VITA

Education

Ph.D. in Biology Aug, 2007 to Aug, 2012

Department of Biology, Indiana State University, Terre Haute, IN 47807, USA

M.Sc. in Plant Biotechnology Dec, 2003 to Mar, 2006

Institute of Agri-Biotechnology, University of Agricultural Sciences, Dharwar, India

B.Sc. in Agriculture Aug, 2009 to Jun, 2004

Agriculture College, Mandya, University of Agricultural Sciences, Bangalore, India

Publications

Seetharam, A., Bai, Y. and Stuart G. W., (2010), A survey of well conserved families of C2H2 zinc-finger genes in Daphnia. *BMC Genomics* 11:276.

Reddy P. S., Fakrudin B., Rajkumar, Punnuri S. M., Arun S. S., Kuruvinashetti M. S., Das I. K. and Seetharama N., (2008), Molecular mapping of genomic regions harboring QTLs for stalk rot resistance in sorghum. *Euphytica*, 159:191-198.

Presentations (selected)

Seetharam, A. and Stuart G. W., Multi-locus *Drosophila* Phylogenomics using TypeIIB enzyme target sites. *Talk presented at Indiana Academy of Sciences annual meeting, Purdue University, West Lafayette, Indiana*. *March 10, 2012*.

Seetharam, A., Keller, V. and Stuart G. W., Study on Conservation and Distribution of C2H2 Zinc-finger Genes in Eukaryotes. *Poster presented at 6th International Symposium on Bioinformatics Research and Applications (ISBRA) University of Connecticut, Storrs, Connecticut. May 23-26, 2010.*

PHYLOGENOMICS: MOLECULAR EVOLUTION IN THE GENOMICS ERA

A dissertation

Presented to

The College of Graduate and Professional Studies

Department of Biology

Indiana State University

Terre Haute, Indiana

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

Arun Somwarpet Seetharam

August 2012

Keywords: Phylogenomics, Singular value decomposition, C2H2 zinc fingers

COMMITTEE MEMBERS

Committee Chair: Gary W. Stuart, Ph.D.

Professor of Biology

Indiana State University

Committee Member: Jennifer K. Inlow, Ph.D.

Associate Professor of Chemistry and Physics

Indiana State University

Committee Member: Allan R. Albig, Ph.D.

Assistant Professor of Biology

Indiana State University

Committee Member: Swapan K. Ghosh, Ph.D.

Professor of Biology

Indiana State University

Committee Member: James P. Hughes, Ph.D.

Professor of Biology

Indiana State University

ABSTRACT

Evolutionary studies in recent years have been transformed by the development of new, powerful techniques for investigating many mechanisms and events of molecular evolution. Large collections of many different complete genomes now available in the public domain offer great advantages to genomic scale evolutionary studies. Phylogenomics, a term often used to describe the use of genomic scale data to infer species phylogeny or to predict protein function through evolutionary history, is greatly benefitted by the revolutionary progress in DNA sequencing technology. In the present study we developed and utilized various phylogenomic methods on large genome-scale data.

In the first study, we applied Singular Value Decomposition (SVD) analysis to re-examine current evolutionary relationships for 12 *Drosophila* species using the predicted proteins from whole genomes. An SVD analysis on unfiltered whole genomes (193,622 predicted proteins) produced the currently accepted *Drosophila* phylogeny at higher dimensions, except for the generally accepted, but difficult to discern, sister relationship between *D. erecta* and *D. yakuba*. Also, in accordance with previous studies, many sequences appear to support alternative phylogenies. In this case, we observed grouping of *D. erecta* with *D. sechellia* when approximately 55% to 95% of the proteins were removed using a filter based on projection values or by reducing resolution by using fewer dimensions.

In the second study, we simulated restriction enzyme digestions on 21 sequenced genomes of various *Drosophila* species. Using the fragments generated by simulated digestion

from the predicted targets of 16 Type IIB restriction enzymes, we sampled a large and effectively arbitrary selection of loci from these genomes. The resulting fragments were then used to compare organisms and to calculate the distance between genomes in pair-wise combination by counting the number of shared fragments between the two genomes. Phylogenetic trees were then generated for each enzyme using this distance measure, and the consensus was calculated. The consensus tree obtained agrees well with the currently accepted tree for these *Drosophila* species. We conclude that multi-locus sub-genomic representation combined with next generation sequencing, especially for individuals and species without previous genome characterization, can improve studies of comparative genomics and the building of accurate phylogenetic trees.

The third study utilized the relatively new *Daphnia* genome in an attempt to identify 40 orthologous groups of C2H2 Zinc-finger proteins that were previously determined to be well conserved in bilaterians. We identified 58 C2H2 ZFP genes in *Daphnia* that belong to these 40 distinct families. The *Daphnia* genome appears to be relatively efficient with respect to these well-conserved C2H2 ZFP, since only 7 of the 40 gene families have more than one identified member. Worms have a comparable number of 6. In flies and humans, C2H2 ZFP gene expansions are more common, since these organisms display 15 and 24 multi-member families respectively. In contrast, only three of the well-conserved C2H2 ZFP families have expanded in *Daphnia* relative to *Drosophila*, and in two of these cases, just one additional gene was found. The KLF/SP family in *Daphnia*, however, is significantly larger than that of *Drosophila*, and many of the additional members found in *Daphnia* appear to correspond to KLF 1/2/4 homologs, which are absent in *Drosophila*, but present in vertebrates.

The last study was conducted to investigate the conservation and distribution of 38 C2H2 ZNF gene families in Eukaryotes. We combined two popular approaches for homolog detection, Reciprocal Best Hit (RBH) and Hidden–Markov model (HMM) profile search, on a diverse set of complete genomes of 124 eukaryotic species ranging from excavates to humans. We succeeded in identifying 3,675 genes as distinct members of the 38 C2H2 gene families. This largely automated technique is much faster than manual methods and is able to detect homologs accurately and efficiently among a diverse set of organisms. Our analysis of the 38 evolutionarily conserved C2H2 ZNF gene families revealed a stepwise appearance of ZNF families, agreeing well with the phylogenetic relationship of the organisms compared and their presumed stepwise increase in complexity.

PREFACE

The aim of this dissertation is to provide insights to the various methods related to phylogenomics that can use large genome-scale data. The study addressed two related areas: phylogenomics as a method to build a species tree using the genome data, and prediction of gene function based on evolutionary analysis. For reconstructing phylogenetic trees using whole genome data, we investigated two different methods. In the first method, we used Singular Value Decomposition (SVD) analysis to re-examine current evolutionary relationships for 12 *Drosophila* species using the predicted proteins from whole genomes. In the second method, we used reduced representations of genomes provided by a novel class of Type IIB restriction endonucleases to reconstruct whole genome phylogenies of 21 *Drosophila* species. For predicting the function of the genes based on evolutionary analysis, we used 40 conserved C2H2 zinc finger genes from bilaterians to uncover zinc finger genes from *Daphnia pulex* and to study the distribution of these families in 124 different species of eukaryotes.

ACKNOWLEDGMENTS

The satisfaction that I feel at the successful completion of my research work would be incomplete if I did not mention the people whose able guidance, suggestion and encouragement crowned my effort with success. I express my deep sense of gratitude to Dr. Gary W. Stuart, esteemed chairman of my committee for selecting me and guiding me with his patience and knowledge throughout these wonderful years. I would also like to thank my committee members Dr. Jennifer Inlow, Dr. Allan Albig, Dr. Swapan Ghosh and Dr. James Hughes for advising me all these years and for critically examining the manuscript, making suitable corrections.

This Dissertation would be incomplete if I did not reckon the sacrifices, love, affection and support of my family members, including wife Mrs. Haninder Kaur, daughter Ms. Avani Seethram, mother Mrs. Premaleela and father Mr. Lakshminaryana Seetharam.

I thank my peers Dr. Yang Bai, Mr. Vincent Keller, Dr. Deepak Kumar, Mrs. Anupama, Mr. Bikram Sharma, Mr. Raghav Pandey and Ms. Amanda Jamison, for inspiration, support and encouragement. I also thank the office staff, Mrs. Tracy McDaniel and Mrs. Sunshine Mack, for their consistent help on administrative issues during my graduate studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The genomics era has been characterized by a massive growth in molecular sequence data, which eventually have led to the growth of many different fields of biology including molecular evolution. Evolutionary studies in recent years have been transformed by the development of new powerful techniques for investigating many mechanisms and events of molecular evolution. The ability to manipulate large datasets has not only resulted in greater precision in revealing evolutionary trends but also contributed to the fundamental understanding of biological systems [1]. Large collections of many different complete genomes now available in the public domain offer great advantages to genomic-scale evolutionary studies including: distinguishing orthologous (genes sharing common ancestral gene) and paralogous (genes originated by duplication, within a species) relationships; identifying the presence or absence of a particular gene in a genome to diagnose species specific gene loss events; studying structural and organizational features of each genome relative to another; and generating whole genome phylogenies. One other field of biology that is benefiting from the accumulating sequence data is functional genomics [2]. This field has expanded from the biological investigation of function for single genes or proteins towards analysis of multiple (or even all) genes or proteins in an organism or group of organisms in a systematic fashion. Functional genomics promises to fill the gap between sequence information and function information.

Although not all major taxa are adequately represented at present, the revolutionary progress in sequencing technologies is ensuring the continuous flow of sequence data for a wide variety of organisms representing a diverse set of taxa. There is also an ever expanding set of databases that store molecular sequence data in a very systematic and easily accessible manner. However, this tremendous amount of data also requires huge advancements in computational and data mining methods. Here we present some of the methods related to phylogenomics that are useful for the analysis of large amounts of genome-scale data.

Phylogenomics

Phylogenomics is a relatively new field of study that has greatly advanced with the revolutionary progress in DNA sequencing technology. The ever increasing number of whole genome sequences in the public domain and the new software tools available for sequence analysis have made this field possible. Phylogenomics in general includes various fields of research with a main emphasis on genomics and evolution. Initially phylogenomics was largely involved in the prediction of protein function through phylogenetic methods using the evolutionary history and orthologous information for each gene [3, 4]. But later, the term phylogenomics was used to describe the use of genomic-scale data to infer species phylogeny [1]. Though much of the recent literature has used this term exclusively for the latter meaning, there have been efforts to unify the various meanings of phylogenomics. Thus, phylogenomics is now defined as a "diverse field of study that mainly depends on molecular phylogenetic analysis of genome-scale data sets for predicting gene function, identifying traces of molecular adaptation, inferring evolutionary patterns of macromolecules, and establishing relationships and divergence times of genes and species" [5].

In our research, we are primarily concerned with new phylogenomic methods for building species trees using whole genome data, and for predicting the function of the genes based on evolutionary analysis. Unification efforts involving these two definitions for phylogenomics are of practical importance for the development of this field. In fact, these two meanings are mutually synergistic because precise functional information about genes genome-wide will improve phylogenetic tree reconstruction, and accuracy in the resulting species trees can aid in functional prediction of genes.

Species tree using genomic scale data

Phylogenetic tree reconstruction was originally performed by comparing morphological features of the organisms (e.g. skeleton structure of animals or flower architecture in plants etc.). But soon, as gene sequences became available, researchers started using the actual molecular sequences that undergo changes during evolution for comparison. Since then, evolutionary biology has come a long way, from analyzing single genes to analyzing several genes from all species to build a species tree. Traditional phylogenetic analysis involves two main steps, first identifying the homologous characters from all the species in question, and second, comparing these characters using standard phylogenetic methods for building the tree. For the first step of identifying homologous characters, characters such as morphological structures, biochemical properties, and molecular sequences (DNA or protein) can be used with the only condition being that they are homologous (sharing common ancestry). In recent times, molecular sequence is the popular choice for evolutionary studies, and in most cases, homologous molecular sequences are identified simply by searching for similar sequences across organisms.

The second step of building phylogenetic trees involves aligning the sequences to score the different character states as identified in various species. This often involves the use of

substitution matrices to score matches, mismatches, gaps and extensions for computing distances. In the next step of tree building, different methods can be used, including distance methods, maximum parsimony (MP), maximum likelihood (ML), and Bayesian methods. Distance methods are the simplest methods, in which the distance computed between the aligned sequences using the scoring method is converted to genetic distance and the evolutionary tree is reconstructed. Popular approaches include neighbor joining (NJ), unweighted pair group method with arithmetic mean (UPGMA) and minimum evolution (ME). In maximum parsimony, the tree is reconstructed from the distance that is calculated based on the minimum number of character changes required to explain the observed data. Maximum likelihood methods are based on the probability that a given tree could have produced the observed data using various probabilistic models. Bayesian methods, which select the tree based on Bayes mathematical formula combining the likelihood function (including model parameters) with prior probabilities [1, 6, 7]. With all these methods there are basically two alternatives for deriving the species tree from the alignment information of collections of multiple distinct sets of orthologous sequences. In one case, all the homologous gene sequences are aligned separately to make character state matrices that can be combined to make a super matrix, which can then be used to construct a single summary tree, or the separate trees for each homologous gene can be built from distinct character state matrices, and then combined to make a single super tree [1, 8].

Traditional methods thus require identification of homologous, and more precisely, orthologous genes in order to estimate relatedness. This process is not without drawbacks[7]. First, identifying the orthologous genes based on similarity alone makes it hard to differentiate them from mere paralogs. Second, pseudogenes with shared similarity to the gene might show up as false positives during the standard search. Finally, potential horizontal gene transfers,

incomplete lineage-sorting, and introgression can further complicate the analysis. One other factor that negatively affects traditional phylogenetic methods for estimating species trees is resolution[9]. Traditional methods can rarely include all the genomic-scale molecular sequence data nor can they include a very wide range of species with broad representation of each group in the analysis. The computational complexity of large scale datasets makes it hard to stretch these methods beyond a fundamental limit.

Consequently, alternate approaches capable of handling large datasets for building reliable species tree have been developed. These approaches are not only novel compared to traditional approaches, but they also can use molecular data other than the primary sequence information. Some of the most successful methods include methods based on gene content [10], gene order [11], or content of protein orthologs and folds [12]. Although these approaches are reasonably effective, they do not necessarily utilize the entire genomic dataset for analysis but instead utilize greatly filtered or preselected datasets. In this study, we have presented two novel methods that can produce a reliable species tree utilizing a large part of the genome. The first approach is based on Singular Value Decomposition (SVD) analysis [13], and is used to resolve the phylogeny of 12 recently sequenced *Drosophila* species. The second method generates phylogenies based on a reduced representation of the genome in which Type IIB enzymes [14] are used to produce random fragments sampled throughout the genome. Both these methods provide accurate comparisons for a high fraction of sequences within whole genomes.

Gene function predictions by evolutionary analysis

High throughput sequencing projects often end up with large amounts of uncharacterized sequence data that is practically inaccessible for most downstream analysis unless annotated with information identifying genes, introns, and other genomic elements. Gene prediction is arguably

the most important part of a genome sequence project; however, gene prediction has not been able to keep pace with rapidly progressing sequencing technologies. There is still a huge gap between these large genome sequence data and the tools required to utilize these data, especially, those needed to predict genes [3]. Many different programs are currently available to predict genes, but they vary in their accuracy and reliability. Most of these methods fall into two main categories: (a) homology methods, where statistically significant similarity information between the unknown gene or protein with the other well studied genes or proteins in a database is used to assign the presumed function and (b) *de novo* or *ab-initio* methods, where the structure of the gene is predicted based on signal sensors (measures that identify the presence of the functional parts, specific for a gene like exon-intron boundaries, open reading frames, promoter and poly A tail), and/or content sensors (measures that differentiates DNA into coding or non-coding, such as compositional bias, codon usage etc.). The latter methods can often be augmented by comparison to gene expression data and/or cDNA sequence information [15]. Both these methods have been widely used either separately or in conjugation for automatically predicting and classifying genes from the sequence information of whole genomes [16].

However, the ability to accurately predict gene function based on similarity to other gene sequences alone is complex and sometimes misleading. Gene duplication can generate paralogous genes that can acquire alternate functions while still retaining much of the similarity to its parent gene. Domain shuffling can lead to novel genes with essentially the same sequence, but with different arrangements. Genes can change function over time in a species-specific manner without undergoing significant change in the sequence. Novel genes without standard structure are fairly common in many species, and incorrectly predicted functions of genes in a database can be easily propagated by homology prediction methods [15, 16].

Molecular phylogenetics, a standard way of inferring relationships among species or genes, can provide valuable information for predicting function for uncharacterized genes even if a gene family has been through a gene duplication and functional divergence or even if the function of a gene has changed in one lineage. The method generally includes: first, the identification of the homologs as with any other gene prediction method; second, phylogenetic tree construction for the identified genes and homologs; third, overlaying functions to the homologs for which functional data are available; and fourth, making functional predictions for the genes of interest. The fourth step of functional prediction from the mapped phylogenetic tree can be carried out many different ways. The simplest method is to detect any gene duplication events in the evolutionary history of a gene, classify the homologs as orthologs and paralogs based on these events and then assign function to the unknown genes based on any ortholog of known function or based on all orthologs having the same function. Other sophisticated methods such as parsimony reconstruction techniques can also be used to predict the likely functions of unknown genes. Here, the function of a given gene is predicted by identifying the evolutionary scenario with the fewest functional changes over time [3].

Although, phylogenomic methods require much more effort and manual labor than homology-based methods, the results produced by them are more accurate [1]. The main reason behind this is that phylogenomics not only uses the sequence similarity information of a gene (as in homology methods), but it also includes the history of the gene. Despite these improvements, it is important to emphasize that all these methods are mere predictions; none of them represents a physical or biochemical test of function, and thus further experimental validation may be needed. However, functional predictions offer great firsthand information about a gene and are

of great value in not only providing direction for biophysical validations, but also for utilizing the tremendous data generated by huge genome projects

Specific aims for this study

In this study, the main emphasis is on utilizing whole genome data for the purpose of reconstructing species trees and to utilize phylogenetic analysis to identify novel genes from multiple genomes. For the first objective, we developed new techniques such as singular value decomposition analysis and type IIB fragments from whole genomes, and for the latter objective, we used standard phylogenomic methods to uncover all bilaterian conserved C2H2 zinc finger genes from the newly sequenced organism *Daphnia pulex*, and subsequently a greatly expanded group of diverse eukaryotic species ranging from protists to humans.

Specific objectives of the study were as follows:

1. To implement our newly improved Singular Value Decomposition (SVD) analysis to re-examine current evolutionary relationships for 12 *Drosophila* species using the whole-genome predicted protein datasets from all the organisms.
2. To build whole genome phylogenies using nucleotide datasets from 21 *Drosophila* species by reduced representation of the genomes using a novel class of Type IIB restriction endonucleases.
3. To identify orthologous members of C2H2type zinc finger genes from the newly sequenced genome *Daphnia pulex,* based on conserved zinc finger gene families in bilaterians.

4. To study the distribution and conservation of 38 bilaterian C2H2zinc finger gene families in various groups of eukaryotes to gain insights into the evolution of zinc finger genes in eukaryotes.

CHAPTER 2

WHOLE GENOME PHYLOGENIES FOR MULTIPLE *DROSOPHILA* SPECIES

Abstract

Reconstructing the evolutionary history of organisms using traditional phylogenetic methods may suffer from inaccurate sequence alignment. An alternative approach, particularly effective when whole genome sequences are available, is to employ methods that do not use explicit sequence alignments. We extend a novel phylogenetic method based on Singular Value Decomposition (SVD) to resolve the phylogeny of 12 recently sequenced *Drosophila* species. SVD analysis provides accurate comparisons for a high fraction of sequences within whole genomes without the prior identification of orthologs or homologous sites. With this method all protein sequences are converted to peptide frequency vectors within a matrix that is decomposed to provide simplified vector representations for each protein of the genome in a reduced dimensional space. These vectors are summed together to provide a vector representation for each species, and the angle between these vectors provides distance measures that are used to construct species trees. An unfiltered whole genome analysis (193,622 predicted proteins) strongly supports the currently accepted phylogeny for 12 Drosophila species at higher dimensions except for the generally accepted, but difficult to discern, sister relationship between *D. erecta* and *D. yakuba*. Also, in accordance with previous studies, many sequences appear to

support alternative phylogenies. In this case, we observed grouping of *D. erecta* with *D. sechellia* when approximately 55% to 95% of the proteins were removed using a filter based on projection values or by reducing resolution by using fewer dimensions. Similar results were obtained when just the *melanogaster* subgroup was analyzed. These results indicate that using our novel phylogenetic method, it is possible to consult and interpret all predicted protein sequences within multiple whole genomes to produce accurate phylogenetic estimations of relatedness between *Drosophila* species. Furthermore, protein filtering can be effectively applied to reduce incongruence in the dataset as well as to generate alternative phylogenies.

Introduction

Methods that determine phylogenies based on a restricted number of genes can be negatively affected by potential horizontal gene transfers, incomplete lineage-sorting, introgression, and the unrecognized comparison of paralogous genes [17]. The recent explosive increase in the number of completely sequenced genomes allows us to consider inferring gene and/or organismal relationships using complete sequence data. Several methods for generating phylogenies based on whole genome information have been explored, and many of these have been applied recently to re-examine the phylogeny of *Drosophila*. These include methods based primarily or exclusively on gene content [10], gene order [11], and detailed comparisons of operationally defined orthologs [18]. However, these methods often fail to provide detailed and unbiased comparisons of a high fraction of sequences and instead produce phylogenies based on greatly filtered, preselected datasets. We recently developed a phylogenetic method that provides accurate comparisons for a high fraction of sequences within whole genomes without the prior identification of orthologous or homologous sites [13]. Our approach allows a relatively comprehensive comparison of complete genome protein sequence, thereby taking into account a

higher fraction of total sequence information and providing comprehensive definitions for the various species of interest. This method has been successfully applied to a number of diverse species including vertebrate mitochondrial genomes, plant viral genomes, and eukaryotic nuclear genomes. [13, 19-21].

Recently, complete genome sequences for 10 additional species of *Drosophila* were added to the sequences already available for *D. melanogaster* and *D. pseudoobscura* to improve the precision and sensitivity of evolutionary inference regarding these organisms [22]. As a result, the currently accepted species phylogeny for these organisms has been further refined and resolved. However, these methods generally continue to utilize greatly filtered data sets primarily comprised of selected single copy orthologous sequences [23-28].

Many such studies have resulted in what is largely considered to be a fully resolved phylogeny for the12 sequenced species of *Drosophila.* However, some doubts remain with respect to the placement of certain members of the melanogaster group: *D. erecta, D. yakuba* and *D. melanogaster*, placement of the Hawaiian species: *D. grimshawi,* and to some extent the *virilis-repleta* group: *D. virilis* and *D. mojavenis* [29-33]. Among these, the placement of *D. erecta* and *D. yakuba* with respect to *D. melanogaster* is perhaps least certain. Though evidence has been presented to support all the possible phylogenies with respect to *D. melanogaster, D. erecta,* and *D. yakuba,* support for each of these phylogenies is not uniformly strong[26]. In this study we apply our more inclusive whole genome phylogenetic method on the 12 genomes of *Drosophila* to further investigate and validate our current understanding of their phylogenetic relationships.

Materials and methods

Datasets

Complete predicted protein sequences for 12 Drosophila species were downloaded from the 'Assembly, Alignment and Annotation of 12 Drosophila species' website (http://rana.lbl.gov/drosophila/) and were compiled into a single dataset. Various distinct subsets of this larger dataset were also constructed. The number of protein sequences found within the genome of each species of *Drosophila* is summarized in Table 1.

Peptide frequencies and SVD

Each protein sequence in the dataset was recoded as overlapping peptide frequency vectors for each of the 160,000 possible tetrapeptides. The resulting matrix "A" was then subjected to Singular value decomposition (SVD), generating three output matrices (left and right singular vectors and their corresponding singular values). Refined (noise-reduced) vector descriptions of proteins are obtained from the truncated right matrix derived via SVD. Two distinct motif/families are frequently identified per triplet, since each triplet describes both a correlated motif/family (positive values) and an anti-correlated motif/family (negative values). The phylogeny studies were conducted under two different SVD settings: one, referred to as "higher dimension," utilized a total of 800 singular triplets as output, and the other, referred to as "lower dimension," utilized just 400 singular triplets as output.

Species trees and branch support

Distance matrices were derived by summing all the SVD derived right protein vectors for a given organism and then comparing the relative orientation of the resulting species vectors using the program cosdist. Species trees were subsequently derived from distance matrices using

Phylip-Neighbor. Two distinct resampling methods were used to provide branch support: a traditional bootstrap procedure and a modified jackknife procedure. For the bootstrap, a fixed number of singular vectors were randomly sampled from the total singular vectors generated and were used to construct 100 species trees. For the successive delete-one jackknife procedure[13], the least dominant singular vector was removed successively (from the total vectors generated, down to 100 vectors) to generate ordered sets of singular vectors, and a new tree was estimated following each removal.

Results

Preliminary studies were conducted using a small dataset comprising only 6 genomes of the melanogaster group (*D. melanogaster, D. sechellia, D. simulans, D. erecta, D. yakuba and D. ananassae*) with a total of 100,851 predicted proteins. Further studies were conducted using a large dataset consisting of all the 12 *Drosophila* spp. genomes with a total of 193,622 proteins (Table 1). An additional 11 genome datasets excluding one of the *melanogaster* group species were also constructed for the detailed analysis of the phylogenies. Among the 12 species, *D. melanogaster* had the highest number of predicted proteins (22,765), and *D. virilis* had the lowest (14,491). Each species' contribution to the dataset was in the range of 7.48% to 8.51%, except for *D. melanogaster*, which contributed about 11.76% for the total. In previous studies, we noted that modest size differences in genomes had little effect on the final outcome of the tree [19].

All the proteins in each of these three datasets were separately recoded as overlapping tetrapeptide frequency vectors. The resulting matrix was then subjected to a truncated SVD analysis that generates three component matrices: the "left" matrix or "peptide" matrix (U), the "right" matrix or "protein" matrix (V) and the central matrix (∑). The original matrix can be

reformed using the relation $A = U \sum V^T$. The "protein" vectors provided in the "right" factor matrix are known to provide reduced dimensional definitions for all proteins in the dataset as linear combinations of the orthogonal "right" singular vectors [20]. The phylogeny studies were conducted under two different SVD settings, one referred to as "higher dimension," where we used a total of 800 singular triplets as output and the other referred to as "lower dimension" using only 400 singular triplets as output. The SVD was then applied to the 12, 11 and 6 species datasets of *Drosophila* separately. The detailed comparative information contained within the hundreds of singular vectors and their corresponding motifs and gene families was subsequently used to build a species phylogeny by summing all the SVD-derived right protein vectors separately for each organism and then comparing the relative orientation of the resulting species vectors [19].

Higher dimension SVD analysis

Figure 1 and Figure 2 show the SVD-based topology obtained via Neighbor-joining for the 6 and 12 genome *Drosophila* species data sets respectively. Two types of resampling methods were used to estimate branch statistics for this tree. The bottom value on each branch was generated using a traditional bootstrap procedure [13] by sampling 800 singular triplets to construct 700 species trees. The top value on each branch was generated using a successive, delete-one jackknife procedure [13] wherein the least dominant singular vector was removed successively (from 800 to 100 vectors) to generate 700 ordered sets of singular vectors, and a new tree was estimated following each removal. Most of the branches were well supported following application of either the modified jackknife procedure or the bootstrap procedure. Bootstrap yielded a slightly lower branch support for the *D. sechellia*, *D. simulans,* and *D. melanogaster* branch, but all other branches were strongly supported by both procedures. The

observed difference was likely due to the uniform use of the 700 most dominant vectors in our modified jackknife procedure, whereas, the standard bootstrap samples randomly over all 800 vectors generated. The end result is a phylogeny that corresponds well to the currently accepted phylogeny [26], except for *D. erecta* and *D. yakuba*, which remain adjacent in the tree but fail to cluster as sister species.

To further examine the robustness of the data supporting the correct tree, we performed a series of analyses by systematically excluding protein sequences that were poorly described by their corresponding singular vectors in terms of projection values. The possible projection values for a given protein range from -1 to +1. In the first step, all protein sequences having projection value less than or equal +0.001 and more than or equal to -0.001 were removed (about 9,500 sequences). The filter was increased stepwise with an increment of 0.001, and each corresponding dataset was used in turn to construct a tree. When about 54.54% (105,596 sequences) of the original dataset was removed (projection value less than or equal to +0.003 and more than or equal to -0.003), a unique clustering of *D. erecta* with *D. sechellia* was observed (Figure 3). Continued successive increases in stringency to remove poorly described proteins failed to alter this novel cluster until more than 95% (185,039) of the total protein sequences were removed. This resulted in a re-clustering of *D. erecta* with *D. yakuba* as sister species, but this re-clustering was accompanied by the movement of *D. melanogaster* to a novel position (Figure 4). Removing a high fraction of poorly described proteins (those with smaller projections on any singular vector) would presumably tend to produce a more highly correlated data set consisting of smaller sets of highly conserved proteins. The tree generated using the modified jackknife procedure, rather than the bootstrap, showed a similar branching pattern. Branch

support values for the tree exceeded 80% in all cases, and only 60% for the *D. yakuba and D. erecta* cluster.

Lower dimension SVD analysis

A corresponding lower dimension analyses of the *Drosophila* spp. was also conducted using the same procedure but with fewer (500) singular triplets. Here the bootstrap branch statistics were generated by sampling 100 random sets of 150 singular triplets to construct 100 species trees. The delete-one jackknife values were generated using 400 ordered sets of singular vectors. Trees were estimated following each successive removal of a least dominant vector from 500 to 100 vectors. The SVD phylogeny obtained for the unfiltered 12 *Drosophila* species dataset (Figure 5) corresponds well to the currently accepted phylogeny, except for *D. erecta*, which shows a novel affinity with *D. sechellia*. It proved possible to disrupt this novel affinity after reducing the number of proteins used in the summation step by 97.57% (Figure 6) by applying a relatively severe filter (projection value less than or equal to +0.035 and more than or equal to -0.035) and thus using only the remaining highly correlated data set consisting of smaller sets of highly conserved proteins. Branch support values for the tree exceeded 80% in all cases, except for the *D. melanogaster, D. yakuba and D. erecta* cluster, which had 70% support.

To study the relationships among members of the *melanogaster* group without the influence of *D. erecta*, a slightly smaller dataset of 11 Drosophila species (178,574 total predicted proteins) was used for analysis. This data set produced the currently accepted phylogeny with strong branch support (Figure 7). The observed relationship was consistent across different levels of protein filtering. Both the bootstrap and the modified jackknife produced strong branch support values for most branches.

A similar result was obtained with an even smaller dataset that included only 6 genomes with 100,851 predicted proteins (Figure 8 and 9). When subjected to SVD analysis, this analysis produced the currently accepted phylogeny for all 6 members of the *melanogaster* group, but only under stringent protein filtering (Figure 9). The effect of including more proteins using a less severe protein filter was similar for both the 12 genome tree and the 6 genome tree: *D. erecta* fails to cluster with *D. yakuba* and instead clusters with *D. sechellia*. However, just as with the 11 *Drosophila* dataset, exclusion of *D. erecta* from the melanogaster group produced the currently accepted phylogeny with strong branch support (Figure 10) without filtering any proteins. The effect of other genomes on the phylogeny was systematically studied by excluding one of the *melanogaster* group species from the original 12 genome dataset. All these analyses showed the novel *D. sechellia* and *D. erecta* clustering (Figures 11, 12, 13 and 14) except for the dataset from which *D. sechellia* was excluded, which produced the currently accepted phylogeny (Figure 15). But, all datasets produced the currently accepted phylogeny under stringent filtering conditions (Figures 16, 17, 18, 19, and 20).

Conclusions

Our results indicate that it is possible to consult and interpret all predicted protein sequences within multiple whole genomes to produce accurate phylogenetic estimations of relatedness between *Drosophila* species. The phylogenetic tree derived for the 6 species of the *melanogaster* group, as well as all 12 species of *Drosophila,* exhibits strong branch support values and corresponds almost exactly to the currently accepted phylogeny. The most recent independent analyses based on whole genome sequence information depends upon filtered data sets in which a restricted number of highly conserved and putatively othologous genes were compared. We conclude that it is possible to produce equivalent results using a novel method

that includes the entire dataset for a more robust analysis. However, this greatly expanded data set appears to contain a strong component of conflicting sequence information that specifically causes *D. erecta* and *D. sechellia* to cluster. This anomaly was observed only when more than 55% (105,596) of the proteins are removed. However, this cluster disappears again when 95% (185,039) of poorly described proteins are removed. At lower dimensions, the *D. erecta* and *D. sechellia* cluster appears to be stable under various filter settings. Only under stringent filtering conditions could the accepted phylogeny be restored. Additionally exclusion of either *D. sechellia* or *D. erecta* from the 12 species dataset yielded the currently accepted phylogeny.

The relative placement of *D. erecta* and *D. yakuba* with respect to *D. melanogaster* was largely uncertain until recent multigene analyses tended to support the same standard tree [23-28, 34]. Previously, single gene analyses supported a variety of distinct trees [29, 31, 35-41], and more recent comprehensive surveys of putative orthologs revealed a high frequency of conflicting trees [25-27]. Depending on the evolutionary model applied, roughly 40% of all orthologous genes examined supported alternative phylogenies within the *melanogaster* subgroup [26]. In this case, the standard *D. erecta/D. yakuba* cluster was specifically examined, and only two alternatives, those in which either of these species specifically clustered instead with *D. melanogaster*, were considered. Two reasons are commonly offered to explain the conflicts observed in these surveys of single-gene phylogenies: incomplete lineage sorting, and introgression. Either of these processes could potentially be at least partly responsible for the novel grouping of *D. erecta* and *D. sechellia* we observed under the special mid-range filtering conditions reported here.

An alternative, but not mutually exclusive, explanation for the conditional novel clustering observed in this work is that the sequence signal causing this clustering exists

primarily outside of a reasonably complete list of identifiable orthologs (Figure 4). Although not a necessity, this signal could easily be interpreted as homoplasious. This interpretation is supported by the fact that the standard clustering of *D. yakuba* and *D. erecta* was observed again when using only protein sequences with the highest projection values, which includes a small subset of proteins that are more likely to represent close homologs or orthologs. It is also possible that the sequence signal responsible might not be exclusively located outside identifiable orthologs, but might also be partly embedded within orthologs as similar subsets of specific sequence changes within these genes. In either case, it would still be interesting to further investigate the source and strength of these presumed homoplasies, given that they specifically and consistently support a single alternative placement for a single species within a complex tree.

Acknowledgements

Support for this work was provided in part by the Biology Department and the School of Graduate Studies at Indiana State University. In addition, help with software development and modification was provided by Yihua Bai from the Center for Instructional and Research Technology at ISU.

Table 1. List of 12 *Drosophila* species used in the analysis, along with the number of predicted proteins.

#	***Species***	***Proteins***
1	*Drosophila simulans*	15415
2	*Drosophila sechellia*	16471
3	*Drosophila melanogaster*	22765
4	*Drosophila erecta*	15048
5	*Drosophila ananassae*	15070
6	*Drosophila yakuba*	16082
7	*Drosophila pseudoobscura*	16308
8	*Drosophila persimilis*	16878
9	*Drosophila willistoni*	15513
10	*Drosophila mojavensis*	14595
11	*Drosophila virilis*	14491
12	*Drosophila grimshawi*	14986

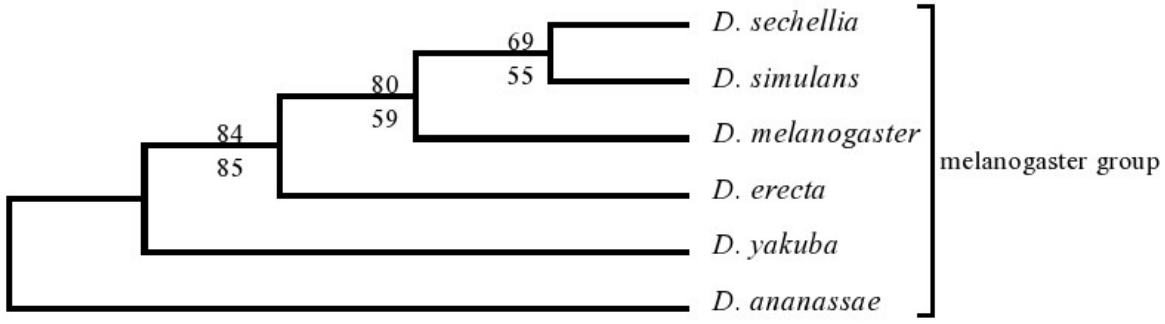

Figure 1. The higher dimension SVD tree for the 6 *Drosophila* spp., using all 700 vectors, without filtering any proteins (upper branch values, modified jackknife and lower branch values, bootstrap procedure for tree generation).

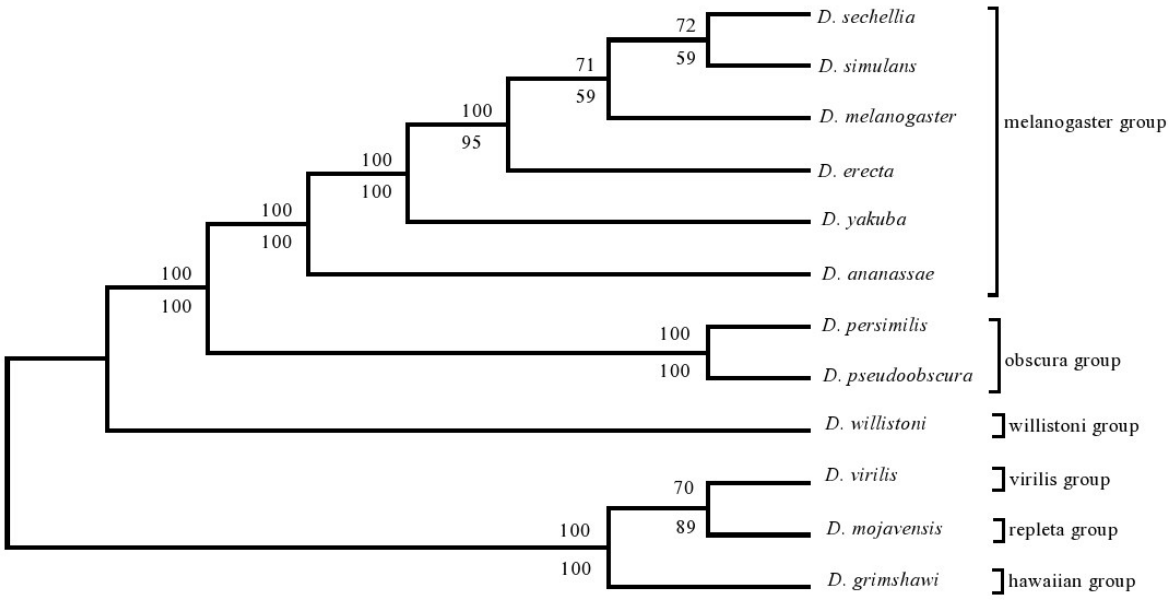

Figure 2. The higher dimension SVD tree for the 12 *Drosophila* spp., using all 700 vectors, without filtering any proteins (upper branch values, modified jackknife and lower branch values, bootstrap procedure for tree generation).

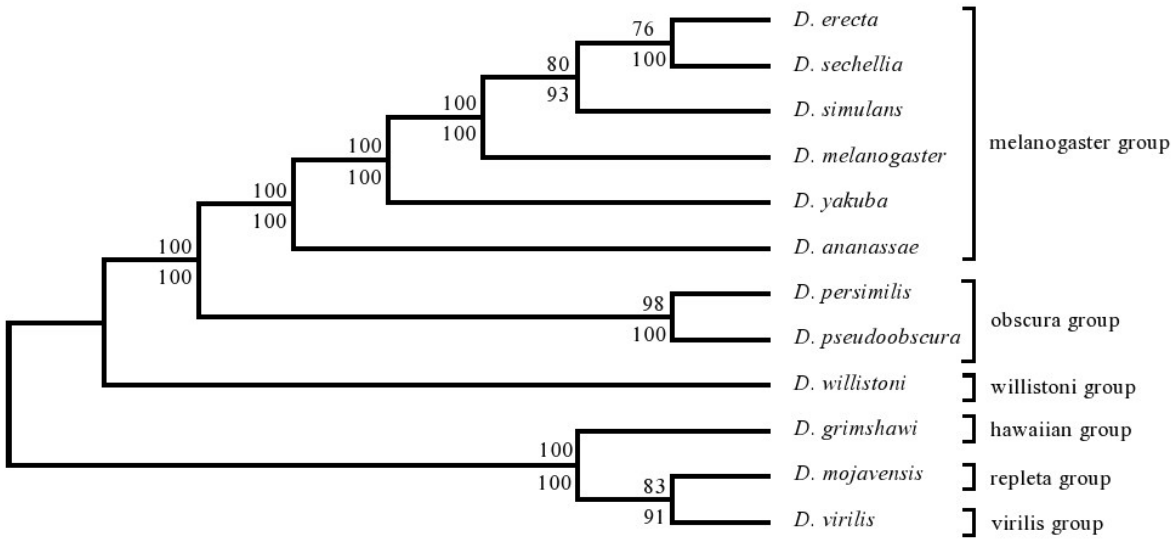

Figure 3. SVD (higher dimension) tree for the 12 *Drosophila* spp., using all 700 vectors, with filtering cut off value of ±0.003, retaining 88,026 (45.46%) protein sequences (upper branch values, modified jackknife and lower branch values, bootstrap procedure for tree generation).

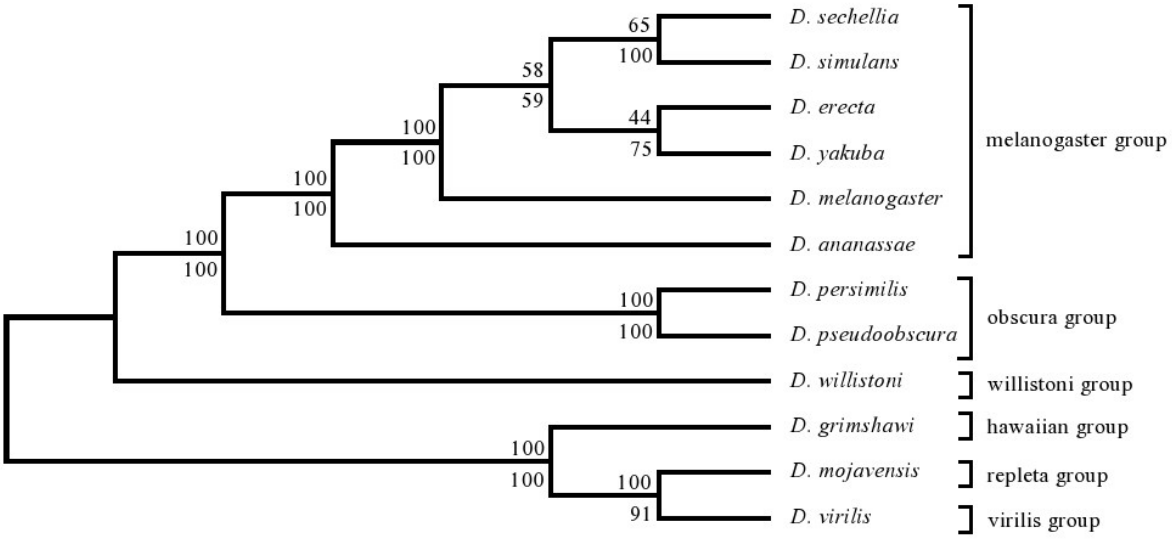

Figure 4. SVD (higher dimension) tree for the 12 *Drosophila* spp., using all 700 vectors, with filtering cut off value of ±0.032, retaining 8,583 (4.43%) protein sequences (upper branch values, modified jackknife and lower branch values, bootstrap procedure for tree generation).

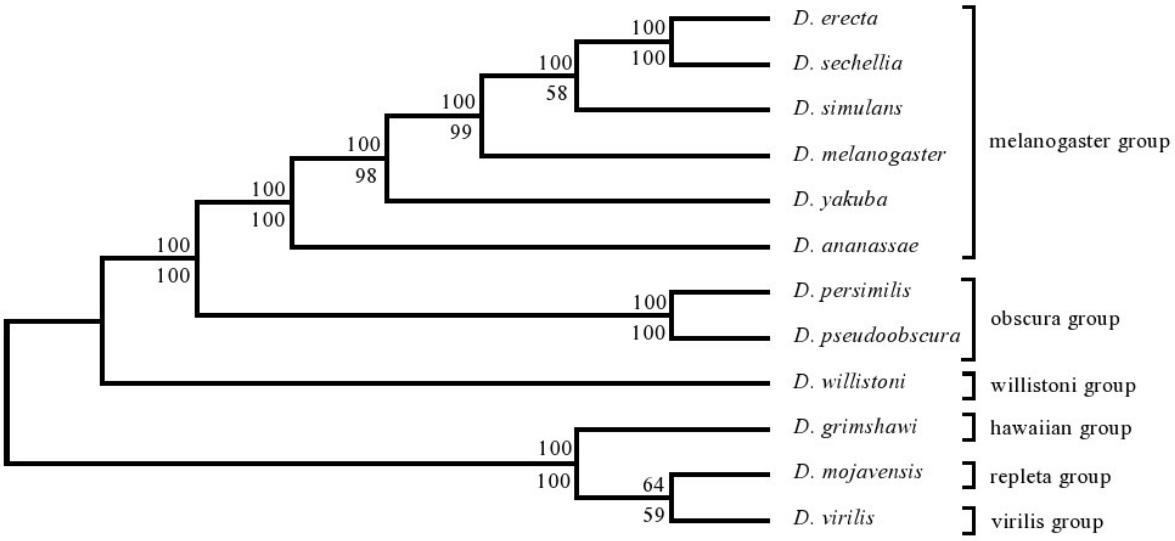

Figure 5. The lower dimension SVD tree for the 12 *Drosophila* spp., using 300 vectors, without filtering any proteins (upper branch values, modified jackknife and lower branch values, bootstrap procedure for tree generation).

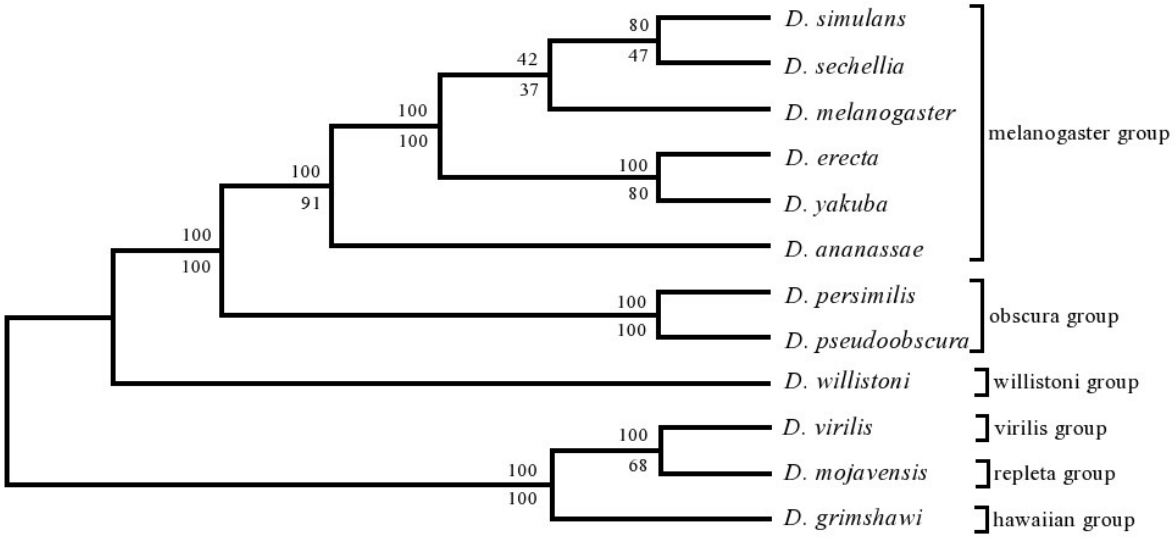

Figure 6. The lower dimension SVD tree for the 12 *Drosophila* spp., using 300 vectors, with heavy filtering of proteins with projection values ≤ ±0.035. A total of 4430 (2.43 %) proteins were used for constructing trees (upper branch values, modified jackknife and lower branch values, bootstrap procedure for tree generation)

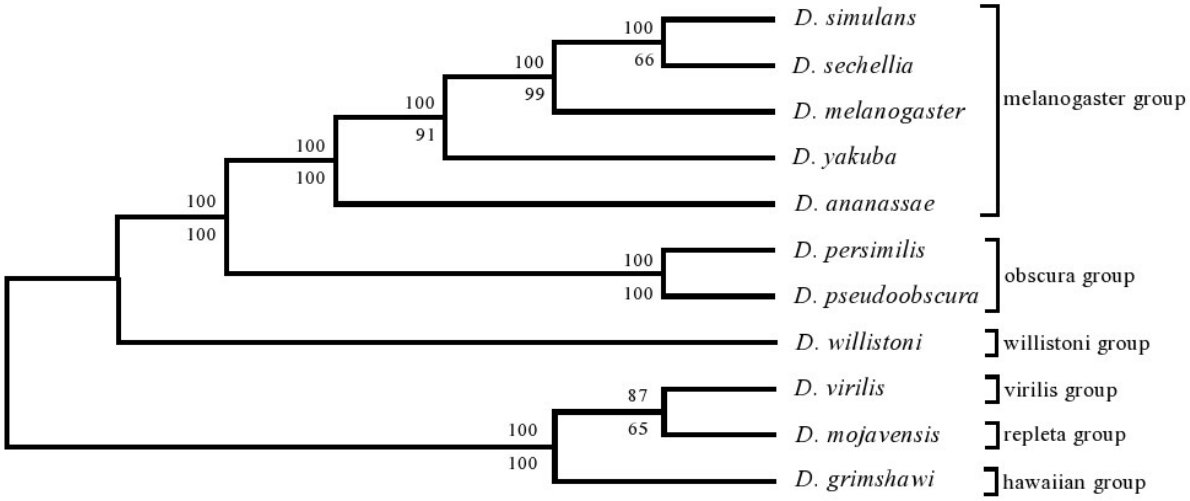

Figure 7. The lower dimension SVD tree for the 11 *Drosophila* species (excluding *D. erecta*) using 300 vectors, without filtering any proteins (upper branch values, modified jackknife and lower branch values, bootstrap procedure for tree generation).

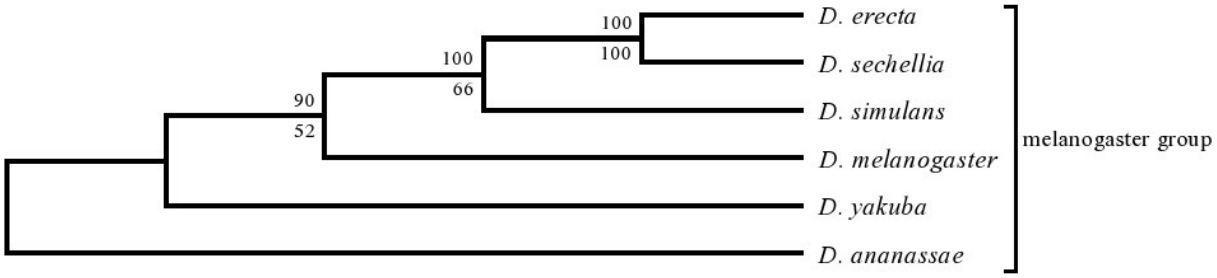

Figure 8. The lower dimension SVD tree for the 6 *Drosophila* species (melanogaster group) using 300 vectors, without filtering any proteins (upper branch values, modified jackknife and lower branch values, bootstrap procedure for tree generation).

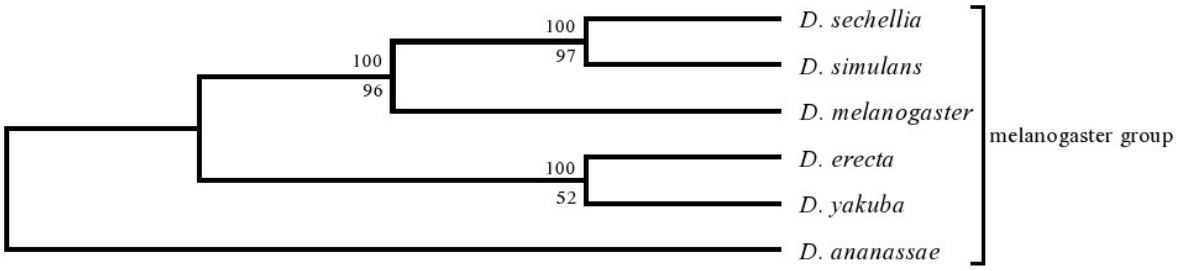

Figure 9. The lower dimension SVD tree for the 6 *Drosophila* spp., using 300 vectors, with heavy filtering of proteins with projection values ≤ ±0.035. A total of 4048 (4.06 %) proteins were used for constructing trees (upper branch values, modified jackknife and lower branch values, bootstrap procedure for tree generation).

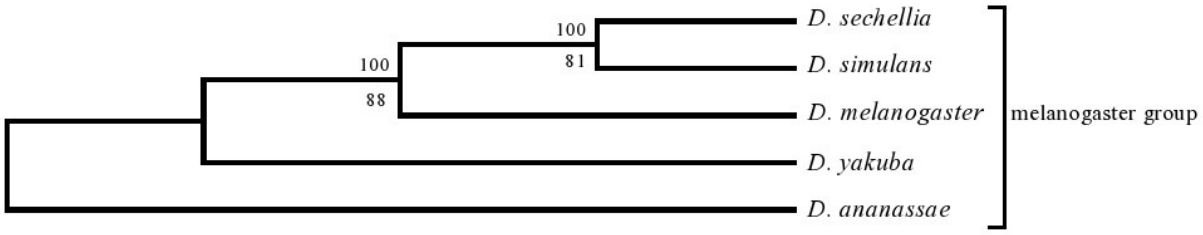

Figure 10. The lower dimension SVD tree for the 5 *Drosophila* species (melanogaster group, excluding *D. erecta*) using 300 vectors, without filtering any proteins (upper branch values, modified jackknife and lower branch values, bootstrap procedure for tree generation).

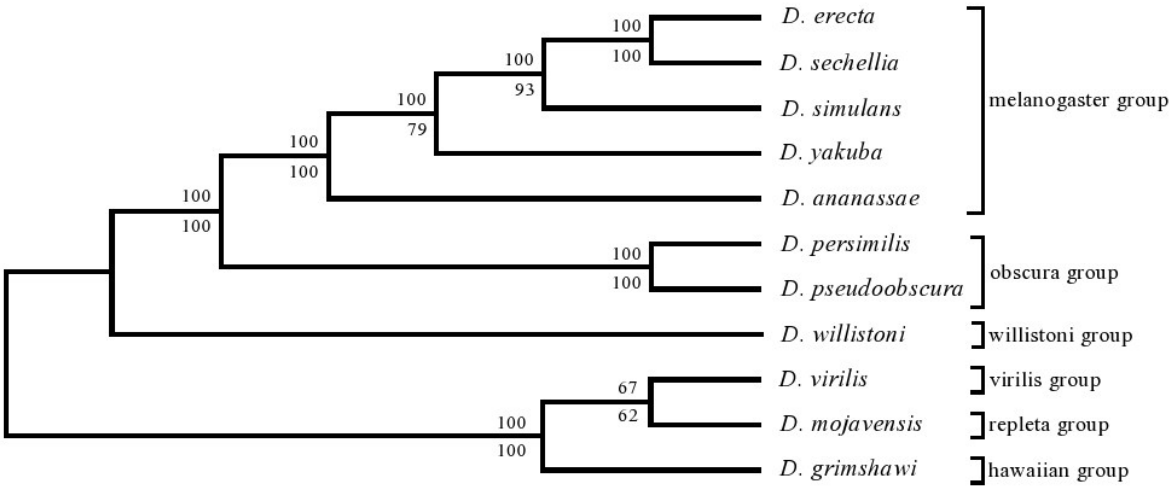

Figure 11. SVD (lower dimension) tree for the 11 *Drosophila* species (excluding *D. melanogaster*), using 300 vectors, without filtering any proteins (upper branch values, modified jackknife and lower branch values, bootstrap procedure for tree generation).

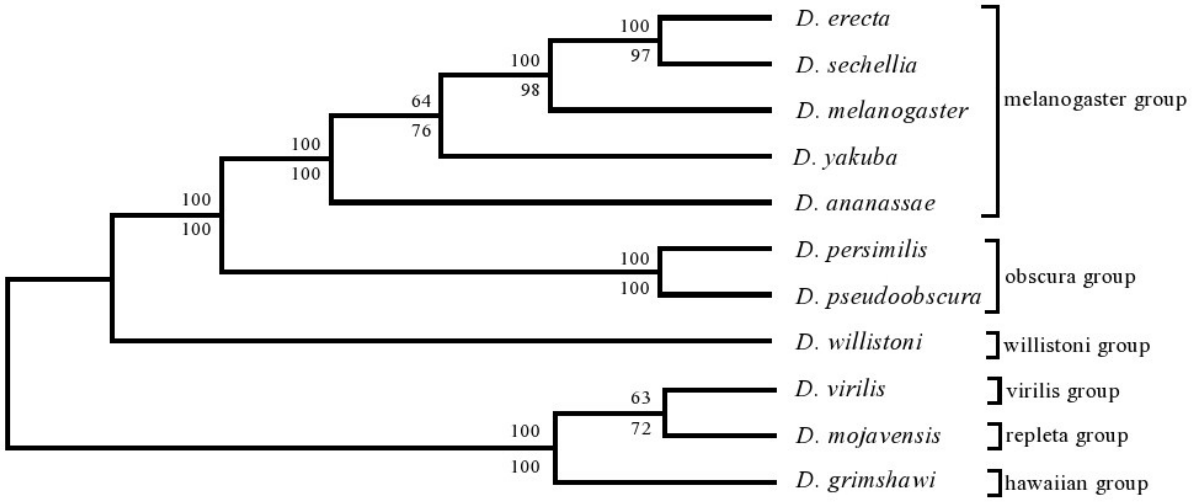

Figure 12. SVD (lower dimension) tree for the 11 *Drosophila* species (excluding *D. simulans* using 300 vectors, without filtering any proteins (upper branch values, modified jackknife and lower branch values, bootstrap procedure for tree generation).

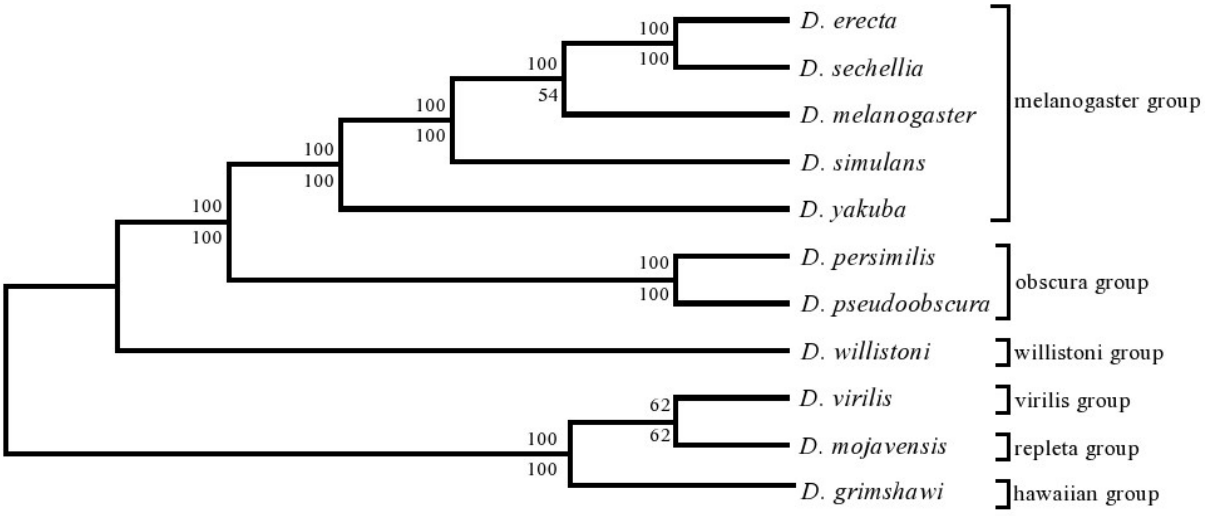

Figure 13. SVD (lower dimension) tree for the 11 *Drosophila* species (excluding *D. ananassae*) using 300 vectors, without filtering any proteins (upper branch values, modified jackknife and lower branch values, bootstrap procedure for tree generation).

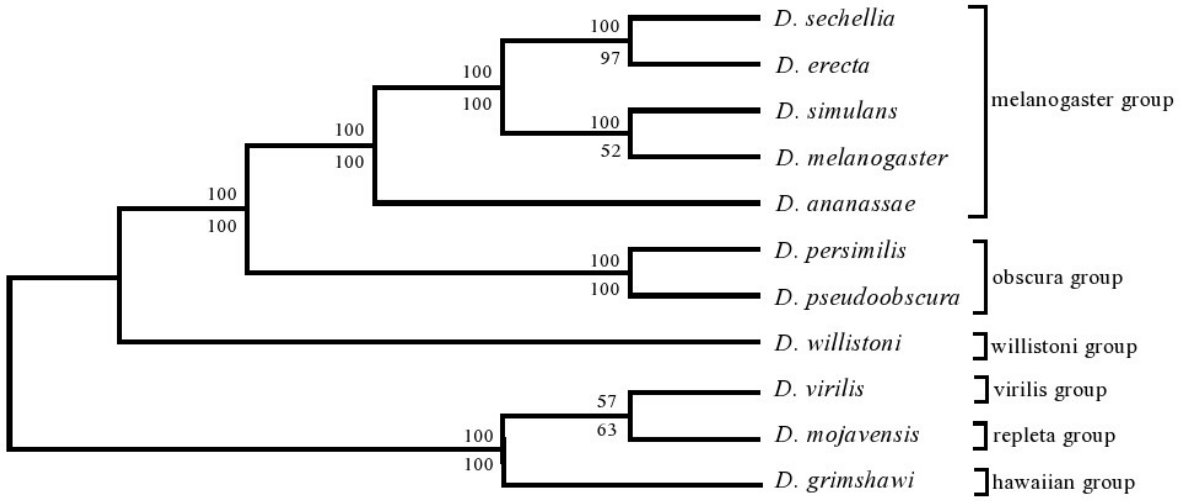

Figure 14. SVD (lower dimension) tree for the 11 *Drosophila* species (excluding *D. yakuba*) using 300 vectors, without filtering any proteins (upper branch values, modified jackknife and lower branch values, bootstrap procedure for tree generation).

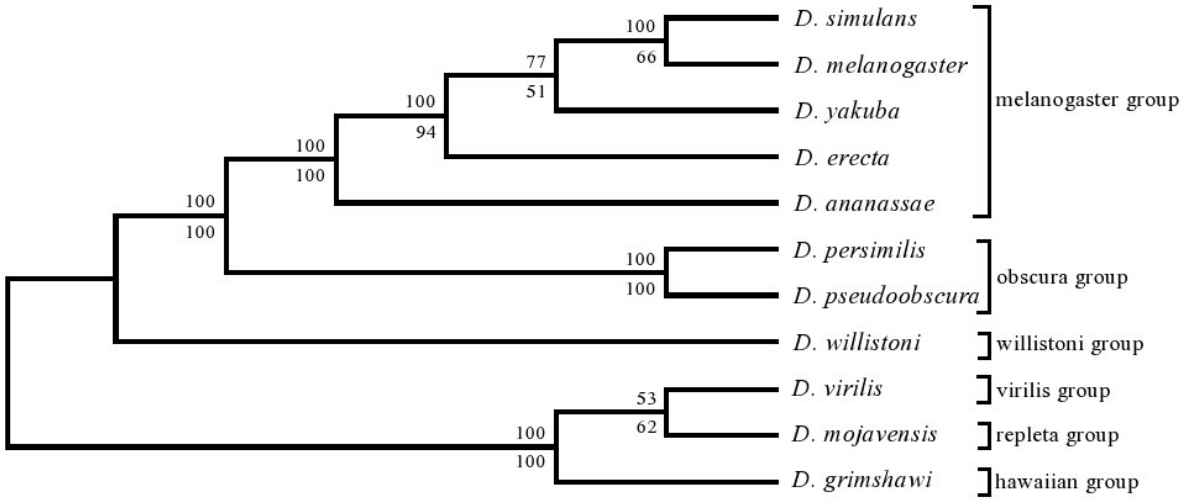

Figure 15. SVD (lower dimension) tree for the 11 *Drosophila* species (excluding *D. sechellia*) using 300 vectors, without filtering any proteins (upper branch values, modified jackknife and lower branch values, bootstrap procedure for tree generation).

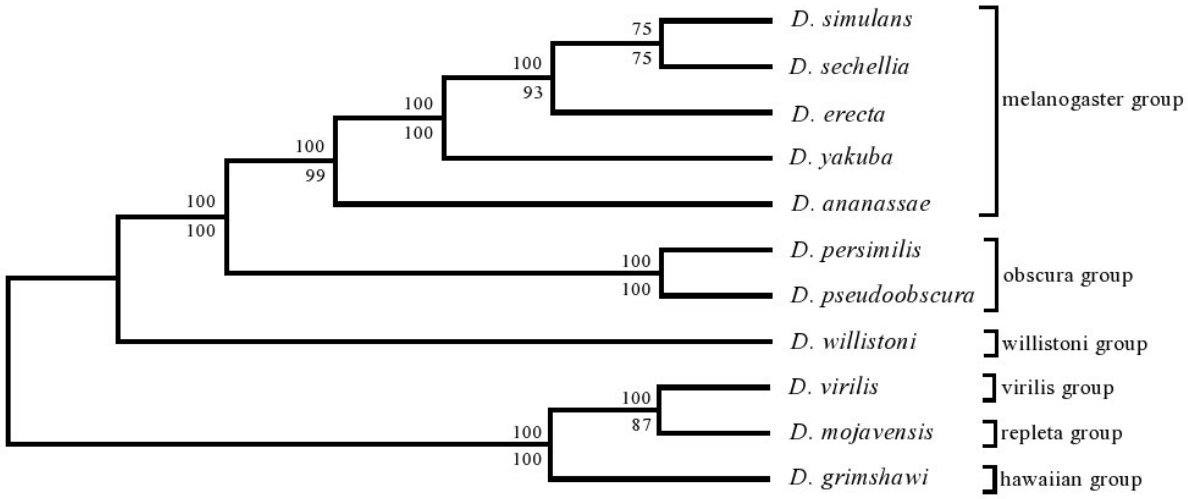

Figure 16. SVD (lower dimension) tree for the 11 *Drosophila* species (excluding *D. melanogaster*), using 300 vectors, with filtering cut off value of ±0.035, retaining 4146 (2.43 %) protein sequences (upper branch values, modified jackknife and lower branch values, bootstrap procedure for tree generation).

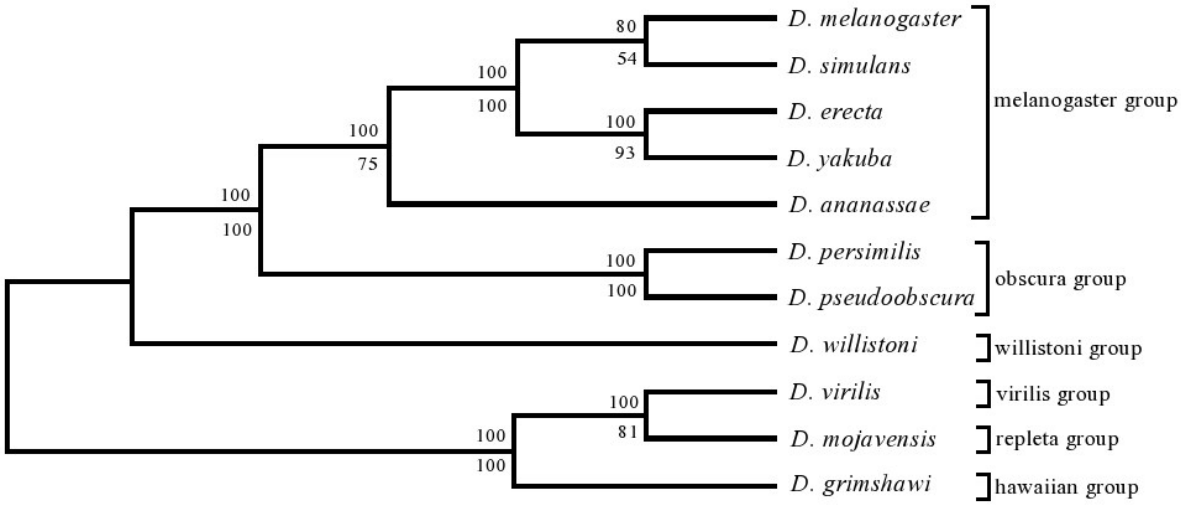

Figure 17. SVD (lower dimension) tree for the 11 *Drosophila* species (excluding *D. sechellia*), using 300 vectors, with filtering cut off value of ±0.035, retaining 4271 (2.43 %) protein sequences (upper branch values, modified jackknife and lower branch values, bootstrap procedure for tree generation).

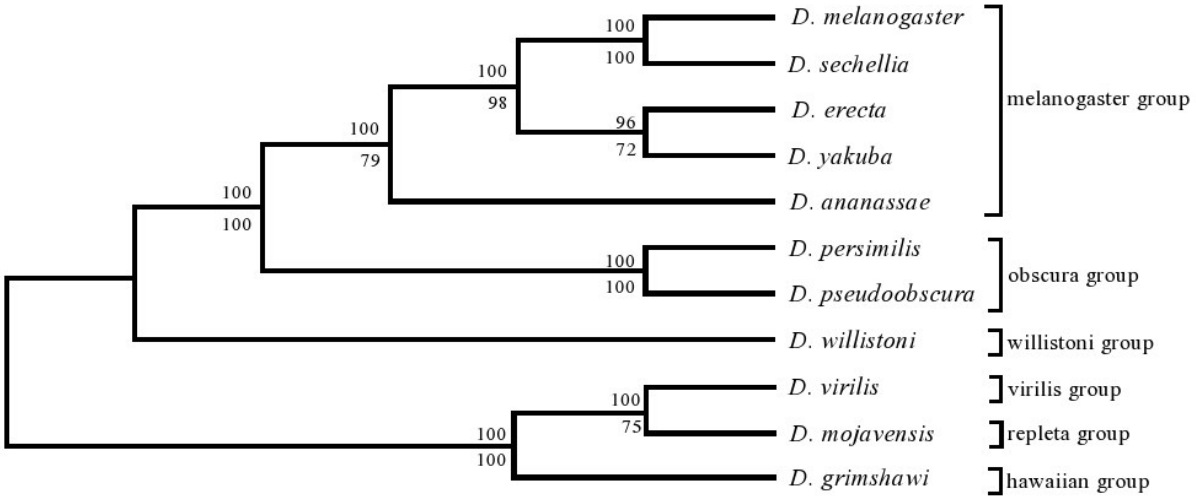

Figure 18. SVD (lower dimension) tree for the 11 *Drosophila* species (excluding *D. simulans*), using 300 vectors, with filtering cut off value of ±0.035, retaining 4611 (2.61 %) protein sequences (upper branch values, modified jackknife and lower branch values, bootstrap procedure for tree generation).

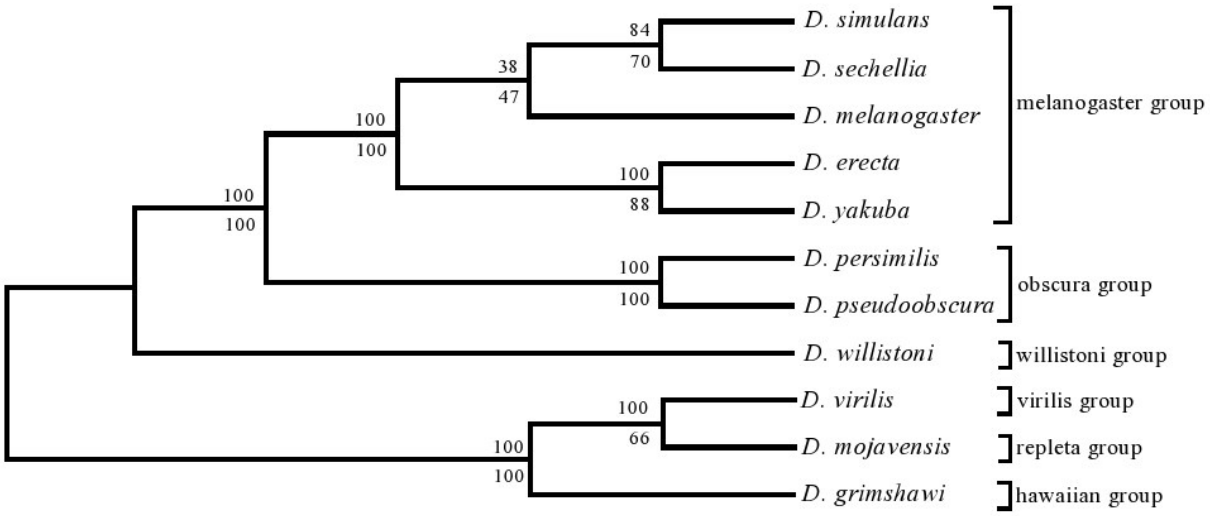

Figure 19. SVD (lower dimension) tree for the 11 *Drosophila* species (excluding *D. ananassae*), using 300 vectors, with filtering cut off value of ±0.035, retaining 4343 (2.45 %) protein sequences (upper branch values, modified jackknife and lower branch values, bootstrap procedure for tree generation).

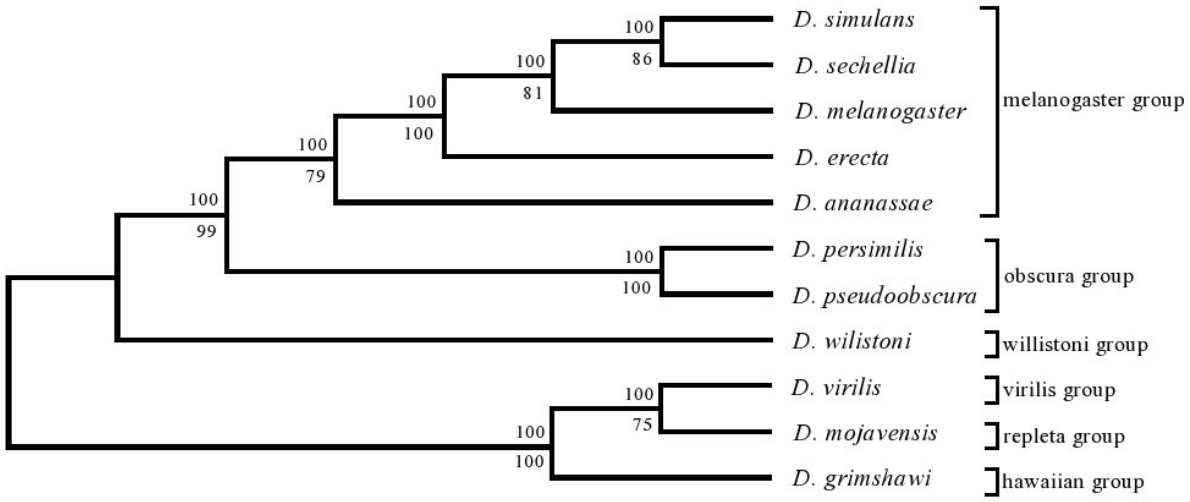

Figure 20. SVD (lower dimension) tree for the 11 *Drosophila* species (excluding *D. yakuba*), using 300 vectors, with filtering cut off value of ±0.035, retaining of 4628 (2.63 %) protein sequences (upper branch values, modified jackknife and lower branch values, bootstrap procedure for tree generation).

CHAPTER 3

WHOLE GENOME PHYLOGENY FOR 21 *DROSOPHILA* SPECIES USING PREDICTED FRAGMENTS EXPECTED FROM TYPE IIB RESTRICTION ENZYME DIGESTION

Abstract

Type IIB restriction endonucleases are site-specific endonucleases that cut both strands of double-stranded DNA upstream and downstream of their recognition sequences. These restriction enzymes have recognition sequences that are generally interrupted and range from 5-7 bases long. They produce DNA fragments which are of uniformly small in length, ranging from 21-33 base pairs (without cohesive ends). The fragments are generated from throughout the entire length of a genomic DNA, providing an excellent fractional representation of the genome. In this study, we simulated restriction enzyme digestions on 21 sequenced genomes of various *Drosophila* species using the predicted targets of 16 Type IIB restriction enzymes to effectively produce a large and arbitrary selection of loci from these genomes. The fragments were then used to compare organisms and to calculate the distance between genomes in pair-wise combination by counting the number of shared fragments between the two genomes. Phylogenetic trees were then generated for each enzyme using this distance measure and the consensus was calculated. The consensus tree obtained agrees well with the currently accepted tree for the *Drosophila* species. We conclude that multi-locus sub-genomic representation

combined with next generation sequencing, especially for individuals and species without previous genome characterization, can improve studies of comparative genomics and the building of accurate phylogenetic trees.

Introduction

Evolutionary relationships of species derived by comparing single orthologous genes or groups of genes can be negatively affected by potential horizontal gene transfers, incomplete lineage-sorting, introgression, and the unrecognized comparison of paralogous genes [1]. However, with the advent of the genomic era, it is now possible for researchers to use the complete genomes of fully sequenced organisms for building trees. Though such trees offer robust analysis, it becomes impractical to use traditional methods for constructing large-scale alignments and for generating trees from these alignments, mainly because of their large size and their highly heterogeneous nature [9]. As a result, there are now sophisticated methods that do not rely on alignment and are optimized for large scale data. These methods generally use vector representation of genes [20, 42] or features such as gene content [43-45], gene order [46, 47], intron positions [48], or protein domain structure [12, 49]. However, these new methods are often not very reliable and are used by few researchers worldwide. Even with the dramatic decrease in the cost of genome sequencing, it is still not attractive to sequence the genomes of those organisms that have little economic value, especially if their genomes are extremely large. On the other hand, the possibility of obtaining a large and representative set of fragments, instead of the whole genome sequence, can be economically feasible even for lesser known species and can provide a valuable alternative for many types of genomic scale studies, including phylogenomics.

Type IIB restriction endonucleases are site-specific endonucleases that cut both strands of double-stranded DNA upstream and downstream of their recognition sequences. These restriction enzymes have recognition sequences that are generally interrupted and range from 5-7 bases long. They produce DNA fragments which are of uniform length, ranging from 21-33 base pairs in length (without cohesive ends) [50]. The fragments are generated from throughout the entire length of genomic DNA, providing an excellent fractional representation of the genome. The fragments generated through this method can be used for various purposes including digital karyotyping [51], for pathogen identification by computational subtraction [52], and for genomic profiling to identify and quantitatively analyze genomic DNAs [53]. This method is particularly useful for unsequenced species and for phylogenetic study of the evolutionary relationships between organisms. In this study, we have tested this method *in silico* and shown that 13 different types of IIB restriction enzymes can be used to accurately reconstruct the phylogeny of 21 *Drosophila* species.

Material and methods

Obtaining datasets

Whole genome, nucleotide sequences for the 21 *Drosophila* species were downloaded from the FlyBase [54], NCBI databases and from Princeton University website[55].

Simulated restriction digestion

The PERL program “Phyper” was used to simulate restriction digestion for all the 16 Type IIB endonuclease enzymes and for processing the obtained fragments. This program generated a representative list of unique fragments for each genome for each enzyme separately

after removing all closely related fragments derived from non-identical members of paralogous gene families.

Fragment comparisons

The representative lists of fragments were then used with another PERL program "Phyppa" for comparative analyses. This program compares each fragment of a genome with every fragment of another genome in order to find identical fragments and similar fragments (fragments with up to 6 mismatches). A total of 210 such comparisons were done in order to generate the full list of shared fragments (identical fragments and similar fragments) for every pair of genomes.

Distance calculations

The number of shared fragments between a pair of genomes was then used to calculate the evolutionary distance by calculating the ratio of shared fragment to the total fragments and then taking the negative natural log of the ratio (Equation 1).

Equation 1

$$Distance = -\ln\left(\frac{Identical\ fragments + Similiar\ fragments}{Total\ fragments\ of\ both\ species}\right)$$

Building trees

Distance measures for all the pairwise comparisons for a particular enzyme were used to build trees using the *neighbor* program from the Phylip package [56]. A consensus tree was them produced by combining trees for all the enzymes with the *consensus* program from Phylip. The flowchart for the entire process is given in Figure 21.

Results

Datasets

The full nucleotide sequences for 21 *Drosophila* species were downloaded from various sources (Table 2). The genome size ranged from 137.82 mb for *D. simulans* to 235.52 mb for *D. willistoni*. *D. willistoni* had the lowest GC content of all (37.89%), and *D. pseudoobscura* had the highest GC content (45.43%). The 21 species of *Drosophila* used here included the subgenus *Sophophora* and the subgenus *Drosophila*. The *Sophophora* group was represented by *melanogaster*, *obscura* and *willistoni* and the *Drosophila* group was represented by *virilis*, *repleta* and *mojavensis*. Out of the 12 subgroups within the *melanogaster* group, 9 subgroups *viz., ananassae, montium, melanogaster, suzukii, takahashii, ficusphila, elegans, rhopaloa* and *eugracilis* were represented by 15 species. Of these, only 2 subgroups had multiple members within our data set, but both displayed a monophyletic arrangement within the final tree shown in Figure 22.

Type IIB restriction enzymes

Table 3 lists he 16 Type IIB restriction endonucleases that could potentially be used for simulating the restriction digestion of *Drosophila* genomes along with their recognition sites, frequency of cuts, and the size of fragment (blunt) that the enzymes leaves behind [52]. Unlike traditional Type II enzymes, Type IIB enzymes cleave on both sides of the recognition sequence (about 7-15 bases upstream and downstream, depending on enzymes) generating a fragment of uniform length. Also, the recognition site is usually split into two parts by some fixed number of random bases. They normally leave 2-3 base overhangs on the generated fragment.

Fragment analyses

The numbers of representative fragments obtained from each genome for each enzyme are listed in Table 4. The most frequent cutting enzymes such as *BslFI* had generally higher numbers of fragments within all genomes compared to other enzymes. Also, *D. pseudoobscura* and *D. persimilis* had relatively higher numbers of fragments compared to other genomes with most of the enzymes.

Conclusions

Fragment Analysis

Following fragment extraction, the original genomic sequences downloaded from various source databases were represented as a collection of fragments of uniform length. For each genome a total of 16 fragment sets were generated by using 16 different type IIB enzymes. The number of fragments generated by each genome was not closely related to the size of the genome but was related to its GC content. Most of the enzymes used in the analysis recognized a GC rich recognition site, which is reflected in the number of fragments generated with GC-rich genomes. The genomes that were GC-rich, such as those of *D. pseudoobscura* and *D. persimilis*, had higher numbers of fragments than other genomes. Similarly, the genomes that had lower GC content, such as *D. willistoni* and *D. grimshawi*, generated fewer fragments. The numbers of fragments generated by the enzymes were also dependent on the frequency of cut sites estimated for those enzymes in random sequence. Most enzymes predicted to be frequent cutters (e.g., *BslFI*) generated large number of fragments. Predicted rare cutters (e.g., *PsrI, PpiI, AloI, CspCI*) generated fewer fragments than other enzymes.

A comparison of fragments between genomes provided a list of fragments that were shared by those genomes. Closely related organisms are expected to share higher numbers of similar fragments (including identical fragments) compared to other distantly related genomes. Similar fragments are defined as those with 6 or fewer mismatches. The pair-wise distance matrices constructed using the similar fragments detected by each enzyme was used to estimate phylogenetic trees. The individual NJ trees obtained for each enzyme were largely consistent with the currently accepted relationships among the various *Drosophila* groups and subgroups, as were the single consensus tree obtained.

Phylogenetic tree

The placement within our tree of the 12 well-studied *Drosophila* species (*D. simulans, D. sechellia, D. melanogaster, D. erecta, D. ananassae, D. yakuba, D. pseudoobscura, D. persimilis, D. willistoni, D. mojavensis, D. virilis and D. grimshawi*) corresponds exactly to the currently accepted phylogeny [57-59]. Overall, the topology of our 21-species tree agrees precisely with those presented by Van der Linde and Houle, 2008 [60] and Yang et al. 2012 [61], except for the placement of the single species *D. eugracilis*, which clustered strongly with the melanogaster subgroup in our tree. These other studies covered different subsets of data and taxa and produced conflicting phylogenetic hypotheses specifically for *D. eugracilis.* In one of these studies (Yang et al. 2012), 7 of the 17 genes supported a subclade formed by (*D. ficusphila,* (*D. eugracilis,*(*D. elegans*, *D. rhopaloa*))) while the other genes switched the position of *D. eugracilis* to another subclade consisting of the *melanogaster* subgroup. Other previous studies also suggested *eugracilis* as a sister group to the *melanogaster* subgroup, as did our analysis [60, 62, 63]. Still other studies placed *D. eugracilis* near the *elegans* subgroup [64],or as a sister group to the *takahashii* and *suzukii* subgroups [65].

We conclude that our method of using multi-locus data obtained from small sub-genomic fragment sets provides good phylogenetic signal and produces a well resolved and well-supported species phylogeny. We note that our sub-genomic sampling method is analogous to previously described methods that use a different class of restriction enzymes to generate “RAD” markers for the comparative analysis of genomes at the population level [66]. We advocate using these approaches for accurate genome representation and combining them with next generation sequencing in order to facilitate comparative genomics and phylogenomic studies on individuals and species, even those lacking previous genetic characterization, at minimal cost.

Table 2. Various *Drosophila* species and source databases used for the analysis. The GC % for each genome was calculated using *infoseq* program from the EMBOSS package [67].

Genome	GC %	Size	Source
D. ananassae	42.56	230.99 mb	FlyBase
D. biarmipes	41.82	168.58 mb	NCBI
D. bipectinata	41.62	166.39 mb	NCBI
D. elegans	40.31	170.51 mb	NCBI
D. erecta	42.65	152.71 mb	FlyBase
D. eugracilis	40.90	156.31 mb	NCBI
D. ficusphila	41.93	151.04 mb	NCBI
D. grimshawi	38.84	200.46 mb	FlyBase
D. kikkawai	41.38	163.57 mb	NCBI
D. melanogaster	42.05	168.73 mb	FlyBase
D. mojavensis	40.22	193.82 mb	FlyBase
D. persimilis	45.29	188.37 mb	FlyBase
D. pseudoobscura	45.43	152.73 mb	FlyBase
D. rhopaloa	40.07	193.90 mb	NCBI
D. santomea	38.52	165.75 mb	Princeton University
D. sechellia	42.53	166.57 mb	FlyBase
D. simulans	43.06	137.82 mb	FlyBase
D. takahashii	40.01	181.00 mb	NCBI
D. virilis	40.80	206.02 mb	FlyBase
D. willistoni	37.89	235.51 mb	FlyBase
D. yakuba	42.43	165.69 mb	FlyBase

Table 3. List of enzymes used for the fragment generation from the 21 *Drosophila* species. Frequency indicates estimated distance between cut sites given a random sequence with all the four bases in equal probability and length refers to blunt tag length

Enzyme	**Recognition sequence**	**Frequency**	**Length**
AlfI	GCANNNNNNTGC	4096	32
AloI	GAACNNNNNNTCC	8192	27
BaeI	ACNNNNGTAYC	4096	28
BcgI	CGANNNNNNTGC	2048	32
BplI	GAGNNNNNCTC	4096	27
BsaXI	ACNNNNNCTCC	2048	27
BslFI	GGGAC	512	21
Bsp24I	GACNNNNNNTGG	2048	27
CjeI	CCANNNNNNGT	512	28
CjePI	CCANNNNNNNTC	512	27
CspCI	CAANNNNNGTGG	8192	33
FalI	AAGNNNNNCTT	4096	27
HaeIV	GAYNNNNNRTC	1024	27
Hin4I	GAYNNNNNVTC	512	27
PpiI	GAACNNNNNCTC	8192	27
PsrI	GAACNNNNNNTAC	8192	27

Table 4. List of fragments generated using 13 different Type IIB restriction enzymes for each of the 21 *Drosophila* genomes.

Genomes	*AlfI*	*AloI*	*BaeI*	*BcgI*	*BplI*	*BsaXI*	*BslFI*	*Bsp24I*	*CspCI*	*FalI*	*HaeIV*	*PpiI*	*PsrI*
D. ananassae	34804	11421	6151	51646	21457	52433	101183	46042	16405	38109	74174	11193	8344
D. biarmipes	41242	12667	6875	63518	22752	51248	109404	44554	18178	41284	75291	12177	10210
D. bipectinata	35642	10893	6616	51208	20363	50001	98937	45563	17131	39286	73197	10545	8622
D. elegans	43207	11314	6068	59905	18764	45496	93763	43259	18466	41866	75238	11027	9753
D. erecta	42781	10517	5914	60434	18119	43684	85735	40020	17793	31931	66412	9979	8677
D. eugracilis	36455	10170	5699	51988	18236	43177	86365	42020	17568	40795	72398	9682	8335
D. ficusphila	38374	11698	5338	60448	20161	47056	89928	39223	17489	37380	69222	11070	8868
D. grimshawi	49667	5891	5212	61420	17341	30379	58175	35658	16642	34409	64560	8062	6977
D. kikkawai	39192	10361	5516	54698	21908	50258	99784	44066	16846	40965	68593	10765	8126
D. melanogaster	39711	9908	6037	59203	16840	41168	81877	39221	17651	31350	68204	9243	8303
D. mojavensis	54782	6294	5234	64186	21048	33289	60708	36674	14774	33071	65210	9090	8012
D. persimilis	43327	10706	7567	59923	25287	53206	113002	48862	16329	31779	76473	12267	8940
D. pseudoobscura	43650	10461	7466	60237	25174	53269	111423	48990	16358	31417	74808	12175	8774
D. rhopaloa	36920	10920	6177	56203	18139	44894	93524	41357	17133	40153	76711	10442	9247
D. santomea	40344	9877	5957	56771	17044	41850	80010	38107	17037	32142	67070	9414	8378
D. sechellia	39876	10371	5808	59204	17430	42659	83936	39380	17276	31541	68359	9792	8289
D. simulans	38549	9815	5547	56820	16777	40735	79826	37436	16666	30304	64321	9148	7773
D. takahashii	37489	11463	5431	58887	19189	45240	91825	39992	26269	37277	74002	10801	8987
D. virrilis	58785	6943	5774	64912	18097	31951	66710	38679	15733	37692	65275	9290	8551
D. willistoni	34033	7083	6177	43299	15103	35578	70085	39996	17240	42202	77102	7941	9626
D. yakuba	42202	10300	6165	59442	17885	43748	83095	39920	18007	33024	69632	9887	8765

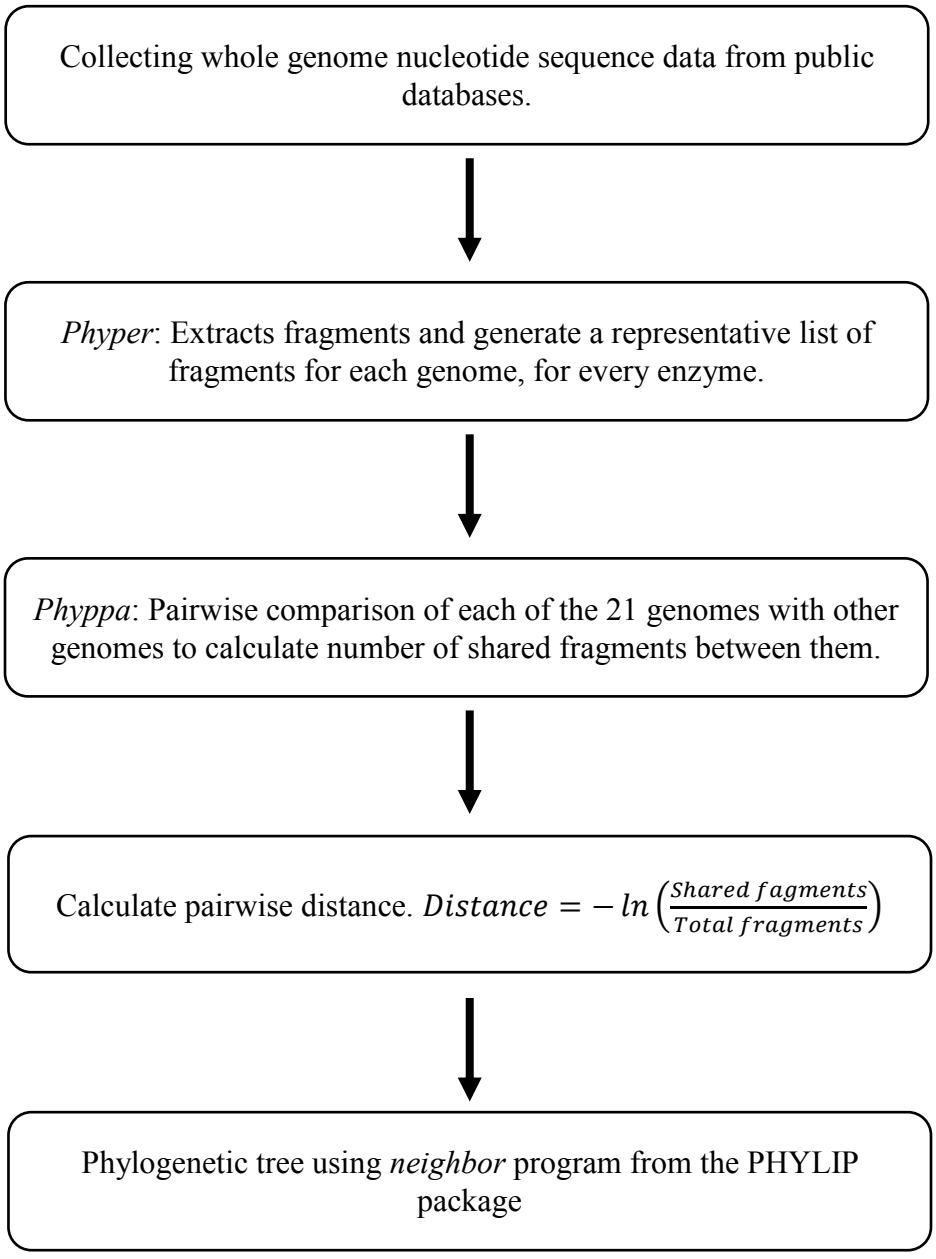

Figure 21. Workflow of the entire process for generating a phylogeny from the Type IIB fragments.

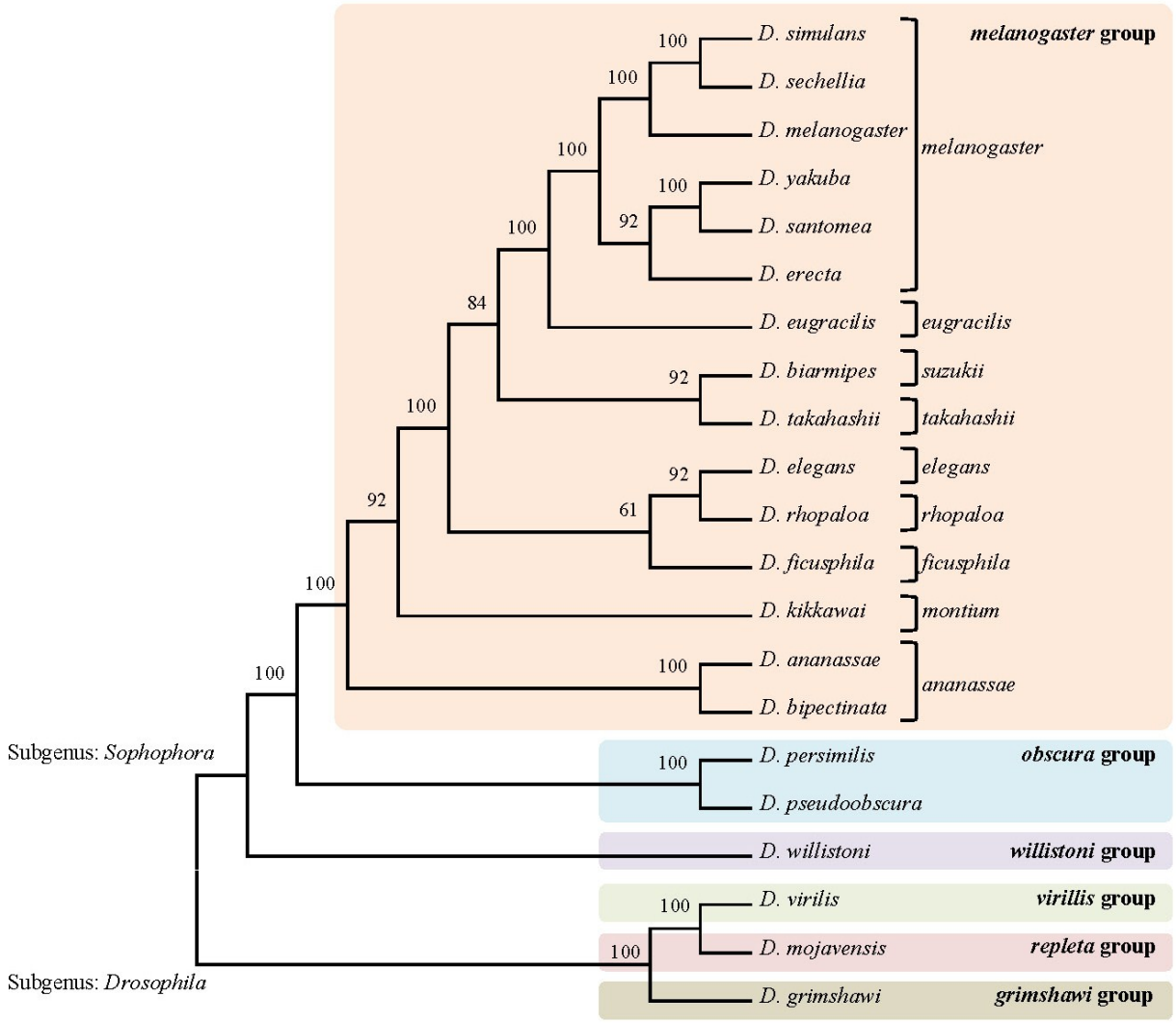

Figure 22. The consensus phylogenetic tree obtained by combining the trees obtained for each of the 13 enzymes. The phylogenetic tree for each enzyme was calculated by extracting the corresponding fragments and then counting the number of shared fragment between every pair of species. The branch support values represent the percentage agreement over 13 enzymes for a given branch.

CHAPTER 4

A SURVEY OF WELL CONSERVED FAMILIES OF C2H2 ZINC-FINGER GENES IN *DAPHNIA PULEX.*

Abstract

A recent comparative genomic analysis tentatively identified roughly 40 orthologous groups of C2H2 Zinc-finger proteins that are well conserved in "bilaterians" (*i.e.* worms, flies, and humans). Here we extend that analysis to include a second arthropod genome from the crustacean, *Daphnia pulex.* Most of the 40 orthologous groups of C2H2 zinc-finger proteins are represented by just one or two proteins within each of the previously surveyed species. Likewise, *Daphnia* were found to possess a similar number of orthologs for all of these small orthology groups. In contrast, the number of Sp/KLF homologs tends to be greater and to vary between species. Like the corresponding mammalian Sp/KLF proteins, most of the *Drosophila* and *Daphnia* homologs can be placed into one of three sub-groups: Class I-III. *Daphnia* were found to have three Class I proteins that roughly correspond to their *Drosophila* counterparts, dSP1, btd, CG5669, and three Class II proteins that roughly correspond to Luna, CG12029, CG9895. However, *Daphnia* have four additional KLF-Class II proteins that are most similar to the vertebrate KLF1/2/4 proteins, a subset not found in *Drosophila*. Two of these four proteins are encoded by genes linked in tandem. *Daphnia* also have three KLF-Class III members, one more

than *Drosophila*. One of these is a likely Bteb2 homolog, while the other two correspond to Cabot and KLF13, a vertebrate homolog of Cabot. Consistent with their likely roles as fundamental determinants of bilaterian form and function, most of the 40 groups of C2H2 zinc-finger proteins are conserved in kind and number in *Daphnia*. However, the KLF family includes several additional genes that are most similar to genes present in vertebrates but missing in *Drosophila*.

Introduction

Zinc-finger proteins (ZFP) represent the largest family of DNA-binding transcription factors in eukaryotes. Although many proteins are predicted to contain single zinc-finger domains, two zinc fingers in close proximity appear to be required for high-affinity DNA binding. There are many diverse subfamilies of zinc-finger proteins in eukaryotes, but the most numerous are the Kruppel-type C2H2 ZFPs. Many of these proteins contain either multiple tandem pairs of zinc-fingers or tandem arrays of three or more zinc-fingers. As transcription factors, they participate generally in the fundamental mechanism of gene expression. However, they usually also play more specific roles in a wide variety of regulated biological processes, including signal transduction, cell growth, differentiation, and development. As part of our collaborative role in annotating the draft genome assembly v1.1 of the *Daphnia pulex* genome [68], we focused our attention on a subset of roughly 40 orthologous groups of C2H2 ZFPs identified in a recent comparative genomic analysis to be well conserved in “bilaterians” (i.e. worms, flies, and humans)[69]. While many of these were known or likely DNA-binding transcription factor encoding proteins with tandem arrays of zinc-fingers (e.g. Zif268, MTF1, TFIIIA, SP1 and KLF), others had only a single zinc finger (e.g. SAP61, SAP62, and Kin17),

which generally lacks a DNA-binding function [69]. Also included are some genes that encode multiple split pairs of C2H2 zinc-fingers, like Disco

Materials and methods

Identification of orthologs in D. pulex

Previously identified orthologs [69] present in the common ancestor of the bilaterians *Homo sapiens*, *Drosophila melanogaster,* and *Caenorhabditis elegans* were used as a focus for the present study. Protein sequences from each of the 3 different species belonging to 39 different classes of C2H2 zinc finger proteins were collected. Each of these sequences was used in turn as a query in a BLAST search against the v1.1 gene model annotations of the draft genome assembly of *D. pulex* to detect homologous protein sequences. High-scoring *Daphnia* sequences were examined to ensure a good overall match, and then used in a reciprocal BLAST against *Homo sapiens*, *Drosophila melanogaster* and *Caenorhabditis elegans.* Only those sequences that detected members of the same family represented by the original query sequence were retained as putative orthologs for those families. Other known C2H2 zinc finger binding genes were also used to search for any new families common to these bilaterians.

After the initial reciprocal BLAST approach, Hidden–Markov model (HMM) searches were conducted using the HMM profiles obtained from TreeFam for each of these 39 gene families to search for additional gene members for the families.. *Daphnia pulex* protein predictions containing all models were used to search with HMM profiles using HMMER 4.0 search. Only domains with an E-value < 0.1 were accepted, and any identified zinc-finger gene was further manually inspected for family characteristics. BLAST search was performed against the non-redundant protein dataset at NCBI to confirm its family association. Only genes that

appeared to be substantially complete and those that had approximately the same number of zinc fingers as other members of the proposed family were considered to be unambiguous members of that family. This approach identified additional homologs of Odd-skipped and Snail families, and validated all the genes that were identified with reciprocal BLAST approach.

Alignments and phylogenetic analyses

Daphnia pulex homologs identified as above were combined with their respective family members from the other three bilaterians (*Homo sapiens*, *Drosophila melanogaster* and *Caenorhabditis elegans*; (see Table 5, 6, 7 and 8) and used to create family-specific and/or multiple family alignments using the *Muscle* program (version 28) with 16 iterations and a standard Clustalw weighting scheme (gap opening extension, closing and separation penalty of 10, 0.2, 4 and 1 respectively) [70]. The obtained alignment was then trimmed and converted to suitable format using trimAI (version 1.2)[71]. Phylogenetic trees were generated using Bayesian inference (*MrBayes*; version 23.2)[72] using WAG amino acid substitution matrix[73], empirically estimated amino acid frequencies plus gamma distribution of eight categories ($WAG+F+\Gamma_8$). Successive runs were executed for a fixed number of generations with a sampling frequency of 100 and a burn-in parameter of 200. Runs were extended in each case until a convergence value of less than 0.03 was achieved. Because the multi-family trees each contained only an exclusive subset of the 40 total C2H2 zinc-finger families, internal branch patterns and statistics could be misleading with respect to the degree of relatedness between families. Hence, internal branches indicating a specific relationship between specific families within multi-family trees were collapsed into polytomies for presentation.

Results

Previously, 39 families of C2H2 ZFP were determined to be present in the common ancestor of bilaterians based on a survey of three organisms: *Homo sapiens*, *Drosophila melanogaster* and *Caenorhabditis elegans*. Although the work described below necessitated the addition of three more conserved families to this collection, two pairs of the original 39 families could also be reasonably combined into single families. Hence, we created a reorganized summary list of 40 orthologous groups of C2H2 ZFP. The resulting list of family members and their accession IDs are provided in Table 5, 6, 7 and 8, while an efficient numerical summary is provided in Table 9. A brief description of the known or proposed function(s) and structural organization for each of these families is provided in series below.

A relatively rigorous assessment of homology/orthology is provided by a set of carefully constructed phylogenetic trees (see Figure 23, 24, 25, 26, 27, 28, 29, 30, 31 and 32). These trees also serve to summarize various evolutionary events of interest, such as presumptive gene losses, duplications, and lineage-specific expansions. Larger C2H2 Zinc-finger families (i.e. Sp and KLF) are presented within separate trees (Figures 23 and 24), while most of the remaining families are displayed in a short series of multi-family summary trees (Figure 25, 26, 27, 28, 29 and 30). The latter provide additional evidence that member gene clusters represent distinct families of correctly identified homologs and putative orthologs. Evidence for the expression of almost all of these genes at the RNA level was obtained using a NimbleGen tiling array [68, 74] and the EST data available at JGI portal [75]. Only three genes lack any evidence of expression to date (KLF2D, ZFam9, and FEZL).

Sp/KLF, the Largest Family of C2H2 ZFP (Figures 23 & 24)

The Sp/KLF proteins are DNA-binding transcription factors each containing 3 zinc-fingers. Although Sp and KLF factors are closely related, they occupy distinct branches on a combined evolutionary tree. To facilitate presentation, however, it was most convenient to separate these families into two distinct trees, each rooted with two members from the other family (Figures 23 and 24). Although frequently described as simple transcription factors (especially Sp), many are known to interact with particular sets of chromatin remodeling complexes to facilitate transcriptional activation or repression [76]. In vertebrates, Sp/KLF proteins are grouped into three classes that tend to correlate with the type of chromatin remodeling complexes they utilize. Conveniently, most of the invertebrate proteins can also be assigned to one of these classes [77]. dSP1, btd, and CG5669 correspond to three distinct subsets of Sp-Class I proteins Sp7/8, Sp5, and Sp1/2/3/4, respectively (Figure 23). Luna, CG12029, and CG9895 correspond to the KLF-Class II proteins KLF6/7, KLF5, and KLF3/8/12, respectively (Figure 24). Bteb2 and Cabot correspond to KLF-Class III proteins KLF15 and KLF9/10/11/13/14/16, respectively (Figure 24). The two remaining fly proteins (CG3065 and hkb) are difficult to place unambiguously. The former shows roughly equal similarity to members of both the Sp and KLF subfamilies, while the latter appears to be a highly diverged and relatively unique member of the Sp/KLF family. Information concerning the functions of many of these proteins are available in a variety of organisms [78, 79]. In vertebrates, some Sp/KLF proteins produce early embryonic lethals when mutated (Sp1, KLF5), some are known to affect behavior and/or the development of structures within the brain (Sp4, Sp8, KLF9), and others affect development of the blood cells (Sp3, KLF1, KLF3), goblet cells in the colon, or bone cells (Sp7). In *Drosophila*, buttonhead (btd) is important for the development of head

structures, and both btd and dSp1 affect development of the mechano-sensory organs [80]. Cabot (cbt) also affects sensory organ development, as well as dorsal closure [81]. Perturbations of luna expression via RNA interference or over-expression during early *Drosophila* embryogenesis leads to developmental arrest at different embryonic stages [82]. *Daphnia* appear to contain three Sp homologs (Sp8, Sp5, Sp4), each of which corresponds to a specific counterpart (SP1, btd, CG5669) in *Drosophila*. In contrast, the KLF gene family in *Daphnia* includes 10 distinct genes, 4 more than the complement in *Drosophila*. Dp-KLF3, Dp-Luna, Dp-Cabot, and Dp-Bteb2 appear to correspond to the fly genes CG9895, Luna, Cabot, and Bteb2, respectively. No clear homologs of KLF17 or CG12029 are apparent in *Daphnia*. In contrast, six KLF genes in *Daphnia* with no direct homologs in *Drosophila* seem to correspond to two subfamilies of KLF found in the vertebrate genome. Five of these, Dp-KLF1A through Dp-KLF1E, may represent a species specific gene expansion that roughly corresponds to an independent expansion in humans that includes KLF1/2/4. The sixth may be a lone homolog of the vertebrate expansion that includes KLF 9/13/14/16.

C2H2 ZFP resistant to deletion/expansion

Some C2H2 ZFP exist as single copy family members in all 4 genomes. Hence, these genes appear to be relatively resistant to deletion or expansion over evolutionary time. ZNF277 is one of several examples (Figure 25). This gene encodes a protein with five C2H2 zinc fingers. The function of this gene is not well understood. In humans, this gene is expressed in early embryonic tissues, parathyroid adenoma, and chronic lymphocytic leukemia suggesting that this gene might be involved in differentiation [83].

ZFAM5 is also known as ZNF622 or Zinc finger related protein 9 (ZRP9) and is highly conserved almost universally in eukaryotes. Most homologs have 4 zinc-fingers (Figure 25).

ZFAM5 was originally identified in mouse as a cellular MPK38 serine/threonine kinase binding protein that may be involved in early T cell activation and embryonic development [84]. In humans, this gene is responsible for interaction with the ubiquitously expressed MYB-B transcriptional regulator. The role of this gene in other organisms is not well understood [85].

ZFAM6 is also known as Zinc finger Matrin Type 2 (ZMAT2). It is a highly conserved gene present in most eukaryotes. Little is known about its function. A single homolog of this gene was also found in *Daphnia* (Figure 25). All family members possess one U1-like zinc finger. This type of C2H2 zinc finger is also present in the protein matrin, the U1 small nuclear ribonucleoprotein C, and other RNA-binding proteins.

ZFAM7 is also known as Zinc Finger 598 (ZNF598) in vertebrates. This gene is highly conserved in most eukaryotes and is present as a single homolog in most species, including *Daphnia* (Figure 25). The gene has five C2H2 zinc fingers. The function of this gene is largely unknown.

SAP61 (Splicosome Associated Protein 61) is also known as Splicosome factor 3a subunit 3 (SF3a3), while SAP62 (Splicosome Associated Protein 62) is also known as Splicosome factor 3a subunit 2 (SF3a2). These sequences do not cluster within a phylogenetic tree of zinc finger genes (Figure 25). However, since they share common function by virtue of being subunits of the same protein complex involved in RNA splicing, we described them together. A single homolog for these genes is present in almost all species of eukaryotes, including *Daphnia*. Both have a highly conserved U1-like zinc finger that is typical for RNA binding proteins. These are essential proteins, required for the formation of SF3a and functional U2 snRNP. Together with SF3b, SF3a binds to the 12S U2 snRNP, which contains a common core of seven Sm proteins and the U2-specific proteins U2-A and U2-B [86, 87].

KIN17 is present in almost all eukaryotic organisms. *Daphnia* have a single homolog for this gene. KIN17 includes a single highly conserved U1-like zinc finger and is likely to be involved in cellular response to DNA damage, gene expression, and DNA replication. The KIN17 protein shares sequence homology with bacterial RecA protein over 40 residues near the c terminus. KIN 17 is ubiquitously expressed in mammals at low levels, but is up regulated after exposure to UV and ionizing radiation. KIN17 binds to DNA targets found in hot spots of illegitimate recombination [88-90].

A single homolog of ZNF207 is present in almost all vertebrates and invertebrates. *Daphnia* also has a single homolog (Figure 28). The N terminus of this protein contains 2 C2H2 type zinc fingers. This gene is expressed ubiquitously in humans [91], but the exact function is still not clear.

C2H2 ZFP with expansions in organisms other than Daphnia

The *Daphnia* genome appears to deploy a relatively efficient set of well conserved C2H2 ZFP because many C2H2 subfamilies have undergone lineage specific expansions in other genomes. In fact, 12 of the 40 well conserved families considered here show expansions in flies or humans, but not in *Daphnia*. For instance, Zfam1 (Figure 26) codes for a small peptide of 70 to 80 residues that contains one C2H2 type zinc finger. This gene is conserved in chordates and insects. Lineage specific independent duplications have generated 2 homologs in *Drosophila* and 2 in *C. elegans*. *Daphnia* have just one homolog for this gene that clusters with the *Drosophila* homolog. The function of these genes is not well understood.

Humans have three ZFAM2/BCL11 homologs: Bcl11A, Bcl11B, and ZNF342/Zfp296. BCL11A has 5 zinc-fingers, and is a homolog of the murine gene Evi9. Evi9 was found to be deregulated in mouse myeloid leukemias induced by proviral integration. Hence Evi9 has

characteristics of a dominant oncogene. Human EVI9/BCL11A is expressed in CD341 myeloid precursors. BCL11A is known to be involved in both Hodgkins and non-Hodgkins B-cell lymphoma [92, 93]. BCL11A acts as a proto-oncogene for B-cell lymphoma, as a recessive oncogene for T-cell lymphoma, and is apparently required for the expression of some globin genes [94]. Bcl11B also appears to act like a recessive oncogene for T-cells [95]. The single *Daphnia* homolog appears closely related to the *Drosophila* version (Figure 26). Apparently, duplication in mammals led to 2 or three versions of this gene, two of which became key regulators in hematopoetic lineages, while the third appears to function in the nervous system. ZNF342 has been indirectly implicated in the suppression of gliomas.

ZFAM4 is also known as Rotund and Squeeze in *Drosophila*, and Lin-29 (abnormal cell LINeage family member 29) in *C. elegans*. There appears to be 3 homologs for this family in humans and one additional homolog in *Drosophila*. Most genes in this family have 5 zinc fingers while two human genes (ZNF384 and ZNF362) and one *Drosophila* gene (CG2052) have an additional zinc finger. The *Daphnia* homolog clusters with *Drosophila* Rotund and squeeze (Figure 26). Roughened eye (Roe) is a part of the rotund gene generated by using a different promoter. They both share the C-terminal region and zinc finger domain but differ in their N-terminal regions. Roe appears to have a role in eye development in the embryos [96]. The *C. elegans* gene *lin-29* is required for terminal differentiation of the lateral hypodermal seam cells during the larval-to-adult molt and proper vulva morphogenesis. CIZ (CAS interacting Protein or ZNF384) is a nucleocytoplasmic shuttling protein that binds to CAS elements found in promoters of matrix metalloproteinases (MMPs) genes that produce enzymes used to degrade the extracellular matrix proteins [97].

ZFAM11 is also known as KCMF (potassium channel modulatory factor). It is present in most vertebrates and invertebrates. *Daphnia*, like most vertebrates, has a single copy. *C. elegans* too has a single member for this family which suggests that the gene has been duplicated specifically in flies/insects (Figure 26). All of these genes have a highly conserved region which includes one ZZ type zinc finger and one C2H2 type zinc finger. The ZZ motif is known to bind two zinc ions and most likely participates in ligand binding or molecular scaffolding. In vertebrates, KCMF1 is shown to have intrinsic E3 ubiquitin ligase activity. Studies indicated that KCMF1 is involved in regulating growth modulators [98]. The function of KCMF1 homologs in *Drosophila* and worms is poorly understood.

Fez typically has 6 zinc-fingers, and is a likely ortholog of the human genes ZNF-312 and 312-like. The forebrain expression pattern for this gene was first described in zebrafish, where there is also a second homolog known as Fezl. Fez expression is first detected in the anterior presumptive neuroectoderm of zebrafish during epiboly. Expression becomes focused in the rostral forebrain region during somitogenesis. By 24hrs, expression is largely restricted to the telencephalon and anterior/ventral region of the diencephalon. Hence Fez is an early marker of anterior neuroectoderm and appears to regulate forebrain development [99]. In mammals, these proteins appear to regulate olfactory-bulb development and neuronal differentiation in the cortex [100, 101]. Double knockouts indicate that together FEZ and FEZL play a role in rostral brain patterning in mouse. *Drosophila* and *Daphnia* appear to have only one homolog of Fez (Figure 26). There is little information about the role of the Fez protein in these organisms.

The zinc-finger E-box binding (ZEB) homeobox genes (Figure 27) were previously described as two separate families in chordates (ZEB1/ZFHX1A/ZFH1 and ZEB2/ZFHX1B/ZFH2) or as the zinc finger axon guidance gene (ZAG1) in *C. elegans*. The gene

is conserved in most bilaterians and usually has a homeodomain flanked by two separate, highly conserved zinc-finger clusters. Most have 6 C2H2 type zinc fingers present as triplets distributed over the length of the gene. The E-box-like target sites overlap with those bound by the Snail family of zinc-finger proteins. ZEB proteins can repress target- gene transcription by recruiting the CtBP (C-terminal-binding protein) co-repressor, which is a component of the larger repressor complex containing HDAC (histone deacetylase) and PcG (polycomb group proteins) [102]. ZEB1 and ZEB2 in humans are expressed in several tissues including muscle and CNS. They are also expressed in T lymphocytes and during skeletal differentiation. They are mediators of epithelial dedifferentiation in mammals through the down-regulation of E-cadherin expression [103]. ZEB2, also known as Smad Interacting Protein 1 (SIP1), is over expressed in cancer cells, causing loss of cell polarity and facilitating migratory and invasive behavior. SIP1 is also involved in the development of the neural-crest, the central nervous system, the septum of the heart, and establishment of the midline [104]. Mutations in SIP1 cause Mowat-Wilson Syndrome, a mental retardation syndrome in humans [105, 106]. In *Drosophila*, ZFH1 is a transcriptional repressor that regulates differentiation of muscle and gonadal cells, but is also expressed in the CNS [107]. ZAG-1 in *C. elegans* also acts as a repressor that regulates multiple, discrete neuron-specific aspects of terminal differentiation, including cell migration, axonal development, and gene expression [108]. *Daphnia* have a single homolog.

ZFHX genes encode zinc-finger homeobox containing proteins previously described as two separate families (ZFH3 and ZFH4) in bilaterians. In vertebrates this gene appears to have undergone duplications generating 2 or more additional homologs (Figure 27). In humans, there are 3 homologs (ZFHX2, ZFHX3 and ZFHX4). Family members usually contain 8 or more C2H2 zinc fingers distributed throughout the gene. The *Daphnia* homolog has 11 C2H2 type

zinc fingers. ZFHX genes are thought to be important regulators of neuronal differentiation [109]. Like most homeotic genes, these genes are also involved in embryonic morphogenesis. ZFHX3, also known as AT motif binding factor 1 (ATBF1), inhibits cell growth and differentiation and may play a role in malignant transformation. It has been shown that it is a potential tumor suppressor gene that represses alpha-fetoprotein (AFP) whose altered expression may lead to development of carcinoma in various tissues [110]. ZFHX4 expression is important for neuronal and muscle differentiation, and in rats it has been shown to be involved in neural cell maturation [111]. The *Drosophila* homolog ZFH2 is involved in establishing proximal-distal domains in the developing wing disc [112]

Spalt-like (SALL) proteins have a variable number of zinc-fingers: the worm homolog has 6, flies and *Daphnia* have 7, and chicken/human have 7 or 9, depending on the homolog. The two fly genes, Spalt-major (SALM) and Spalt-related (SALR), appear to have duplicated independently from the ancestral gene that also gave rise to the four human homologs (Figure 27). SAL in flies is required for proper development of the trachea, for vein patterning in wing imaginal discs, and for bristle formation in the thorax. In the latter case, SAL acts through regulation of pro-neural gene expression [113]. Nervous system expression is a well-conserved aspect of SAL gene function from worms to man. Mutations in the worm homolog (SEM4) affect development of neurons and sensory organs, while mutations in a human homolog (SALL1) result in sensorineural hearing loss and mental retardation (but also anal, genital, and limb malformation). Flies lacking both SALM and SALR are also deaf, with limb and genital malformations potentially analogous to those in humans [114].

ZEP homologs (Figure 29) are referred to as Schnurri (Shn) in flies and SMA9 in worms. There are three or more homologs in vertebrates, but just one in worms, flies, and *Daphnia*,

suggesting a lineage-specific expansion exclusive to vertebrates. ZEP proteins generally have five C2H2 zinc-fingers divided into two pairs and a solitary medial finger (missing in some homologs). The worm homologs have an additional C-terminal zinc finger pair with little similarity to the other family members, implying a unique function. The C terminus of this protein is also unique to worm [115]. ZEP/Shn/SMA9 homologs are involved in BMP signaling. Bone morphogenetic proteins (BMPs) are members of the transforming growth factor β (TGF β) family that regulate various biological process including embryonic axes, cell fate determination, proliferation and apoptosis in both invertebrate and vertebrate model systems. In mouse, Shn-2 is required for efficient transcription of *PPARγ2*, which in turn drives the expression of several genes involved in adipocyte differentiation [116]. Shn3 in mouse is a transcriptional regulator of Runx2, which in turn activates several osteoblast differentiation genes. In humans Shn3 is involved in T-cell proliferation, cytokine production, effector function, and inflammatory response [117]. In worms, SMA9/SMAD affects body size regulation and male tail patterning in worms [118]. In *Drosophila*, Shn binds to SMAD to form the repression complex controlling brinker (Brk), which is a transcriptional repressor of the Dpp gene. Dpp is involved in anterior-posterior patterning and cell proliferation in the wing blade [119].

The Insulinoma Associated gene (IA1) of vertebrates is called Nerfin in flies and Egg laying 46 (EGL46) in worms. *Daphnia* and C. elegans appear to have just one homolog with two conserved zinc-fingers. This gene appears to have undergone independent duplications in the human and fly lineages, giving rise to two paralogs in each (Figure 29). IA homologs are involved in various aspects of neuronal differentiation including cell fate specification, axon guidance decisions and cell migration. In humans IA1 promotes pancreatic and intestinal

endocrine cell development [120]. Recent reports for mice and zebra fish imply that its role in neurogenesis is conserved across vertebrates as well as invertebrates [121, 122].

Hamlet is also called PR domain zinc finger protein 16 (PRDM16) or ecotropic virus integration site 1 (EVI-1) homolog in vertebrates, Hamlet in *Drosophila* and Egg laying 43 (EGL43) in *C. elegans*. *Daphnia* has one homolog. Independent duplications in insects and vertebrates appear to have generated two paralogs each in their respective clades (Figure 29). All homologs contain an N-terminal PR (PRD1-BF1-RIZ1) homology domain followed by a group of six zinc fingers and a group of three additional ZFs at the C-terminus. In *Drosophila*, Hamlet functions as a binary genetic switch specifically affecting the dendritic branching structure of external sensory (ES) neurons in the peripheral nervous system [123]. In C. elegans, egl-43 encodes two transcription factors that act to control HSN migration and phasmid neuron development, presumably by regulating other genes that function directly in these processes [124]. The murine homolog Evi-1 was obtained from a common site of viral integration in murine myeloid leukemia. The human homolog MDS1/EVI1 is transcriptionally activated in diseases like acute myeloid leukemia (AML) and myelodysplastic syndrome (MDS) that are associated with several recurrent chromosomal aberrations. Recently, another homolog of HAM called *MEL1* (*MDS1/EVI1*-like gene 1) was identified as a member of the *EVI1* gene family and also as a PR domain member (PRDM16), all of which are implicated in neural development [125]. The disruption of the PR domain of this gene can cause leukemia. A partial disruption of the *Mds1/Evi1* locus in mouse leads to multiple defects causing mid-gestation lethality, including defects of hypocellularity in the neuroectoderm and a failure of peripheral nerve formation [126].

The stripe gene (Sr) in *Drosophila* functions in the epidermis to facilitate cellular recognition of myotubules (Figure 30). Hence, stripe mutants exhibit a disruption in myotubule

patterning. Stripe is a member of the EGR (early growth response) family of transcription factors. The Egr transcription factors are rapidly induced by diverse extracellular physiological/chemical stimuli within the vertebrate nervous system. These proteins possess 3 zinc fingers. Another member of this family, Krox20, is known to be involved in development of the hindbrain and neural crest in mammals. Analysis of mouse knockouts has demonstrated that Egr2/Krox-20 is important for hindbrain segmentation and development, peripheral nervous system (PNS) myelination, and Schwann cell differentiation [127]. Krox20 expression correlates with the onset of myelination in the PNS. Egr-1 and egr-3 are also both implicated in learning and memory [48]. EGR-1 was shown to be induced in specific subregions of the brain during retrieval of fear memories. Knockout mice further showed that egr-1 was essential for the transition from short- to long-term plasticity and for the formation of long-term memories. In T-cells of the immune system, egr-3 and egr-4 work together with NF-kappaB to control transcription of genes encoding inflammatory cytokines. Egr-2 and Egr-3 can also inhibit T cell activation [128]. The egr genes are distantly related to the Wilm's tumor (WT) gene. The latter, like the distantly related klumpfuss gene in *Drosophila*, has 4 zinc fingers rather than 4. The single *Daphnia* homolog is seen to be most similar to that in *Drosophila* (Figure 30).

Disco homologs have 4 to 6 zinc fingers in a paired arrangement. In humans, Basonuclin 1 and 2 (BNC1 & 2) correspond to the Disco and Disco-r genes present in flies (Figure 30). In humans basonuclin is expressed in keratinocytes, germ cells, cornea, and lens epithelia. BNC2 mRNA is abundant in cell types that possess BNC1 but is also found in tissues that lack BCN1, such as kidney, intestine, and uterus [129]. In keratinocytes, BNC maintains proliferative capacity and prevents their terminal differentiation [130]. Recently, it has been suggested that BNC2 is also involved in mRNA export, nonsense-mediated decay, and/or polyadenylation

[129]. Both BNC's activate the expression of rRNA genes. Disconnected *(Disco)* and disco-related *(Disco-r)* are two functionally redundant, neighboring genes localized on the fly X chromosome that may act in combination with the homeotic genes *deformed (dfd)* and Sex Combs Reduced *(SCR)* to specify gnathal structures in *Drosophila* [131, 132]. The ancestral Disco gene appears to have undergone independent duplication events in the human and fly lineages. *Daphnia* appear to have just one homolog.

C2H2 ZNF homologs duplicated in Daphnia but not Drosophila

Lineage Specific duplications of well conserved C2H2 ZFP in *Daphnia* appear to be rare. GPS homologs (GFI/PAG/Sens) generally have 6 tandem zinc-fingers (Figure 29). The C elegans homolog (PAG3) has only 5 fingers (missing the first one). The *Drosophila* homolog (SENS) has only 4 fingers (missing the first two). The two *Daphnia* homologs cluster together and thus appear to be a recent duplication independent of that producing the two GFI homologs in humans. One *Daphnia* homolog (GPSa) has a unique 5aa insertion between fingers 1 and 2. GPS proteins are involved in hematopoesis and neurogenesis. The hematopoetic functions of vertebrate GFI-1 and GFI-1B (6 fingers) appear to be exchangeable, but distinct, due to differential cell-type specific expression. They differ in their ability to facilitate late maturation of inner ear neurons [133]. While generally known as transcriptional repressors, some act as transcriptional activators and/or conditional repressors (like Sens). The single worm homolog serves to repress touch neuron gene expression in interneuron cells [134]. The *Drosophila* Senseless gene is required for normal sensory organ development [135]. The two *Daphnia* GPS genes are closely linked on the same scaffold. A domain required for transcriptional repression, the SNAG domain, is found only in vertebrate GFI proteins to date; hence no SNAG domain is seen in the *Daphnia* homologs [136].

PRDM/Blimp proteins are Putative Positive Regulatory Domain/B-lymphocyte induced maturation proteins. These proteins have 5 fingers, but the 5th finger is relatively poorly conserved and has a C2HC structure. A single PRDM1/Blimp gene is found in humans, one in flies, and one in worms (Figure 29). Blimp1 expression in the tracheal system of *Drosophila* embryos was found to be important for the development of this tissue. Blimp1 is also induced by ecdysone, and reduced Blimp1 expression results in prepupal lethality [137]. Blimp1 is expressed in many other tissues of *Drosophila*. Blimp is similarly expressed in many different tissues in vertebrates, where it is known to play important roles in embryogenesis, germ cell determination, specification in nerve and muscle cells, linage determination in epidermis, and B-cell maturation [138-140]. There appear to be two Blimp homologs in *Daphnia*. However, in one, the 5th finger is missing, and the third finger has two serines replacing the two cysteines of the finger. Consequently, the function of this finger (and the entire protein) may have changed substantially.

C2H2 ZNF absent from one or more organisms

Zinc finger X-linked duplicated (ZXD) is a newly described C2H2 zinc finger family in bilaterians present in most chordates and has undergone duplication specifically in mammals. Among the bilaterians, humans and other mammals have 3 homologs for this family while nematodes, water fleas, sea urchins, chicken and frog all have one homolog including *Daphnia* (not shown). Interestingly, no homologs have been detected in insect lineage that has a sequenced genome and in *C. elegans*. Zinc finger X-linked duplicated family member C (ZXDC) along with its binding partner ZXDA, forms a complex that interacts with CIITA and regulates MHC II transcription [141, 142]. The function of the other paralog ZXDB as well as the homolog in *C. elegans* is little understood.

CTCF and CTCFL (Figure 28) have 11 zinc fingers arranged in tandem. They act in part as “enhancer blockers” in vertebrates by binding to insulator elements. In flies, dCTCF binds to the Fab-8 insulator element between iab-7 and iab-8 [143]. Mammalian CTCF’s are also involved in reading gene imprinting marks (i.e. Boris) at a high fraction of imprinted genes [144]. Recently, CTCF (along with YY1) has also been implicated in the global repression mechanism known as X-inactivation [145].There are two human homologs, CTCF and CTCFL, but only one in *Daphnia* and *Drosophila*, and none in worms.

Yin Yang 1 (YY1) generally has 4 zinc-fingers. Recent phylogenetic analyses proposed that the YY1 gene has undergone independent duplication events in different lineages through retro-transposition. Two duplication events in placental mammals are believed to have given rise to the YY1, YY2, and REX1 (Reduced EXpression). A similar duplication event in flies produced the pleiohomeotic (Pho) and Pho-like genes [146]. The *Daphnia* Pho gene clusters with the *Drosophila* Pho (Figure 28). YY1 acts to activate or repress transcription in different contexts. In mammals, this gene appears to play multiple roles, including induction and patterning of the embryonic nervous system, differentiation within blood cell lineages, cell-cycle control, cell proliferation, differentiation, and apoptosis, DNA synthesis and packaging, and X-inactivation [112]. The fly homologs, Pho and phol, are classified as PcG (poly comb group) proteins that bind to PREs (PcG response elements that regulate homeotic genes). Pho and Phol act redundantly to repress homeotic gene expression in imaginal discs of the fly [147].

Hindsight (Hnt) in *Drosophila* is a homolog of Ras-Responsive Element Binding protein 1 (RREB1) in vertebrates. No homolog has been detected in worms, but *Daphnia* appears to have a single homolog (Figure 32). The number of zinc fingers varies from species to species: 15 in humans, but only 10 in *Daphnia*. The *Pebbled (peb)* gene encodes the Hindsight protein,

involved in morphogenetic processes and is expressed in several kinds of epithelial cells during development including extra embryonic amnioserosa, midgut, trachea, and the photoreceptor cells of the developing adult retina. In the amnioserosa, Hnt is required for embryonic germ band retraction and embryonic dorsal closure [148]. In tracheal development, it is required for the maintenance of epithelial integrity and assembly of apical extracellular structures known as taenidia [149]. During eye development, it is required for the accumulation of F actin in the apical tip of photoreceptor precursor cells in the ommatidial clusters, as well as in the developing rhabdomere during the pupal period [150]. HNT expression is also essential for maintaining epithelial integrity for amnioserosa, and retinal epithelium. Recently HNT has been shown to regulate Notch signaling in follicular epithelial development, which in turn alters cell differentiation and cell division. It is responsible for repressing String, Cut, and Hedgehog signaling, which are essential for regulating follicular cell proliferation [151]. The human homolog of HNT, RREB1 acts as a transcription factor that binds specifically to the RAS-responsive elements (RRE) of gene promoters. Recent investigations indicate that RREB1 is essential for spreading and migration of MCF-10A breast epithelial cells [152].

OAZ was apparently duplicated giving rise to two homologs in many vertebrates, including humans. OAZ is also called ZF423, while the closely related protein EHZF1 is called ZF521. No worm homolog has been detected, but a single homolog exists in *Daphnia* (Figure 27). Human and mouse homologs for this family have 30 zinc fingers, while *Drosophila* and *Daphnia* have fewer. OAZ/ZNF 423 and EHZF/ZNF521 are implicated in the control of olfactory epithelium, in B-lymphocyte differentiation, and in signal transduction by bone morphogenic protein (BMP). They are known to activate the BMP target genes *vent-2* (*Xenopus*) and *ventx2* (human) via interaction with SMAD [153]. ZNF521, in humans is known to regulate

ontogenesis of the hemato-vascular system through BMP pathways. OAZ/ZNF423 can also repress BMPs by activating repressors of BMPs [154, 155]. OAZ can apparently use different clusters of zinc fingers to interact with DNA, RNA or Protein [156]. In flies, 21 Zinc fingers are grouped into 4 clusters; the cluster near the amino terminus is assumed to bind DNA. DmOAZ is expressed throughout the life of flies and is strongest in posterior spiracles. Recent studies have shown that OAZ is involved in controlling posterior structure by regulating specific genes [157].

ZFAM9 is also known as Positive regulatory domain 13 (PRDM13) and is present in most vertebrates. A likely homolog of this gene is also present in *Drosophila* and *Daphnia* (Figure 28). However, the *C. elegans* genome had no homolog for this gene. All orthologs have 4 C2H2 zinc fingers. The function of these genes is unclear.

MTF (Metal-responsive Transcription Factor) have 6 tandem zinc-fingers. MTF activates metallothionein promoters in metazoans (Figure 25). MTF binds to the metal responsive element (MRE) and is involved in metal homeostasis and heavy-metal detoxification [158]. MTF in *Drosophila* appears to have a greater role in copper homeostasis than seen in vertebrates [159]. A single MTF homolog exists in *Daphnia* (not shown), but there appear to be no homologs in worms.

TFIIIA is a DNA-binding transcription factor that also binds RNA. It is generally required for 5sRNA gene expression in metazoans. TF3A's are poorly conserved between distantly related organisms. Vertebrate, insect, fungal, and plant sequences show within group similarity but only weak between group similarity [160, 161]. TFIIIA usually has 9 zinc-fingers, as does the single homolog in *Daphnia* (not shown). In yeast, *S. pombe* has a 10th zinc finger following a long spacer, while *S. cereviseae* has a long spacer between the 8th and 9th fingers [162]. No homolog of TFIIIA has yet been identified in worms.

C2H2 ZNF homologs of Drosophila developmental control genes

Daphnia have single homologs of the following well-known developmental control genes in *Drosophila*: Ovo, CI, LMD, OPA, SCRT, Slug, and ESG. In addition, *Daphnia* have three genes similar to the Odd gene family members Bowl, Bowel, and SOB (Figures 31 and 32). Although these genes encode C2H2 ZNF, other companion papers from the *Daphnia* consortium are intended to cover these in detail (personal communication).

Conclusions

Zinc-finger proteins probably represent the largest class of DNA-binding transcription factors in metazoan organisms, and as such, are likely to play critical roles in determining the extent to which various aspects of form and function are shared among taxa. The majority of these proteins are C2H2 zinc-finger proteins, many of which are already known to affect development and/or differentiation through a more or less direct effect on gene activation and/or repression.

A recent comparison of the full complement of C2H2 zinc-finger proteins observed or predicted within worm, fly, and human genomes has led to the tentative identification of nearly 40 orthologous groups shared between humans and invertebrates. From a phylogenetic perspective, the *Daphnia* genome would be expected to contain identifiable members for most of these groups. Using the reciprocal blast hit approach for estimating orthology, we uncovered 58 genes in *Daphnia* that appeared to be members of one of the 40 families conserved in bilaterians (Table 5, 6, 7 and 8). At least one member was identified for almost all families; only the JAZ family appeared to be absent from both *Daphnia* and *C. elegans*. All but two families had 3 or fewer members; only the SP and KLF families had more than three members each. Only the Odd-skipped and Snail families had 4 members each, while GLI, GFI, and Blimp had two

members each. For all other families (33), a single conserved member was identified in *Daphnia*. For 9 of these, a single conserved member was also present in each of the three other genomes; hence the latter genes appear to be relatively resistant to lineage specific deletion or expansion. Only three of the 40 families (including the most notable example, KLF) exhibit duplication or expansion in *Daphnia* relative to flies, but in many cases, gene duplications or expansions observed in flies and/or humans appear to be absent in *Daphnia*. Thus the *Daphnia* genome appears to be relatively efficient with respect to the number of C2H2 ZNF homologs per family.

Updating a previous analysis of C2H2 ZFP present in the common ancestor of bilaterians based on a survey of *Homo sapiens*, *Drosophila melanogaster* and *Caenorhabditis elegans*, we identified 58 well conserved C2H2 ZFP genes in *Daphnia* that belong to 40 distinct families. The *Daphnia* genome appears to be relatively efficient with respect to these well conserved C2H2 ZFP, since only 7 of the 40 gene families have more than one identified member. Worms have a comparable number of 6. In flies and humans, C2H2 ZFP gene expansions are more common, because these organisms display 15 and 24 multi-member families, respectively. In contrast, only three of the well conserved C2H2 ZFP families have expanded in *Daphnia* relative to *Drosophila*, and in two of these cases, just one additional gene was found. The KLF/SP family in *Daphnia*, however, is significantly larger than that of *Drosophila*, and many of the additional members found in *Daphnia* appear to correspond to KLF 1/2/4 homologs present in vertebrates.

Acknowledgements

The sequencing and portions of the analyses were performed at the DOE Joint Genome Institute under the auspices of the U.S. Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National

Laboratory under Contract No. DE-AC02-05CH11231, Los Alamos National Laboratory under Contract No. W-7405-ENG-36 and in collaboration with the Daphnia Genomics Consortium (DGC) http://daphnia.cgb.indiana.edu. Additional analyses were performed by wFleaBase, developed at the Genome Informatics Lab of Indiana University with support to Don Gilbert from the National Science Foundation and the National Institutes of Health. Coordination infrastructure for the DGC is provided by The Centre for Genomics and Bioinformatics at Indiana University, which is supported in part by the METACyt Initiative of Indiana University, funded in part through a major grant from the Lilly Endowment, Inc. Our work benefits from, and contributes to the Daphnia Genomics Consortium.

Table 5. The updated list of SP and KLF homologs with their accession numbers found in *Homo sapiens* (Build 36.3) *Drosophila melanogaster* (Build 4.1), *Caenorhabditis elegans* (Build 7.1) and *Daphnia pulex* (Version 1.1).

Family	*Homo sapiens*	*Drosophila melanogaster*	*Caenorhabditis elegans*	*Daphnia pulex*	Location of the *D. pulex* genes on v1.1 of the draft assembly
SP	SP1 (P08047) SP2 (Q02086) SP3/SPR2 (Q02447) SP4/SPR1 (Q02446) SP5 (Q6BEB4) SP6/KLF14 (Q3SY56) SP7 (Q8TDD2) SP8 (Q8IXZ3)	BTD (Q24266) DSp1 (NP_727360) CG5669 (NP_651232)	SPTF1 (NP_001021466) SPTF2 (NP_495833) SPTF3 (NP_493353)	SP5 (Dappu-114437) SP1 (Dappu-315784) SP8 (Dappu-106303)	scaffold_130:263756-265967 scaffold_15:792601-795915 scaffold_42:141432-144224
KLF	KLF1 (Q13351) KLF2 (Q9Y5W3) KLF3 (P57682) KLF4 (O43474) KLF5 (Q13887) KLF6 (Q99612) KLF7 (O75840) KLF8 (O95600) KLF9 (Q13886) KLF10 (Q13118) KLF11 (O14901) KLF12 (NP_009180) KLF13 (NP_057079) KLF14 (Q8TD94) KLF15 (Q9UIH9) KLF16 (Q9BXK1) KLF17 (Q5JT82)	CG12029 (NP_647822) CG9895 (NP_611747) CG3065 (NP_726393) CABUT (NP_608529) LUNA (NP_995811) BTEB2 (NP_572185)	KLF1 (NP_497632) F53F8.1 (NP_507995) MUA1 (AAU20846)	KLF1A (Dappu-48391) KLF1B (Dappu-51551) KLF1C (Dappu-243802) KLF1D (Dappu-262353) KLF3 (Dappu-27999) LUNA (Dappu-310992) CABUT (Dappu-312628) KLF9 (Dappu-315814) BTEB2 (Dappu-50068) KLF1E (Dappu-262162)	scaffold_16:1551074-1551469 scaffold_26:196741-197325 scaffold_26:237754-238940 scaffold_168:128271-129531 scaffold_1:1214746-1215144 scaffold_3:2311937-2324470 scaffold_6:1962325-1965508 scaffold_15:985228-986723 scaffold_21:551156-551927 scaffold_164:255538-257421

Table 6. Updated list of C2H2 ZNP families with expansion in all lineages.

Family	*Homo sapiens*	*Drosophila melanogaster*	*Caenorhabditis elegans*	*Daphnia pulex*	Location
EGR	EGR1 (NP_001955) EGR2 (NP_000390) EGR3 (NP_004421) EGR4 (NP_001956)	SR (NP_524395)	EGRH1 (NP_510467) EGRH2 (NP_500019) EGRH3 (NP_001041062)	SR (Dappu-96734)	5:1579836-1581770
ZFH1/2	ZEB1 (P37275) ZEB2 (O60315)	ZFH1 (P28166)	ZAG1 (Q94196)	ZFH1 (Dappu-225224)	31:791001-795287
ZFH3/4	ZFHX2 (Q9C0A1) ZFHX3 (Q15911) ZFHX4 (Q86UP3)	ZFH2 (P28167)	ZC124.3 (O45019)	ZFH2 (Dappu-233159)	2:1855545-1862285
SPALT	SALL1 (Q9NSC2) SALL2 (Q9Y467) SALL3 (Q9BXA9) SALL4 (Q9UJQ4)	SPALTm (P39770) SPALTr (NP_523548)	SEM4(NP_491997)	SALL (Dappu-111734)	88:97830-103728
DISCO	BNC1 (Q01954) BNC2 (Q6ZN30)	DISCO (P23792) DISCO-r (NP_727938)	F55C5.11 (Q1ZXU0)	DISCO (Dappu-442650)	91:404649-427757
GFI	GFI1 (Q99684) GFI1b (NP_004179)	SENS (NP_524818)	PAG3 (O02265)	GPS-A (Dappu-113215) GPS-B (Dappu-113216)	106:388895-391820 106:410992-414351
BLIMP1	PRDM1 (O75626) ZNF683 (Q8IZ20)	BLIMP1 (NP_647982)	BLIMP1 (NP_492723)	PRDM1A (Dappu-319330) PRDM1B (Dappu-333601)	29:157096-160885 247:75181-89061
ZEP	HIV-EP1 (P15822) HIV-EP2 (P31629) HIV-EP3 (NP_078779)	SHN (NP_476724)	SMA9 (CAF31491)	SHN (Dappu-226641)	60:513635-519628
IA1	INSM1 (Q01101) INSM2 (NP_115983)	NERFIN1 (NP_524783) NERFIN2 (NP_524300)	EGL46 (NP_504694)	NERFIN (Dappu-95880)	3:3754452-3756797
EVI1	PRDM16 (Q9HAZ2) EVI1 (Q03112)	HAM (Q8I7Z8) CG10348 (NP_609904)	EGL46 (CAA91353)	HAM (Dappu-113201)	106:272281-278224
FEZ	FEZF1(NP_001019784) FEZF2 (NP_060478)	CG31670 (NP_608631)	Y38H8A.5 (NP_502594)	FEZL(Dappu-40822)	2:1567364-1568302
ZFAM1	ZNF706 (Q9Y5V0)	CG18081 (NP_648807) CG15715 (NP_648808)	C01F6.9 (NP_501583) K10B4.1b (NP_001024783)	ZFAM706 (Dappu-230733)	17:391722-392672
ZFAM2	ZNF342 (Q8WUU4) BCL11A (Q9H165) BCL11B (Q9C0K0)	CG9650 (NP_727173)	F13H6.1b (NP_001122913)	BCL11 (Dappu-323911)	57:749423-752628
ZFAM4	ZNF384 (Q8TF68) ZNF362 (NP_689706)	CG2052 (NP_726568) RN (NP_996178) SQZ (NP_524403)	LIN29 (NP_496545)	RN (Dappu-104384)	30:275315-279360
ZFAM11 &12	KCMF (NP_064507)	CG11984 (NP_731306) CG31642 (NP_723159) CG31835 (NP_723881) CG15286 (NP_609706)	ZK652.6 (NP_001023029)	KCMF (Dappu-310981)	3:2220789-2222830
ZIC	ZIC1 (Q15915) ZIC2 (O95409) ZIC3 (O60481) ZIC4(Q8N9L1) ZIC5(Q96T25)	OPA (P39768) SUG (Q7K0S9)	REF2 (Q94178)	OPA (Dappu-290567)	104:131019-135418
OVO	OVOL1 (O14753) OVOL2 (Q9BRP0)	OVOrb (P51521)	LIN48 (Q19996)	OVO*(Dappu-290491)	191:57-1323
SNAIL	SNAIL3 (NP_840101) SNAIL2 (O43623) SNAIL1 (O95863) hSCRT1 (Q9BWW7) hSCRT2 (Q9NQ03)	SNAIL (P08044) ESG (P25932) WOR (NP_476601) SCRT (Q24140) CG12605 (NP_995996) CG12391 (NP_610639) CG17181 (NP_612040)	K02D7 (NP_499902) SCRT1(NP_491001) CES1 (NP_492338)	Dappu- 53927 Dappu- 129982 ESG (Dappu-347447) Dappu- 61957	39:954341-955431 110:193847-194734 23:1247838-1249532 110:238640-239641
GLI	GLI1 (P08151) GLI2 (P10070) GLI3 (P10071) GLIS1 (Q8NBF1) GLIS2 (Q9BZE0) GLIS3 (Q8NEA6)	CI (P19538) LMD (NP_732811) SUG (NP_996057)	TRA1(NP_001022880)	CI (Dappu-346973) LMD (Dappu-118558)	3:880659-885822 374:17043-18830
ODD Skipped	OSR1 (Q8TAX0) OSR2 (Q8N2R0)	ODD (P23803) SOB (Q9VQS7) BOWL (Q9VQU9)	ODD1 (NP_498552) ODD2 (NP_509032)	Dappu-238529 Dappu-335367 BOWL (Dappu-347540) Dappu- 323619	11:2090652-2095755 1:2333632-2339817 11:2120184-2122082 55:200488-243259

Table 7. The updated list of C2H2 zinc finger protein families that are resistant to expansion or deletion in *Homo sapiens, Drosophila melanogaster, Caenorhabditis elegans* and *Daphnia pulex* along with their accession numbers.

Family	*Homo sapiens*	*Drosophila melanogaster*	*Caenorhabditis elegans*	*Daphnia pulex*	Location of the *D. pulex* genes on v1.1 of the draft assembly
SAP61	SF3A3 (Q12874)	NOI (O46106)	T13H5.4 (NP_495799)	SF3A3 (Dappu-216576)	scaffold_86:295015-297217
SAP62	SF3A2 (Q15428)	CG10754 (NP_648603)	F11A10.2 (NP_502290)	SF3A2 (Dappu-226064)	scaffold_47:92393-93703
KIN17	KIN17 (O60870)	KIN17 (NP_649212)	Y52B11A.9 (NP_492860)	KIN17 (Dappu-187099)	scaffold_2:3390169-3391601
TF3A	TF3A (Q92664)	TF3A (NP_573161)	TF3A (NP_498067)	TF3A (Dappu-309275)	scaffold_94:449510-451421
ZNF207	ZNF207 (O43670)	CG17912 (NP_609808)	B0035.1 (NP_502124)	ZNF207 (Dappu-225978)	scaffold_45:119210-125698
ZNF277	ZN277 (Q9NRM2)	CG9890 (NP_611750)	ZTF7 (NP_505526)	ZNF277 (Dappu-187894)	scaffold_16:264655-266301
ZFAM5	ZNF622 (Q969S3)	CG6769 (NP_573252)	C16A4.4 (NP_498397)	ZNF622 (Dappu-194021)	scaffold_11:102282-103757
ZFAM6	ZMAT2 (Q96NC0)	CG11586 (NP_647881)	ZK686.4 (NP_498692)	ZMAT2 (Dappu-229015)	scaffold_139:171223-172629
ZFAM7	ZNF598 (Q86UK7)	CG11414 (NP_611932)	C52E12.1 (NP_495439)	ZNF598 (Dappu-323704)	scaffold_56:13693-16737

Table 8. The updated list of C2H2 zinc finger families that are absent from one or more organisms along with their accession numbers.

Family	***Homo sapiens***	***Drosophila melanogaster***	***Caenorhabditis elegans***	***Daphnia pulex***	**Location of the *D. pulex* genes on v1.1 of the draft assembly**
YY1	YY1 (P25490) YY2 (O15391) ZFP42 (Q96MM3)	YY1/PHO (NP_648317.1) PHOL (NP_648317)		PHO (Dappu-59123)	scaffold_78:192562-193513
HNT	RREB1 (Q92766)	PEB (NP_476674)		PEB (Dappu- 98615)	scaffold_10:170443-175752
MTF	MTF (Q14872)	MTF (NP_729491)		MTF (Dappu-227205)	scaffold_72:130738-133279
OAZ	EBF (Q2M1K9) ZNF521 (NP_056276)	OAZ-PB (NP_001097315)		OAZ (Dappu- 95503)	scaffold_3:1704248-1709703
ZFAM8	JAZF1 (Q86VZ6)	CG12054 (NP_651853)			
ZFAM9	PRDM13 (Q9H4Q3)	CG13296 (NP_648032)		PRDM13 (Dappu-111472)	scaffold_86:123582-125034
CTCF	CTCF (P49711) CTCFL (Q8NI51)	CTCF (NP_648109)		CTCF (Dappu-302037)	scaffold_158:218357-221909
ZXD	ZXDA (P98168) ZXDB (P98169) ZXDC (Q2QGD7)			ZXD (Dappu-54047)	scaffold_39:175871-177445

Table 9. Number of C2H2 genes identified in *Daphnia* (Dp) as compared to the updated list of C2H2 gene families found in humans (Hs), flies (Dm), and worm (Ce).

#	*Family*	*Fullname*	*Hs*	*Dm*	*Ce*	*Dp*
1	SP	Specificity protein	8	3	3	3
2	ZIC*	Zinc finger protein of the cerebellum/Sugarbabe	5	2	1	1
3	OVO*	Protein ovo/Protein shaven baby	2	1	1	1
4	SNAIL*	Neural crest transcription factor Slug/Snail/Escargot/Worniu/Scratch	5	7	3	4
5	GLI*	Glioma-associated oncogene/cubitus interruptus	6	2	1	2
6	EGR/KROX	Early growth response 1/Transcription factor Zif268/Stripe	4	1	3	1
7	KLF	Kruppel-like zinc finger protein	18	6	3	10
8	ZFH1/2	Zn finger homeobox protein 1/Smad-interacting protein	2	1	1	1
9	ZFH3/4	Zn finger homeodomain protein 3-4	3	1	1	1
10	OSR*	odd-skipped-related 2/Sob/Odd/Bowl	2	3	2	4
11	SPALT	Sal-like protein 1/Spalt-like transcription factor	4	2	1	1
12	DISCO	Zinc finger protein basonuclin	2	2	1	1
13	GFI	Growth factor independent protein	2	1	1	2
14	YY1	Yin and yang 1/Delta transcription factor/NF-E1/Pho	3	2	0	1
15	BLIMP	Beta-interferon gene positive regulatory domain I-binding factor	2	1	1	2
16	ZEP	HIV type I enhancer-binding protein 1/Schnurri	3	1	1	1
17	IA1	Insulinoma-associated protein 1/Nerfin	2	2	1	1
18	EVI1	Ecotropic virus integration site 1/Hamlet	2	2	1	1
19	SAP61	Splicing factor 3A subunit 3/Spliceosome-associated protein 61	1	1	1	1
20	SAP62	Splicing factor 3A subunit 2/Spliceosome-associated protein 62	1	1	1	1
21	KIN17	KIN antigenic determinant of recA protein	1	1	1	1
22	HNT	RAS responsive element binding protein 1	1	1	0	1
23	MTF	Metal regulatory element-binding transcription factor 1	1	1	0	1
24	TF3A	Transcription Factor III A	1	1	1	1
25	ZNF207	Zinc finger protein 207	1	1	1	1
26	ZNF277	Zinc finger protein 277	1	1	1	1
27	FEZ	Forebrain embryonic zinc finger protein	2	1	1	1
28	OAZ	Smad- and Olf-interacting zinc finger protein	2	1	0	1
29	ZFAM1	Zinc finger protein 706	1	2	2	1
30	ZFAM2	B-cell lymphoma/leukemia 11A	3	1	1	1
31	ZFAM4	Zinc finger protein 384/Nuclear matrix transcription factor 4	2	3	1	1
32	ZFAM5	Zinc finger protein 622 / Zinc finger-like protein 9	1	1	1	1
33	ZFAM6	Zinc finger matrin-type protein 2	1	1	1	1
34	ZFAM7	zinc finger protein 598	1	1	1	1
35	ZFAM8	Juxtaposed with another zinc finger protein; JAZ	1	1	0	0
36	ZFAM9	Zinc finger protein family 9	1	1	0	1
37	ZFAM10	Bromodomain and PHD finger-containing protein	3	1	1	1
38	ZFAM11/12	Potassium channel modulatory factor	1	4	1	1
39	CTCF	CCCTC-binding factor	2	1	0	1
40	ZXD	Zinc finger X-linked Duplicated protein	3	0	1	1

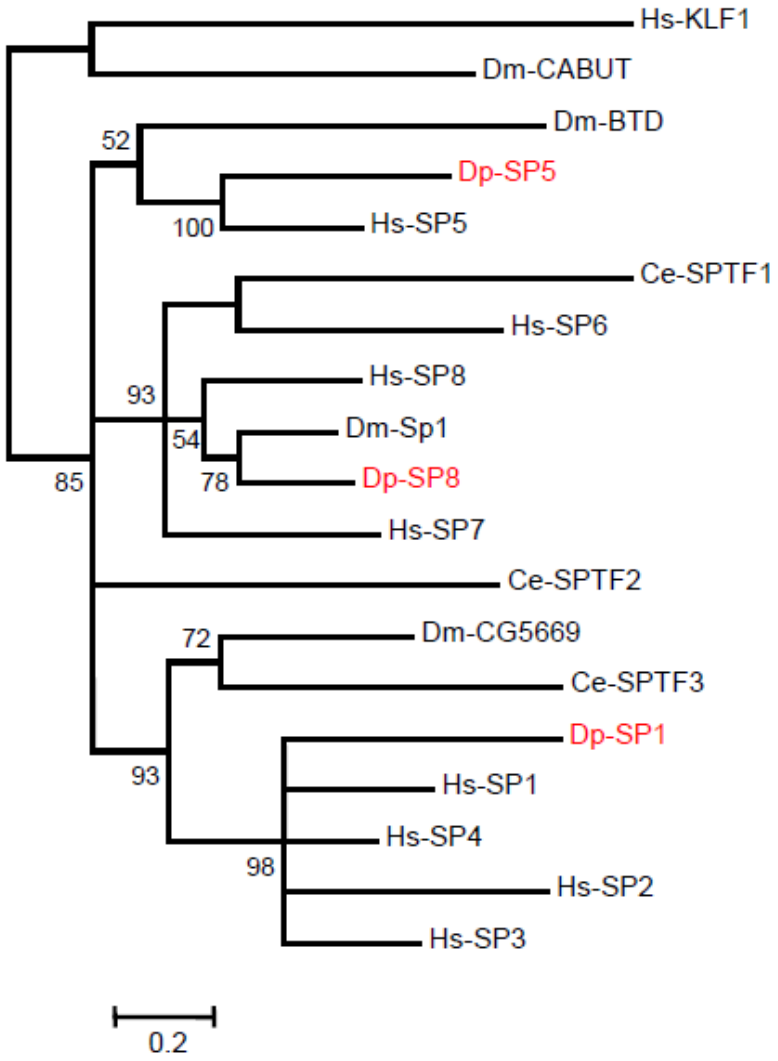

Figure 23. SP homologs in *Daphnia* correspond to those in *Drosophila*. Bayesian phylogenetic analysis of all SP proteins from Humans, *Drosophila*, *Daphnia* and *C. elegans* rooted with two KLF homologs, Hs-KLF1 and Dm-Cabot. The branch values indicate posterior probability and values greater than 50 are shown (Hs- *Homo sapiens*, Dm – *Drosophila melanogaster,* Ce - *Caenorhabditis elegans* and Dp – *Daphnia pulex*).

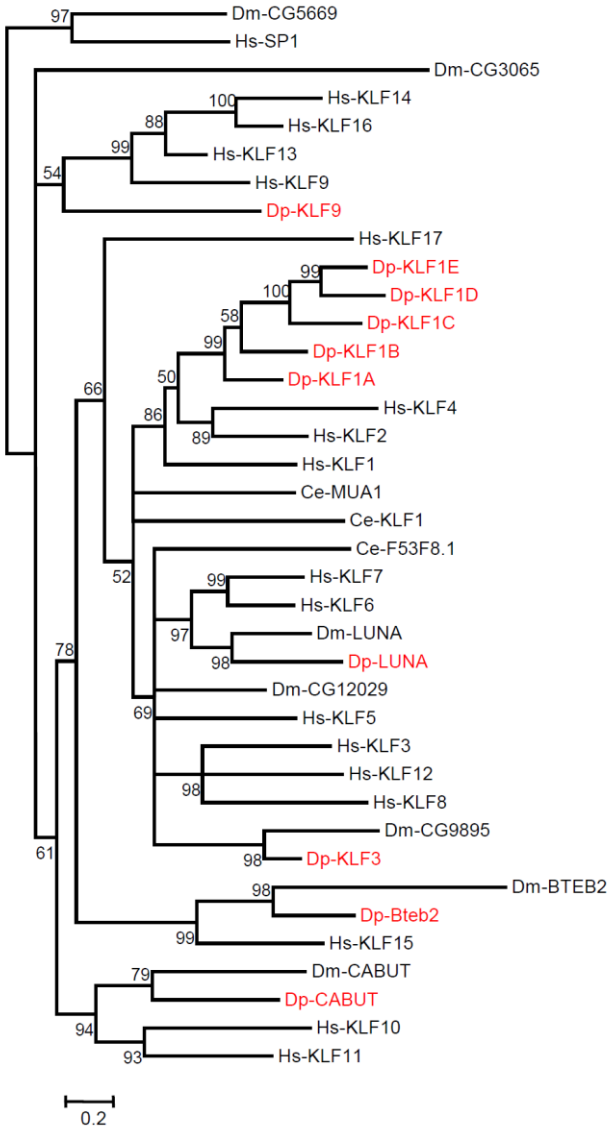

Figure 24. Additional KLF homologs in *Daphnia* relative to *Drosophila*. Bayesian phylogenetic analysis of all KLF proteins from Humans, *Drosophila*, *Daphnia* and *C. elegans* rooted with two SP homologs, Hs-SP1 and Dm-CG5669. The branch values indicate posterior probability and values greater than 50 are shown (Hs- *Homo sapiens*, Dm – *Drosophila melanogaster,* Ce - *Caenorhabditis elegans* and Dp – *Daphnia pulex*).

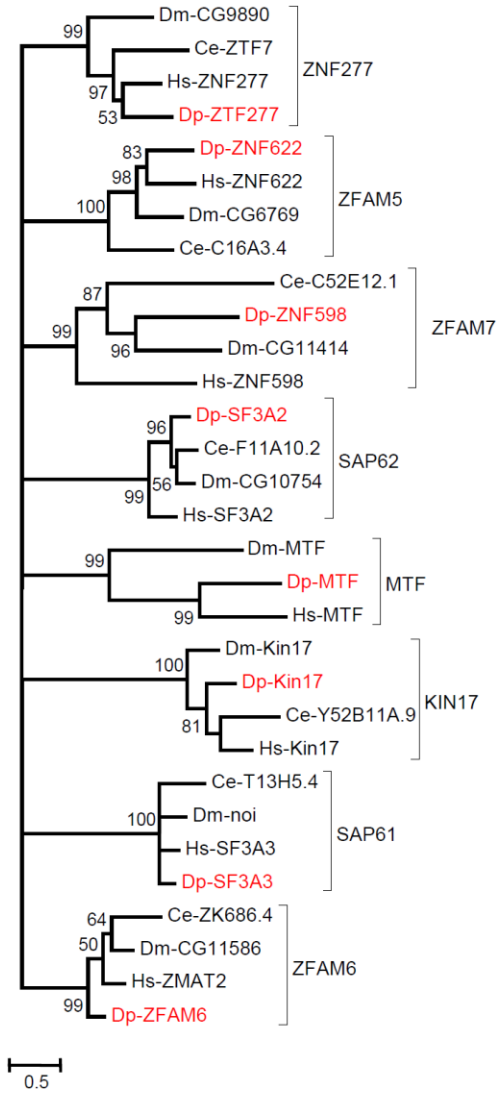

Figure 25. Bayesian phylogenetic analysis of C2H2 ZNF families that appear to be resistant to deletion/expansion in bilaterians (other than MTF). Proteins of ZNF277, Zfam5, Zfam7, SAP62, KIN17, SAP61 and Zfam6 families that have one member in each family and the MTF family that has missing member in *C. elegans,* were used to construct phylogenetic tree. The branch values indicate posterior probability and values greater than 50 are shown (Hs- *Homo sapiens*, Dm – *Drosophila melanogaster,* Ce -*Caenorhabditis elegans* and Dp – *Daphnia pulex*).

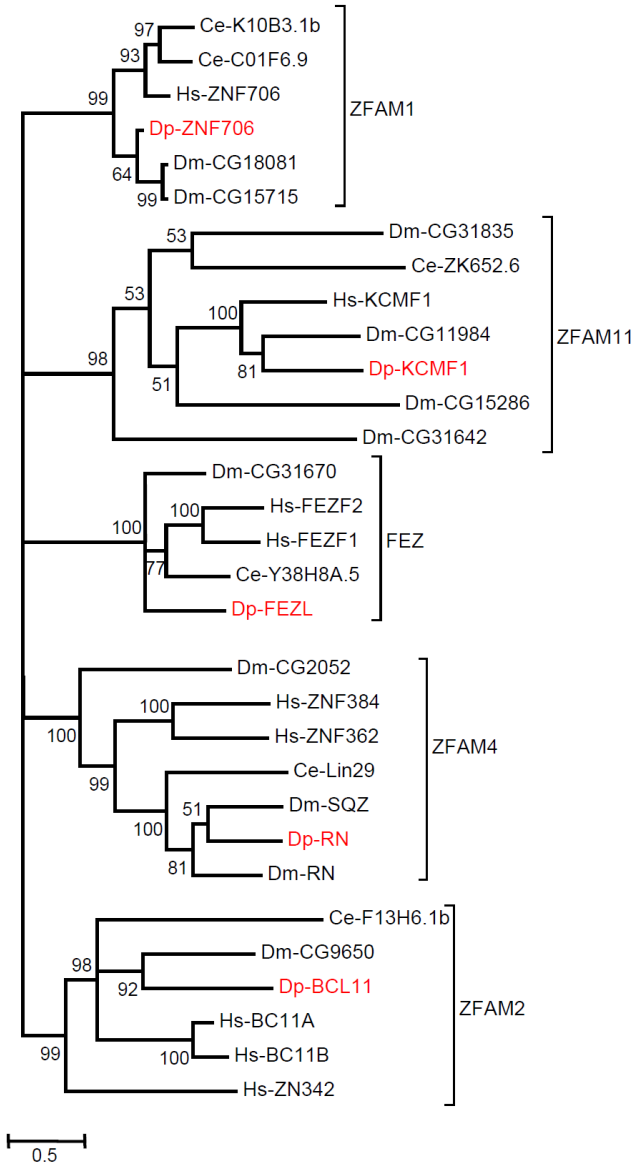

Figure 26. Bayesian phylogenetic analysis of C2H2 ZNF families with expansions in organisms other than *Daphnia*. Proteins of Zfam1, Zfam11, Fez, Zfam4 and Zfam2 family all having one member in *Daphnia* but more than one member in other genomes. The branch values indicate posterior probability and values greater than 50 are shown (Hs- *Homo sapiens*, Dm – *Drosophila melanogaster,* Ce -*Caenorhabditis elegans* and Dp – *Daphnia pulex*).

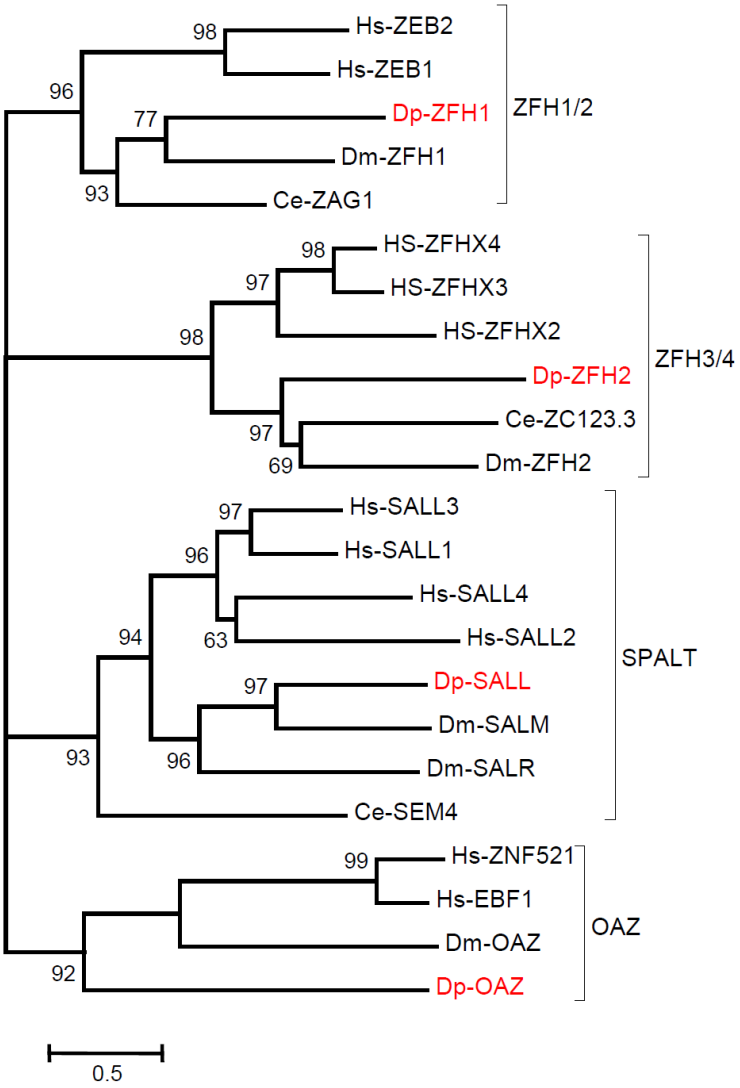

Figure 27. Bayesian phylogenetic analysis of C2H2 ZNF families with expansions in organisms other than *Daphnia*. Families ZFH1/2, ZFH3/4 and OAZ have additional homologs only for Humans, Spalt family has additional homologs for both humans and *Drosophila* and all families have one homolog for the *Daphnia* genome. Oaz family has no homolog for *C. elegans.* The branch values indicate posterior probability and values greater than 50 are shown (Hs- *Homo sapiens*, Dm – *Drosophila melanogaster,* Ce -*Caenorhabditis elegans* and Dp – *Daphnia pulex*).

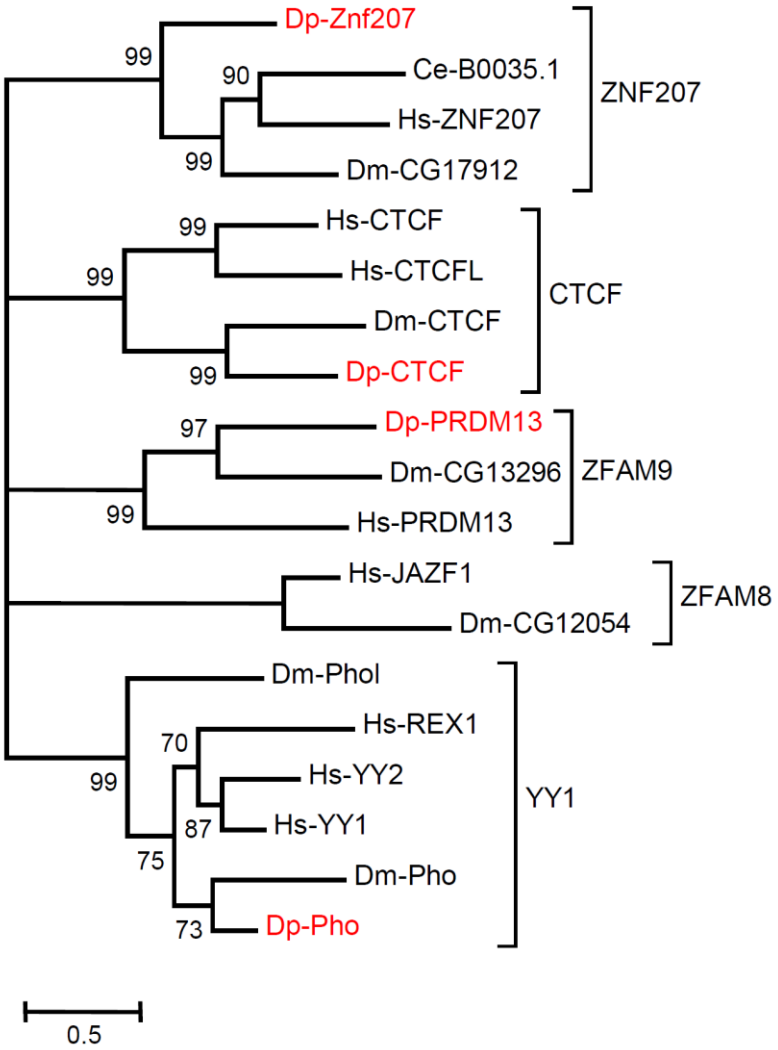

Figure 28. C2H2 ZNF absent from one or more organisms. Except for the family ZNF207 all other families in this tree is missing homolog in at least one genome. Families CTCF, Zfam9 and YY1 have a missing member for *C. elegans* and family Zfam8 is missing homolog in both *Daphnia* and *C. elegans*. The branch values indicate posterior probability and values greater than 50 are shown (Hs- *Homo sapiens*, Dm – *Drosophila melanogaster,* Ce -*Caenorhabditis elegans* and Dp – *Daphnia pulex*)

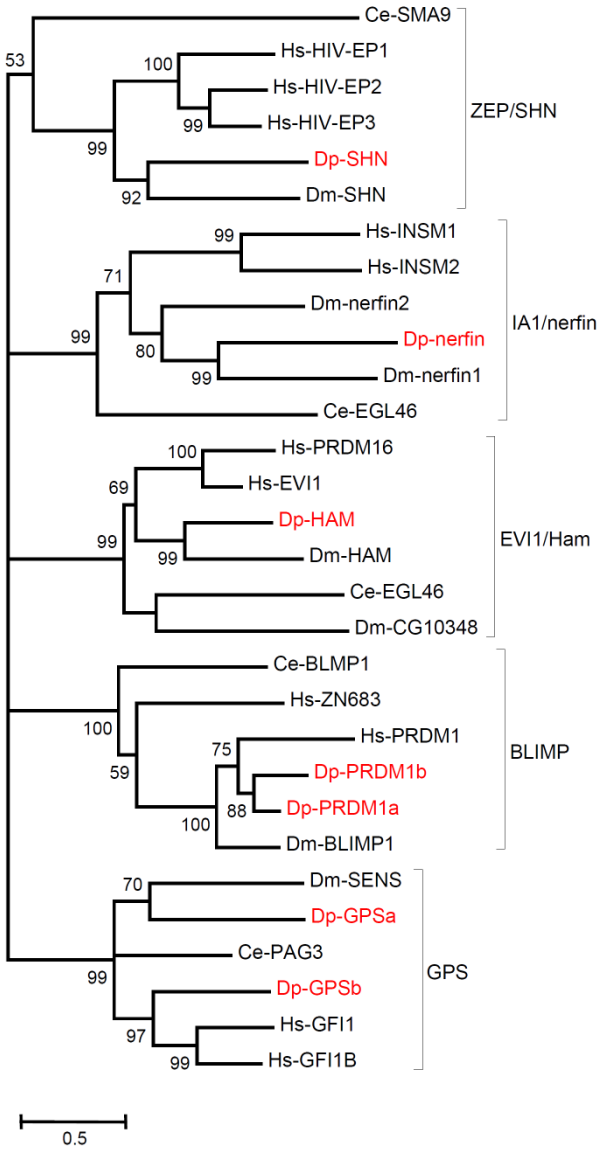

Figure 29. Bayesian phylogenetic analysis of C2H2 ZNF families with expansions in bilaterians. Family Zep/Shn has additional homologs only in humans, families IA1/Nerfin and Evi1/Ham have additional homologs in humans and *Drosophila* and families Blimp and GPS have additional homologs in humans and *Daphnia* but not in *Drosophila*. The branch values indicate posterior probability and values greater than 50 are shown (Hs- *Homo sapiens*, Dm – *Drosophila melanogaster,* Ce -*Caenorhabditis elegans* and Dp – *Daphnia pulex*).

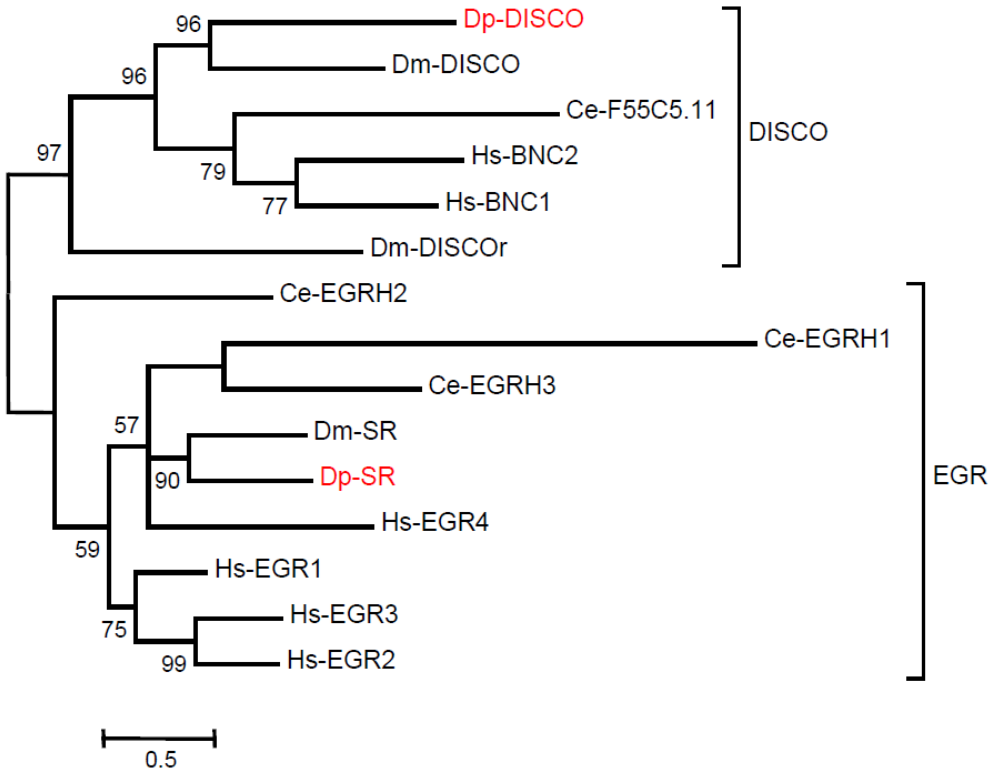

Figure 30. Bayesian phylogenetic analysis of C2H2 ZNF families Disco and EGR. Family Disco has additional homologs in humans and *Drosophila* and EGR has additional homologs in humans and *C. elegans*. The branch values indicate posterior probability and values greater than 50 are shown (Hs- *Homo sapiens*, Dm – *Drosophila melanogaster*, Ce -*Caenorhabditis elegans* and Dp – *Daphnia pulex*).

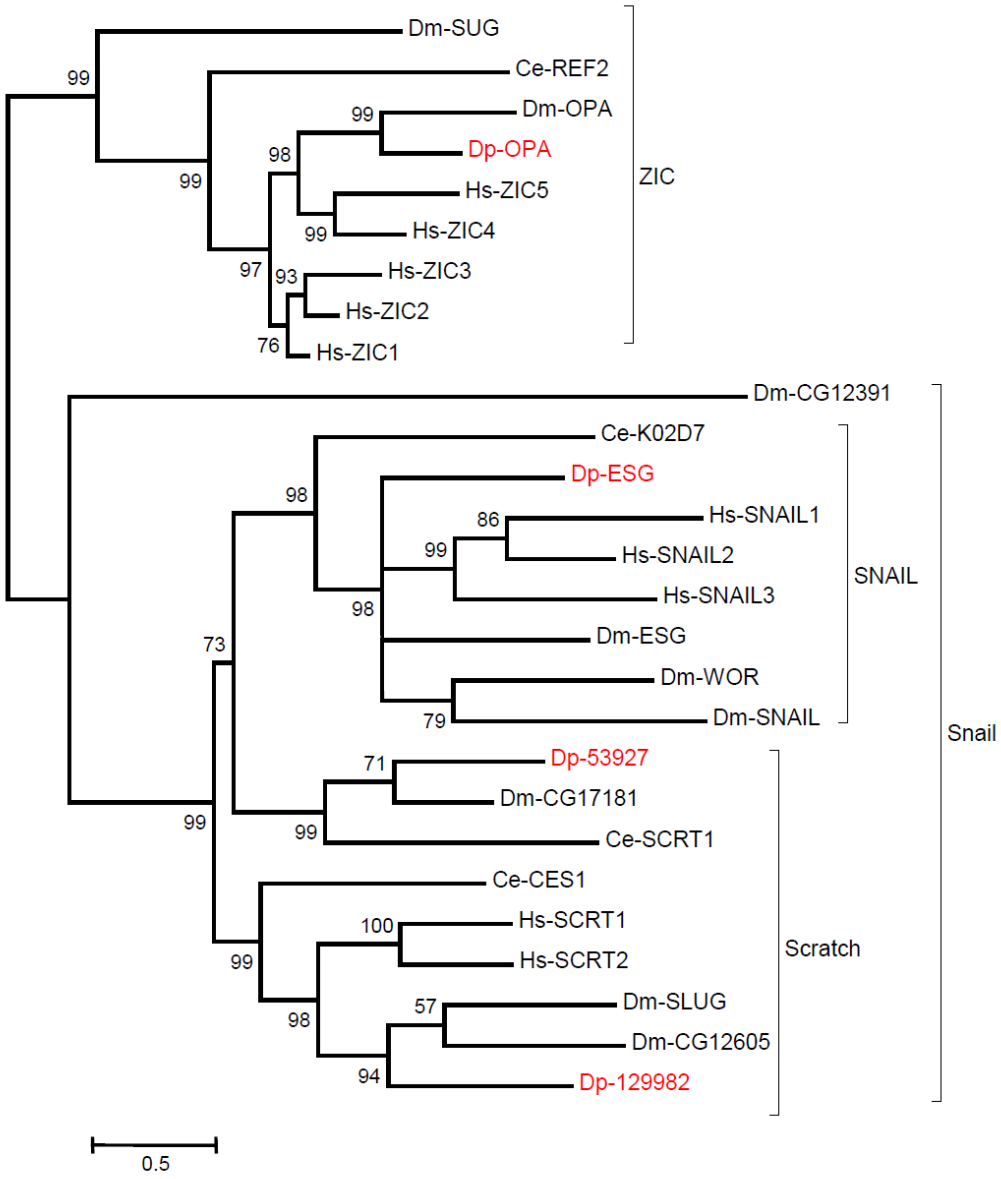

Figure 31. Genes likely to be involved in oogenesis and/or pattern formation showing no expansion in *Drosophila* relative to *Daphnia*.

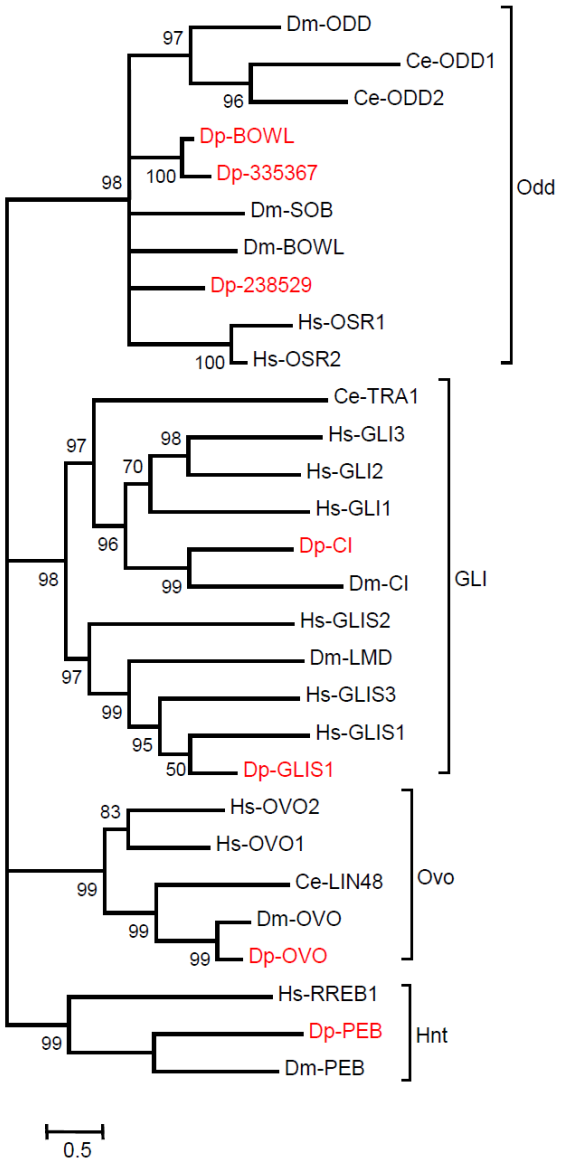

Figure 32. Genes likely to be involved in oogenesis and/or pattern formation showing expansions in *Drosophila* relative to *Daphnia.*

CHAPTER 5

A STUDY ON CONSERVATION AND DISTRIBUTION OF C2H2 ZINC FINGER GENES IN EUKARYOTES.

Abstract

The C2H2 zinc-finger (ZNF) containing protein family is one of the largest and most complex gene families in metazoan genomes. These genes are known to exist in almost all eukaryotes, and they constitute a major subset of eukaryotic transcription factors. The genes of this family usually occur as clusters in genomes and are thought to have undergone a massive expansion in vertebrates by multiple tandem duplication events [163]. In this study, we combined two popular approaches for homolog detection, Reciprocal Best Hit (RBH) [164] and Hidden–Markov model (HMM) profiles search, [165] on a diverse set of complete genomes of 124 eukaryotic species ranging from excavates to humans to identify all detectable members of 38 C2H2 ZNF gene families. We succeeded in identifying 3,675 genes as distinct members of 38 C2H2 gene families. These 38 families are distributed among the eukaryotes as progressive additions of gene blocks with increasing complexity of the organisms. The first block featuring the protists had 7 families, the second block featuring plants had 2 families, the third block featuring the fungi had 2 families (one of which was also present in plants) and the final block consisted of metazoans with 26 families. Among the metazoans, the simpler unicellular

metazoans had just 14 of the 26 families while most of the bilaterians had all 26 families making up a total of 38 families. Multiple potential examples of lineage-specific gene duplications and gene losses were also observed. Our hybrid approach combines features of the both RBH and HMM methods for homolog detection. This largely automated technique is much faster than manual methods and is able to detect homologs accurately and efficiently among a diverse set of organisms. Our analysis of the 38 evolutionarily conserved C2H2 ZNF gene families revealed a stepwise appearance of ZNF families, agreeing well with the phylogenetic relationship of the organisms compared and their presumed stepwise increase in complexity [166]

Introduction

The morphological complexity of organisms can be, to a certain extent, assigned to the transcription factors that control expression of various genes such as those that control signal transduction, cell growth, differentiation, and development [167]. One such family of transcription factors is the Zinc Finger Protein (ZFP) family, which is the largest family of DNA-binding transcription factors in eukaryotes. Of these ZFPs, the C2H2 type of zinc finger proteins remains the largest group [69]. This group is characterized by zinc fingers, consisting of 20-30 amino acid residues with a zinc ion coordinated by 2 cysteine and 2 histidine residues. C2H2 ZFPs often contain more than one such finger as tandem repeats. These proteins are known to exist in prokaryotes and eukaryotes and are most common in mammals. It is estimated that more than 700 C2H2 genes exist in humans accounting for more than 2 per cent of the total human genes [163]. Most of these C2H2 ZFPs act by binding DNA duplexes using their zinc finger motifs and are involved in controlling expression of their target genes. Some C2H2 ZFPs also play roles as either subunits of transcription proteins, splicing factors, or DNA damage repair proteins. We assume that as morphologically simpler organisms evolved increasing numbers of

genes, they must also have developed new control genes, including additional zinc finger genes, to evolve into more complex organisms.

With the advent of new “next generation” sequencing methods and the explosive growth of sequence databases, faster and more reliable methods for identification of gene family members, including the C2H2 ZNP genes, are of great interest. The study of the evolution of the C2H2-ZNP genes in various genomes may help to elucidate their possible role in the functions associated with speciation. Homolog prediction is one of the most vital steps in the functional annotation of genomes. The correct identification of homologs and putative orthologs greatly facilitates the accuracy of downstream analysis such as phylogenetic tree construction, protein structure prediction, prediction of protein-protein interaction, and species classification [168]. An effective and commonly used method of homolog/ortholog prediction is Reciprocal-best-BLAST-hits (RBH) [164, 169], where genes from two species are homologs and potential orthologs if they are both best BLAST hits when the gene from one genome is used to search the other genome. Although RBH is an effective procedure, potential homologs in multi-member families might be missed due to the restricted amount of information about the gene family in question that is present in just two sequences. More sophisticated methods based on Hidden–Markov models (HMM) [165] can also be applied and are easily automated for homolog detection [168, 170]. In the HMM method, each family is typically described by one or more information-rich HMM profiles that can be used to efficiently scan entire genomes for matches. This approach in general is very sensitive in detecting homologs and can be applied for large-scale, genome-level detection[168]. Homolog prediction is especially difficult when multiple related gene families are considered, as exemplified by the many diverse C2H2 ZNF gene families [69]. The high baseline of similarity among the families and subfamilies of C2H2 ZFPs,

along with their large numbers makes automated detection and assignment of C2H2 ZNF genes a challenge [69].

In our approach of automated gene homolog detection, we combined both RBH and HMM methods. The procedure requires minimal manual input, and the results are rapidly and accurately obtained. The method is analogous to the existing method of orthology detection in expressed sequence tags (EST) called HaMStR (Hidden Markov Model based Search for Orthologs using Reciprocity) [171]. Like our method HaMStR also uses the forward Hidden Markov Model and reverse BLAST search to extend existing ortholog cluster with sequences from further taxa. However, unlike HaMStR that used the large number of core orthologs as the reference set our method only used targeted set of ortholog families that were manually identified from 4 different species proteomes.

To understand the complex evolution of these zinc finger family genes, we undertook a survey to identify the different members of zinc finger family genes from all the eukaryotes that represent different taxa in the Tree of Life. Our previous work on the zinc-finger proteins of *Daphnia* compared with those of human, worm and fly provided the starting point of the current work on eukaryotes [172]. For the present study, we used a large subset of partially edited and augmented HMM profiles representing 38 C2H2 ZNF gene families within the bilaterian organisms and then used these profiles to predict gene family memberships from an extensive variety of 124 completely sequenced eukaryotic species.

Materials and methods

Generating HMM profiles

Previously identified putative orthologs that were present in the common ancestor of the bilaterians *Homo sapiens, Drosophila melanogaster*, and *Caenorhabditis elegans* were used as a focus for the present study. The Hidden–Markov model (HMM) profiles for each of these families were created with the HMMER 3.0 [173] package, using the NCBI reference sequences [174] belonging to the respective family described in the literature [163, 172, 175-177]. First, specific keywords describing the families were used to retrieve sequences for those families from various databases such as NCBI refseq [178] and Swissprot [179], and then these sequences were aligned using the *muscle* multiple sequence aligner program. Finally, the *hmmbuild* option of the HMMER package was used to build the profile. The reference sequences were obtained from diverse taxa in order to make the profiles more representative of the genomes chosen for study.

Obtaining eukaryotic protein datasets

The protein datasets of completely sequenced organisms representing all major eukaryotic clades were downloaded from NCBI, Ensembl, JGI, and Sanger. The downloaded genomes were then categorized into various class/phyla based on NCBI taxonomy information. Complete lists of the species under each phylum class are given in Table 10 and 11, and the sources for these genomes along with their build numbers are provided in Table 12, 13, 14 and 16. The obtained genomes were sorted taxonomically into 4 groups as protists, plants, fungi and metazoans.

HMM profile search

Whole predicted proteomes of the various species were scanned with all created HMM profiles using the *hmmsearch* option of the HMMER 3.0 package using minimum e-value threshold of 0.001. A loop written in Bash script was used to complete the reiterative *hmmsearch* procedure and the processing of results. For each HMM-genome pair, sequence hits were sorted based on the score for the full sequence and then on the best domain score. Only those sequences that had scores greater than 100 were chosen to be used in BLAST searches (standalone BLAST version 2.2.25. from NCBI) [169].

BLAST search

Standalone BLAST was performed using the chosen sequences against a local sequence database consisting only of the well annotated, complete set of genes from *Homo sapiens, Drosophila melanogaster*, and *Caenorhabditis elegans.* Subsequently, no more than 3 best hits from these focused BLAST results were scanned for accession numbers that matched a master list of such numbers. This master list was constructed using only those genes from the three reference organisms that were members of a given HMM profile/family. This pairwise process was repeated for each profile and each genome. Only the sequences that identified the correct family as verified by the master list accession numbers were chosen as family members.

Increasing specificity

The process was repeated two more times after adding the identified members from the previous round to generate a new HMM profile for the family. In order to increase the specificity of ortholog detection, during the second round, separate HMM profiles were generated for each of four taxonomic clades protists, fungi, plants, and metazoans. For those families for which the

sequence data was not available for different clades, general profiles were again used in the second round after updating the HMM profile with new sequences.

All the sequences identified as orthologs in the respective family were then catalogued. Those families that had multiple members were then analyzed to determine whether they were truly parologous or just duplicate sequences by aligning them using clustalw software [70].

Results

The hybrid method developed for homolog detection is largely automated, rapid, and efficient for identifying members of C2H2 ZFP genes. This method utilizes HMM profiles of the gene families for initial sensitive detection of putative homologs from a variety of genomes and then validates these putative homologs using a focused BLAST search of a restricted set of well annotated genomes and comparison to a master list of known homologs. This method is logically extensible to any number of gene families represented by an HMM, and any number of complete genomes (and predicted proteomes) available for analysis. The NCBI refseq database and Swiss-prot provided an excellent resource for the C2H2 ZFP gene sequences used to generate HMM profiles after alignment of reference sequences. Since the entire analysis was dependent on the HMM profile, the quality of the profiles used is crucial. Care was taken to choose only families that have accurate profiles. A total of 38 HMM profiles were generated for C2H2 ZFP gene families based on the existing information on those families.

During the first round of ortholog detection, 124 protein datasets belonging to various eukaryotic groups were scanned using 38 C2H2 ZFP HMM profiles separately. The output obtained consisted of potential homologs recognized for each profile within each genome. In the next step, a focused local BLAST used these potential profile-derived homologs, individually as queries against a set of well annotated, reference genomes. The BLAST outputs generated were

then scanned for the presence of "master list" genes as the top hits in order to decide unambiguous membership in the gene families represented by the HMMs used. The sequences found this way were used to further refine the HMM profiles to increase the specificity, and two more rounds of this process were performed. The final list of presumed gene family members was catalogued in a spreadsheet.

The final output with 38 HMM profiles on 124 eukaryotic genomes identified 3,675 members of a relatively complex subset of C2H2 ZNF families. Although initial HMM profiles were biased with more bilaterian sequences, subsequent scans employed separate HMM profiles for various eukaryotic groups derived from the sequences belonging to the respective groups. In the present study, 124 genomes were classified as 4 different groups. The first group of protists had 30 species belonging to excavates (including phyla Parabasilia, Fornicate and Euglenozoa), Chromalveolates (including phyla Apicomplexa, Ciliophora, Rhizaria, Heterokontophyta and Cryptophyta) and Amebozoa. The second group consisted of 16 different plant species belonging to Cyanidiophyta, Chlorophyta and Streptophyta. The third group had 28 species of fungi with phyla Basidiomycota, Ascomycota and Microsporidia (Table 11). The last and the largest group consisted of metazoans with 50 species consisting of Choanozoa, Placozoa, Porifera, Cnidarian, Nematoda, Annelida, Arthropoda, Echinodermata, Tunicata, Cephalochordata and chordate (Table 12). Heterogeneous representation of various groups was mainly due to either a lack of genome sequence or non-availability of the proteome datasets. Despite the breadth of the organisms scanned, the results (Figure 33) indicate a clear pattern of gene block conservation within closely related organisms as well as a reasonable progression of gene family additions that correlates well with a presumed increase in organismal complexity. This nearly uniform block pattern was occasionally disrupted by the presence of "holes" within the blocks (perhaps

representing a lineage or organism specific gene loss) and the presence of "loner" genes (genes that appear to be absent from almost all other closely related organisms). The latter may represent putative horizontal gene transfer events.

Gene families present in all Eukaryotes

Among the 38 C2H2 ZNP gene families, seven families (SF3A2, SF3A3, KIN17, ZFP598, ZFP622, ZMAT2 and ZNF207) appear to be present in almost all eukaryotes. Some exceptions include Discoba, which lacked families ZMAT2 and ZNF207, Rizaria, which lacked the ZNF207 family and microsporidia, which lacked ZMAT2, ZNF207, ZF622 and KIN17 families. A phylum/class represented by multiple species would be considered to have a particular family even if one organism belonging to the phylum lacked that gene family. Missing family members in some species could merely represent the absence of gene models from the data set due to error, incomplete sequencing coverage, or incorrect gene model prediction.

All these 7 gene families have just one homolog in almost all the species scanned. SF3A3 (Splicosome factor 3a subunit 3), SF3A2 (Splicosome factor 3a subunit 2), Kin17, and ZMAT2 (Zinc finger Matrin Type 2) all encode single highly conserved U1-like C2H2 zinc fingers. ZNF598 (Zinc Finger 598) has five C2H2 zinc fingers, ZNF622 (Zinc finger 622) has four C2H2 zinc fingers and ZNF207 has 2 C2H2 type zinc fingers. SF3A3 and SF3A2 are known to act as subunits for RNA splicing machinery [86-88], Kin17 is believed to be involved in the cellular response to DNA damage, gene expression, and DNA replication, and ZNF622 is known to be involved in early T cell activation and embryonic development in mouse. The exact functions of the other gene families (ZNF207, ZNF598 and ZMAT2) are not clearly understood.

The gene families added in plants and amoebozoans

The next expansion of gene families occurred in plants with an addition of the 2 families TFIIIA and YY1. Although, lower plants belonging to phylum Chlorophyta (green algae) lacked both these families, but the families were present in all higher plants belonging to phylum Streptophyta. Both families occurred as single homologs in most of the species, except YY1 which had 2 homologs in class Lillopsida. In addition to TFIIIA and YY1, Amoebozoa also had two more families, ZNF277 and ZIC. Though these families were not present in any other closely related groups (fungi or plants), they were present in lower metazoans.

TFIIIA (Transcription factor III A) with 9 zinc-fingers, is a DNA-binding transcription factor known to bind RNA and required for 5sRNA gene expression in metazoans [172]. YY1 (Yin Yang 1) generally has 4 zinc-fingers and has multiple functions, both as repressor and as an activator of gene expression [180]. In metazoans, they play roles in induction and patterning of the embryonic nervous system, differentiation within blood cell lineages, cell-cycle control, cell proliferation, differentiation, apoptosis, DNA synthesis and packaging, and X-inactivation [180]. The exact role of both these families in plants is not well understood.

Gene family additions in fungi

Expansion of gene families in fungus included the addition of 2 families (TFIIIA and GLI) to the original 7 families present in all the eukaryotes. Of the 2 families, TFIIIA was also present in plants, while GLI was not. TFIIIA has just one homolog in all the fungus species, as is true for plants and other eukaryotes,. Although GLI (Glioma-associated oncogene) occurs as a multi-gene family in most metazoans, it has one homolog in all fungus species. In Humans, the GLI family is known to regulate various aspects of early development of the central nervous system.

Gene family additions in Metazoans

The final massive expansion of C2H2 ZFP gene families occurred in metazoans with the addition of the remaining 26 gene families (Figure 33). Lower metazoans including Choanozoa, Porifera, Cnideria, and Placozoa only have a partial representation of these 26 families. Choanozoa, considered to be the most basal among the metazoans [166], have just 4 families added (KLF, JAZ, BLIMP and EVI1). They also lacked the families that were added in plants and fungi (GLI, TFIIA and YY1). The Porifera phylum, has an additional 4 families (SP, EGR, ZXD and MTF1), but compared to Choanozoa, they share just one family (KLF) and lack 3 families (JAZ, BLIMP and EVI1). Cniderians have all the families present in Choanozoa and Porifera except BLIMP and EVI1. They also have additional 6 families (SNAIL, OVO, GFI, IA1, FEZ and PRDM13) not present in Choanozoa and Porifera. Placozoa have all the families present in cniderians except MTF1, GLI and TFIIIA. Most of the bilaterians have almost all the 26 families represented except for a few phyla/classes that lack one or more families. The prominent phyla/classes lacking some of the families are arachnids (lacking ZNF 384), some insects (lacking ZXD), nematodes (lacking TFIIIA, ZXD, MTF1, JAZ, PRDM13, and CTCF), echinoderms (lacking DISCO), urochordates (lacking JAZ, OVO, FEZ, PRDM13, ZNF384, ZEP and OAZ), cephalochordates (lacking RREB, ZNF 384 and ZNF362), neoavians (lacking IA1) and monotremes (lacking ZNF 384 and ZNF362). The complete list of bilaterian specific zinc finger families are ZNF384, ZNF362, ZEP, DISCO (Disconnected), RREB (RAS responsive element binding protein), OAZ (Smad- and Olf-interacting zinc finger protein), CTCF, OSR (odd-skipped-related), SPALT, ZFHX1 and ZEB.

Conclusions

Our approach combined features of the both RBH and HMM methods of homolog detection. The largely automated technique is much faster than manual methods and is able to detect homologs accurately when compared to RBH alone. Furthermore, this method can be easily applied to new gene families that can be represented by an HMM, and to any number of completed genomes (and predicted proteomes) available for analysis. A total of 3,675 genes was identified from 124 completely sequenced eukaryotic genomes that belong to 38 members of a relatively complex subset of C2H2 ZNF families. These gene families in eukaryotes revealed a stepwise evolutionary process of gene block additions, which agrees well with the phylogenetic relationship of the organisms [166], as well as a presumed increase in organismal complexity.

Out of the 38 total families, 7 families are present in all eukaryotes. The increased morphological complexity from primitive protists to plants or fungi involved addition of two families, with one family common to both fungus and plants. The final expansion in metazoans added 26 families to those present in other groups (protists, plants and fungi) and this expansion correlates with the large increase in morphological complexity of these organisms. Most gene families resistant to expansion (single member gene families) are highly conserved and are represented in most of the eukaryotic species. We assume that these families are present in the common ancestor of eukaryotes as they are involved in fundamental processes such as DNA damage repair and intron splicing. The remarkable conservation of these gene families with respect to sequence, as well their ability to resist expansion, is consistent with previous observations [181-184]. Those functioning as structural proteins, pathogen response proteins, stress related proteins, signaling proteins, and proteins acting as transcription factors are often more prone to lineage specific expansions than are proteins that are involved in basic cellular

functions like DNA modification and RNA metabolism [185]. It is still unclear why specific gene families undergo massive expansion while some remain unchanged across evolutionary distances. It has been hypothesized that lineage-specific expansions are a principle means of adaptation and one of the most important sources of organizational and regulatory diversity in many organisms during transitions towards higher complexity[185].

Acknowledgements

We acknowledge the Center for Instruction, Research, and Technology (CIRT) at Indiana State University for computer cluster usage and other computational resources. Dr. Yihua Bai, CIRT, provided critical programming assistance.

Table 10. List of species represented in different phyla/class of the protists, plants and fungus.

Group	Species
Discoba	*Leishmania major, Leishmania braziliensis, Leishmania infantum, Trypanosoma brucei-427, Trypanosoma brucei-Gambiense, Trypanosoma brucei-Treu927, Trypanosoma cruzi-EsmeraldoLike, Trypanosoma cruzi-NonEsmeraldoLike*
Metamonada	*Trichomonas vaginalis, Giardia intestinalis*
Rizaria	*Bigelowiella natans*
Diatoms	*Phytophthora ramorum, Phytophthora sojae, Thalassiosira pseudonana, Phaeodactylum tricornutum, Guillardia theta*
Ciliates	*Paramecium tetraurelia*
Apicomplexeans	*Babesia bovis, Toxoplasma gondii, Theileria parva, Theileria annulata, Plasmodium berghei, Plasmodium chabaudi, Plasmodium falciparum, Plasmodium falciparum-ITA, Plasmodium knowlesi, Plasmodium vivax, Plasmodium yoelii*
Red Algae	*Cyanidioschyzon merolae*
GreenAlgae	*Chlamydomonas reinhardtii, Coccomyxa spcies, Micromonas species, Ostreococcus lucimarinus, Ostreococcus tauri*
Mosses	*Physcomitrella patens*
Malpighiales	*Brachypodium distachyon*
Monocots	*Sorghum bicolor, Zea mays, Eucalyptus grandis*
Dicots	*Cucumis sativus, Arabidopsis lyrata, Arabidopsis thaliana, Glycine max, Vitis vinifera*
Amoebozoans	*Dictyostelium discoideum, Entamoeba histolytica*
Microsporidians	*Encephalitozoon cuniculi*
Ascomycetes	*Mycosphaerella fijiensis, Aspergillus fumigatus, Aspergillus nidulans, Aspergillus niger, Aspergillus oryzae, Penicillium chrysogenum, Blumeria graminis, Botrytis cinerea, Ashbya gossypii, Candida albicans, Candida glabrata, Debaryomyces hansenii, Kluyveromyces lactis, Saccharomyces cerevisiae, Yarrowia lipolytica, Schizosaccharomyces pombe, Neurospora crassa, Chaetomium globosum, Magnaporthe grisea*
Basidiomycetes	*Agaricus bisporus, Laccaria bicolor, Phanerochaete chrysosporium, Serpula lacrymans, Sporisorium rellianum, Ustilago maydis, Puccinia graminis, Cryptococcus neoformans*

Table 11. List of species represented in different phyla/class of the metazoans.

Group	Species
Choanozoas	*Monosiga brevicollis*
Porifera	*Amphimedon queenslandica*
Cniderians	*Nematostella vectensis*
Placozoans	*Trichoplax adhaerens*
Other insects	*Bombyx mori, Tribolium castaneum, Acyrthosiphon pisum, Pediculus humanus*
Hymenopterans	*Apis mellifera, Nasonia vitripennis*
Dipterean	*Aedes aegypti, Anopheles gambiae, Culex quinquefasciatus, Drosophila melanogaster, Drosophila pseudoobscura*
Arachinids	*Ixodes scapularis*
Nematodes	*Caenorhabditis briggsae, Caenorhabditis elegans, Meloidogyne hapla, Pristionchus pacificus*
Annelids	*Capitella teleta, Helobdella robusta*
Crustaceans	*Daphnia pulex*
Echinoderms	*Strongylocentrotus purpuratus*
Urochordates	*Ciona intestinalis*
Cephalochordates	*Branchiostoma floridae*
Fishes	*Danio rerio, Takifugu rubripes, Tetraodon nigroviridis*
Amphibianns	*Xenopus silurana tropicalis*
Neoavians	*Gallus gallus, Meleagris gallopavo, Taeniopygia guttata*
Iguanas	*Anolis carolinensis*
Monotremes	*Ornithorhynchus anatinus*
Marsupials	*Monodelphis domestica*
Eutherians	*Ailuropoda melanoleuca, Loxodonta africana, Bos taurus, Equus caballus, Cavia porcellus, Canis lupus familiaris, Callithrix jacchus, Oryctolagus cuniculus, Mus musculus, Rattus norvegicus, Macaca mulatta, Pan troglodytes, Pongo pygmaeus, Homo sapiens*

Table 12. List of species belonging to group Protists and Amoebozoans with genome builds and source information.

Organism	Genome source and build
Bigelowiella natans CCMP2755	*JGIv1.0*
Guillardia theta CCMP2712	*JGIv1.0*
Phaeodactylum tricornutum	*JGI v2.0*
Phytophthora ramorum	*JGI v1.1*
Phytophthora sojae	*JGI v3.0*
Pythium ultimum BR144	*Pythium Genome DB (final release)*
Thalassiosira pseudonana	*JGI v3.0*
Babesia bovis T2Bo	*PiroplasmaDB v1.1*
Cryptosporidium hominis	*CryptoDB v4.5*
Cryptosporidium parvum IowaII	*CryptoDB v4.5*
Paramecium tetraurelia	*ParameciumDB v1.63*
Plasmodium berghei	*PlasmoDB v8.1*
Plasmodium chabaudi	*PlasmoDB v8.1*
Plasmodium falciparum	*PlasmoDB v8.1*
Plasmodium knowlesi	*PlasmoDB v8.1*
Plasmodium vivax	*PlasmoDB v8.1*
Plasmodium yoelii yoelii	*PlasmoDB v8.1*
Tetrahymena thermophila	*Ciliate.org (final release)*
Theileria annulata Ankara	*PiroplasmaDB v1.1*
Theileria parva Muguga	*PiroplasmaDB v1.1*
Toxoplasma gondii VEG	*ToxoDB v7.1*
Giardia intestinalis isolate A	*GiardiaDB v2.5*
Leishmania braziliensis	*TriTrypDB v4.0*
Leishmania infantum	*TriTrypDB v4.0*
Leishmania major Friedlin	*TriTrypDB v4.0*
Trichomonas vaginalis	*TrichDB v1.3*
Trypanosoma brucei Treu927	*TriTrypDB v4.0*
Trypanosoma brucei 427	*TriTrypDB v4.0*
Trypanosoma brucei Gambiense	*TriTrypDB v4.0*
Trypanosoma cruzi Esmeraldo	*TriTrypDB v4.0*
Dictyostelium discoideum	*DictyBase v1.0? (No Version info)*
Entamoeba histolytica	*AmoebaDB v1.7*

Table 13. List of species belonging to group plants with genome build and source information.

Organism	Genome source and build
Arabidopsis lyrata	*Phytozome v7.0*
Arabidopsis thaliana Columbia	*Phytozome v7.0*
Brachypodium distachyon	*Phytozome v7.0*
Cucumis sativus	*Phytozome v7.0*
Glycine max	*Phytozome v7.0*
Physcomitrella patens	*Phytozome v7.0*
Sorghum bicolor	*Phytozome v7.0*
Vitis vinifera	*Phytozome v7.0*
Zea mays mays	*Phytozome v7.0*
Eucalyptus grandis	*Phytozome v7.0*
Chlamydomonas reinhardtii	*Phytozome v7.0*
Cyanidioschyzon merolae 10D	*CMGP (Final release)*
Micromonas sp. RCC299	*JGI v3.0*
Ostreococcus lucimarinus CCE9901	*JGI v2.0*
Ostreococcus tauri	*JGI v2.0*

Table 14. List of species belonging to group fungus with genome build and source information.

Organism	Genome source and build
Agaricus bisporus var. bisporus H97	*JGI v2.0*
Ashbya gossypii ATCC 10895	*EMBL-UniProtKB*
Aspergillus fumigatus	*BROAD institute (Final release)*
Aspergillus nidulans	*BROAD institute (Final release)*
Aspergillus niger	*BROAD institute (Final release)*
Aspergillus oryzae	*BROAD institute (Final release)*
Blumeria graminis ffsp hordei	*blugen.org (Final release)*
Botrytis cinerea	*BROAD institute (Final release)*
Candida albicans SC5314	*EMBL-UniProtKB*
Candida glabrata	*EMBL-UniProtKB*
Chaetomium globosum NBRC_6347	*EMBL-UniProtKB*
Cryptococcus neoformans B-3501A	*EMBL-UniProtKB*
Debaryomyces hansenii	*EMBL-UniProtKB*
Encephalitozoon cuniculi	*EMBL-UniProtKB*
Kluyveromyces lactis	*EMBL-UniProtKB*
Laccaria bicolor	*EMBL-UniProtKB*
Magnaporthe grisea	*BROAD institute (Final release)*
Mycosphaerella fijiensis	*JGI v2.0*
Neurospora crassa	*EMBL-UniProtKB*
Penicillium chrysogenum Wisconsin	*EMBL-UniProtKB*
Phanerochaete chrysosporium	*JGI v2.1*
Puccinia graminis f. sp. tritici	*EMBL-UniProtKB*
Saccharomyces cerevisiae Lalvin_EC1118	*EMBL-UniProtKB*
Schizosaccharomyces pombe	*EMBL-UniProtKB*
Sporisorium rellianum	*EMBL-UniProtKB*
Ustilago maydis	*EMBL-UniProtKB*
Yarrowia lipolytica	*EMBL-UniProtKB*

Table 15. List of species belonging to group metazoans with genome build and source information.

Organism	Genome source and build
Ailuropoda melanoleuca	*NCBI current release (1/12/2012)*
Bos primigenius taurus	*NCBI current release (1/12/2012)*
Callithrix jacchus	*NCBI current release (1/12/2012)*
Canis lupus familiaris	*NCBI current release (1/12/2012)*
Cavia porcellus	*NCBI current release (1/12/2012)*
Equus ferus caballus	*NCBI current release (1/12/2012)*
Homo sapiens	*NCBI current release (1/12/2012)*
Loxodonta africana	*NCBI current release (1/12/2012)*
Macaca mulatta	*NCBI current release (1/12/2012)*
Monodelphis domestica	*NCBI current release (1/12/2012)*
Mus musculus C57BL/6J	*NCBI current release (1/12/2012)*
Ornithorhynchus anatinus	*NCBI current release (1/12/2012)*
Oryctolagus cuniculus	*NCBI current release (1/12/2012)*
Pan troglodytes	*NCBI current release (1/12/2012)*
Pongo pygmaeus/Pongo abelii	*NCBI current release (1/12/2012)*
Rattus norvegicus	*NCBI current release (1/12/2012)*
Anolis carolinensis	*NCBI current release (1/12/2012)*
Danio rerio	*NCBI current release (1/12/2012)*
Gallus gallus	*NCBI current release (1/12/2012)*
Meleagris gallopavo	*NCBI current release (1/12/2012)*
Takifugu rubripes	*JGI v4.0*
Taeniopygia guttata	*NCBI current release (1/12/2012)*
Tetraodon nigroviridis	*Genoscope v8.2*
Xenopus tropicalis	*NCBI current release (1/12/2012)*
Acyrthosiphon pisum	*NCBI current release (1/12/2012)*
Aedes aegypti Liverpool	*VectorBase v1.2*
Anopheles gambiae PEST	*VectorBase v3.6*
Nasonia vitripennis	*NCBI current release (1/12/2012)*
Apis mellifera	*NCBI current release (1/12/2012)*
Bombyx mori p50T	*SilkDB (Final release)*
Culex quinquefasciatus	*VectorBase v1.2*
Drosophila melanogaster	*FlyBase v5.42*
Drosophila pseudoobscura	*FlyBase v2.25*
Pediculus humanus	*VectorBase v1.2*
Tribolium castaneum GA-2	*NCBI current release (1/12/2012)*
Caenorhabditis briggsae	*Sanger-Wormbase*
Caenorhabditis elegans	*Sanger-Wormbase*
Meloidogyne hapla	*Sanger-Wormbase*
Pristionchus pacificus	*Sanger-Wormbase*
Amphimedon queenslandica	*NCBI current release (1/12/2012)*
Branchiostoma floridae	*JGI v1.0*
Capitella teleta	*JGI v1.0*
Ciona intestinalis	*NCBI current release (1/12/2012)*
Daphnia pulex	*JGI v1.1*
Helobdella robusta	*JGI v1.0*
Ixodes scapularis	*VectorBase v1.1*
Nematostella vectensis	*JGI v1.0*
Strongylocentrotus purpuratus	*NCBI current release (1/12/2012)*
Trichoplax adhaerens	*JGI v1.0*
Monosiga brevicollis	*JGI v1.0*

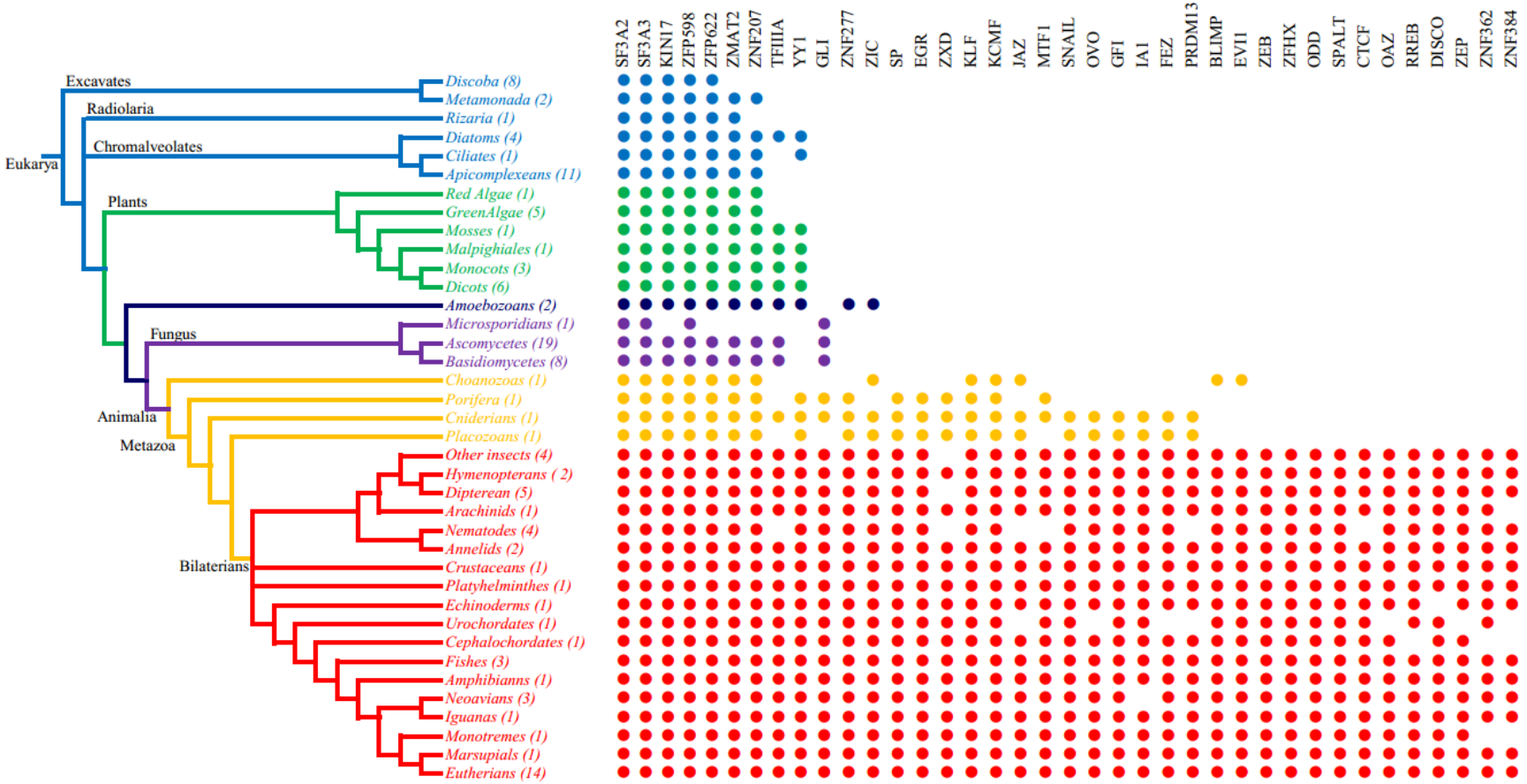

Figure 33. Summarized representation of the distribution of 38 gene families among various groups of eukaryotes. Numbers in parentheses indicate the number of species sampled from each phylum. Colors indicate various taxonomical groups (Light blue: Protists, Green: Plants, Dark blue: Amoebozoans, Purple: Fungus; Yellow: Lower metazoans and Red: Bilaterians).

CHAPTER 6

DISCUSSION AND CONCLUSIONS

In the present study we used phylogenomic methods that can handle large genome-scale data. The study addressed two related areas: phylogenomics as a method to build a species tree using the genome data, and prediction of gene function based on evolutionary analysis. For reconstructing phylogenetic trees using whole genome data, we investigated two different methods. In the first method we used Singular Value Decomposition (SVD) analysis to re-examine current evolutionary relationships for 12 *Drosophila* species using the predicted proteins from whole genomes. In the second method, we used reduced representations of genomes provided by a novel class of Type IIB restriction endonucleases to reconstruct whole genome phylogenies of 21 *Drosophila* species. For predicting the function of the genes based on evolutionary analysis, we used 40 conserved C2H2 zinc finger genes from bilaterians to uncover zinc finger genes from *Daphnia pulex* and to study the distribution of these families in 124 different species of eukaryotes. Below are the conclusions for each of the studies.

Whole genome phylogenies for multiple *Drosophila* species

Here we extended a novel phylogenetic method based on Singular Value Decomposition (SVD) to resolve the phylogeny of 12 recently sequenced *Drosophila* species. This method provides accurate comparisons for a high fraction of sequences within whole genomes without

the prior identification of orthologs or homologous sites. Our results indicate that it is possible to consult and interpret all predicted protein sequences within multiple whole genomes to produce accurate phylogenetic estimations of relatedness among *Drosophila* species. The phylogenetic tree derived from the 6 species of the melanogaster group, as well as all 12 species of *Drosophila,* exhibits strong branch support values and corresponds almost exactly to the currently accepted phylogeny. The most recent independent analyses based on whole genome sequence information depends upon filtered data sets in which a restricted number of highly conserved and putatively othologous genes are compared. We conclude that it is possible to include the entire dataset for a more robust analysis using a novel method to produce equivalent results.

Whole genome phylogeny of 21 *Drosophila* species using predicted fragments expected from Type IIB restriction enzyme digestion

In this study we simulated restriction enzyme digestions on 21 sequenced genomes of various *Drosophila* species using the predicted targets of 16 Type IIB restriction enzymes to effectively produce a large and arbitrary selection of loci from these genomes. These fragments provided an excellent fractional representation for those genomes. Following fragment extraction, the original genomic sequences downloaded from various source databases were represented as a collection of fragments of uniform length. For each genome a total of 16 fragment sets was generated by using 16 different type IIB enzymes. The number of fragments generated by each genome were related to the GC content and the frequency of cut sites estimated for those enzymes in random sequence. A comparison of fragments between genomes provided a list of fragments that were shared by those genomes. Closely related organisms had higher numbers of similar fragments (including identical fragments) compared to more distantly

related genomes. Similar fragments are defined as those with 6 or fewer mismatches. These similar fragments were then used to construct pairwise distance matrices for each enzyme, which was then used to construct phylogenetic trees. The individual NJ trees obtained for each enzyme and the consensus tree from all the enzymes were largely consistent with the currently accepted relationships among the various *Drosophila* groups and subgroups. The topology of the tree agrees precisely with those presented by Van der Linde and Houle, 2008 [60] and Yang et al. 2012 [61], except for the placement of the single species *D. eugracilis*, which formed a novel clustering with the melanogaster subgroup in our tree. Thus we conclude that our method of using multi-locus data obtained from small sub-genomic fragments provides good phylogenetic signal and produces a well resolved and well-supported species phylogeny. We note that our sub-genomic sampling method is analogous to previously described methods that use a different class of restriction enzymes to generate "RAD" markers for the comparative analysis of genomes at the population level [66]. We conclude that multi-locus, sub-genomic representation combined with next generation sequencing, especially for individuals and species without previous genome characterization, can improve knowledge of comparative genomics and the building of accurate phylogenetic trees, and not just between members of a population within species, but also between distantly related species as well.

A survey of well conserved families of C2H2 zinc-finger genes in *D. pulex*

In this study, we extended the previous analysis of bilaterians, where they identified 40 orthologous groups of C2H2 zinc-finger proteins in man, fly, and worm to include a second arthropod genome from the crustacean, *Daphnia pulex.* From a phylogenetic perspective, the *Daphnia* genome would be expected to contain identifiable members for most of these groups. Not surprisingly, *Daphnia* was found to be relatively efficient with respect to these well

conserved C2H2 ZFP and they possessed a similar number of orthologs for all these small orthology groups. In *Daphnia*, 7 of the 40 gene families had more than one identified member. Worms have a comparable number of 6 families. In flies and humans, C2H2 ZFP gene expansions are more common, displaying 15 and 24 multi-member families, respectively. In contrast, only three of the well conserved C2H2 ZFP families have expanded in *Daphnia* relative to *Drosophila*, and in two of these cases, just one additional gene was found. The KLF/SP family in *Daphnia*, however, is significantly larger than that of *Drosophila*, and many of the additional members found in *Daphnia* appear to correspond to KLF 1/2/4 homologs, which are absent in *Drosophila*, but present in vertebrates. Thus we conclude that the *Daphnia* genome appears to be relatively efficient with respect to the number of C2H2 ZNF homologs per family, with the exception of KLF/SP.

A study of conservation and distribution of C2H2 zinc finger genes in eukaryotes

In this study, we combined two popular approaches for homolog detection, Reciprocal Best Hit (RBH) and Hidden–Markov model (HMM) profiles to search a diverse set of complete genomes from 124 eukaryotic species ranging from excavates to humans to identify 38 C2H2 ZNF gene families. The largely automated technique was much faster than manual methods and was able to detect homologs accurately when compared to RBH alone. A total of 3,675 genes were identified from 124 completely sequenced eukaryotic genomes that belonged to 38 members of a relatively complex subset of C2H2 ZNF families. Most gene families that were resistant to expansion (single member gene families) were highly conserved and were represented in most of the eukaryotic species. We assume that these families were present in the common ancestor of eukaryotes as they were involved in fundamental processes such as DNA damage repair and intron splicing. The increased morphological complexity expected with the

transition from primitive protists to plants or fungi involved the addition of two families each, one of which was common among both fungi and plants. The final expansion in metazoans further added 26 families to the existing number, which correlates with morphological complexity of these organisms. These gene families in eukaryotes revealed a stepwise evolutionary process of gene block additions, which agrees well with the phylogenetic relationship of the organisms compared as well as a presumed increase in organismal complexity.

REFERENCES

1. Delsuc, F., H. Brinkmann, and H. Philippe, *Phylogenomics and the reconstruction of the tree of life.* Nat Rev Genet, 2005. **6**(5): p. 361-75.
2. Sjolander, K., *Phylogenomic inference of protein molecular function: advances and challenges.* Bioinformatics, 2004. **20**(2): p. 170-9.
3. Eisen, J.A., *Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis.* Genome Res, 1998. **8**(3): p. 163-7.
4. Eisen, J.A., D. Kaiser, and R.M. Myers, *Gastrogenomic delights: a movable feast.* Nat Med, 1997. **3**(10): p. 1076-8.
5. Kumar, S., et al., *Statistics and truth in phylogenomics.* Mol Biol Evol, 2012. **29**(2): p. 457-72.
6. Philippe, H., et al., *Phylogenomics of eukaryotes: impact of missing data on large alignments.* Mol Biol Evol, 2004. **21**(9): p. 1740-52.
7. Yang, Z. and B. Rannala, *Molecular phylogenetics: principles and practice.* Nat Rev Genet, 2012. **13**(5): p. 303-14.
8. Bininda-Emonds, O.R., *Inferring the Tree of Life: chopping a phylogenomic problem down to size?* BMC Biol, 2011. **9**: p. 59.
9. Jeffroy, O., et al., *Phylogenomics: the beginning of incongruence?* Trends Genet, 2006. **22**(4): p. 225-31.

10. Dutilh, B.E., et al., *Signature genes as a phylogenomic tool.* Mol Biol Evol, 2008. **25**(8): p. 1659-67.

11. Belda, E., A. Moya, and F.J. Silva, *Genome rearrangement distances and gene order phylogeny in gamma-Proteobacteria.* Mol Biol Evol, 2005. **22**(6): p. 1456-67.

12. Lin, J. and M. Gerstein, *Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels.* Genome Res, 2000. **10**(6): p. 808-18.

13. Stuart, G.W. and M.W. Berry, *An SVD-based comparison of nine whole eukaryotic genomes supports a coelomate rather than ecdysozoan lineage.* BMC Bioinformatics, 2004. **5**: p. 204.

14. Marshall, J.J. and S.E. Halford, *The type IIB restriction endonucleases.* Biochem Soc Trans, 2010. **38**(2): p. 410-6.

15. Brent, M.R., *How does eukaryotic gene prediction work?* Nat Biotechnol, 2007. **25**(8): p. 883-5.

16. Mathe, C., et al., *Current methods of gene prediction, their strengths and weaknesses.* Nucleic acids research, 2002. **30**(19): p. 4103-4117.

17. Thornton, J.W. and R. DeSalle, *Gene family evolution and homology: genomics meets phylogenetics.* Annu Rev Genomics Hum Genet, 2000. **1**: p. 41-73.

18. Ciccarelli, F.D., et al., *Toward automatic reconstruction of a highly resolved tree of life.* Science, 2006. **311**(5765): p. 1283-7.

19. Stuart, G.W., K. Moffett, and S. Baker, *Integrated gene and species phylogenies from unaligned whole genome protein sequences.* Bioinformatics, 2002. **18**(1): p. 100-8.

20. Stuart, G.W., K. Moffett, and J.J. Leader, *A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes.* Mol Biol Evol, 2002. **19**(4): p. 554-62.

21. Stuart, G.W., P.K. Moffett, and R.F. Bozarth, *A comprehensive open reading frame phylogenetic analysis of isometric positive strand ssRNA plant viruses.* Arch Virol, 2006. **151**(6): p. 1159-77.

22. Drosophila.12.Genomes.Consortium, *Evolution of genes and genomes on the Drosophila phylogeny.* Nature, 2007. **450**(7167): p. 203-218.

23. Bhutkar, A., W.M. Gelbart, and T.F. Smith, *Inferring genome-scale rearrangement phylogeny and ancestral gene order: a Drosophila case study.* Genome Biol, 2007. **8**(11): p. R236.

24. Clark, A.G., et al., *Evolution of genes and genomes on the Drosophila phylogeny.* Nature, 2007. **450**(7167): p. 203-18.

25. Machado, C.A. and J. Hey, *The causes of phylogenetic conflict in a classic Drosophila species group.* Proc Biol Sci, 2003. **270**(1520): p. 1193-202.

26. Pollard, D.A., et al., *Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting.* PLoS genetics, 2006. **2**(10): p. e173.

27. Wong, A., et al., *Phylogenetic incongruence in the Drosophila melanogaster species group.* Mol Phylogenet Evol, 2007. **43**(3): p. 1138-50.

28. Rosenfeld, J.A., et al., *Using whole genome presence/absence data to untangle function in 12 Drosophila genomes.* Fly (Austin), 2008. **2**(6): p. 291-9.

29. Russo, C.A., N. Takezaki, and M. Nei, *Molecular phylogeny and divergence times of drosophilid species.* Mol Biol Evol, 1995. **12**(3): p. 391-404.

30. Eanes, W.F., *Progress and prospects in evolutionary biology: The Drosophila model.* Science, 1998. **280**(5365): p. 850-851.

31. Lewis, R.L., A.T. Beckenbach, and A.O. Mooers, *The phylogeny of the subgroups within the melanogaster species group: likelihood tests on COI and COII sequences and a Bayesian estimate of phylogeny.* Mol Phylogenet Evol, 2005. **37**(1): p. 15-24.

32. O'Grady, P.M. and M.G. Kidwell, *Phylogeny of the subgenus Sophophora (Diptera : Drosophilidae) based on combined analysis of nuclear and mitochondrial sequences.* Molecular Phylogenetics and Evolution, 2002. **22**(3): p. 442-453.

33. Remsen, J. and P. O'Grady, *Phylogeny of Drosophilinae (Diptera : Drosophilidae), with comments on combined analysis and character support.* Molecular Phylogenetics and Evolution, 2002. **24**(2): p. 249-264.

34. Kopp, A. and J.R. True, *Phylogeny of the Oriental Drosophila melanogaster species group: a multilocus reconstruction.* Syst Biol, 2002. **51**(5): p. 786-805.

35. Arhontaki, K., et al., *Functional constraints of the Cu,Zn superoxide dismutase in species of the Drosophila melanogaster subgroup and phylogenetic analysis.* J Mol Evol, 2002. **55**(6): p. 745-56.

36. Gailey, D.A., et al., *A phylogeny of the Drosophilidae using the sex-behaviour gene fruitless.* Hereditas, 2000. **133**(1): p. 81-3.

37. Jeffs, P.S., E.C. Holmes, and M. Ashburner, *The molecular evolution of the alcohol dehydrogenase and alcohol dehydrogenase-related genes in the Drosophila melanogaster species subgroup.* Mol Biol Evol, 1994. **11**(2): p. 287-304.

38. Ko, W.Y., R.M. David, and H. Akashi, *Molecular phylogeny of the Drosophila melanogaster species subgroup.* J Mol Evol, 2003. **57**(5): p. 562-73.

39. Matsuo, Y., *Molecular evolution of the histone 3 multigene family in the Drosophila melanogaster species subgroup.* Mol Phylogenet Evol, 2000. **16**(3): p. 339-43.

40. Schlotterer, C., et al., *Comparative evolutionary analysis of rDNA ITS regions in Drosophila.* Mol Biol Evol, 1994. **11**(3): p. 513-22.

41. Shibata, H. and T. Yamazaki, *Molecular evolution of the duplicated Amy locus in the Drosophila melanogaster species subgroup: concerted evolution only in the coding region and an excess of nonsynonymous substitutions in speciation.* Genetics, 1995. **141**(1): p. 223-36.

42. Qi, J., H. Luo, and B. Hao, *CVTree: a phylogenetic tree reconstruction tool based on whole genomes.* Nucleic acids research, 2004. **32**(Web Server issue): p. W45-7.

43. Huson, D.H. and M. Steel, *Phylogenetic trees based on gene content.* Bioinformatics, 2004. **20**(13): p. 2044-9.

44. Snel, B., P. Bork, and M.A. Huynen, *Genome phylogeny based on gene content.* Nat Genet, 1999. **21**(1): p. 108-10.

45. Tekaia, F., A. Lazcano, and B. Dujon, *The genomic tree as revealed from whole proteome comparisons.* Genome Res, 1999. **9**(6): p. 550-557.

46. Korbel, J.O., et al., *SHOT: a web server for the construction of genome phylogenies.* Trends in Genetics, 2002. **18**(3): p. 158-162.

47. Bourque, G. and P.A. Pevzner, *Genome-scale evolution: Reconstructing gene orders in the ancestral species.* Genome Res, 2002. **12**(1): p. 26-36.

48. Roy, S.W. and W. Gilbert, *Resolution of a deep animal divergence by the pattern of intron conservation.* Proc Natl Acad Sci U S A, 2005. **102**(12): p. 4403-4408.

49. Yang, S., R.F. Doolittle, and P.E. Bourne, *Phylogeny determined by protein domain content.* Proc Natl Acad Sci U S A, 2005. **102**(2): p. 373-378.

50. Roberts, R.J., et al., *A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes.* Nucleic acids research, 2003. **31**(7): p. 1805-12.

51. Stebbins, G.L., *Variation and evolution in plants*, 1950, Columbia University Press N.Y.

52. Tengs, T., et al., *Genomic representations using concatenates of Type IIB restriction endonuclease digestion fragments.* Nucleic acids research, 2004. **32**(15): p. e121.

53. Dunn, J.J., et al., *Genomic signature tags (GSTs): a system for profiling genomic DNA.* Genome Res, 2002. **12**(11): p. 1756-65.

54. McQuilton, P., S.E. St Pierre, and J. Thurmond, *FlyBase 101--the basics of navigating FlyBase.* Nucleic acids research, 2012. **40**(Database issue): p. D706-14.

55. Rebeiz, M., et al., *Evolution of the tan locus contributed to pigment loss in Drosophila santomea: a response to Matute et al.* Cell, 2009. **139**(6): p. 1189-96.

56. Felsenstein, J., *PHYLIP (Phylogeny Inference Package)* D.b. author, Editor 2005.

57. Hahn, M.W., M.V. Han, and S.G. Han, *Gene family evolution across 12 Drosophila genomes.* PLoS Genet, 2007. **3**(11): p. e197.

58. Stark, A., et al., *Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures.* Nature, 2007. **450**(7167): p. 219-32.

59. Clark, A., et al., *Evolution of genes and genomes on the Drosophila phylogeny.* Nature, 2007. **450**: p. 203 - 218.

60. van der Linde, K. and D. Houle, *A supertree analysis and literature review of the genus Drosophila and closely related genera (Diptera, Drosophilidae).* Insect Systematics & Evolution, 2008. **39**(3): p. 241-267.

61. Yang, Y., et al., *Increasing the data size to accurately reconstruct the phylogenetic relationships between nine subgroups of the Drosophila melanogaster species group (Drosophilidae, Diptera).* Mol Phylogenet Evol, 2012. **62**(1): p. 214-23.

62. van der Linde, K., et al., *A supermatrix-based molecular phylogeny of the family Drosophilidae.* Genet Res (Camb), 2010. **92**(1): p. 25-38.

63. Pelandakis, M. and M. Solignac, *Molecular phylogeny of Drosophila based on ribosomal RNA sequences.* J Mol Evol, 1993. **37**(5): p. 525-43.

64. Inomata, N., H. Tachida, and T. Yamazaki, *Molecular evolution of the amy multigenes in the subgenus Sophophora of Drosophila.* Mol Biol Evol, 1997. **14**(12): p. 1338.

65. Yang, Y., et al., *Phylogenetic relationships of Drosophila melanogaster species group deduced from spacer regions of histone gene H2A-H2B.* Mol Phylogenet Evol, 2004. **30**(2): p. 336-43.

66. Baird, N.A., et al., *Rapid SNP discovery and genetic mapping using sequenced RAD markers.* PLoS One, 2008. **3**(10): p. e3376.

67. Rice, P., I. Longden, and A. Bleasby, *EMBOSS: the European Molecular Biology Open Software Suite.* Trends Genet, 2000. **16**(6): p. 276-7.

68. Colbourne, J.K., V.R. Singan, and D.G. Gilbert, *wFleaBase: the Daphnia genome database.* BMC Bioinformatics, 2005. **6**: p. 45.

69. Knight, R.D. and S.M. Shimeld, *Identification of conserved C2H2 zinc-finger gene families in the Bilateria.* Genome Biol, 2001. **2**(5): p. RESEARCH0016.

70. Thompson, J.D., T.J. Gibson, and D.G. Higgins, *Multiple sequence alignment using ClustalW and ClustalX.* Curr Protoc Bioinformatics, 2002. **Chapter 2**: p. Unit 2 3.

71. Capella-Gutierrez, S., J.M. Silla-Martinez, and T. Gabaldon, *trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses.* Bioinformatics, 2009: p. btp348.

72. Ronquist, F. and J.P. Huelsenbeck, *MrBayes 3: Bayesian phylogenetic inference under mixed models.* Bioinformatics, 2003. **19**(12): p. 1572-4.

73. Whelan, S. and N. Goldman, *A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach.* Molecular Biology and Evolution, 2001. **18**(5): p. 691-699.

74. Stillman, J.H., et al., *Recent advances in crustacean genomics.* Integr. Comp. Biol., 2008: p. icn096.

75. Tang, Z., et al., *ESTPiper - a web-based analysis pipeline for expressed sequence tags.* BMC Genomics, 2009. **10**(1): p. 174.

76. Lomberk, G. and R. Urrutia, *The family feud: turning off Sp1 by Sp1-like KLF proteins.* Biochem J, 2005. **392**(Pt 1): p. 1-11.

77. Kaczynski, J., T. Cook, and R. Urrutia, *Sp1- and Kruppel-like transcription factors.* Genome Biol, 2003. **4**(2): p. 206.

78. Brown, J.L., et al., *An Sp1/KLF binding site is important for the activity of a Polycomb group response element from the Drosophila engrailed gene.* Nucleic Acids Res, 2005. **33**(16): p. 5181-9.

79. Kawakami, Y., et al., *Sp8 and Sp9, two closely related buttonhead-like transcription factors, regulate Fgf8 expression and limb outgrowth in vertebrate embryos.* Development, 2004. **131**(19): p. 4763-74.

80. Wimmer, E.A., et al., *buttonhead and D-Sp1: a novel Drosophila gene pair.* Mech Dev, 1996. **59**(1): p. 53-62.

81. Munoz-Descalzo, S., J. Terol, and N. Paricio, *Cabut, a C2H2 zinc finger transcription factor, is required during Drosophila dorsal closure downstream of JNK signaling.* Dev Biol, 2005. **287**(1): p. 168-79.

82. De Graeve, F., et al., *Identification of the Drosophila progenitor of mammalian Kruppel-like factors 6 and 7 and a determinant of fly development.* Gene, 2003. **314**: p. 55-62.

83. Liang, H., W. Guo, and L. Nagarajan, *Chromosomal mapping and genomic organization of an evolutionarily conserved zinc finger gene ZNF277.* Genomics, 2000. **66**(2): p. 226-8.

84. Seong, H.A., et al., *Phosphorylation of a novel zinc-finger-like protein, ZPR9, by murine protein serine/threonine kinase 38 (MPK38).* Biochem J, 2002. **361**(Pt 3): p. 597-604.

85. Seong, H.A., K.T. Kim, and H. Ha, *Enhancement of B-MYB transcriptional activity by ZPR9, a novel zinc finger protein.* J Biol Chem, 2003. **278**(11): p. 9655-62.

86. Kramer, A., et al., *Structure-function analysis of the U2 snRNP-associated splicing factor SF3a.* Biochem Soc Trans, 2005. **33**(Pt 3): p. 439-42.

87. Tanackovic, G. and A. Kramer, *Human splicing factor SF3a, but not SF1, is essential for pre-mRNA splicing in vivo.* Mol Biol Cell, 2005. **16**(3): p. 1366-77.

88. Biard, D.S., et al., *Ionizing radiation triggers chromatin-bound kin17 complex formation in human cells.* J Biol Chem, 2002. **277**(21): p. 19156-65.

89. Miccoli, L., et al., *The human stress-activated protein kin17 belongs to the multiprotein DNA replication complex and associates in vivo with mammalian replication origins.* Mol Cell Biol, 2005. **25**(9): p. 3814-30.

90. Masson, C., et al., *Global genome repair is required to activate KIN17, a UVC-responsive gene involved in DNA replication.* Proc Natl Acad Sci U S A, 2003. **100**(2): p. 616-21.

91. Pahl, P.M., et al., *ZNF207, a ubiquitously expressed zinc finger gene on chromosome 6p21.3.* Genomics, 1998. **53**(3): p. 410-2.

92. Medina, K.L. and H. Singh, *Gene regulatory networks orchestrating B cell fate specification, commitment, and differentiation.* Curr Top Microbiol Immunol, 2005. **290**: p. 1-14.

93. Satterwhite, E., et al., *The BCL11 gene family: involvement of BCL11A in lymphoid malignancies.* Blood, 2001. **98**(12): p. 3413-20.

94. Sankaran, V.G., et al., *Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A.* Science, 2008. **322**(5909): p. 1839-42.

95. Karlsson, A., et al., *Bcl11b mutations identified in murine lymphomas increase the proliferation rate of hematopoietic progenitor cells.* BMC Cancer, 2007. **7**: p. 195.

96. St Pierre, S.E., et al., *Control of Drosophila imaginal disc development by rotund and roughened eye: differentially expressed transcripts of the same gene encoding functionally distinct zinc finger proteins.* Development, 2002. **129**(5): p. 1273-81.

97. Martini, A., et al., *Recurrent rearrangement of the Ewing's sarcoma gene, EWSR1, or its homologue, TAF15, with the transcription factor CIZ/NMP4 in acute leukemia.* Cancer Res, 2002. **62**(19): p. 5408-12.

98. Kreppel, M., et al., *Suppression of KCMF1 by constitutive high CD99 expression is involved in the migratory ability of Ewing's sarcoma cells.* Oncogene, 2006. **25**(19): p. 2795-800.

99. Yang, Z., N. Liu, and S. Lin, *A zebrafish forebrain-specific zinc finger gene can induce ectopic dlx2 and dlx6 expression.* Dev Biol, 2001. **231**(1): p. 138-48.

100. Hirata, T., et al., *Zinc-finger genes Fez and Fez-like function in the establishment of diencephalon subdivisions.* Development, 2006. **133**(20): p. 3993-4004.

101. Jeong, J.Y., et al., *Patterning the zebrafish diencephalon by the conserved zinc-finger protein Fezl.* Development, 2007. **134**(1): p. 127-36.

102. Liu, Y., et al., *The zinc finger transcription factor ZFHX1A is linked to cell proliferation by Rb-E2F1.* Biochem J, 2007. **408**(1): p. 79-85.

103. Liu, Y., et al., *Zeb1 links epithelial-mesenchymal transition and cellular senescence.* Development, 2008. **135**(3): p. 579-88.

104. Vandewalle, C., et al., *SIP1/ZEB2 induces EMT by repressing genes of different epithelial cell-cell junctions.* Nucleic Acids Res, 2005. **33**(20): p. 6566-78.

105. Wakamatsu, N., et al., *Mutations in SIP1, encoding Smad interacting protein-1, cause a form of Hirschsprung disease.* Nat Genet, 2001. **27**(4): p. 369-70.

106. Cacheux, V., et al., *Loss-of-function mutations in SIP1 Smad interacting protein 1 result in a syndromic Hirschsprung disease.* Hum Mol Genet, 2001. **10**(14): p. 1503-10.

107. Postigo, A.A. and D.C. Dean, *Differential expression and function of members of the zfh-1 family of zinc finger/homeodomain repressors.* Proc Natl Acad Sci U S A, 2000. **97**(12): p. 6391-6.

108. Wacker, I., et al., *zag-1, a Zn-finger homeodomain transcription factor controlling neuronal differentiation and axon outgrowth in C. elegans.* Development, 2003. **130**(16): p. 3795-805.

109. Jung, C.G., et al., *Homeotic factor ATBF1 induces the cell cycle arrest associated with neuronal differentiation.* Development, 2005. **132**(23): p. 5137-45.

110. Kim, C.J., et al., *Down-regulation of ATBF1 is a major inactivating mechanism in hepatocellular carcinoma.* Histopathology, 2008. **52**(5): p. 552-9.

111. Hemmi, K., et al., *A homeodomain-zinc finger protein, ZFHX4, is expressed in neuronal differentiation manner and suppressed in muscle differentiation manner.* Biol Pharm Bull, 2006. **29**(9): p. 1830-5.

112. Terriente, J., et al., *The Drosophila gene zfh2 is required to establish proximal-distal domains in the wing disc.* Dev Biol, 2008. **320**(1): p. 102-12.

113. Rusten, T.E., et al., *Spalt modifies EGFR-mediated induction of chordotonal precursors in the embryonic PNS of Drosophila promoting the development of oenocytes.* Development, 2001. **128**(5): p. 711-22.

114. Dong, P.D., et al., *Drosophila spalt/spalt-related mutants exhibit Townes-Brocks' syndrome phenotypes.* Proc Natl Acad Sci U S A, 2003. **100**(18): p. 10293-8.

115. Liang, J., et al., *The Caenorhabditis elegans schnurri homolog sma-9 mediates stage- and cell type-specific responses to DBL-1 BMP-related signaling.* Development, 2003. **130**(26): p. 6453-64.

116. Jin, W., et al., *Schnurri-2 controls BMP-dependent adipogenesis via interaction with Smad proteins.* Dev Cell, 2006. **10**(4): p. 461-71.

117. Jones, D.C., et al., *Regulation of adult bone mass by the zinc finger adapter protein Schnurri-3.* Science, 2006. **312**(5777): p. 1223-7.

118. Liang, J., et al., *Transcriptional repressor and activator activities of SMA-9 contribute differentially to BMP-related signaling outputs.* Dev Biol, 2007. **305**(2): p. 714-25.

119. Yao, L.C., et al., *Schnurri transcription factors from Drosophila and vertebrates can mediate Bmp signaling through a phylogenetically conserved mechanism.* Development, 2006. **133**(20): p. 4025-34.

120. Gierl, M.S., et al., *The zinc-finger factor Insm1 (IA-1) is essential for the development of pancreatic beta cells and intestinal endocrine cells.* Genes Dev, 2006. **20**(17): p. 2465-78.

121. Kuzin, A., et al., *Nerfin-1 is required for early axon guidance decisions in the developing Drosophila CNS.* Dev Biol, 2005. **277**(2): p. 347-65.

122. Lukowski, C.M., R.G. Ritzel, and A.J. Waskiewicz, *Expression of two insm1-like genes in the developing zebrafish nervous system.* Gene Expr Patterns, 2006. **6**(7): p. 711-8.

123. Moore, A.W., L.Y. Jan, and Y.N. Jan, *hamlet, a binary genetic switch between single- and multiple- dendrite neuron morphology.* Science, 2002. **297**(5585): p. 1355-8.

124. Baum, P.D., et al., *The Caenorhabditis elegans gene ham-2 links Hox patterning to migration of the HSN motor neuron.* Genes Dev, 1999. **13**(4): p. 472-83.

125. Nishikata, I., et al., *A novel EVI1 gene family, MEL1, lacking a PR domain (MEL1S) is expressed mainly in t(1;3)(p36;q21)-positive AML and blocks G-CSF-induced myeloid differentiation.* Blood, 2003. **102**(9): p. 3323-32.

126. Hoyt, P.R., et al., *The Evi1 proto-oncogene is required at midgestation for neural, heart, and paraxial mesenchyme development.* Mech Dev, 1997. **65**(1-2): p. 55-70.

127. Warner, L.E., et al., *Mutations in the early growth response 2 (EGR2) gene are associated with hereditary myelinopathies.* Nat Genet, 1998. **18**(4): p. 382-4.

128. Wieland, G.D., et al., *Early growth response proteins EGR-4 and EGR-3 interact with immune inflammatory mediators NF-kappaB p50 and p65.* J Cell Sci, 2005. **118**(Pt 14): p. 3203-12.

129. Vanhoutteghem, A. and P. Djian, *Basonuclins 1 and 2, whose genes share a common origin, are proteins with widely different properties and functions.* Proc Natl Acad Sci U S A, 2006. **103**(33): p. 12423-8.

130. Vanhoutteghem, A. and P. Djian, *Basonuclin 2: an extremely conserved homolog of the zinc finger protein basonuclin.* Proc Natl Acad Sci U S A, 2004. **101**(10): p. 3468-73.

131. Mahaffey, J.W., C.M. Griswold, and Q.M. Cao, *The Drosophila genes disconnected and disco-related are redundant with respect to larval head development and accumulation of mRNAs from deformed target genes.* Genetics, 2001. **157**(1): p. 225-36.

132. Robertson, L.K., et al., *An interactive network of zinc-finger proteins contributes to regionalization of the Drosophila embryo and establishes the domains of HOM-C protein function.* Development, 2004. **131**(12): p. 2781-9.

133. Fiolka, K., et al., *Gfi1 and Gfi1b act equivalently in haematopoiesis, but have distinct, non-overlapping functions in inner ear development.* EMBO Rep, 2006. **7**(3): p. 326-33.

134. Jia, Y., et al., *The C. elegans gene pag-3 is homologous to the zinc finger proto-oncogene gfi-1.* Development, 1997. **124**(10): p. 2063-73.

135. Acar, M., et al., *Senseless physically interacts with proneural proteins and functions as a transcriptional co-activator.* Development, 2006. **133**(10): p. 1979-89.

136. Jafar-Nejad, H. and H.J. Bellen, *Gfi/Pag-3/senseless zinc finger proteins: a unifying theme?* Mol Cell Biol, 2004. **24**(20): p. 8803-12.

137. Agawa, Y., et al., *Drosophila Blimp-1 is a transient transcriptional repressor that controls timing of the ecdysone-induced developmental pathway.* Mol Cell Biol, 2007. **27**(24): p. 8739-47.

138. Davis, M.M., *Blimp-1 over Budapest.* Nat Immunol, 2007. **8**(5): p. 445-7.

139. Kallies, A., et al., *Transcriptional repressor Blimp-1 is essential for T cell homeostasis and self-tolerance.* Nat Immunol, 2006. **7**(5): p. 466-74.

140. Martins, G.A., et al., *Transcriptional repressor Blimp-1 regulates T cell homeostasis and function.* Nat Immunol, 2006. **7**(5): p. 457-65.

141. Jambunathan, S. and J.D. Fontes, *Sumoylation of the zinc finger protein ZXDC enhances the function of its transcriptional activation domain.* Biol Chem, 2007. **388**(9): p. 965-72.

142. Al-Kandari, W., et al., *The zinc finger proteins ZXDA and ZXDC form a complex that binds CIITA and regulates MHC II gene transcription.* J Mol Biol, 2007. **369**(5): p. 1175-87.

143. Moon, H., et al., *CTCF is conserved from Drosophila to humans and confers enhancer blocking of the Fab-8 insulator.* EMBO Rep, 2005. **6**(2): p. 165-70.

144. Loukinov, D.I., et al., *BORIS, a novel male germ-line-specific protein associated with epigenetic reprogramming events, shares the same 11-zinc-finger domain with CTCF, the insulator protein involved in reading imprinting marks in the soma.* Proc Natl Acad Sci U S A, 2002. **99**(10): p. 6806-11.

145. Boumil, R.M., et al., *Differential methylation of Xite and CTCF sites in Tsix mirrors the pattern of X-inactivation choice in mice.* Mol Cell Biol, 2006. **26**(6): p. 2109-17.

146. Kim, J.D., C. Faulk, and J. Kim, *Retroposition and evolution of the DNA-binding motifs of YY1, YY2 and REX1.* Nucleic Acids Res, 2007. **35**(10): p. 3442-52.

147. Brown, J.L., et al., *The Drosophila pho-like gene encodes a YY1-related DNA binding protein that is redundant with pleiohomeotic in homeotic gene silencing.* Development, 2003. **130**(2): p. 285-94.

148. Stronach, B.E. and N. Perrimon, *Investigation of leading edge formation at the interface of amnioserosa and dorsal ectoderm in the Drosophila embryo.* Development, 2001. **128**(15): p. 2905-13.

149. Wilk, R., et al., *The hindsight gene is required for epithelial maintenance and differentiation of the tracheal system in Drosophila.* Dev Biol, 2000. **219**(2): p. 183-96.

150. Pickup, A.T., et al., *Control of photoreceptor cell morphology, planar polarity and epithelial integrity during Drosophila eye development.* Development, 2002. **129**(9): p. 2247-58.

151. Sun, J. and W.M. Deng, *Hindsight mediates the role of notch in suppressing hedgehog signaling and cell proliferation.* Dev Cell, 2007. **12**(3): p. 431-42.

152. Mukhopadhyay, N.K., et al., *The zinc finger protein ras-responsive element binding protein-1 is a coregulator of the androgen receptor: implications for the role of the Ras pathway in enhancing androgenic signaling in prostate cancer.* Mol Endocrinol, 2007. **21**(9): p. 2056-70.

153. Ku, M., et al., *OAZ regulates bone morphogenetic protein signaling through Smad6 activation.* J Biol Chem, 2006. **281**(8): p. 5277-87.

154. Bond, H.M., et al., *Early hematopoietic zinc finger protein-zinc finger protein 521: a candidate regulator of diverse immature cells.* Int J Biochem Cell Biol, 2008. **40**(5): p. 848-54.

155. Bond, H.M., et al., *Early hematopoietic zinc finger protein (EHZF), the human homolog to mouse Evi3, is highly expressed in primitive human hematopoietic cells.* Blood, 2004. **103**(6): p. 2062-70.

156. Hata, A., et al., *OAZ uses distinct DNA- and protein-binding zinc fingers in separate BMP-Smad and Olf signaling pathways.* Cell, 2000. **100**(2): p. 229-40.

157. Krattinger, A., et al., *DmOAZ, the unique Drosophila melanogaster OAZ homologue is involved in posterior spiracle development.* Dev Genes Evol, 2007. **217**(3): p. 197-208.

158. Wang, Y., et al., *Metal-responsive transcription factor-1 (MTF-1) selects different types of metal response elements at low vs. high zinc concentration.* Biol Chem, 2004. **385**(7): p. 623-32.

159. Balamurugan, K., et al., *Copper homeostasis in Drosophila by complex interplay of import, storage and behavioral avoidance.* EMBO J, 2007. **26**(4): p. 1035-44.

160. Hanas, J.S., et al., *cDNA cloning, DNA binding, and evolution of mammalian transcription factor IIIA.* Gene, 2002. **282**(1-2): p. 43-52.

161. Mathieu, O., et al., *Identification and characterization of transcription factor IIIA and ribosomal protein L5 from Arabidopsis thaliana.* Nucleic Acids Res, 2003. **31**(9): p. 2424-33.

162. Schulman, D.B. and D.R. Setzer, *Identification and characterization of transcription factor IIIA from Schizosaccharomyces pombe.* Nucleic Acids Res, 2002. **30**(13): p. 2772-81.

163. Tadepally, H.D., G. Burger, and M. Aubry, *Evolution of C2H2-zinc finger genes and subfamilies in mammals: species-specific duplication and loss of clusters, genes and effector domains.* BMC Evol Biol, 2008. **8**: p. 176.

164. Rivera, M.C., et al., *Genomic evidence for two functionally distinct gene classes.* Proc Natl Acad Sci U S A, 1998. **95**(11): p. 6239-44.

165. Eddy, S.R., *Profile hidden Markov models.* Bioinformatics, 1998. **14**(9): p. 755-63.

166. Pennisi, E., *Drafting a tree.* Science, 2003. **300**(5626): p. 1694.

167. Valentine, J.W. and A.G. Collins, *The significance of moulting in Ecdysozoan evolution.* Evol Dev, 2000. **2**(3): p. 152-6.

168. Soding, J., *Protein homology detection by HMM-HMM comparison.* Bioinformatics, 2005. **21**(7): p. 951-60.

169. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

170. Chen, F., et al., *Assessing performance of orthology detection strategies applied to eukaryotic genomes.* PLoS One, 2007. **2**(4): p. e383.

171. Ebersberger, I., S. Strauss, and A. von Haeseler, *HaMStR: profile hidden markov model based search for orthologs in ESTs.* BMC Evol Biol, 2009. **9**: p. 157.

172. Seetharam, A., Y. Bai, and G.W. Stuart, *A survey of well conserved families of C2H2 zinc-finger genes in Daphnia.* BMC Genomics, 2010. **11**: p. 276.

173. *HMMER3: a new generation of sequence homology search software.* 2010.

174. Pruitt, K.D., et al., *NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy.* Nucleic Acids Res, 2012. **40**(Database issue): p. D130-5.

175. Englbrecht, C.C., H. Schoof, and S. Bohm, *Conservation, diversification and expansion of C2H2 zinc finger proteins in the Arabidopsis thaliana genome.* BMC Genomics, 2004. **5**(1): p. 39.

176. Haerty, W., et al., *Comparative analysis of function and interaction of transcription factors in nematodes: extensive conservation of orthology coupled to rapid sequence evolution.* BMC Genomics, 2008. **9**: p. 399.

177. Materna, S.C., et al., *The C2H2 zinc finger genes of Strongylocentrotus purpuratus and their expression in embryonic development.* Dev Biol, 2006. **300**(1): p. 108-20.

178. Pruitt, K.D., et al., *NCBI Reference Sequences: current status, policy and new initiatives.* Nucleic Acids Res, 2009. **37**(Database issue): p. D32-6.

179. Boutet, E., et al., *UniProtKB/Swiss-Prot.* Methods Mol Biol, 2007. **406**: p. 89-112.

180. Iuchi, S. and N. Kuldell, *Zinc finger proteins : from atomic contact to cellular function*. Molecular biology intelligence unit2005, Georgetown, Tex.

New York: Landes Bioscience ;

Kluwer Academic/Plenum Publishers. 276 p.

181. Lespinet, O., et al., *The role of lineage-specific gene family expansion in the evolution of eukaryotes.* Genome Res, 2002. **12**(7): p. 1048-59.

182. Rubin, G., et al., *Comparative genomics of the eukaryotes.* Science, 2000. **287**: p. 2204 - 2215.

183. Chervitz, S.A., et al., *Comparison of the complete protein sets of worm and yeast: orthology and divergence.* Science, 1998. **282**(5396): p. 2022-8.

184. Fumasoni, I., et al., *Family expansion and gene rearrangements contributed to the functional specialization of PRDM genes in vertebrates.* BMC Evol Biol, 2007. **7**: p. 187.

185. Demuth, J.P., et al., *The evolution of mammalian gene families.* PLoS ONE, 2006. **1**: p. e85.