \( Y \).

**Maximum weighted bipartite matching**: as shown in Figure 3, given a bipartite graph \( G (V (X, Y), E) \) and the weights \( W (E) \) associated with the edges \( E \), maximum weighted bipartite matching is the matching \( M (M \subseteq E) \) that maximizes \( W (M) \), the sum of weights of edges in \( M \).

Part 2: Description of the method

**Maximum cardinality bipartite matching**: a bipartite matching in which cardinality reaches maximum.

**Maximum weighted bipartite matching**: a bipartite matching in which global weight reaches maximum.
The two types of bipartite matching differ in the quantitative measure from which they seek maximum. One is cardinality (the number of matched pairs) and the other is global weight (the sum of weights that are associated with matched pairs). In some cases, there might be a trade-off between cardinality and global weight (Figure 4). So to have a legitimate/appropriate balance between cardinality and global weight, it is necessary to associate them together in a way that the maximum of either of them can be considered along the scale of the other. In our method, we considered global weight along the scale of cardinality.

For a given a bipartite graph $G(V(X,Y), E)$, its bipartite matching as $M_G (M_G \subseteq E)$ is a subset of edges $E$ that contains only non-adjacent edges. The number of edges in $M_G$ is called its cardinality $k_M \in [k_{min}, k_{max}]$, where $k_{min}=0$ and $k_{max}$ =maximum cardinality that $M_G$ can have. The sum of weights that are associated with all edges in $M_G$ is its global weight $W_{ms}$. There are two things worth noting here:

1. It is possible there are multiple $M_G$ with the same $k_M$. For example, for bipartite graph in figure 5, there are two $M_G$ that have the same cardinality 3.

2. It is possible that there are multiple $M_G$ with the same $k_M$ and different $W_{ms}$.

For example, the two bipartite matchings in figure 5 have different $W_{ms}$.

In other words, for a given bipartite graph and for any cardinality $k_M \in [k_{min}, k_{max}]$, there could be multiple bipartite matchings $M_G$ that could have different global weights $W_{ms}$. Based on this, a good way to consider
global weight along the scale of cardinality would be to rate the maximum of global weight against the scale of cardinality. That is, for each \( K_m \), to consider only the \( M_g \) that has the largest global weight. This could be dissected into two parts: the first is to look for the \( M_g \) that has the largest global weight for each \( k_M \). The second is to \( k_M \) has to be increased from \( k_{\min} \) to \( k_{\max} \) by one each time. The legitimacy of incremental computation has been proven in previous works on bipartite matching, for more technical details in computer science, we refer the reader to the cited works [404-407].

Part 3: Execution of the algorithm

1. Load bipartite graph \( G(V(X,Y),E) \), convert \( G \) to \( G' \) by :
   1.1 Making undirected edges to directed edges (\( e_{x-y} \) to \( e_{x\rightarrow y} \)), and for each edge \( e \in E \), \( w_{adj} = -w_m + \max(w_m) \).
   1.2 Source vertex \( s \) is connected to vertex \( x \in X \), and target vertex \( t \) is connected to vertex \( y \in Y \), and make \( w_{s\rightarrow x} = 0 \) and \( w_{y\rightarrow t} = 0 \).

2. Set \( k_M = 0 \), \( M_{\max} = 0 \), \( D = \{ \text{key: } k_M, \text{value: } M_{\max} \} \)

3. While (there exists augmenting path with respect to \( M_{\max} \) ) do:
   1. Search for an augmenting path \( p \) with minimal weight using Dijkstra’s algorithm (compute the shortest path from \( s \) to \( t \) in \( G' \), the corresponding path in \( G \) is augmenting path \( p \) with minimal weight).
   2) Invert \( p \) and increases \( M_{\max} \) by one edge:
3) Recalibrate all directed edges (assuming from vertex \( u \) to vertex \( v \)) in \( G' \) that are not included in \( p \) by function \( w_{u\rightarrow v} = w_{u\rightarrow v} + d_u - d_v \) (this function makes \( w_{u\rightarrow v} \geq 0 \) be satisfied so Dijkstra’s algorithm can be used).

4) \( D[k_M] = M_{w_{max}} \)

5) \( k_M = k_M + 1 \)

4. Output \( D \)
Figure 1. Bipartite Graph. The blue nodes represent one party while the orange nodes the other party.
Figure 2. Maximum cardinality bipartite matching. The nodes in blue and orange represent nodes in a bipartite graph and lines drawn between them are edges in this bipartite graph. Numbers on the top of lines are weights associated with these edges. Strokes in red denote a maximum cardinality bipartite matching.
Figure 3. Maximum weighted bipartite matching. The nodes in blue and orange represent nodes in a bipartite graph and lines drawn between them are edges in this bipartite graph. Numbers on the top of lines are weights associated with these edges. Strokes in red denote a maximum weighted bipartite matching.
Figure 4. The trade-off between cardinality and weight. The nodes in blue and orange represent nodes in a bipartite graph and lines drawn between them are edges in this bipartite graph. Numbers on the top of lines are weights associated with these edges. Strokes in red denote the bipartite matching. Red strokes in A represent a bipartite matching that has a cardinality of 4 and a global weight of 55. Red strokes in B represent a bipartite matching that has a cardinality of 3 and a global weight of 150.
Figure 5. Bipartite matchings with the same cardinality. The nodes in blue and orange represent nodes in a bipartite graph and lines drawn between them are edges in this bipartite graph. Numbers on the top of lines are weights associated with these edges. Strokes in red denote the bipartite matching. Red strokes in A represent a bipartite matching that has a cardinality of 3, and red strokes in B represent another bipartite matching that also has a cardinality of 3. The matching in A has a global weight different from that of the matching in B.