2010

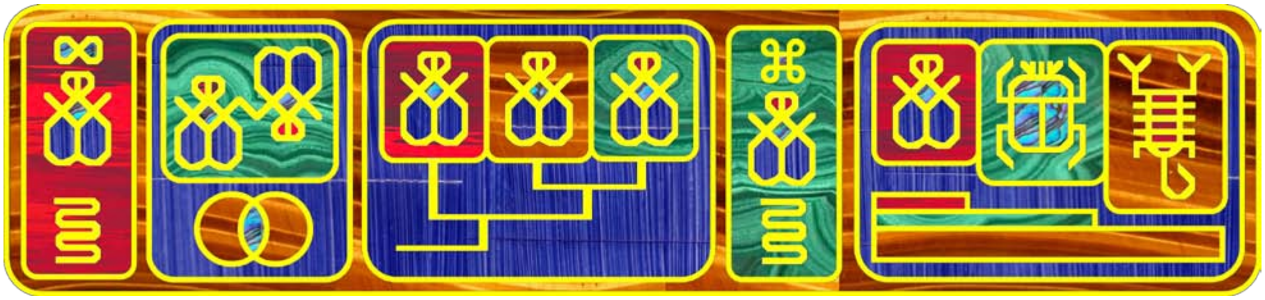
Network Analysis and Comparative Phylogenomics of MicroRNAs and their Respective Messenger RNA Targets Using Twelve Drosophila species

M Ryan Woodcock
Virginia Commonwealth University

Network Analysis and Comparative Phylogenomics of MicroRNAs

and their Respective Messenger RNA Targets

Using Twelve *Drosophila* species



By M. Ryan Woodcock, B.S.
Fall 2010

Integrative Life Science Ph.D. Program,
Virginia Commonwealth University (VCU)

Chairperson and Mentor:
Danail G. Bonchev, Ph.D., D.Sc.,

Committee Members:
Gregory M. Plunkett, Ph.D.,
Lemont Kier, Ph.D.,
Maria Rivera, Ph.D.,
J. M.Turbeville, Ph.D.,
Tarynn Witten, Ph.D.

TABLE of CONTENTS 2

ABSTRACT 5

THESIS OVERVIEW 7

ACKNOWLEDGEMENTS 9

CHAPTER I. MicroRNA Target Prediction and Network Properties in *Drosophila*

Abstract 13

Introduction 14

Methods 16

Results 23

Discussion 36

CHAPTER II. Drosophilid Patterns of MicroRNA Network Conservation across Interactions, Targets, and Species

Abstract 44

Introduction 45

Methods 46

Results 47

Discussion 68

CHAPTER III. Reconstruction of *Drosophila* Phylogeny using MicroRNA-Target Network Edges

Abstract 76

Introduction 77

Methods 78

Results 79

Discussion 85

CHAPER IV. MicroRNA-Target Network Topology across Regions of Chromosomal Synteny (Muller Elements) and Genes linked to species Diagnostic Phenotype in *Drosophila*.

Abtsract

Introduction 89

Methods 90

Results 91

Discussion 107

CHAPTER V. Nested Hierarchal Organization of Conservation for MicroRNAs and their Putative Targets to *Drosophila melanogaster*.

Abstract 117

Introduction 118

Methods 119

Results & Discussion 123

SUMMARY 136

REFERENCES 140

APPENDIX I. Selection of Multiple Sequence Alignment using Criteria from Phylogenetic Reconstruction 149

APPENDIX II. Example MicroRNA-Target Interactome Network 151

APPENDIX III. Network Adjacency, Distance Quantification, and Parametric Score Tables for Select MicroRNA-Target Networks of *Drosophila* 152

APPENDIX IV. Phylogenetic Reconstruction from Weighted MicroRNA-Target Network Edges 159

APPENDIX V. Phylogenetic Reconstructions using Protein-Protein Interactome Data 161

APPENDIX VI. Table of Life Stage, FlyBase Terms and Description of Features Available to Diagnose *Drosophila* species 166

APPENDIX VII. Phylogenies of Taxa Surveyed for Conservation of MicroRNAs and Targets to *Drosophila melanogaster* 172

Abstract

NETWORK ANALYSIS AND COMPARATIVE PHYLOGENOMICS OF MICRORNAS AND THEIR RESPECTIVE MESSENGER RNA TARGETS USING TWELVE *DROSOPHILA* SPECIES

By M. Ryan Woodcock, Ph.D.

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University

Virginia Commonwealth University, 2010

Danail G. Bonchev, Ph.D., D.Sc.,
Professor of Mathematical Sciences

MicroRNAs represent a special class of small (~21–25 nucleotides) non-coding RNA molecules which exert powerful post-transcriptional control over gene expression in eukaryotes. Indeed microRNAs likely represent the most abundant class of regulators in animal gene regulatory networks. This study describes the recovery and network analyses of a suite of homologous microRNA targets recovered through two different predicition methods for whole gene regions across twelve *Drosophila* species. Phylogenetic criteria under an accepted tree topology were used as a reference frame to 1) make inference into microRNA-target predictions, 2) study mathematical properties of microRNA-gene regulatory networks, 3) and conduct novel phylogenetic analyses using character data derived from weighted edges of the microRNA-target networks. This study investigates the evidences of natural selection and phylogenetic signatures inherent within the microRNA regulatory networks and quantifies time and mutation necessary to rewire a microRNA regulatory network. Selective factors that appear to operate upon seed aptamers include cooperativity (redundancy) of interactions and transcript length. Topological analyses of microRNA regulatory networks recovered significant enrichment for a motif possessing a redundant link in all twelve species sampled. This would suggest that optimization of the whole interactome topology itself has been historically subject to natural selection where resilience to attack have offered selective advantage. It seems that only a modest number of microRNA–mRNA interactions exhibit conservation over *Drosophila* cladogenesis. The

decrease in conserved microRNA-target interactions with increasing phylogenetic distance exhibited a cure typical of a saturation phenomena. Scale free properties of a network intersection of microRNA target predictions methods were found to transect taxonomic hierarchy.

THESIS OVERVIEW

MicroRNAs represent a special class of small (~21–25 nucleotides) non-coding RNA molecules which exert powerful post-transcriptional control over gene expression in eukaryotes. Indeed microRNAs likely represent the most abundant class of regulators in animal gene regulatory networks. In animals, microRNAs bind with partial to complete complementarity to target regions in the sequence of messenger RNAs. Subsequently, these interactions can forcibly regulate gene expression by stalling protein production from messenger RNAs or inducing wholesale breakdown of the messenger RNA itself. The outcome of these regulatory interactions appears to play key roles in processes of cellular differentiation, maintenance of tissue identity, and optimize genetic programming to confer robust tolerance against environmental fluctuations. MicroRNA regulatory networks are dense, with most messenger RNAs targeted by multiple microRNAs, which in turn enables precise coordinated control of targets and wide regulatory versatility.

MicroRNA genes have been broadly conserved across various animal groups and the ancestral acquisition of microRNAs appear strongly correlated to physical modifications of animal groups across geologic time. Therefore it seems that microRNA-mediated gene regulation in animals has likely played an essential role in the origins of complex body plans. The microRNA repertoire has no doubt modulated the use of a substantial fraction of the animal transcriptome and provides a crucial operator upon the sequence evolution of all messenger RNAs (Bartel & Chen, 2004). But while microRNA genes are themselves strongly conserved, the microRNA–target interactions seem to exhibit high plasticity across animal groups. Thus, while both microRNAs and target genes may be individually conserved, the interaction between the two elements may not be conserved.

In light of the proceeding, several global questions arise regarding the development of microRNA regulatory networks through natural history. What role may the length of the entire transcript play in the natural selection for microRNA targets? What respective relationships exist for conservation of regulatory network structure, is the network consistent throughout natural history or does structure alter in some fashion consistent with phylogenetic history? If the later is true, then is the signature of phylogeny found embedded within network structure?

The central purpose of this research thesis has been to study microRNA gene regulation, combining principles and bioinformatic tools of networks biology, comparative genomics, phylogeny, and whole genome data available for twelve closely related fruit fly (*Drosophila*) species. The fruit fly model (order Diptera: family Drosophilidae) represent as logical choice for study as it is arguably the best studied multi-cellular animal. It has formed the core of more than a century of biological study in phylogenetics and population genetics. Moreover, the available *Drosophila* genome sequences provide an unprecedented dataset to contrast conservation history, coding genes, and regulatory genes across the well-defined phylogeny of the sequenced species (*Drosophila* 12 Genomes Consortium, 2007). This study, presented in five CHAPTERS, investigates the evidences of natural selection and phylogenetic signatures inherent within the microRNA regulatory networks in twelve *Drosophila* species. It is among the goals of this project to address quantification of time and mutation necessary to rewire a microRNA regulatory network. Phylogenetic criteria under an accepted tree topology were used as a reference frame to 1) make inference into microRNA-target predictions, 2) study mathematical properties of microRNA-gene regulatory networks, 3) and conduct novel phylogenetic analyses using character data derived from weighted edges of the microRNA-target networks.

ACKNOWLEDGEMENTS

There is a Hawaiian saying that *'the knowledge is in the doing'*. Thus it seems that in most life projects there is no way to be told exactly how to accomplish the task beforehand; the necessary skills must be uncovered as one goes. To any given task you may already have most of the needed tools; but only half of the knowledge necessary to finish the purpose (Job 28:18).
In turn, scientific research progresses this way:

1) The first time the researcher attempts a task, it does not work; and he has no idea why.
2) The second time the researcher attempts the task, the task will not work; but at least the major design flaw has been found.
3) The third time the researcher attempts the task, the task still will not work; but at least the minor design flaw has been found.
4) The fourth time the researcher attempts the task, it finally works; but halfway through its working, he discovers that there was an easier way to accomplish the purpose form the outset!

Even so, while everything is two steps forward and one step back at the start; by the time of the end, the work gains momentum and moves, one step backward, then rolls three steps forward. Some of the best written inspiration comes at 3AM. These have been my research experiences.

It is a humbling prospect to think that all historical awareness of my person may come solely from the pages of this document. So for the sake of science, history, and equity (and hoping that somewhere, someone in the distant future may read these words), I endeavor to record a representative testimony and acknowledge the contributions of all those who breathed life into this project. Meanwhile the work continues with a view to last a lifetime; and I have other data which are not of this run; those also I must bring, and they must listen to my query, and they will become one table, one researcher (John 10:16).

This work has been my *Opus magnum muscae*. It has taken a full part of my vital energy (Ecclesiastes 9:10), a sacrifice of life, and still it has enriched me as a blessing. I have always trusted that the truth attends to itself and wisdom is proved righteous by its works; and time may tell where I for my part have chosen the good portion (Matthew 11:19; Luke 10:42). I am indebted to the many people who put their trust in my scientific potential and invested in me as

person. This work is dedicated to all those who saw me through every crisis of time, life, and faith (Luke 22:28).

Faith and reason are as the shoes on one's feet: a person will get farther in life with both of them than with just one or the other. And so I give first thanks in faith to my God, who by Jesus Christ answered my prayers and blessed the work of my hand. During this time I have lived and learned that as one of His Witnesses, I will experience things that a hundred ordinary lives could not accommodate. The hand of God is never too short to find good use for you exactly where you stand; and that hand will never place you where it cannot protect you (Psalm 139:8-10). Where I drew a line in the sand, He made it into a fortification; and even curses He turned into blessings (Nehemiah 13:2).

Everything begins with family and so to my father Melvin and my mother Linda Woodcock, and my sister Michelle Smith, and all the rest of my family, and my brothers and sisters in faith, I am grateful for their loving support. Friends are the only family that one can choose; and at times the only family to be chosen. And so I give thanks to my friends Danny Davis, Greg Melia, David Housman, Eric Wallenberg, Jared Hattaway, Chris Owens, and Charles Hicks. I have seven who call me their '*best friend'* , and I am the richest of all men (Proverbs 17:17). And to the love of my life who is not here to see the end of this work, wherever she may be, I declare that I continue to believe in a love so great that it will be worth the wait (Proverbs 31:10).

The Virginia Commonwealth University Integrative Life Science Doctoral program has provided a truly unique forum in which to build an integrative thesis project. I can not imagine that this special work could be conducted anywhere else; and I am truly grateful for the opportunities and support provided me. Special thanks are owed to Jean McNeil who worked behind the scenes for my aid; I may never know the full of the good work that she has done in my behalf (Matthew 6:3-4). My thanks also go out a host of fellow graduate students and allies namely: Aaron Aunins, Billy Budd, Dan Carr, Dr. Antoine Nicolas, Dr. Pedro Fiaschi, Dr. Ryan Garrick, Morgan Gostel, Melisse Ilhan, Sarah Chase Rothschild, Sterling Thomas, Chris Waggener, and VernWilliamson. Additional gratitude is due for the aid of Dr. Buck, Dr. Allison Johnson, and the entire Center for the Study of Biological Complexity. My thanks also goes out to Carlisle Childress and Nihar Sheth for cheerful technical support. Most singularly this endeavor is especially indebted to Logan Voegtly who provided invaluable programming support. Two people shorten a road and it was a great privilege to act as his student research mentor (Philippians 2:22).

Many thanks are due to Dr. Danail Bonchev, Dr. Greg Plunkett, Dr. Lemont Kier, Dr. Maria Rivera, Dr. Clint Turbeville, and Dr. Tarynn Witten for their involvement as graduate thesis committee members. Each of these individuals contributed a unique view, specialized expertise, and personal support. Doctor Danail Bonchev has served in an outstanding capacity. He is the best mentor for which a graduate student could ever ask. I am grateful to Greg Plunkett for providing me an epiphyte existence in his araliaceous lab; he gave me a welcoming place to lodge when I was without a home for research (Matthew 25:35). Doctor Lemont Kier has always been my advocate and through his unconquerable optimism he offers continued refreshment. I am thankful to Dr. Maria Rivera who has always provided a good-spirited filter of reality (Proverbs 9:8). Great appreciation is due to Dr. Clint Turbeville for his wise guidance in teaching (Proverbs 22:6). And Doctor Tarynn Witten has been a readily-available source of encouragement through our continuing series of philosophical discussions (Malachi 3:16).

Lastly it seems appropriate to acknowledge the fruit fly itself. For the record I hate fruit flies. But one does not need to love something to respect it. This wholly unremarkable insect was domesticated, not for its virtues as a model organism, but as a creature of last resort; and yet even so, it has become the work horse of modern biology and is arguably the world's best known organism. *Drosophila* has become living proof that the meek do inherit the Earth (Matthew 5:5).

CHAPTER I.

MicroRNA Target Prediction and Network Properties in *Drosophila*

Keywords & concepts: Bipartite Network; GRAFMAN Software; *Drosophila*; Messenger RNA Transcript Length; MicroRNA Aptamer Sequence Evolution; MicroRNA Target Prediction; MicroRNA-Target Network; MiRanda; TargetScan.

ABSTRACT

MicroRNA regulatory networks are dense with most target genes targeted by multiple microRNAs, and exhibit precise combinatorial control of targets giving increased regulatory versatility. This study describes the recovery and network analyses of a suite of homologous microRNA targets recovered through two different predicition methods for whole gene regions across twelve *Drosophila* species. Data recovered from microRNA target prediction were integrated with data from taxonomic hierarchical conservation and molecular phylogeny through a MySQL database of linked tables called "*musca*". TargetScan output (61.9GB) recovered a network of 14,860 targets, 1,090,221 microRNA-target interactions, 11,302,034 unique aptamer site interactions, and 112 microRNA families. Output form the MiRanda algorithm (2.96GB) recovered a network of 14,583 targets, 241,861 microRNA-target interactions, 390,560 unique aptamer site interactions, and 121 microRNA families. The network intersection of target prediction methods recovered a network of 12,616 targets, 78,280 microRNA-target interactions, 226,270 unique aptamer site interactions, and 112 microRNAs. The intersection of microRNA target prediction methods produced networks of increased potential biological relevance compared to respective parent networks. The sizable target datasets produced in this study are applicable for continuing research in *Drosophila* molecular biology and could be biochemically verified using whole genome microarray analyses and miRNP immunopurification. Moreover differential microRNA enrichment patterns by prediction method would seem to indicate that selective factors presiding over regulation by compensatory aptamers (MiRanda) and seed regions aptamers (TargetScan) are different. Selective factors that appear to operate upon seed aptamers include cooperativity (redundancy) of interactions and transcript length. As transcript length increases the likelihood of acquisition of a seed-type aptamer binding site also increases.

INTRODUCTION

MicroRNAs represent short (generally 21–25 nucleotides) endogenously expressed single-stranded non-coding RNA molecules derived from larger stem-loop precursors. In animals, microRNAs bind with variable complementarity to the aligned 3’ untranslated regions (3’UTR) and other sites of target messenger RNAs (mRNAs); subsequently gene expression is regulated by mRNA cleavage induction or translational rate control using the cells innate RNA-Interference (RNAi) pathway (Lu *et al.*, 2008; Stark *et al.*, 2007b). Each microRNA may target transcripts for hundreds of genes (that are unrelated to the loci that encode the microRNAs themselves), but multiple microRNAs might need to bind to a particular transcript to achieve repression; in principle, this could even be accomplished by the combinatorial action of different microRNA species (Bartel & Chen, 2004; Lewis *et al.*, 2005). Thus the microRNA milieu, unique to each cell type, productively dampens the expression of thousands of gene transcripts and provides important natural selective force operative upon all metazoan messenger RNA sequences (Bartel & Chen, 2004). The biological importance of microRNA is evidenced by high conservation across phylogeny and by the many life processes in which they are implicated; including developmental timing, cell proliferation, apoptosis, metabolism, cell differentiation, and morphogenesis (Ambros, 2004; Bartel & Chen, 2004; Stark *et al.*, 2005). Indeed, MicroRNAs probably represent the most abundant classes of regulators of animal gene networks (Sempere *et al.*, 2007, Stark *et al.*, 2007a). Approximately 20% of transcription in *Drosophila melanogaster* seems to be unassociated with protein-coding genes; and microRNA expression would be included among the later value (*Drosophila* 12 Genomes Consortium, 2007).

Genome-wide metazoan microRNA target predictions indicate that thousands of genes (perhaps 20-60% of all genes) are likely to come under regulation (Friedman *et al.*, 2009; Lewis *et al.*, 2005; Stark *et al.*, 2005). Experimental evidence indicates that the most crucial aspect to microRNA-target hybridization is a seed region (~7 nucleotides) on the 5’end of a mature microRNA (Grün *et al.*, 2005; Lewis *et al.*, 2005; Lu *et al.*, 2008; Rajewsky, 2006; Wang *et al.*, 2008). Admittedly, there is a continuum of 3' pairing quality between and within microRNA aptamers, but principally microRNA-target site interactions can be classified into three type classes: 5’-dominant canonical, 5’-dominant seed only and 3’-compensatory (Brennecke *et al*., 2005; Sethupathy, *et al.*, 2006). Canonical sites have perfect complementarity to the seed portion of the 5’ end of the microRNA and extensive base pairing along the 3’ end of the microRNA (Brennecke *et al*., 2005; Sethupathy, *et al.*, 2006). Conversely, seed-type interactions have perfect base pairing to the seed portion of the 5’ end of the microRNA but limited base pairing along the 3’ end of the microRNA (Brennecke *et al*., 2005). Lastly, the 3’-compensatory sites have extensive base pairing along the 3’ end of the microRNA to compensate for imperfect or a shorter stretch of base pairing to the seed portion of the microRNA (Brennecke *et al*., 2005; Sethupathy, *et al.*, 2006).

Available multiple species microRNA target predictions for *Drosophila* have been extensively drawn from alignments of 3’UTRs (Grün *et al.*, 2005; Huynh *et al.*, 2006; Megraw *et al.*, 2007). There are, however, reasons to suspect that these predicted targets represent only a small fraction of the total targets and have likely overlooked a sizeable body of important microRNA targets present in protein coding regions and potentially present in 5’UTRs and introns (Bartel & Chen, 2004; Grün *et al.*, 2005, Kheradpour *et al.*, 2007, Lytle *et al.*, 2007, Smalheiser & Torvik, 2006, Stark *et al.*, 2007b). Indeed drosophilid heptamers complementary to different positions in mature microRNAs demonstrate a distinctive conservation pattern; indicative of functional targeting in coding regions and similar to that found in 3’UTRs (correlation coefficient 0.96; Stark *et al.*, 2007b). These microRNA motifs exhibit high conservation in all three reading frames; suggesting that they are specifically selected within coding regions for their RNA-level function. Likewise, other studies have shown that microRNA motifs in coding regions are preferentially conserved in vertebrates, can lead to repression in experimental assays, and are avoided in genes co-expressed with the microRNA (Farh *et al.*, 2005; Grimson *et al.*, 2007; Kloosterman *et al.*, 2004; Lewis *et al.*, 2005).

This study details the recovery and network analyses of a suite of homologous microRNA targets recovered for whole gene regions across twelve *Drosophila* species. In particular these data will be valuable for comparison to the microRNA predictions for seven *Drosophila* species drawn from older genome alignments of 3’UTRs, to complement twelve species comparisons using only 3’UTR regions, and to complement whole genome microRNA target predictions prepared under other methods (Grün *et al.*, 2005; Kheradpour *et al.*, 2007; Stark *et al.*, 2007b). To these ends, the central software selected for microRNA target prediction were MiRanda and TargetScan (Enright *et al.*, 2003; Lewis *et al.*, 2005). These tools were chosen in light of performance review comparing sensitivity and specificity of five microRNA target prediction methods using verified interaction data where TargetScan, PicTar and MiRanda recovered best performance, with sensitivity values ranging between 65% and 68% (Grün *et al.*, 2005, Maziere & Enright, 2007, Sethupathy, *et al.*, 2006).

METHODS

Selection of a *Drosophila* Multiple Sequence Alignment. The initiation of this project required selection of a suitable multiple sequence alignment set for the twelve *Drosophila* species using phylogenetic criteria (APPENDIX I, TABLE 8). Nine separate multiple sequence alignments were produced for a trial data set of eleven *Drosophila* genes coding for enzymes involved in glycolysis (Clark & Wang, 1994). These genes were selected as a proxy sample of the *Drosophila* genome on the basis of their established use in allozyme studies, expected selective neutrality, and general consistency between molecular phylogeny and metabolic character data (Burkhart *et al.*, 1984; Clark & Wang, 1994; Ko, *et al.*, 2003; Pollard *et al.*, 2006). Gene regions were extracted directly from three published multiple sequence alignments for twelve *Drosophila* species; namely PECAN/Mercator, MAVID/Mercator, and MULTIZ alignments available from UCSC (Blanchette *et al.*, 2004; Dewey, 2007, Kent *et al.*, 2002; Paten *et al.*, 2008; Stark *et al.*, 2007b). PECAN is a consistency based multiple-alignment program that favors global optimization of alignment while still inherently working in a pairwise fashion (Paten *et al.*, 2008). The MAVID program utilizes a progressive-alignment approach incorporating maximum-likelihood inference of ancestral sequences, automatic guide-tree construction, and constraints derived from a global homology map of the sequences (Bray & Pachter, 2004). Mercator represents an orthology mapping method designed to identify blocks of synteny (conserved gene order) from pairwise similarity scores between sets of non-overlapping genome (Dewey & Pachter, 2006). The MULTIZ approach uses pairwise alignments of orthologous sequences produced by BLASTZ, filters these to select the best matches to specified reference sequences, and conducts guided alignment of multiple reference blocksets in order to recover a union blockset of the original query sequences (Blanchette *et al.*, 2004).

Novel reconstructions were produced to isolate alignment method variables, enable comparison of published alignments, and to evaluate the prospective value of generating novel large-scale genome reconciliations for the molecular analyses of the project. Three additional multiple sequence alignments were produced through 1) novel re-alignment of MULTIZ sequence data using the multiple sequence alignment program MAFFT, 2) production of a three-way reconciliation of whole published alignments in one-step with production of a consensus sequence, and 3) through three-way reconciliation of published alignments in a species-by-species (twelve separate sub-alignments) manner and recovering a sequence consensus (Katoh & Toh, 2008; see APPENDIX I, TABLE 8). Multiple sequence alignment reconciliation was implemented through ClustalW using the BioEdit freeware package where input gaps were locked and gap insertion penalties set to zero (Hall, 1999). Any resultant gap-only positions were extracted using the MEGA software package (Kumar *et al.*, 2008). These methods recovered a reconciled alignment geometry accommodating the original structure of all input alignments.

Given the fundamental premise that alignment position of infers homology, tree metrics were utilized to provide a quantitative means to evaluate alignments against one another on the basis of internal consistency, resolution of tree topology, support for the reference tree (CHAPTER III, FIGURE 22). Criteria from phylogenetic reconstruction considered in selection of a multiple sequence alignment were evaluated for sequence information content and for statistical best fit to one of 56 models of nucleotide sequence evolution through PAUP* and Modeltest (Posada & Crandall, 1998; Swofford, 2002). Phylogenetic reconstructions from all alignments were conducted under distance with neighbor-joining and standard parsimony using PAUP*, and under Bayesian inference through MrBayes software (Huelsenbeck & Ronquist, 2005; Swofford, 2002). The suite of resulting phylogenetic trees were statistically evaluated to high confidence limits using bootstrap, jackknife, posterior probability, tree consistency indices, likelihood score, sensitivity to concavity-parameter alteration (k=0, 10, 100), topology-dependent permutation test against the established *Drosophila* phylogeny, and partition homogeneity of component genes in the dataset (Bull *et al.*, 1993; Faith, 1991; Goloboff, 1993).

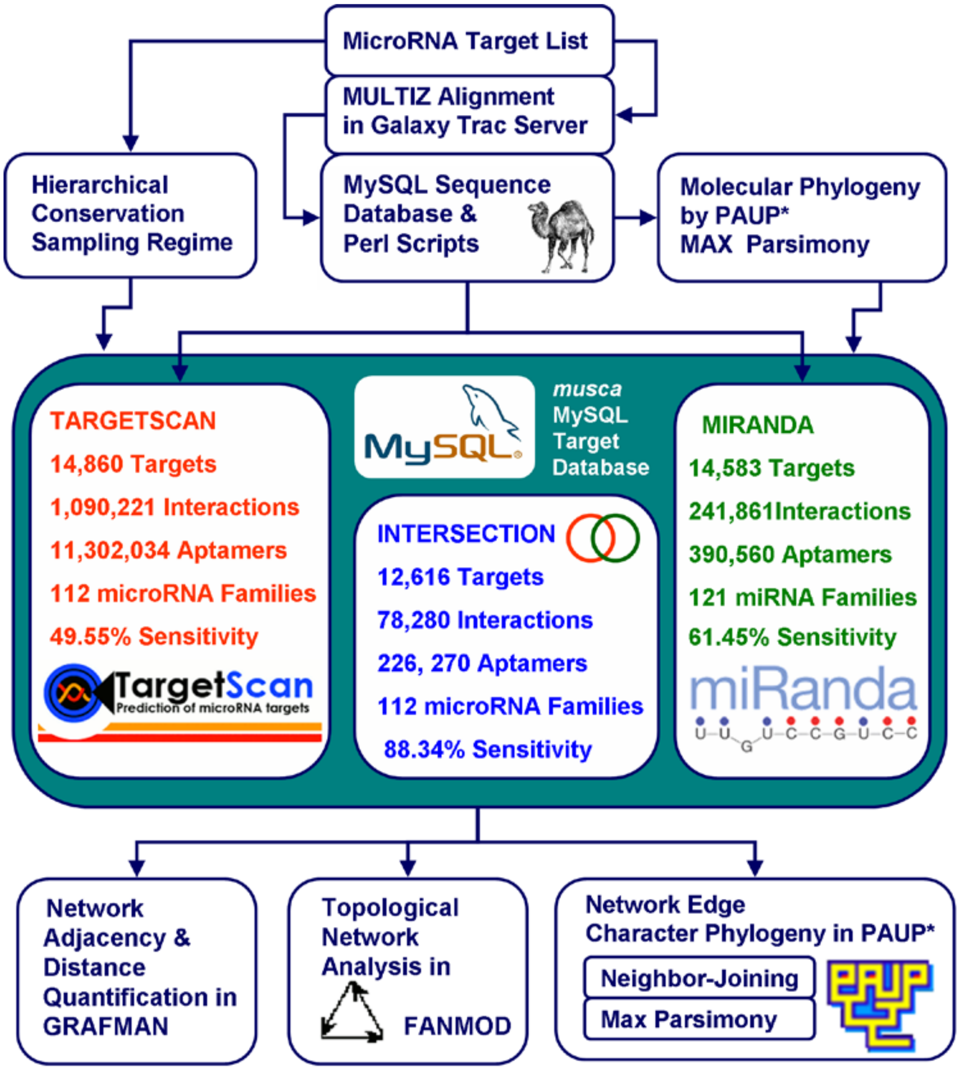


FIGURE 1. Flowchart of bioinformatic tools and processing for microRNA-target data. Methods appear boxed to match specific project aims and arrows indicate dataflow originating from a list of 14,925 microRNA targets. Criteria from phylogenetic reconstruction considered in selection of the MULTIZ multiple sequence alignment appear in APPENDIX I, TABLE 8. A MySQL database of linked tables called "*musca*" acted as central repository for parameters recovered from microRNA target prediction, hierarchical conservation, and molecular phylogeny. Elementary microRNA-Target network properties are described in CHAPTER I, whereas CHAPTERS II, III, IV, and V of this document detail properties of microRNA-Target conservation, natural history and phylogeny. Data output from network adjacency and distance quantification in GRAFMAN are extensively utilized in CHAPTER I, II, & V. Natural historical commentary relevant to analysis of network topology through FANMOD are given in

CHAPTER IV. Output data of molecular phylogenetic reconstructions are utilized in CHAPTERS II & IV, while phylogenetic reconstructions using network edges are described in CHAPTER III. Both molecular and network edge phylogenies using PAUP* made comparison to the reference tree illustrated in CHAPTER III, FIGURE 22. The hierarchical conservation sampling regime of microRNAs and targets is displayed in FIGURE 27 and discussed in CHAPTER V.

Primary Data Manipulation. Project methodology followed the generalized flowchart illustrated in FIGURE 1 subsequent to multiple sequence alignment selection.
The number of predicted targets varies considerably according to method with only limited overlap in the top-ranking targets, indicating that individual methods might only capture subsets of real targets and/or may include a high number of background matches (Brennecke *et al.*, 2005). Thus all accessible microRNA target predictions were treated as potentially complementary and a total non-redundant set 16,204 putative microRNA targets was generated (Ambros, 2004; Megraw *et al.*, 2007; Rajewsky, 2006). The total dataset included the union of nearly 50 published target prediction sets from *D. melanogaster* alone (15,016 genes recovered; 92.67% of the total). The prospective microRNA target gene list was further supplemented through interologous extrapolation from cross species interactant comparison using nearly 300 individual microRNA target prediction sets published for 22 different organisms compiled and converted into *Drosophila* homologs with the aid of BioMART server for Ensembl (Flannick *et al.*, 2006; Matthews *et al.,* 2001; Sharan & Ideker, 2006; Smedley, *et al.*, 2009; Suthram *et al.*, 2005). Chromosomal locations of known or putative microRNAs were isolated in the whole genome sequence for all twelve *Drosophila* species through the DroSpeGe database and genes adjacent to microRNA loci within 50 kilobases (1,690 genes; 121 clusters found) were likewise catalogued into the putative microRNA target list (Gilbert, 2007). Impetus for the later search strategy was taken from experimental evidence for frequent co-expression of microRNAs with neighboring genes (Baskerville & Bartel, 2005).

Large-scale bioinformatic data processing was initiated to extract multiple sequence alignments data for all putative microRNA target genes. Genomic coordinates of *D. melanogaster* for gene regions (including 5’UTRs, coding sequence, introns, and 3’UTRs) for 15,082 non-redundant genes of the total putative microRNA target were extracted through FlyBase batch download and the BioMart server (Smedley, *et al.*, 2009, Wilson *et al.*, 2008). These data were used to define extraction regions from the whole *Drosophila* genome alignment accessible through the MULTIZ table of the UCSC Genome Bioinformatics Site (Kuhn *et al.*, 2007). The extracted output were uploaded directly to the Galaxy server (Giardine *et al.*, 2005). This server is a refined interface allowing users to conduct and save independent queries of genomic data from different sources, and perform sequential large-scale calculations using operators such as join, union, intersection, and subtraction (Giardine *et al.*, 2005). A gene region output was recovered where gap-only columns and non-*Drosophila* species were removed and only species of interest were included, ordered consistently, and oriented according to open reading frames. Thereafter, 1,202,106 KB of Galaxy output was modified using a MySQL database and Perl scripts to extract alignment data for user-defined regions of interest and produce files formatted for microRNA target prediction (FIGURE 1; Giardine *et al.*, 2005; Sun Microsystems, Inc. 2008-2009).

MicroRNA Target Prediction. Sequence data for each putative target gene was subjected to batch microRNA target predication for 121 drosophilid microRNA families using MiRanda and TargetScan (Enright *et al.*, 2003; Lewis *et al.*, 2005). Both target prediction methods were produced through the Fenn supercomputing cluster of Virginia Commonwealth University. There were 830 MB of MULTIZ extracted FASTA sequences input into MiRanda. Likewise, 7.04 GB of tab delimited data were input for TargetScan. The dynamic programming algorithm MiRanda examined canonical microRNA interactions by optimizing and recovering all non-overlapping hybridization alignments between microRNA and input sequence according to the nucleotide complementarity score set to some user-defined cutoff value (Enright *et al.*, 2003). The default setting for Gibbs free energy of nucleotide hybridization in MiRanda is ΔG = -20 kcal/mol (Enright *et al.*, 2003). To ensure higher target specificity in these analyses, the hybridization energy threshold for MiRanda was set to ΔG = -25 kcal/mol. There were 146 mature microRNA sequences input into MiRanda and correspondingly there were 121 microRNA families input into TargetScan. The TargetScan algorithm predicted microRNA targets by searching for the presence of conserved octamer and heptamer sites matching a microRNA seed region (Lewis *et al.*, 2005; Sethupathy, *et al.*, 2006). Notably, TargetScan features an efficient reduction in the false-positive rate, but an increased false-negative rate due to requirements of strict complementation in the seed region (Maziere & Enright, 2007). There were 2.9 GB and 12.0 GB of microRNA target prediction data recovered from MiRanda and TargetScan respectively.

Additional microRNA target predictions were conducted through MiRanda and TargetScan using the entire putative target gene and 13 deuterostome-specific microRNAs from *Homo sapiens* with no known homologs to *Drosophila* (Berezikov *et al.*, 2010; Gilbert, 2007; Griffiths-Jones *et al.*, 2006; Hertel *et al.*, 2006; Lu *et al.*, 2008; Sempere *et al.*, 2007; Sethupathy, *et al.*, 2006). Specifically, the deuterostome microRNAs input were: *hsa-miR-126, hsa-miR-135a, hsa-miR-141, hsa-miR-148a, hsa-miR-153, hsa-miR-183, hsa-miR-200b, hsa-miR-21, hsa-miR-216a, hsa-miR-217, hsa-miR-338-3p, hsa-miR-93,* and *hsa-miR-96*. These analyses functioned as a provisional negative control of the target prediction methods where any *Drosophila* target to a *hsa-miR* would represent a false positive. MiRanda and TargetScan respectively produced 164 KB and 2.96 MB of output. Computational performance of target prediction methods was gauged for sensitivity from the proceeding output. Sensitivity was defined by dividing the average targets recovered per endogenous *Drosophila* microRNAs by the sum of the averages retrieved both alien and endogenous microRNAs ({sensitivit*y* = average targets per *dme-miR* / (average targets per *dme-miR* + average targets per *hsa-miR*) }; compare Sethupathy, *et al.*, 2006).

MicroRNA Regulatory Network Adjacency & Distance Quantification. A MySQL database of linked tables called *"musca*" was created to act as central repository for microRNA target prediction data and hierarchical conservation, and molecular phylogeny parameters recovered in other analyses (Sun Microsystems, Inc. 2008-2009). Collectively, 31 GB of data were input data into *musca*, recovering a database of 33,818,624 KB. A total of 41 regulatory networks in 111 MB were formatted out of the *musca* database for TargetScan, MiRanda and the intersection of methods and analyzed for descriptive properties. The network quantification and distance analyses were performed using in-house GRAFMAN software available under Linux on the Watson supercomputer cluster of Virginia Commonwealth University (FIGURE 1; APPENDIX III, TABLE 9; Karabunarliev & Bonchev, 2002). Connectivity-based network descriptors calculated through GRAFMAN included: total number of vertices, total number of edges, total adjacency, average vertex degree, network connectedness, and information index for vertex degree distribution. Likewise, distance-based measurements calculated through GRAFMAN included: total distance, average distance per node (network radius), average distance, Shannon information index, and information index for distance degree distribution. The vertex degree defines the number of interactions (or the number of nearest neighbors). Two vertices are adjacent where an edge exists between them. An adjacency matrix represents a table that encodes the directed structure of a network. Network connectedness quantifies the density of a network. Node distance or vertex distance degree represents the sum of distances from an individual node to all other vertices in the network. Thus the distance between two non-neighboring nodes is equal to the number of edges along the shortest path that connects them. Consequently the total distance of the graph is defined as the sum of distances between all pairs of vertices. The radius of a network is the smallest eccentricity of any vertex where the eccentricity of a vertex is the length of the longest minimal path from that vertex to some vertex in the graph. A path is not minimal if the two vertices at its endpoints could be connected by a shorter path. Information is innate to any system in which elements can be grouped according to one or more criteria. This Information is a measure of system's diversity; the more complex a system is the larger its innate information content (Bonchev, 1983; Shannon & Weaver, 1949). Total GRAFMAN output was represented in 37.8 KB.

RESULTS

Molecular source data for this project was selected from the MULTIZ sequence alignment published for *Drosophila* on the basis of the strict consensus of tree metrics recovered for a trial data set of eleven coding genes coding for glycolytic enzymes (Blanchette *et al.*, 2004 Clark & Wang, 1994). The consensus of all test criteria favored the MULTIZ alignment over other published and novel reconciled alignments (APPENDIX I, TABLE 8). A non-redundant set of 16,204 putative microRNA targets was generated for *Drosophila* through extensive literature review and the compilation of 1572 separate datasets. A set of 15,082 non-redundant genes of the total putative microRNA target could be extracted through FlyBase batch download and the BioMart server. There were 14,925 genes which recovered target prediction data through MiRanda and/or TargetScan and of these, 95.11% (14,195) correspond to protein coding genes.

Networks generated were strictly bipartite in which a node (or vertex or point) may represent an individual microRNA or a target gene transcript. The flow of information *in vivo* progresses from microRNA to target gene transcript and therefore all networks recovered must be represented as a directed graph (or directed network). Accordingly, all interactions (links or edges) were directed solely from microRNA to target. An example illustration of a microRNA-target interaction network a representing 1.66% of the total microRNA target dataset is contained in FIGURE 37 of APPENDIX II. The term "*aptamer*" describes an individual binding site interaction between a microRNA and a select target gene transcript region. Thus, there may be multiple aptamers per individual target gene transcript and microRNA-aptamer networks describe each unique microRNA-aptamer interaction. Conversely, microRNA-target networks consider only the presence or absence of any interaction (inferred regulation) regardless of transcript binding region(s). Consequently, all aptamer interactions per target transcript become synonymized in microRNA-target networks. The distribution of network nodes according to the number of their connections is illustrated for targets and aptamers in FIGURE 2 and FIGURE 4 respectively. A double-logarithmic plot of data illustrated in FIGUREs 2C and 2F is presented in FIGURE 3. A comparison of numbers of unique microRNAs to numbers of aptamer sites observed per target transcript is presented by method across the union of twelve *Drosophila* species in FIGURE 5. Likewise a comparison of target transcript nucleotide length to the numbers of unique microRNA regulators observed is presented by method across the union of twelve *Drosophila* species in FIGURE 6. It is of note that similar analyses have only considered the 3'UTR of *Drosophila*; conversely this study examines nucleotide length for the entire messenger RNA transcripts (Stark et al., 2005). Similarly a comparison of target transcript nucleotide length to numbers of aptamer sites observed is presented in FIGURE 7. Additionally, a percent target distribution profile of microRNAs by prediction method is illustrated in FIGURE 8.

TargetScan output (61.9GB) recovered a network of 14,860 targets, 1,090,221 microRNA-target interactions, 11,302,034 unique aptamer site interactions, and 112 microRNA families across the union of twelve *Drosophila* species (FIGURE 1). TargetScan percent network composition per single microRNA ranged from 0.59 to 1.12%. When considering the twelve species individually, the numbers of targets under regulation per single microRNA ranged from 1,145 to 7,166 with an average of 3,885.55. However the union of twelve *Drosophila* species was substantially enriched with a range of 6,379 to 12,193 targets per microRNA; with 9,734.11 as an average. There were 9 microRNA families not recovered through TargetScan; namely: *dme-miR-1003/1004, dme-miR-10-3p/1006, dme-miR-275/306, dme-miR-279/286/996, dme-miR-285/995/998, dme-miR-2a-1/6/11/13/306, dme-miR-2a-2/2c, dme miR-3/309/318,* and *dme-miR-92/310/311/312/31.*

Network descriptors for TargetScan across the union of twelve *Drosophila* species are presented in APPENDIX III, TABLE 9. The total network adjacency was 2180442. The average TargetScan target vertex degree distribution was 146.732. The TargetScan network connectivity was 0.00987498. The Shannon information index for the *Drosophila* union TargetScan network was 4.56 x10^8 bits. The total TargetScan network distance was 445,660,532 and the network radius was 29,990.6. Thus a minimum path of 29,990.6 steps are required to transect the entire network. The average distance per target in the TargetScan network was 2.02; thus any given set of target transcripts are removed from one another by an average of roughly two microRNA regulators.

The overlap for genes undergoing microRNA regulation of the TargetScan dataset to previously published microRNA prediction sets ranged from 94.86 to 96.81%. These later data were for *Drosophila melanogaster* from MiRGen, PicTar, and RNA22 represented 6.37 to 23.13% of the total candidate microRNA targets (Grün *et al.*, 2005; Huynh *et al.*, 2006; Megraw *et al.*, 2007). There were 10.82% of the TargetScan targets possessing an aptamer sites in 3'UTR for *D. melanogaster*. Likewise, 4.91% of TargetScan targets in *D. melanogaster* uncovered an aptamer in the 5'UTR. Control data for the TargetScan union of twelve *Drosophila* species using 13 deuterostome microRNAs recovered a network of 14,358 vertices, and 128,821 interactions (APPENDIX III, TABLE 9). The numbers of targets per alien microRNA ranged from 6,050 to 11,535, with 9,909.32 as an average. Thus the calculated computational sensitivity of TargetScan was 49.55%.

MiRanda output (2.96GB) recovered a network of 14,583 targets, 241,861 microRNA-target interactions, 390,560 unique aptamer site interactions, and 121 microRNA families across the union of twelve *Drosophila* species (FIGURE 1). The percent network composition per single microRNA ranged from 0.01 to 3.23%. For individual *Drosophila* species, the numbers of targets under regulation per single microRNA ranged from 0 to 1,052 with an average of 205.92. The union of twelve *Drosophila* species was enriched with a range of 14 to 8,624 targets per microRNA, with an average of 2384.75. Using a hybridization energy cutoff of -25 kcal/mol, an average ΔG of -26.4 kcal/mol was recovered across the union of twelve *Drosophila* species. For *Drosophila melanogaster*, 3.94% of MiRanda targets held an aptamer binding region within 3UTR. Conversely, 4.24% of *D. melanogaster* targets had aptamers recovered within 5UTRs. There was a 25.04 to 27.81% overlap for MiRanda to published microRNA prediction data (Grün *et al.*, 2005; Huynh *et al.*, 2006; Megraw *et al.*, 2007). Control data by MiRanda for the union of twelve *Drosophila* species using 13 microRNAs of *Homo sapiens* recovered a network of 10,242 vertices, and, 19,450 interactions (APPENDIX III, TABLE 9). The numbers of targets per alien microRNA ranged from 357 to 3,892, with 1,496.15 as an average. Thus the calculated computational sensitivity of MiRanda was 61.45 %.

Network descriptors for MiRanda across the union of twelve *Drosophila* species are presented in APPENDIX III, TABLE 9. The total network adjacency was 483,722. The average MiRanda target vertex degree was 33.1703. The MiRanda network connectivity was 0.00227474. The Shannon information index for the *Drosophila* union MiRanda network was 5.23 x10^8 bits. The total MiRanda network distance was 458,029,284 and the network radius was 31,408.4.The average distance per target in the MiRanda network was 2.15392.

Intersection of target prediction methods across the union of twelve *Drosophila* species recovered a network of 12,616 targets, 78,280 microRNA-target interactions, 226,270 unique aptamer site interactions, and 112 microRNAs (APPENDIX III, TABLE 9). Intersection percent network composition per single microRNA ranged from 0.014 to 3.71%. Notably, both parent methods and their intersection recovered the same average network composition per single microRNA regulator of an average of 0.89%. The numbers of targets under regulation per single microRNA in individual species ranged from 0 to 412 with an average of 376.84. The union of twelve *Drosophila* species was enriched with a target range of 11 to 2,901 targets per microRNA; with 698.93 as an average. Aptamer binding sites in the 3UTR were recovered for 2.02% the targets in the *D. melanogaster* intersection set. Similarly, aptamer binding sites in 5UTRs of *D. melanogaster* were represented in 2.51% of the intersection target dataset.

Network descriptors for the intersection of methods across the union of twelve *Drosophila* species are located in APPENDIX III, TABLE 9. The total network adjacency was 156,560; this was a respective 327,162 and 2,023,882 step decrease for MiRanda and TargetScan. The average intersection target vertex degree was 12.41; a 20.76 and 134.32 degree reduction over MiRanda and TargetScan respectively. The intersection network connectivity reduced from parent methods to 0.000983721. The total intersection network distance was 52,6424,944; this was an increase of 68,395,660 and 80,764,412 steps for MiRanda and TargetScan respectively. The intersection network radius was 41,726.8; this represented a 10318.4 and 11,736.2 step increase from MiRanda and TargetScan. The average distance per target increased from roughly 2 to 3.31 with the intersection of methods. Thus, any two given target transcripts are removed from one another by an average of three microRNA regulators. The Shannon information index for the *Drosophila* union TargetScan network was 9.42x10^8 bits. The later value represents a 4.19 x10^8 and 4.86 x 0^8 bit information (complexity) increase over the MiRanda and TargetScan networks.

There was a 6.24% overlap in aptamer sites between the TargetScan and MiRanda datasets and a 74.97% overlap recovered for the number of genes undergoing microRNA regulation. The seed regions of multiple microRNAs in the same family would be homologous (and presumably functionally equivalent) and thus the network intersection recovers multiple hits for MiRanda (from mature microRNAs of individual species) to TargetScan aptamers (microRNA seed regions). Three-way intersection of MiRanda and TargetScan to other microRNA target datasets recovered a 16.70 to 18.64% overlap (Grün *et al.*, 2005; Huynh *et al.*, 2006; Megraw *et al.*, 2007). The numbers of targets per *hsa-miR* ranged from 33 to 211, with an average of 92.23. The calculated computational sensitivity of network intersection of MiRanda and TargetScan was 88.34%; a 26.89 and 38.79% increase over parent respective method sensitivity.

The TargetScan prediction dataset include microRNA-target interactions of both the 5'-dominant canonical and 5'-dominant seed only types (Sethupathy, *et al.*, 2006). MiRanda target predictions would include both the 5'-dominant canonical and 3'-compensatory microRNA-target interactions having extensive base pairing along the 3' end of the microRNA (Sethupathy, *et al.*, 2006). The network intersection of MiRanda and TargetScan prediction methods defines a set of 78,280 microRNA interactions with perfect complementarity to the seed portion of the 5' end of the microRNA and extensive base pairing along the 3' end of the microRNA; namely the 5'-dominant canonical microRNA targets. Conversely, those 163,581 microRNA interactions included in MiRanda but not in the intersection of methods would be those having extensive base pairing to the 3' end of the microRNA to compensate for imperfect or a shorter stretch of base pairing to the seed portion of the microRNA; namely, to the 3'-compensatory sites. Likewise, those 1,011,941 microRNA interactions included in TargetScan but excluded from the intersection of methods would be of the 5'-dominant seed only type.

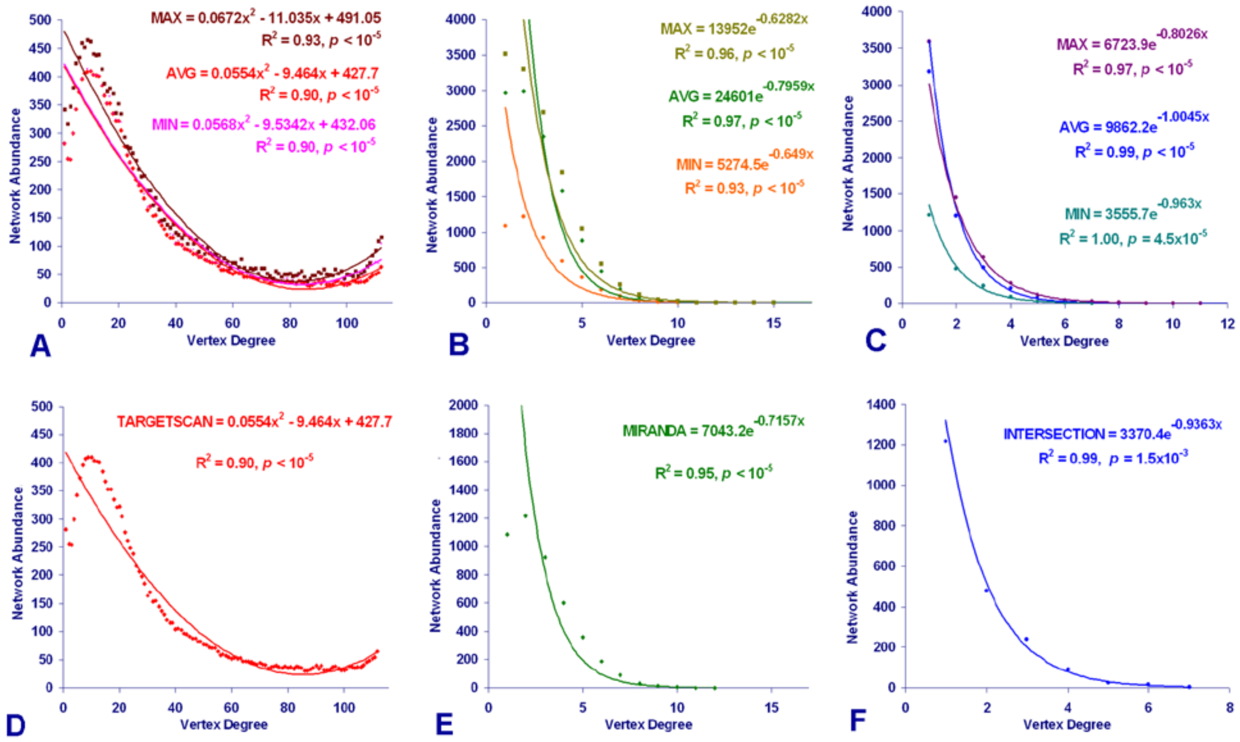


FIGURE 2. Vertex degree distribution and network abundance of microRNAs per target gene. MicroRNA targets are predicted across twelve *Drosophila* species according to (A) TargetScan (B) MiRanda and (C) the network intersection of methods. Likewise microRNA targets networks predicted for *D. melanogaster* alone are also presented for (D) TargetScan (E) MiRanda and (E) the intersection of methods. All data are unbinned. MicroRNA targets predicted across the union of twelve *Drosophila* species included:
(A) 14,760 from TargetScan; (B) 14,462 from MiRanda; and (C) 12,498 targets from the network intersection of methods. Unique microRNA-target interactions predicted across the union of twelve *Drosophila* species included: (A) 1,090,221 from TargetScan; (B) 241,861 from MiRanda; and (C) 78, 280 from the network intersection of methods. Power-law trend lines with functions for the target maxima across comparison of twelve *Drosophila* species recovered non-linear regression coefficients of (A) TargetScan: $R^2 = 0.77$, $p < 10^{-5}$; (B) MiRanda: $R^2 = 0.81$, $p < 10^{-5}$; and (C) for the network intersection of methods: $R^2 = 0.90$, $p < 10^{-5}$. Trend line functions and non-linear regression coefficient of determination were recovered for the target maxima across comparison of twelve *Drosophila* species for (A) TargetScan: $y = 0.0672\, x^2 - 11.035x + 491.05$, $R^2 = 0.93$, $p < 10^{-5}$; (B) MiRanda: $y = 13952e^{-0.6282x}$, $R^2 = 0.96$, $p < 10^{-5}$; and (C) for the network intersection of methods: $y = 6723.9e^{-0.8026x}$, $R^2 = 0.97$, $p < 10^{-5}$. Power-law trend lines

with non-linear regressions for target minima across comparison of twelve *Drosophila* species were as follows: (A) TargetScan: $R^2 = 0.7372$, $p < 10^{-5}$; (B) MiRanda: $R^2 = 0.71$, $p < 10^{-5}$; and (C) for the network intersection of methods: $R^2 = 0.90$, $p < 10^{-5}$. Target minima across comparison of twelve *Drosophila* species recovered trend lines with functions and non-linear regressions as follows: (A) TargetScan: $y = 0.0568\, x^2 - 9.5342\, x + 432.06$, $R^2 = 0.90$, $p < 10^{-5}$; (B) MiRanda: $y = 5274.5e^{-0.649x}$, $R^2 = 0.93$, $p < 10^{-5}$; and (C) for the network intersection of methods: $y = 3555.7e^{-0.963x}$, $R^2 = 1.00$, $p = 4.5x10^{-5}$. The average MicroRNA targets predicted across twelve *Drosophila* species generated the following non-linear regressions under power-law trendlines: (A) TargetScan: $R^2 = 0.77$, $p < 10^{-5}$; (B) MiRanda: $R^2 = 0.77$, $p < 10^{-5}$; and (C) for the network intersection of methods: $R^2 = 0.879$, $p < 10^{-5}$. MicroRNA targets predicted for the average across twelve *Drosophila* species generated the following trend lines and non-linear regression: (A) TargetScan: $y = 0.0554\, x^2 - 9.464\, x + 427.7$, $R^2 = 0.90$, $p < 10^{-5}$; (B) MiRanda: $y = 24601e^{-0.7959x}$, $R^2 = 0.97$, $p < 10^{-5}$; and (C) for the network intersection of methods: $y = 9862.2e^{-1.0045x}$, $R^2 = 0.99$, $p < 10^{-5}$. MicroRNA targets networks predicted for *D. melanogaster* alone produced trend lines and non-linear regressions of (D) $y = 0.0554\, x^2 - 9.464\, x + 427.7$, $R^2 = 0.90$, $p < 10^{-5}$, from TargetScan; (E) $y = 7043.2e^{-0.7157x}$, $R^2 = 0.95$, p $< 10^{-5}$, from MiRanda; and (F) $y = 23370.4e^{-0.9363x}$, $R^2 = 0.99$, $p = 1.5\ x10^{-3}$, from the network intersection of methods. Power-law trend non-linear regressions for MicroRNA targets networks predicted for *D. melanogaster* alone were (D) $R^2 = 0.77$, $p < 10^{-5}$, from TargetScan; (E) $R^2 = 0.73$, $p < 10^{-5}$, from MiRanda; and (F) $R^2 = 0.90$, $p < 10^{-5}$, from the network intersection of methods.

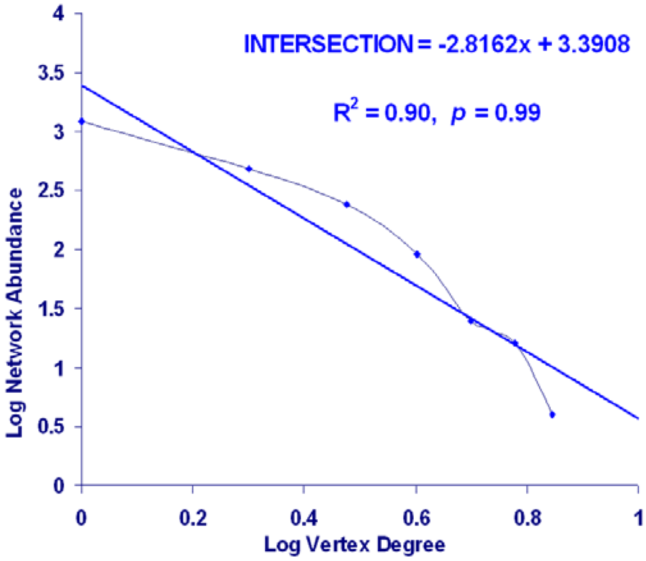


FIGURE 3. Double-logarithmic plot of vertex degree distribution and network abundance of microRNAs per target gene. A linear case was to be expected but predicted for *Drosophila melanogaster* alone produced a power-law trend line and non-linear regression of y = -2.8162 x + 3.3908, R^2 = 0.90 which was not statistically significant.

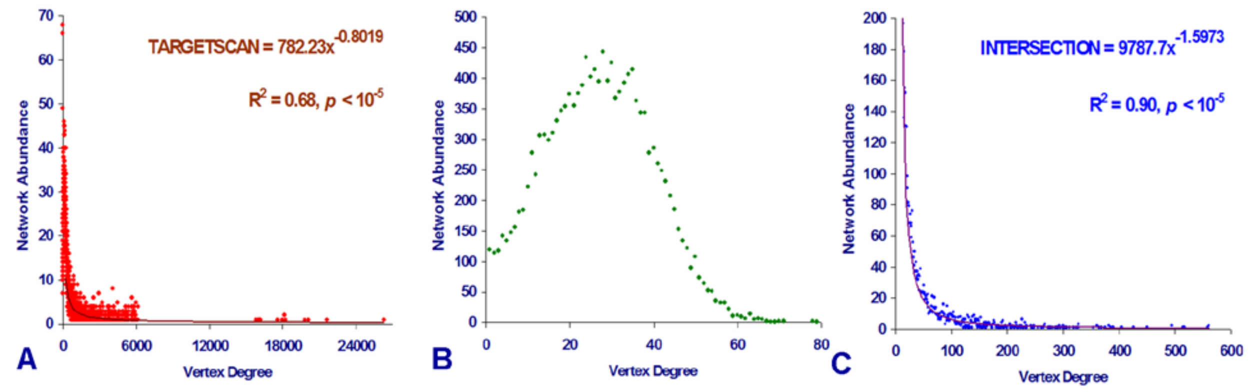


FIGURE 4. Distribution of aptamers per target transcript according to their abundance in microRNA networks. MicroRNA targets are predicted across the union of twelve *Drosophila* species according to (A) TargetScan (B) MiRanda and (C) the network intersection of methods. Accordingly for 112 microRNA families, unique microRNA aptamer sites predicted included: (A) 11,302,034 from TargetScan; (B) 390,560 from MiRanda; and (C) 226,270 aptamers from the network intersection of methods. All data are unbinned. The plotted curves represent power-law trend lines with functions and non-linear regression coefficient of determination for (A) TargetScan: $y = 782.23\,x^{-0.8019}$, $R^2 = 0.68$, $p < 10^{-5}$; and (C) for the network intersection of methods: $y = 9787.7\,x^{-1.5973}$, $R^2 = 0.90$, $p < 10^{-5}$. Pearson correlation coefficients according to method were as follows: (A) TargetScan: -0.42; (B) MiRanda: -0.62; (C) network intersection of methods: -0.33. Note that the vertex degree range for aptamer binding sites appears greater in intersection of methods than in the MiRanda alone due to the overlap of multiple TargetScan aptamers (microRNA seed regions) to aptamer sites recovered from MiRanda.

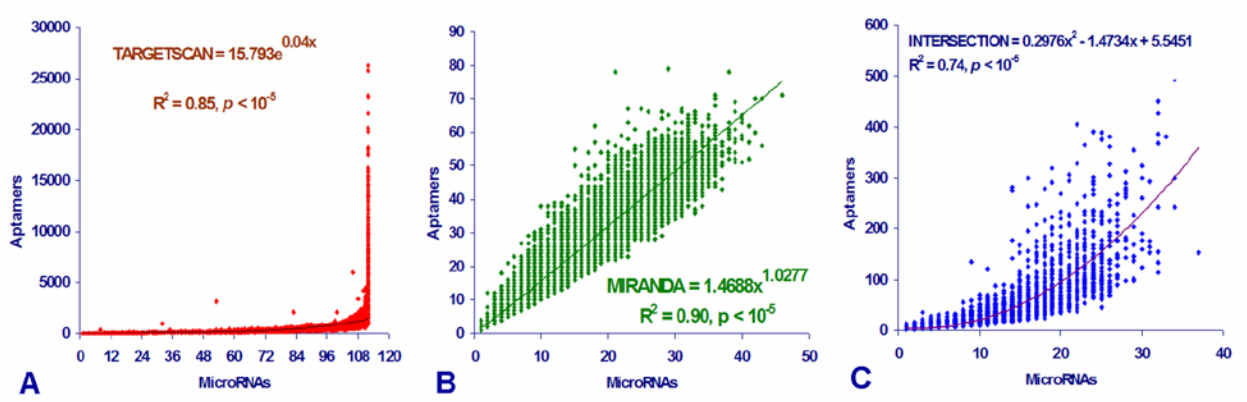


FIGURE 5. Comparison numbers of unique microRNAs to numbers of aptamer sites observed per target transcript. MicroRNA targets are predicted across the union of twelve *Drosophila* species according to (A) TargetScan (B) MiRanda and (C) the network intersection of methods. Accordingly for 112 microRNA families, unique microRNA aptamer sites predicted included: (A) 11,302,034 from TargetScan; (B) 390,560 from MiRanda; and (C) 226,270 aptamers from the network intersection of methods. The plotted curves demonstrate power-law trend lines with functions and non-linear regression coefficient of determination for (A) TargetScan: $y = 0.0428\, x^{2.1138}$, $R^2 = 0.73$; (B) MiRanda: $y = 1.4688\, x^{1.0277}$, $R^2 = 0.90$, $p < 10^{-5}$; and (C) for the network intersection of methods:$y = 1.1291\, x^{1.3003}$, R^2= 0.89. Additionnally TargetScan data A) recovered better fot with an exponential trend line of $y = 15.793\, e^{0.04\,x}$, $R^2 = 0.85$, $p < 10^{-5}$. The network intersection data C) could also be described by a polynomial function of $y = 0.2976\, x^2 - 1.4734x + 5.5451$, $R^2 = 0.74$, $p < 10^{-5}$. Pearson correlation coefficients according to target prediction method were as follows: (A) TargetScan: 0.47; (B) MiRanda: 0.92; (C) network intersection of methods: 0.78. Note that the range of the numbers of aptamer binding sites appears greater in intersection of methods than in the MiRanda alone due to the overlap of multiple TargetScan aptamers (microRNA seed regions) to aptamer sites recovered from MiRanda.

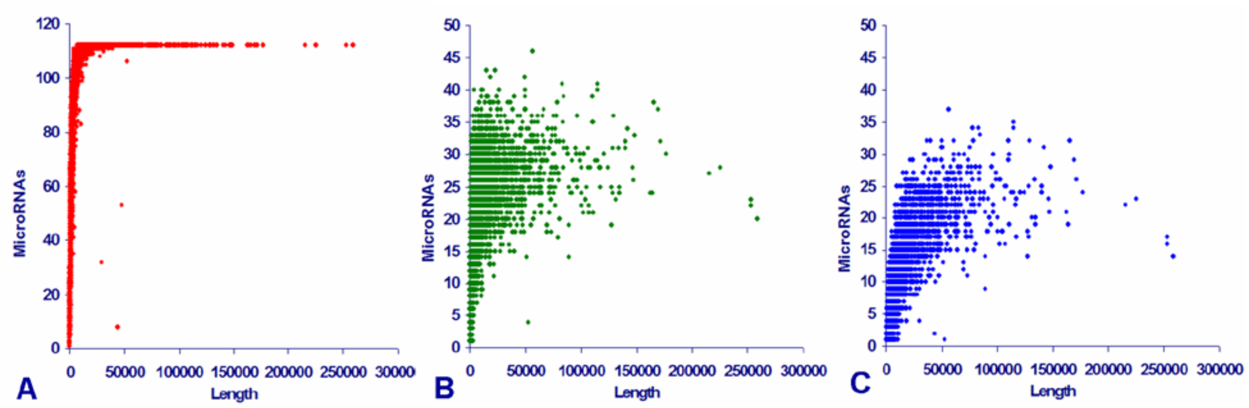


FIGURE 6. Comparison of target transcript nucleotide length to numbers of unique microRNA regulators observed. MicroRNA targets are predicted for 14195 protein coding genes across the union of twelve *Drosophila* species according to (A) TargetScan (B) MiRanda and (C) the network intersection of methods. Accordingly for 112 microRNA families, unique microRNA-protein coding target interactions predicted included: (A) 1,082,224 from TargetScan; (B) 240,394 from MiRanda; and (C) 77,900 from the network intersection of methods. Pearson correlation coefficients retrieved according to method were as follows: (A) TargetScan: 0.45; (B) MiRanda: 0.44; (C) network intersection of methods: 0.72.

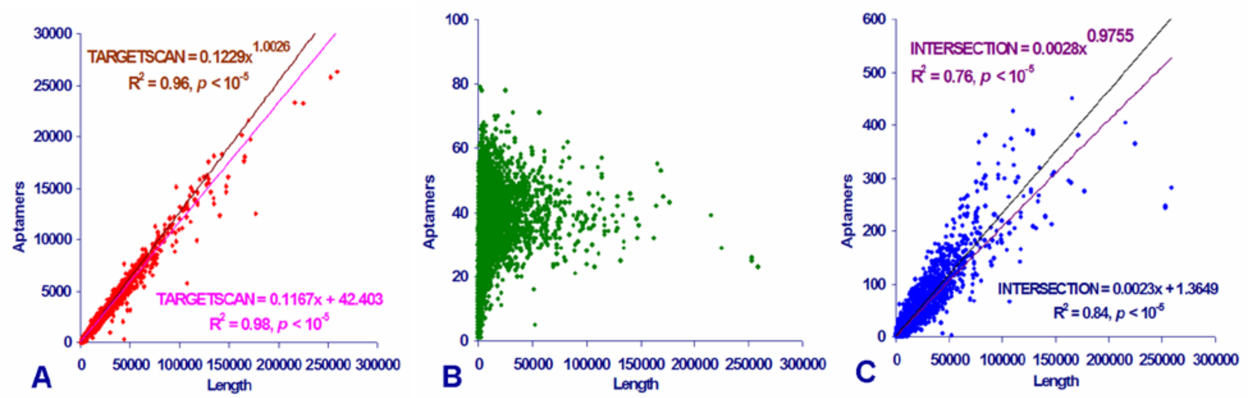


FIGURE 7. Comparison of target transcript nucleotide length to numbers of aptamer sites observed. MicroRNA targets are predicted for 14195 protein-coding genes across the union of twelve *Drosophila* species according to (A) TargetScan (B) MiRanda and (C) the network intersection of methods. Accordingly, for 112 microRNA families, aptamer sites predicted included: (A) 11,246,285 from TargetScan; (B) 386,954 from MiRanda; and (C) 225,152 aptamers from the network intersection of methods. (A) TargetScan data recovered a power-law trend lines with functions and non-linear regression coefficient of $y = 0.1229\ x^{1.0026}$, $R^2 = 0.96$, $p < 10^{-5}$. Additionally TargetScan data could also be well described by a linear trend line of $y = 0.1167\ x + 42.403$; $R^2 = 0.98$, $p < 10^{-5}$. Likewise the network intersection data could also be described by a linear trend line of $y = 0.0023\ x + 1.3649$; $R^2 = 0.84$. (C) The network intersection of methods was fitted to a power-law trend lines of $y = 0.0028\ x^{0.9755}$, $R^2 = 0.76$, $p < 10^{-5}$. Respective Pearson correlation coefficients recovered according to method were as follows: (A) TargetScan: 0.99; (B) MiRanda: 0.33; (C) network intersection of methods: 0.92.

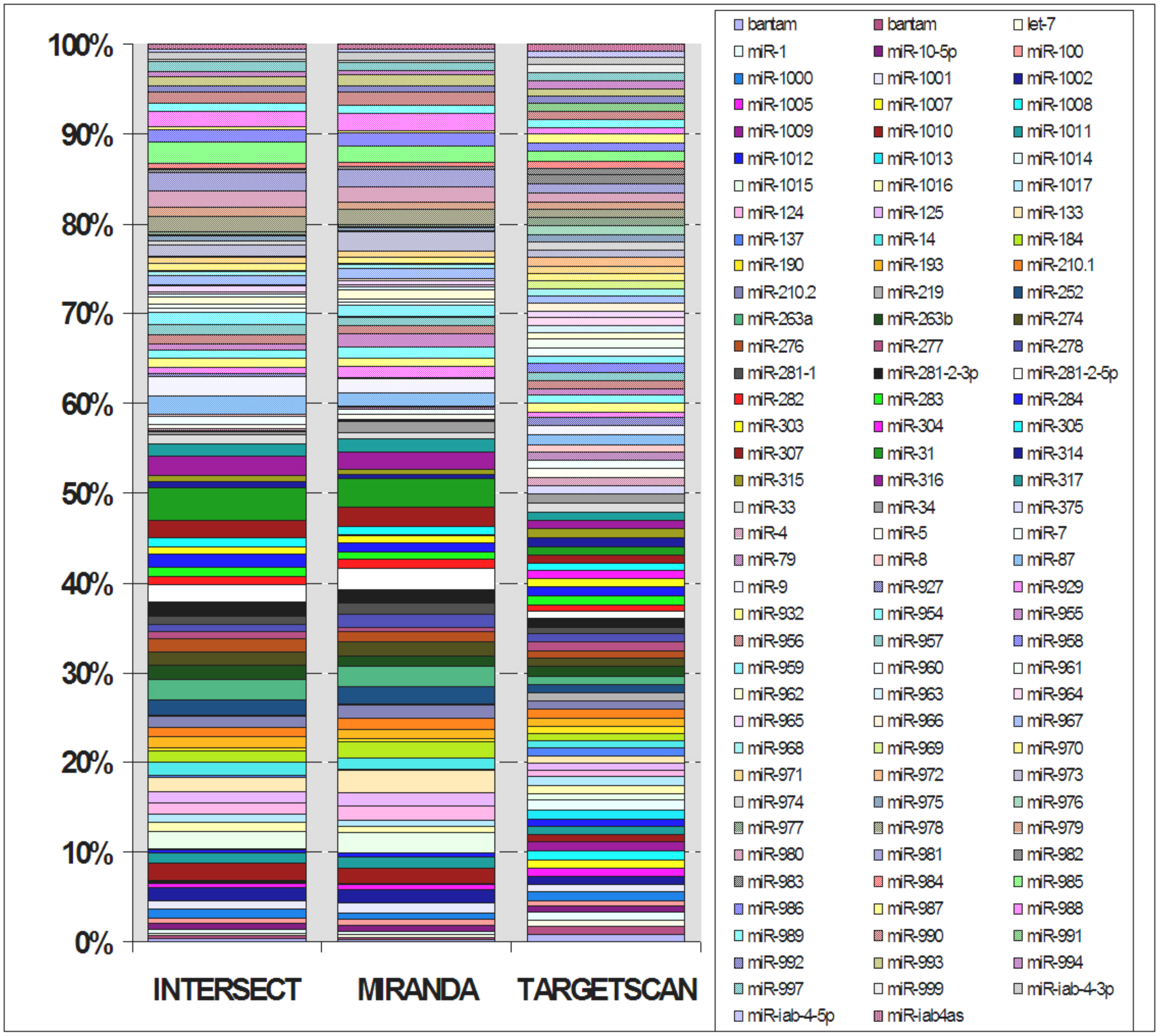


FIGURE 8. Percent target distribution profile of microRNAs by prediction method. Results are color-coded by microRNA. Each category represents the target dataset union across twelve *Drosophila* species. The Intersect category represents the network intersection of MiRanda and TargetScan microRNA target prediction data. Accordingly, targets predicted included: (A) 14,760 from TargetScan; (B) 14,462 from MiRanda; and (C) 12,498 targets from the network intersection of methods. Unique microRNA-target interactions predicted Included: (A) 1,090,221 from TargetScan; (B) 241,861 from MiRanda; and (C) 78, 280 from the network intersection of methods.

DISCUSSION

MicroRNA Regulatory Network Properties. Complex natural patterns may display underlying simplicity through scale-invariance (Albert, 2005; Barabasi & Albert, 1999; Gavin *et al.*, 2006; Hastings *et al.*, 1993). Scale-invariance indicates that a pattern remains unchanged regardless of magnification or contraction and in turn, scaling rules follow forms characterized under power law functions. Given a function, $f(x) = A * x^{b}$, where "*A*" and "*b*" are constants: any scaling of the variable "*x*" by a constant "*A*" incurrs proportionate scaling of the function itself. Thus, all power laws of set scaling exponent "*b*" are equivalent up to constant factors; each is essentially a scaled version of the others. Many natural networks are scale free having constant properties irregardless of network size and demonstrate power-law behavior for their vertex degree frequency distribution (Barabasi & Albert, 1999). In such cases node network abundance may decline expontially with increasing vertex degree. Typically, the constant "*b*" within biological networks is within a range of -2 to -3 (Barabasi & Albert, 1999). Moreover, where such power law behavior is present then a double-logarithmic plot of *log f(x)* against *log x* is linear. Nevertheless other functions asides power-law may accurately expression the non-linear scale-free nature of the network (Dorogovtsev & Mendes, 2003).

The nature of network power-law behavior varied according to microRNA prediction method. Nevertheless, node network abundance displayed a general exponential decline as the number of targets or aptamers per transcript increases (FIGUREs 2 & 4). Power-law trend lines values for "*b*" from TargetScan aptamer-degree-frequency and target-degree-frequency distributions fall below normal biological range (*b* = -0.74 to -0.83; FIGUREs 4A, 2A, & 2D). MiRanda aptamer and target-degree-frequency distributions recovered power-law trend lines values for "*b*" that partially cross the biological range of -2 to -3 (*b* = -0.23 to -4.06; FIGUREs 4B, 2B & 2E). Likewise target-degree-frequency distribution for the intersection of microRNA target prediction methods produce exponential values falling within biological range (*b* = -2.82 to -4.22; FIGUREs 2C & 2F). Aptamer-degree-frequency distribution for the intersection of target prediction methods exponential values falling below biological range (*b* = -1.5973; FIGUREs 4C). Coefficients of determination for fit to a power-law function displayed an increase for the intersection (R^2 = 0.88 to 0.90) over both parent methods (MiRanda, R^2 = 0.71-0.81; TargetScan, R^2 = 0.73-0.77) across minima, maxima and averages of twelve *Drosophila* species (FIGURE 2A, FIGUREs 2B & 2C). Similarly for *Drosophila melanogaster* data alone, power-law trend line coefficients of determination increase for the intersection (R^2 = 0.90) over parent methods (MiRanda, R^2 = 0.73; TargetScan, R^2 = 0.77). Likewise coefficients of determination for power-law trend line fit using apatamer-degree-freqeuncy distribution increased for the intersection (R^2 = 0.90) over parent methods (MiRanda, R^2 = 0.23; TargetScan, R^2 = 0.68) across the union of twelve *Drosophila* species (FIGURE 4). It should be recalled that all these microRNA target and aptamer data are unbinned and further binning the data into frequency groups of 3 or 5 could bring the exponential interval into natural range. Nevertheless, a clear linear relationship between the two axes indicative of power-law behavior was expected and visible for the intersection of MiRanda and TargetScan networks in *Drosophila melanogaster* alone (R^2 = 0.90; FIGURE 3) but is not statsically signifcant.

From the proceeding it can be reasoned that, compared to respective parent networks, the intersection of microRNA target prediction methods produces target networks of increased potential biological relevance. The intersection network outperformed parent methods exhibiting a 1.44 to 1.78 fold increase in target sensitivity. Therefore the network intersection of methods appears to have increased discriminatory power over its parent methods. Moreover the network intersection encoded 9.42×10^8 bits of information. This was nearly double in innate complexity of the parent methods. Thus the network intersection is a smaller, but substantially richer set of data compared to its parent methods.

Prospectives for MicroRNA-Target Verification. Genome-wide microRNA target predictions in animals have estimated that between 20 to 60% of all genes are likely to come under regulation (Friedman *et al.*, 2009; Lewis *et al.*, 2005; Stark *et al.*, 2005). This study recovered target prediction data though MiRanda and/or TargetScan for 14,925 genes (>90% *Drosophila* genome); and of these, 95.11% correspond to protein coding transcripts. The sizable target datasets produced in this study are applicable for continuing research in *Drosophila* molecular biology. Accordingly, it is suggested that the vast numbers of potential interactions proposed for all methods by these bioinformatic studies could be biochemically verified using whole genome microarray analyses and miRNP immunopurification. The later method effectively identifies microRNA targets based on their *in vivo* physical association with mature microRNAs bound to Argonaute protein-containing effector complexes (Easow, Teleman, & Cohen, 2007).

MicroRNA Regulatory Network in the Context of Natural Selection. The animal microRNA repertoire has operated upon the sequence evolution of all messenger RNAs, modulating the use of a substantial fraction of the transcriptome (Bartel & Chen, 2004). Notably, quantitative relationships between microRNA regulators and aptamer binding sites per target transcripts differed between microRNA target prediction methods. A positive correlation between numbers of unique microRNAs and numbers of aptamer sites was weak for TargetScan (Pearson correlation coefficient = 0.47), moderate for network intersection of methods (0.78) and strong for MiRanda data (0.92; FIGURE 5). These values likely reflect differential selective forces operative upon the seed- and canonical-type microRNA interactions predicted by TargetScan and MiRanda, respectively (Lewis *et al.*, 2005). Moreover, differences could be expected where the respective microRNA aptamer types participate in different molecular responses; potentially where canonical interactions induce RNA transcript cleavage while seed-type interactions stall translation of protein by physically blocking the ribosome assembly (Lu *et al.*, 2008; Stark *et al.*, 2007b). Increasing diversity of microRNA regulators per transcript requires increasing diversity of unique canonical-type aptamer binding sites (FIGURE 5B). Conversely, even with increased diversity of microRNA regulators per transcript, there is substantial redundancy of seed-type aptamer binding sites (FIGURE 5A). Stringent regulation by a single microRNA is rare: single sites contained in most targets do not appear to be sufficient to confer strong repression (Stark *et al.*, 2005). Moreover this makes a switch-like relationship unlikely and microRNAs might not primarily be involved in developmental decision-making (Stark *et al.*, 2005). Accordingly, the findings of this study are consistent with the principle of canalization: *Drosophila* genes selected for greatest microRNA coordinate control with seed-type interactions also have greater aptamer redundancy wired into the regulation (Hornstein, & Shomron, 2006).

Messenger RNA Transcript Length. The length of an entire target transcript appears to represent another factor exerting selective influence upon the microRNA interactome. Increase in target transcript length and increased number of interacting microRNAs targeting the transcript were only weakly correlated for all methods (Pearson correlation coefficients range 0.44 to 0.72; FIGURE 6). In contrast, a strong Pearson correlation coefficients was observed between increasing target transcript length and number of aptamer binding sites per transcript was observed for TargetScan data (0.99) and network intersection of methods (0.92; FIGURES 7A & 7B). Thus it stands to reason that as transcript length increases the likelihood of acquisition of a seed-type aptamer binding site also increases. These findings for the entire messenger RNA transcript agree with conclusions of other detailed analysis which have considered only the 3'UTR of *Drosophila* messenger RNAs (Stark *et al.*, 2005). In the later case, genes with more microRNA sites have both average longer 3'UTRs and significantly more binding sites per kilobase of 3'UTR sequence. Reciprocally, a large set of genes involved in basic cellular processes avoid microRNA regulation due to short 3'UTRs that are specifically depleted of microRNA binding sites (Stark *et al.*, 2005). Moreover, the lengthening of 3'UTRs by alternative polyadenylation during mouse embryonic development is predicted to significantly augment microRNA-mediated posttranscriptional control (Ji *et al.*, 2009). Rapidly proliferating murid T-lymphoctye cells express messenger RNAs with shortened 3'UTRS and fewer available microRNA aptamers (Sanberg *et al.*, 2008). Such trends collectively indicate that transcripts operate under length selection to acquire or eliminate microRNA aptamer sites (Stark *et al.*, 2005). Over the course of natural history functional aptamers will appear randomly and genes presented with microRNA regulation come under selection to specifically avoid or utilize regulation. Regulatory avoidance would be expected for genes normally expressed at high levels in which microRNA-mediated repression would be detrimental (Bartel & Chen, 2004; Stark *et al.*, 2005).

A Model of MicroRNA Aptamer Sequence Evolution. Representation of individual microRNAs in the regulatory interactome differed according to method. Cursory visual scan of bars in FIGURE 8 reveals a nearly uniform distribution of microRNAs across the TargetScan dataset while the MiRanda dataset displays differential target enrichment by microRNA. These findings for TargetScan data contradict a hypothesis that the number of seed sites might grow over natural history such that ancient microRNAs would tend to have more targets than those more recently acquired (Brennecke *et al*., 2005). The percent target distribution profile for the intersection of prediction methods reflects similar differential microRNA enrichment pattern to that presented by the parent MiRanda interactome. As with quantitative relationships between microRNA regulators and aptamer binding sites (FIGURE 5), differential microRNA enrichment patterns by method may reflect separate selective forces operating upon the seed- and canonical-type microRNA interactions (Lewis *et al.*, 2005).

A widespread natural occurrence of heptamers matching microRNA seed regions in the *Drosophila* genome may account for the even microRNA-target distribution profile of TargetScan data. Indeed other genomic analyses of *Drosophila* have indicated that a majority of microRNA target sites lack substantial pairing in the 3' end in nearby sequences (Brennecke *et al.*, 2005). Given this, then it seems reasonable to accept a model of microRNA target acquisition that initiates with the chance acquisition of a functional seed region with only seven or eight bases complementary to a microRNA (Brennecke *et al.*, 2005). Over the course of sequence evolution, each aptamer site adapts to increase or decrease its pairing to the microRNA regulator, and in this way, fine-tunes the degree of repression in cells that express the corresponding microRNA species (Bartel & Chen, 2004). Subsequently, if the initial seed-type microRNA regulatory interaction is positively reinforced by natural selection, then later mutations may expand the binding site region along the 3’ end of the microRNA in order to confer stronger repression and optimize the regulation (Brennecke *et al.*, 2005). Thus a 5’-dominant canonical interaction may be produced. Further mutation along the aptamer may alter the original seed region and a 3’ compensatory microRNA target would remain (Sethupathy, *et al.*, 2006). Canonical sites can thus be seen as an extension of the seed type with enhanced 3' pairing in addition to a sufficient 5' seed or as an extension of the 3' compensatory type with improved 5' seed quality in addition to sufficient 3' pairing (Brennecke *et al*., 2005).

According to the previous model, 5' dominant canonical and seed sites should be responsive to all members of a given microRNA family, whereas 3' compensatory sites should differ in their sensitivity to different microRNA family members depending on the degree of 3' complementarity (Brennecke *et al*., 2005). Thus the 3' compensatory class of target sites may be utilized *in vivo* to discriminate among individual microRNA family members (Brennecke *et al*., 2005). Experimental results in *Drosophila* suggest that a functional seed requires a continuous sequence of at least 4 or 5 nucleotides and that there is some position dependence to the pairing, since sites that produce comparable pairing energies differ in their ability to function (Brennecke *et al*., 2005). Conversely for 3’ compensatory sites, extensive 3' pairing of up to 17 nucleotides in the absence of the minimal 5' element is not sufficient to confer regulation (Brennecke *et al*., 2005). This suggests that the sequences that could pair to the 3' end of the microRNAs are not important for regulation as they do not appear to be under selective pressure (Brennecke *et al*., 2005). Nevertheless, deep conservation of 3' complementarity suggests that these aptamer sites are likely to be a functionally important in some sense (Brennecke *et al*., 2005). It is here prosposed that many of the 3'compensatory sites are likely to be biologically functional and that many predicted 3'compensatory targets of MiRanda data could potentially recover a 5' seed (and thus would need to be reclassified as 5' dominant cannonical type aptamers) if one were also to consider microRNAs that have undergone adenosine-to-inosine (A-to-I) editing (Habig, Dale, & Bass, 2007).

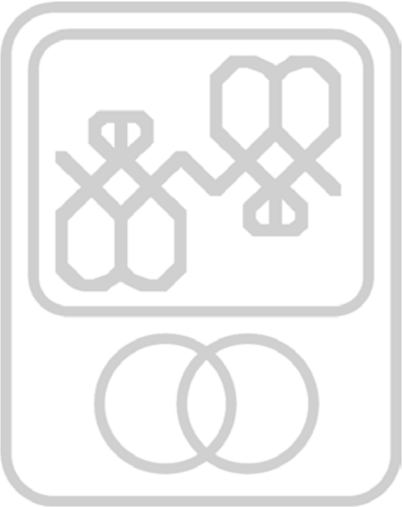
Future work from these data should empirically examine aptamer site sequence evolution in greater detail. To these ends, sequence regions from the *Drosophila* multiple sequence alignment exhibiting all three classes of microRNA-target sites will be isolated. Thereafter, molecular phylogenetic analysis will be conducted to map nucleotide character states changes within sequence regions against the established *Drosophila* phylogeny. In this way molecular phylogenetics may verify the historical transition from 5'-dominant seed to 5'-dominant canonical to 3'-compensatory for microRNA aptamers.

Continuing Research. The preceeding analyses provide abundant source material for continuing research. The methodology outlined in FIGURE 1 can be readily reproduced for other organisms. To date 14.0 GB of microRNA target data have already been output for MiRanda and TargetScan using 525, 731 KB of sequence data for 28,123 genes and 67 microRNAs for six species of nematode, namely: *Caenorhabditis brenneri*, *C. briggsae*, *C. elegans*, *C. japonica*, *C. remanei*, and *Pristionchus pacificus*. Future research will examine the influence of incremental increase of threshold values for hybridization energy upon the node-degree-frequency distribution plot of MiRanda microRNA target networks: the average Gibbs Free energy value for the MiRanda of -26.4 kcal/mol indicates that the majority of MicroRNA targets recovered approach the threshold ΔG value of -25. It has been previously hypothesized that ranking microRNA target sites according to overall complementarity or free energy of duplex formation might not reflect their biological activity (Brennecke *et al*., 2005). Moreover the network behavior of discrete classes of microRNA aptamers should be considered further. MicroRNA networks for MiRanda and TargetScan targets not conserved in the other method should be generated. These later networks together with the network intersection of methods would represent exlcusive networks of the 5’-seed, 3’-compensatory, and 5’-dominant canonical type aptamers. Comparison of these three networks may provide further insights into natural selective forces operative upon microRNA interactions. Network properties in this study may also have been influenced by the presence of 47 microRNAs exhibiting lineage specific expansion for in *Drosophila* (Berezikov *et al.*, 2010). Restriction of the networks under microRNAs strictly conserved within the genus *Drosophila* also increased calculated sensitivity rates for MiRanda and the intersection of both prediction methods to 65.29 and 89.94%, respectively.

Further study for deuterostome specific microRNAs with no homologs in *Drosophila* is also warranted. Natural selective principles governing network expansion with *de novo* microRNA acquisition might be discerned though a contrast of descriptive and topological network properties for microRNAs endogenous and alien to *Drosophila* (APPENDIX III, TABLE 8). Given that a transcriptome cannot have target selection operative for genes not part of their microRNA repertoire, it is reasonable to assume that each alien would effectively behave in the *Drosophila* regulatory network much as a newly acquired microRNA. The relative simplicity of microRNA genes, together with their gene family sequence diversity and their independent acquisition in plant and animal lineages, indicates that acquisition of de novo microRNA genes (with their consequent impact on transcript regulation) might occur with relative frequency over course of natural history (Bartel & Chen, 2004).

CHAPTER II.

Drosophilid Patterns of MicroRNA Network Conservation across Interactions, Targets, and Species



Keywords & concepts: Dobzhansky-Muller hypothesis; *Drosophila*; Interology; Messenger RNA Transcript Length; MicroRNA Aptamer Sequence Evolution; MicroRNA-Target Interaction Conservation; Molecular Clocks; Regulatory Element Conservation; Regulatory Network Conservation.

ABSTRACT

Strict patterns of conservation of interaction between microRNAs and targets remain unclear. This study considers the interrelated topics of 1) aptamer sequence evolution, 2) microRNA-target interaction conservation (interology), 3) the conservation of microRNA target regulation across species and 4) time since species divergence. Molecular phylogenetic reconstructions using standard parsimony (MP) were conducted through PAUP* under a constrained tree topology matching the expected phylogeny of the genus *Drosophila* for a total of 14,929 putative microRNA target genes. A molecular clock rate of 1,579,192 unweighted maximum parsimony steps per million years was recovered. This represents the first comprehensive study to directly relate molecular sequence evolution and phylogeny with microRNA regulatory network interology in *Drosophila*. MicroRNA regulatory network structure was found to change over time and across species. The decrease in conserved microRNA-target interactions with increasing phylogenetic distance exhibited a cure typical of a saturation phenomena. It seems that only a modest number of microRNA–mRNA interactions exhibit conservation over *Drosophila* cladogenesis. The minimal numbers of conserved microRNA-target interactions retained throughout all taxa were 1,839 from MiRanda, 13,357 from TargetScan, and 135 for the intersection of both methods. These later values likely represent the presence of a functionally-constrained core of microRNA-target interactions essential to *Drosophila.* These findings represent the first comprehensive study to directly relate molecular sequence evolution and phylogeny with microRNA regulatory network interology in *Drosophila*.

INTRODUCTION

MicroRNA regulatory networks are dense, with most target genes targeted by multiple microRNAs, and they exhibit precise combinatorial control of targets giving increased regulatory versatility (Enright *et al.*, 2003; Grün *et al.*, 2005; Lewis *et al.*, 2005; Sempere *et al.*, 2007; Stark *et al.*, 2005; Stark *et al.*, 2007a). The ease by which novel microRNAs target sites can be altered or lost, coupled to the profound consequences in developmental processes, provides powerful source variation upon which selection can operate (Stark *et al.*, 2005). Strict patterns of conservation of interaction between microRNAs and targets remain unclear. MicroRNA seed sequences remain largely invariant over large phylogenetic distance and likewise, about 50% of octanucleotide blocks in vertebrate 3'UTRs are conserved and complementary to known microRNAs (Lu *et al.*, 2008; Stark *et al.*, 2005; Xie *et al.*, 2005). Bayesian phylogenetic modeling for microRNA target sites in mammals suggests that target repertoires of some microRNAs have been largely conserved since mammalian origin, while other target repertoires of microRNAs have accumulated significant changes (Gaidatzis *et al.*, 2007). Lineage specific microRNAs exhibit far fewer conserved targets and lower expression than do the more broadly conserved microRNAs; even when considering only more recently emerged targets (Friedman *et al.*, 2009; Grün *et al.*, 2005; Stark *et al.*, 2007a). Thus it seems that although both microRNA and target genes may be conserved, the interactions between them may not (Lee *et al.*, 2007, Matthews *et al.*, 2001). Further refinement in the study of dual microRNAs and target conservation is necessary.

Selective pressures for conserving functional target sites between related species may differ significantly: functional target sites for one microRNA may be preferentially conserved in one species, while functional sites for another microRNA may be preferentially conserved in another species (Gaidatzis *et al.*, 2007). For instance, the polymorphism pattern of *miR-310s* in *Drosophila* indicates lineage specific differentiation under positive selection for *D. melanogaster* in comparison to other species (Lu *et al.*, 2008). The degree of divergence of microRNA regulatory networks could likely dictate clade-specific reproductive isolation, and principally, the conservation between microRNA-target interactions may be viewed to indicate what kinds of genetic programs have been conserved between species (Lee *et al.*, 2007; Prochnik *et al.*, 2007). Moreover, microRNAs are also likely involved in adaptive regulatory circuit extension where organisms expand the functional portion of their genome as they also incorporate survival information about their niche (Lee *et al.*, 2007). Under this model, it is possible that recently acquired species-specific microRNAs would be most involved in fine-tuning gene expression to adapt organisms to different environments, rather than supporting more ancient developmental programs (Stark *et al.*, 2005). Among the 78 microRNAs reported in *Drosophila* before 2007, only 5 are confirmed to be newly emerged in the genus, but these have likely accumulated many adaptive changes during a surprisingly long period (roughly 55 million years) of natural history (Lu *et al.*, 2008; Stark *et al.*, 2007a). Effectively the divergence spanned by the genus *Drosophila* exceeds that of the entire mammalian radiation when generation time is considered (*Drosophila* 12 Genomes Consortium). This study considers the interrelated topics of 1) aptamer sequence evolution, 2) microRNA-target interaction conservation (interology), 3) the conservation of microRNA target regulation across species and 4) time since species divergence. Notably, the findings presented here represent the first comprehensive study to directly relate molecular sequence evolution and phylogeny with microRNA regulatory network interology in *Drosophila*. Likewise, this study addresses the relationships between regulation conservation and nucleotide length for the entire messenger RNA transcripts where similar analyses prior have only briefly considered regulatory conservation relative the 3'UTR of *Drosophila* (Stark et al., 2005).

METHODS

Multiple sequence alignments from MULTIZ available from UCSC for a total of 14929 putative microRNA target genes were formatted into a Nexus file with command line for phylogenetic analyses (16.2 GB data produced) and subjected in batch to molecular phylogenetic reconstructions through PAUP* (CHAPTER I, FIGURE 1; Blanchette *et al.*, 2004; Swofford, 2002). Molecular phylogenetic reconstructions using standard parsimony (MP) were conducted under a constrained tree topology matching the expected phylogeny of the genus *Drosophila* (CHAPTER III, FIGURE 22) with accelerated transformation of characters optimized on the tree(s) in memory (ACCTRAN) with gaps coded as a 5th character state and rooted assigning *D. grimshawi* as an outgroup. Parametric scores for Consistency Index (CI), Retention Index (RI), Rescaled Consistency Index (RC), Homoplasy Index (HI), and Goloboff-fits (G-fit) and tree length (L) were extracted from 1.05 GB of PAUP* output files using Perl scripts. Further molecular phylogenies constrained to the reference tree were conducted under maximum likelihood (ML) criteria under a general time reversible model with gamma distributed rates and invariant sites (GTR + I + G Model; Lanave *et al.*, 1984; Swofford, 2002). Likelihood scores and base frequencies were extracted from 4.81 GB of PAUP* output files using Perl scripts. There were a total of 30.3 KB of parametric score data from MP and ML molecular phylogenetic analyses extracted. These were input incorporated into the *musca* MySQL database and integrated to microRNA target data from TargetScan, MiRanda and the network intersection of target predictions method through joint table queries (CHAPTER I, FIGURE 1). Query of the *musca* database recovered percent conservation of microRNA-target interactions, microRNA-aptamer site interactions, species, and transcript length. Moreover, interspecific comparisons of microRNA network interactions and target data were conducted directly through the *musca* database using the network intersection for all unique (66) cross-species combinations of *Drosophila* sampled (164 network Files recovered in 1.04 KB). Additionally, molecular clock estimates of *Drosophila* divergence times were extrapolated using cross-species comparisons from phylogenetic reconstruction under maximum parsimony across 14926 genes, and a calibrated divergence time of 12.8 MYA for *D. melanogaster* from *D. yakuba* based on inference from African biogeography (Lachaise *et al.*, 1988).

RESULTS

Molecular Phylogenetics. The entire dataset of 14,925 genes represented 91,915,264 total characters; of which 64,748,176 (70.44 %) were potentially parsimony informative, 14,690,262 (15.98 %) characters were constant, and 12,217,058 (13.29 %) characters were variable but uninformative. The parsimony based reconstructions for the total dataset recovered average parametric scores per target gene as follows: Consistency Index (CI) = 0.721, Retention Index (RI) = 0.50, Rescaled Consistency Index (RC) = 0.277, and Homoplasy Index (HI) = 0.277 (see APPENDIX III, TABLE 12). Reconstruction using maximum likelihood recovered an average percent GC content per gene of 50.1%, an average percent invariable rate of 20.8%, and an average gamma rate of 2.578. The observed total GC content was on par with previously established estimates for the genus *Drosophila* (Rodrigiuez-Trelles, Tarroi, & Ayala, 2000). The average *ln* Likelihood score was -38815.36. There was an average of 0.017 bits of information per base position per gene.

The total target data recovered total parametric scores for parsimony based reconstructions as follows: CI = 0.724, RI = 0.462, RC = 0.276, HI = 0.276 (see APPENDIX III, TABLE 13). The total tree length of the entire dataset was 203,694,546 steps. Reconstruction using maximum likelihood recovered a total GC content was 0.50%, total invariable rate pf 20.4%, and a total gamma of 2.126. The average –*ln* Likelihood score was 137,036.648. The total overall bits per base position 0.016. A total of 245 genes were recovered with a consistency index greater than or equal to 90% to the reference phylogeny (CHAPTER III, FIGURE 22). Within this later set there were 52 protein coding genes, 4 processed psuedogenes, 146 t-RNAs, and 11 small nuclear RNAs, and 27 microRNAs. In finer scale, the potential phylogenetic utility for whole gene regions under the control tree topology could be addressed through comparison of likelihood and parametric scores like consistency index and homoplasy index, the number of potentially parsimony informative sites out of the total gene region (see CHAPTER IV).

Comparisons of mutational steps under maximum parsimony to numbers of microRNA target and network edges recovered are presented in FIGURE 20. Maximum parsimony steps between species represent interspecific difference between the sum of the fit of 91,915,264 characters across 14,925 genes to the reference phylogeny (of combined length 203,694,546; see topology CHAPTER III, FIGURE 22) as inferred through a branch-bound search using accelerated character state transformation with Gaps coded as a 5th character state. The range of maximum parsimony steps ranged from 9,731,085 (*Drosophila simulans* vs. *D. sechellia*) to 98,103,533 (*Drosophila simulans vs. D. willistoni*) with an average of 67,253,867. Comparisons of calculated divergence time to numbers of microRNA target and network edges recovered are diagramed in FIGURE 21. The recovered molecular clock rate was 1,579,192 unweighted maximum parsimony steps per million years

Interspecific Regulatory Network Comparison under TargetScan data. The range of TargetScan targets across species was from 14,001 in *Drosophila grimshawi* to 14,543 in *D. melanogaster* with an average of 14,325 targets per individual species regulatory network. A comparison of microRNA targets to regulatory network edges for TargetScan across 66 interspecific network comparisons in *Drosophila* is displayed in FIGURE 9A. The percent conservation of microRNA targets to percent conservation of regulatory network edges across all *Drosophila* interspecific network comparisons is displayed in FIGURE 11. Relationships conservation for both microRNA targets and regulatory network edges were well described using a subtle curvilinear plot for both TargetScan and MiRanda (R^2 = 0.99; FIGURE 11). TargetScan exhibited a much higher conservation profile than MiRanda for both microRNA targets and regulatory network edges (FIGURE 11).

The minimal number of cross-species conserved TargetScan targets retrieved through direct cross-species comparison through network intersection was 12,383 for *Drosophila mojavensis vs. D. persimilis.* This later species pair had a 77.68% conservation of targets. The maximum number of shared targets observed for TargetScan was 14,437 for the intersection of *Drosophila melanogaster* vs. *D. sechellia*. This later species combination had a 98.62% target conservation. Direct cross-species comparison through network intersection across all TargetScan networks returned an average of 2,066 targets with a 83.62% conservation between species.

Drosophila species conservation of microRNA regulation per target gene by TargetScan is represented in the orange bars of FIGURE 12. These data present the regulation of a given target gene by some microRNA; regardless of the identity of the microRNA. Across all *Drosophila* species, there were only 89 (0.6%) targets apomorphic to one species. Conversely, there were 13357 (90.76%) targets fully in regulation by some microRNA for all twelve *Drosophila* species. A conservation comparison for *Drosophila* species to microRNA regulators for TargetScan is presented in FIGURE 13A. Likewise a conservation comparison for *Drosophila* species to aptamer sites predicted through TargetScan is contained in FIGURE 14A.

The TargetScan conservation for gene regulatory interaction per individual microRNA was 27.25 to 50.51%. This later conservation for gene regulatory interaction was had a 39.44% average per individual microRNA. A comparasion of unique microRNAs observed through TargetScan to percent conservation of microRNA interactions is presented in FIGURE 15A. Similarly, comparison of aptamer sites observed to maximum percent conservation of aptamer sites is presented in FIGURE 17A. Additionally, a comparison of aptamer sites observed through TargetScan to percent conservation of microRNA interactions is drawn in FIGURE 16A.

TargetScan regulatory network microRNA-target interactions ranged from 373,916 in *Drosophila mojavensis* to 494,131 interactions in *D. melanogaster*. There was an average of 434,423 interactions across twelve *Drosophila* species. Direct cross-species comparision through network intersection revealed that the lowest numbers of conserved microRNA-target regulatory interactions were 204,984 in *Drosophila mojavensis* vs. *D. persimilis* with a 35.56% conservation between species. Conversely, the highest number of conserved microRNA-target regulatory interactions for TargetScan were 428,713 in *Drosophila simulans vs. D. sechellia* (compare tree topology in CHAPTER III, FIGURE 22). There was an 81.52% conservation of microRNA-target interaction between these later to sister species. TargetScan regulatory networks recovered an average of 255,602 microRNA-Target interactions with a 42.07% conservation rate across twelve *Drosophila* species.

Interspecific Regulatory Network Comparison using MiRanda data. MiRanda targets across species ranged from 4,509 in *D. melanogaster* to 12,632 in *D. simulans* with an average of 11,460 targets per individual species regulatory network. A comparison of microRNA targets to regulatory network edges for MiRanda across 66 interspecific network comparisons in *Drosophila* is displayed in FIGURE 9B. The percent conservation of microRNA targets to percent conservation of regulatory network edges across all *Drosophila* interspecific network comparisons is displayed in FIGURE 11.

The minimal number of cross-species conserved MiRanda targets recovered by direct cross-species comparision through network intersection was 416 for *Drosophila melanogaster vs. D. willistoni*. This later species pair had a 2.75% conservation of targets. The maximum number of shared targets observed for MiRanda was 10,561 for the intersection of *Drosophila pseudoobscura vs. D. persimilis*. This later pairing of sister species from the *D. obscura* group had a 75.47% target conservation. Direct cross-species comparision through network intersection across all MiRanda networks returned an average of 2,066 targets with a 10.72% conservation between species.

Drosophila species conservation of microRNA regulation per target gene by MiRanda is represented in the green bars of FIGURE 12. Across all *Drosophila* species, there were only 144 (1.0%) targets apomorphic to a single species. Conversely, there were 1,839 (12.71%) targets fully conserved in regulation by some microRNA. A conservation comparison for *Drosophila* species to microRNA regulators for MiRanda is presented in FIGURE 13B. Likewise a conservation comparison for *Drosophila* species to aptamer sites predicted through MiRanda is contained in FIGURE 14B.

The MiRanda conservation for individual target interaction per microRNA was much lower than TargetScan at 2.40 to 10.99% with 8.21% as an average. A comparasion of unique microRNAs observed through MiRanda to percent conservation of microRNA interactions is held in FIGURE 15B. Similarly, comparison of aptamer sites observed to maximum percent conservation of aptamer sites is presented in FIGURE 17B. Additionally, a comparison of aptamer sites observed through MiRanda to percent conservation of microRNA interactions is drawn in FIGURE 16B.

MiRanda regulatory network microRNA-target interactions ranged from 13,044 in *Drosophila mojavensis* to 36,844 interactions in *D. melanogaster*. There was an average 32,550 interactions across all twelve *Drosophila* species. Direct cross-species comparision through network intersection revealed that the lowest numbers of conserved microRNA-target regulatory interactions were 462 in *Drosophila melanogaster vs. D. willistoni*; with a 1.13% conservation between species. Conversely, the highest number of conserved microRNA-target regulatory interactions for MiRanda was 23,505 in *Drosophila pseudoobscura vs. D. persimilis.* There was an 47.74% conservation of microRNA-target interaction between these later to sister species. MiRanda regulatory networks recovered an average of 2789 microRNA-Target interactions with a 4.74% conservation rate across twelve *Drosophila* species. It is of potential biological significance that this average conservation rate for MiRanda is almost an order of magnitude lower than the rate recovered from TargetScan data (FIGURE 11).

Additional Interspecific Network Comparisons. Interspecific network comparisons byTargetScan and MiRanda according to microRNA target and network edges recovered are presented in FIGURE 10. A species–specific trend is observed where R^2 values decline with decreasing relatedness to *Drosophila melanogaster* in FIGURE 10 (see CHAPTER III, FIGURE 22). It is also notable that among four sets of closest sister species of taxa sampled, (namely *Drosophila erecta - D. yakuba, D. mojavensis – D. virilis, D. persimilis – D. pseudoobscura,* and *D. sechelia – D. similans*; compare CHAPTER III, FIGURE 22). The interspecific conservation of microRNA interactions ranged from 4.5 to 40.1% for MiRanda and 31.3 to 81.5% for

TargetScan. Thus the recovered ranges of shared interactions were 9,648 to 20,813 for MiRanda and 220,436 to 428,713 interactions for TargetScan. Conversely, non-conserved microRNA interactions between closest sampled *Drosophila* sister species ranged 445 to 5,558 interactions for MiRanda and from 97,170 to 356,004 interactions for TargetScan.

Drosophila species conservation of microRNA regulation per target gene by network intersection of prediction methods is represented in the blue bars of FIGURE 12. Across all *Drosophila* species, there were 1,676 (13.4%) targets apomorphic to a single species. There were only 135 (1.08%) targets fully conserved in regulation by some microRNA across all species. A conservation comparison for *Drosophila* species to microRNA regulators for network intersection of prediction methods is presented in FIGURE 13C. Likewise a conservation comparison for *Drosophila* species to aptamer sites predicted is contained in FIGURE 14C. A comparasion of unique microRNAs observed through the network intersection of prediction methods to percent conservation of microRNA interactions is held in FIGURE 15C. Similarly, comparison of aptamer sites observed to maximum percent conservation of aptamer sites is presented in FIGURE 17C. Additionally, a comparison of aptamer sites observed through network intersection of prediction methods to percent conservation of microRNA interactions is drawn in FIGURE 16C.

A comparison of target transcript nucleotide length to maximum percent conservation of predicted aptamer sites is illustrated in FIGURE 18. Likewise target transcript nucleotide length is compared to percent conservation of microRNAs in FIGURE 19. Increasing length and conservation were positively correlated for TargetScan but weakly negatively correlated for data from MiRanda and network intersection of prediction methods (FIGURE 19). Likewise, aptamer site conservation was weakly negatively correlated to transcript length (FIGURE 18)

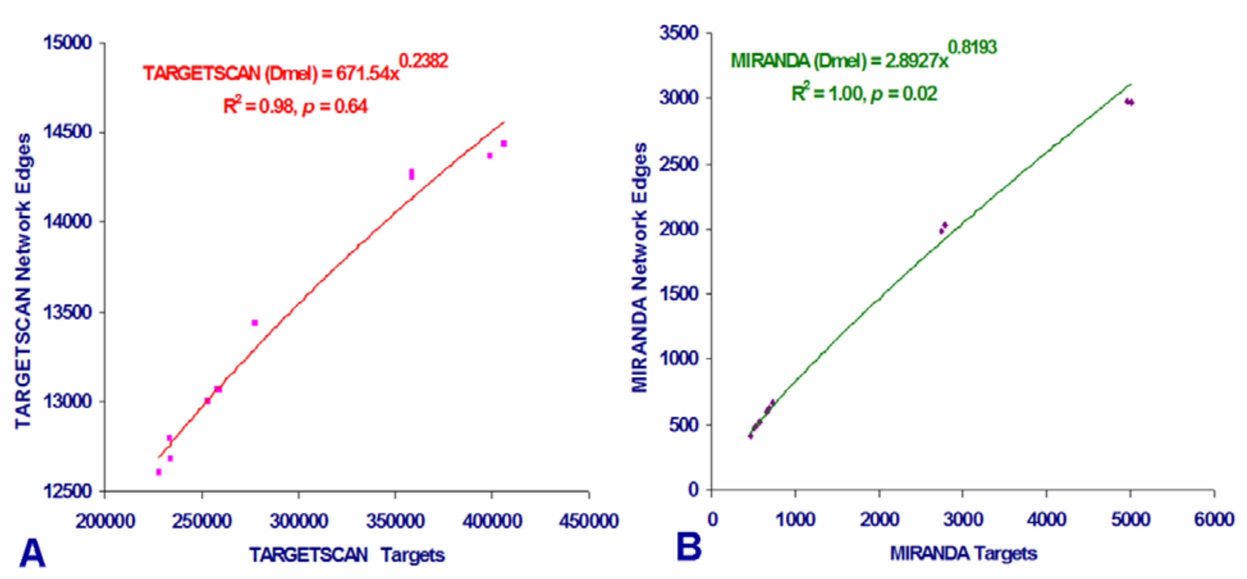


FIGURE 9. Comparison of microRNA targets to regulatory network edges by prediction method across twelve interspecific network comparisons to *Drosophila melanogaster*. Trend line plots functions and coefficients of determination recovered are presented for *D. melanogaster* alone (Dmel). (A) TargetScan data recovered a power-law trend line with functions and non-linear regression coefficient of (Dmel) $y = 671.54\, x^{0.2382}$, $R^2 = 0.98$, $p = 0.64$. (B) MiRanda data recovered a power-law trend line with functions and non-linear regression coefficient of (Dmel) $y = 2.8927\, x^{0.8193}$, $R^2 = 0.997$, $p = 0.02$. Pearson correlation coefficients recovered according to method were as follows: (A) TargetScan: 0.98; (B) MiRanda: 0.98.

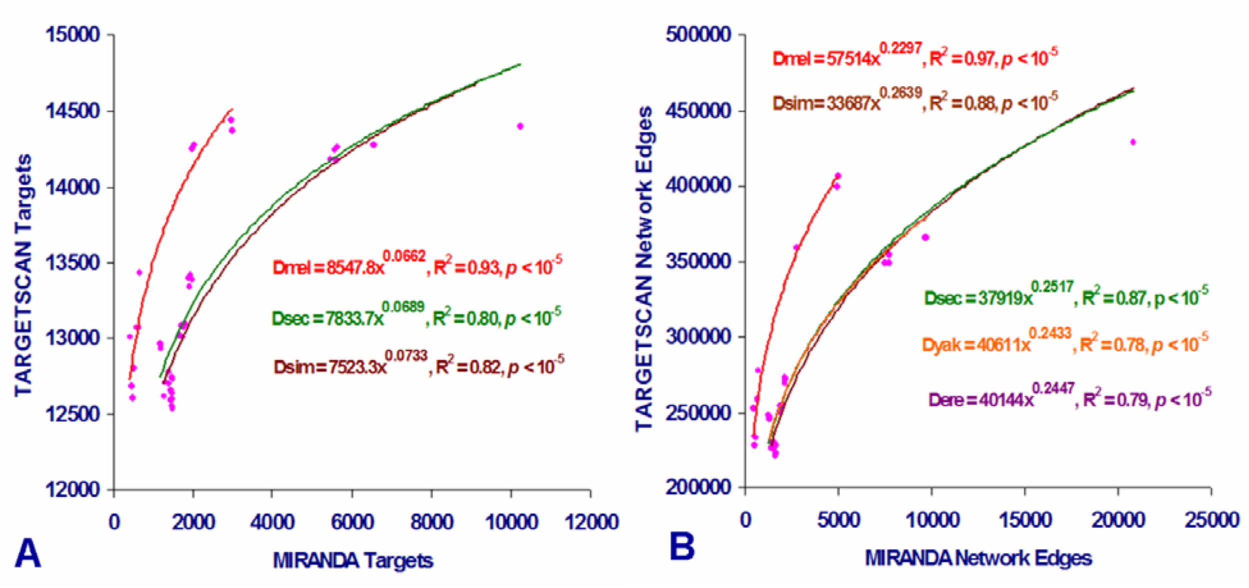


FIGURE 10. Interspecific network comparisons by TargetScan and MiRanda according to numbers of microRNA target and network edges recovered. MicroRNA target prediction method results are directly compared across 66 interspecific comparisons of twelve *Drosophila* species according to (A) the numbers of targets and (B) the numbers of network edges output. Trend line plots functions and coefficients of determination recovered are color coded according to select species. *Drosophila* species are abbreviated respectively: Dere) *erecta*; Dmel) *melanogaster*; Dsec) *sechellia*; Dsim) *simulans*; Dyak) *yakuba*. (A) Data for targets by species recovered power-law trend lines with functions and non-linear regression as follows:

melanogaster) $y = 8547.8\,x^{0.0662}$, $R^2 = 0.93$, $p < 10^{-5}$;

sechellia) $y = 7833.7\,x^{0.0689}$, $R^2 = 0.80$, $p < 10^{-5}$;

simulans) $y = 7523.3\,x^{0.0733}$, $R^2 = 0.81$, $p < 10^{-5}$;

(B) Data for microRNA-target interactions by select species recovered power-law trend lines with functions and non-linear regression as follows:

erecta) $y = 40144\,x^{0.2447}$, $R^2 = 0.79$, $p < 10^{-5}$;

melanogaster) $y = 57514\,x^{0.2297}$, $R^2 = 0.97$, $p < 10^{-5}$;

sechellia) $y = 37919\,x^{0.2517}$, $R^2 = 0.87$, $p < 10^{-5}$;

simulans) $y = 33687\,x^{0.2639}$, $R^2 = 0.88$, $p < 10^{-5}$;

yakuba) $y = 40611\,x^{0.2433}$, $R^2 = 0.78$, $p < 10^{-5}$;

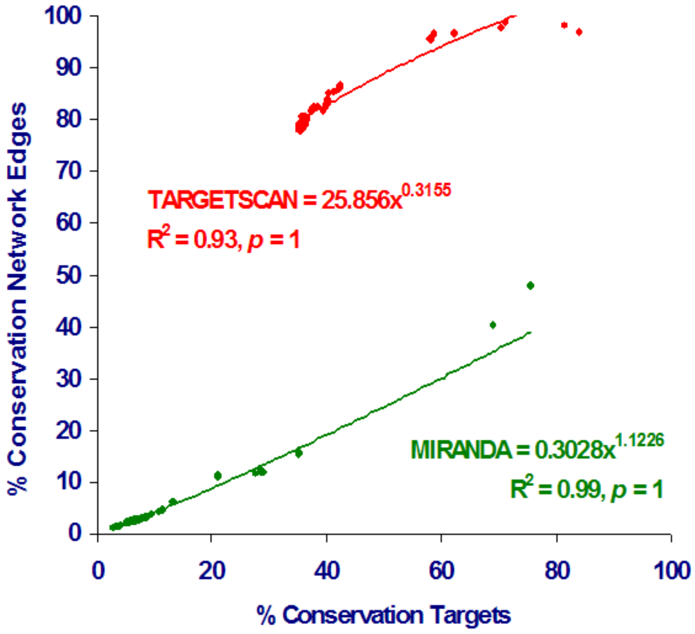


FIGURE 11. Percent conservation of microRNA targets to percent conservation of regulatory network edges across 66 interspecific network comparisons in *Drosophila*. Accordingly, microRNA targets predicted across the union of twelve *Drosophila* species included: 14,760 from TargetScan and 14,462 from MiRanda. Accordingly for 112 microRNA families, unique microRNA-target interactions species predicted included 1,090,221 from TargetScan and 241,861 from MiRanda. TargetScan network data demonstrated a power-law trend lines with functions and non-linear regression coefficient of determination of $y = 25.856\, x^{0.3155}$, $R^2 = 0.93$, $p = 1$. MiRanda network data recovered a power-law trend line function and non-linear regression coefficient of determination of $y = 0.3028\, x^{1.1226}$, $R^2 = 0.99$, $p = 1$. Pearson correlation coefficients recovered according to method were as follows: TargetScan) 0.96 and MiRanda) 0.26.

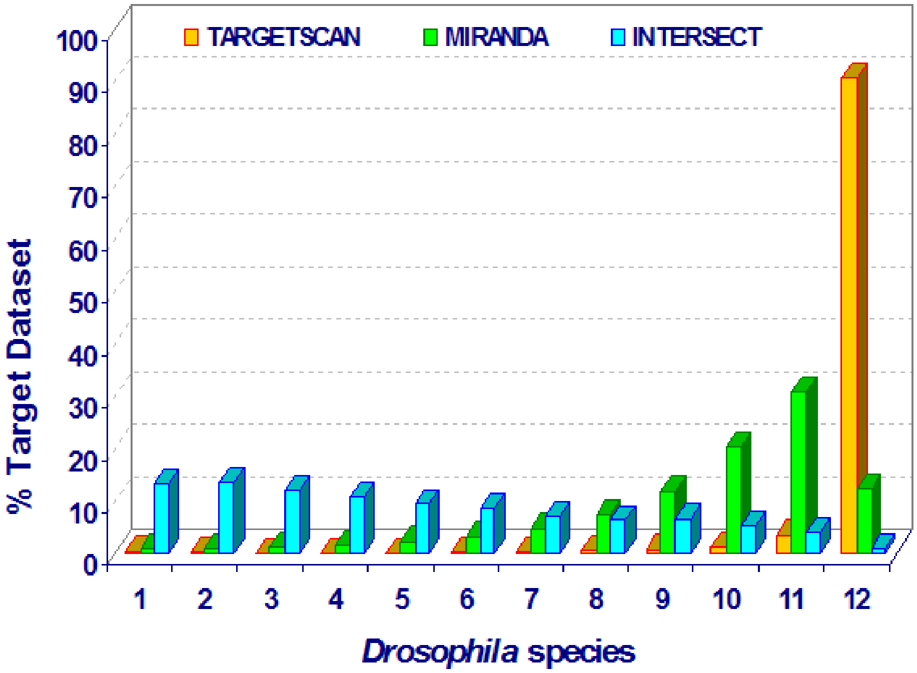


FIGURE 12. *Drosophila* species conservation of microRNA regulation per target gene by prediction method. Counts of *Drosophila* species exhibiting any microRNA regulation for individual target genes appear on the x-axis and are color-coded by microRNA-target prediction method. Representative percentages of each species category per microRNA-target prediction method each appear on the y-axis. Accordingly, total microRNA targets predicted included: (A) 14,760 from TargetScan; (B) 14,462 from MiRanda; and (C) 12,498 targets from the network intersection of methods.

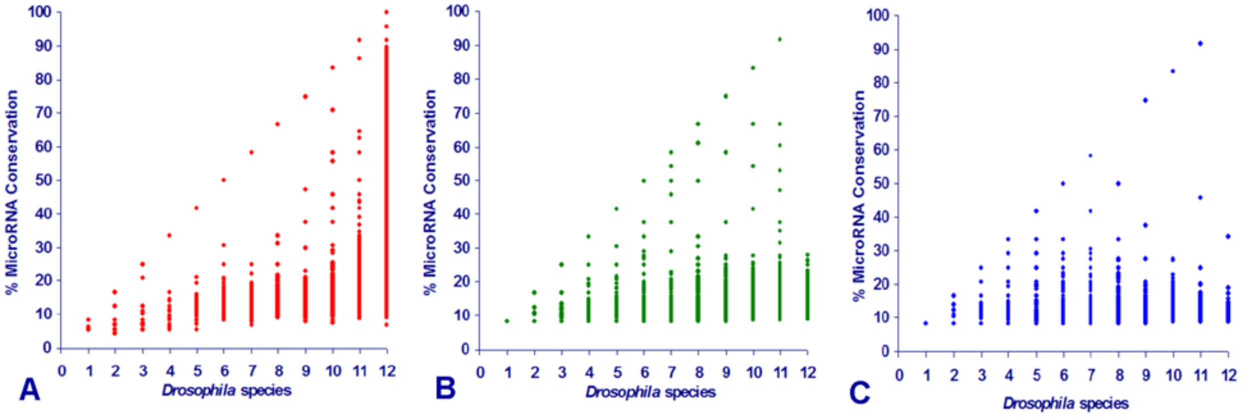


FIGURE 13. Conservation comparison for *Drosophila* species to microRNA regulators. The count of *Drosophila* species exhibiting any microRNA regulation for individual target genes appear on the x-axes, while the percent conservation of unique microRNA-target interaction combinations appear on the y-axes. Accordingly for 112 microRNA families, unique microRNA-target interactions predicted included: (A) 1,090,221 from TargetScan; (B) 241,861 from MiRanda; and (C) 78, 280 from the network intersection of methods. Respective Pearson correlation coefficients recovered according to method were as follows: (A) TargetScan: 0.21; (B) MiRanda: 0.15; (C) network intersection of methods: 0.21.

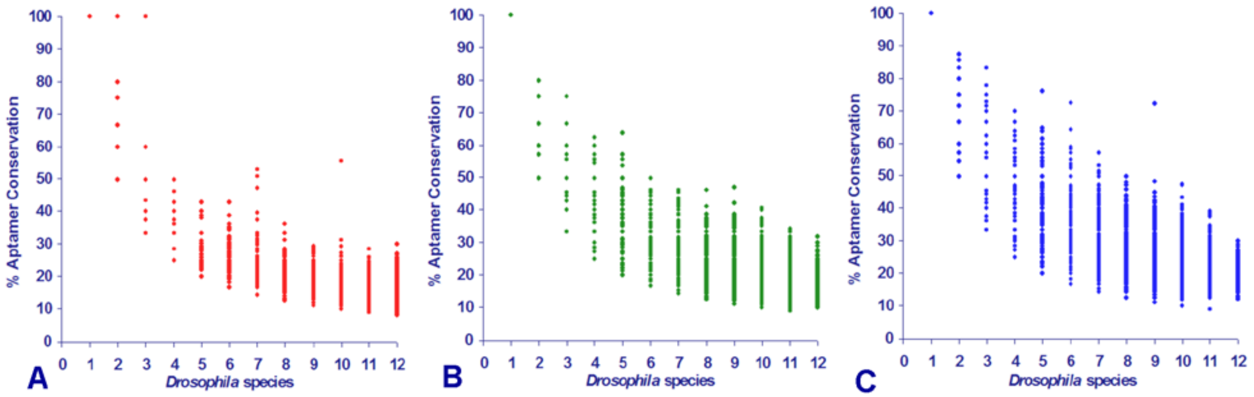


FIGURE 14. Conservation comparison for *Drosophila* species to aptamer sites. The count of *Drosophila* species exhibiting any microRNA regulation for individual target genes appear on the x-axes, while the maximum percent conservation of unique microRNA-aptamer interaction combinations appear on the y-axes. There may be many unique aptamer sites interactions per individual microRNA-target combination. Accordingly for 112 microRNA families, unique microRNA aptamer sites predicted included: (A) 11,302,034 from TargetScan; (B) 390,560 from MiRanda; and (C) 226,270 aptamers from the network intersection of methods. Pearson correlation coefficients retrieved according to method were as follows: (A) TargetScan: -0.85; (B) MiRanda: -0.76; (C) network intersection of methods: -0.76.

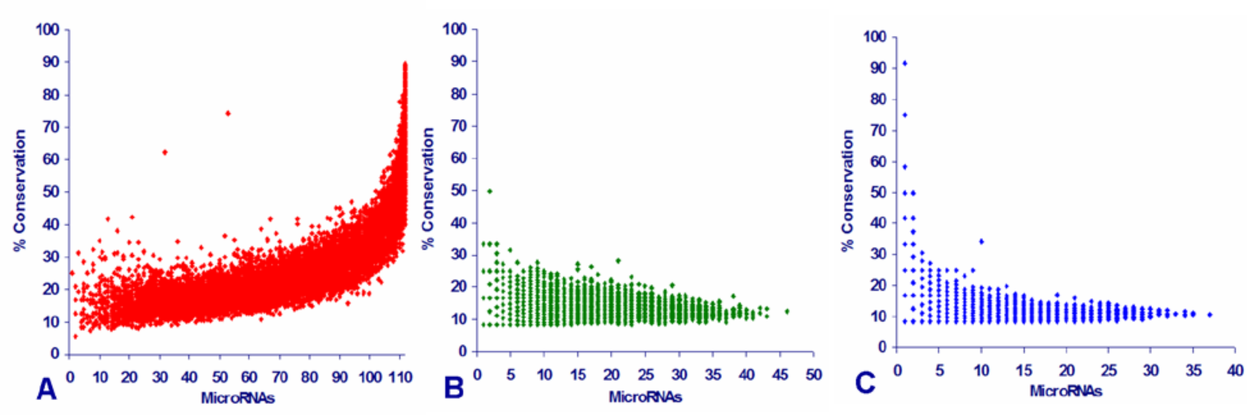


FIGURE 15. Comparasion of unique microRNAs observed to percent conservation of microRNA interactions. MicroRNA targets are predicted for 14195 protein coding genes across the union of twelve *Drosophila* species according to (A) TargetScan (B) MiRanda and (C) the network intersection of methods. Accordingly for a total 112 microRNA families, unique microRNA-protein coding target interactions predicted included: (A) 1,082,224 from TargetScan; (B) 240,394 from MiRanda; and (C) 77,900 from the network intersection of methods. Pearson correlation coefficients retrieved according to method were as follows: (A) TargetScan: 0.753752; (B) MiRanda: -0.021; (C) network intersection of methods: -0.009.

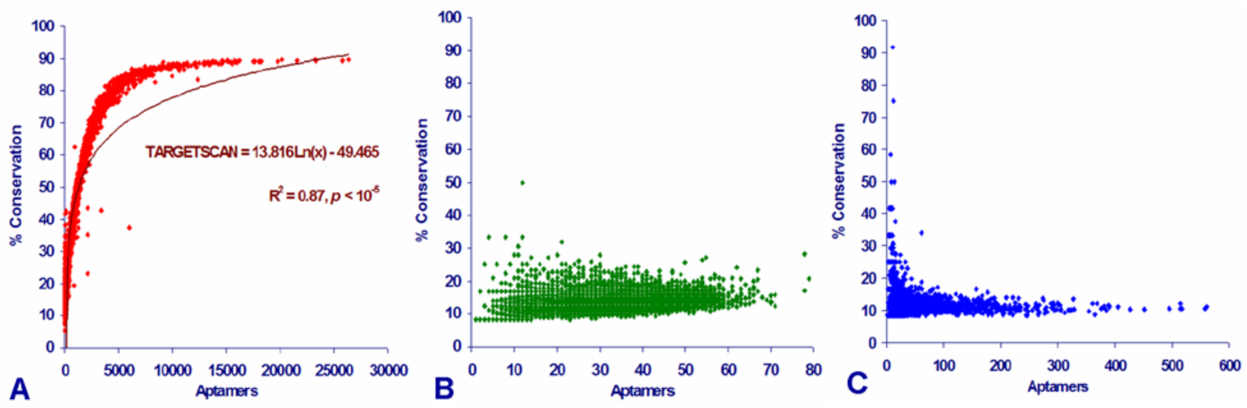


FIGURE 16. Comparison of aptamer sites observed to percent conservation of microRNA interactions. MicroRNA target sites are predicted for 14195 protein coding genes across the union of twelve *Drosophila* species according to (A) TargetScan (B) MiRanda and (C) the network intersection of methods. Accordingly for 112 microRNA families, unique microRNA-protein coding target interactions predicted included: (A) 1,082,224 from TargetScan; (B) 240,394 from MiRanda; and (C) 77,900 from the network intersection of methods. Likewise, aptamer sites on protein-coding transcripts predicted included: (A) 11,246,285 from TargetScan; (B) 386,954 from MiRanda; and (C) 225,152 aptamers from the network intersection of methods. There may be many unique aptamer sites interactions per individual microRNA-target combination. A) TargetScan data could also be described with a logarithmic trend line of y = 13.816 $Ln(x)$ - 49.465; $R^2 = 0.87$, $p < 10^{-5}$. Pearson correlation coefficients retrieved according to method were as follows: (A) TargetScan: 0.79811; (B) MiRanda: 0.31; (C) network intersection of methods: 0.038.

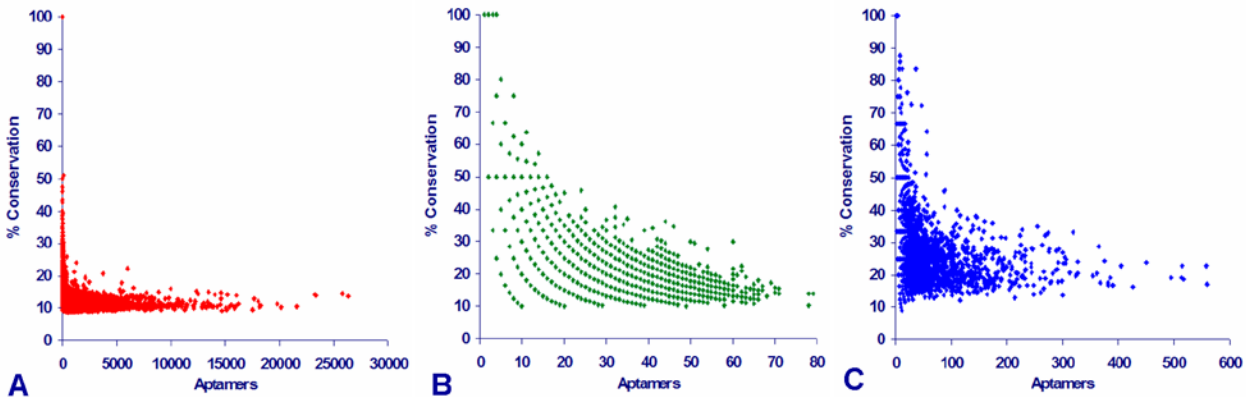


FIGURE 17. Comparison of aptamer sites observed to maximum percent conservation of aptamer sites. MicroRNA target sites are predicted for 14195 protein coding genes across the union of twelve *Drosophila* species according to (A) TargetScan (B) MiRanda and (C) the network intersection of methods. There may be many unique aptamer sites interactions per individual microRNA-target combination. Accordingly, aptamer sites on protein-coding transcripts predicted included: (A) 11,246,285 from TargetScan; (B) 386,954 from MiRanda; and (C) 225,152 aptamers from the network intersection of methods. Pearson correlation coefficients retrieved according to method were as follows: (A) TargetScan: -0.16; (B) MiRanda: -0.55; (C) network intersection of methods: -0.32.

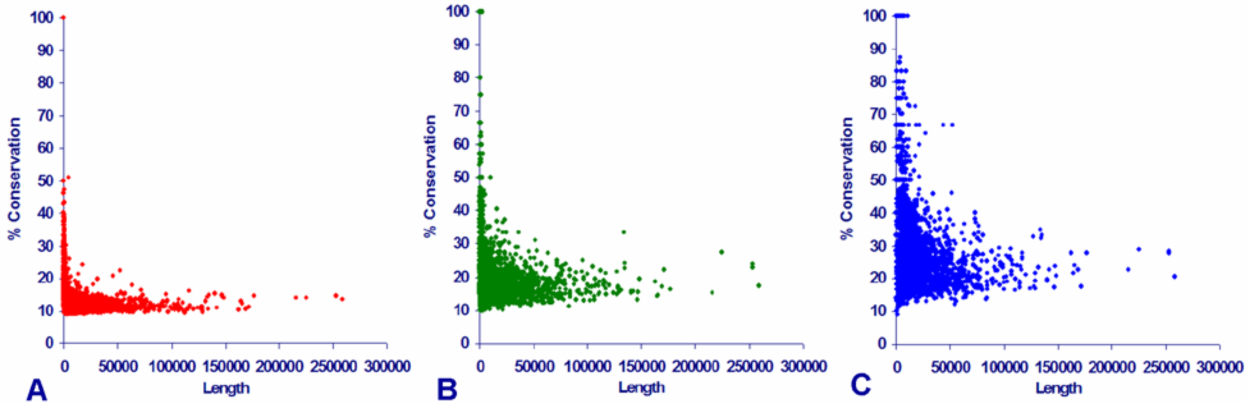


FIGURE 18. Comparison of target transcript nucleotide length to maximum percent conservation of aptamer sites. MicroRNA target sites are predicted for 14195 protein coding genes across the union of twelve *Drosophila* species according to (A) TargetScan (B) MiRanda and (C) the network intersection of methods. Accordingly, for 112 microRNA families, aptamer sites on protein-coding transcripts predicted included: (A) 11,246,285 from TargetScan; (B) 386,954 from MiRanda; and (C) 225,152 aptamers from the network intersection of methods. There may be many unique aptamer sites interactions per individual microRNA-target combination. Pearson correlation coefficients retrieved according to method were as follows: (A) TargetScan: -0.13; (B) MiRanda: -0.14; (C) network intersection of methods: -0.28.

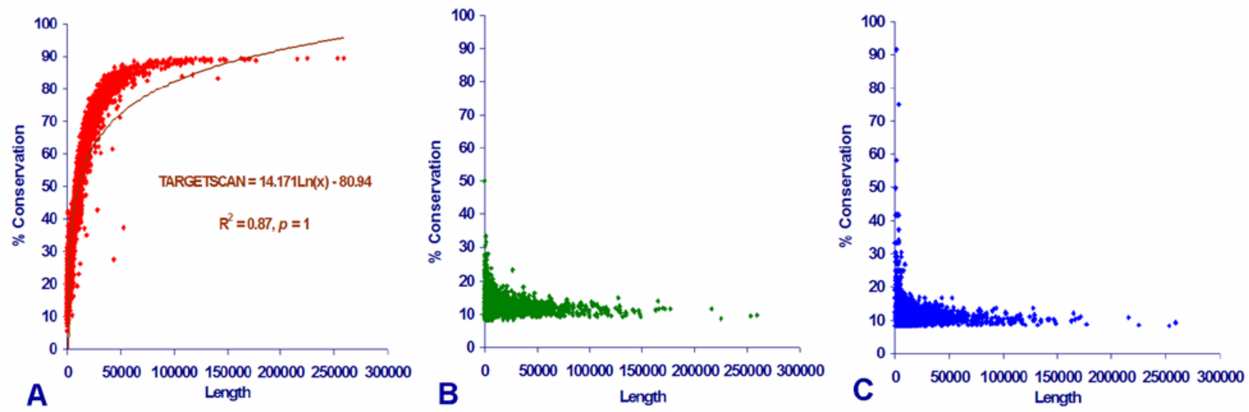


FIGURE 19. Comparison of target transcript nucleotide length to percent conservation of microRNA interactions. MicroRNA target sites are predicted for 14195 protein coding genes across the union of twelve *Drosophila* species according to (A) TargetScan (B) MiRanda and (C) the network intersection of methods. Accordingly for 112 microRNA families, unique microRNA-protein coding target interactions predicted included: (A) 1,082,224 from TargetScan; (B) 240,394 from MiRanda; and (C) 77,900 from the network intersection of methods. A) TargetScan data could be described with a logarithmic trend line of y = 14.171 $Ln(x)$ - 80.94; R^2 = 0.87, p = 1. Pearson correlation coefficients retrieved according to method were as follows: (A) TargetScan: 0.78; (B) MiRanda: -0.12; (C) network intersection of methods: -0.02.

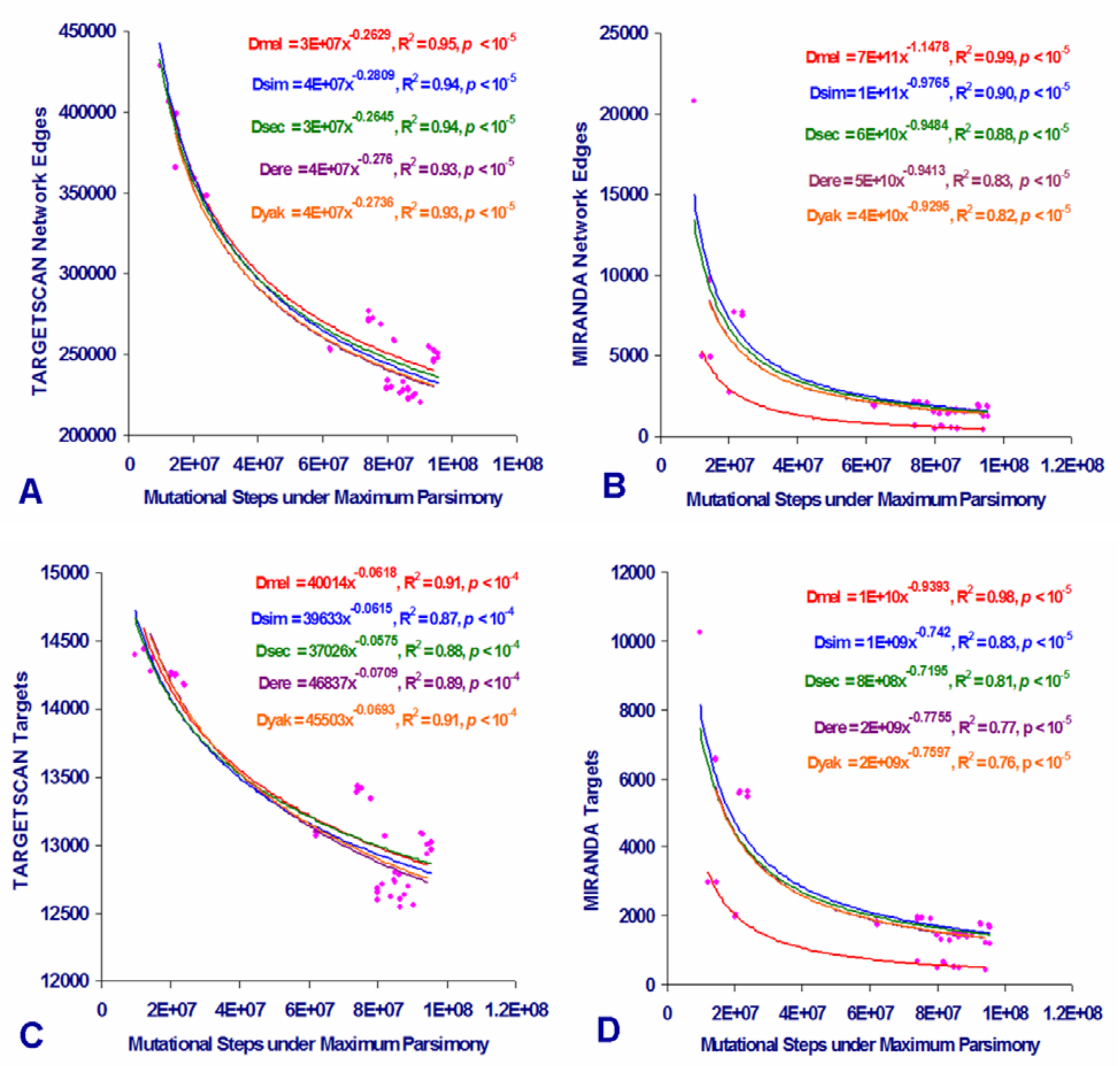


FIGURE 20. Comparisons of mutational steps under maximum parsimony to numbers of microRNA target and network edges recovered. Conserved microRNA regulatory network edges (A & B) edges and (C & D) nodes are presented across 66 interspecific comparisons of twelve *Drosophila* species. Trend line plots functions and coefficients of determination recovered are color coded according to select species. *Drosophila* species are abbreviated respectively: Dere) *erecta*; Dmel) *melanogaster*; Dsec) *sechellia*; Dsim) *simulans*; Dspp) the union of twelve *Drosophila* species; Dyak) *yakuba.*

(A) TargetScan target data for species recovered power-law trend lines with functions and non-linear regression as follows:

erecta) y = 5x10^{10} x $^{-0.9413}$, R^2 = 0.83, p < 10^{-5};

melanogaster) y = 7x10^{11} x $^{-1.1478}$, R^2 = 0.99, p < 10^{-5};

sechellia) y = 4x10^{10} $x^{-0.9295}$, R^2 = 0.82, p < 10^{-5};

simulans) y = 1x10^{11} x $^{-0.9765}$, R^2 = 0.897, p < 10^{-5};

yakuba) y = 40611 x $^{0.2433}$, R^2 = 0.78, p < 10^{-5};

(B) MiRanda target data for species recovered power-law trend lines with functions and non-linear regression as follows:

erecta) y = 2x10^9 x $^{-0.7755}$, R^2 = 0.77, p < 10^{-5};

melanogaster) y = 1x10^{10} x $^{-0.9393}$, R^2 = 0.98, p < 10^{-5};

sechellia) y = 8x10^8 x $^{-0.7195}$, R^2 = 0.81, p < 10^{-5};

simulans) y = 1x10^9 x $^{-0.742}$, R^2 = 0.83, p < 10^{-5};

yakuba) y = 2x10^9 x $^{-0.7597}$, R^2 = 0.76, p < 10^{-5};

(C) TargetScan network interaction data for species recovered power-law trend lines with functions and non-linear regression as follows:

erecta) y = 46837 x $^{-0.0709}$, R^2 = 0.89, p < 10^{-4};

melanogaster) y = 40014 x $^{-0.0618}$, R^2 = 0.91, p < 10^{-4};

sechellia) y = 37026 x $^{-0.0575}$, R^2 = 0.88, p < 10^{-4};

simulans) y = 39633 x $^{-0.0615}$, R^2 = 0.87, p < 10^{-4};

yakuba) y = 45503 x $^{-0.0693}$, R^2 = 0.91, p < 10^{-4};

(D) TargetScan network interaction data for species recovered power-law trend lines with functions and non-linear regression as follows:

erecta) y = 4x10^7 x $^{-0.276}$, R^2 = 0.93, p < 10^{-5};

melanogaster) y = 3x10^7 x $^{-0.2629}$, R^2 = 0.95, p < 10^{-5};

sechellia) y = 3x10^7 x $^{-0.2645}$, R^2 = 0.94, p < 10^{-5};

simulans) y = 4x10^7 x $^{-0.2809}$, R^2 = 0.94; , p < 10^{-5}

yakuba) y = 4x10^7 x $^{-0.2736}$, R^2 = 0.93, p < 10^{-5};

(A-D) All other species when fitted to power law function presented linear regression R^2 values less than 0.75.

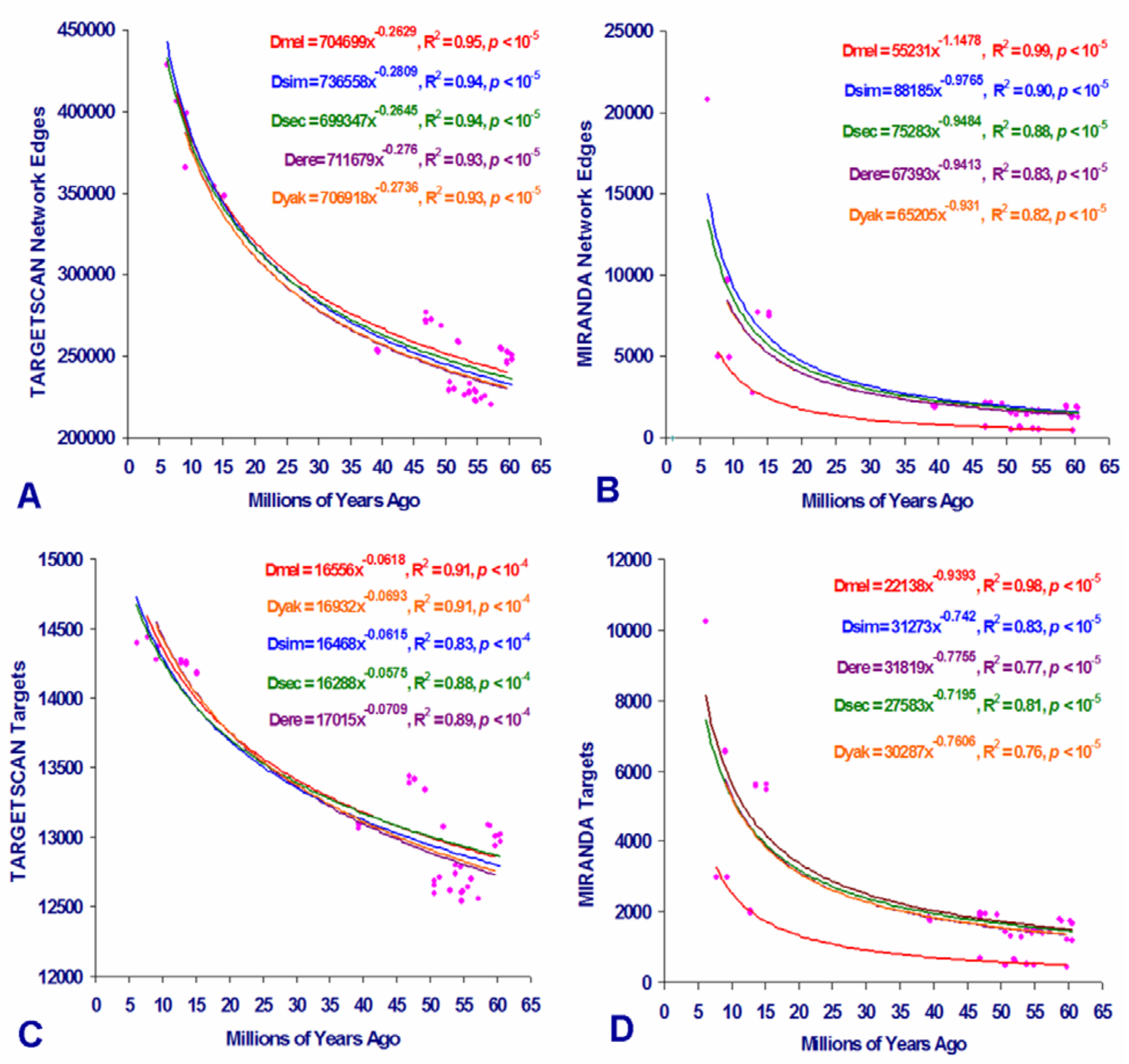


FIGURE 21. Comparisons of calculated divergence time to numbers of microRNA target and network edges recovered. Conserved microRNA regulatory network edges (A & B) edges and (C & D) nodes are presented across 66 interspecific comparisons of twelve *Drosophila* species. Trend line plots functions and coefficients of determination recovered are color coded according to select species. *Drosophila* species are abbreviated respectively: Dere) *erecta*; Dmel) *melanogaster*; Dsec) *sechellia*; Dsim) *simulans*; Dspp) the union of twelve *Drosophila* species; Dyak) *yakuba.*

(A) TargetScan target data for species recovered power-law trend lines with functions and non-linear regression as follows:

erecta) $y = 67393\, x^{-0.9413}$, $R^2 = 0.83$, $p < 10^{-5}$;

melanogaster) $y = 55231\, x^{-1.1478}$, $R^2 = 0.99$, $p < 10^{-5}$;

sechellia) $y = 75283\, x^{-0.9484}$, $R^2 = 0.88$, $p < 10^{-5}$;

simulans) $y = 88185\, x^{-0.9765}$, $R^2 = 0.90$, $p < 10^{-5}$;

yakuba) $y = 65205\, x^{-0.931}$, $R^2 = 0.82$, $p < 10^{-5}$;

(B) MiRanda target data for species recovered power-law trend lines with functions and non-linear regression as follows:

erecta) $y = 31819\, x^{-0.7755}$, $R^2 = 0.77$, $p < 10^{-5}$;

melanogaster) $y = 22138\, x^{-0.9393}$, $R^2 = 0.98$, $p < 10^{-5}$;

sechellia) $y = 27583\, x^{-0.7195}$, $R^2 = 0.81$, $p < 10^{-5}$;

simulans) $y = 31273\, x^{-0.742}$, $R^2 = 0.83$, $p < 10^{-5}$;

yakuba) $y = 30287\, x^{-0.7606}$, $R^2 = 0.76$, $p < 10^{-5}$;

(C) TargetScan network interaction data for species recovered power-law trend lines with functions and non-linear regression as follows:

erecta) $y = 17015\, x^{-0.0709}$, $R^2 = 0.89$, $p < 10^{-4}$;

melanogaster) $y = 16556\, x^{-0.0618}$, $R^2 = 0.91$, $p < 10^{-4}$;

sechellia) $y = 16288\, x^{-0.0575}$, $R^2 = 0.88$, $p < 10^{-4}$;

simulans) $y = 16468\, x^{-0.0615}$, $R^2 = 0.873$, $p < 10^{-4}$;

yakuba) $y = 16932\, x^{-0.0693}$, $R^2 = 0.91$, $p < 10^{-4}$;

(D) TargetScan network interaction data for species recovered power-law trend lines with functions and non-linear regression as follows:

erecta) $y = 711679\, x^{-0.276}$, $R^2 = 0.93$, $p < 10^{-5}$;

melanogaster) $y = 704699\, x^{-0.2629}$, $R^2 = 0.95$, $p < 10^{-5}$;

sechellia) $y = 699347\, x^{-0.2645}$, $R^2 = 0.94$, $p < 10^{-5}$;

simulans) $y = 736558\, x^{-0.2809}$, $R^2 = 0.94$, $p < 10^{-5}$;

yakuba) $y = 706918\, x^{-0.2736}$, $R^2 = 0.93$, $p < 10^{-5}$;

(A-D) All other species when fitted to power law function presented linear regression R^2 values less than 0.75.

DISCUSSION

Patterns & Models of MicroRNA Aptamer Site Evolution. The exquisite simplicity of microRNAs and their shared stem-loop structure makes these non-coding RNAs particularly amenable to phylogenetic analysis (*Drosophila* 12 Genomes Consortium, 2007). Pre-microRNA sequences are also highly conserved, evolving at about 10% of the rate of synonymous sites (*Drosophila* 12 Genomes Consortium, 2007). It is notable then that there were 27 microRNAs among a total 245 genes recovered with a consistency greater than or equal to 90% for the reference tree topology (CHAPTER III, FIGURE 22). These findings are consistent with previous analyses where microRNA conservation enables ready reconstruction of sequence evolution such that microRNAs may be utilized as molecular markers to resolve taxonomic disputes, and phylogenetic shadowing can be used to elucidate new microRNA genes (Boffelli *et al.*, 2003; Heimberg *et al.*, 2008; Rota-Stabelli *et al*., 2010; Sempere *et al.*, 2007).

The patterns of counts of *Drosophila* species exhibiting any microRNA regulation for individual target genes appear reversed for the network intersection over parent methods (FIGURE 12). An interplay of complex factors appears to operate in conservation for microRNA regulation per target gene (FIGURE 12). It may be readily discerned that a seed is well conserved and likewise a compensatory region is also well conserved; but the possession of regions in an aptamer is not well conserved (FIGURE 12). Moreover, it would seem that selective factors presiding over regulation by compensatory aptamers (MiRanda) and seed regions aptamers (TargetScan) are different (CHAPTER I; Lewis *et al.*, 2005). Consequently, the different aptamer classes (seed *vs.* compensatory) have also differentially wired the microRNA regulatory network in abundance, density and regulatory circuit conservation (CHAPTER I FIGUREs 8 & 4; CHAPTER II FIGUREs 11 & 16). Selective factors that appear to operate upon seed aptamers include cooperativity (redundancy) of interactions and transcript length. Applying the proposed model of aptamer sequence evolution (5’-seed ↔ 5’-dominant ↔ 3’-compensatory; CHAPTER I; Brennecke *et al*., 2005), this could indicate that the 5’-dominant microRNA aptamers represented in the intersection are highly species-specific (hence the high observed apomorphy), acquired with difficulty and/or easily lost to become 3’-compensatory sites when a mutation alters the seed region of the aptamer.

Seed-type regulations appear to work cooperatively: as numbers and diversity of microRNAs increase, the density and conservation of aptamers is increased (FIGURE 15A & 16A). While complementarity of seven or more bases to the 5' end microRNA is sufficient to confer regulation, seed sites are expected to be more effective when present in more than one copy due to their lower hybridization energies (Brennecke *et al*., 2005). Similarly experimental results in *Drosophila* observed that the magnitude of regulation for octamer and heptamer seeds was strongly increased when two copies of the site were introduced into a 3’UTR (Brennecke *et al*., 2005).

Likewise there is increasing selection pressure to acquire microRNA regulation through a seed-type interaction with increasing transcript length (FIGURE 19A). Indeed as noted previously, transcript length increases, the likelihood of acquisition of a seed-type aptamer binding site also increases and targets transcript may their alter in selective response to a need for up- or down-regulation (CHAPTER I, FIGURE 7). Therefore increasing length increases the likelihood of conservation of microRNA regulation using some seed-type aptamer, but the conservation of any given seed-type aptamer site is negatively correlated to transcript length (FIGURE, 18). The aforementioned study also hypothesized a widespread natural occurrence of heptamers matching microRNA seeds to account for an even microRNA-target distribution profile of TargetScan data (CHAPTER I, FIGURE 8). If the later hypothesis is correct, then TargetScan data (FIGURE14A) also indicates that all such microRNA seed matching heptamers are moderately conserved; with between < 10 to 30% percent of all individual aptamer sites conserved across all twelve *Drosophila* species studied.

It would seem that regulation through seed-type interactions is highly favored once acquired with 90.76% of such targets fully in regulation by some microRNA for all twelve *Drosophila* species (FIGURE 12). These findings indicate that seed sites are a biologically meaningful subgroup within the 5' dominant site category (Brennecke *et al*., 2005). It is of great potential biological significance that this average conservation rate for MiRanda is almost an order of magnitude lower than the rate recovered from TargetScan data (FIGURE 11). Indeed even the minimum cross-species conservation (77.68%) is higher than the maximum range (75.47%) of conservation with compensatory type regulations (FIGURE 11). Likewise, direct comparison of prediction methods across all unique *Drosophila* species combinations clearly revealed that more targets and network interactions were conserved for microRNAs employing seed-type aptamers (TargetScan) over compensatory binding sites (FIGURE 10). It has been suggested from the results of functional assays in *Drosophila* that any site with a heptamer or octamer seed should to be regarded with high likelihood of validity —especially when the seed is conserved through phylogeny (Brennecke *et al*., 2005).

Compensatory type interactions appear to have selection pressures relaxed with increasing participation of microRNAs (FIGURE 15B, 16B). This is observed where conservation declines as numbers of microRNAs increases (FIGURE 15B). Likewise, an increase in target transcript length has a moderate reduction in the conservation of compensatory type interactions (Pearson correlation -0.13; FIGURE 19B). Nevertheless, deep conservation of regulation with 3' complementarity suggests that some of these aptamer sites are likely to be functionally important (FIGURE 11 & 12; Brennecke *et al*., 2005). Moreover the compensatory nature of a 5’-dominant canonical regulation appears to outweigh its seed-type nature in terms of selective pressure (compare FIGURE 15 & 16 where profile of C matches B). The acquision of a compensatory region could confer an advantage by allowing a site to become differentially regulated by microRNA family members, but the findings of this study contradict a hypothesis that compensatory sites are acquired to allow a target gene to acquire a dependence on inputs from multiple microRNAs (Brennecke *et al*., 2005). Thus from a natural selective perspective, target regulation is functionally adequate provided one or a few compensatory aptamers are operational. Individually, canonical sites are likely to be more effective than other site types because of their higher pairing energy, and may be functional with a single aptamer (Brennecke *et al*., 2005). Between two-thirds to one-half of the seed sites seem biologically relevant (Brennecke *et al*., 2005). The later values are on par with the predicted sensitivity of TargetScan at 49.55% (see Results CHAPTER I).

Weak positive correlation was observed for conservation of microRNA regulation by species across all prediction methods (Pearson coefficient 0.15 to 0.21; FIGURE 13). But conversely, moderate to strong negative correlations for conservation of individual aptamers are observed for all methods (-0.85 to -0.76; FIGURE 14). Moreover weak negative correlation was observed across all prediction methods for conservation of aptamer interactions with numbers of aptamer sites (Pearson coefficient -0.16 to -0.55; FIGURE 17). Therefore any given aptamer site is not strictly conserved. Thus is seems that once microRNA regulation of a target transcript has been acquired, it comes under strong selection to maintain this regulation in some fashion

(FIGURE 13, 16); but the individual microRNA regulator and site of regulation involved need not be strictly conserved (FIGURE 14, 17, 18). This pragmatic biological strategy could be succinctly described: '*once [transcript] regulated, continually regulated; regardless of how regulated'*. Such a patterns would be predicted where the microRNA regulatory network must develop as an integrated whole rather than an assemblage of independent interactions (see DISCUSSION, CHAPTER III) and exherts force for phenotype canalization (see DISCUSSION, CHAPTER IV). Future work in microRNA-target conservation history should examine the aptamer sequence mutation rate for evidence of positive selection by comparing frequency of mutations synonymous to microRNA binding against mutations non-synonymous to microRNA binding (*i.e.* mutation eliminates microRNA binding site).

Regulatory Network Conservation across *Drosophila* species. A consistent set relationship of targets and interactions was retained for both microRNA target prediction methods across all interspecific comparisons: a strong positive correlation was observed for the increase of regulatory network edges relative to targets across interspecific comparison for both MiRanda and TargetScan (Pearson correlations = 0.98; FIGURE 9). Relationships for these datasets could be well explained using subtle curvilinear plots (values R^2 = 0.96 to 0.99; FIGURE 9). The recovered molecular clock rate was 1,579,192 unweighted maximum parsimony steps per million years. Estimated timings for most drosophilid divergence events under a molecular clock were within the ranges of previously published molecular times scales, inferences from biogeography, and fossil records (Tamura, Subramanian, and Kumar, 2004). Notable expectations, however, were the substantially older estimates for divergences of *D. melanogaster vs. D. simulans* at 9.32 MYA, and *D. psueudoobscura vs. D. persimilis* at 22.1 MYA. These dates which deviate from the expected divergence occur in natural history after the calibration point event. Previous molecular study has estimated the *D. melanogaster vs. D. simulans* divergence at 5.4 ± 1.1 MYA and the divergence of *D. psueudoobscura* from *D. persimilis* at 0.85 ± 0.27 MYA (Tamura, Subramanian, and Kumar, 2004).

The curvilinear shape recovered for trend lines in interspecific comparisons indicates (implies) that network change over time is constrained to retain some core functionality (FIGURES 20, FIGURE 21). This kind of curve with increasing phylogenetic distance is typical of a saturation phenomena. In such a case, networks of seed-type microRNA aptamers

(TargetScan data) saturates less rapidly than networks of compensatory microRNA aptamers (MiRanda data) (FIGURES 20, FIGURE 21). Additionally, this saturation difference may impact utility of microRNA interaction data as a phylogenetic marker (See CHAPTER III). It seems that only a modest number of microRNA–mRNA interactions may exhibit conservation over *Drosophila* cladogenesis (Grün *et al.*, 2005). Labile interactions irrespective of microRNA or target conservation history challenge interologous cross-species comparison as premise for microRNA target prediction (Matthews *et al.*, 2001). It is expected that lineage-specific genes arising in some drosophilid species subsets are those that exhibit greatest microRNA interaction plasticity (*Drosophila* 12 Genomes Consortium, 2007).

The degree of microRNA regulatory network divergence could likely dictate clade-specific reproductive isolation; and principally, the conservation between microRNAs may be viewed to indicate what kinds of genetic programs have been conserved between species (Lee *et al.*, 2007; Prochnik *et al.*, 2007). The minimal numbers of conserved microRNA-target interactions retained throughout all taxa were 1,839 from MiRanda, 13,357 from TargetScan, and 135 for the intersection of both methods. These later values likely represent the presence of a functionally-constrained core of microRNA-target interactions essential to *Drosophila*. Conversely, the minimal number of divergent microRNA-target interactions acquired since speciation of sampled *Drosophila* sister species was 445 and 97,170 for MiRanda and TargetScan, respectively. Notably, hybrids are viable among some of these combinations of sister species and future studies of microRNA network interaction conservation could consider relationships to quantified postzygotic isolation between *Drosophila* species (Coyne and Orr, 1989; Coyne and Orr, 1997). Any significant observations from such a study would have implications toward the Dobzhansky-Muller hypothesis in *Drosophila* hybridization (Dobzhansky, 1936; Muller, 1942). According to the Dobzhansky–Muller model in interspecific hybridizations, gene interactions will be disrupted by different incompatible allele combinations that are fixed in diverged species, consequently leading to hybrid sterility or inviability (Dobzhansky, 1936; Johnson and Porter, 2000; Orr, 1997; Muller, 1942). Clearly, the alteration of regulatory genetic pathways plays an important role in speciation, and these pathways can provide a plausible source for the epistatic variation implicated in the acquisition of postzygotic reproductive isolation (Johnson and Porter, 2000). Indeed computational models indicate that hybrid fitness reduction occurs more often as the number of loci in the pathway increase and as the binding site interactions become more complex; conversely, less hybrid fitness reduction is observed when the populations start with imperfect pathway bindings (Johnson and Porter, 2000).

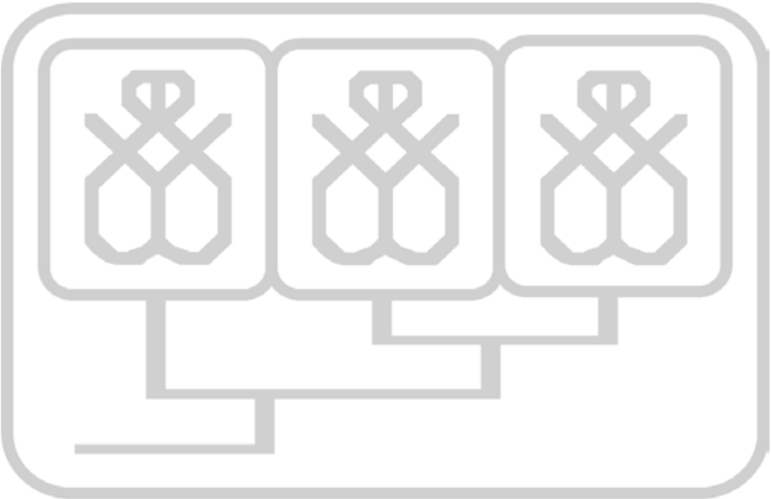
A powerful taxonomic bias is observed in FIGURES 10, 20 and 21 where R^2 values decline with decreasing relatedness to *Drosophila melanogaster* (CHAPTER III, FIGURE 22). Factors potentially compounding these analyses include the original genome assembly, predicted gene content, lineage specific expansion of the microRNA repertoire, and maturation rate heterogeneity. Of the twelve *Drosophila* genomes which have reached comparative assembly freeze 1 (CAF1) status, *D. sechellia* and *D. persimilis* were sequenced at 4X coverage level, and most others at 8X level. The *D. simulans* assembly was an exception, that is a mosaic from several different strains at 1X to 4X coverage (*Drosophila* 12 Genomes Consortium, 2007; Gilbert, 2007 Wilson *et al.*, 2008). These assemblies are very similar but differ subtly. The manual curation and reconciliation process has clearly improved areas of each assembly, but may have altered some manually curated regions and these assembly qualities affect results of comparative genomic interpretations (*Drosophila* 12 Genomes Consortium, 2007; Gilbert, 2007 Wilson *et al.*, 2008). Additionally, the non *melanogaster* species have many more predicted lineage-specific genes than *Drosophila melanogaster*; total protein-coding sequence ranges from 38.9 Megabases in *D. melanogaster* to 65.4 Megabases in *D. willistoni* (*Drosophila* 12 Genomes Consortium, 2007). Network properties may also be influenced by the presence of 47 microRNAs exhibiting specific expansion for palearctic *Drosophila* (10 microRNAs), the *melanogaster* subgroup (20 microRNAs), and *melanogaster* alone (17 microRNAs; Berezikov *et al.*, 2010). Removal of lineage-enriched microRNAs from further network analysis may mitigate the observed *melanogaster* species lineage bias. Future microRNA network conservation studies will limit analyses to 65 microRNA regulators strictly conserved within the genus *Drosophila* (Berezikov *et al.*, 2010). Additionally, mutational biases in *Drosophila* have been shown to be unequal and to fluctuate broadly among even relatively closely related species; in turn these factors have generated extensive nucleotide composition differences and differences in major codon preferences (Akashi, Kliman, and Eyre-Walker, 1998; Rodrigiuez-Trelles, Tarroi, & Ayala, 2000). Non-coding genome regions are likely to reflect greater compositional affects from mutational biases than other regions that are constrained by natural selection to maintain functionality of encoded proteins (Sueoka, 1988; Rodrigiuez-Trelles, Tarroi, & Ayala, 2000).

Future work will examine *Drosophila* conservation across multiple scales comparing microRNA regulatory network and protein-protein interaction data across species (see APPENDIX V).

Further research will examine microRNA regulatory network conservation through study of reconstructed ancestral states of microRNA-target interactions. Specifically, microRNA targets observed for reconstructed ancestral sequences from molecular data will be contrasted against expected microRNA regulatory interactions derived from internal nodes of microRNA regulatory network phylogenies (CHAPTER III). To date 71 sequence sets (in 5.05 GB) for 9 hypothetical ancestors have been prepared. These taxa include ancestors to 1) the subgenus *Drosophila*, 2) the subgenus *Sophophora*, 3) palearctic *Drosophila*, 3) the *Drosophila obscura* group, 4) the *Drosophila melanogaster* subgroup, 5) the *Drosophila melanogaster* group, and 6) the ancestor to *Drosophila erecta* + *yakuba*, 7) the ancestor to *Drosophila sechellia* + *D. simulans*; 8) the ancestor to *Drosophila melanogaster* + *D. sechellia* + *D. simulans,* and 9) the ancestor of the *Drosophila virilis* + *repleta* groups. Some ancestral state sequences were derived from PAUP* output of the interior nodes of molecular phylogenies of putative target genes using maximum likelihood with a GTR+I+G model. Other ancestral state sequences were derived with Gaps coded as a 5th character state under six variants of maximum parsimony, namely: 1) DOLLO parsimony, 2) irreversible character states (IRREV), 3, 4) Lundberg and Midpoint rooting methods, 5) standard parsimony with outgroup rooting using *Drosophila grimshawi* and "accelerated transformation" (ACCTRAN), and 6) and "delayed transformation" (DELTRAN) of characters optimized on the tree(s) in memory. Moreover, TargetScan and MiRanda predictions have been completed for reconstructed hypothetical ancestors produced from both standard parsimony with (ACCTRAN) and maximum likelihood (9.0 GB TargetScan data; 15.1 GB MiRanda data).

CHAPTER III.

Reconstruction of *Drosophila* Phylogeny using MicroRNA-Target Network Edges



Keywords & concepts: *Drosophila;* Maximum Parsimony; MicroRNA Regulatory Network Structure; MicroRNA-Target Network; Neighbor-Joining; Phylogenetic Reconstruction.

ABSTRACT

Trace evidences of natural history are persevered, not only in the sequences of genes andproteins, but also in the functional wiring of biological networks. Indeed as complements to gene-based phylogenies, there is a vast potential to accurately reconstruct phylogeny using abstract, modular representations of regulatory interactions. The specific aim of this project was to advance the methodology of making phylogenetic inference directly from network structure using microRNA interaction network data reconstructed from fruit flies (*Drosophila*). The presence or absence of individual microRNA aptamers were coded in a binary character state (0 or 1) using source data for microRNA-target aptamer predictions retrieved from the *musca* MYSQL database. Phylogenetic analyses were conducted through PAUP* under standard parsimony (MP) and distance using neighbor joining (NJ) algorithm. The signature of *Drosophila* phylogeny was found embedded within the microRNA regulatory network structure. TThe phenetic approach of Neighbor Joining recovers better signal for the reference tree toplogy over character-based standard parsimony. Consistent congruence of regulatory network phylogenies to reference species tree topology also has strong implications to understanding microRNA-target natural history that phylogenetic history were best represented when the regulatory network was treated as single entity rather than a series of separable parts. The findings of this study represent the first documented inference of phylogeny from microRNA regulatory network structure and demonstrate the potential to accurately reconstruct phylogeny using abstract representations from network architecture. It is expected that microRNA interactome network data could serve as a useful counterpart to complement or supplement DNA sequence and morphology for phylogeny.

INTRODUCTION

Evidence of common descent is persevered not only in gene and protein sequences, but also in the details of functional module wiring (Mazurie *et al.*, 2008). Multiple network alignment has been advanced as a promising means by which to infer homology between species using higher order functional data (Hartwell *et al.*, 1999). Research along these lines has steadily progressed, starting from manual alignments of metabolic pathways, to pairwise BLAST hit score based guided-alignments, to probabilistic formulations for alignment and multiple-species identification of conserved functional modules (Altschul *et al.*, 1997; Dandekar *et al.*, 1999; Flannick *et al.*, 2006; Forst & Schulten, 2001;Kelley *et al.*, 2003; Koyuturk *et al.*, 2005; Ogata *et al.*, 2000; Matthews *et al.*, 2001; Stuart *et al.*, 2003). Similarly, with the proliferation of genomic data from multiple organisms, higher level functional components are being advanced as complements to gene-based phylogenies (Mazurie *et al.*, 2008). For example, phylogenetic inference has been demonstrated from the presence or absence of enzymes in the genomes (either alone or in combination with the metabolic network structure), from the similarity or functional annotation of enzyme sequences in combination with the comparison of their direct neighbors, from the presence or absence of pathways across species, from metabolic network graph-kernels, and from the completeness of pathways across species (Clemente *et al.*, 2007; Forst *et al.*, 2006; Forst & Schulten, 2001; Heymans & Singh, 2003; Hong *et al.*, 2004; Liao *et al.*, 2002; Ma & Zeng, 2004; Mazurie *et al.*, 2008; Oh *et al.*, 2006; Zhang *et al.*, 2006). In one *Drosophila*-specific case, weighted edge network data derived from glycolytic enzyme interactions had success in recovering a phenogram largely consistent with the expected phylogeny (Clark & Wang, 1994).

These preceeding findings demonstrate that network structure is not just strongly correlated to phylogeny but underscore the vast potential to accurately reconstruct phylogeny using abstract, modular representations of metabolic reactions (Mazurie *et al.*, 2008, Suthram *et al.*, 2005). By extension, the presence of phylogenetic information within microRNA-target network architecture is anticipated. Large sets of microRNA-target interactions of a regulatory network could represent valuable phylogenetic markers because while individual regulatory reactions may be subject to strong positive or negative selection, it is hypothesized based on microArray expression data that a neutral selective property presides over the entire interactome, where the rate of transcriptome change is proportional to time and the majority of gene expression differences within and between species are not functional adaptations, but selectively neutral or nearly neutral. (Khaitovich *et al.*, 2004; Rajewsky, 2006). Consistent congruence of regulatory network phylogenies to a reference species tree topology will have strong implications to understanding microRNA-target natural history. Thus far, pilot studies have recovered phylogenetic information from the weighted edges of microRNA networks and produced species tree topologies fully (or mostly) congruent with an expected topology (APPENDIX IV, FIGURE 38). Moreover, a whole genome regulatory network phylogeny advances the methodology of making phylogenetic inference directly from network structure, and provides a valuable medium to investigate gene regulatory interactions in *Drosophila* speciation (Gompel *et al.*, 2005, Mazurie *et al.*, 2008; McGregor *et al.*, 2007; Sucena & Stern, 2000; Suthram *et al.*, 2005). Therefore, the purpose of this study was to recover novel phylogenetic reconstructions from binary character data of microRNA-target regulatory networks. Indeed the findings presented here represent the first documented inference of phylogeny from microRNA regulatory network structure.

METHODS

The presence or absence of individual microRNA aptamers were coded in a binary character state (0 or 1) using source data for microRNA-target aptamer predictions (described in CHAPTER I) retrieved from the *musca* MYSQL database (FIGURE 1). Data extracted from *musca* database were formatted into Nexus file with command line for phylogenetic analyses. Nexus file lengths were of 241,939 lines for MiRanda, 1,090,302 lines for TargetScan, and 783,56 for the network intersection of prediction methods. Phylogenetic analyses conducted through PAUP* under standard parsimony (MP) and distance using neighbor joining (NJ) algorithm. The single most parsimonious rooted tree retrieved from parsimony reconstruction under a branch-and-bound search with outgroup rooting using *Drosophila grimshawi* and accelerated transformation of characters optimized on the tree(s) in memory. Other reconstructions were conducted for select datasets under DOLLO parsimony and parsimony with irreversible character states. Additional tests performed under parsimony criteria included a constrained-tree topology-dependent permutation tail probability test to the reference tree (Faith, 1991) where 10,000 randomized matrices were used to generate a null distribution. The T-PTP test statistic is calculated by subtracting minimum tree length under constrained monophyly from minimal unconstrained tree length (ΔL = range of steps; *L = length difference for unpermuted data), can be interpreted as significant support for a specified monophyly (Carpenter *et al.*, 1998; Faith and Trueman, 1996, Swofford *et al.*, 1996). Branch supports of trees were evaluated by nonparametric bootstrap (BP), third- and half-delete jackknife (JK) calculated to high confidence levels using 10,000 replicates.

RESULTS

TargetScan MicroRNA-Target Network Data generated an aligned matrix of 1,090,221 binary characters; of which 82,388 (7.56%) characters were constant, 293,774 (26.95 %) characters were variable but parsimony-uninformative, and 714,059 (65.50%) characters were potentially parsimony-informative. The sum of minimum possible lengths was 1,007,833 and the sum of maximum possible lengths was 2,846,378. The parsimony based reconstructions for TargetScan data recovered 1 shortest tree of 194,5742 steps in length (FIGURE 23) with Consistency Index (CI) = 0.518, Retention Index (RI) = 0.490, Rescaled Consistency Index (RC) = 0.254, Homoplasy Index (HI) = 0.482, and Goloboff-fits (G-fit) of -518,490.732. Likewise the distance-based reconstructions for the TargetScan dataset recovered NJ tree (FIGURE 24) of 1,945,742 steps in length and parsimony scores of CI = 0.518, RI = 0.490, RC = 0.254, HI = 0.482, and G-fit = -518,490.732. Notably the trees recovered for maximum parsimony and neighbor joining were identical in topology, length, and parametric scores. The recovered tree topology of both reconstruction methods was extensively congruent to the established drosophilid phylogenies (FIGURE 22) and differing only in the placement of *Drosophila erecta* and *D. yakuba*. These trees were also well supported with bootstap and jackknife frequencies of 100% for all nodes. Additionally, in comparison to the other datasets, the TargetScan data demonstrate the best reconstruction to the reference phylogeny (FIGURE 22) under maximum parsimony criteria

MiRanda MicroRNA-Target Network Data recovered an aligned matrix of 241,861 binary characters. There was 1 (4.13×10^{-4}%) constant character, 157,494 (65.12 %) characters which were variable but parsimony-uninformative, and 84,366 (34.88%) characters that were potentially parsimony-informative. The sum of minimum possible lengths was 241,860 and the sum of maximum possible lengths was 375,468. The parsimony based reconstructions for MiRanda data recovered 1 shortest tree of 310,994 steps in length (FIGURE 23) with CI = 0.778, RI = 0.483, RC = 0.375, HI = 0.222, and G-fit = –68,707.139. The parsimony tree was largely incongruent to the established *Drosophila* phylogeny. It differed in 1) the placement of *Drosophila ananassae* as sister to the *D. obscura* group, 2) the placement of *D. melanogaster* as sister to other species of the *melanogaster* subgroup, 3) the placement of *Drosophila willistoni* with species of the subgenus *Drosophila*, 4) and in the placement of *Drosophila erecta* and *D. yakuba* nearest sisters. Nevertheless, this tree shape was highly reproducible with bootstrap and jackknife frequencies of 99 to 100% for all nodes.

Distance-based reconstructions for the MiRanda dataset recovered NJ tree (FIGURE 24) of 312,404 steps in length. This tree recovered parsimony scores of CI = 0.774, RI = 0.472, RC = 0.365, HI = 0.226, and G-fit = -68,374.746. In contrast to maximum parsimony, the neighbor-joining method produced a tree that was largely congruent to the reference tree for *Drosophila* and differing only in the union of *D. erecta* and *D. yakuba* into a clade. This tree was highly supported with bootstrap and jackknife frequencies of 100% for all nodes.

Intersection of MicroRNA Target prediction methods Network Data produced an aligned binary matrix of 78,280 characters. Of these, 3,405 (4.35%) characters were constant, 61,184 (78.16%) characters were variable but parsimony-uninformative, and 13,691 (17.49%) characters were potentially parsimony-informative. The sum of minimum possible lengths was 74,875 and the sum of maximum possible lengths was 93,785. The parsimony based reconstructions for network intersection data recovered 1 shortest tree of 85,050 steps in length (FIGURE 23) with CI = 0.880, RI = 0.462, RC = 0.407, HI = 0.120, and G-fit = -11,276.436. This parsimony tree was largely incongruent to the reference phylogeny and substantially differed in 1) the dissolution of the *Drosophila obscura* group clade, 2) the placement of *Drosophila ananassae* as sister to the *D. persimilis*, 3) and in the placement of *D. erecta* and *D. yakuba* nearest sisters, and 4) the placement of the *Drosophila mojavensis and D. virilis* clade embedded into the subgenus

Sophophora as sister to *Drosophila persimilis* and *D. ananassae.* Node supports were variable by bootstrap and jackknife ranging from < 75 to 100 %. Furthermore, tree topology did not alter with reconstruction by other character-based settings; namely DOLLO parsimony and parsimony with irreversible character states. A topology-dependent permutation test for the dataset against the established *Drosophila* phylogeny indicated significant incongruence of the data for the expected topology (T-PTP *L= -1,470, ΔL = - 803 to -1,121, *P* = 1).

Neighbor-Joining reconstructions for the network intersection dataset recovered a tree (FIGURE 24) of 85,223 steps in length and parsimony scores of CI = 0.879, RI = 0.453, RC = 0.398, HI = 0.121, and G-fit = -11,230.950. Congruence to the reference phylogeny improved with neighbor-joining over standard parsimony. Nevertheless, the NJ tree deviated from the expected phylogeny in 1) the dissolution of the *Drosophila obscura* group clade, 2) the placement of *Drosophila ananassae* as sister to the *D. persimili*s, 3) the placement of *D. melanogaster* as sister to other species of the *melanogaster* subgroup, 4) and in the placement of *Drosophila erecta* and *D. yakuba* nearest sisters. Bootstrap and jackknife support values were variable across the tree, ranging for < 75 to 100 %.

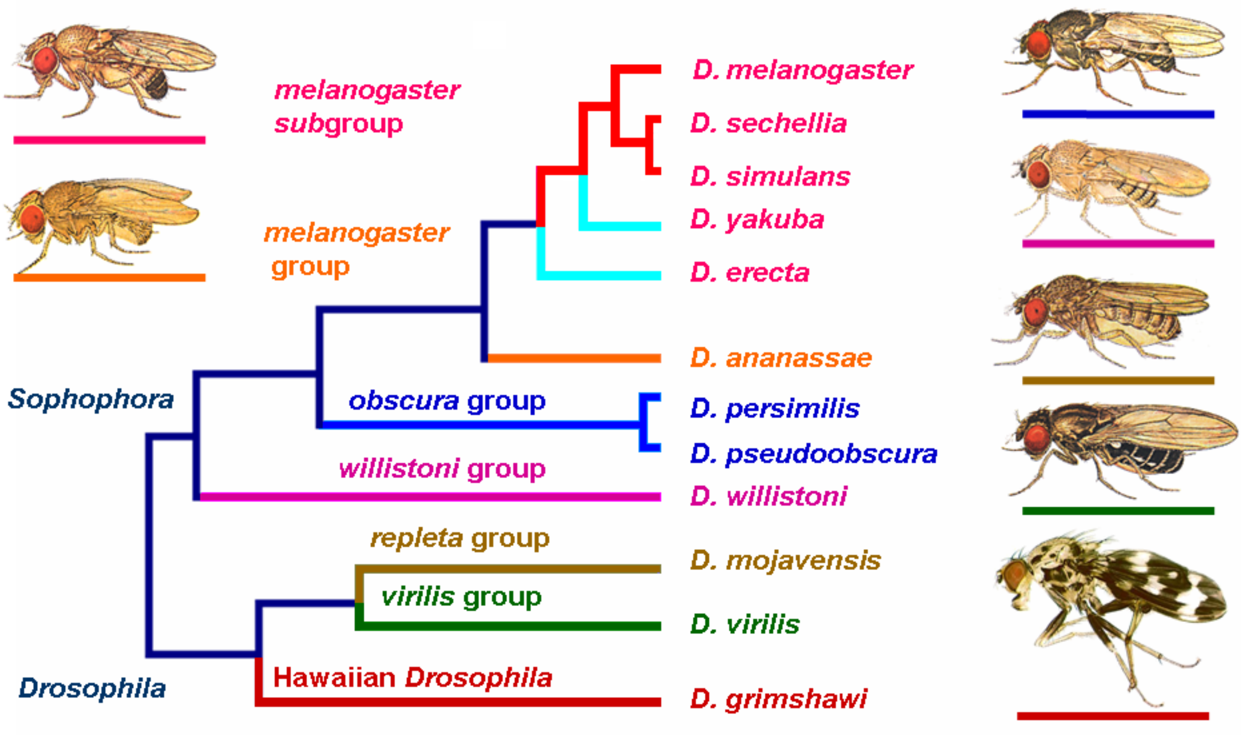


FIGURE 22. Reference phylogeny of the genus *Drosophila*. Included species are those with complete genome sequences available. Illustrated example species are underscored in color according to their respective branch (Gilbert, 2007, Wilson *et al.*, 2008). This phylogeny is consistent with data from chromosome homology, species morphology, and concatenated gene sequence data. Alternate phylogenies place *Drosophila erecta* and *D. yakuba* in a clade sister to *D. melanogaster*, *D. sechellia*, and *D. simulans* (Ko, *et al.*, 2003; Pollard *et al.*, 2006).

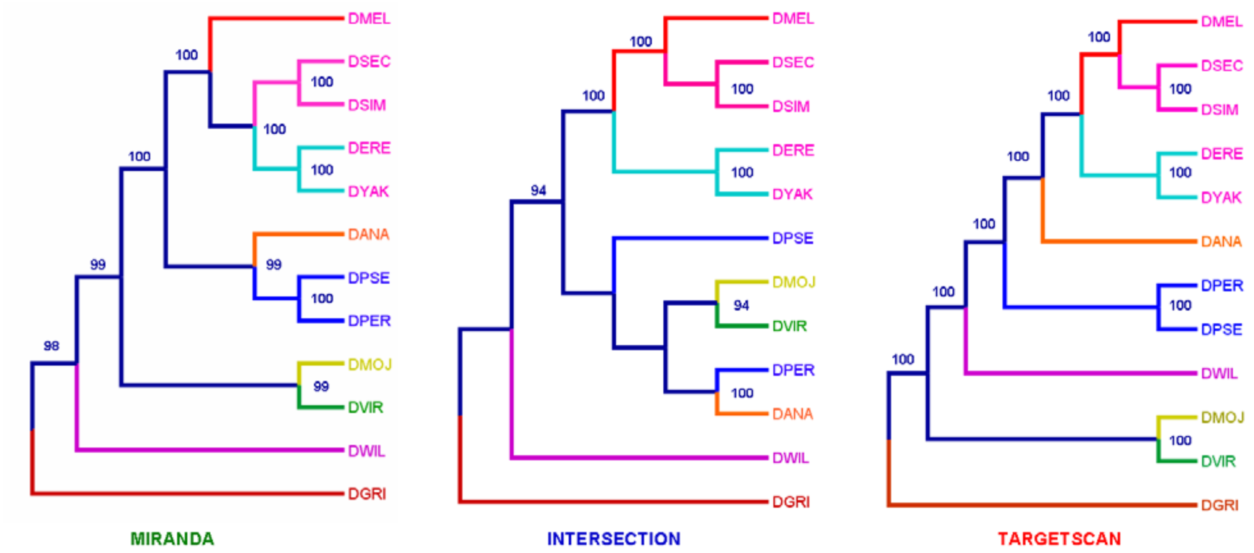


FIGURE 23. Cladograms of the shortest tree recovered from maximum parsimony using microRNA-target interaction data from MiRanda, TargetScan and the network intersection of prediction methods. These cladogram represent the single most parsimonious rooted tree retrieved from parsimony reconstruction under a branch-and-bound search with outgroup rooting using *Drosophila grimshawi* and accelerated transformation of characters optimized on the tree(s) in memory. *Drosophila* species are abbreviated respectively: DANA) *ananassae*; DERE) *erecta*; DMEL) *melanogaster*; DMOJ) *mojavensis*; DPER) *persimilis*; DPSE) *pseudoobscura*; DSEC) *sechellia*; DSIM) *simulans*; DVIR) *virilis*; DYAK) *yakuba.* Bootstrap values >75% appear near their respective branches. Branches are color-coded to match the reference phylogeny of the genus *Drosophila* presented in FIGURE 22. The cladograms derived MiRanda and TargetScan network data are congruent to the reference tree expect for the cladal placement of *D. yakuba* and *D.erecta* as nearest sisters.

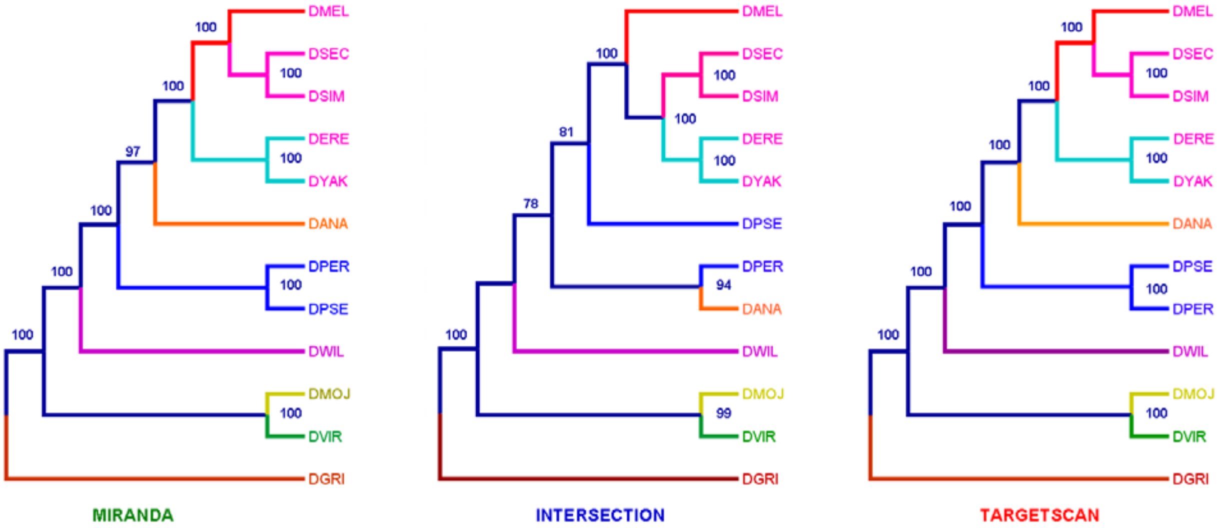


FIGURE 24. Cladograms recovered from Neighbor-Joining using microRNA-target interaction data from MiRanda, TargetScan and the network intersection of prediction methods. *Drosophila* species are abbreviated respectively: DANA) *ananassae*; DERE) *erecta*; DMEL) *melanogaster*; DMOJ) *mojavensis*; DPER) *persimilis*; DPSE) *pseudoobscura*; DSEC) *sechellia*; DSIM) *simulans*; DVIR) *virilis*; DYAK) *yakuba*. Bootstrap values >75% appear near their respective branches. Branches are color-coded to match the reference phylogeny of the genus *Drosophila* presented in FIGURE 22. The cladograms derived MiRanda and TargetScan network data are congruent to the reference tree expect for the cladal placement of *Drosophila yakuba* and *D. erecta* as nearest sisters.

DISCUSSION

The findings of this study represent the first documented inference of phylogeny from unweighted microRNA regulatory network structure and demonstrate the potential to accurately reconstruct phylogeny using abstract representations from network architecture (Mazurie *et al.*, 2008, Suthram *et al.*, 2005). Likewise pilot studies using character data derived from weighted edges of a microRNA-Target network also recovered support a hypothesis that weighted edge microRNA network structure itself can be directly utilized for phylogenetic inference (see APPENDIX IV, FIGURE 38). Differences in recovered tree topologies likely reflected the influence of underlying methodological biases incurred when analyzing numerical character of data. As much (99% or more) of the apparent support for an "optimal" tree can originate from an inherent methodological bias rather than actual phylogenetic signal (Swofford *et al.*, 2001). The only recourse then is to sample a wide range of methods (Swofford *et al.*, 2001). Nevertheless, consistent congruence of regulatory network phylogenies to a reference species tree topology has strong implications to understanding microRNA-target natural history.

The consistent recovery of *Drosophila yakuba* and *D. erecta* as nearest sisters for all datasets is not unexpected. The exact phylogenetic relationship between the later species has been a subject of some controversy (Ko, *et al.*, 2003; Pollard *et al.*, 2006). Alternate phylogenies to the reference tree place *Drosophila erecta* and *D. yakuba* in a clade sister to *D. melanogaster*, *D. sechellia*, and *D. simulans* (Ko, *et al.*, 2003; Pollard *et al.*, 2006). Indeed it would seem that the later phyloegentic hypothesis is strongly favored by phylogenetic reconstruction using regulatory network data. Molecular phylogenetic analyses with special focus on *Drosophila erecta* and *D. yakuba* recovered widespread, statistically significant, and robust incongruence in nucleotide and amino acid substitutions, insertions and deletions. These results are consistent with a hypothesis of incomplete lineage sorting between *Drosophila erecta* and *D. yakuba* where the same ancestral polymorphisms became fixed during the two rapid speciation events that led to these species (Pollard *et al.*, 2006).

The neighbor-joining method is rooted in phenetic philosophy and is based on the minimum-evolution criterion for phylogenetic trees (Saitou & Nei, 1987). Phenetics, also known as taximetrics, attempts organismal classification through overall similarity (or variation; Sneath & Sokal, 1973). The neighbor-joining method has become the most widely used distance method for building phylogenetic trees; indeed the original paper has been cited about 13,000 times (Gascuel & Steel, 2006). But because it is a greedy algorithm that constructs the tree in a step-wise fashion and may not find a tree topology with least total branch length, the neighbor-joining method has been extensively superseded in phylogenetics by methods that do not rely on distance measures (Mihaescu, Levy, & Pachter, 2006; Saitou & Nei, 1987). Even so, where the method has it has been extensively tested it usually finds a tree that is quite close to the optimal tree (Mihaescu, Levy, & Pachter, 2006). In the case of this study it is apparent that the phenetic approach of Neighbor Joining recovers better signal for the reference tree toplogy (FIGURE 22) over the character-based approach of standard parsimony (compare MiRanda in FIGURE 23 & 24). This would only be expected if phylogenetic history were best reconstructed when the regulatory network was treated as the sum of total interactions rather than a series of separated interactions. Indeed, along these lines it has been hypothesized that while individual regulatory reactions may be subject to strong positive or negative selection, a property of selective neutrality presides over the entire interactome (Khaitovich *et al.*, 2004; Rajewsky, 2006). Thus it seems that microRNA networks have formed as a tightly integrated unit rather than an assemblage of independent interactions. This view harmonizes to earlier findings for considering conservation of microRNA regulation and conservation of individual aptamers (see CHAPTER II commentary on FIGURE 13 & 14).

Separate consideration of seed (TargetScan) and compensatory (MiRanda) aptamer data appears to perform better for phylogenetic reconstruction than consideration of aptamers in both categories (5'dominant cannonical; network intersection data). The results of the topology dependent permutation test indicate that that network intersection interactome data could not be reconciled to accommodate the shape of the reference phylogeny. The source of the incongruence is not clear at this time. Much phylogenetic signal appears to be lost and relationships are with the overall network conservation at less than 10% across all species (CHAPTER II FIGURE 12). Nevertheless, relationships of some nearest sister species were resolvable for network intersection data. It is anticipated though that reduction of the dataset to 65 microRNAs strictly conserved in all *Drosophila* may improve topological congruence of an intersection network phylogeny to the reference tree (Berezikov *et al.*, 2010). Future research will also examine the influence of incremental increase of threshold values for hybridization energy upon phylogenetic reconstruction using MiRanda microRNA target networks. It is hypothesized that congruence to reference tree topology will improve with more stringent values for Gibbs free energy. Further phylogenetic reconstructions should also be conducted using Bayesian inference through MrBayes software (Huelsenbeck & Ronquist, 2005; Swofford, 2002). Differing methods of reconstruction are expected to recover phylogenetic topologies with the differing levels of resolution.

The utility of microRNA-target interactions as characters for phylogenetic analyses may also be applied to selection of novel molecular markers. Polymerase Chain Reaction (PCR) amplification using primers matching conserved microRNA seeds could be produced and utilized in a manner similar to the use of retrotransposon marker data for the reconstruction of phylogeny (Shedlock and Okada, 2000). In the later case, PCR oligonucleotide primers facing outwards from retrotransposons are made to amplify between two retroelements inserted into the genome; the number and sizes of fragments amplified differ between lineages and inter-retrotransposons amplified polymorphisms (IRAPs) can be used as phylogenetic markers (see Flavell *et al.* 1999; Kalendar *et al.* 1999; Kumar & Hirochika 2001).

It is expected that microRNA interactome network data could serve as a useful counterpart to complement DNA sequence and morphology for phylogeny. Additionally microRNA regulatory network phylogenies may be combined and contrasted to other phylogenies derived from protein-protein interactions. To date additional phylogenetic analyses have been were initiated for a sample of protein interaction and paralogy data constructed from reciprocal BLAST data of 12 *Drosophila* species using whole mRNA libraries, EISE_exonerate, EISE_genemapper, EISE_genewise, and GLEANR computationally predicted annotations accessible through FlyBase and DroSpeGe (Birney, *et al.*, 2004; Chatterji and Pachter, 2006; Heger and Ponting, 2006; Mackey *et al.*, 2006; Slater and Birney, 2005). Pilot studies for phylogenetic reconstruction using protein-protein network interaction data are described in APPENDIX V. These approaches continue to advance the methodology of making phylogenetic inference directly from network structure, and provides a valuable medium to further investigate gene regulatory interactions in *Drosophila* speciation (Gompel *et al.*, 2005, Mazurie *et al.*, 2008; McGregor *et al.*, 2007; Sucena & Stern, 2000; Suthram *et al.*, 2005).

CHAPTER IV.

MicroRNA-Target Network Topology across Regions of Chromosomal Synteny (Muller Elements) and Genes linked to species Diagnostic Phenotype in *Drosophila*.

Keywords & concepts: Chromosome Synteny; *Drosophila* systematics; FANMOD; Fast-X Hypothesis; GRAFMAN Software; Haldane’s Rule; MicroRNA-Target Network; Molecular Markers; Muller Elements; Network Motifs; Network Topology Species Diagnostic Phenotype.

ABSTRACT

While a strong case can be substantiated for microRNA-moderated control over basic of animal anatomy, roles of microRNA regulation for details of fly anatomy remain largely unexplored. Data form molecular phylogeny and network topology were integrated with select subnetworks and emergent implications to natural selection were considered. To these ends, associations for regions of major chromosomal synteny across twelve *Drosophila* species were prepared and literature review for genes linked to anatomical features and physiological processes features used to diagnose species within the genus *Drosophila*. Directed networks from the intersection of MicroRNA prediction methods were enumerated under FANMOD for subgraphs of size 3 and 4 using 1000 replicates and 10,000 random network samples. Topological analyses of microRNA regulatory networks recovered significant enrichment for the S2T2 motif possessing a redundant link (motif-204) in all twelve species sampled for many Muller elements. The network enrichment of motifs possessing partial internal redundancy would have powerful implications toward understanding *Drosophila* speciation at the level of microRNA-gene regulatory interactions: this would suggest that optimization of the whole interactome topology itself has been historically subject to natural selection where resilience to attack have offered selective advantage. The findings presented in this study represent a novel intergration of microRNA regulatory network topology to chromsomal synteny and genes linked to species diagnostic phenotypes. Collective patterns observed indicate that respective Muller element networks have developed within the *Drosophila* transcriptome as separate regulatory modules. The repeating motif patterns across elements observed would not be expected if Muller elements were not a natural subdivision of the total *Drosophila* regulatory network. Given the results of this study, implications toward the genetic basis of Haldane’s rule are discussed.

INTRODUCTION

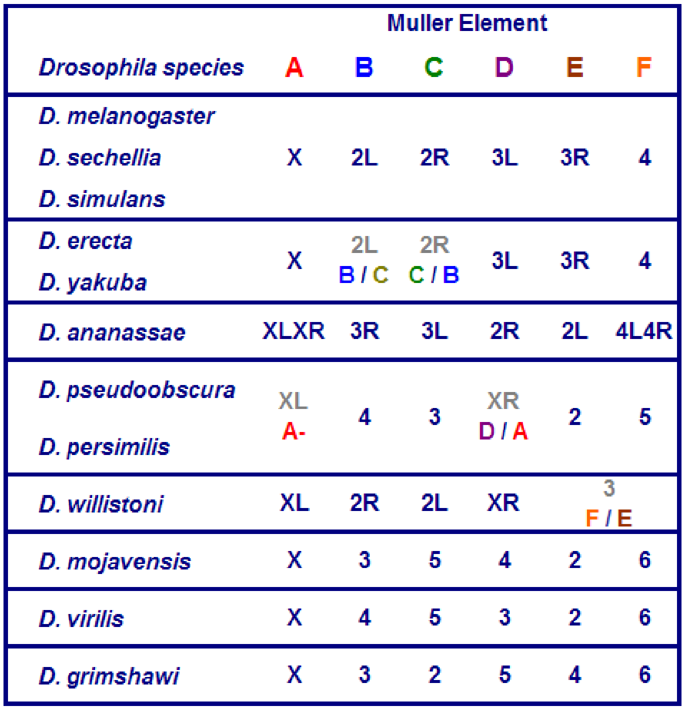
It is hypothesized that certain reoccurring subgraphs will be over-represented for all microRNA-target networks in *Drosophila*. These later subgraphs represent network *motifs* where they occur in complex networks at frequencies much higher than those in randomized networks; potentially any network subgraph can be a motif if it obeys this definition. Such motifs are present in networks from biochemistry, neurobiology, ecology, and engineering (Milo *et al.*, 2002). Notably, the larger the network, the more significant the presence of the motif and these motifs may serve as a fingerprint of network functionality; thus for instance, conserved proteome network motifs allow for prediction of protein-protein interactions (Albert & Albert, 2004). Motifs may thus define universal classes of networks (Milo *et al.*, 2002).

Previous network based approaches have included a genome-scale *Caenorhabditis elegans* microRNA regulatory networks that contained experimentally-mapped computationally predicted transcription factor and microRNA interactions (Martinez *et al.*, 2008). The integrated microRNA network recovered 23 high flux capacity composite feedback loops in which a transcription factor controls a microRNA that is itself regulated by that same microRNA; such loops occurred more frequently than expected by chance and likely constitute a genuine network motif (Martinez *et al.*, 2008). Similarly, 17 significant motifs, of which the regulated feedback loop was the most significantly overrepresented, were recovered from microRNA and transcription factor integrated networks built from MiRanda and PicTar mammalian prediction data (Yu *et al.*, 2008).

While the aforementioned studies have revealed several network motifs relating microRNAs to other regulatory factors, there has been considerably less emphasis on examining the selective functionality of network motifs (Yu *et al.*, 2008). To these ends, data form molecular phylogeny and network topology were integrated with select subnetworks and emergent implications to natural selection were considered. In the case of drosophilid regulatory networks, the motifs expected to exhibit enrichment are those possessing partial internal redundancy (Bonchev *et al.*, 2009; Martinez *et al.*, 2008; Tsang *et al.*, 2007). The presence of such motifs would have powerful implications toward understanding *Drosophila* speciation at the level of microRNA-gene regulatory interactions: this would suggest that optimization of the whole interactome topology itself has been historically subject to natural selection where resilience to attack have offered selective advantage (Hornstein, & Shomron, 2006; Martinez *et al.*, 2008; Tsang *et al.*, 2007; Yu *et al.*, 2008). Specific focus was directed to *Drosophila* chromosome elements and genes linked to species diagnostic phenotypes. Notably, the findings presented in this study represent a novel intergration of microRNA regulatory network topology to chromsomal synteny and genes linked to species diagnostic phenotypes. Moreover, these analyses in network architecture will complement the existing whole genome network analyses of Mammalian and *Caenorhabditis* microRNA targets, and further the study of drosophilid speciation within the scope of gene-regulatory networks (Hornstein, & Shomron, 2006; Martinez *et al.*, 2008; Tsang *et al.*, 2007; Yu *et al.*, 2008). Additionally, the evaluation of gene region phylogenetic utility under the reference tree would have great potential applications to rationalizing selection of molecular markers or morphological characters for *Drosophila* systematics.

METHODS

Associations for regions of major chromosomal synteny across twelve *Drosophila* species were prepared for the total putative microRNA target dataset of using genome scaffold information available through DroSpeGe database (14570 genes; Gilbert, 2007). Although *Drosophila* species vary in their number of chromosomes, there are six fundamental chromosome arms common to all species (*Drosophila* 12 Genomes Consortium, 2007; Sturtevant & Novitski, 1941). Most pairs of orthologous genes are found on the same Muller element, but there is extensive gene shuffling within Muller elements between even moderately diverged genomes (Sturtevant & Novitski, 1941; *Drosophila* 12 Genomes Consortium, 2007). The nomenclature suggested by Muller (1940) remains the prevailing convention. According to this system the recognizable elements are lettered, in the sequence familiar in melanogaster: the X-chromosome of that species becomes A; 2L, B; 2R, C; 3L, D; 3R, E; 4, F (TABLE 1; Sturtevant & Novitski, 1941). A configuration like that found in *D. virilis* (where all six elements are separate) likely reflects the ancestral karyotype of *Drosophila* (Sturtevant & Novitski, 1941). The provisional chromosome “U” contains 34,630 small scaffolds produced by the Celera shotgun assembler which could not be consistently joined with larger scaffolds (Wilson *et al.*, 2008). Genes relegated to the U-Muller element are to be regarded as unsorted. Because their position is not assigned, U-Muller associated target genes were regarded as a control or baseline for the transcriptome irrespective of chromosome loci.



	Muller Element					
Drosophila species	A	B	C	D	E	F
D. melanogaster *D. sechellia* *D. simulans*	X	2L	2R	3L	3R	4
D. erecta *D. yakuba*	X	2L B / C	2R C / B	3L	3R	4
D. ananassae	XLXR	3R	3L	2R	2L	4L4R
D. pseudoobscura *D. persimilis*	XL A-	4	3	XR D / A	2	5
D. willistoni	XL	2R	2L	XR	3 F / E	
D. mojavensis	X	3	5	4	2	6
D. virilis	X	4	5	3	2	6
D. grimshawi	X	3	2	5	4	6

TABLE 1. Standard chromosome numbering and Muller element associations for regions of conserved synteny across twelve species of *Drosophila*. Chromosomes are numbered in descending order by length where chromosome 1 is also the X heterosome for all *Drosophila* species. The designations L and R represent the long and short chromosomal arms respectively. The haploid number of chromosomes in *Drosophila* varies from three to six and five acrocentric rod chromosomes is the ancestral state for the genus. Recombinant chromosomes relative to *D. melanogaster* are colored grey and participating Muller elements are indicated. *Drosophila erecta* and *D. yakuba* exhibit a pericentric inversion of chromosome 2 where Muller elements B and C reordered B / C and C / B from telomere to centromere. Muller elements A and D are fused and pericentrically inverted to form a metacentric X for *Drosophila pseudoobscura* and *D. persimilis.* A similar A / D Muller element fusion lacking inversion/transposition is observed for *Drosophila willistoni*. Additionally for *Drosophila willistoni* there is a fusion of Muller element F into the distal end of the E element (*Drosophila* 12 Genomes Consortium, 2007; Muller, 1940; Sturtevant & Novitski, 1941; Wilson *et al.*, 2008).

Literature review for anatomical features and physiological processes features used to diagnose species within the genus *Drosophila*, recovered a novel list of 118 FlyBase anatomy terms (FBbt) and 93 gene ontology (GO) categories (see APPENDIX VI; Grimaldi, 1990; Markow & O'Grady, 2005; Wilson *et al.*, 2008). Batch downloads were conducted for each FBbt and GO through FlyBase for genes known to influence phenotypes specific to the term (2,331 genes recovered; matching 14.38% of the total potential microRNA target dataset; Wilson *et al.*, 2008). By convention of this paper the data for suite of genes associated to species diagnostic phenotypes are abbreviated as "FBbt". Gene associations by Muller element and FBbt data were imported as tables into *musca* MySQL database. Parametric scores derived from molecular phylogeny (see CHAPTER II) for these user-defined subnetworks of interest were recovered through the *musca* database (see APPENDIX III, TABLE 12).

Using source data for the intersection of microRNA-Target methods detailed in CHAPTER I, network data for was downloaded from the *musca* database for the FBbt and Muller element datasets. Network topolgy analyses were conducted through FANMOD using the Godel supercomputing cluster in the VCU Center for the Study of Biological Complexity and 451 KB of data were output (CHAPTER I, FIGURE 1). FANMOD is a tool for network motif detection that implements at orders of magnitude faster than any other existing algorithm for this task; facilitating the detection of larger motifs in bigger networks than otherwise feasible (Wernicke & Rasche, 2006). By convention of this paper, motifs are identified by the number of starting (S) and terminating points (T) and an adjective identification defined from FANMOD. The network motifs are those patterns for which the probability of appearing in a randomized network an equal or greater number of times than in the real network is lower than a cutoff value (Milo *et al.*, 2002). Additionally FANMOD makes use of a Z-score parameter to quantify the difference from normal of a recovered motif. Directed networks from the intersection of MicroRNA prediction methods were enumerated under FANMOD for subgraphs of size 3 and 4 using 1000 replicates and 10,000 random network samples. Study was limited to data from the intersection of target prediction methods due to the computational limits of FANMOD. Further network quantification and distance analyses were performed using in-house GRAFMAN software available under Linux on the Watson supercomputer cluster of Virginia Commonwealth University (FIGURE 1; Karabunarliev & Bonchev, 2002). Data recovered from the *musca* database in the form of 143 network files (5.71MB) were input to GRAFMAN and 360KB of data were output (see APPENDIX III, TABLE 9).

RESULTS

Network adjacency, distance quantification, and parametric score tables for microRNA regulatory from Muller element and FBbt data are presented in APPENDIX III, TABLEs 9 & 12. Informative characters sets made up 70% of the *Drosophila* genome, with 58 to 82% informative characters for user defined subnetworks of interest (APPENDIX III, TABLE 12). Molecular phylogenies with these subnetworks of genes would recover < 1 to 42% of total tree length. Likewise subnetworks of genes of interest gave <1 to 90.4 % coverage of total gene content. The drosophilid life cycle marking stage-specific numbers of FlyBase anatomy terms (FBbt) and Gene Ontology (GO) categories features available to diagnose *Drosophila species* is presented in FIGURE 25. Likewise individual FlyBase terms are described and cross-listed with life stage in APPENDIX VI, TABLE 13**.** It is also of note that the 1355 genes of the FBbt dataset (12% total genome characters) recovered the following average parametric scores per gene when molecular phylogeny was constrained to the topology of the reference tree: CI = 0.726, HI = 0.274, RC = 0.274, RI = 0.504, Goloboff-fit = -4461.012, gamma rate = 2.944 gamma rate, invariable sites rate = 0.227 invariable sites rate, -*ln* likelihood score = 47128.40, and 0.018 average bits of information per base (See Methods of CHAPTER II; FIGURE 22; APPENDIX III, TABLE 12). Comparable parametric score values are displayed in TABLE 12 (APPENDIX III) for 155 molecular markers taken from 590 operational taxonomic units (OTUs, typically species) available through Genbank (NCBI; 187 FlyBase records available, 1% total dataset). Network node distribution frequencies according to Muller elements and FBbt data are illustrated for microRNA targets of *Drosophila melanogaster* in FIGURE 26.

Topologies and conversion pathways for motifs observed in *Drosophila* microRNA-target interaction networks are presented in FIGURE 27. Two types of size 3 and four types of size 4 motifs were observable in strictly bipartite microRNA-target interaction networks. The S2T1 motif-36 is superimposable upon the S3T1 motif-2184, the S2T2 motif-140 and the S2T2 motif-204. Conversely, the S1T2 motif-6 is superimposable upon the S2T2 motif-140, the S1T3 motif-14, and S2T2 motif-204. Motifs of subgraph size 4 are produced from the S2T1 motif-36 or the S1T2 motif-6 with the addition of a microRNA regulator or target node to either size 3 subgraph.

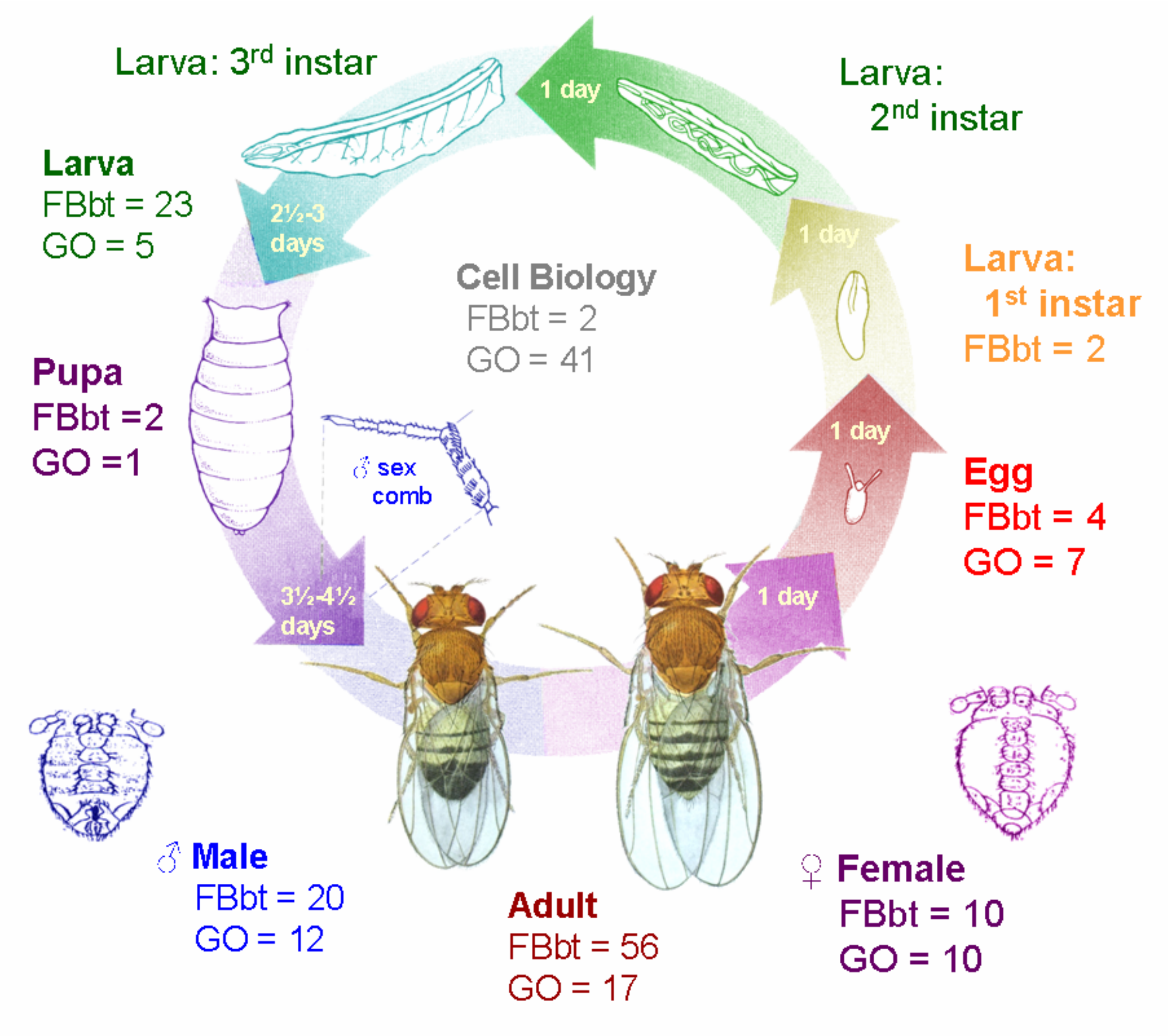


FIGURE 25. Drosophilid life cycle marking stage-specific numbers of 118 FlyBase anatomy terms (FBbt) and 93 Gene Ontology (GO) categories features available to diagnose *Drosophila species* (Wilson *et al.*, 2008). For each FBbt and GO, batch downloads were conducted through FlyBase for genes known to influence phenotypes specific to the term: 2331 genes were recovered and 1355 of these recovered microRNA targets representing 8.94% of total potential microRNA target dataset. Redrawn from Mertens and Hammersmith (2007).

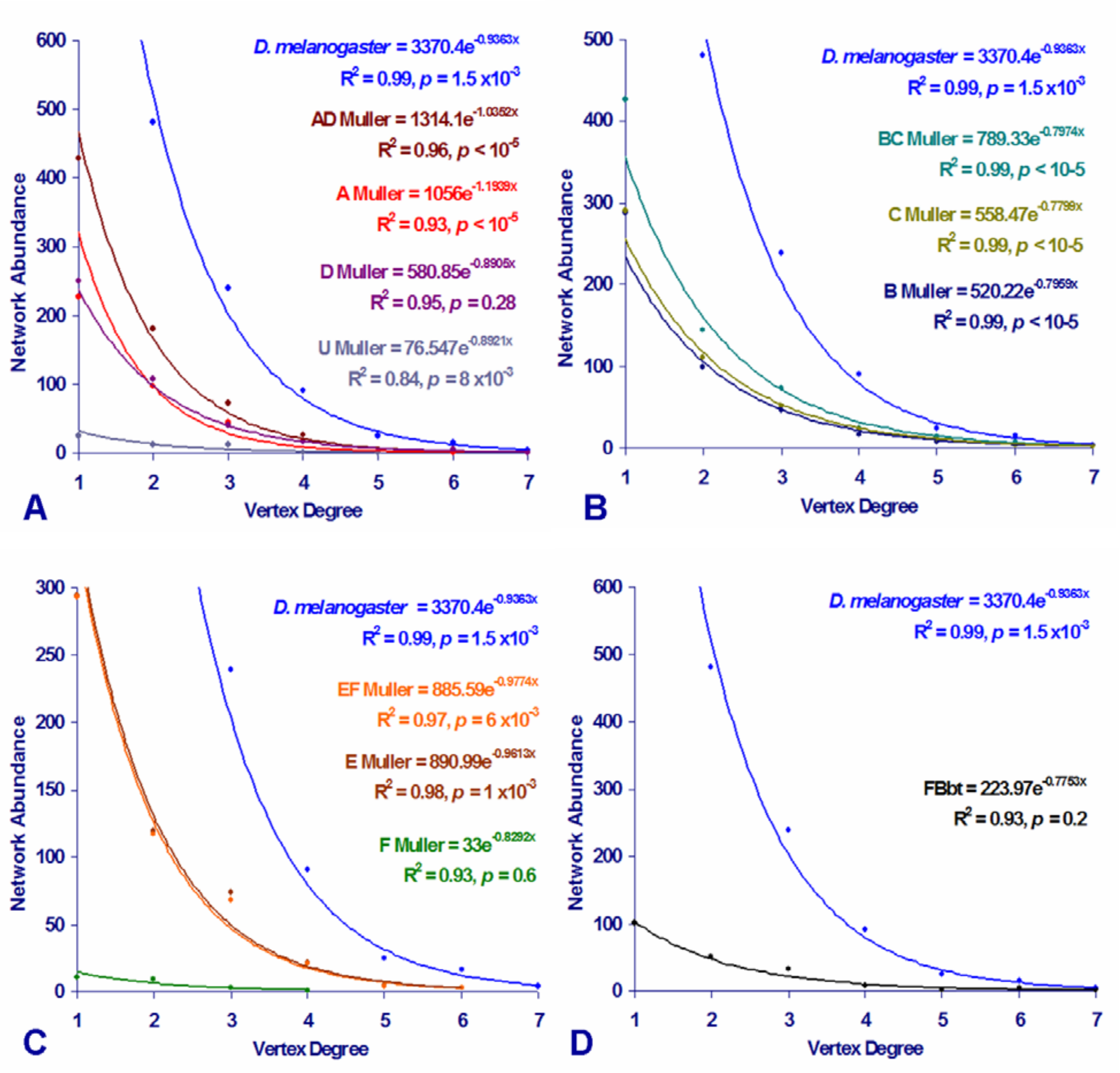


FIGURE 26. Vertex degree distribution and network abundance of microRNAs per target gene for Muller element and FBbt data of *Drosophila melanogaster*. Networks were defined using target prediction data derived from an intersection of MiRanda and TargetScan. Power-law function trend lines and data points are colored according to subnetwork of interest. All data are unbinned. The red trend line and data for *Drosophila* are equivalent to the network intersection of *D. melanogaster* illustrated in CHAPTER I, FIGURE 2C.

(A) Select Muller elements A, D, AD, and U (chromosomes X, 3L, the union of X-3L, and Unknown respectively) recovered trend lines with functions, non-linear regressions and statistical support as follows:

Drosophila) y 3370.4e $^{-0.9363x}$, $R^2 = 0.99$, $p = 1.5 \text{ x}10^{-3}$;

A Muller) $y = 1056e^{-1.1939x}$, $R^2 = 0.93$, $p < 10^{-5}$;

AD Muller) $y = 1314.1e^{-1.0352x}$, $R^2 = 0.96$, $p < 10\text{-}^{5}$;

D Muller) $y = 580.85e^{-0.8905x}$, $R^2 = 0.95$, $p = 0.28$;

U Muller) $y = 76.547e^{-0.8921x}$, $R^2 = 0.84$, $p = 8 \text{ x}10^{-3}$.

(B) Select Muller elements B, C, and BC (chromosomes 2L, 2R and 2, respectively) recovered power-law trend lines with functions and non-linear regressions as follows:

B Muller) $y = 520.22e^{-0.7959x}$, $R^2 = 0.99$, $p < 10^{-5}$;

BC Muller) $y = 789.33e^{-0.7974x}$, $R^2 = 0.99$, $p < 10^{-5}$;

C Muller) $y = 558.47e^{-0.7799x}$, $R^2 = 0.99$, $p < 10^{-5}$.

(C) Select Muller elements E, F, and the union of E and F (chromosomes 3R, 4, the union of 4-3R, respectively) recovered power-law trend lines with functions and non-linear regressions as follows:

E Muller) $y = 890.99e^{-0.9613x}$, $R^2 = 0.98$, $p = 1 \text{ x}10^{-3}$;

EF Muller) $y = 885.59e^{-0.9774x}$, $R^2 = 0.97$, $p = 6 \text{ x}10^{-3}$;

F Muller) $y = 33e^{-0.8292x}$, $R^2 = 0.93$, $p = 0.6$.

(D) The **FBbt** dataset recovered a power-law trend line function and non-linear regression of

$y = 223.97e^{-0.7753x}$, $R^2 = 0.93$, $p = 0.2$.

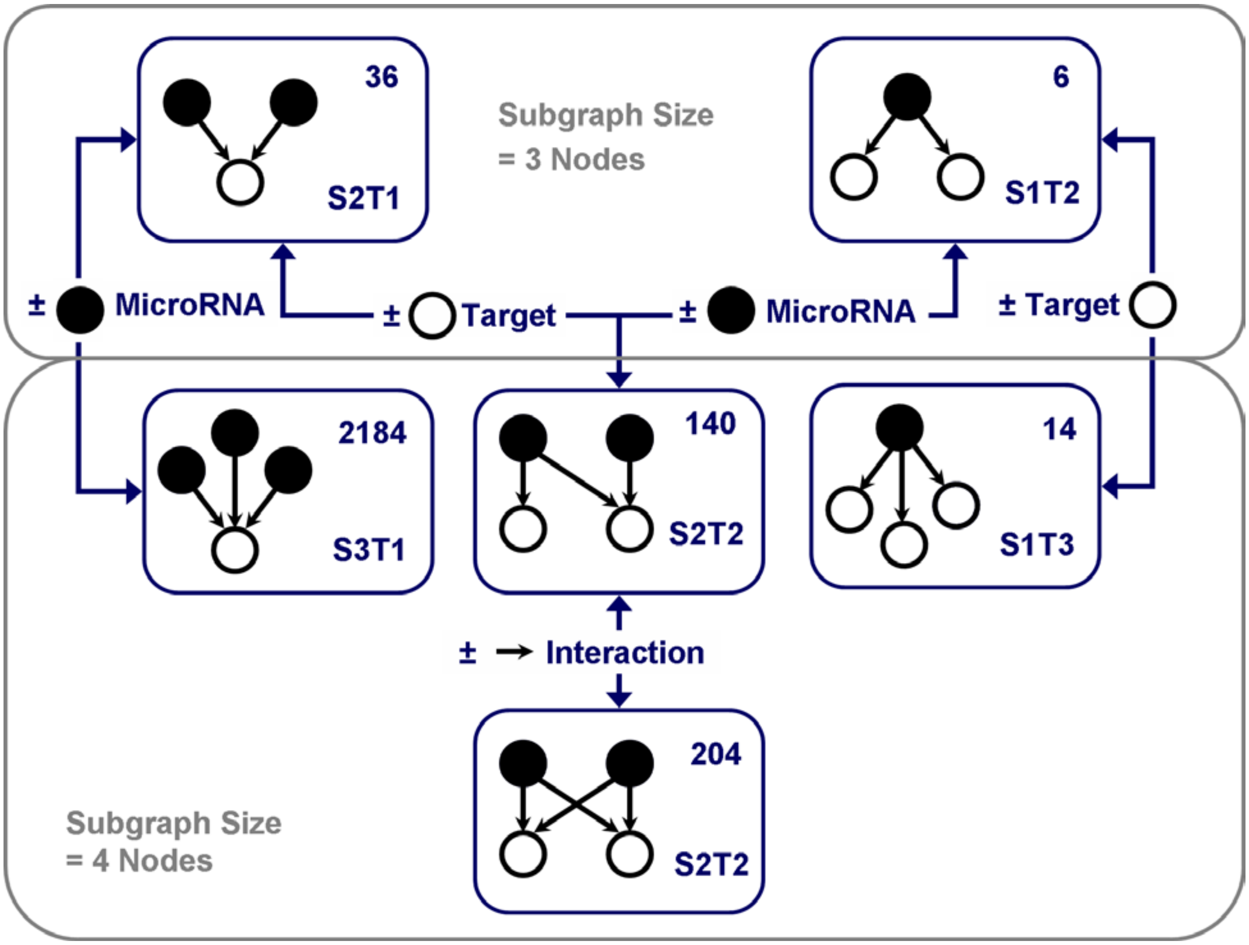


FIGURE 27. Topologies and conversion pathways for motifs observed in *Drosophila* microRNA-target interaction networks. MicroRNA interaction networks were strictly bipartite and only these motifs were present for network subgraphs composed of three and four nodes. Each motif topology is drawn to reflect the two nodal types; where black circles illustrate microRNAs and white circles represent targets. Motif adjective identifications and start-to-termination flow descriptors are reported in the upper and lower right corner of each respective motif box. Connectors diagram motif conversion pathways; specifically the loss or acquisition of a node type or interaction necessary for motif-to-motif topology conversion.

S2T1 motif-36 & S1T2 motif-6. Significant scores were recovered for the S1T2 motif-6 in all twelve species sampled for all microRNA networks analyzed (TABLE 2). The S1T2 motif-6 was typically in the representative majority for all size 3 subgraphs in all microRNA networks. Significant scores were also recovered for the S2T1 motif-36 across all twelve species sampled for all microRNA networks analyzed (TABLE 3). While the S2T1 motif-6 was typically in the representative minority for all size 3 subgraphs in all microRNA networks, scores indicate enrichment of this motif type. Negligible z-scores ranging from -1 to 1 were recovered for motifs of subgraph size 3 and thus three node cases are not particularly informative for networks considered in this study.

S1T3 motif-14. Addition of a target node produces the S1T3 motif-14 from the S1T2 motif-6 (FIGURE 27). Representation of the S1T3 motif-14 was variable by species and dataset (TABLE 4). The S1T3 motif-14 recovered significant network presence in twelve species sampled for all chromosome regions elements excluding Muller element F. Three species in the later dataset recovered significant S3T1 motif-2184 presence. The unsorted genes in the U-Muller element had seven species with significant scores for S3T1 motif-2184 while eleven species had significant S1T3 motif-2184 representation in the FBbt dataset network.

S3T1 motif-2184. The S3T1 motif-2184 is acquired only from the S2T1 motif-36 with the addition of a microRNA regulator (FIGURE 27). Representation of the S3T1 motif-2184 varied by species and dataset (TABLE 5). The S3T1 motif-2184 recovered significant network presence in twelve species sampled for all chromosome regions elements excluding Muller element F. The later dataset recovered significant S3T1 motif-2184 presence in three species. The unsorted genes in the U-Muller element had seven species with significant scores for S3T1 motif-2184 while eleven species had significant S3T1 motif-2184 representation in the FBbt dataset network.

S2T2 motif-140. The S2T2 motif-140 is acquired from either the S2T1 motif 36 or the S1T2 motif 6 respectively with the addition of either a target node or a microRNA regulator (FIGURE 27). The S2T2 motif-140 presented no significant score for any species across all microRNA networks analyzed (TABLE 6). Thus while subgraphs S2T2 was present in networks at variable frequency, their representation was not significantly enriched.

S2T2 motif-204. The S2T2 motif-204 or “bi-fan” can only be acquired from the S2T2 motif-140 with the addition of an interaction between a microRNA regulator and a target node (FIGURE 27). Great diversity is observed across datasets and species for the representation of the S2T2 motif-204 (TABLE 7). Among all motif observed the S2T2 motif-204 has relatively highest degree of enrichment with greatest average Z-scores; but in terms of absolute values the frequency of these motifs is low. The S2T2 motif-240 recovered significant network presence in all twelve species sampled for Muller elements A, B, D, E and EF. Significant network presence of the S2T2 motif-240 for ten species was recovered in Muller elements C and EF. The number of species recovering significant network representation of the S2T2 motif-240 declined from twelve to four when chromosome regions were combined in the Muller elements AD and BC. The unsorted genes in the U-Muller element had only five species with significant scores for S2T2 motif-240, while only one species had significant S2T2 motif-240 representation in the FBbt dataset network.

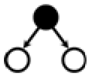
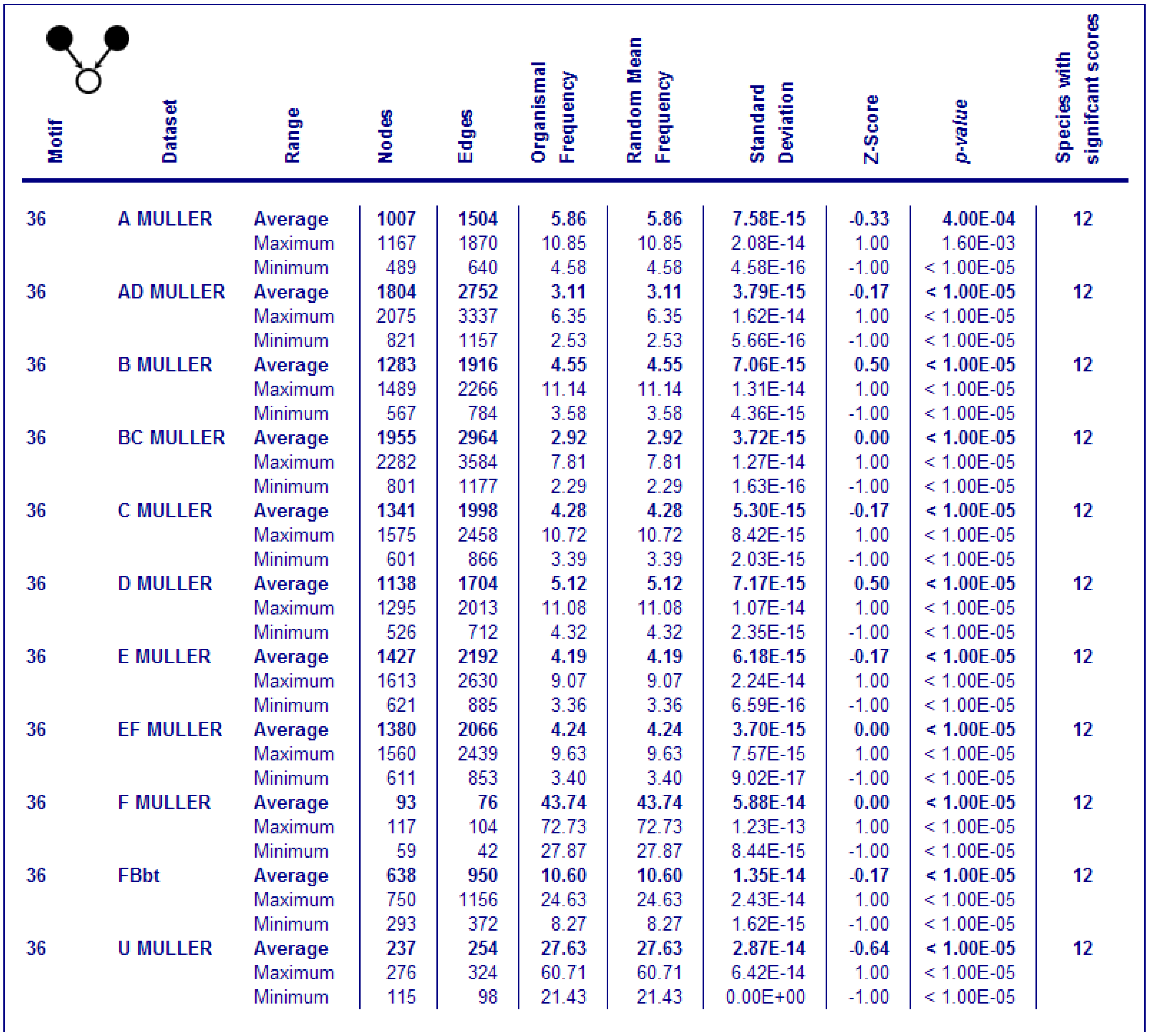
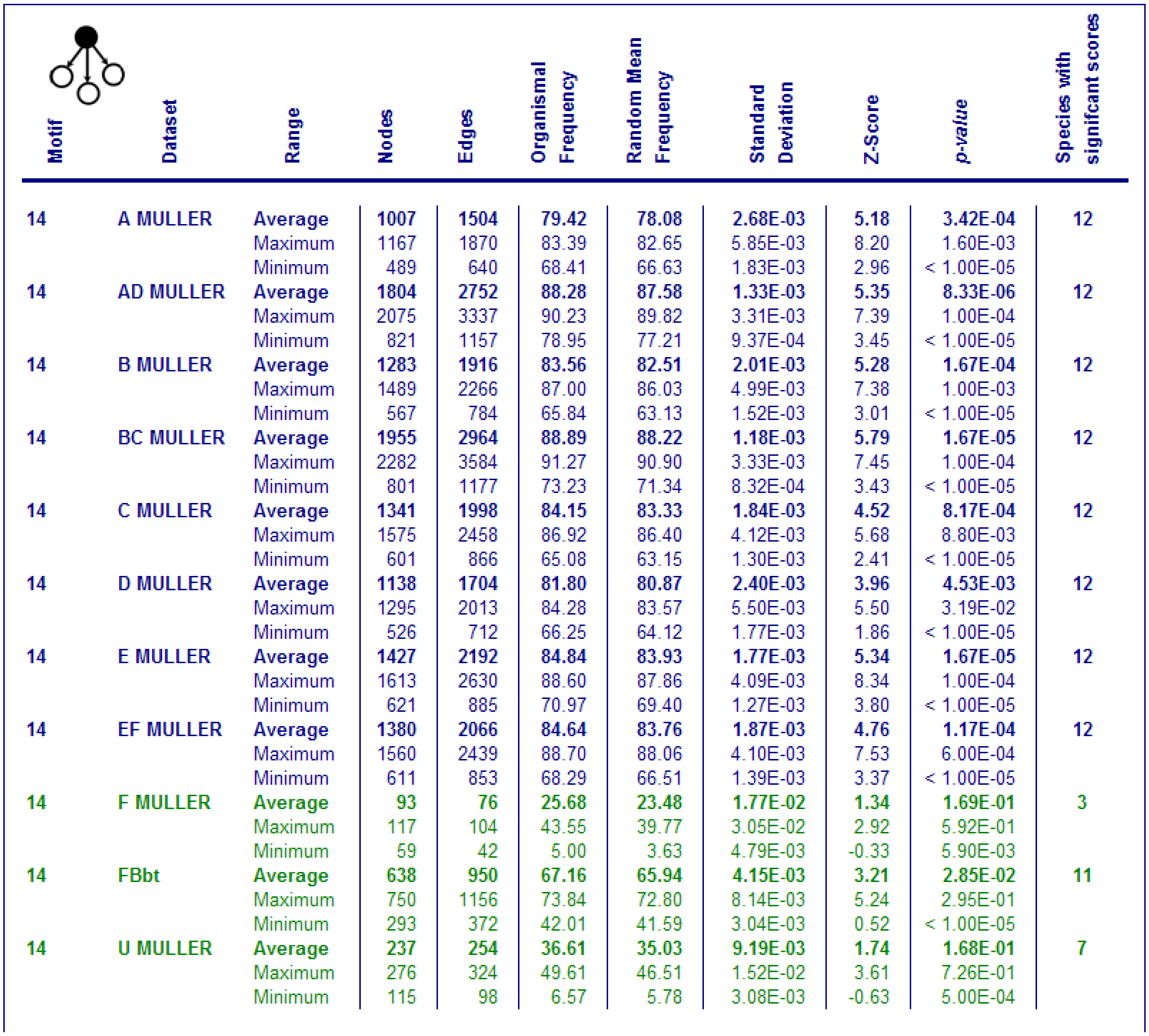
 Motif	Dataset	Range	Nodes	Edges	Organismal Frequency	Random Mean Frequency	Standard Deviation	Z-Score	*p-value*	Species with signifcant scores
6	A MULLER	**Average**	**1007**	**1504**	**94.14**	**94.14**	**1.02E-13**	**0.00**	**< 1.00E-05**	12
		Maximum	1167	1870	95.42	95.42	2.26E-13	1.00	< 1.00E-05	
		Minimum	489	640	89.15	89.15	2.28E-14	-1.00	< 1.00E-05	
6	AD MULLER	**Average**	**1804**	**2752**	**96.89**	**96.89**	**1.18E-13**	**0.00**	**< 1.00E-05**	12
		Maximum	2075	3337	97.47	97.47	2.40E-13	1.00	< 1.00E-05	
		Minimum	821	1157	93.65	93.65	3.33E-14	-1.00	< 1.00E-05	
6	B MULLER	**Average**	**1283**	**1916**	**95.45**	**95.45**	**7.00E-14**	**0.00**	**< 1.00E-05**	12
		Maximum	1489	2266	96.42	96.42	2.13E-13	1.00	< 1.00E-05	
		Minimum	567	784	88.86	88.86	2.55E-15	-1.00	< 1.00E-05	
6	BC MULLER	**Average**	**1955**	**2964**	**97.08**	**97.08**	**6.06E-14**	**-0.17**	**< 1.00E-05**	12
		Maximum	2282	3584	97.71	97.71	2.11E-13	1.00	< 1.00E-05	
		Minimum	801	1177	92.19	92.19	1.11E-16	-1.00	< 1.00E-05	
6	C MULLER	**Average**	**1341**	**1998**	**95.72**	**95.72**	**8.94E-14**	**0.00**	**< 1.00E-05**	12
		Maximum	1575	2458	96.61	96.61	2.42E-13	1.00	< 1.00E-05	
		Minimum	601	866	89.28	89.28	2.33E-15	-1.00	< 1.00E-05	
6	D MULLER	**Average**	**1138**	**1704**	**94.88**	**94.88**	**1.12E-13**	**-0.17**	**< 1.00E-05**	12
		Maximum	1295	2013	95.68	95.68	2.27E-13	1.00	< 1.00E-05	
		Minimum	526	712	88.92	88.92	2.10E-14	-1.00	< 1.00E-05	
6	E MULLER	**Average**	**1427**	**2192**	**95.81**	**95.81**	**7.61E-14**	**-0.50**	**< 1.00E-05**	12
		Maximum	1613	2630	96.64	96.64	2.25E-13	1.00	< 1.00E-05	
		Minimum	621	885	90.93	90.93	9.33E-15	-1.00	< 1.00E-05	
6	EF MULLER	**Average**	**1380**	**2066**	**95.76**	**95.76**	**8.16E-14**	**0.00**	**< 1.00E-05**	12
		Maximum	1560	2439	96.60	96.60	1.75E-13	1.00	< 1.00E-05	
		Minimum	611	853	90.37	90.37	6.66E-15	-1.00	< 1.00E-05	
6	F MULLER	**Average**	**93**	**76**	**56.26**	**56.26**	**7.25E-14**	**-0.17**	**< 1.00E-05**	12
		Maximum	117	104	72.13	72.13	1.29E-13	1.00	< 1.00E-05	
		Minimum	59	42	27.27	27.27	5.33E-15	-1.00	< 1.00E-05	
6	FBbt	**Average**	**638**	**950**	**89.40**	**89.40**	**8.41E-14**	**-0.17**	**< 1.00E-05**	12
		Maximum	750	1156	91.73	91.73	2.23E-13	1.00	< 1.00E-05	
		Minimum	293	372	75.37	75.37	2.55E-15	-1.00	< 1.00E-05	
6	U MULLER	**Average**	**237**	**254**	**72.38**	**72.38**	**8.35E-14**	**-0.64**	**< 1.00E-05**	12
		Maximum	276	324	78.57	78.57	1.99E-13	1.00	< 1.00E-05	
		Minimum	115	98	39.29	39.29	0.00E+00	-1.00	< 1.00E-05	

TABLE 2. MicroRNA regulatory network topology statistics recovered from FANMOD for the S1T2 motif-6 using target prediction data derived from an intersection of MiRanda and TargetScan. Rows correspond to the average, maximum, and minimum scores recovered from select subnetworks of interest (dataset) across a comparision of twelve *Drosophila* species. Rows are color coded according to the numbers of species in the dataset recovering significant *P*-values. Datasets recovering significant scores for all species sampled are colored in **dark blue**.



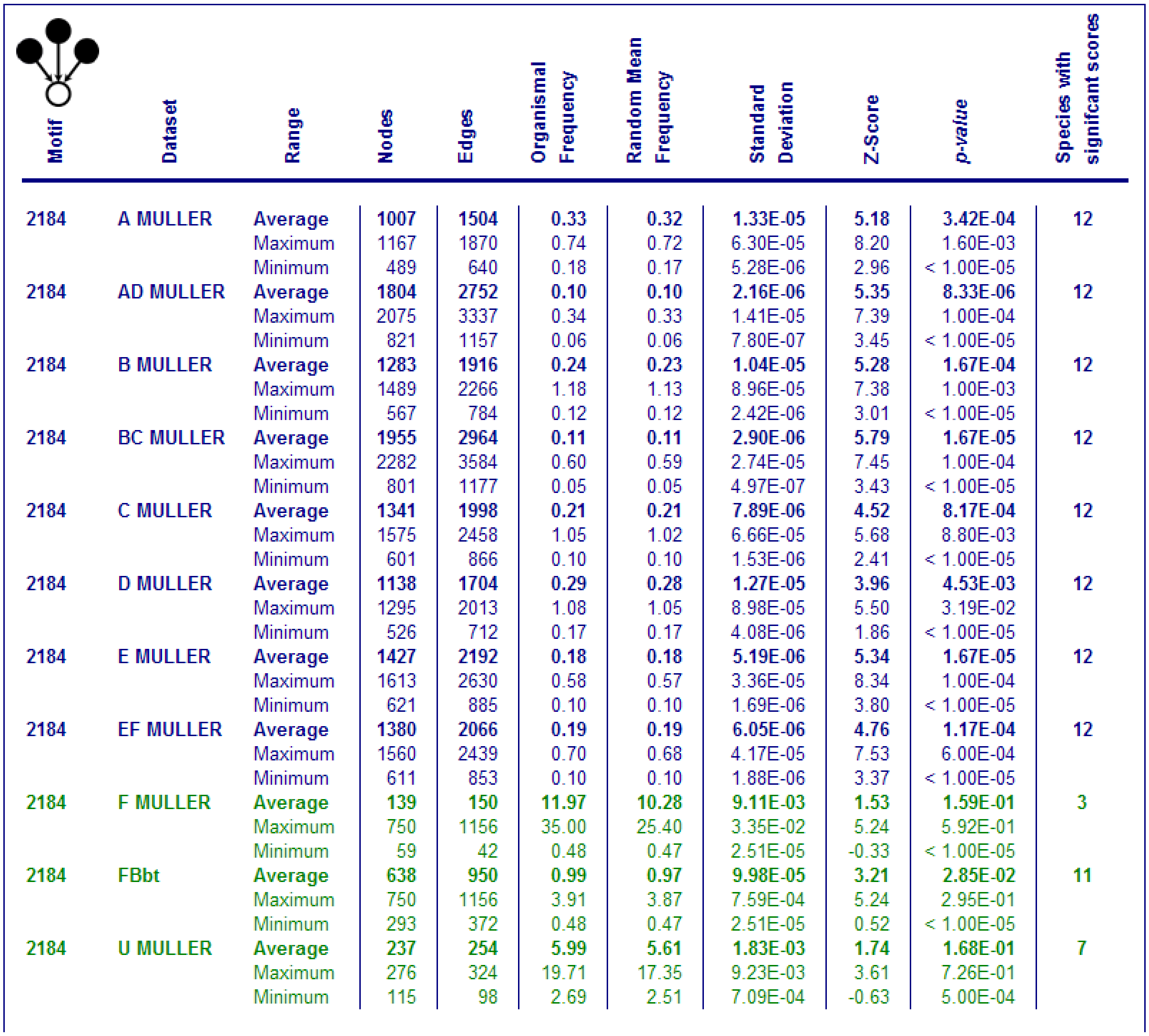
Motif	Dataset	Range	Nodes	Edges	Organismal Frequency	Random Mean Frequency	Standard Deviation	Z-Score	*p-value*	Species with signifcant scores
36	**A MULLER**	**Average**	**1007**	**1504**	**5.86**	**5.86**	**7.58E-15**	**-0.33**	**4.00E-04**	**12**
		Maximum	1167	1870	10.85	10.85	2.08E-14	1.00	1.60E-03	
		Minimum	489	640	4.58	4.58	4.58E-16	-1.00	< 1.00E-05	
36	**AD MULLER**	**Average**	**1804**	**2752**	**3.11**	**3.11**	**3.79E-15**	**-0.17**	**< 1.00E-05**	**12**
		Maximum	2075	3337	6.35	6.35	1.62E-14	1.00	< 1.00E-05	
		Minimum	821	1157	2.53	2.53	5.66E-16	-1.00	< 1.00E-05	
36	**B MULLER**	**Average**	**1283**	**1916**	**4.55**	**4.55**	**7.06E-15**	**0.50**	**< 1.00E-05**	**12**
		Maximum	1489	2266	11.14	11.14	1.31E-14	1.00	< 1.00E-05	
		Minimum	567	784	3.58	3.58	4.36E-15	-1.00	< 1.00E-05	
36	**BC MULLER**	**Average**	**1955**	**2964**	**2.92**	**2.92**	**3.72E-15**	**0.00**	**< 1.00E-05**	**12**
		Maximum	2282	3584	7.81	7.81	1.27E-14	1.00	< 1.00E-05	
		Minimum	801	1177	2.29	2.29	1.63E-16	-1.00	< 1.00E-05	
36	**C MULLER**	**Average**	**1341**	**1998**	**4.28**	**4.28**	**5.30E-15**	**-0.17**	**< 1.00E-05**	**12**
		Maximum	1575	2458	10.72	10.72	8.42E-15	1.00	< 1.00E-05	
		Minimum	601	866	3.39	3.39	2.03E-15	-1.00	< 1.00E-05	
36	**D MULLER**	**Average**	**1138**	**1704**	**5.12**	**5.12**	**7.17E-15**	**0.50**	**< 1.00E-05**	**12**
		Maximum	1295	2013	11.08	11.08	1.07E-14	1.00	< 1.00E-05	
		Minimum	526	712	4.32	4.32	2.35E-15	-1.00	< 1.00E-05	
36	**E MULLER**	**Average**	**1427**	**2192**	**4.19**	**4.19**	**6.18E-15**	**-0.17**	**< 1.00E-05**	**12**
		Maximum	1613	2630	9.07	9.07	2.24E-14	1.00	< 1.00E-05	
		Minimum	621	885	3.36	3.36	6.59E-16	-1.00	< 1.00E-05	
36	**EF MULLER**	**Average**	**1380**	**2066**	**4.24**	**4.24**	**3.70E-15**	**0.00**	**< 1.00E-05**	**12**
		Maximum	1560	2439	9.63	9.63	7.57E-15	1.00	< 1.00E-05	
		Minimum	611	853	3.40	3.40	9.02E-17	-1.00	< 1.00E-05	
36	**F MULLER**	**Average**	**93**	**76**	**43.74**	**43.74**	**5.88E-14**	**0.00**	**< 1.00E-05**	**12**
		Maximum	117	104	72.73	72.73	1.23E-13	1.00	< 1.00E-05	
		Minimum	59	42	27.87	27.87	8.44E-15	-1.00	< 1.00E-05	
36	**FBbt**	**Average**	**638**	**950**	**10.60**	**10.60**	**1.35E-14**	**-0.17**	**< 1.00E-05**	**12**
		Maximum	750	1156	24.63	24.63	2.43E-14	1.00	< 1.00E-05	
		Minimum	293	372	8.27	8.27	1.62E-15	-1.00	< 1.00E-05	
36	**U MULLER**	**Average**	**237**	**254**	**27.63**	**27.63**	**2.87E-14**	**-0.64**	**< 1.00E-05**	**12**
		Maximum	276	324	60.71	60.71	6.42E-14	1.00	< 1.00E-05	
		Minimum	115	98	21.43	21.43	0.00E+00	-1.00	< 1.00E-05	

TABLE 3. MicroRNA regulatory network topology statistics recovered from FANMOD for the S2T1 motif-36 using target prediction data derived from an intersection of MiRanda and TargetScan. Rows correspond to the average, maximum, and minimum scores recovered from select subnetworks of interest (dataset) across a comparision of twelve *Drosophila* species. Rows are color coded according to the numbers of species in the dataset recovering significant *P*-values. Datasets recovering significant scores for all species sampled are colored in **dark blue**.



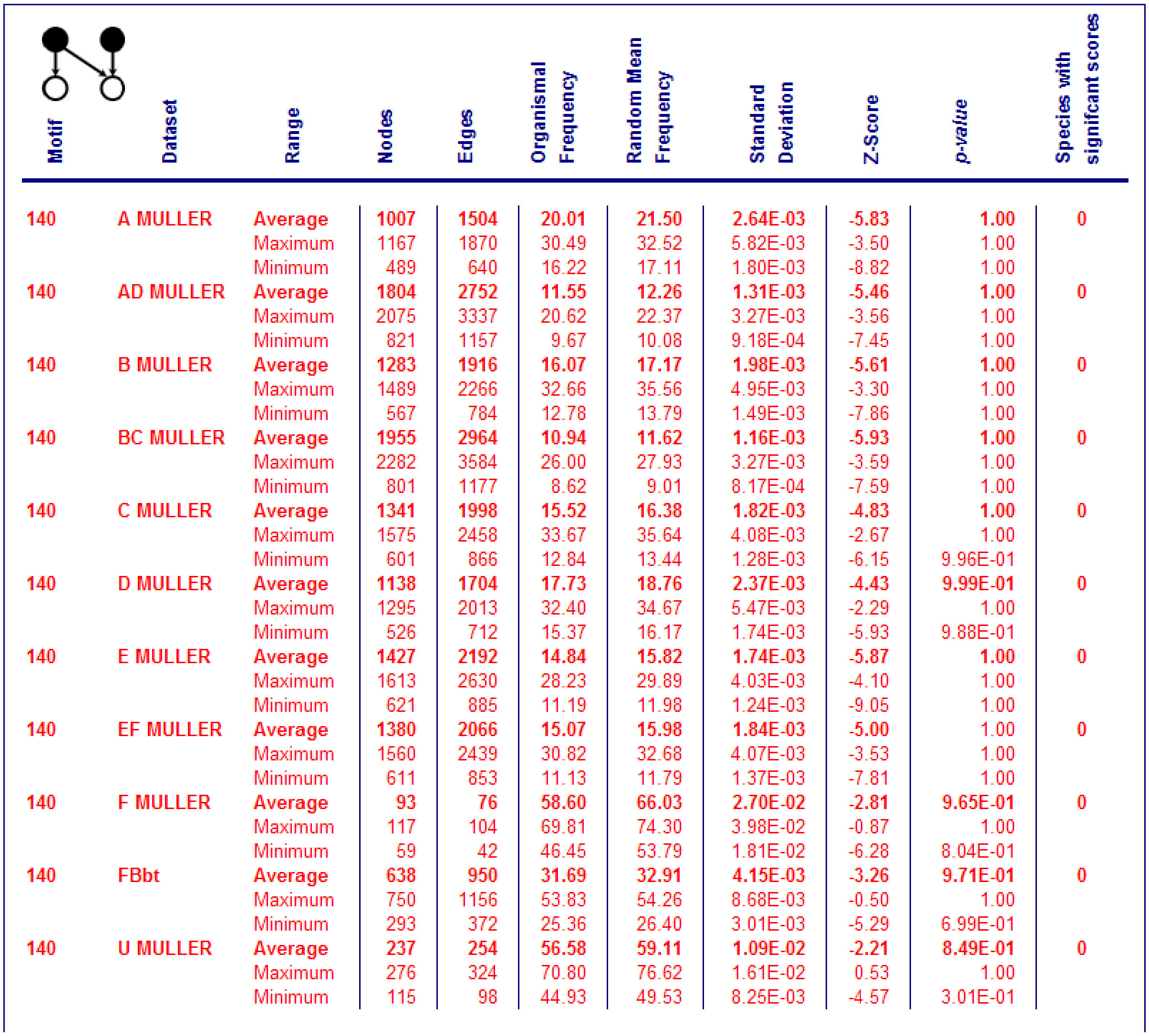
Motif	Dataset	Range	Nodes	Edges	Organismal Frequency	Random Mean Frequency	Standard Deviation	Z-Score	*p-value*	Species with signifcant scores
14	**A MULLER**	**Average**	**1007**	**1504**	**79.42**	**78.08**	**2.68E-03**	**5.18**	**3.42E-04**	**12**
		Maximum	1167	1870	83.39	82.65	5.85E-03	8.20	1.60E-03	
		Minimum	489	640	68.41	66.63	1.83E-03	2.96	< 1.00E-05	
14	**AD MULLER**	**Average**	**1804**	**2752**	**88.28**	**87.58**	**1.33E-03**	**5.35**	**8.33E-06**	**12**
		Maximum	2075	3337	90.23	89.82	3.31E-03	7.39	1.00E-04	
		Minimum	821	1157	78.95	77.21	9.37E-04	3.45	< 1.00E-05	
14	**B MULLER**	**Average**	**1283**	**1916**	**83.56**	**82.51**	**2.01E-03**	**5.28**	**1.67E-04**	**12**
		Maximum	1489	2266	87.00	86.03	4.99E-03	7.38	1.00E-03	
		Minimum	567	784	65.84	63.13	1.52E-03	3.01	< 1.00E-05	
14	**BC MULLER**	**Average**	**1955**	**2964**	**88.89**	**88.22**	**1.18E-03**	**5.79**	**1.67E-05**	**12**
		Maximum	2282	3584	91.27	90.90	3.33E-03	7.45	1.00E-04	
		Minimum	801	1177	73.23	71.34	8.32E-04	3.43	< 1.00E-05	
14	**C MULLER**	**Average**	**1341**	**1998**	**84.15**	**83.33**	**1.84E-03**	**4.52**	**8.17E-04**	**12**
		Maximum	1575	2458	86.92	86.40	4.12E-03	5.68	8.80E-03	
		Minimum	601	866	65.08	63.15	1.30E-03	2.41	< 1.00E-05	
14	**D MULLER**	**Average**	**1138**	**1704**	**81.80**	**80.87**	**2.40E-03**	**3.96**	**4.53E-03**	**12**
		Maximum	1295	2013	84.28	83.57	5.50E-03	5.50	3.19E-02	
		Minimum	526	712	66.25	64.12	1.77E-03	1.86	< 1.00E-05	
14	**E MULLER**	**Average**	**1427**	**2192**	**84.84**	**83.93**	**1.77E-03**	**5.34**	**1.67E-05**	**12**
		Maximum	1613	2630	88.60	87.86	4.09E-03	8.34	1.00E-04	
		Minimum	621	885	70.97	69.40	1.27E-03	3.80	< 1.00E-05	
14	**EF MULLER**	**Average**	**1380**	**2066**	**84.64**	**83.76**	**1.87E-03**	**4.76**	**1.17E-04**	**12**
		Maximum	1560	2439	88.70	88.06	4.10E-03	7.53	6.00E-04	
		Minimum	611	853	68.29	66.51	1.39E-03	3.37	< 1.00E-05	
14	**F MULLER**	**Average**	**93**	**76**	**25.68**	**23.48**	**1.77E-02**	**1.34**	**1.69E-01**	**3**
		Maximum	117	104	43.55	39.77	3.05E-02	2.92	5.92E-01	
		Minimum	59	42	5.00	3.63	4.79E-03	-0.33	5.90E-03	
14	**FBbt**	**Average**	**638**	**950**	**67.16**	**65.94**	**4.15E-03**	**3.21**	**2.85E-02**	**11**
		Maximum	750	1156	73.84	72.80	8.14E-03	5.24	2.95E-01	
		Minimum	293	372	42.01	41.59	3.04E-03	0.52	< 1.00E-05	
14	**U MULLER**	**Average**	**237**	**254**	**36.61**	**35.03**	**9.19E-03**	**1.74**	**1.68E-01**	**7**
		Maximum	276	324	49.61	46.51	1.52E-02	3.61	7.26E-01	
		Minimum	115	98	6.57	5.78	3.08E-03	-0.63	5.00E-04	

TABLE 4. MicroRNA regulatory network topology statistics recovered from FANMOD for the S1T3 motif-14 using target prediction data derived from an intersection of MiRanda and TargetScan. Rows correspond to the average, maximum, and minimum scores recovered from select subnetworks of interest (dataset) across a comparision of twelve *Drosophila* species. Rows are color coded according to the numbers of species in the dataset recovering significant *P*-values. Datasets recovering significant scores for all species sampled are colored in **dark blue**. **Green** rows have mixed representation of significant and non-significant scores among species.



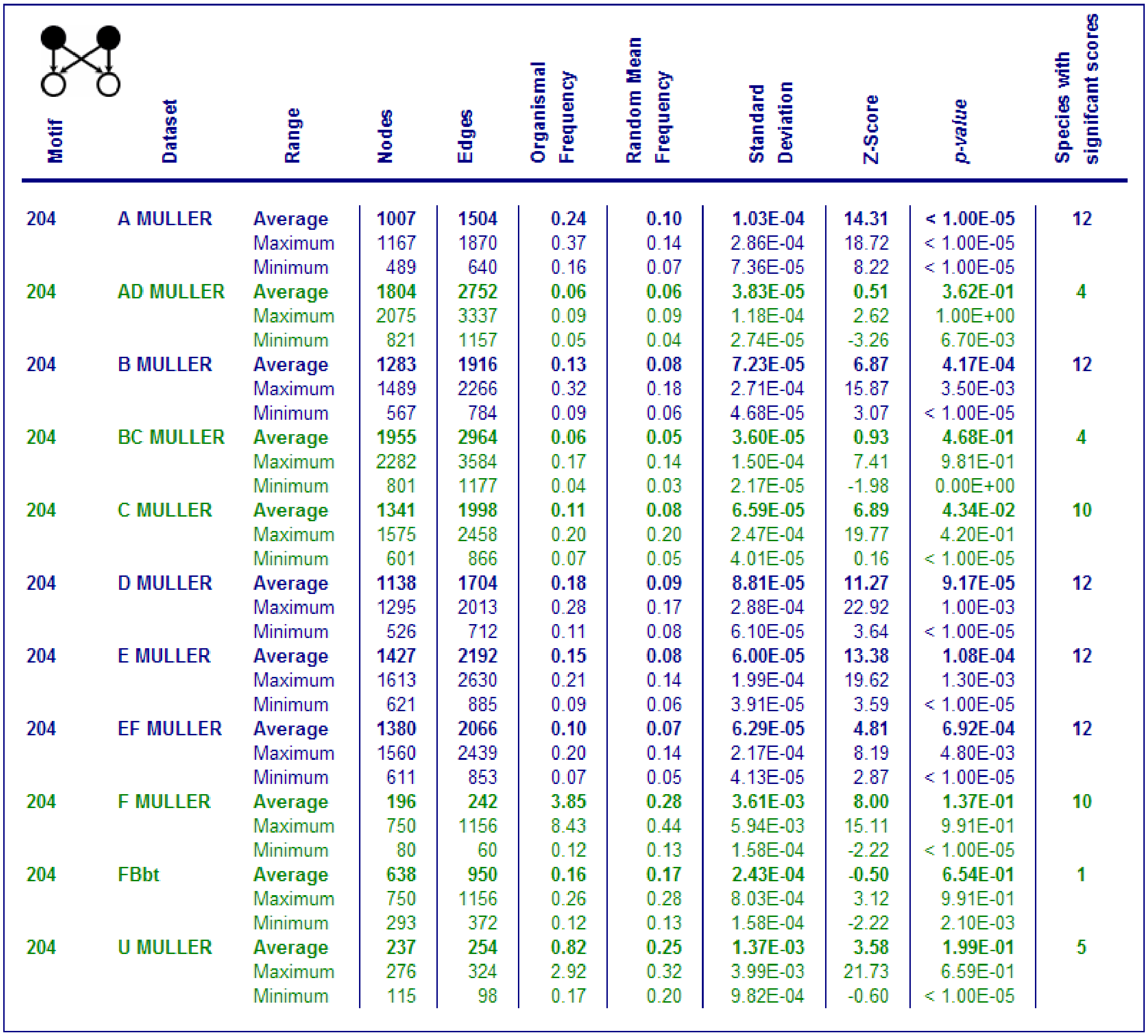
Motif	Dataset	Range	Nodes	Edges	Organismal Frequency	Random Mean Frequency	Standard Deviation	Z-Score	*p-value*	Species with significant scores
2184	A MULLER	Average	1007	1504	0.33	0.32	1.33E-05	5.18	3.42E-04	12
		Maximum	1167	1870	0.74	0.72	6.30E-05	8.20	1.60E-03	
		Minimum	489	640	0.18	0.17	5.28E-06	2.96	< 1.00E-05	
2184	AD MULLER	Average	1804	2752	0.10	0.10	2.16E-06	5.35	8.33E-06	12
		Maximum	2075	3337	0.34	0.33	1.41E-05	7.39	1.00E-04	
		Minimum	821	1157	0.06	0.06	7.80E-07	3.45	< 1.00E-05	
2184	B MULLER	Average	1283	1916	0.24	0.23	1.04E-05	5.28	1.67E-04	12
		Maximum	1489	2266	1.18	1.13	8.96E-05	7.38	1.00E-03	
		Minimum	567	784	0.12	0.12	2.42E-06	3.01	< 1.00E-05	
2184	BC MULLER	Average	1955	2964	0.11	0.11	2.90E-06	5.79	1.67E-05	12
		Maximum	2282	3584	0.60	0.59	2.74E-05	7.45	1.00E-04	
		Minimum	801	1177	0.05	0.05	4.97E-07	3.43	< 1.00E-05	
2184	C MULLER	Average	1341	1998	0.21	0.21	7.89E-06	4.52	8.17E-04	12
		Maximum	1575	2458	1.05	1.02	6.66E-05	5.68	8.80E-03	
		Minimum	601	866	0.10	0.10	1.53E-06	2.41	< 1.00E-05	
2184	D MULLER	Average	1138	1704	0.29	0.28	1.27E-05	3.96	4.53E-03	12
		Maximum	1295	2013	1.08	1.05	8.98E-05	5.50	3.19E-02	
		Minimum	526	712	0.17	0.17	4.08E-06	1.86	< 1.00E-05	
2184	E MULLER	Average	1427	2192	0.18	0.18	5.19E-06	5.34	1.67E-05	12
		Maximum	1613	2630	0.58	0.57	3.36E-05	8.34	1.00E-04	
		Minimum	621	885	0.10	0.10	1.69E-06	3.80	< 1.00E-05	
2184	EF MULLER	Average	1380	2066	0.19	0.19	6.05E-06	4.76	1.17E-04	12
		Maximum	1560	2439	0.70	0.68	4.17E-05	7.53	6.00E-04	
		Minimum	611	853	0.10	0.10	1.88E-06	3.37	< 1.00E-05	
2184	F MULLER	Average	139	150	11.97	10.28	9.11E-03	1.53	1.59E-01	3
		Maximum	750	1156	35.00	25.40	3.35E-02	5.24	5.92E-01	
		Minimum	59	42	0.48	0.47	2.51E-05	-0.33	< 1.00E-05	
2184	FBbt	Average	638	950	0.99	0.97	9.98E-05	3.21	2.85E-02	11
		Maximum	750	1156	3.91	3.87	7.59E-04	5.24	2.95E-01	
		Minimum	293	372	0.48	0.47	2.51E-05	0.52	< 1.00E-05	
2184	U MULLER	Average	237	254	5.99	5.61	1.83E-03	1.74	1.68E-01	7
		Maximum	276	324	19.71	17.35	9.23E-03	3.61	7.26E-01	
		Minimum	115	98	2.69	2.51	7.09E-04	-0.63	5.00E-04	

TABLE 5. MicroRNA regulatory network topology statistics recovered from FANMOD for the S3T1 motif-2184 using target prediction data derived from an intersection of MiRanda and TargetScan. Rows correspond to the average, maximum, and minimum scores recovered from select subnetworks of interest (dataset) across a comparision of twelve *Drosophila* species. Rows are color coded according to the numbers of species in the dataset recovering significant p-values. Datasets recovering significant scores for all species sampled are colored in **dark blue**. **Green** rows have mixed representation of significant and non-significant scores among species.



Motif	Dataset	Range	Nodes	Edges	Organismal Frequency	Random Mean Frequency	Standard Deviation	Z-Score	*p-value*	Species with signifcant scores
140	A MULLER	Average	1007	1504	20.01	21.50	2.64E-03	-5.83	1.00	0
		Maximum	1167	1870	30.49	32.52	5.82E-03	-3.50	1.00	
		Minimum	489	640	16.22	17.11	1.80E-03	-8.82	1.00	
140	AD MULLER	Average	1804	2752	11.55	12.26	1.31E-03	-5.46	1.00	0
		Maximum	2075	3337	20.62	22.37	3.27E-03	-3.56	1.00	
		Minimum	821	1157	9.67	10.08	9.18E-04	-7.45	1.00	
140	B MULLER	Average	1283	1916	16.07	17.17	1.98E-03	-5.61	1.00	0
		Maximum	1489	2266	32.66	35.56	4.95E-03	-3.30	1.00	
		Minimum	567	784	12.78	13.79	1.49E-03	-7.86	1.00	
140	BC MULLER	Average	1955	2964	10.94	11.62	1.16E-03	-5.93	1.00	0
		Maximum	2282	3584	26.00	27.93	3.27E-03	-3.59	1.00	
		Minimum	801	1177	8.62	9.01	8.17E-04	-7.59	1.00	
140	C MULLER	Average	1341	1998	15.52	16.38	1.82E-03	-4.83	1.00	0
		Maximum	1575	2458	33.67	35.64	4.08E-03	-2.67	1.00	
		Minimum	601	866	12.84	13.44	1.28E-03	-6.15	9.96E-01	
140	D MULLER	Average	1138	1704	17.73	18.76	2.37E-03	-4.43	9.99E-01	0
		Maximum	1295	2013	32.40	34.67	5.47E-03	-2.29	1.00	
		Minimum	526	712	15.37	16.17	1.74E-03	-5.93	9.88E-01	
140	E MULLER	Average	1427	2192	14.84	15.82	1.74E-03	-5.87	1.00	0
		Maximum	1613	2630	28.23	29.89	4.03E-03	-4.10	1.00	
		Minimum	621	885	11.19	11.98	1.24E-03	-9.05	1.00	
140	EF MULLER	Average	1380	2066	15.07	15.98	1.84E-03	-5.00	1.00	0
		Maximum	1560	2439	30.82	32.68	4.07E-03	-3.53	1.00	
		Minimum	611	853	11.13	11.79	1.37E-03	-7.81	1.00	
140	F MULLER	Average	93	76	58.60	66.03	2.70E-02	-2.81	9.65E-01	0
		Maximum	117	104	69.81	74.30	3.98E-02	-0.87	1.00	
		Minimum	59	42	46.45	53.79	1.81E-02	-6.28	8.04E-01	
140	FBbt	Average	638	950	31.69	32.91	4.15E-03	-3.26	9.71E-01	0
		Maximum	750	1156	53.83	54.26	8.68E-03	-0.50	1.00	
		Minimum	293	372	25.36	26.40	3.01E-03	-5.29	6.99E-01	
140	U MULLER	Average	237	254	56.58	59.11	1.09E-02	-2.21	8.49E-01	0
		Maximum	276	324	70.80	76.62	1.61E-02	0.53	1.00	
		Minimum	115	98	44.93	49.53	8.25E-03	-4.57	3.01E-01	

TABLE 6. MicroRNA regulatory network topology statistics recovered from FANMOD for the S2T2 motif-140 using target prediction data derived from an intersection of MiRanda and TargetScan. Rows correspond to the average, maximum, and minimum scores recovered from select subnetworks of interest (dataset) across a comparision of twelve *Drosophila* species. Rows are color coded according to the numbers of species in the dataset recovering significant *P*-values. Rows for datasets with non-significant scores for all species sampled are colored in **red**.



Motif	Dataset	Range	Nodes	Edges	Organismal Frequency	Random Mean Frequency	Standard Deviation	Z-Score	*p-value*	Species with signifcant scores
204	A MULLER	Average	1007	1504	0.24	0.10	1.03E-04	14.31	< 1.00E-05	12
		Maximum	1167	1870	0.37	0.14	2.86E-04	18.72	< 1.00E-05	
		Minimum	489	640	0.16	0.07	7.36E-05	8.22	< 1.00E-05	
204	AD MULLER	Average	1804	2752	0.06	0.06	3.83E-05	0.51	3.62E-01	4
		Maximum	2075	3337	0.09	0.09	1.18E-04	2.62	1.00E+00	
		Minimum	821	1157	0.05	0.04	2.74E-05	-3.26	6.70E-03	
204	B MULLER	Average	1283	1916	0.13	0.08	7.23E-05	6.87	4.17E-04	12
		Maximum	1489	2266	0.32	0.18	2.71E-04	15.87	3.50E-03	
		Minimum	567	784	0.09	0.06	4.68E-05	3.07	< 1.00E-05	
204	BC MULLER	Average	1955	2964	0.06	0.05	3.60E-05	0.93	4.68E-01	4
		Maximum	2282	3584	0.17	0.14	1.50E-04	7.41	9.81E-01	
		Minimum	801	1177	0.04	0.03	2.17E-05	-1.98	0.00E+00	
204	C MULLER	Average	1341	1998	0.11	0.08	6.59E-05	6.89	4.34E-02	10
		Maximum	1575	2458	0.20	0.20	2.47E-04	19.77	4.20E-01	
		Minimum	601	866	0.07	0.05	4.01E-05	0.16	< 1.00E-05	
204	D MULLER	Average	1138	1704	0.18	0.09	8.81E-05	11.27	9.17E-05	12
		Maximum	1295	2013	0.28	0.17	2.88E-04	22.92	1.00E-03	
		Minimum	526	712	0.11	0.08	6.10E-05	3.64	< 1.00E-05	
204	E MULLER	Average	1427	2192	0.15	0.08	6.00E-05	13.38	1.08E-04	12
		Maximum	1613	2630	0.21	0.14	1.99E-04	19.62	1.30E-03	
		Minimum	621	885	0.09	0.06	3.91E-05	3.59	< 1.00E-05	
204	EF MULLER	Average	1380	2066	0.10	0.07	6.29E-05	4.81	6.92E-04	12
		Maximum	1560	2439	0.20	0.14	2.17E-04	8.19	4.80E-03	
		Minimum	611	853	0.07	0.05	4.13E-05	2.87	< 1.00E-05	
204	F MULLER	Average	196	242	3.85	0.28	3.61E-03	8.00	1.37E-01	10
		Maximum	750	1156	8.43	0.44	5.94E-03	15.11	9.91E-01	
		Minimum	80	60	0.12	0.13	1.58E-04	-2.22	< 1.00E-05	
204	FBbt	Average	638	950	0.16	0.17	2.43E-04	-0.50	6.54E-01	1
		Maximum	750	1156	0.26	0.28	8.03E-04	3.12	9.91E-01	
		Minimum	293	372	0.12	0.13	1.58E-04	-2.22	2.10E-03	
204	U MULLER	Average	237	254	0.82	0.25	1.37E-03	3.58	1.99E-01	5
		Maximum	276	324	2.92	0.32	3.99E-03	21.73	6.59E-01	
		Minimum	115	98	0.17	0.20	9.82E-04	-0.60	< 1.00E-05	

TABLE 7. MicroRNA regulatory network topology statistics recovered from FANMOD for the S2T2 motif-240 using target prediction data derived from an intersection of MiRanda and TargetScan. Rows correspond to the average, maximum, and minimum scores recovered from select subnetworks of interest (dataset) across a comparision of twelve *Drosophila* species. Rows are color coded according to the numbers of species in the dataset recovering significant *P*-values. Datasets recovering significant scores for all species sampled are colored in **dark blue**. **Green** rows have mixed representation of significant and non-significant scores among species.

DISCUSSION

Natural Selection across Network Topology. The findings presented in this study represent a novel intergration of microRNA regulatory network topology to chromsomal synteny and genes linked to species diagnostic phenotypes. Previous comprehensive research has been conducted using FANMOD where whole organismal networks of interacting metabolites were analyzed in parallel for 107 species from the database of Ma and Zheng and 251 species available from KEGG (Kyoto Encyclopedia of Genes; December 2007 release; Kanehisa *et al.*, 2008; Kanehisa *et al.*, 2000; Kanehisa *et al.*, 2006; Ma & Zeng, 2003). Additional analyses were also conducted using the 251 species data from KEGG for entire networks of interacting pathways. Most significantly, these analyses recovered evidence high statistical support for the over-representation of certain feed-forward and bi-parallel motifs (subgraphs) with 3 and 4 nodes (Bonchev *et al.*, 2009). These studies provide methodological demonstration and philosophical justification relevant to this study of drosophilid speciation using microRNA-regulatory networks. The motifs exhibiting considerable enrichment were those having a redundant link. Similarly, microRNA network recovered significant enrichment for the S2T2 motif possessing a redundant link (motif-204) in all twelve species sampled for many Muller elements (TABLE 7). Conversely, the S2T2 motif lacking a redundant link, motif-140, presented no significant score for any species across all microRNA networks analyzed (TABLE 6). Notably, the S2T2 bi-fan motif-204 is often represented in networks that perform information processing, even though they describe elements as different as biomolecules within a cell and synaptic connections between neurons (Milo *et al.*, 2002). In the context of adaptive significance, these results indicated that the need of higher network resilience against attacks not only compensates the energy price for the extra link formation but also exceeded the potential benefit of a faster performance (Bonchev *et al.*, 2009).

Representation of the S1T3 motif-14 was variable by species and dataset (TABLE 4), but respective S1T3 motifs enrichment might be expected to occur where there has been strong selection for broad regulation of suites of functionally interrelated targets. Alternatively the S1T3 motif-14 may be acquired from the S1T2 motif-6 with the addition of a target node as an artifact of paralogous gene duplication (FIGURE 27). Target gene duplication allows for an increase in phenotypic possibilities while using the same genetic repertoire by providing more variability of gene expression levels (Lee *et al.*, 2007). In either case a single microRNA could effectively exert simultaneous regulatory control in cases where motif participant targets are synchronously expressed *in vivo*. This regulatory approach would be most advantageous in circumstances where control is required to decisively shut down an entire biological pathway at several key points by enacting enzyme down-regulation at the translational level (He & Hannon, 2004; Stark *et al.*, 2003). Such regulatory strategies would likely play important roles in the generating tissue specific differentiation, and misregulation of microRNA expression itself could potentially have severe developmental consequences (Li *et al.*, 2006). Congruently, over-reaching biochemical control for enzymatic pathways has been well documented in *Drosophila* (He & Hannon, 2004; Stark *et al.*, 2003). Likewise, human microRNAs are evidenced to have exclusive GO term association by selectively targeting functionally distinct population of genes according to their transcript AT and GC content (Robins & Press, 2005).

Representation of the S3T1 motif-2184 varied by species and dataset (TABLE 5) Expansion of a microRNA repertoire through paralogy could enrich S3T1 subgraphs in some microRNA networks. Novel microRNAs acquired through duplication may be a simple mechanism for increasing microRNA cellular dosage (Lu *et al.*, 2008; Prochnik *et al.*, 2007; Stark *et al.*, 2007a). Sister microRNA genes would likely share ancestral target transcripts following their initial divergence. However, the later scenario is not the case for the selected networks considered in these analyses. These data represent network intersection of microRNA target prediction methods, and mature microRNA regulators are grouped according to distinct microRNA families. Therefore, all S3T1 type subgraphs observed would have arisen through target interactions involving a trio of unrelated microRNA families.

MicroRNAs acquired from novel families enable fine-tuning to particular target suites (Lu *et al.*, 2008; Prochnik *et al.*, 2007; Stark *et al.*, 2007a). Observed enrichment in S3T1 motifs might be expected to occur where there has been strong selection for coordinate microRNA control of a given target (Hornstein, & Shomron, 2006). The nature of the coordinated target control would depend upon the relative strengths of separate microRNA-target interactions. Temporal ontogeny further complicates matters. Given that microRNA-mediated targeting depends on the expression of both microRNA and targets, distinct microRNAs that have the same seed sequence may still have different target sets simply due to differences in their expression profile; even though they principally recognize the same set of target sites (Gaidatzis *et al.*, 2007). Current research favors a ‘‘restrictive model’’ of gene regulation, where the level of genome regulation steadily increases throughout development: the mRNA-regulating function of microRNAs suggests that this trend may be due to the effects of increased expression of microRNA genes (Olsen & Ambros, 1999; Strauss *et al.*, 2006).

In cases where motif participant microRNAs are not co-expressed *in vivo,* a common target could be separately regulated according to different tissue lines (Stark *et al.,* 2005). Alternatively, where multiple motif participant microRNAs are co-expressed, an increase in S3T1 motif abundance would imply increased selection for redundancy of control over a given target. The latter case could happen where different environmental circumstances produce a default regulatory reaction: different microRNAs could be cued to become operative and regulate a shared target under separate regulatory stimuli. Indeed, both life-stage tissue-specific expression and microRNA mediated regulation of metabolic responses to the environment have been observed in *Drosophila* (Aravin *et al.*, 2003; Lai *et al.*, 2003; Stark *et al.*, 2003). Additionally, where individual interactions were of variable strength (and individual regulatory interactions target could thus be ‘leaky’), redundant control using coexpression of multiple microRNAs may be necessary to keep the common target under biologically proper regulation (Stark *et al.*, 2005).

In the context of *Drosophila*, there is documented precedent for redundancy of control under co-expressed microRNAs (Lai *et al.*, 2003; Stark *et al.*, 2003). Enrichment of S3T1 motifs would be consistent with the design principle of canalization (Hornstein, & Shomron, 2006). Mathematical modeling shows that microRNA motifs in mammals may stabilize feedback loops to resist environmental perturbation and this provides one mechanism to explain the robust nature of microRNA controlled developmental programs (Yu *et al.*, 2008). Specifically, this mechanism has been suggested to contribute to the canalization of genetic programs; where canalized traits have an increased capacity to absorb mutational variance, resist radical change, and effectively maintain the phenotypic reproducibility of development (Hornstein & Shomron, 2006; Yu *et al.*, 2008). Canalized genotypes give rise to the same phenotype in different enviroments; presumably because the product (character state) of development is of adaptive significance. In a canalized state, genetic variations do not readily affect phenotype (and are therefore not subject to the same rigorous natural selection), but upon loss of canalization, genetic variations are uncovered. Thus, canalization not only contributes to developmental robustness, but potentially to adaptive innovations (Hornstein & Shomron, 2006).

Selection for coordinate control through enrichment of S3T1 motifs could have strong implications toward understanding molecular mechanisms underlying the acquisition of phenotypic plasticity (Bartel & Chen, 2004; Hornstein & Shomron, 2006). MicroRNAs most likely have a critical involvement in adaptive regulatory circuit extension; where organisms expand the functional portion of their genome as they also incorporate survival information about their niche (Lee *et al.*, 2007). Under this model, it is possible that recently acquired species-specific microRNAs would be most involved in fine-tuning gene expression to adapt organisms to different environments, rather than supporting more ancient developmental programs (Stark *et al.*, 2005). Given suggested correlations between microRNA acquisition and morphological innovation, it is here hypothesized that among sets of closely related organisms those select taxa exhibiting greatest phenotypic plasticity among would have the greatest S3T1 type microRNA regulatory network motif enrichment (Sempere *et al.*, 2006). Future network topology research using *Drosophila* should compare motif patterns for regulatory networks of 65 conserved microRNAs against species-specific thermal tolerance ranges and other quantifiable indicators of species phenotypic plasticity (Berezikov *et al.*, 2010; FIGURE 25).

MicroRNA Regulatory Networks for Regions of Chromosomal Synteny (Muller Elements). All microRNA regulatory networks for Muller elements in *Drosophila melanogaster* (excluding the F-element) were well-described by power law trend lines (R^2 = 0.82 to 0.95) with exponents of greater absolute value than the unsorted genes (FIGIRE 26). Moreover the power law behavior of these latter networks recovered an exponential constant “b” within (or approaching) the biological range of -2 to -3 (Barabasi & Albert, 1999). Thus, these user-defined Muller element regulatory networks would appear biologically relevant. Notably, the mixing of chromosome regions for Muller element AD produces many insignificant scores through FANMOD (TABLE 6). Alteration of S2T2 motif-240 enrichment patterns with the union of individual Muller elements (A and D vs. AD; B and C vs. BC) may be indicative that Muller elements operate as discrete entities within microRNA regulatory networks. Indeed, the high degree of enrichment observed in the A-Muller elements is lost with the combination of the D-Muller element. The repeating motif patterns across elements observed would not be expected if

Muller elements were not a natural subdivision of the total *Drosophila* network. Thus, the collective patterns observed indicate that respective Muller element networks have developed within the *Drosophila* transcriptome as separate regulatory modules.

Relationships between chromosome loci and microRNA regulation have been previously documented on a smaller scale. MicroRNA genes are frequently co-expressed along with their targets and these targets are often located within 50 kilobases (Baskerville & Bartel, 2005). Assuming this is an optimized genetic program, then this would in turn imply natural selection operative for genomic position. Future integration of karyotype loci to microRNA target data may shed some insights into the adaptive forces presiding over the frequent recapitulation of chromosome inversion groups observed across geographic clines (Hoffmann, 2004). In these cases, reoccurrence of chromosome inversion races regardless of lineage sorting is attributed to natural selection for sets of co-adapted gene complexes; although these complexes have yet to be identified (Hoffmann, 2004). Here it is proposed that microRNA genes are likely candidates involved in this co-adaptation. In the observed cases of this study there were 112 microRNA families found participant in every Muller element regulatory network; except element F with 99 microRNA families. Therefore any regulatory wiring as a discrete unit for Muller element targets seems to operate irrespective of the location of the microRNA regulator’s source chromosome locus. Future investigation will consider microRNA locus as a variable by limiting network topology analyses to microRNAs and targets of a single shared Muller element.

The results of this study for regions of major chromosome synteny have powerful implications toward the genetic basis of Haldane’s rule. This principle states that the heterogametic sex of an F1 cross of two different animal species is the most likely to be absent, rare, or sterile (Haldane, 1922). This rule applies to mammals, lepidopterans, birds, orthopterans and dipterans; thus in *Drosophila*, 142 documented interspecific hybridizations yield sterile XY males and fertile XX females (Coyne, 1985). One interpretation of Haldane’s rule proposes that X-chromosomes are tachytelic (exhibit a faster rate of change) over autosomes and in turn rapidly accumulate epistatic incompatibilities contributing to male-specific vs. female-specific sterility (Coyne, 1985). Results of FANMOD analyses which show a similar representations of regulatory motif enrichment for heterosomes (Muller element A) and autosomes (Muller elements B-E) disagree with the later fast-X model; at least at the functional level of regulatory network architecture (TABLES 2 to 6). Additionally, parametric scores from molecular phylogeny for Muller elements demonstrated no outstanding elevations of average Homoplasy Index relative to average Retention Index (APPENDIX III, TABLE 11). Comparable average gamma rate range (2.44-2.94) and average -*ln* Likelihood Score (30693.62 to 38035.50) were observed for (excluding Muller element F) under a general time reversible sequence evolution model with gamma rate variation and invariable sites (GTR + I + G). Excluding the F-Muller region, only the invariable site rate was slightly lower for the A-Muller element (0.18) over heterosome regions (0.20 to 0.21; APPENDIX III, TABLE 11). Perhaps this aberrant pattern is related to the innate accelerated genetic drift, lower codon usage, and low levels of recombination documented along the F chromosome element (Kliman & Hey, 1993). Nevertheless, if some divergence rate deviation occurred for heterosomes relative to the autosomes, then a deviation of parametric scores for A-Muller relative to the other elements would have been anticipated. Thus a fast-X hypothesis of Haldane's rule may also be contradicted at the level of nucleotide sequence evolution. Future research should follow the methodology of CHAPTER II and conduct direct microRNA network interspecific comparisons across separate Muller elements.

MicroRNA Regulatory Networks for Genes linked to species Diagnostic Phenotype in *Drosophila*. *Drosophila* species share a distinctive body plan and life cycle, but vary considerably in their morphology, ecology and behavior. The twelve sequenced species represent indigenes from Africa, Asia, Pacific Islands, and North and South America. There are cosmopolitan species that have colonized the entire planet (*Drosophila melanogaster* and *D. simulans*) and closely related species endemic to single islands (*D. sechellia*; *Drosophila* 12 Genomes Consortium, 2007). Drosophilid flies may be encountered living in deserts, in the tropics, on volcanic islands and, often as human commensals. A variety of behavioral strategies is also encompassed by the sequenced species; ranging from feeding generalist like as *Drosophila ananassae,* to species such as *D. sechellia,* specialized to feed on the fruit of a single plant species (*Drosophila* 12 Genomes Consortium, 2007). The interplay of gene regulatory systems to major morphological features in *Drosophila* species diagnosis is already well established (Gompel *et al.*, 2005; McGregor *et al.*, 2007; Sucena & Stern, 2000). Nevertheless, while the case for microRNA-mediated control over basic bilaterally-symmetric morphology can be well substantiated for animals, the role of microRNAs in drosophilid morphology has remained largely unexplored (Lu *et al.*, 2008; Stark *et al.*, 2007a). Likewise, scutellar bristles represent a morphological trait for drosophilid species diagnosis (FBbt:00004312, APPENDIX VI,TABLE 13), and notably microRNAs are suggested to canalize the ontological pathway controlling these bristle numbers (Grimaldi, 1990; Hornstein & Shomron, 2006; Markow & O'Grady, 2005). These FBbt dataset of genes associated to anatomical features and physiological processes used to diagnose species within *Drosophila* could potentially represent a genome sample of the microRNA regulatory core underlying species diagnostic phenotypes (FIGURE 25; APPENDIX VI, TABLE 13).

While individual genes have each experienced their own independent history and widespread inconsistency of gene trees is expected, nevertheless, not all divergent gene trees will be mutually incompatible to one another or to a reference tree (*Drosophila* 12 Genomes Consortium, 2007; Kopp & True, 2002; Pollard *et al.*, 2006). Phylogenetic utility may be addressed within this framework and comparable parametric score values are displayed in APPENDIX III, TABLE 12. The FBbt dataset recovered parametric scores comparable to *Drosophila* molecular markers available through NCBI database. Indeed the FBbt data exhibited higher percentage of parsimony informative characters, lower average homoplasy and higher average consistency to the reference tree than NCBI molecular markers. Nevertheless the average likelihood score for NCBI data is more optimal over the FBbt dataset; where *–ln* Likelihood score represents the sum of the probability of the data given the tree and the tree with lowest negative log-transformed likelihood is preferred. Nevertheless it seems reasonable to propose that the FBbt are a representative genomic sample more congruent to the reference tree topology than other phylogenetic markers in use for *Drosophila*.

While extensive molecular studies of drosophilids have already been conducted, there have been few successful attempts to revise taxonomic schema to fit phylogeny; despite the overwhelming accumulation of evidence against traditional phenetic groups erected on the basis of genital structure (O'grady *et al.*, 1998; Kwiatowski & Ayala, 1999; Kopp & True, 2002; Robe *et al.*, 2005). The Drosophilidae remain an agglomeration of non-monophyletic groups; where molecular data regularly present genera emerging from subgenera as a normal feature of the taxonomy (Robe *et al.*, 2005). The most notable offenders are the paradigm genus and subgenus of *Drosophila*: these have become catchall designations and are utterly useless in designating any biologically meaningful clade (Remsen and O'grady, 2002; Robe *et al.*, 2005). Drosophilid taxonomy will require exhaustive systematic overhaul to accommodate monophyletic groups recovered from the congruencies of molecular datasets (Robe *et al.*, 2005). But in application, groupings based on morphological characters need not be discarded outright, but should rather be evaluated for systematic utility based on consistency to molecular datasets. Where such measures have been provisionally implemented, remarkable consensus between molecular and morphological reconstructions have been recovered (Cameron *et al.*, 2007; Kopp & True, 2002; O'grady *et al.*, 1998). Here it is proposed that genes of the FBbt dataset linked species diagnostic phenotype could be useful in rationalizing selection of suitable molecular markers or morphological characters for *Drosophila* phylogeny. Those genes of the dataset exhibiting most optimum phylogenetic support to the reference topology could be used to propose suites of species-diagnostic morphological characters best suited for accurate phylogeny. Any novel molecular address to morphological features in drosophilid species diagnosis will have great potential applications to *Drosophila* systematics.

The phenotypic traits encoded by the genes of the FBbt regulatory network are likely to have adapted in relation to one another: selection for single phenotypic traits may exert section for a suite of other phenotypic traits through the action of shared regulatory elements. This principle has been demonstrated with foxes (*Vulpes vulpes*) bred for tamability in a 40-year experiment. Selection for tolerance to human socialization incurred significant physical changes in coat color, ears, limbs, and developmental timing of these foxes (Trut, 1999). These remarkable transformations essentially recapitulated the domestication of the dog; *Canis lupus* into *Canis familiaris* (Trut, 1999). In the case of *Drosophila*, selection for one single phenotypic trait may exert section for a number of other traits through the regulatory force of microRNA regulators held in common. Along these lines, the FBbt target dataset recovered enrichment for the S1T3 motif-14, and the S3T1 motif-2184 in eleven species and significant scores in S2T2 motif-240 for one species (TABLE 4 & 5). The patterns of species recovering significant scores for the FBbt target suite were different for the species profile of the unsorted (U Muller) targets. Indeed the FBbt dataset have average Z-scores higher than those observed for Muller elements A-F for S1T3 and S3T1 motifs. These findings indicate that the FBbt network has developed under a selective regime different from those presiding over individual Muller elements or the overall trascriptome. Moreover considering the S2T2 motif-240 enrichment for one species, lineage-specific selection seems to have been operative upon the regulation of genes linked to species diagnostic phenotypic traits.

CHAPTER V.

Nested Hierarchal Organization of Conservation for MicroRNAs and their Putative Targets to *Drosophila melanogaster*.

Keywords & concepts: Biological Complexity; GRAFMAN Software; MicroRNA Repertoire Expansion; MicroRNA-Target Network; Regulatory Network Conservation; Taxonomic Hierarchy.

ABSTRACT

It is hypothesized that microRNA-mediated transcript regulation has likely played a critical role in the primordial origins of complex animal body plans. Phylogenic gain in microRNA gene expression observed alongside the acquisition of organismal complexity is likely related to an increase in gene regulatory network complexity. This study examined microRNA network properties traced through taxonomic hierarchy considering both the acquisition of potential network targets and regulators. Primary literature review and database analysis were conducted to establish modules of conserved microRNAs across metazoan taxonomy. A hierarchical schema for the conservation of microRNAs and their putative targets to *Drosophila melanogaster* was engineered through comprehensive meta-analysis combing 1131 datasets from 325 species tracing through 207 subclades, gene homology assertion data from 12 databases, and 160 supercomputing BLAST-N searches (E-value $1e^{-5}$) of genome trace or expressed sequence tag (EST) data. Conservation history of 90.39 % of the total *Drosophila* dataset could be resolved through this hierarchical sampling regime; tracing from taxonomic order down to empire. The findings presented in this study represent the first documentations of *Drosophila* microRNA regulatory network behavior thorough taxonomic heirarchy. Scale free properties of a network intersection of microRNA target predictions methods were found to transect taxonomic hierarchy. Newly acquired microRNAs from novel families reinforce the pre-existing regulatory network and expand the targetset incrementally to include a small number of novel genes. Lineage specific microRNAs were found to exhibit far fewer conserved targets than do the more broadly conserved microRNAs; even when considering only more recently emerged targets. There was a dramatic expansion in network complexity with the expansion of the microRNA repertoire and this corresponds to the expansion in biological complexity

INTRODUCTION

MicroRNAs are unusual in comparison to other genetic elements in that they have been continually added to animal genomes. Indeed, the observed hierarchical conservation structure of microRNAs over phyletic distance is only possible if acquired microRNAs become fixed in an animal genome and are not lost secondarily (Heimberg *et al.*, 2008). There is a strong positive correlation between microRNA acquisition and morphological complexity observed in animals (Lee *et al.*, 2007). It is hypothesized that microRNA-mediated transcript regulation has likely played a critical role in the primordial origins of complex animal body plans (Sempere *et al.*, 2006). For instance, the dramatic expansion of the microRNA repertoire in bilaterally-symmetric animals relative to poriferans (sponges) and jellyfish (cnidarians) suggests that increased microRNA-mediated gene regulation accompanied the advent of organ-containing body plans drawn from three primary tissue types (FIGURE 29; Prochnik *et al.*, 2007). Likewise, character inference from deep phylogenetic analyses indicate potential functional causality between novel microRNA family acquisition and the foundation of the vertebrate phenotype. In the later scenario, novel microRNA acquisition would have had to proceed any genome duplication event (Heimberg *et al.*, 2008).

The phylogenic gain in microRNA gene expression observed alongside the acquisition of organismal complexity is likely related to an increase in gene regulatory network complexity (Heimberg *et al.*, 2008; Lee *et al.*, 2007; Sempere *et al.*, 2006). Newly acquired microRNAs would likely reinforce the pre-existing regulatory network, but will also target a small number of novel genes (Sempere *et al.*, 2007; Stark *et al.*, 2007a). Novel microRNAs acquired through *miR* gene duplication may be a simple mechanism for increasing microRNA cellular dosage, while microRNAs from novel families will enable fine-tuning to particular target suites (Lu *et al.*, 2008, Prochnik *et al.*, 2007,Stark *et al.*, 2007a). In either case, newly acquired microRNAs will have a significant impact on the overall microRNA regulatory network; but a large number of adaptive changes over a long period of time may be essential for full network integration (Sempere *et al.*, 2007; Stark *et al.*, 2007a).

Once integrated into a gene regulatory network, the mature microRNA sequences come under intense stabilizing selection ensuring conservation (Heimberg *et al.*, 2008, Sempere *et al.*, 2007). For example, strong conservation can be found in arthropods despite great phyletic distance, and similarly some 30 microRNAs remain conserved across all bilaterally symmetric animals (Lee *et al.*, 2007; Prochnik *et al.*, 2007). Generally speaking, microRNAs conserved in sequence are often expressed within identical tissues during analogous developmental stages in different organisms (Lee *et al.*, 2007). In considering conserved microRNAs expression patterns and computationally predicted targets in vertebrates and flies, it is likely that most of these microRNA mediated regulations control developmental pathways fundamental to bilaterally-symmetric animals (Enright *et al.*, 2003; Griffiths-Jones *et al.*, 2006; Prochnik *et al.*, 2007).

The purpose of this study was to examine microRNA network properties traced through taxonomic hierarchy considering both the acquisition of potential network targets and regulators. Network quantification and distance calculation comparisons of microRNA target network modules provided a valuable reference framework within which to evaluate the hierarchical conservation of microRNA-target interactions (FIGURE 29). The phyletic comparison of microRNA acquisition to patterns of target gene acquisition was a subject of particular interest. Consistent with the premise of interologous cross-species comparison, it is predicted that sets of well-conserved genes and ancient microRNAs will exhibit greatest interaction stability (Friedman *et al.*, 2009; Matthews *et al.*, 2001). The observed expansion of organismal complexity over natural history while utilizing a very similar genetic repertoire suggests that the complexity of genome regulation present in the organism also correlates with its complexity (Lee *et al.*, 2007; Sempere *et al.*, 2006). With this in view this study focused on relationships between microRNA acquisition and measures of network complexity. Notably, the findings presented in this study represent the first documentations of *Drosophila* microRNA regulatory network behavior thorough taxonomic heirarchy.

METHODS

Primary literature review and database analysis were conducted to establish modules of conserved microRNAs across metazoan taxonomy (Gilbert, 2007, Griffiths-Jones *et al.*, 2006; Hertel *et al.*, 2006; Sempere *et al.*, 2006; Sethupathy, *et al.*, 2006). Likewise a hierarchical schema for the conservation of microRNAs and their putative targets to *Drosophila melanogaster* was engineered through a comprehensive meta-analysis combing 1131 datasets from 325 species tracing through 207 subclades, gene homology assertion data from 12 databases, and 160 supercomputing BLAST-N searches (E-value $1e^{-5}$) of genome trace or expressed sequence tag (EST) data (Altschul *et al.*, 1990; McCarter, *et al.*, 2005; Benson *et al.*, 2008; Birney, *et al.*, 2006; Chen *et al.*, 2006 Gauthier *et al.*, 2007, Lawson *et al.*, 2009, Lee *et al.*, 2002, Marchler-Bauer *et al.*, 2009, Mulder *et al.*, 2002, Nègre *et al.*, 2006, O'Brien *et al.*, 2005, O'Brien *et al.*, 2004, Remm *et al.*, 2001, Sonnhammer, *et al.*, 1997, Tatusov *et al.*, 2003, Wheeler *et al.*, 2006, Wu *et al.*, 2006). Collectively, the conservation history of 90.39 % of the total *Drosophila* dataset could be resolved through this hierarchical sampling regime; tracing from taxonomic order down to empire (FIGURE 29). Phylogenies of taxa surveyed for conservation of microRNAs and targets to *Drosophila melanogaster* were rendered through TreeFam database and are presented in FIGURES 41 to 47 of APPENDIX VII (Li *et al.*, 2006; Ruan *et al.*, 2008).

Taxonomic Rank	Name	# Conserved MicroRNAs	# Targets	% Total Putative Targets	# Taxa Sampled	# Subclades	# Data Modules
Species	*Drosophila melanogaster*	112	14925	100.00	1	0	164
Species subgroup	Union to *D. simulans*	95	14925	100.00	2	1	89
Palearctic clade	Union to *D. pseudoobscura*	75	14925	100.00	2	1	89
Genus	*Drosophila*	65	14925	100.00	4	2	132
Order	Diptera	60	13495	90.42	6	3	24
Superorder	Endopterygota	55	13341	89.39	5	1	11
Phylum [clade]	Arthropoda	41	13283	89.00	3	2	5
Superphylum	Ecdysozoa	40	12997	87.08	11	3	54
Subkingdom [clade]	Protostomia *s.t.* (Union to Eutrochozoa)	39	12997	87.08	6	3	6
Subkingdom [clade]	Coelomata *s.t.* (Union to Deuterostomia)	37	12997	87.08	84	58	275
Subkingdom [clade]	Bilateria *s.t.* (Union to Platyhelminthes)	28	12997	87.08	5	2	5
Subkingdom [clade]	Nephrozoa	18	12997	87.08	111	69	388
Subkingdom [clade]	Triploblastica	6	11314	75.81	8	0	13
Subkingdom [clade]	Eumetazoa (Union to Cnidaria)	2	11020	73.84	2	1	4
Kingdom	Metazoa (Union to Porifera)	0	11008	73.76	1	0	3
Superkingdom [clade]	Opisthokonta (Union to Fungi)	0	9659	64.72	19	6	49
Superkingdom	Eukaryota	0	9163	61.39	60	27	47
Empire	Biota (Union to Archaea + Eubacteria)	0	1669	11.18	21	6	3
TOTAL	Putative MicroRNA Targets		14925	100.00	325	207	1572

FIGURE 29. Nested hierarchical organization of conservation for microRNAs and their putative targets to *Drosophila melanogaster*. The taxonomic rank and its name are given (left). The abbreviation *s.t.* represents the historical sense (*sensu traditionalis*) of specified taxonomic name. The numbers of nodes represent the numbers of genes of the total microRNA target dataset known to be conserved at each rank. Taxa typically indicate species, but in some cases represented compilations of sets of species groups in the raw data (see APPENDIX VII). Subclades represent the natural groupings present for the taxa sampled at each taxonomic rank. Data modules indicate the number of separate sets of data sampled per each rank. MicroRNAs are inferred according to the deepest rank at which they are conserved (Berezikov *et al.*, 2010; Gilbert, 2007; Griffiths-Jones *et al.*, 2006; Hertel *et al.*, 2006; Lu *et al.*, 2008; Sempere *et al.*, 2007; Sethupathy, *et al.*, 2006).

Taxonomic ranks were formalized by the following systematic conventions and criteria. The empire **Biota** defines union of all organic life apart from viruses (Brands, 2005; APPENDIX VII, FIGURE 47). The superkingdom **Eukaryota** defines life-forms with a cellular nucleus and membrane bound organelles (Chatton, 1925; APPENDIX V, FIGURE 47). The superkingdom clade **Opisthokonta** comprise a grouping of eukaryotes with flagellate cells including both the kingdoms of animals and fungi (Cavalier-Smith, 1987; APPENDIX VII, FIGURE 43). Sponges (phylum Porifera) and other multicellular animals are included in the kingdom **Metazoa** (Haeckel, 1896 APPENDIX VII, FIGURE 43). The subkingdom rank **Eumetazoa** includes hydra, corals, and jellyfish (phylum Cnidaria) and other organisms an embryonic development progressing through gastrulation (Brands, 2005; APPENDIX VII, FIGURE 43).

The **Triploblastica** represents the subkingdom clade including all organisms with tissue specific differentiation into three primary germ layers; endoderm, mesoderm, and ectoderm (Lankester, 1877). The subkingdom clade **Nephrozoa** is a grouping exclusive to Nemertodermatida and Acoela flatworms but inclusive of 23 bilaterian phyla; most of which contain some sort of excretory nephridial structures (Jondelius *et al.*, 2002). The subkingdom rank **Bilateria** in the traditional sense includes flatworms (phylum Platyheminthes) and other organisms with a bilateral axis of body symmetry (Hatschek, 1888; APPENDIX VII, FIGURE 45). The subkingdom rank **Coelomata**, in the traditional sense, contains those organisms with a fluid filled body cavity (coelom); including vertebrates, echinoderms (starfish, sea urchins, *etc*.), and other deuterostomes. (Hyman 1951; APPENDIX VII, FIGURE 43 to 45). The subkingdom rank **Protostomia**, in the traditional sense, includes of those organisms in which the blastopore deepens during gastrulation to become the archenteron as the first phase in the growth of the gut (Grobben, 1908; APPENDIX VII, FIGURE 45). The later group is also inclusive of molluscs (phylum Mollusca) and segmented worms (phylum Annelida) comprising the Eutrochozoa (Ghiselin, 1988).

The superphylum **Ecdysozoa** includes those animals which shed their exoskeleton (undergo ecdysis) including water bears and round worms; the phyla Tardigrada and Nematoda respectively (Aguinaldo *et al.*, 1997). The rank inclusive of arachnids, centipedes, crustaceans, insects, and millipedes represents the phylum **Arthropoda** and these organisms are distinguished by their exoskeleton, segmented body, and jointed appendages (Latreille, 1829; APPENDIX VII, FIGURE 42). The superorder **Endopterygota** contains those winged insects undergoing complete metamorphosis through a pupal stage and includes beetles (order Coleoptera) wasps (order Hymenoptera), caddisflies (order Trichoptera), butterflies and moths (order Lepidoptera), and fleas (order Siphonaptera; Sharp, 1898**;** APPENDIX VII, FIGURE 41). Lastly, true flies, mosquitoes, and gnats are included in the order **Diptera** (Linnaeus, 1758).

Each taxonomic module in was inclusive of the microRNAs and target data of the subordinate rank when tracing up from Biota to *Drosophila*. These conservation data were imported into the *musca* MySQL database. Further network quantification and distance analyses were performed using in-house GRAFMAN software available under Linux on the Watson supercomputer cluster of Virginia Commonwealth University (FIGURE 1; Karabunarliev & Bonchev, 2002). Data recovered from the *musca* database in the form of 575 network files (551.4 MB) were input to GRAFMAN and 2.4 MB of data were output (see APPENDIX III, TABLE 10 & 11). Likewise parametric scores from molecular phylogeny were recovered (see CHAPTER II) through the *musca* MySQL database and are presented in APPENDIX III, TABLE 13.

RESULTS & DISCUSSION

MicroRNA Regulatory Network Properties across Taxonomic Hierarchy. Vertex degree distribution and network abundance of microRNAs per target gene are presented by method in FIGURES 30, 31, & 32 for taxonomic rank specific subnetworks of microRNA-target network data. A power-law function was a poor mathematical descriptor of TargetScan data traced through taxonomic hierarchy (R^2 = 0.19 to 0.66; FIGURE 30). These functions for TargetScan aptamer-degree-frequency and target-degree-frequency distributions fell below normal biological range of -2 to -3 (*b* = -0.4647 to -0.696). Polynomial trend line functions were superior to power-law functions as descriptors of TargetScan data traced through taxonomic hierarchy (R^2 = 0.49 to 0.85; FIGURE 30). Similarly, the network data from MiRanda predictions cannot be cannot be precisely described by a power law when tracing through taxonomic rank (R^2 = 0.68 to 0.73; FIGURE 31). Nevertheless, MiRanda target-degree-frequency distributions consistently recovered power-law exponential values approaching or within biological range (*b* = -1.5237 to -2.9952). Power-law functions were suitable mathematical descriptors of target intersection network data traced through taxonomic hierarchy ($R^2 \approx 90$; FIGURE 32), while exponential trend line functions recoverd improved statistical support. These later data retrieved power-law function exponents that fell within biological range for all taxonomic ranks (*b* = -2.388 to -2.9294). Thus it appears that the scale free properties of the network intersection transect taxonomic hierarchy. Moreover, eponential trend line functions represented a better descriptor of target intersection network data traced through taxonomic hierarchy (R^2 = 0.97-0.99; FIGURE 32).

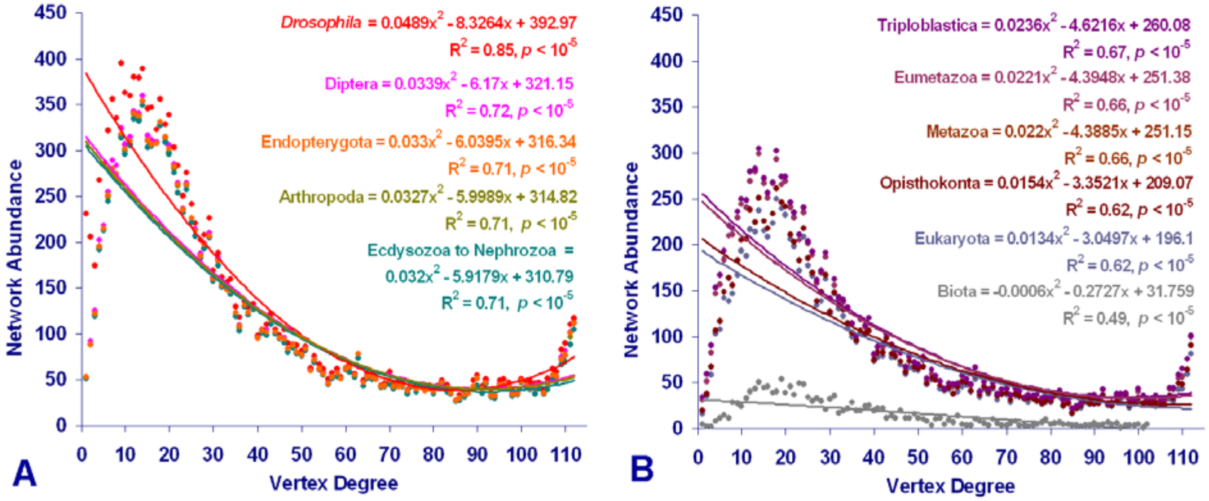


FIGURE 30. Vertex degree distribution and network abundance of microRNAs per target gene for taxonomic rank specific subnetworks of TargetScan microRNA-target network data. MicroRNA targets are predicted across the union of twelve *Drosophila* species. All data are unbinned. Taxonomic ranks from empire to genus were defined according to the schema presented in FIGURE 29 and trend line plots, functions, and coefficients of determination recovered for select taxonomic ranks are color-coded accordingly. Power-law trend line functions recovered non-linear regressions of 0.19 to 0.68, p < 10^{-4}. Polynomial trend line functions, non-linear regressions, and statistical significance recovered by rank were as follows:

Drosophila) $y = 0.0489\,x^2 - 8.3264\,x + 392.97$ $R^2 = 0.85$, $p < 10^{-5}$;

Diptera) $y = 0.0339\,x^2 - 6.17\,x + 321.15$, $R^2 = 0.72$, $p < 10^{-5}$;

Endopterygota) $y = 0.033\,x^2 - 6.0395\,x + 316.34$, $R^2 = 0.71$, $p < 10^{-5}$;

Arthropoda) $y = 0.0327\,x^2 - 5.9989\,x + 314.82$, $R^2 = 0.71$, $p < 10^{-5}$;

Ecdysozoa to Nephrozoa) $y = 0.032\,x^2 - 5.9179\,x + 310.79$, $R^2 = 0.71$, $p < 10^{-5}$;

Triploblastica) $y = 0.0236\,x^2 - 4.6216\,x + 260.08$, $R^2 = 0.67$, $p < 10^{-5}$;

Eumetazoa) $y = 0.0221\,x^2 - 4.3948\,x + 251.38$, $R^2 = 0.66$, $p < 10^{-5}$;

Metazoa) $y = 0.022\,x^2 - 4.3885\,x + 251.15$, $R^2 = 0.66$, $p < 10^{-5}$;

Opisthokonta) $y = 0.0154\,x^2 - 3.3521\,x + 209.07$;

Eukaryota) $y = 0.0134\,x^2 - 3.0497\,x + 196.1$, $R^2 = 0.62$, $p < 10^{-5}$;

Biota) $y = -0.0006\,x^2 - 0.2727\,x + 31.759$, $R^2 = 0.49$, $p < 10^{-5}$.

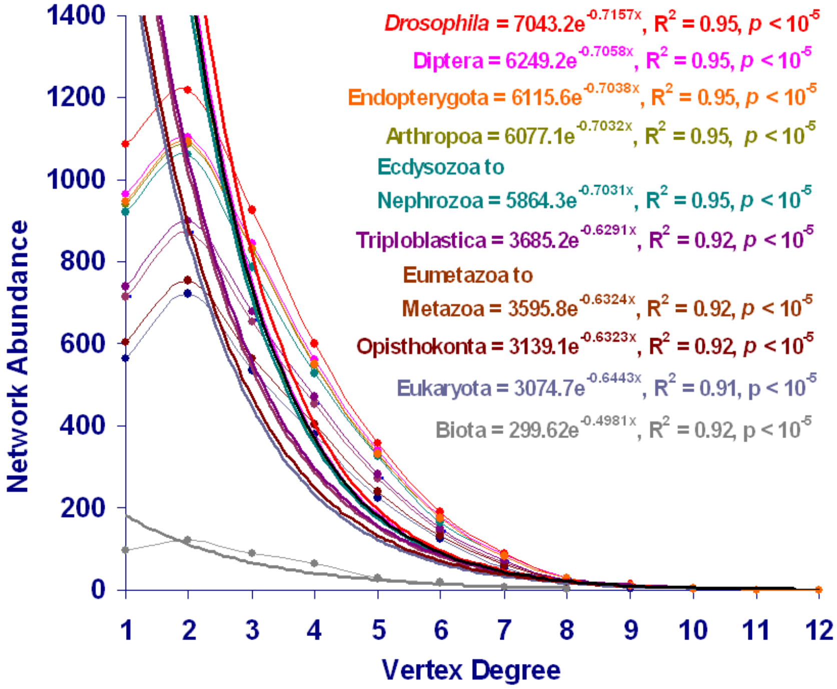


FIGURE 31. Vertex degree distribution and network abundance of microRNAs per target gene for taxonomic rank specific subnetworks of MiRanda microRNA-target network data. MicroRNA targets are predicted across the union of twelve *Drosophila* species. All data are unbinned. Taxonomic ranks from empire to genus were defined according to the schema presented in FIGURE 29 and trend line plots, functions, and coefficients of determination recovered for select taxonomic ranks are color-coded accordingly. Power-law trend line functions recovered non-linear regressions of 0.68 to 0.73, $p < 10^{-4}$. Exponential trend line functions, non-linear regressions, and statistical significance recovered by rank were as follows: ***Drosophila***) $y = 7043.2e^{-0.7157x}$, $R^2 = 0.95$, $p < 10^{-5}$; **Diptera**) $y = 6249.2e^{-0.7058x}$, $R^2 = 0.95$, $p < 10^{-5}$; **Endopterygota**) $y = 6115.6e^{-0.7038x}$, $R^2 = 0.95$, $p < 10^{-5}$; **Arthropoda**) $y = .1e^{-0.7032x}$, $R^2 = 0.95$, $p < 10^{-5}$; **Ecdysozoa to Nephrozoa**) $y = 5864.3e^{-0.7031x}$, $R^2 = 0.95$, $p < 10^{-5}$; **Triploblastica**) $y = 3685.2e^{-0.6291x}$, $R^2 = 0.92$, $p < 10^{-5}$; **Eumetazoa to Metazoa**) $y = 3595.8e^{-0.6324x}$, $R^2 = 0.92$, $p < 10^{-5}$; **Opisthokonta**) y = 3139.1e-0.6323x, $R^2 = 0.92$, $p < 10^{-5}$; **Eukaryota**) $y = 3074.7e^{-0.6443x}$, $R^2 = 0.91$, $p < 10^{-5}$; **Biota**) $y = 299.62e^{-0.4981x}$, $R^2 = 0.92$, $p < 10^{-5}$.

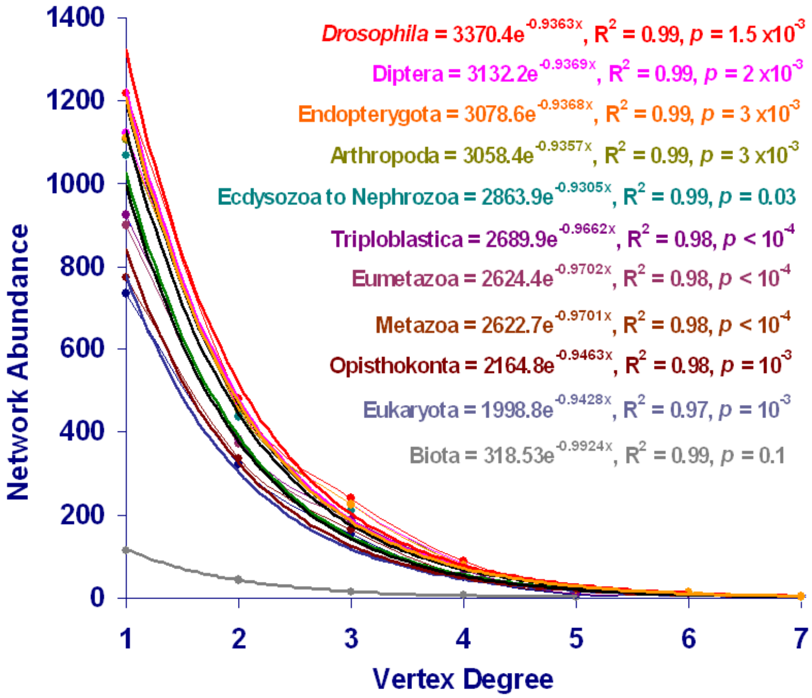


FIGURE 32. Vertex degree distribution and network abundance of microRNAs per target gene for taxonomic rank specific subnetworks from the network intersection of prediction methods. MicroRNA targets are predicted across the union of twelve *Drosophila* species using the network intersection of MiRanda and TargetScan. All data are unbinned. Taxonomic ranks from empire to genus were defined according to the schema presented in FIGURE 29 and trend line plots, functions, and coefficients of determination recovered for select taxonomic ranks are color-coded accordingly. Power-law trend line functions recovered non-linear regressions of 0.90 to 0.93, p < 10^{-4}. Exponential trend line functions, non-linear regressions, and statistical significance recovered by rank were as follows:

***Drosophila*)** $y = 3370.4e^{-0.9363x}$, $R^2 = 0.99$, $p = 1.5 \text{ x}10^{-3}$;

Diptera) $y = 3132.2e^{-0.9369x}$, $R^2 = 0.99$, $p = 2 \text{ x}10^{-3}$;

Endopterygota) $y = 3078.6e^{-0.9368x}$, $R^2 = 0.99$, $p = 3 \text{ x}10^{-3}$;

Arthropoda) $y = 3058.4e^{-0.9357x}$, $R^2 = 0.99$, $p = 3 \text{ x}10^{-3}$;

Ecdysozoa to Nephrozoa) $y = 2863.9e^{-0.9305x}$, $R^2 = 0.99$, $p = 0.03$;

Triploblastica) $y = 2689.9e^{-0.9662x}$, $R^2 = 0.98$, $p < 10^{-4}$;

Eumetazoa) $y = 2624.4e^{-0.9702x}$, $R^2 = 0.98$, $p< 10^{-4}$; **Metazoa)** $y = 2622.7e^{-0.9701x}$, $R^2 = 0.98$, $p< 10^{-4}$;

Opisthokonta) $y =2164.8e^{-0.9463x}$, $R^2 = 0.98$, $p = 10^{-3}$;

Eukaryota) $y = 1998.8e^{-0.9428x}$, $R^2 = 0.97$, $p = 10^{-3}$; **Biota)** $y = 318.53e^{-0.9924x}$, $R^2 = 0.99$, $p = 0.1$.

Regulatory Network Targets, Interactions, and MicroRNA Regulators through Taxonomic Hierarchy. The data presented here are essentially a history of the expansion of the *Drosophila* microRNA network. MicroRNA targets, and numbers of unique microRNAs, are compared to regulatory network edges conserved through taxonomic hierarchy across the union of twelve *Drosophila* species in FIGURE 33. Likewise, numbers of unique microRNAs are compared to microRNA targets conserved through taxonomic hierarchy across the union of twelve *Drosophila* species in FIGURE 34. Homology sampling of taxa sharing the rankings from Nephrozoa to Ecdysozoa recovered the same number of targets; while the microRNA repertoire expands from 18 to 40 genes (FIGURE 29). Additionally, there is a sizable gap in putative targets between the empire Biota and the superkingdom Eukaryota (FIGURE 29). Nevertheless growth of network interactions with the expansion network of putative target genes through taxonomic hierarchy could be well described using subtle curvilinear plots for all prediction methods (values R^2 = 0.99; FIGURE 33A). Likewise total network interactions for all target prediction methods increased with the expansion of the microRNA repertoire through taxonomic hierarchy. These data could be well accommodated using power-law trend lines (values R^2 = 0.98 to 99; FIGURE 33B). Moreover, when cross-comparing regulatory interactions under every method, scaling constants and exponent values of the power-law trend lines for the microRNAs are larger than for targets genes. From this information it can be reasoned that that newly acquired microRNAs from novel families reinforce the pre-existing regulatory network (Sempere *et al.*, 2007; Stark *et al.*, 2007a). These data also match a hypothesis that microRNA-mediated regulations control developmental pathways primordial to bilaterians. The later hypothesis was derived from comparison of microRNAs expression patterns and computationally predicted targets in vertebrates and flies (Enright *et al.*, 2003; Griffiths-Jones *et al.*, 2006; Prochnik *et al.*, 2007).

Direct comparison of expansion of the microRNA repertoire to expansion of putative target genes through taxonomic hierarchy recovered relationships well described with logarithmic trend lines (R^2 = 0.92 to 0.99; FIGURE 34). Thus, it appears that newly acquired microRNAs from novel families expand the targetset incrementally to include a small number of novel genes (Sempere *et al.*, 2007; Stark *et al.*, 2007a). Indeed, each prediction method did not reach regulation of all target nodes until the full microRNA complement was included in the genus *Drosophila* (FIGURE 34). The logarithmic plots of microRNA and target data for all prediction methods (FIGURE 34) are also consistent with the hypothesis that lineage specific microRNAs exhibit far fewer conserved targets than do the more broadly conserved microRNAs; even when considering only more recently emerged targets (Friedman *et al.*, 2009; Grün *et al.*, 2005; Stark *et al.*, 2007a).

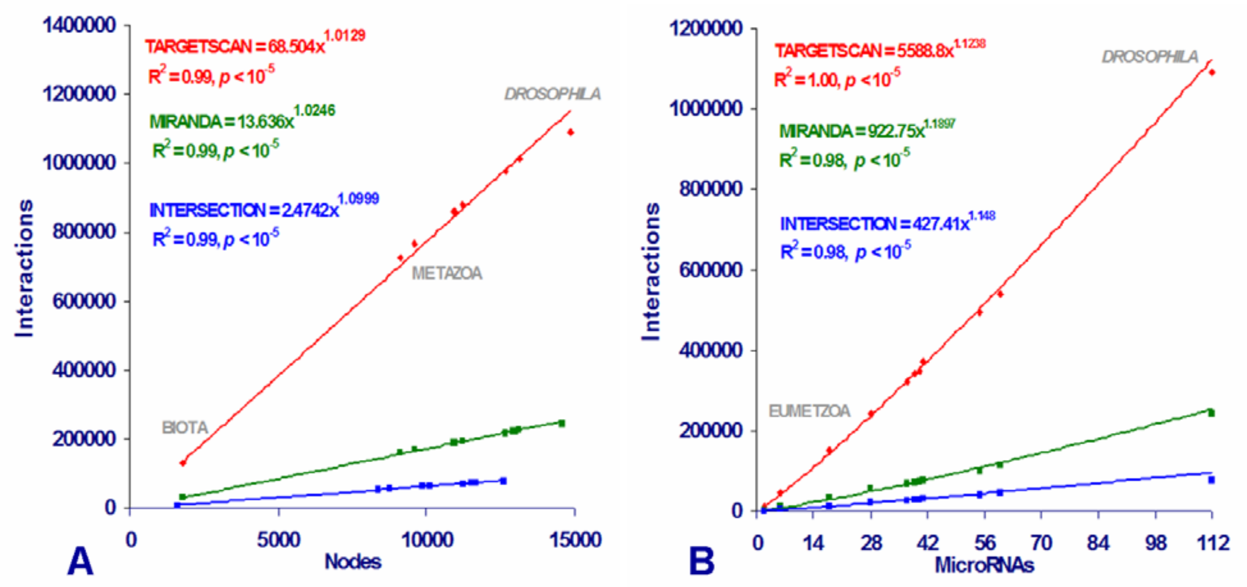


FIGURE 33. Comparison of microRNA targets, and unique microRNAs to regulatory network edges conserved through taxonomic hierarchy across the union of twelve *Drosophila* species. Source network data are presented in APPENDIX III, TABLES 10 & 11. Part (A) represents the expansion of putative target genes through taxonomic hierarchy where networks at each rank were produced with 112 microRNA families held constant. From these data the following trend lines and non-linear regressions were recovered: **TargetScan**) $y = 68.504\ x^{1.0129}$, R^2 = 0.99, $p < 10^{-5}$; **MiRanda**) $y = 13.636x^{1.0246}$, R^2 = 0.99, $p < 10^{-5}$; and for the network **intersection** of methods) $y = 2.4742\ x^{1.0999}$, R^2 = 0.99, $p < 10^{-5}$. Part (B) represents the expansion of the microRNA repertoire through taxonomic hierarchy where a total pool of 14925 available targets was held constant. The plotted curves represent power-law trend lines with functions and non-linear regression coefficient of determination for: **TargetScan**) $y = 5588.8x^{1.1238}$, R^2 = 1.00, $p < 10^{-5}$; **MiRanda**) $y = 922.75\ x^{1.1897}$, R^2 = 0.98, $p < 10^{-5}$; and for the network **intersection** of methods) $y = 427.41\ x^{1.148}$, R^2 = 0.98, $p < 10^{-5}$.

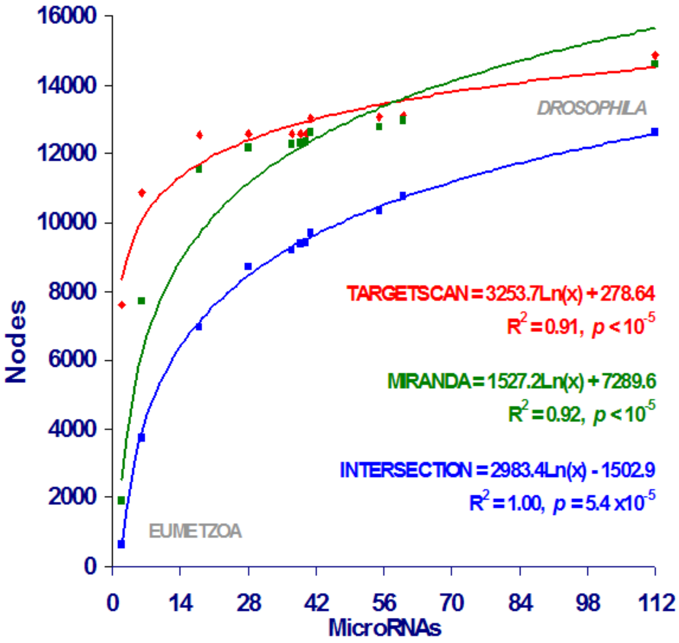


FIGURE 34. Comparison of numbers unique microRNAs to microRNA targets conserved through taxonomic hierarchy across the union of twelve *Drosophila* species. Original network data are presented in APPENDIX III, TABLES 10 & 11. The plotted curves demonstrate fit to logarithmic trend lines with functions and non-linear regression coefficient of determination as follows:

TargetScan) y = 3253.7 *Ln(x)* + 278.64, $R^2 = 0.91$, $p < 10^{-5}$;

MiRanda) y = 1527.2 *Ln(x)* + 7289.6, $R^2 = 0.92$, $p < 10^{-5}$;

and for the network **intersection** of methods) y = 2983.4 *Ln*(x) - 1502.9, $R^2 = 1.00$, $p < 10^{-5}$.

MicroRNA Regulatory Network Information Content and Biological Complexity through Taxonomic Hierarchy. A comparison of microRNA targets, and unique microRNAs to Shannon information index through taxonomic hierarchy is represented in FIGURE 35. The *Drosophila* microRNA interaction network for the network intersection of methods encoded over 9 billion bits of information (FIGURE 35A). This was nearly double the innate complexity of the parent methods and these patterns are maintained as network information content is traced through taxonomic hierarchy. Thus while the network intersection is a smaller dataset, it is substantially richer compared to its parent methods; and this quality is preserved throughout taxonomic hierarchy. Futhermore these information index data may be related to structural increase in biological complexity with the expansion of the microRNA repertoire (Lee *et al.*, 2007; Sempere *et al.*, 2006). The increase in gene regulatory network complexity is likely related to phylogenic gain in the acquisition of organismal complexity (Heimberg *et al.*, 2008; Lee *et al.*, 2007; Sempere *et al.*, 2006). There is a dramatic expansion in network complexity with the expansion of the microRNA repertoire and this corresponds to the expansion in biological complexity from Eumetazoa to Triploblastica (FIGURE 35B). These results harmonize to theories that increased microRNA-mediated gene regulation accompanied the advent of organ-containing body plans drawn from three primary tissue types (Prochnik *et al.*, 2007).

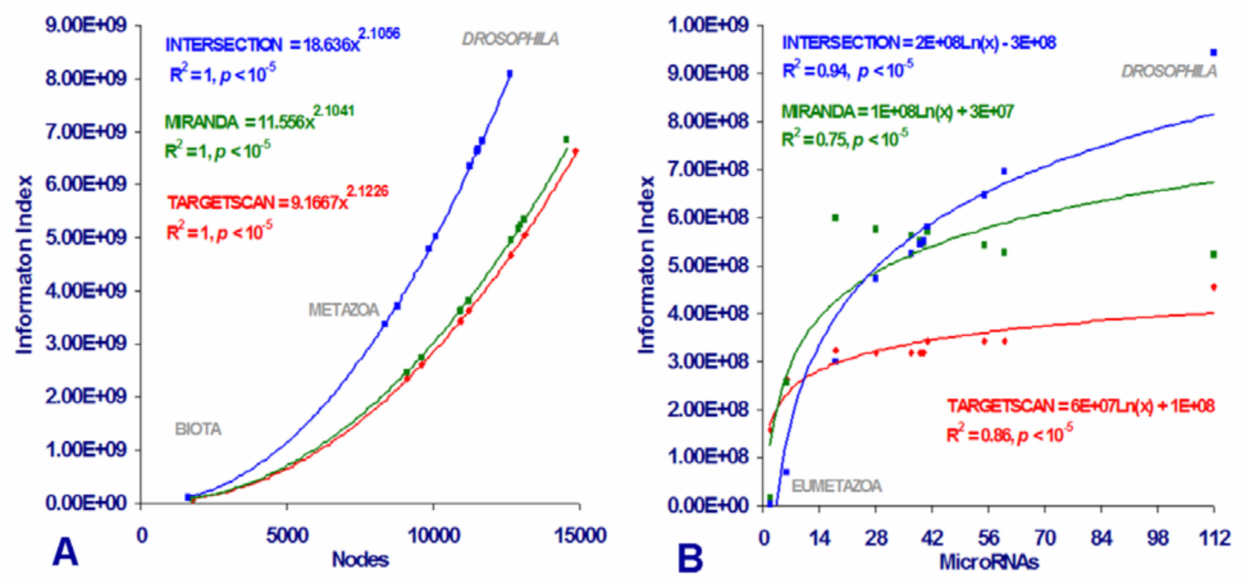


FIGURE 35. Comparison of microRNA targets, and unique microRNAs to Shannon information index for regulatory networks conserved through taxonomic hierarchy across the union of twelve *Drosophila* species. Original network quantifications are presented in APPENDIX III, TABLES 10 & 11.

Part (A) represents the expansion of putative target genes through taxonomic hierarchy. Networks at each rank were produced with 112 microRNA families held constant. The plotted curves represent power-law trend lines with functions and non-linear regression coefficient of determination for: **TargetScan**) $y = 9.1667\, x^{2.1226}$, $R^2 = 1$, $p < 10^{-5}$;

MiRanda) $y = 11.556\, x^{2.1041}$, $R^2 = 1.00$, $p < 10^{-5}$;

and for the network **intersection** of methods) $y = 18.636\, x^{2.1056}$, $R^2 = 1$, $p < 10^{-5}$.

Part (B) represents the expansion of the microRNA repertoire through taxonomic hierarchy where a total pool of 14925 available targets was held constant. The plotted curves demonstrate fit to logarithmic trend lines with functions and non-linear regression coefficient of determination as follows: **TargetScan**) $y = 6\text{x}10^7\, Ln(x) + 1\text{x}10^8$, $R^2 = 0.86$, $p < 10^{-5}$;

MiRanda) $y = 1\text{x}10^8\, Ln(x) + 3\text{x}10^7$, $R^2 = 0.74$, $p < 10^{-5}$;

and for the network **intersection** of methods) $y = 2\text{x}10^8\, Ln(x) - 3\text{x}10^8$, $R^2 = 0.94$, $p < 10^{-5}$.

MicroRNA Regulatory Network Connectedness through Taxonomic Hierarchy. MicroRNA targets, and unique microRNAs are compared to regulatory network connectedness conserved through taxonomic hierarchy in FIGURE 36. Network connectedness declines through taxonomic hierarchy with the expansion of putative target genes for all methods (FIGURE 36A). Thus, as the network was expanded with increasing available targets, the regulatory network density decreased. Conversely, network connectedness was variable through taxonomic hierarchy according to methods with the expansion of the microRNA repertoire (FIGURE 36B). Notably, the connectivity behavior of TargetScan data were well described with a linear function ($R^2 = 0.98$); indicating that TargetScan network density increased with the expansion of the microRNA repertoire. The differing behaviors for TargetScan and MiRanda data according to network connectivity likely reflect different selection profiles for seed-type and compensatory aptamers through taxonomic rank (see CHAPTER II).

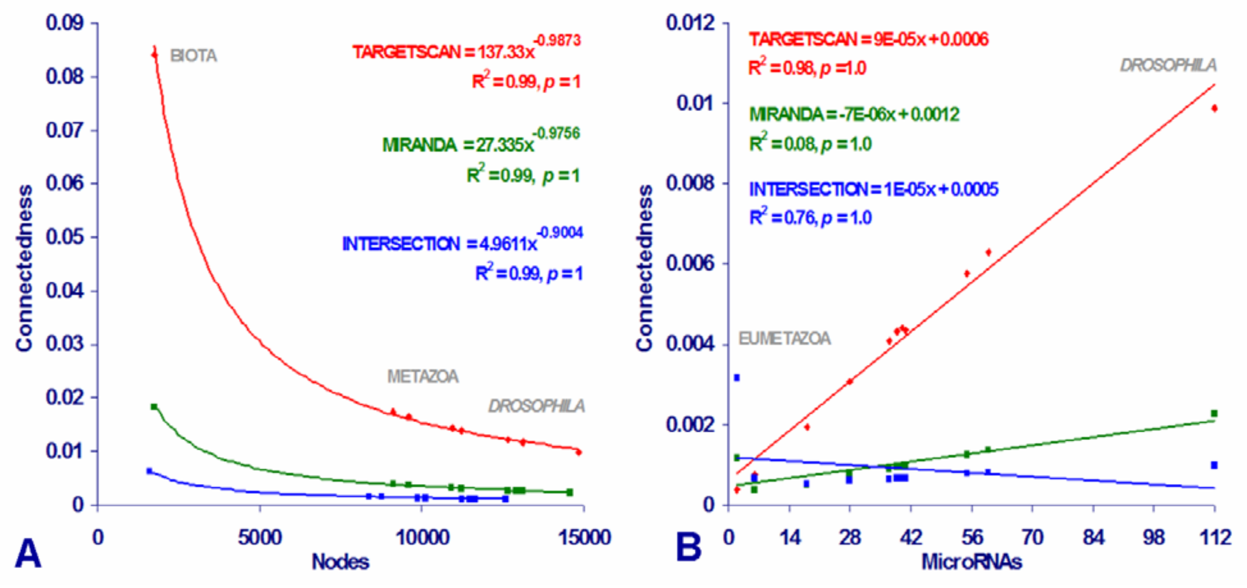


FIGURE 36. Comparison of microRNA targets, and unique microRNAs to network connectedness for regulatory networks conserved through taxonomic hierarchy across the union of twelve *Drosophila* species. Original network connectivity data are presented in APPENDIX III, TABLES 10 & 11. Part (A) represents the expansion of putative target genes through taxonomic hierarchy. Networks at each rank were produced with 112 microRNA families held constant. The plotted curves represent power-law trend lines with functions and non-linear regression coefficient of determination for:

TargetScan) y = 137.33 $x^{-0.9873}$, R^2 = 0.99, p = 1;

MiRanda) y = 27.335 $x^{-0.9756}$, R^2 = 0.99, p = 1;

and for the network **intersection** of methods) y = 4.9611 $x^{-0.9004}$, R^2 = 0.99, p = 1.

Part (B) represents the expansion of the microRNA repertoire through taxonomic hierarchy where a total pool of 14925 available targets was held constant. The plotted curves demonstrate fit to linear trend lines with functions and non-linear regression coefficient of determination as follows: **TargetScan**) y = 9 x10^{-5} x + 0.0006, R^2 = 0.98, p = 1;

MiRanda) y = 1x10¯5 x + 0.0005, R^2 = 0.76, p = 1;

and for the network **intersection** of methods) y = -7x10^{-6} x + 0.0012, R^2 = 0.08, p = 1.

Continuing Research. Further study should examine the conservation regime (FIGURE 29) of *Drosophila* microRNA-target networks for properties of bottom-up hierarchical (nested) modularity (Ravasz *et al.*, 2002). High average clustering coefficient is required for modular network organization and these modules represent discrete entities of elementary components and (presumed) functionality. Modular description of individual genes in regulatory networks allows for characterization of operon and regulon structures (Ravasz *et al.*, 2002). It is likely that most microRNA mediated regulations control developmental pathways fundamental to bilaterians (Enright *et al.*, 2003; Griffiths-Jones *et al.*, 2006; Prochnik *et al.*, 2007). And given that microRNAs conserved in sequence are often expressed within identical tissues during analogous developmental stages in different organisms, future work will consider the overlap of subnetworks of the nested hierarchical conservation regime to gene ontology expression data from microarrays and gauge with percent overlap if ontogeny recapitulates phylogeny at a molecular level (Gaidatzis *et al.*, 2007; Haeckel, 1867; Lee *et al.*, 2007). In support of this hypothesis, certain phylogenetic signal parameters were observed to increase with taxonomic depth. Total Goloboff Fit and likelihood scores to the reference tree generally became more optimal tracing downward from *Drosophila* to Biota (CHAPTER III, FIGURE 22). Indeed the likelihood score for the empire Biota is most optimal over all the higher taxonomic ranks; where –*ln* Likelihood score represents the sum of the probability of the data given the tree and the tree with lowest negative log-transformed likelihood is preferred (APPENDIX III, TABLE 13).

SUMMARY

MicroRNA regulatory networks are dense with most target genes targeted by multiple microRNAs, and exhibit precise combinatorial control of targets giving increased regulatory versatility. This study detailed the recovery and network analyses of a suite of homologous microRNA targets recovered through two different predicition methods for whole gene regions across twelve *Drosophila* species. TargetScan output (61.9GB) recovered a network of 14,860 targets, 1,090,221 microRNA-target interactions, 11,302,034 unique aptamer site interactions, and 112 microRNA families. Output form the MiRanda algorithm (2.96GB) recovered a network of 14,583 targets, 241,861 microRNA-target interactions, 390,560 unique aptamer site interactions, and 121 microRNA families. The network intersection of target prediction methods recovered a network of 12,616 targets, 78,280 microRNA-target interactions, 226,270 unique aptamer site interactions, and 112 microRNAs. Data recovered from microRNA target prediction were integrated with data from taxonomic hierarchical conservation and molecular phylogeny through a MySQL database of linked tables called "*musca*". It is notable then that there were 27 microRNAs among a total 245 genes recovered with a consistency greater than or equal to 90% for the reference tree topology. The methodology of this research outlined in FIGURE 1 can be readily reproduced for other organisms. The sizable target datasets produced in this study are applicable for continuing research in *Drosophila* molecular biology and could be biochemically verified using whole genome microarray analyses and miRNP immunopurification.

The intersection of microRNA target prediction methods produced networks of increased potential biological relevance compared to respective parent networks. This later network contained nearly double the innate complexity of the parent methods and these patterns are maintained as network information content is traced through taxonomic hierarchy. Thus while the network intersection is a smaller dataset, it is substantially richer compared to its parent methods; and this quality is preserved throughout taxonomic hierarchy. Futhermore this increase in network complexity was well correlated to structural increase in biological complexity with the expansion of the microRNA repertoire. These findings represent a novel documentation of *Drosophila* microRNA regulatory network behavior thorough taxonomic heirarchy.

MicroRNA regulatory network structure was found to change over time and across species. The decrease in conserved microRNA-target interactions with increasing phylogenetic distance exhibited a curve typical of a saturation phenomena. It seems that only a modest number of microRNA–mRNA interactions exhibit conservation over *Drosophila* cladogenesis. The minimal numbers of conserved microRNA-target interactions retained throughout all taxa were 1,839 from MiRanda, 13,357 from TargetScan, and 135 for the intersection of both methods. These latter values likely represent the presence of a functionally-constrained core of microRNA-target interactions essential to *Drosophila*. Networks may also have been influenced by the presence of 47 microRNAs exhibiting lineage specific expansion for in *Drosophila.* These collective findings represent the first comprehensive study to directly relate molecular sequence evolution and phylogeny with microRNA regulatory network interology in *Drosophila*.

An interplay of complex factors appears to operate in species conservation for microRNA regulation per target gene (FIGURE 12). Moreover, differential microRNA enrichment patterns by prediction method would seem that selective factors presiding over regulation by compensatory aptamers (MiRanda) and seed regions aptamers (TargetScan) are different (CHAPTER I). Selective factors that appear to operate upon seed aptamers include cooperativity (redundancy) of interactions and transcript length. Notably, these novel findings for entire messenger RNA transcripts are in accord with conclusions of other detailed analyses which have considered only the 3’UTR of *Drosophila* messenger RNAs (Stark *et al.*, 2005). As transcript length increases, the likelihood of acquisition of a seed-type aptamer binding site also increases. Support for a basic model of aptamer sequence evolution is addressed, where:
5’-seed↔5’-dominant↔ 3’-compensatory (Brennecke *et al*., 2005).

The signature of *Drosophila* phylogeny was found embedded within the microRNA regulatory network structure. The findings of this study represent the first documented inference of phylogeny from microRNA regulatory network structure and demonstrate the potential to accurately reconstruct phylogeny using abstract representations from network architecture. It is expected that microRNA interactome network data could serve as a useful counterpart to complement or supplement DNA sequence and morphology for phylogeny. Consistent congruence of regulatory network phylogenies to reference species tree topology also has strong implications to understanding microRNA-target natural history. Apparently, the phenetic approach of Neighbor Joining recovers better signal for the reference tree toplogy (FIGURE 22) over character-based standard parsimony (compare MiRanda in FIGURE 23 & 24). This would only be expected if phylogenetic history were best represented when the regulatory network was treated as single entity rather than a series of separable parts.

While a strong case can be substantiated for microRNA-moderated control over the basics of animal anatomy, the roles of microRNA regulation for details of fly anatomy remain largely unexplored. Any resulting integration of microRNA gene regulatory networks to chromosome or anatomical data for *Drosophila* species diagnosis represents an important step to broadening an understanding of the mechanics of speciation. The findings presented in this study represent a novel intergration of microRNA regulatory network topology to chromsomal synteny and genes linked to species diagnostic phenotypes. Topological analyses of microRNA regulatory networks recovered significant enrichment for the S2T2 motif possessing a redundant link (motif-204) in all twelve species sampled for many Muller elements (TABLE 7). The network enrichment of motifs possessing partial internal redundancy would have powerful implications toward understanding *Drosophila* speciation at the level of microRNA-gene regulatory interactions: this would suggest that optimization of the whole interactome topology itself has been historically subject to natural selection where resilience to attack have offered selective advantage. The repeating motif patterns across elements observed would not be expected if Muller elements were not a natural subdivision of the total *Drosophila* regulatory network. Collective patterns observed indicate that respective Muller element networks have developed within the *Drosophila* transcriptome as separate regulatory modules. The results of this study for regions of major chromosome synteny also have powerful implications toward the genetic basis of Haldane's rule. A fast-X hypothesis of Haldane's rule may be contradicted at the level of network topology and nucleotide sequence evolution.

Literature review for genes linked to anatomical features and physiological processes features used to diagnose species within the genus *Drosophila*, recovered a novel list of 2,331 genes (14.38% total target dataset) from 118 FlyBase anatomy terms (FBbt) and 93 gene ontology (GO) categories. Notably, these FBbt dataset of genes could potentially represent a genome sample of the microRNA regulatory core underlying species diagnostic phenotypes. Moreover, genes of the FBbt dataset linked species diagnostic phenotype could be useful in rationalizing selection of suitable molecular markers or morphological characters for *Drosophila* phylogeny. Motif enrichment patterns indicate that lineage-specific selection seems to have been operative upon the regulation of genes linked to species diagnostic phenotypic traits.

REFERENCES

1. Aguinaldo, A. M. A., Turbeville, J. M., Linford, L. S., Rivera, M. C., Garey, J. R., Raff, R. A. & Lake, J. A. (1997). "Evidence for a clade of nematodes, arthropods and other moulting animals". *Nature*, 489–493.
2. Akashi, H., Kliman, R. M. & Eyre-walker, A. (1998). Mutation pressure, natural selection, and the evolution of base composition in *Drosophila. Genetica*, 102/103, 49–60.
3. Albert, R., (2005) Scale-free networks in cell biology. *Journal of Cell Science*, 118, 4947.
4. Albert, I., & Albert, R. (2004). Conserved network motifs allow protein-protein interaction prediction. *arXiv.org>q-bio> q-bio/0406042*.
5. Altschul S. F, Gish W., Miller, W., Myers, E.W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215, 403–410.
6. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25, 3389–3402.
7. Ambros, V. (2004). The functions of animal microRNAs. *Nature*, 431, 350-355.
8. Aravin A. A., M. Lagos-Quintana, A. Yalcin, M. Zavolan, D. Marks, B. Snyder, T. Gaasterland, J. Meyer, & T. Tuschl. (2003). The small RNA profile during *Drosophila melanogaster* development. *Developmental Cell Biology*, 5, 337-350.
9. Baskerville, S. & Bartel, D. P. (2005). Microarray profiling of microRNAs reveals frequent coexpression with neighboring microRNAs and host genes. *RNA*, 11, 241-247.
10. Barabasi, A.-L., Albert, R. E. (1999) Emergence of Scaling in Random Networks. Science 1999, 286, 509-512.
11. Bartel, D. P., & Chen, C. Z. (2004). Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. *Nature Reviews Genetics*, 5, 396–400.
12. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. (2008). GenBank. *Nucleic Acids Research*, 36, D25-D30.
13. Berezikov, E., Liu, N., Flynt, A.S., Hodges, E., Rooks, M. Hannon, G.J., & Lai, E.C. (2010). Evolutionary flux of canonical microRNAs and mirtrons in *Drosophila*, Nature Genetics, 42, 6-9.
14. Birney, E., Andrews, D., Caccamo, M., *et al.*, (2006). Ensembl 2006. *Nucleic Acids Research*, 34, D556–D561.
15. Birney, E., Clamp, M. and Durbin, R. (2004). GeneWise and Genomewise. *Genome Research*, 14, 988-995.
16. Blanchette M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D., Haussler, D., Miller, W. (2004). Aligning multiple genomic sequences with the threaded blocks*et al*igner. *Genome Research*, 14, 708-715.
17. Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L. & Rubin, E.M. (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, 299, 1391-1394.
18. Bonchev, D. G. (1983) Information-Theoretic Indices for Characterization of Chemical Structures. Research Studies Press: Chichester, UK.
19. Bonchev, D. G., *et al.*, (2009). Networks Motif Topology and Evolution. *in prep.*
20. Brands, S. J. (comp.) (1989-2005). *Systema Naturae* 2000. Amsterdam, The Netherlands. http://sn2000.taxonomy.nl/>
21. Bray N., & Pachter L. (2004). MAVID: Constrained ancestral alignment of multiple sequences. *Genome Research*, 4:693-699.
22. Brennecke, J., Stark, A., Russell, R. B., & Cohen, S. M. (2005). Principles of MicroRNA–Target Recognition. PLoS Biology, 3, 404-418.
23. Bull, J. J, Huelsenbeck, J. P, Cunningham, C. W., Swofford, D. L., & Waddell, P. J. (1993). Partitioning and combining data in phylogenetic analysis. *Systematic Biology*, 42, 384–397.
24. Burkhart, B. D., Montgomery, E., Langley, C. H., Voelker, R. A. (1984). Characterization of allozyme null and low activity alleles from two natural populations of *Drosophila* MELANOGASTER. *Genetics*, 107, 295-306.
25. Cameron, S. L., Lambkin, C. L., Barker, S. C., & Whiting, M. F. 2007. A Mitochondrial genome phylogeny of Diptera: whole genome sequence data accurately resolve relationships over broad timescales with high precision. *Systematic Entomology*, 32, 40–59.
26. Carpenter, J. M., Goloboff, P. A., & Farris, J. S. (1998). PTP is meaningless, T-PTP is contradictory: A reply to Trueman. *Cladistics*, 14, 105–116.
27. Cavalier-Smith, T. (1987). "The origin of fungi and pseudofungi". in Rayner, Alan D. M. (ed.). *Evolutionary biology of Fungi*. Cambridge: Cambridge University Press, 339–353.

28. Chatterji, S. & Pachter, L. (2006). Reference based annotation with GeneMapper, *Genome Biology*, 7, R29.
29. Chatton, É. (1925). *Pansporella perplexa, amoebiens à spores protégées parasite de daphnies. Réflections sur la biologie et la phylogénie des protozoaires*. Annals Science National Zoology (Sér. 10), 8, 5–84
30. Chen, F., Mackey, A. J., Stoeckert, C. J. Jr, & Roos, D. S. (2006). OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Research*, 34, D363-D388.
31. Clark, A. G. & Wang, L. (1994). Comparative evolutionary analysis of metabolism in nine *Drosophila* species. *Evolution*, 48, 1230-1243.
32. Clemente, J. C., Satou, K., & Valiente G.l. (2007). Phylogenetic reconstruction from non-genomic data. *Bioinformatics*, 23, e110–e115.
33. Coyne, J.A. (1985). The genetic basis of Haldane's rule. *Nature*, 314, 736-738.
34. Coyne, J. A., & Orr, H. A. (1989). Patterns of speciation in *Drosophila*. *Evolution*, 43, 362–381.
35. Coyne, J. A., & Orr, H. A. (1997). "Patterns of speciation in *Drosophila*" revisited. *Evolution*, 51, 295–303.
36. Dandekar, T., Schuster, S., Snel, B., Huynen, M., & Bork, P. (1999). Pathway alignment: Application to the comparative analysis of glycolytic enzymes. *The Biochemical Journal*, 343, 115–124.
37. de Wit, E., Linsen, S. E., Cuppen, E., & Berezikov, E. (2009). Repertoire and evolution of microRNA genes in four divergent nematode species. *Genome Research*, 19, 2064-2074.
38. Dewey, C. N. (2007). Aligning multiple whole genomes with Mercator and MAVID. *Methods in Molecular Biology*, 395, 221-236.
39. Dewey, C., & Pachter, L. (2006). Mercator: Multiple whole-genome orthology map construction. Available: http://bio.math.berkeley.edu/mercator/.
40. Dobzhansky, T. (1936). Studies on hybrid sterility. II. Localization of sterility factors in *Drosophila pseudoobscura* hybrids. *Genetics* 21:113–35.
41. Dorogovtsev, S.N. & Mendes, J.F.F. (2003). Evolution of Networks: from biological networks to the Internet and WWW, Oxford University Press, ISBN 0-19-851590-1.
42. *Drosophila* 12 Genomes Consortium (2007). Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*, 450, 203-218.
43. Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C., & Marks, D. S. (2003). MicroRNA targets in *Drosophila*. *Genome Biology*, 5, R1-R14.
44. Easow, G., Aurelio A. Teleman, A. A.,& Cohen, S. M. (2007). Isolation of microRNA targets by miRNP immunopurification. *RNA*, 13, 1–7
45. Faith, D. P. (1991). Cladistic permutation tests for monophyly and nonmonophyly. *Systematic Zoology*, 40, 366-375.
46. Faith, D. P., & Trueman, J. W. H. (1996). When the topology-dependent permutation test (T-PTP) for monophyly returns significant support for monophyly, should that be equated with (a) rejecting a null hypothesis of nonmonophyly, (b) rejecting a null hypothesis of "no structure", (c) failing to falsify a hypothesis of monophyly, or (d) none of the above? *Systematic Biology*, 45, 580–586.
47. Farh, K. K. *et al.* (2005) The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science*, 310, 1817–1821.
48. Flannick, J. Novak, A., Srinivasan, B. S., McAdams, H. H., Batzoglou, S. (2006). Graemlin: general and robust alignment of multiple large interaction networks. *Genome Research*, 16, 1169-81.
49. Flavell, A. J., Knox, M. R., Pearce, S. R. & Ellis, T. H. N. (1999) Retrotransposon-based insertion polymorphisms (RBIP) for high-throughput marker analysis. *Plant Journal*, 16, 643-650.
50. Forst, C.V., Flamm, C., Hofacker, I. L., & Stadler, P. F. (2006). Algebraic comparison of metabolic networks, phylogenetic inference, and metabolic innovation. *BMC Bioinformatics*, 7, 67.
51. Forst, C.V., & Schulten, K. (2001). Phylogenetic analysis of metabolic pathways. *Journal Molecular Evolution*, 52, 471–489.
52. Friedman, R. C., Farh, K. K-H., Burge, C. B., & Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs *Genome Research*, 19, 92-105.
53. Gaidatzis, D., van Nimwegen, E., Hausser, J., & Zavolan, M. (2007). Inference of microRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics*, 8, 69.
54. Gascuel, O., & Steel, M. (2006). "Neighbor-joining revealed". *Molecular Biology & Evolution*, 23, 1997–2000.
55. Gauthier, J.P., Legeai, F., Zasadzinski, A., Claude Rispe C., & Tagu, D. (2007). AphidBase: a database for aphid genomic resources. *Bioinformatics,* 23, 783-784.
56. Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Impelfeld, D., B., Edelmann, A., Heurtier, M.-A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A.-M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari,

G., Drewes, G., Neubauer, G., Rick, J. M., Kuster, B., Bork, P., Russell, R. B., Superti-Furga, G. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440, 631.

57. Ghiselin, M. T. (1988). The origin of molluscs in the light of molecular evidence. pp. 66-95 in l H. Harvey and L. Partridge, eds. Oxford surveys in evolutionary biology. Vol. 5. Oxford University Press, New York.

58. Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W. J., Nekrutenko, A. (2005). Galaxy: A platform for interactive large-scale genome analysis. *Genome Research*, 15, 1451–1455.

59. Gilbert, D.G., (2007). DroSpeGe: rapid access database for new *Drosophila* species genomes. *Nucleic Acids Research*, 35, D480-D485.

60. Goloboff, P. A. (1993). Estimating character weights during tree search. *Cladistics*, 9, 83–91.

61. Gompel, N., Prud'homme, B., J. Wittkopp, P. J., Kassner, V. A., & Carroll, S. B. (2005). Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in *Drosophila*. *Nature*, 433, 481-487.

62. Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A., & Enright, A. J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*, 34, D140-D144.

63. Grimaldi, D. A. (1990). A phylogenetic, revised classification of genera in the Drosophilidae (Diptera). *Bulletin of the American Museum of Natural History*, 197, pp.139.

64. Grimson, A. *et al.* (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular Biology of the Cell*, 27, 91–105.

65. Grobben, K. (1908). *Die systematische Einteilung des Tierreichs*. Verh. Zool. Bot. Ges. Wien 58 (1908), pp. 491–511.

66. Grün, D., Wang, Y. L., Langenberger, D., Gunsalus, K. C., & Rajewsky, N. (2005). MicroRNA Target Predictions across Seven *Drosophila* Species and Comparison to Mammalian Targets. *PLoS Computational Biology*, 1, e13.

67. Habig,J. W., Dale,T., & Bass, B. L.(2007). miRNA Editing—We Should Have Inosine This Coming. *Molecular Cell*, 25, 792-793.

68. Haeckel, E. 1867. *Generelle Morphologie der Organismen*. Georg Reimer, Berlin.

69. Haeckel, E. (1896). *Systematische Phylogenie der Wirbellosen Thiere*; Teil 2. Georg Reimer, Berlin.

70. Haldane, J. B. S. (1922). Sex ratio and unisexual sterility in hybrid animals. *Journal of Genetics*, 12, 101-109.

71. Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, 41, 95–98.

72. Hartwell, L.H., Hopfield, J.J., Leibler, S., & Murray, A.W. (1999). From molecular to modular cell biology. *Nature*, 402, 47–52.

73. Hastings, H. M., Sugihara, G. (1993). *Fractals: A Users Guide for the Natural Sciences.* Oxford University Press, Oxford, U.K.,

74. Hatschek, B. (1888). *Lehrbuch der Zoologie*, 1. Lieferung, Gustav Fischer, Jena. pp. 1—144.

75. He, L. and G. J. Hannon. (2004). MicroRNAs: Small RNAs with a big role in gene regulation. *Genetics*, 5, 522-531.

76. Heger, A. & Ponting, C. (2006) *Drosophila* gene prediction pipeline with Exonerate. URL: http://wwwfgu.anat.ox.ac.uk:8080/flies/documentation.html

77. Heimberg, A. M., Sempere, L. F., Moy, V. N., Donoghue, P. C., & Peterson, K. J. (2008). MicroRNAs and the advent of vertebrate morphological complexity. *Proceedings of the National Academy of Science, USA*, 105, 2946-2950.

78. Hertel, J. Lindemeyer, M., Missal, K., Fried, C., Tanzer, A., Flamm, C., Hofacker, I. L., Stadler, P. F., & the Students of Bioinformatics Computer Labs 2004 and 2005. (2006). The Expansion of the Metazoan MicroRNA Repertoire. *BMC Genomics*, 7, 1-26.

79. Heymans, M. & Singh, A.K. (2003). Deriving phylogenetic trees from the similarity analysis of metabolic pathways. *Bioinformatics*, 19 (Suppl. 1), i138–i146.

80. Hillis, D. M., & Bull, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology*, 42, 182–192.

81. Hoffmann, A. A., Sgrn, C. M., & Weeks, A. R. (2004) Chromosomal inversion polymorphisms and adaptation. *Trends in Ecology & Evolution*, 19, 482-488.

82. Hong, S.H., Kim, T.Y., & Lee, S.Y. (2004). Phylogenetic analysis based on genome-scale metabolic pathway reaction content. *Applied Microbiology and Biotechnology*, 65, 203–210.

83. Hornstein, E. & Shomron, N. (2006). Canalization of development by microRNAs. *Nature Genetics*, 38, S20-S24.

84. Huelsenbeck, J. P., & Ronquist, F. (2005). MrBayes v3.1.2: Bayesian Analysis of Phylogeny Distributed by the author. San Diego: Section of Ecology, Behavior and Evolution, Division of Biological Sciences University of California.

85. Huynh, T., MiRanda, K., Tay, Y., Ang, Y.-S., Tam, W.-L., Thomson, A. M., Lim, B., & Rigoutsos, I. (2006). A pattern-based method for the identification of microRNA-target sites and their corresponding RNA/RNA complexes. *Cell*, 126, 1203-1217.

86. Hyman, L. H. (1951). The invertebrates. Vol. 2: Platyhelminthes and Rhynchocoela. McGraw-Hill, New York.

87. Ji, Z., Lee, J. Y., Pan, Z., Jiang, B., & Tian, B. (2009) Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proceedings of the National Academy of Science*, 106, 7028–7033

88. Johnson, N. A., Porter, A. H. (2000). Rapid speciation via parallel, directional selection on regulatory genetic pathways. *Journal of Theoretical Biology*, 205, 527–542.

89. Jondelius, U., Ruiz-Trillo, I., Baguñà, J. & Riutort, M. (2002). The Nemertodermatida are basal bilaterians and not members of the Platyhelminthes. *Zoologica Scripta*, 31, 201–215.

90. Kalaev, M., Smoot, M,. Ideker, T., Sharan, R. (2008). NetworkBLAST: comparative analysis of protein networks. *Bioinformatics*, 24, 594-596.

91. Kalendar, R., Grob, T., Regina, M., Suomeni, A., & Schulman, A. (1999). IRAP and REMAP: two new retrotransposon-based DNA fingerprinting techniques. *Theoretical and Applied Genetics*, 98, 704–711.

92. Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., & Yamanishi, Y. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, 36, D480-D484.

93. Kanehisa, M. & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28, 27-30.

94. Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., & Hirakawa, M. (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research*, 34, D354-357.

95. Karabunarliev, S. & Bonchev (2002). Grafman software package.

96. Katoh, K, & Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics*, 9, 286-298.

97. Kelley, B.P., Sharan, R., Karp, R.M., Sittler, T., Root, D.E., Stockwell, B.R., & Ideker, T. (2003). Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proceedings of the National Academy of Science*, 100, 11394–11399.

98. Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Research*, 12, 996–1006.

99. Khaitovich, P., Weiss, G., Lachmann, M., Hellmann, I., Enard, W., Muetzel, B., Wirkner, U., Ansorge, W., Pääbo, S. A neutral model of transcriptome evolution. (2004). *PLoS Biology*, 5, E132.

100. Kheradpour P., Stark A., Roy S., & Kellis M. (2007). Reliable prediction of regulator targets using 12 *Drosophila* genomes. *Genome Research*, 1, 1919-1931.

101. Kliman, R. M., & Hey, J. (1993). Reduced natural selection associated with low recombination in *Drosophila melanogaster. Molecular Biology and Evolution*, 10, 1239–1258.

102. Kloosterman, W. P., Wienholds, E., Ketting, R. F. & Plasterk, R. H. Substrate requirements for let-7 function in the developing zebrafish embryo. (2004). *Nucleic Acids Research*, 32, 6284–6291 .

103. Ko, W. Y., David, R. M., & Akashi H. (2003). Molecular phylogeny of the *Drosophila melanogaster* species subgroup. *Journal of Molecular Evolution*, 57, 562-573.

104. Kopp, A. & True, J. R. (2002). Phylogeny of the Oriental *Drosophila melanogaster* Species Group: A Multilocus Reconstruction. *Systematic Biology*, 51, 786–805.

105. Koyuturk, M., Grama, A., & Szpankowski, W. (2005). Pairwise local alignment of protein interaction networks guided by models of evolution. *Lecture Notes in Bioinformatics*, 3500, 48–65.

106. Kuhn, R. M., Karolchik, D., Zweig, A. S., Trumbower, H., Thomas, D. J., Thakkapallayil, A., Sugnet, C. W., Stanke, M., Smith, K. E., Siepel, A., Rosenbloom, K. R., Rhead B., Raney, B. J., Pohl A., Pedersen, J. S., Hsu, F., Hinrichs, A. S., Harte, R. A., Diekhans, M., Clawson, H., Bejerano, G., Barber, G. P., Baertsch, R., Haussler, D., Kent WJ. (2007). The UCSC Genome Browser database: Update 2007. *Nucleic Acids Research*, 35, D668–D673.

107. Kumar A, Hirochika H. (2001). Applications of retrotransposons as genetic tools in plant biology. *Trends in Plant Sciences*, 6, 127–134.

108. Kumar S, Dudley, J., Nei, M., & Tamura, K. (2008). MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. *Briefings in Bioinformatics*, 9, 299-306.

109. Kwiatowski, J., & Ayala, F. J. (1999). Phylogeny of *Drosophila* and Related Genera: Conflict between Molecular and Anatomical Analyses. *Molecular Phylogenetics and Evolution*, 13, 319–328.

110. Lachaise, D., Cariou, M. L., David, J. R., Lemeunier, F., Tsacas, L., *et al.* (1988). Historical biogeography of the *Drosophila melanogaster* species subgroup. *Evolutionary Biology*, 22, 159–225.

111. Lai, E. C., P. Tomancak, R. W. Williams, & Rubin, G. M. (2003). Computational identification of *Drosophila* microRNA genes. Genome Biology 4, R42.

112. Lanave, C., Preparata, G., Saccone, C., & Serio, G. (1984). A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution*, 20, 86-93.

113. Lankester, E. Ray. (1877). Notes on the embryology and classification of the animal kingdom: Comprising a revision of speculations relative to the origin and significance of germ layers. Quarterly Journal of the Microscience Socitey, 17, 399-454.

114. Latreille, P. A. (1829). *Les crustacés, les arachnides et les insectes, distribués en familles naturelles, ouvrage formant les tomes 4 et 5 de celui de M. le Baron Cuvier sur le règne animal* (deuxième édition). Tome second Paris Déterville. xxiv-556.

115. Lawson, D., Arensburger, P., Atkinson, P., Besansky, N. J., Bruggner, R. V., Butler, R., Campbell, K. S., Christophides, G. K., Christley, S., Dialynas. E., Hammond, M., Hill, C. A., Konopinski, N., Lobo, N. F., MacCallum, R. M., Madey, G., Megy, K., Meyer, J., Redmond, S., Severson, D. W., Stinson, E. O., Topalis, P., Birney, E., Gelbart, W. M., Kafatos, F. C., Louis, C., & Collins, F. H. (2009). VectorBase: a data resource for invertebrate vector genomics. *Nucleic Acids Research*, 37, D583-587.

116. Lee, C. T., Risom T., & Strauss, W. M. (2007). Evolutionary conservation of microRNA regulatory circuits: an examination of microRNA gene complexity and conserved microRNA-target interactions through metazoan phylogeny. *DNA and Cell Biology*, 26, 209-218.

117. Lee, Y., Sultana, R., Pertea, G., Cho, J., Karamycheva, S., Tsai, J., Parvizi, B., Cheung, F., Antonescu V., White, J., Holt, I., Liang, F., & Quackenbush (2002). Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). *Journal Genome Research*, 2, 493-502.

118. Lewis, B. P., Burge, C. B., & Bartel, D. P. (2005). Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets. *Cell*, 120:15-20 (2005).

119. Li H., Coghlan, A., Ruan, J., Coin, L. J., Hériché, J. K., Osmotherly, L., Li, R., Liu, T., Zhang, Z., Bolund, L., Wong, G. K., Zheng, W., Dehal, P., Wang. J., & Durbin, R. (2006). TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Research*, 1, D572-80.

120. Liao, L. *et al.*, (2002). Genome comparisons based on profiles of metabolic pathways. *In Proceedings of the 6th International Conference on Knowledge-Based Intelligent Information and Engineering Systems*. Cream, Italy.

121. Linnaeus, C. (1758). *Systema naturae per regna tria naturae, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis*. *Holmiae*. 10, 1, 824 pp.

122. Lu, J., Fu, Y., Kumar, S., Shen, Y., Zeng, K., Xu, A., Carthew, R., Wu, C.I. (2008). Adaptive Evolution of Newly Emerged Micro-RNA Genes in *Drosophila*. *Molecular Biology and Evolution*, 25, 929-938.

123. Lytle, J. R., Yario, T. A., Steitz, J. A. (2007). Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5’ UTR as in the 3’ UTR. *Proceedings of the National Academy of Science*, 104, 9667-9672.

124. Ma, H. W., Zeng, A.P. (2003). The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics*, 19, 423-30.

125. Ma, H.-W. & Zeng, A.-P. (2004). Phylogenetic comparison of metabolic capacities of organisms at genome level. *Molecular Phylogenetics and Evolution*, 31, 204–213.

126. Mackey A.J., Pereira F.C.N., & Roos, D.S. (2006). GLEAN: improved eukaryotic gene prediction by tatistical consensus of gene evidence. Manuscript in draft.

127. Marchler-Bauer, A., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., Fong, J. H., Geer, L. Y., Geer, R. C., Gonzales, N. R., Gwadz, M., He, S., Hurwitz, D. I., Jackson, J. D., Ke, Z., Lanczycki, C. J., Liebert, C. A., Liu, C., Lu, F., Lu, S., Marchler, G. H., Mullokandov, M., Song, J. S., Tasneem, A., Thanki, N., Yamashita, R. A., Zhang, D., Zhang, N., & Bryant, S. H. (2009). CDD: specific functional annotation with the Conserved Domain Database. *Nucleic Acids Research*, 37, D205-D210.

128. Markow T, & O’Grady, P. (2005). *Drosophila: a guide to species identification and use*. Academic Press. 259 pp.

129. Martinez, N. J., Ow, M. C., Barrasa, M. I., Hammell, M., Sequerra, R., Doucette-Stamm, L., Roth, F. P., Ambros, V. R., Walhout, A. J. (2008). A *C. elegans* genome-scale microRNA network contains composite feedback motifs with high flux capacity. *Genes & Development*, 22, 2535-2549.

130. Matthews, L.R., Vaglio, P., Reboul, J., Ge, H., Davis, B.P., Garrels, J., Vincent, S., & Vidal, M. (2001). Identification of potential interaction networks using sequence-based searches for conserved protein–protein interactions or "interologs." *Genome Research*, 11, 2120–2126.

131. Maziere, P., & Enright, A. J. (2007). Prediction of microRNA targets. *Drug Discovery Today*, 12, 452-458.

132. Mazurie, A., Bonchev, D., Schwikowski, B., Buck, G. A. (2008). Phylogenetic distances are encoded in networks of interacting pathways. *Bioinformatics*, 24, 2579-2585.

133. McGregor, A. P., Orgogozo, V., Delon, I., Zanet, J., Srinivasan1, D. G., Payre, F. & Stern, D. L.1 Morphological evolution through multiple cis-regulatory mutations at a single gene. (2007). *Nature*, 448, 587-591.

134. McCarter, J.P., Bird, D. McK., & Mitreva, M. (2005) Nematode Gene Sequences: Update for December 2005. *Journal of Nematology*, 37, 417–421.

135. Megraw, M., Sethupathy, P., Corda, B., & Hatzigeorgiou, A. G. (2007). miRGen: a database for the study of animal microRNA genomic organization and function. *Nucleic Acids Research*, 2007, 35, D149-55.

136. Mertens, T. R., & Hammersmith, R. L. (2007). *Genetics Laboratory Investigations*. 13th Ed. MacMillan Publishing. Co, NY.

137. Mihaescu, R., Levy, D., & Pachter, L. (2006). "Why neighbor-joining works". *arXiv:cs/0602041v3*.

138. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U. (2002) Network Motifs: Simple Building Blocks of Complex Networks. *Science*, 298, 824-827.

139. Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R., Courcelle, E., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Griffith-Jones S., Haft, D., Hermjakob, H., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Orchard, S., Pagni, M., Peyruc, D., Ponting, C. P., Servant, F., Sigrist, C. J.; & InterPro Consortium. (2002). InterPro: an integrated documentation resource for protein families, domains and functional sites. *Briefings in Bioinformatics*, 3, 225-235.

140. Muller, H. J. (1940). Bearings of the *Drosophila* work on systematics. *The New Systematics*, 185-268. Oxford: Clarendon Press.

141. Muller, H. J. (1942). Isolating mechanisms, evolution, and temperature. *Biology Symposium*. 6, 71–125.

142. Nègre V, Hôtelier T, Volkoff AN, Gimenez S, Cousserans F, Mita K, Sabau X, Rocher J, López-Ferber M, d'Alençon E, Audant P, Sabourault C, Bidegainberry V, Hilliou F, Fournier P. (2006). SPODOBASE: an EST database for the lepidopteran crop pest *Spodoptera*. *BMC Bioinformatics*, 7, 322.

143. Nikitin, A., Egorov, S., Daraselia, N., & Mazo, I. (2003). Pathway studio—the analysis and navigation of molecular networks. *Bioinformatics*, 19, 1–3.

144. O'Brien, E. L. L, Remm M, & Sonnhammer, E. L. L. (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Research*, 33, D476–D480.

145. O'Brien, K. P., Westerlund, I., & Sonnhammer, E. L. (2004). OrthoDisease: a database of human disease orthologs. *Human Mutation*. 24, 112-119.

146. Ogata, H., Fujibuchi, W., Goto, S., & Kanehisa, M. (2000). A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Research*, 28, 4021–4028.

147. O'grady, P. M., J. B. Clark, & M. G. Kidwell. (1998). Phylogeny of the *Drosophila saltans* Species Group Based on Combined Analysis of Nuclear and Mitochondrial DNA Sequences. *Molecular Biology and Evolution*, 15, 656–664.

148. Oh, S. J., Joung, J. G., Chang, J. H., & Zhang, B. T. (2006). Construction of phylogenetic trees by kernel-based comparative analysis of metabolic networks. *BMC Bioinformatics*, 7, 284.

149. Olsen, P. H., & Ambros, V. (1999). The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation. *Developmental Biology*, 216, 671–680.

150. Orr, A. H. (1997). Haldane's rule. *Annual Review Ecology & Systematics*. 28, 195–218.

151. Paten, B., Herrero, J., Beal, K., Fitzgerald, S., Birney, E. (2008). Enredo and Pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Research*, 18, 1814-1828.

152. Pollard, D. A., Iyer, V. N., Moses, A. M., Eisen, M. B. (2006). Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. *PLoS Genetics*, 27, e173.

153. Posada, D., & Crandall, K.A. (1998). MODELTEST: testing the model of DNA substitution. *Bioinformatics*, 14, 817-818.

154. Prochnik, S. E., Rokhsar, D. S., & Aboobaker, A. A. (2007). Evidence for a microRNA expansion in the bilaterian ancestor. *Development Genes and Evolution*, 217, 73–77.
155. Rajewsky, N. (2006). MicroRNA target predictions in animals. *Nature Genetics*, 38, S8 - S13.
156. Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai Z. N., & Barabási, A.-L. Hierarchical organization of modularity in metabolic networks. *Science*, 2002, 297, 1551.
157. Remm, M., Storm, C. E. V., & Sonnhammer, E. L. L. (2001). Automatic Clustering of Orthologs and In-paralogs from Pairwise Species Comparisons. *Journal Molecular Biology*, 314, 1041–1052.
158. Remsen, J. & O'grady, P. (2002). Phylogeny of Drosophilinae (Diptera: Drosophilidae), with comments on combined analysis and character support. *Molecular Phylogenetics and Evolution*, 24, 249–264.
159. Robe, L. J., Valente, V. L. S., Budnik, M., & Loreto, E. L. S.. (2005). Molecular phylogeny of the subgenus *Drosophila* (Diptera, Drosophilidae) with an emphasis on Neotropical species and groups: A nuclear versus mitochondrial gene approach. *Molecular Phylogenetics and Evolution*, 36, 623–640.
160. Robins, H. and Press, W. H. (2005) Human microRNAs target a functionally distinct population of genes with AT-rich 3'UTRs. *Proceedings of the National Academy of Science*, USA, 102, 15557–15562.
161. Rodrigiuez-Trelles, F., Tarrio, R., & Ayala, F. J. Evidence for a High Ancestral GC Content in *Drosophila*. (2000) *Molecular Biology and Evolution*, 17, 1710–1717.
162. Rota-Stabelli, O., Campbell, L., Brinkmann, H., Edgecombe, G. D., Longhorn, S. J., Peterson, K. J., Pisani, D., Philippe, H., & Telford, M. J., (2010). A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. *Proceedings of the Royal Society Biological Sciences*. doi: 10.1098/rspb.2010.0590.
163. Ruan, J., Li, H., Chen, Z., Lachlan A. C., Coin, J. M., Guo1, Y., Hériché, J-K., Hu1, Y., Kristiansen, K., Li, R.., Liu, T., Moses, A., Qin, J., Vang, S., Vilella, A. J., Ureta-Vidal, A., Bolund, L., Wang, J., & Durbin, R. (2008). TreeFam: 2008 Update. *Nucleic Acids Research*, 36, D735-D740.
164. Saitou, N., & Nei, M.. (1987). "The neighbor-joining method: a new method for reconstructing phylogenetic trees". *Molecular Biology & Evolution*, 4, 406–425.
165. Sandberg,R., Neilson, J. R., Sarma, A., Sharp, P. A., & Burge, C. B.(2008). Proliferating Cells Express mRNAs with Shortened 3' Untranslated Regions and Fewer MicroRNA Target Sites. *Science*, 320, 1643
166. Sempere, L. F., Cole, C. N. McPeek, M. A, & Peterson, K.J. (2006). The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. *Journal of Experimental Zoology, (Molecular and Developmental Evolution)*, 306B, 1-14.
167. Sempere, L. F., Martinez, P., Cole, C., Baguñà, J., & Peterson, K. J. (2007). Phylogenetic distribution of microRNAs supports the basal position of acoel flatworms and the polyphyly of Platyhelminthes. *Evolution & Development*, 5, 409-415.
168. Sethupathy, P., Corda, B., &. Hatziegeorgiou, A. G. (2006). TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA*, 12, 192-197.
169. Shannon, C., & Weaver, W. (1949) *Mathematical Theory of Communications*. University of Illinois Press: Urbana, MI.
170. Sharan R. & Ideker T. (2006). Modeling cellular machinery through biological network comparison, *Nature Biotechnology* 24, 427-433.
171. Sharp, D. (1898) The Cambridge Natural History, Insects. vol. 2. Macmillan and Co., London.
172. Shedlock, A. M., & Okada, N. (2000). SINE insertions: Powerful tools for molecular systematics. *Bioessays*, 22, 148–160.
173. Slater, G. St. C. & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6, 31.
174. Smalheiser, N. R., Torvik, V. I. (2006). Alu elements within human mRNAs are probable microRNA targets. *Trends in Genetics*, 10, 532-536.
175. Smedley, D. Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., & Kasprzyk, A. (2009). BioMart - biological queries made easy. *BMC Genomics*, 10, 22.
176. Sneath, P. H. A., & Sokal, R. R. (1973). *Numerical taxonomy — The principles and practice of numerical classification*. W. H. Freeman, San Francisco. xv + 573 p.
177. Sonnhammer, E. L., Eddy, S. R., & Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*. 1997, 28, 405-420.
178. Stark, A., Brennecke, J., Bushati, N., Russell, R. B., & Cohen, S. M. (2005). Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell*, 123, 1133-46.
179. Stark, A., Brennecke, J., Russell, R. B., & Cohen, S. M. (2003). Identification of *Drosophila* MicroRNA targets, *PLoS Biology*, 1, E60

180. Stark, A., Kheradpour, P., Parts, L., Brennecke, J., Hodges, E., Hannon, G. J., Kellis, M. (2007). Systematic discovery and characterization of fly microRNAs using 12 *Drosophila* genomes. *Genome Research*, 12, 1865-1879.

181. Stark, A., Lin, M. F., Kheradpour, P., Pedersen, J. S., Parts, L., Carlson, J. W., Crosby, M. A., Rasmussen, M. D., Roy, S., Deoras, A. N., Ruby, J. G., Brennecke, J.; Harvard FlyBase curators; Berkeley *Drosophila* Genome Project, Hodges, E., Hinrichs, A. S., Caspi, A., Paten, B., Park, S. W., Han, M.V., Maeder, M. L., Polansky, B. J., Robson, B. E., Aerts, S., van Helden, J., Hassan, B., Gilbert, D. G., Eastman, D. A., Rice, M., Weir, M., Hahn, M. W., Park, Y., Dewey, C. N., Pachter, L., Kent, W. J., Haussler, D., Lai, E. C., Bartel, D. P., Hannon, G. J., Kaufman, T. C., Eisen, M. B., Clark, A. G., Smith, D., Celniker, S. E., Gelbart, W. M., &Kellis, M. (2007). Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature*, 450, 219-232.

182. Strauss, W. M., Chen, C., Lee, C. T., and Ridzon, D. (2006). Nonrestrictive developmental regulation of microRNA gene expression. *Mammal Genome*, 17, 833–840.

183. Stuart, J.M., Segal, E., Koller, D., and Kim, S.K. 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302, 249–255.

184. Sturtevant A. H. & Novitski, E. (1941) The homologies of the chromosome elements in the genus *Drosophila*. *Genetics*, 36, 517.

185. Sucena, E. & Stern, D. L. (2000). Divergence of larval morphology between *Drosophila sechellia* and its sibling species caused by cis-regulatory evolution of ovo/shaven-baby. *Proceedings of the National Academy of Sciences*, 97, 4530–4534.

186. Sueoka, N. (1988). Directional mutation pressure and neutral molecular evolution. *Proceedings of the National Academy of Science, USA*, 85, 2653–2657.

187. Sun Microsystems, Inc. (2008-2009). MySQL AB.

188. Suthram S., Sittler, T., Ideker, T. (2005). The *Plasmodium* protein network diverges from those of other Eukaryotes. *Nature*, 438, 108-112.

189. Swofford, D. L. (2002). *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*, Version 4. Sinauer Associates, Sunderland, Massachusetts.

190. Swofford, D. L., Thorne, J. L., Felsenstein, J., & Wiegmann, B. M. (1996). The topology dependent permutation test for monophyly does not test for monophyly. *Systematic Biology*, 46, 575-579.

191. Swofford, D. L., Waddell, P. J., Huelsenbeck, J. P., Foster, P. G., Paul O. Lewis, P. O., & Rogers, J. S. (2001). Bias in Phylogenetic Estimation and Its Relevance to the Choice between Parsimony and Likelihood Methods. *Systematic Biology*, 50, 525–539.

192. Tamura, K., Subramanian, S., & Kumar, S. Temporal Patterns of Fruit Fly (*Drosophila*) Evolution Revealed by Mutation Clocks. (2004). *Molecular Biology and Evolution,* 21, 36–44.

193. Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan S., Wolf, Y. I., Yin, J. J., & Natale, D.A. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4, 41.

194. Trut, L. N. (1999) Early Canid Domestication: The Farm-Fox Experiment. *American Scientist*, 87, 160-169.

195. Tsang, J., Zhu, J., & van Oudenaarden A. (2007). MicroRNA-Mediated Feedback and Feed forward Loops Are Recurrent Network Motifs in Mammals. *Molecular Cell*, 26, 753-67.

196. Wang, X., Gu, J., Zhang, M. Q., & Li, Y. (2008). Identification of phylogenetically conserved microRNA cis-regulatory elements across 12 *Drosophila* species. *Bioinformatics*, 24, 165-171.

197. Wernicke S., & Rasche, F. (2006). FANMOD: a tool for fast network motif detection. *Bioinformatics,* 22, 1152–1153.

198. Wheeler, D. L., Barrett, T., Benson, D. A., *et al.*, (2006). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 34, D173–D180.

199. Wilson, R.J., Goodman, J.L., Strelets, V.B., & the FlyBase Consortium. (2008). FlyBase: integration and improvements to query tools. *Nucleic Acids Research*, 36, D588-D593.

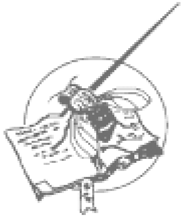
200. Wu, C. H., Apweiler, R., Bairoch, A., Natale, D. A., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Mazumder, R., O'Donovan, C., Redaschi, N., & Suzek B. (2006). The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Research*, 34, D187-D191.

201. Xie, X., Lu, J., Kulbokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S., & Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3'UTRs by comparison of several mammals. *Nature*, 434, 338–345.

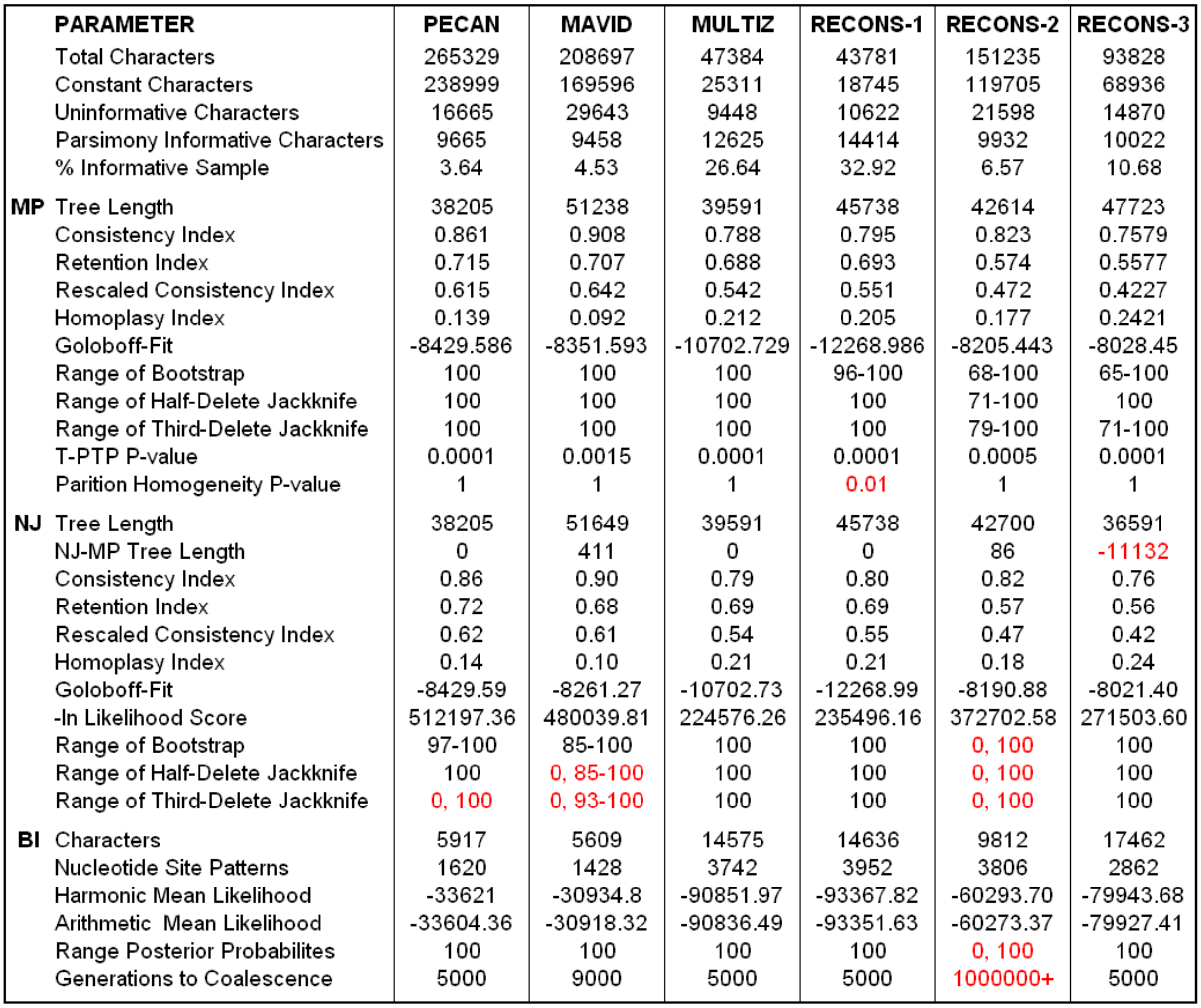
202. Yan L., Wang, F., Lee, J-A., & Gao, F-B. (2006). *MicroRNA-9a* ensures the precise specification of sensory organ precursors in *Drosophila. Genes & Development*, 20, 2793-2805.

203. Yu, X., Lin, J., Zack, D. J., Mendell J. T, & Qian J. (2008). Analysis of regulatory network topology reveals functionally distinct classes of microRNAs. *Nucleic Acids Research*, 36, 6494–6503.

204. Zhang, Y., Li, S., Skogerbø, G., Zhang, Z., Zhu, X., Zhang, Z., Sun, S., Lu, H., Shi, B., & Chen, R. (2006). Phylophenetic properties of metabolic pathway topologies as revealed by global analysis. *BMC Bioinformatics*, 7, 252.



APPENDIX I. Selection of Multiple Sequence Alignment using Criteria from Phylogenetic Reconstruction



	PARAMETER	PECAN	MAVID	MULTIZ	RECONS-1	RECONS-2	RECONS-3
	Total Characters	265329	208697	47384	43781	151235	93828
	Constant Characters	238999	169596	25311	18745	119705	68936
	Uninformative Characters	16665	29643	9448	10622	21598	14870
	Parsimony Informative Characters	9665	9458	12625	14414	9932	10022
	% Informative Sample	3.64	4.53	26.64	32.92	6.57	10.68
MP	Tree Length	38205	51238	39591	45738	42614	47723
	Consistency Index	0.861	0.908	0.788	0.795	0.823	0.7579
	Retention Index	0.715	0.707	0.688	0.693	0.574	0.5577
	Rescaled Consistency Index	0.615	0.642	0.542	0.551	0.472	0.4227
	Homoplasy Index	0.139	0.092	0.212	0.205	0.177	0.2421
	Goloboff-Fit	-8429.586	-8351.593	-10702.729	-12268.986	-8205.443	-8028.45
	Range of Bootstrap	100	100	100	96-100	68-100	65-100
	Range of Half-Delete Jackknife	100	100	100	100	71-100	100
	Range of Third-Delete Jackknife	100	100	100	100	79-100	71-100
	T-PTP P-value	0.0001	0.0015	0.0001	0.0001	0.0005	0.0001
	Parition Homogeneity P-value	1	1	1	0.01	1	1
NJ	Tree Length	38205	51649	39591	45738	42700	36591
	NJ-MP Tree Length	0	411	0	0	86	-11132
	Consistency Index	0.86	0.90	0.79	0.80	0.82	0.76
	Retention Index	0.72	0.68	0.69	0.69	0.57	0.56
	Rescaled Consistency Index	0.62	0.61	0.54	0.55	0.47	0.42
	Homoplasy Index	0.14	0.10	0.21	0.21	0.18	0.24
	Goloboff-Fit	-8429.59	-8261.27	-10702.73	-12268.99	-8190.88	-8021.40
	-ln Likelihood Score	512197.36	480039.81	224576.26	235496.16	372702.58	271503.60
	Range of Bootstrap	97-100	85-100	100	100	0, 100	100
	Range of Half-Delete Jackknife	100	0, 85-100	100	100	0, 100	100
	Range of Third-Delete Jackknife	0, 100	0, 93-100	100	100	0, 100	100
BI	Characters	5917	5609	14575	14636	9812	17462
	Nucleotide Site Patterns	1620	1428	3742	3952	3806	2862
	Harmonic Mean Likelihood	-33621	-30934.8	-90851.97	-93367.82	-60293.70	-79943.68
	Arithmetic Mean Likelihood	-33604.36	-30918.32	-90836.49	-93351.63	-60273.37	-79927.41
	Range Posterior Probabilites	100	100	100	100	0, 100	100
	Generations to Coalescence	5000	9000	5000	5000	1000000+	5000

TABLE 8. Selection of Multiple Sequence Alignment using Criteria from Phylogenetic Reconstruction. Selection criteria used to eliminate an alignment are colored in red. Novel reconstructions are abbreviated respectively: RECONS-1) re-alignment of MULTIZ sequence data using the multiple sequence alignment program MAFFT; RECONS-2) one-step three-way reconciliation of PECAN, MAVID, and MULTIZ alignments with production of a consensus sequence; and RECONS-3) species-by-species three-way reconciliation and consensus of the later three published alignments (Dewey, 2007; Katoh & Toh, 2008; Kent *et al.*, 2002; Paten *et al.*, 2008; Stark *et al.*, 2007b). Phylogenetic methods are abbreviated respectively: BI) Bayesian Inference under general time reversible model with gamma distributed rates and invariant sites;

MP) standard parsimony; NJ) neighbor joining under general time reversible model with gamma distributed rates and invariant sites; and T-PTP) topology-dependent permutation test against the established *Drosophila* phylogeny (Faith, 1991; Huelsenbeck & Ronquist, 2005; Swofford, 2002). Ranges of bootstrap, jackknife, and posterior probability refer to support refer to tree topology for the drosophilid reference phylogeny (CHAPTER III, FIGURE 22).

APPENDIX II. Example MicroRNA-Target Interactome Network

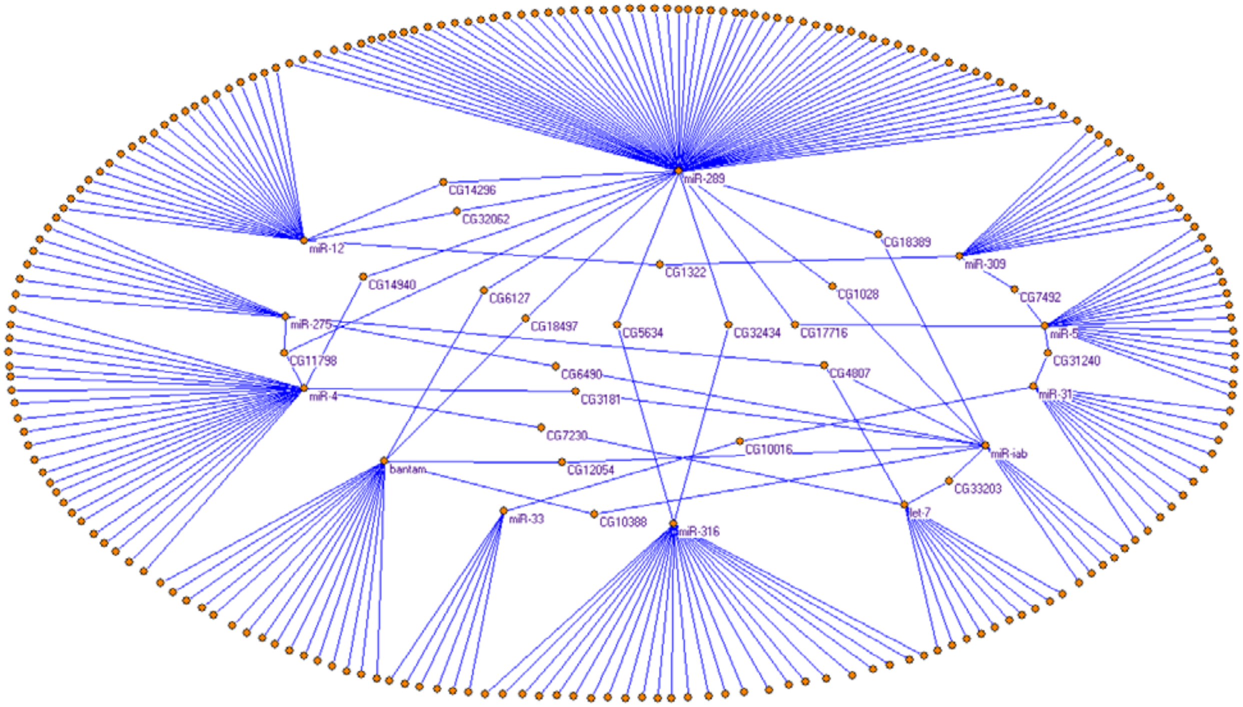
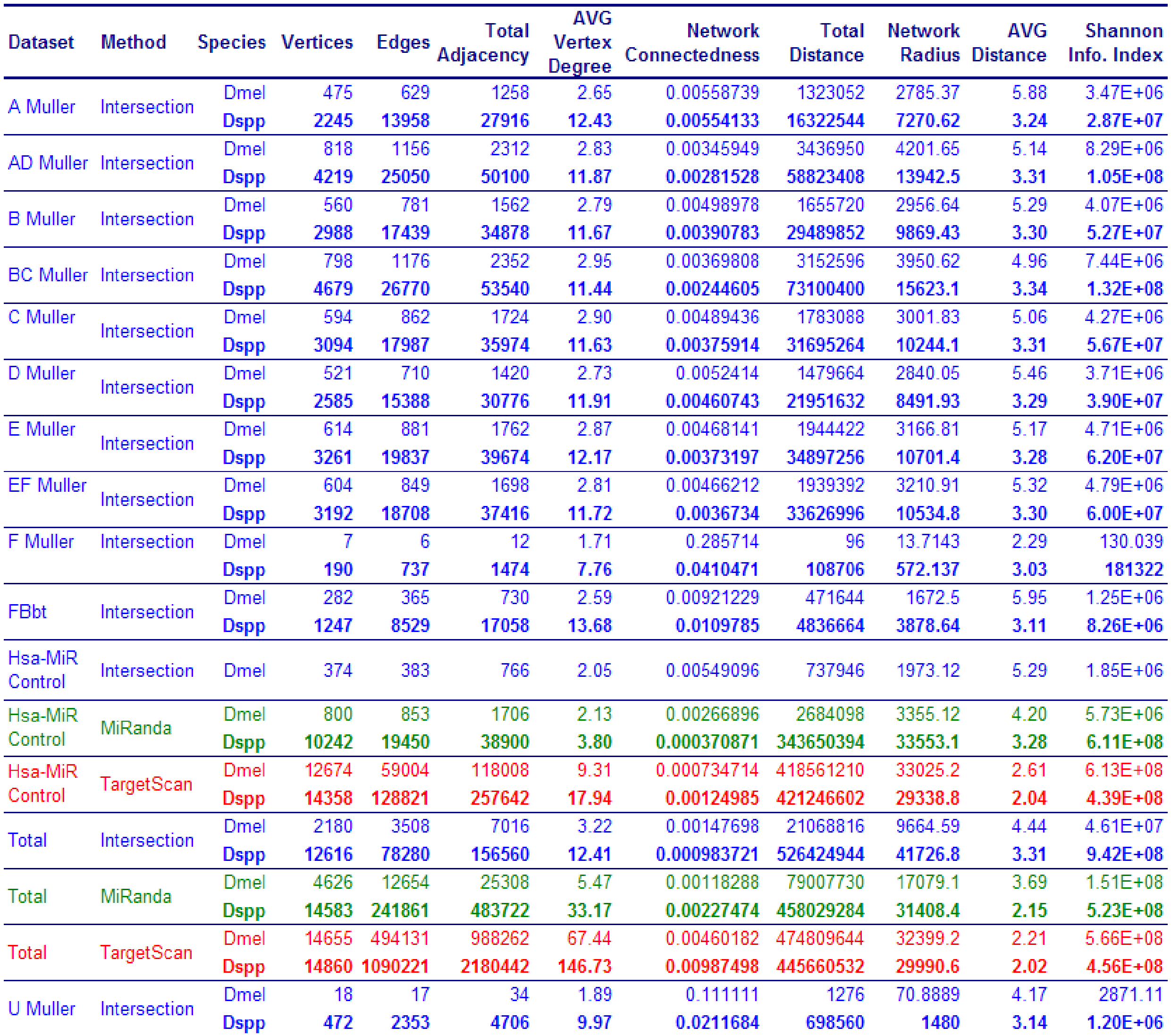


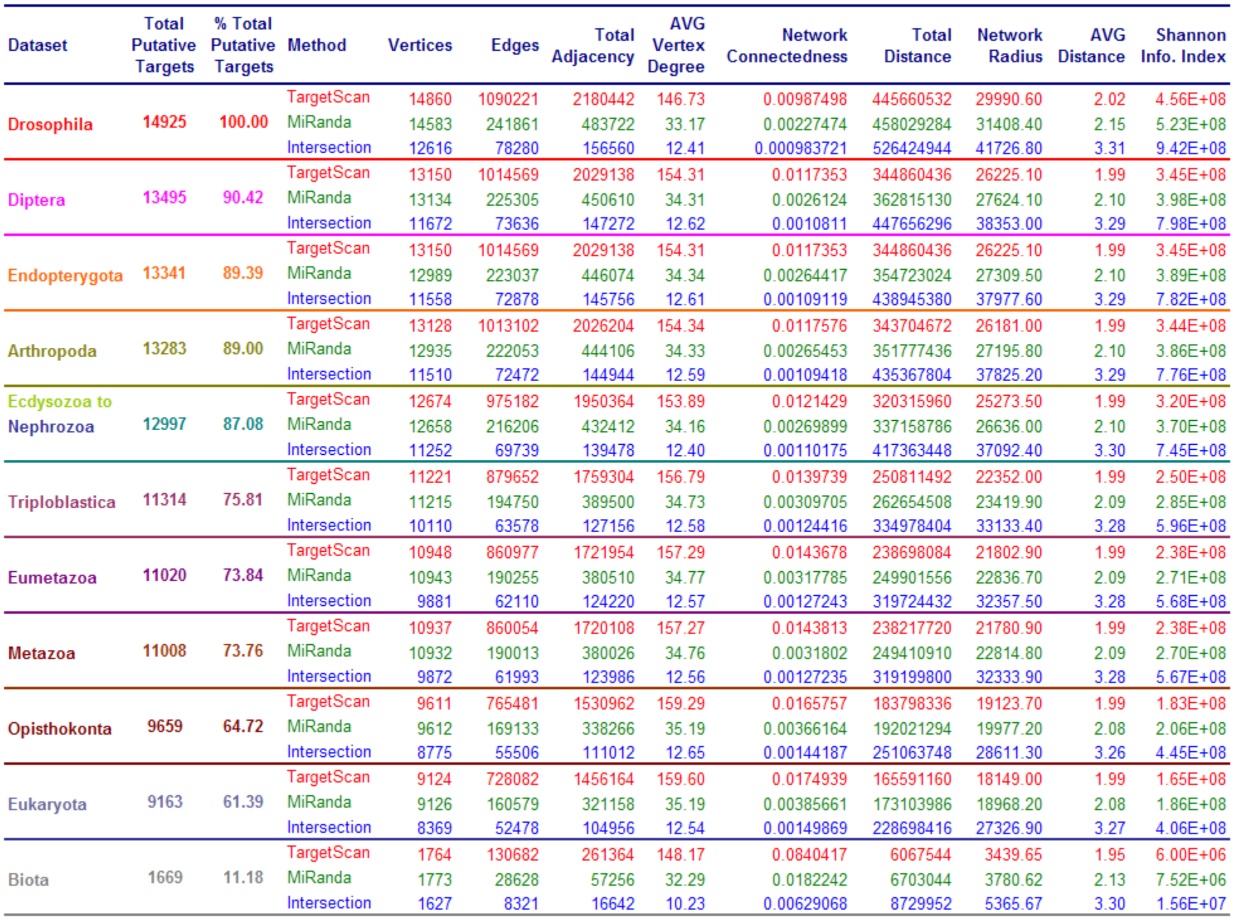
FIGURE 37. Example Interactome Network of MicroRNA-Target Interactions representing 1.66% of the total microRNA target dataset. This network includes 248 nodes and represents the intersection (overlap) of two datasets for 12 microRNAs (Enright *et al.*, 2003; Grün, *et al.*, 2005). Only microRNAs (designated by miR-#) and targets (designated with CG#) predicted to interact with multiple microRNAs are labeled; unlabeled nodes represent predicted targets to a single microRNA

APPENDIX III. Network Adjacency, Distance Quantification, and Parametric Score Tables for Select MicroRNA-Target Networks of *Drosophila*



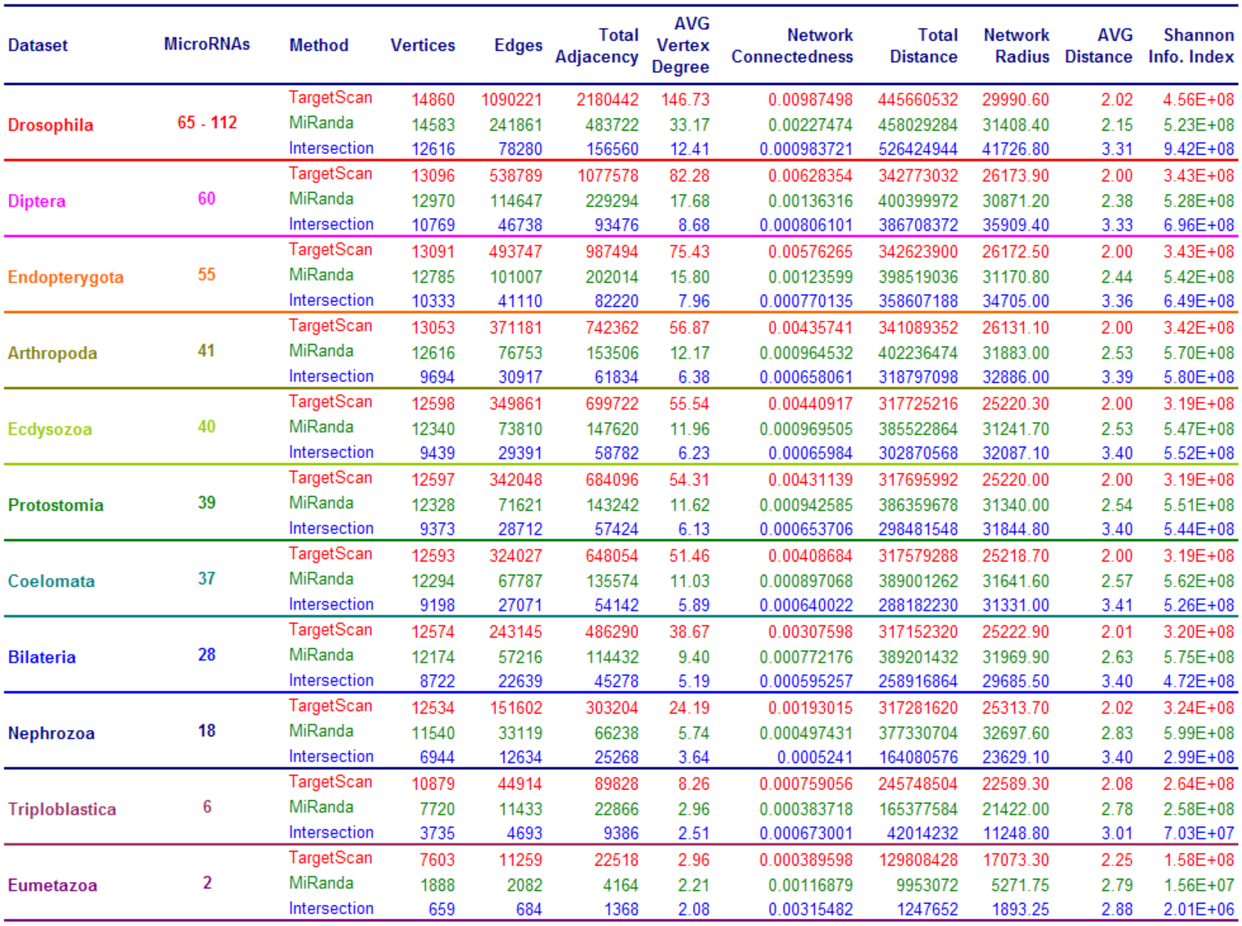
Dataset	Method	Species	Vertices	Edges	Total Adjacency	AVG Vertex Degree	Network Connectedness	Total Distance	Network Radius	AVG Distance	Shannon Info. Index
A Muller	Intersection	Dmel	475	629	1258	2.65	0.00558739	1323052	2785.37	5.88	3.47E+06
		Dspp	**2245**	**13958**	**27916**	**12.43**	**0.00554133**	**16322544**	**7270.62**	**3.24**	**2.87E+07**
AD Muller	Intersection	Dmel	818	1156	2312	2.83	0.00345949	3436950	4201.65	5.14	8.29E+06
		Dspp	**4219**	**25050**	**50100**	**11.87**	**0.00281528**	**58823408**	**13942.5**	**3.31**	**1.05E+08**
B Muller	Intersection	Dmel	560	781	1562	2.79	0.00498978	1655720	2956.64	5.29	4.07E+06
		Dspp	**2988**	**17439**	**34878**	**11.67**	**0.00390783**	**29489852**	**9869.43**	**3.30**	**5.27E+07**
BC Muller	Intersection	Dmel	798	1176	2352	2.95	0.00369808	3152596	3950.62	4.96	7.44E+06
		Dspp	**4679**	**26770**	**53540**	**11.44**	**0.00244605**	**73100400**	**15623.1**	**3.34**	**1.32E+08**
C Muller	Intersection	Dmel	594	862	1724	2.90	0.00489436	1783088	3001.83	5.06	4.27E+06
		Dspp	**3094**	**17987**	**35974**	**11.63**	**0.00375914**	**31695264**	**10244.1**	**3.31**	**5.67E+07**
D Muller	Intersection	Dmel	521	710	1420	2.73	0.0052414	1479664	2840.05	5.46	3.71E+06
		Dspp	**2585**	**15388**	**30776**	**11.91**	**0.00460743**	**21951632**	**8491.93**	**3.29**	**3.90E+07**
E Muller	Intersection	Dmel	614	881	1762	2.87	0.00468141	1944422	3166.81	5.17	4.71E+06
		Dspp	**3261**	**19837**	**39674**	**12.17**	**0.00373197**	**34897256**	**10701.4**	**3.28**	**6.20E+07**
EF Muller	Intersection	Dmel	604	849	1698	2.81	0.00466212	1939392	3210.91	5.32	4.79E+06
		Dspp	**3192**	**18708**	**37416**	**11.72**	**0.0036734**	**33626996**	**10534.8**	**3.30**	**6.00E+07**
F Muller	Intersection	Dmel	7	6	12	1.71	0.285714	96	13.7143	2.29	130.039
		Dspp	**190**	**737**	**1474**	**7.76**	**0.0410471**	**108706**	**572.137**	**3.03**	**181322**
FBbt	Intersection	Dmel	282	365	730	2.59	0.00921229	471644	1672.5	5.95	1.25E+06
		Dspp	**1247**	**8529**	**17058**	**13.68**	**0.0109785**	**4836664**	**3878.64**	**3.11**	**8.26E+06**
Hsa-MiR Control	Intersection	Dmel	374	383	766	2.05	0.00549096	737946	1973.12	5.29	1.85E+06
Hsa-MiR Control	MiRanda	Dmel	800	853	1706	2.13	0.00266896	2684098	3355.12	4.20	5.73E+06
		Dspp	**10242**	**19450**	**38900**	**3.80**	**0.000370871**	**343650394**	**33553.1**	**3.28**	**6.11E+08**
Hsa-MiR Control	TargetScan	Dmel	12674	59004	118008	9.31	0.000734714	418561210	33025.2	2.61	6.13E+08
		Dspp	**14358**	**128821**	**257642**	**17.94**	**0.00124985**	**421246602**	**29338.8**	**2.04**	**4.39E+08**
Total	Intersection	Dmel	2180	3508	7016	3.22	0.00147698	21068816	9664.59	4.44	4.61E+07
		Dspp	**12616**	**78280**	**156560**	**12.41**	**0.000983721**	**526424944**	**41726.8**	**3.31**	**9.42E+08**
Total	MiRanda	Dmel	4626	12654	25308	5.47	0.00118288	79007730	17079.1	3.69	1.51E+08
		Dspp	**14583**	**241861**	**483722**	**33.17**	**0.00227474**	**458029284**	**31408.4**	**2.15**	**5.23E+08**
Total	TargetScan	Dmel	14655	494131	988262	67.44	0.00460182	474809644	32399.2	2.21	5.66E+08
		Dspp	**14860**	**1090221**	**2180442**	**146.73**	**0.00987498**	**445660532**	**29990.6**	**2.02**	**4.56E+08**
U Muller	Intersection	Dmel	18	17	34	1.89	0.111111	1276	70.8889	4.17	2871.11
		Dspp	**472**	**2353**	**4706**	**9.97**	**0.0211684**	**698560**	**1480**	**3.14**	**1.20E+06**

TABLE 9. Network descriptors for adjacency and distance quantification in select microRNA-target networks. The network quantification and distance analyses were performed using in-house GRAFMAN software available under Linux on the Watson supercomputer cluster of Virginia Commonwealth University (Karabunarliev & Bonchev, 2002). Datasets are presented in alphabetical order and subgrouped separately for the union of twelve *Drosophila* species (Dspp) and *D. melanogaster* alone (Dmel). The union of twelve *Drosophila* species appears in bold. Rows are further colored according to target prediction method with MiRanda in green, TargetsScan in red, and the network intersection of methods in blue.



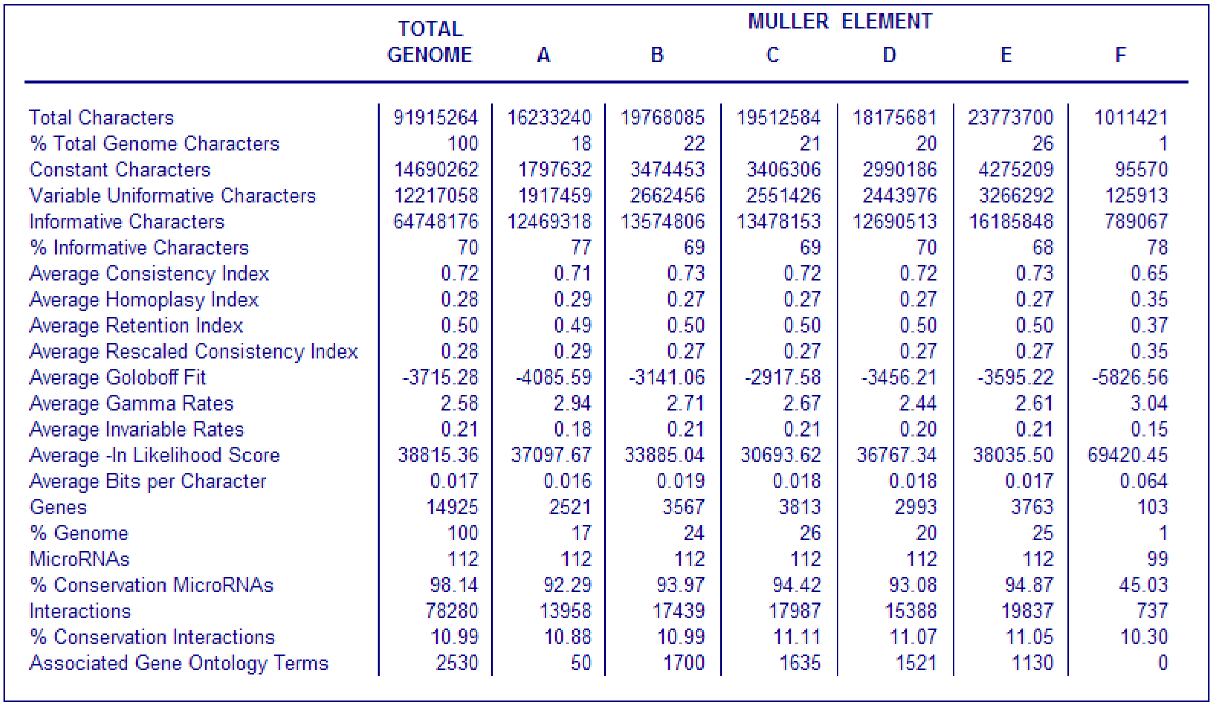
Dataset	Total Putative Targets	% Total Putative Targets	Method	Vertices	Edges	Total Adjacency	AVG Vertex Degree	Network Connectedness	Total Distance	Network Radius	AVG Distance	Shannon Info. Index
Drosophila	14925	100.00	TargetScan	14860	1090221	2180442	146.73	0.00987498	445660532	29990.60	2.02	4.56E+08
			MiRanda	14583	241861	483722	33.17	0.00227474	458029284	31408.40	2.15	5.23E+08
			Intersection	12616	78280	156560	12.41	0.000983721	526424944	41726.80	3.31	9.42E+08
Diptera	13495	90.42	TargetScan	13150	1014569	2029138	154.31	0.0117353	344860436	26225.10	1.99	3.45E+08
			MiRanda	13134	225305	450610	34.31	0.0026124	362815130	27624.10	2.10	3.98E+08
			Intersection	11672	73636	147272	12.62	0.0010811	447656296	38353.00	3.29	7.98E+08
Endopterygota	13341	89.39	TargetScan	13150	1014569	2029138	154.31	0.0117353	344860436	26225.10	1.99	3.45E+08
			MiRanda	12989	223037	446074	34.34	0.00264417	354723024	27309.50	2.10	3.89E+08
			Intersection	11558	72878	145756	12.61	0.00109119	438945380	37977.60	3.29	7.82E+08
Arthropoda	13283	89.00	TargetScan	13128	1013102	2026204	154.34	0.0117576	343704672	26181.00	1.99	3.44E+08
			MiRanda	12935	222053	444106	34.33	0.00265453	351777436	27195.80	2.10	3.86E+08
			Intersection	11510	72472	144944	12.59	0.00109418	435367804	37825.20	3.29	7.76E+08
Ecdysozoa to Nephrozoa	12997	87.08	TargetScan	12674	975182	1950364	153.89	0.0121429	320315960	25273.50	1.99	3.20E+08
			MiRanda	12658	216206	432412	34.16	0.00269899	337158786	26636.00	2.10	3.70E+08
			Intersection	11252	69739	139478	12.40	0.00110175	417363448	37092.40	3.30	7.45E+08
Triploblastica	11314	75.81	TargetScan	11221	879652	1759304	156.79	0.0139739	250811492	22352.00	1.99	2.50E+08
			MiRanda	11215	194750	389500	34.73	0.00309705	262654508	23419.90	2.09	2.85E+08
			Intersection	10110	63578	127156	12.58	0.00124416	334978404	33133.40	3.28	5.96E+08
Eumetazoa	11020	73.84	TargetScan	10948	860977	1721954	157.29	0.0143678	238698084	21802.90	1.99	2.38E+08
			MiRanda	10943	190255	380510	34.77	0.00317785	249901556	22836.70	2.09	2.71E+08
			Intersection	9881	62110	124220	12.57	0.00127243	319724432	32357.50	3.28	5.68E+08
Metazoa	11008	73.76	TargetScan	10937	860054	1720108	157.27	0.0143813	238217720	21780.90	1.99	2.38E+08
			MiRanda	10932	190013	380026	34.76	0.0031802	249410910	22814.80	2.09	2.70E+08
			Intersection	9872	61993	123986	12.56	0.00127235	319199800	32333.90	3.28	5.67E+08
Opisthokonta	9659	64.72	TargetScan	9611	765481	1530962	159.29	0.0165757	183798336	19123.70	1.99	1.83E+08
			MiRanda	9612	169133	338266	35.19	0.00366164	192021294	19977.20	2.08	2.06E+08
			Intersection	8775	55506	111012	12.65	0.00144187	251063748	28611.30	3.26	4.45E+08
Eukaryota	9163	61.39	TargetScan	9124	728082	1456164	159.60	0.0174939	165591160	18149.00	1.99	1.65E+08
			MiRanda	9126	160579	321158	35.19	0.00385661	173103986	18968.20	2.08	1.86E+08
			Intersection	8369	52478	104956	12.54	0.00149869	228698416	27326.90	3.27	4.06E+08
Biota	1669	11.18	TargetScan	1764	130682	261364	148.17	0.0840417	6067544	3439.65	1.95	6.00E+06
			MiRanda	1773	28628	57256	32.29	0.0182242	6703044	3780.62	2.13	7.52E+06
			Intersection	1627	8321	16642	10.23	0.00629068	8729952	5365.67	3.30	1.56E+07

TABLE 10. **Network descriptors for adjacency and distance quantification in for taxonomic rank specific target variable subnetworks of MiRanda microRNA-target network data**. These network data represent the expansion of putative target genes through taxonomic hierarchy where networks at each rank were produced with 112 microRNA families held constant. The network quantification and distance analyses were performed using in-house GRAFMAN software available under Linux on the Watson supercomputer cluster of Virginia Commonwealth University (Karabunarliev & Bonchev, 2002). Dataset labels are color-coded to match the schema of selected taxonomic ranks presented in CHAPTER V, FIGURE 29. Rows are further subgrouped separately and colored according to target prediction method with MiRanda in green, TargetsScan in red, and the network intersection of methods in blue.

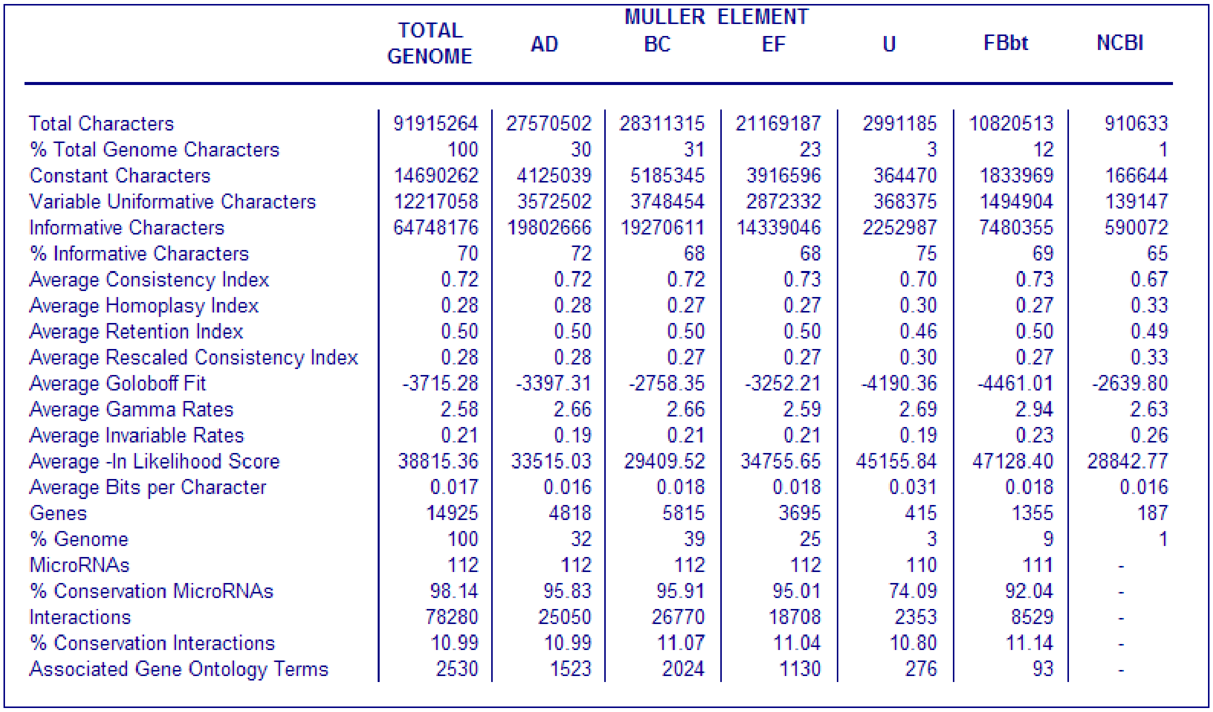


Dataset	MicroRNAs	Method	Vertices	Edges	Total Adjacency	AVG Vertex Degree	Network Connectedness	Total Distance	Network Radius	AVG Distance	Shannon Info. Index
Drosophila	65 - 112	TargetScan	14860	1090221	2180442	146.73	0.00987498	445660532	29990.60	2.02	4.56E+08
		MiRanda	14583	241861	483722	33.17	0.00227474	458029284	31408.40	2.15	5.23E+08
		Intersection	12616	78280	156560	12.41	0.000983721	526424944	41726.80	3.31	9.42E+08
Diptera	60	TargetScan	13096	538789	1077578	82.28	0.00628354	342773032	26173.90	2.00	3.43E+08
		MiRanda	12970	114647	229294	17.68	0.00136316	400399972	30871.20	2.38	5.28E+08
		Intersection	10769	46738	93476	8.68	0.000806101	386708372	35909.40	3.33	6.96E+08
Endopterygota	55	TargetScan	13091	493747	987494	75.43	0.00576265	342623900	26172.50	2.00	3.43E+08
		MiRanda	12785	101007	202014	15.80	0.00123599	398519036	31170.80	2.44	5.42E+08
		Intersection	10333	41110	82220	7.96	0.000770135	358607188	34705.00	3.36	6.49E+08
Arthropoda	41	TargetScan	13053	371181	742362	56.87	0.00435741	341089352	26131.10	2.00	3.42E+08
		MiRanda	12616	76753	153506	12.17	0.000964532	402236474	31883.00	2.53	5.70E+08
		Intersection	9694	30917	61834	6.38	0.000658061	318797098	32886.00	3.39	5.80E+08
Ecdysozoa	40	TargetScan	12598	349861	699722	55.54	0.00440917	317725216	25220.30	2.00	3.19E+08
		MiRanda	12340	73810	147620	11.96	0.000969505	385522864	31241.70	2.53	5.47E+08
		Intersection	9439	29391	58782	6.23	0.00065984	302870568	32087.10	3.40	5.52E+08
Protostomia	39	TargetScan	12597	342048	684096	54.31	0.00431139	317695992	25220.00	2.00	3.19E+08
		MiRanda	12328	71621	143242	11.62	0.000942585	386359678	31340.00	2.54	5.51E+08
		Intersection	9373	28712	57424	6.13	0.000653706	298481548	31844.80	3.40	5.44E+08
Coelomata	37	TargetScan	12593	324027	648054	51.46	0.00408684	317579288	25218.70	2.00	3.19E+08
		MiRanda	12294	67787	135574	11.03	0.000897068	389001262	31641.60	2.57	5.62E+08
		Intersection	9198	27071	54142	5.89	0.000640022	288182230	31331.00	3.41	5.26E+08
Bilateria	28	TargetScan	12574	243145	486290	38.67	0.00307598	317152320	25222.90	2.01	3.20E+08
		MiRanda	12174	57216	114432	9.40	0.000772176	389201432	31969.90	2.63	5.75E+08
		Intersection	8722	22639	45278	5.19	0.000595257	258916864	29685.50	3.40	4.72E+08
Nephrozoa	18	TargetScan	12534	151602	303204	24.19	0.00193015	317281620	25313.70	2.02	3.24E+08
		MiRanda	11540	33119	66238	5.74	0.000497431	377330704	32697.60	2.83	5.99E+08
		Intersection	6944	12634	25268	3.64	0.0005241	164080576	23629.10	3.40	2.99E+08
Triploblastica	6	TargetScan	10879	44914	89828	8.26	0.000759056	245748504	22589.30	2.08	2.64E+08
		MiRanda	7720	11433	22866	2.96	0.000383718	165377584	21422.00	2.78	2.58E+08
		Intersection	3735	4693	9386	2.51	0.000673001	42014232	11248.80	3.01	7.03E+07
Eumetazoa	2	TargetScan	7603	11259	22518	2.96	0.000389598	129808428	17073.30	2.25	1.58E+08
		MiRanda	1888	2082	4164	2.21	0.00116879	9953072	5271.75	2.79	1.56E+07
		Intersection	659	684	1368	2.08	0.00315482	1247652	1893.25	2.88	2.01E+06

TABLE 11. Network descriptors for adjacency and distance quantification for taxonomic rank specific microRNA variable subnetworks of MiRanda microRNA-target network data. These network data represent the expansion of the microRNA repertoire through taxonomic hierarchy where a total pool of 14925 available targets was held constant. The network quantification and distance analyses were performed using in-house GRAFMAN software available under Linux on the Watson supercomputer cluster of Virginia Commonwealth University (Karabunarliev & Bonchev, 2002). Dataset labels are color-coded to match the schema of selected taxonomic ranks presented in CHAPTER V, FIGURE 29. Rows are further subgrouped separately and colored according to target prediction method with MiRanda in green, TargetsScan in red, and the network intersection of methods in blue.



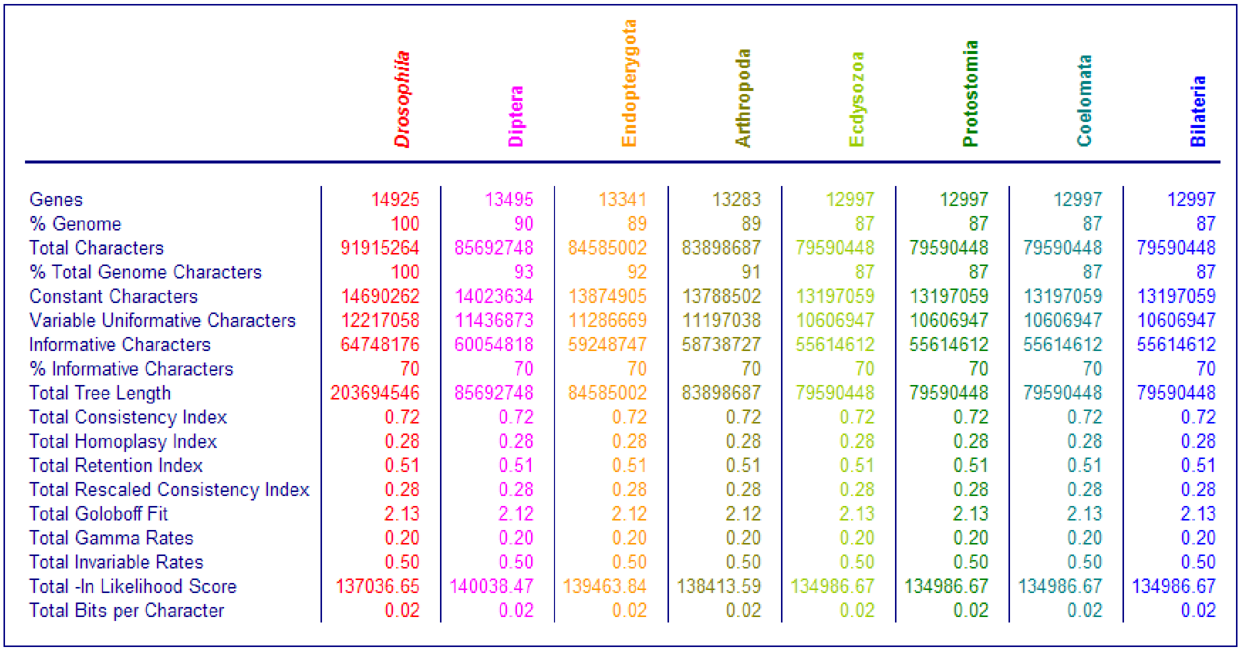
	TOTAL	MULLER ELEMENT					
	GENOME	A	B	C	D	E	F
Total Characters	91915264	16233240	19768085	19512584	18175681	23773700	1011421
% Total Genome Characters	100	18	22	21	20	26	1
Constant Characters	14690262	1797632	3474453	3406306	2990186	4275209	95570
Variable Uniformative Characters	12217058	1917459	2662456	2551426	2443976	3266292	125913
Informative Characters	64748176	12469318	13574806	13478153	12690513	16185848	789067
% Informative Characters	70	77	69	69	70	68	78
Average Consistency Index	0.72	0.71	0.73	0.72	0.72	0.73	0.65
Average Homoplasy Index	0.28	0.29	0.27	0.27	0.27	0.27	0.35
Average Retention Index	0.50	0.49	0.50	0.50	0.50	0.50	0.37
Average Rescaled Consistency Index	0.28	0.29	0.27	0.27	0.27	0.27	0.35
Average Goloboff Fit	-3715.28	-4085.59	-3141.06	-2917.58	-3456.21	-3595.22	-5826.56
Average Gamma Rates	2.58	2.94	2.71	2.67	2.44	2.61	3.04
Average Invariable Rates	0.21	0.18	0.21	0.21	0.20	0.21	0.15
Average -ln Likelihood Score	38815.36	37097.67	33885.04	30693.62	36767.34	38035.50	69420.45
Average Bits per Character	0.017	0.016	0.019	0.018	0.018	0.017	0.064
Genes	14925	2521	3567	3813	2993	3763	103
% Genome	100	17	24	26	20	25	1
MicroRNAs	112	112	112	112	112	112	99
% Conservation MicroRNAs	98.14	92.29	93.97	94.42	93.08	94.87	45.03
Interactions	78280	13958	17439	17987	15388	19837	737
% Conservation Interactions	10.99	10.88	10.99	11.11	11.07	11.05	10.30
Associated Gene Ontology Terms	2530	50	1700	1635	1521	1130	0



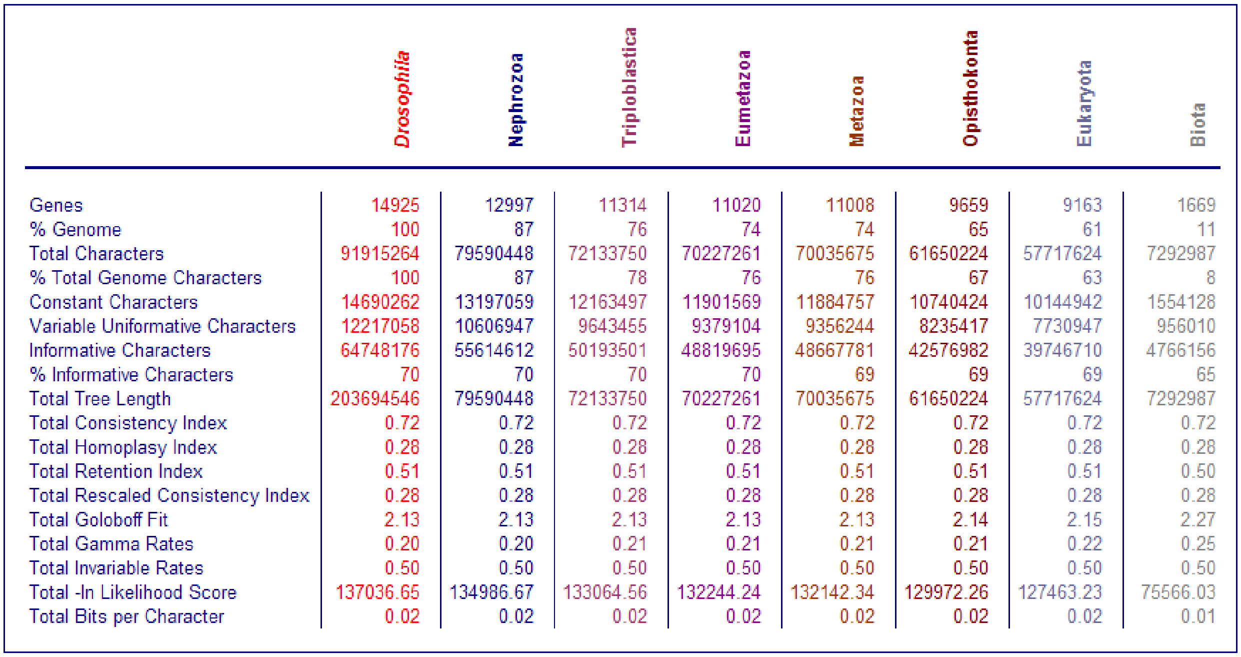
	TOTAL	MULLER ELEMENT					
	GENOME	AD	BC	EF	U	FBbt	NCBI
Total Characters	91915264	27570502	28311315	21169187	2991185	10820513	910633
% Total Genome Characters	100	30	31	23	3	12	1
Constant Characters	14690262	4125039	5185345	3916596	364470	1833969	166644
Variable Uniformative Characters	12217058	3572502	3748454	2872332	368375	1494904	139147
Informative Characters	64748176	19802666	19270611	14339046	2252987	7480355	590072
% Informative Characters	70	72	68	68	75	69	65
Average Consistency Index	0.72	0.72	0.72	0.73	0.70	0.73	0.67
Average Homoplasy Index	0.28	0.28	0.27	0.27	0.30	0.27	0.33
Average Retention Index	0.50	0.50	0.50	0.50	0.46	0.50	0.49
Average Rescaled Consistency Index	0.28	0.28	0.27	0.27	0.30	0.27	0.33
Average Goloboff Fit	-3715.28	-3397.31	-2758.35	-3252.21	-4190.36	-4461.01	-2639.80
Average Gamma Rates	2.58	2.66	2.66	2.59	2.69	2.94	2.63
Average Invariable Rates	0.21	0.19	0.21	0.21	0.19	0.23	0.26
Average -ln Likelihood Score	38815.36	33515.03	29409.52	34755.65	45155.84	47128.40	28842.77
Average Bits per Character	0.017	0.016	0.018	0.018	0.031	0.018	0.016
Genes	14925	4818	5815	3695	415	1355	187
% Genome	100	32	39	25	3	9	1
MicroRNAs	112	112	112	112	110	111	-
% Conservation MicroRNAs	98.14	95.83	95.91	95.01	74.09	92.04	-
Interactions	78280	25050	26770	18708	2353	8529	-
% Conservation Interactions	10.99	10.99	11.07	11.04	10.80	11.14	-
Associated Gene Ontology Terms	2530	1523	2024	1130	276	93	-

TABLE 12. Average parametric scores per gene from molecular phylogeny of targets represented in Muller element and FBbt data. Parametric scores were calculated through PAUP* where molecular phylogeny was constrained to the topology of the reference tree (CHAPTER III, FIGURE 22; Swofford, 2002). Likelihood scores were calculated under a under

a under a general time reversible sequence evolution model with gamma rate variation and invariable sites (GTR+I+G). The NCBI dataset represents 155 molecular markers taken from 590 operational taxonomic units (OTUs, typically species) available through Genbank (187 FlyBase records available, 1% total dataset).



	Drosophila	Diptera	Endopterygota	Arthropoda	Ecdysozoa	Protostomia	Coelomata	Bilateria
Genes	14925	13495	13341	13283	12997	12997	12997	12997
% Genome	100	90	89	89	87	87	87	87
Total Characters	91915264	85692748	84585002	83898687	79590448	79590448	79590448	79590448
% Total Genome Characters	100	93	92	91	87	87	87	87
Constant Characters	14690262	14023634	13874905	13788502	13197059	13197059	13197059	13197059
Variable Uniformative Characters	12217058	11436873	11286669	11197038	10606947	10606947	10606947	10606947
Informative Characters	64748176	60054818	59248747	58738727	55614612	55614612	55614612	55614612
% Informative Characters	70	70	70	70	70	70	70	70
Total Tree Length	203694546	85692748	84585002	83898687	79590448	79590448	79590448	79590448
Total Consistency Index	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72
Total Homoplasy Index	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28
Total Retention Index	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51
Total Rescaled Consistency Index	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28
Total Goloboff Fit	2.13	2.12	2.12	2.12	2.13	2.13	2.13	2.13
Total Gamma Rates	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20
Total Invariable Rates	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
Total -ln Likelihood Score	137036.65	140038.47	139463.84	138413.59	134986.67	134986.67	134986.67	134986.67
Total Bits per Character	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02



	Drosophila	Nephrozoa	Triploblastica	Eumetazoa	Metazoa	Opisthokonta	Eukaryota	Biota
Genes	14925	12997	11314	11020	11008	9659	9163	1669
% Genome	100	87	76	74	74	65	61	11
Total Characters	91915264	79590448	72133750	70227261	70035675	61650224	57717624	7292987
% Total Genome Characters	100	87	78	76	76	67	63	8
Constant Characters	14690262	13197059	12163497	11901569	11884757	10740424	10144942	1554128
Variable Uniformative Characters	12217058	10606947	9643455	9379104	9356244	8235417	7730947	956010
Informative Characters	64748176	55614612	50193501	48819695	48667781	42576982	39746710	4766156
% Informative Characters	70	70	70	70	69	69	69	65
Total Tree Length	203694546	79590448	72133750	70227261	70035675	61650224	57717624	7292987
Total Consistency Index	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72
Total Homoplasy Index	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28
Total Retention Index	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.50
Total Rescaled Consistency Index	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28
Total Goloboff Fit	2.13	2.13	2.13	2.13	2.13	2.14	2.15	2.27
Total Gamma Rates	0.20	0.20	0.21	0.21	0.21	0.21	0.22	0.25
Total Invariable Rates	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
Total -ln Likelihood Score	137036.65	134986.67	133064.56	132244.24	132142.34	129972.26	127463.23	75566.03
Total Bits per Character	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01

TABLE 13. Total sum parametric scores from molecular phylogeny of target genes represented in taxonomic rank specific subnetworks of microRNA-target network data. These network data represent the expansion of putative target genes through taxonomic hierarchy where networks at each rank were produced with 112 microRNA families held constant. Rows and coloring correspond to the datasets of selected taxonomic ranks presented in CHAPTER V, FIGURE 29. Parametric scores were calculated through PAUP* where molecular phylogeny

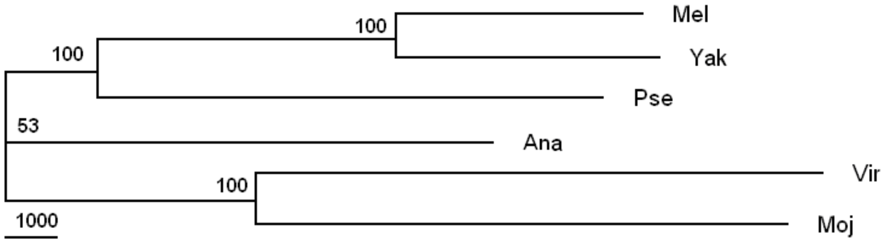
was constrained to the topology of the reference tree (CHAPTER III, FIGURE 22; Swofford, 2002). Likelihood scores were calculated under a under a under a general time reversible sequence evolution model with gamma rate variation and invariable sites (GTR+I+G).

APPENDIX IV. Phylogenetic Reconstruction from Weighted MicroRNA-Target Network Edges

Phylogenetic analyses have been conducted for a weighted edge network of 441 target genes linked to microRNA-277 predicted from the PicTar search algorithm of aligned 3’UTRs of target genes with medium sensitivity and specificity (setting S3; Grün *et al.*, 2005). Specifically, network edges for the free energies of microRNA-target duplex hybridization (-ΔG kcal/mol) were coded into a species matrix as absolute integer values and scored proportionally (by 1, 10, 100, or 1000) according to decimal place. Methods of phylogenetic reconstruction employed included standard parsimony (MP) under a heuristic search and distance criteria using the neighbor-joining algorithm (NJ) as implemented through PAUP* (Swofford, 2002). Branch supports of trees were evaluated by nonparametric bootstrap (BP), third- and half-delete jackknife (JK) calculated to high confidence levels. Both reconstruction methods recovered a single tree topology largely congruent to established drosophilid phylogenies and differed only in the placement of *D. ananassae* (FIGURE 38). Nodal supports for the placement of the later taxa were low and itinerant (MP-BP = 57, MP-JK½ = 58, MP-JK⅓ = 63; NJ-BP = 53, NJ-JK½ = 50, NJ-JK⅓ = 62), but otherwise this tree topology was well supported with bootstrap and jackknife values of 100.

Under favorable conditions with roughly equal rates of change and symmetric branches, bootstrap values greater than 70% correspond to a probability of greater than 95% that the true phylogeny has been found (Hillis and Bull, 1993). On the basis of visual inspection it is hypothesized that the problematic placement of *D. ananassae* in this analysis may stem from missing data or mis-inferred target site homology in the original alignment data used by PicTar for target prediction (Grün *et al.*, 2005). Recovery of phylogenetic information from the weighted edges of networks and the production of species tree topologies fully (or mostly) congruent with an expected topology under the previously described pilot studies is most notable. These findings provide evidence to support a hypothesis that weighted edge microRNA network structure itself can be directly utilized for phylogenetic inference (Mazurie *et al.*, 2008; Suthram *et al.*, 2005).

FIGURE 38. Phylogram recovered using weighted edge *dme-miR-227*-target data derived from PicTar (Grün *et al.*, 2005). This phylogram represents single most parsimonious midpoint-rooted tree retrieved from weighted standard parsimony reconstruction with free energies of hybridization (-ΔG kcal/mol) for 441 mRNA transcripts targeted to *miR-277*. Bootstrap proportions indicated near respective branches. This tree recovered parsimony scores of length 61640 steps, CI=0.865, RI=0.592, RC=0.512, HI=0.135, and G-fit=-201.206. *Drosophila* species are abbreviated respectively: Ana) *ananassae*; Mel) *melanogaster*; Moj) *mojavensis*; Pse) *pseudoobscura*; Vir) *virilis*; Yak) *yakuba.*



APPENDIX V. Phylogenetic Reconstruction using Protein-Protein Interactome Data

Novel phylogenetic analyses were conducted using character data derived from weighted edges of protein-protein interactions networks (FIGURE 39). This methodology was developed using data from DroSpeGe database for genes exhibiting statistically significant expansion or depletion in the twelve *Drosophila* species according to in functional categories by gene ontology (GO; Gilbert, 2007). The published dataset consisted of copy numbers (paralogs) for a list of 1473 genes and 100 GO terms. Gene identifications were used as input for PathWay Studio 6.0 (Ariadne Genomics) software, and a direct protein-protein interaction network (interactome) was resolved with 1000 edges and 613 nodes (Nikitin *et al.*, 2003). With the aid of a MySQL database, the direct protein-protein interactome was used as a cipher to key into the gene paralogy table for the original dataset (Sun Microsystems, Inc. 2008-2009). Species-specific weights for 716 protein interactome edges were recovered assuming a one-to-one second order reaction mechanism according to the product of the copy numbers of the protein-coding genes. Each network edge weight was coded as a numeric character for phylogenetic analyses and scored proportionally (by 1, 10, 100, or 1000) according to decimal position.

Phylogenetic analyses were performed separately and in union for gene copy numbers, GO terms, and edge weights of the protein-interactome. Methods for phylogenetic reconstruction employed included Bayesian inference through MrBayes, standard parsimony (MP) and distance criteria using the neighbor-joining (NJ) algorithm as implemented through PAUP* (Huelsenbeck & Ronquist, 2005, Swofford, 2002). Partition homogeneity test for gene copy number, GO terms, and weighted protein-protein interactions recovered no significant incongruence (P = 1.0); thus all data could be evaluated as part of a shared phylogenetic history (Bull *et al.*, 1993) Nevertheless, the tree topology and resolution recovered for all separate and combined datasets was found to differ depending upon the phylogenetic methods employed. Bayesian inference of the entire dataset of gene copy numbers, GO terms, and protein-interactome edge weights recovered a tree topology fully congruent to the reference tree with strong posterior probability (PP) branch supports of 97-100. A topology-dependent permutation test (T-PTP) for the entire dataset against the established *Drosophila* phylogeny indicated significant support of the data for the expected topology (P = 0.0168), however the most parsimonious tree differed from the reference tree in the placement of *D. erecta* and *D. yakuba* within the *D. melanogaster* subgroup, and the *D. obscura* clade outside of the subgenus *Sophophora* (CHAPTER III, FIGURE 22; Faith, 1991). Branch supports under standard parsimony were variable and inconsistent between statistical measures: nonparametric bootstrap (BP) < 50-81; third-delete jackknife (JK⅓) = 24-82; half-delete jackknife (JK½) = 34-93. In contrast under NJ, the entire dataset recovered a strongly supported (BP = 100, JK⅓ = 100, JK½ = 100) tree nearly congruent to the expected topology differing only in the placement of *Drosophila yakuba* relative to *D. erecta.*

In comparison to the entire dataset, separate phylogenetic analyses under MP and NJ using only gene copy number or GO term data repeatedly recovered the same nearly congruent topology to the expected tree (differing only in *Drosophila yakuba* to *D. erecta* relative placement) with consistent BP, JK½, and JK⅓ frequencies at 100. Similarly, separate T-PTP analyses of gene copy number or GO term data indicated significant support for the reference phylogeny (P < 10^{-4} and P = 0.0224), Bayesian inference for gene copy number retrieved a strongly supported tree topology (PP = 86-100) fully congruent to the reference tree (Faith, 1991). Conversely, GO term data retrieved a poorly resolved (PP = <50-100) tree incongruent to the expected topology.

The protein-interactome edge weight dataset consisted of 1172 unordered characters with 586 each of weights of 10 and 1. There were 577 (49.2% of dataset) constant characters, 106 (9.0% of dataset) variable but parsimony-uninformative, and 489 (41.7% of dataset) potentially parsimony-informative characters recovered from this data. Bayesian inference recovered a well-supported tree mostly congruent to the reference topology (PP = 100) except for the unusual placement of *D. willistoni* as nearest sister to *D. melanogaster* (PP = 71). Branch supports under standard parsimony were variable (BP = 55-100, JK⅓ = 52-100, JK½ =5 1-100) but relatively consistent per node, and the most parsimonious trees differed from the reference topology in the placement of *D. erecta* within the *D. melanogaster* subgroup. Nevertheless, a T-PTP study retrieved significant support of the data for the reference *Drosophila* tree (P < 10^{-4}; Faith, 1991). The NJ tree topology was strongly supported (BP=100, JK⅓ = 100, JK½ = 100) and nearly congruent to the expected tree; differing only in the placement of *Drosophila yakuba* relative to *D. erecta* (CHAPTER III, FIGURE 22; FIGURE 39). These findings provide evidence to support a hypothesis that weighted edge protein interaction network structure itself can be directly utilized for phylogenetic inference.

Additional phylogenetic analyses were initiated for an expanded sample of paralogy data. A paralogy table was constructed from reciprocal BLAST data of 12 *Drosophila* species using whole mRNA libraries, EISE_exonerate, EISE_genemapper, EISE_genewise, and GLEANR computationally predicted annotations accessible thorough FlyBase and DroSpeGe (Birney, *et al.*, 2004; Chatterji and Pachter, 2006; Heger and Ponting, 2006; Mackey *et al.*, 2006; Slater and Birney, 2005). Cytoscape freeware with BioNetBuilder and DroID plugins was utilized to build networks of interacting proteins for *Drosophila melanogaster* this network of interacting proteins were used as a cipher to key into each gene paralogy table. Species-specific network edge weights were recovered from the product of the copy numbers of the protein-coding genes. The edge weights were used as characters for phylogenetic reconstruction through PAUP* (FIGURE 40). Separate phylogenetic analyses of strict paralogy, presence/absence of network edges, and proteome weighted edges for messenger RNA derived data under parsimony and neighbor joining consistently recovered tree topologies nearly congruent topology to the expected tree (CHAPTER III, FIGURE 22). Branch supports by bootstrap were highly variable (>50-100%) while recovered jackknife robust and largely consistent between statistical measures with frequencies between 75 and 100. Recovered trees differed from the reference tree due to itinerant placement of *Drosophila ananassae* and *D. melanogaster* (FIGURE 40). While differences in recovered tree topologies from the reference tree may stem from missing data or mis-inferred target site homology in the source data or may indicate underlying methodological biases, nevertheless, the recovery of phylogenetic information from the weighted edges of networks and the production of species tree topologies largely congruent to the expected topology is most notable. Collectively, the findings of these studies indicate that while these GO terms have likely been subject to positive selection, there is nevertheless the likely presence of phylogenetic information available in gene copy number and the inferable weighted-edge network structure (*Drosophila* 12 Genomes Consortium, 2007; Gilbert, 2007; Khaitovich *et al.*, 2004).

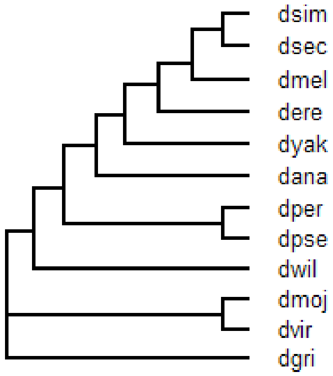


FIGURE 39. Cladogram recovered from Neighbor-Joining using weighted edge protein-protein interactome data derived from the DroSpeGe database (Gilbert, 2007). This tree recovered parsimony scores of length 3709 with parametric scores of CI=0.582, RI=0.353, RC=0.205, HI=0.418, and G-fit= -419.074. The NJ tree topology was strongly supported (BP=100, JK⅓ =100, JK½=100). All bootstrap, half- and third-delete jackknife frequencies retrieved for this topology were 100. *Drosophila* species are abbreviated respectively: dana) *ananassae*; dere) *erecta*; dgri) *grimshawi*; dmel) *melanogaster*; dmoj) *mojavensis*; dper) *persimilis*; dpse) *pseudoobscura*; dsec) *sechellia*; dsim) *simulans*; dvir) *virilis*; dwil) *willistoni*; dyak) *yakuba.*

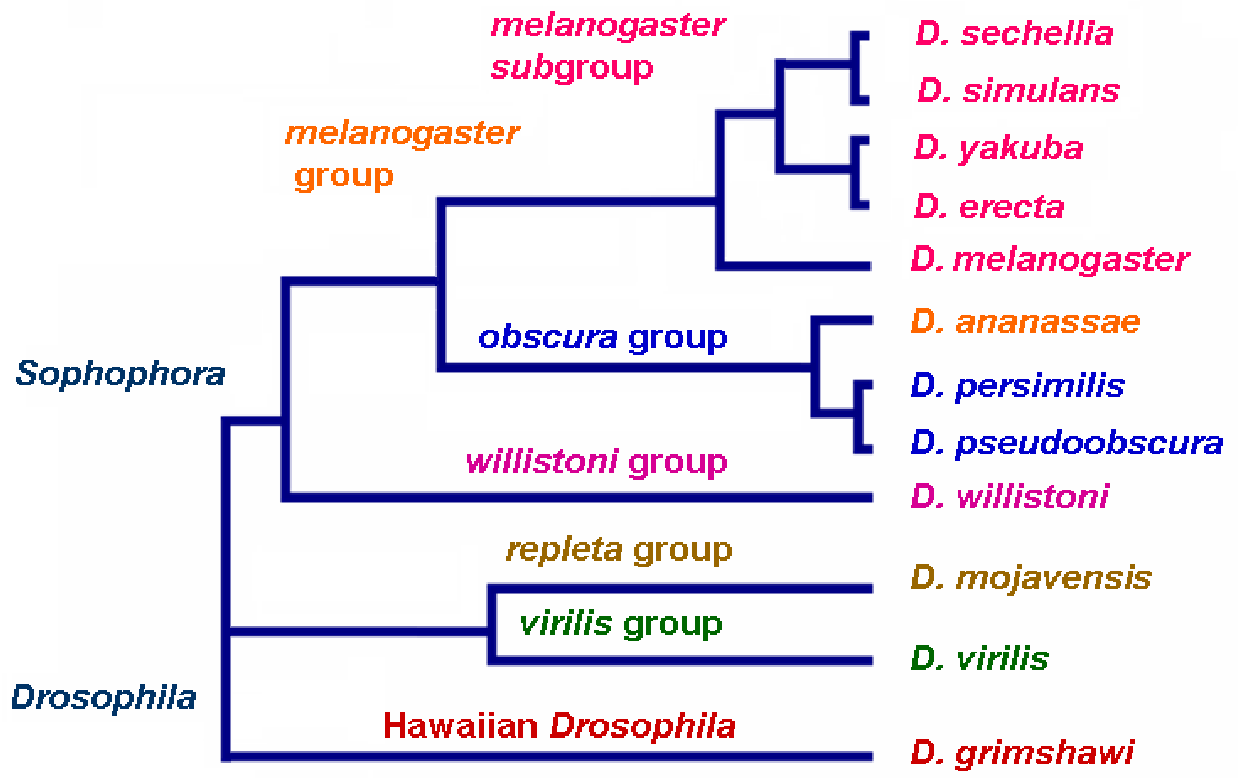
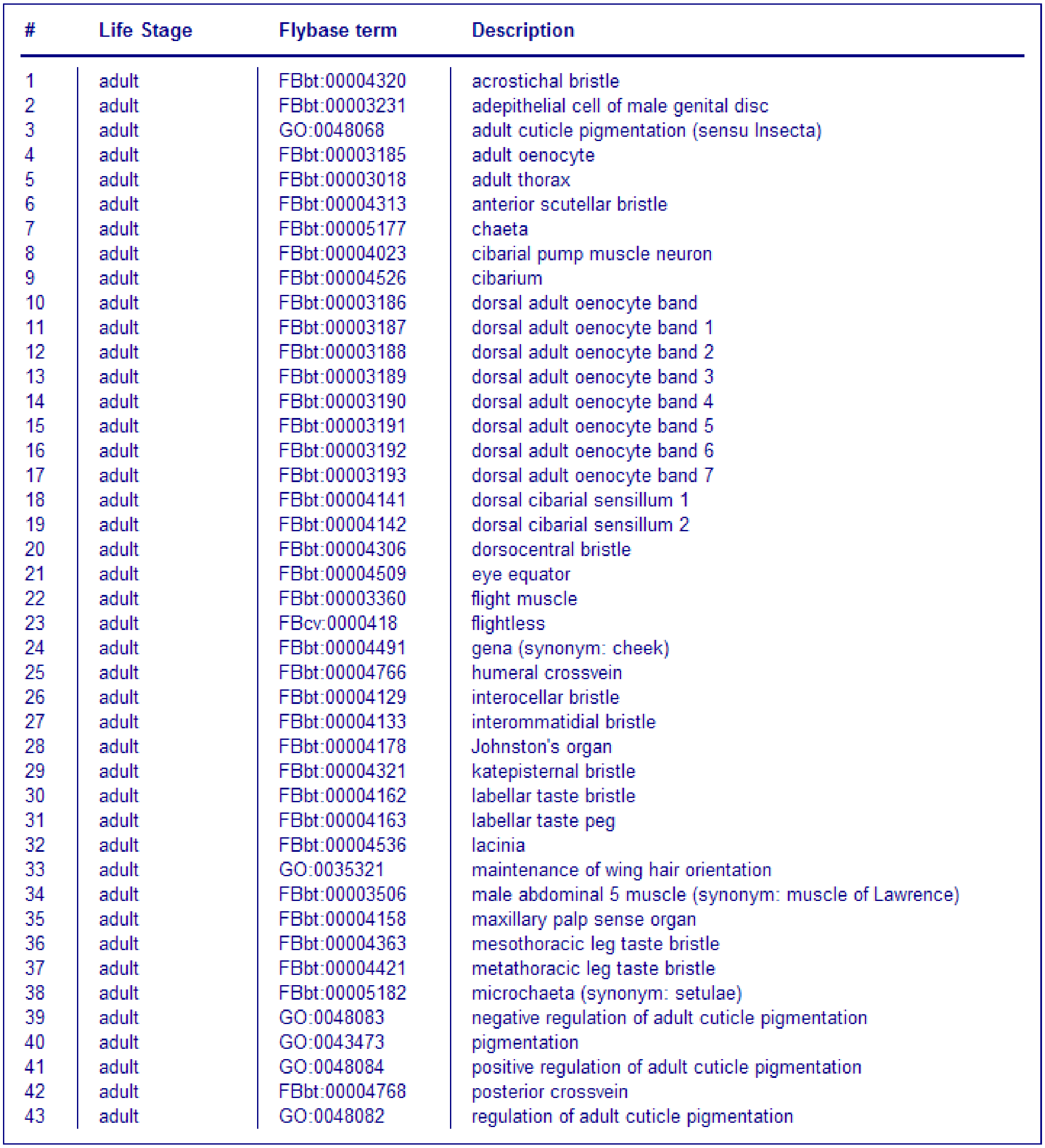
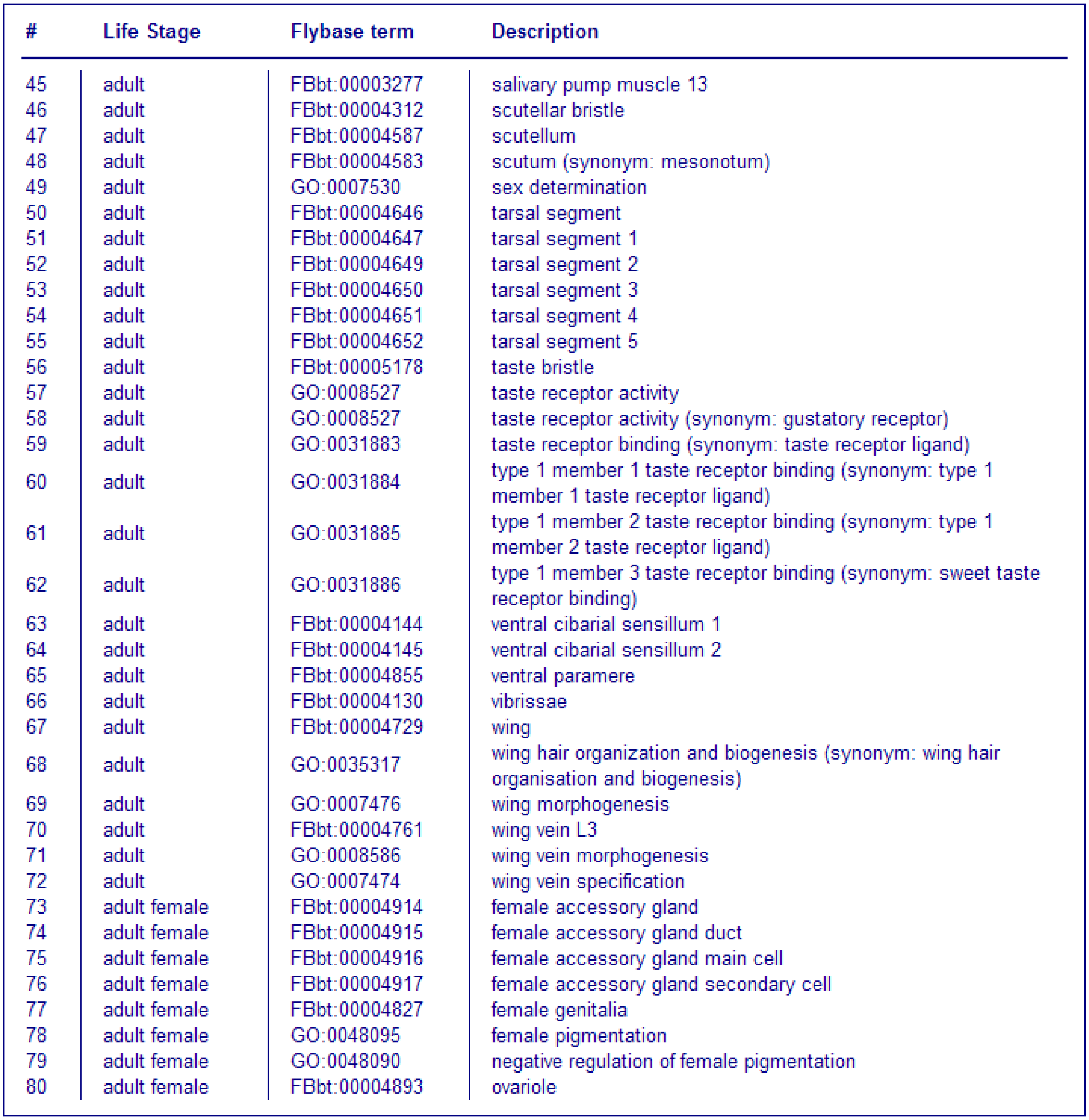


FIGURE 40. Cladogram recovered from using weighted edge protein-protein interactome data derived from reciprocal messenger RNA BLASTs of 12 *Drosophila* species. This cladogram represents the single most parsimonious rooted tree retrieved from weighted parsimony reconstruction under a branch-and-bound search. This tree recovered a length of 96,887 for 215,394 total characters; of which 20,761 (9.6% total sample) were potentially parsimony informative. Recovered parsimony scores included: Consistency Index (CI) of 0.875, Retention Index (RI) of 0.714, Rescaled Consistency Index (RC) of 0.625, Homoplasy index (HI) of 0.125, and Goloboff-fit (G-fit) of -19034.447. Nodal support values by bootstrap were highly variable (>50-100%) while recovered jackknife frequencies retrieved were high (95-100%) for all clades expect *Sophophora* (61-64).

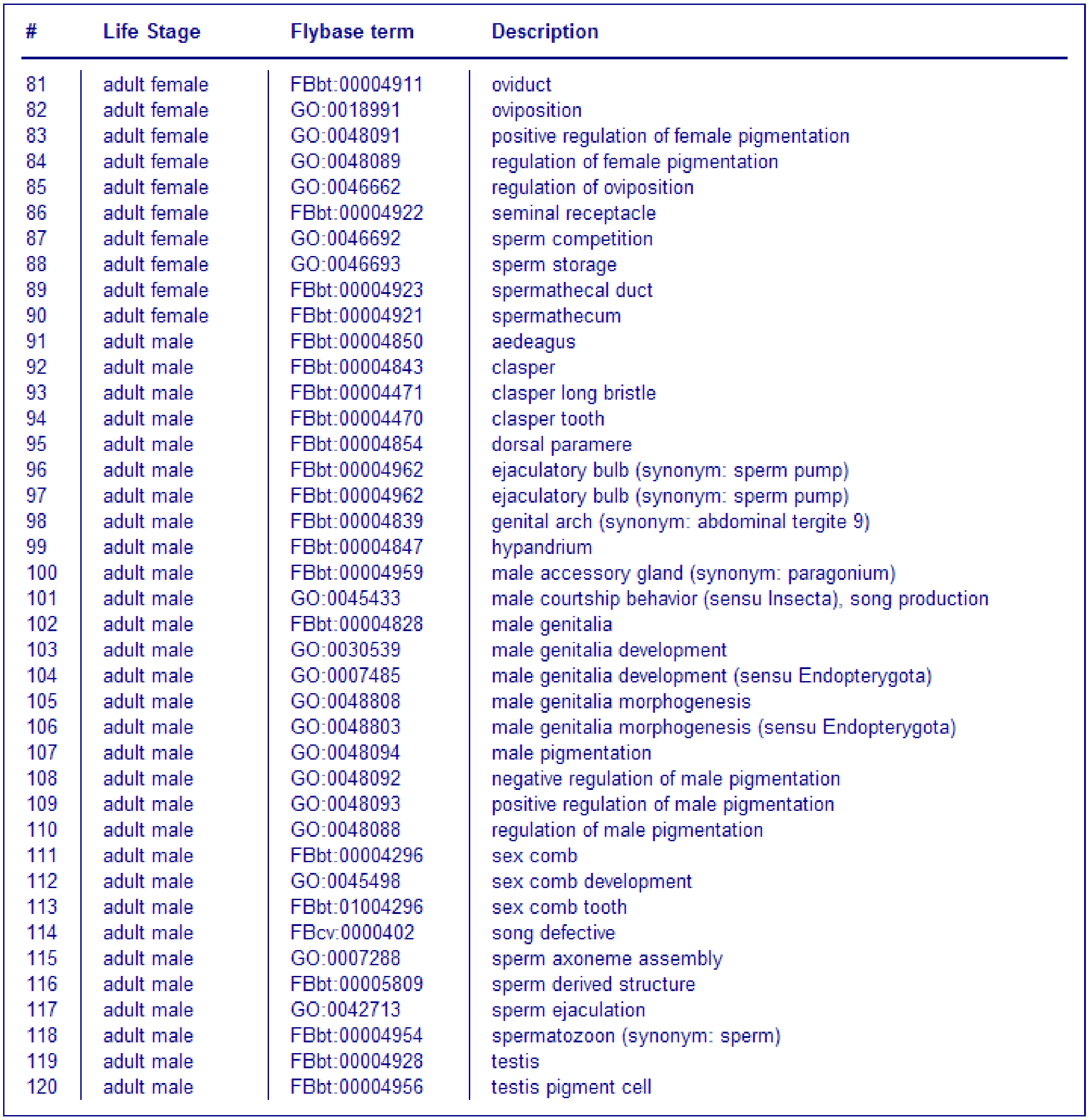
APPENDIX VI. Table of Life Stage, FlyBase Terms and Description of Features Available to Diagnose *Drosophila* species.



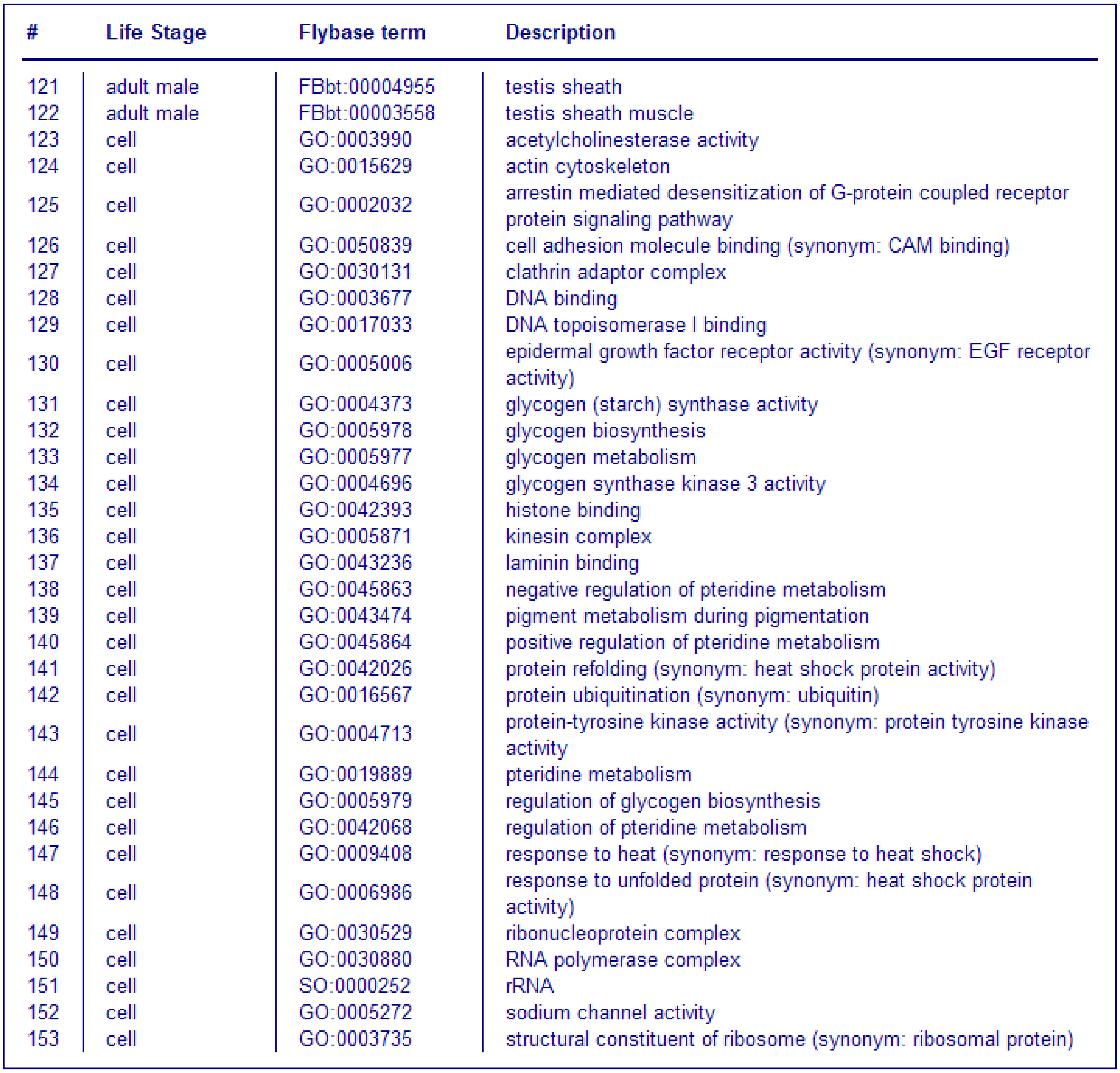
#	Life Stage	Flybase term	Description
1	adult	FBbt:00004320	acrostichal bristle
2	adult	FBbt:00003231	adepithelial cell of male genital disc
3	adult	GO:0048068	adult cuticle pigmentation (sensu Insecta)
4	adult	FBbt:00003185	adult oenocyte
5	adult	FBbt:00003018	adult thorax
6	adult	FBbt:00004313	anterior scutellar bristle
7	adult	FBbt:00005177	chaeta
8	adult	FBbt:00004023	cibarial pump muscle neuron
9	adult	FBbt:00004526	cibarium
10	adult	FBbt:00003186	dorsal adult oenocyte band
11	adult	FBbt:00003187	dorsal adult oenocyte band 1
12	adult	FBbt:00003188	dorsal adult oenocyte band 2
13	adult	FBbt:00003189	dorsal adult oenocyte band 3
14	adult	FBbt:00003190	dorsal adult oenocyte band 4
15	adult	FBbt:00003191	dorsal adult oenocyte band 5
16	adult	FBbt:00003192	dorsal adult oenocyte band 6
17	adult	FBbt:00003193	dorsal adult oenocyte band 7
18	adult	FBbt:00004141	dorsal cibarial sensillum 1
19	adult	FBbt:00004142	dorsal cibarial sensillum 2
20	adult	FBbt:00004306	dorsocentral bristle
21	adult	FBbt:00004509	eye equator
22	adult	FBbt:00003360	flight muscle
23	adult	FBcv:0000418	flightless
24	adult	FBbt:00004491	gena (synonym: cheek)
25	adult	FBbt:00004766	humeral crossvein
26	adult	FBbt:00004129	interocellar bristle
27	adult	FBbt:00004133	interommatidial bristle
28	adult	FBbt:00004178	Johnston's organ
29	adult	FBbt:00004321	katepisternal bristle
30	adult	FBbt:00004162	labellar taste bristle
31	adult	FBbt:00004163	labellar taste peg
32	adult	FBbt:00004536	lacinia
33	adult	GO:0035321	maintenance of wing hair orientation
34	adult	FBbt:00003506	male abdominal 5 muscle (synonym: muscle of Lawrence)
35	adult	FBbt:00004158	maxillary palp sense organ
36	adult	FBbt:00004363	mesothoracic leg taste bristle
37	adult	FBbt:00004421	metathoracic leg taste bristle
38	adult	FBbt:00005182	microchaeta (synonym: setulae)
39	adult	GO:0048083	negative regulation of adult cuticle pigmentation
40	adult	GO:0043473	pigmentation
41	adult	GO:0048084	positive regulation of adult cuticle pigmentation
42	adult	FBbt:00004768	posterior crossvein
43	adult	GO:0048082	regulation of adult cuticle pigmentation



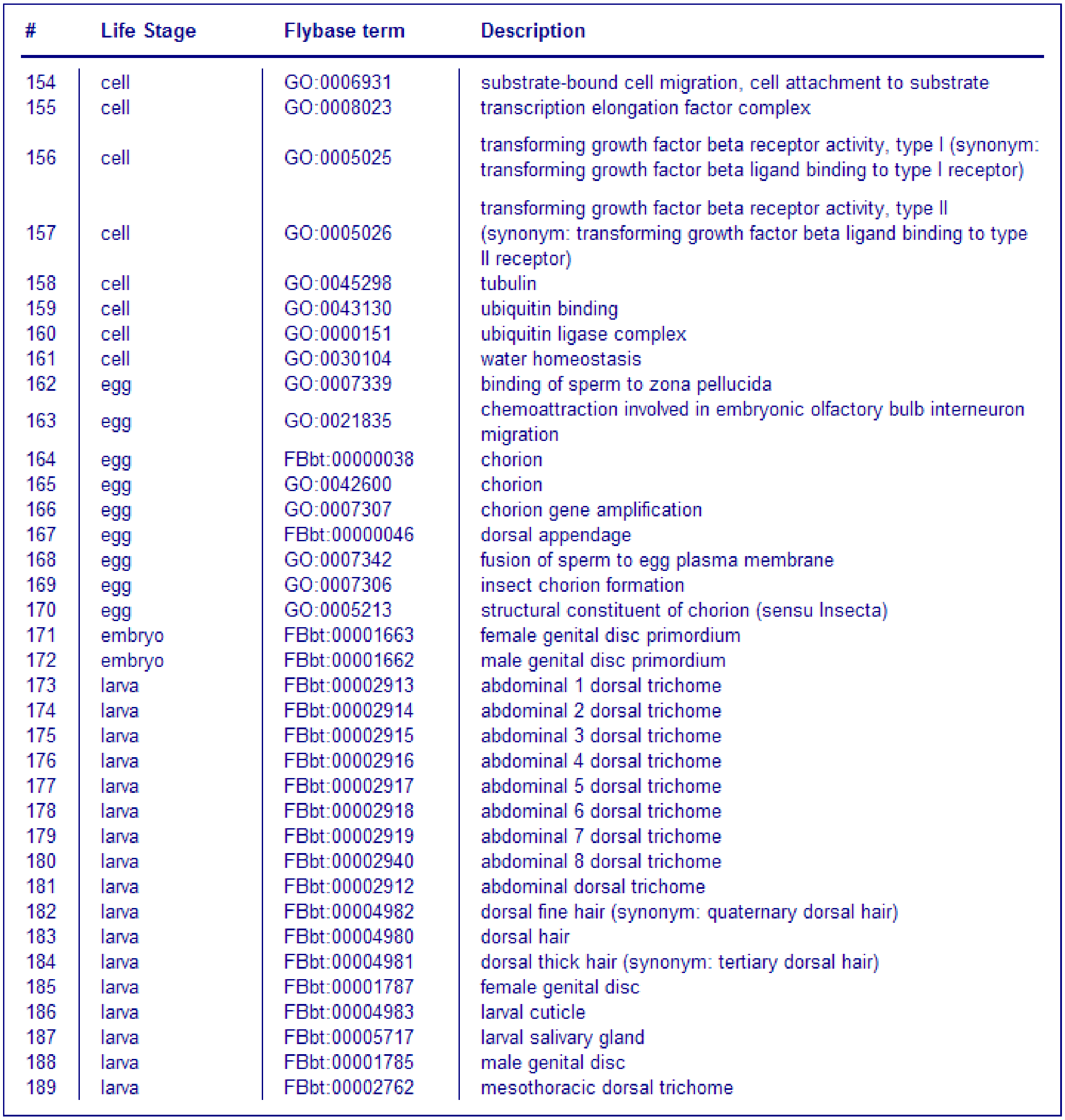
#	Life Stage	Flybase term	Description
45	adult	FBbt:00003277	salivary pump muscle 13
46	adult	FBbt:00004312	scutellar bristle
47	adult	FBbt:00004587	scutellum
48	adult	FBbt:00004583	scutum (synonym: mesonotum)
49	adult	GO:0007530	sex determination
50	adult	FBbt:00004646	tarsal segment
51	adult	FBbt:00004647	tarsal segment 1
52	adult	FBbt:00004649	tarsal segment 2
53	adult	FBbt:00004650	tarsal segment 3
54	adult	FBbt:00004651	tarsal segment 4
55	adult	FBbt:00004652	tarsal segment 5
56	adult	FBbt:00005178	taste bristle
57	adult	GO:0008527	taste receptor activity
58	adult	GO:0008527	taste receptor activity (synonym: gustatory receptor)
59	adult	GO:0031883	taste receptor binding (synonym: taste receptor ligand)
60	adult	GO:0031884	type 1 member 1 taste receptor binding (synonym: type 1 member 1 taste receptor ligand)
61	adult	GO:0031885	type 1 member 2 taste receptor binding (synonym: type 1 member 2 taste receptor ligand)
62	adult	GO:0031886	type 1 member 3 taste receptor binding (synonym: sweet taste receptor binding)
63	adult	FBbt:00004144	ventral cibarial sensillum 1
64	adult	FBbt:00004145	ventral cibarial sensillum 2
65	adult	FBbt:00004855	ventral paramere
66	adult	FBbt:00004130	vibrissae
67	adult	FBbt:00004729	wing
68	adult	GO:0035317	wing hair organization and biogenesis (synonym: wing hair organisation and biogenesis)
69	adult	GO:0007476	wing morphogenesis
70	adult	FBbt:00004761	wing vein L3
71	adult	GO:0008586	wing vein morphogenesis
72	adult	GO:0007474	wing vein specification
73	adult female	FBbt:00004914	female accessory gland
74	adult female	FBbt:00004915	female accessory gland duct
75	adult female	FBbt:00004916	female accessory gland main cell
76	adult female	FBbt:00004917	female accessory gland secondary cell
77	adult female	FBbt:00004827	female genitalia
78	adult female	GO:0048095	female pigmentation
79	adult female	GO:0048090	negative regulation of female pigmentation
80	adult female	FBbt:00004893	ovariole



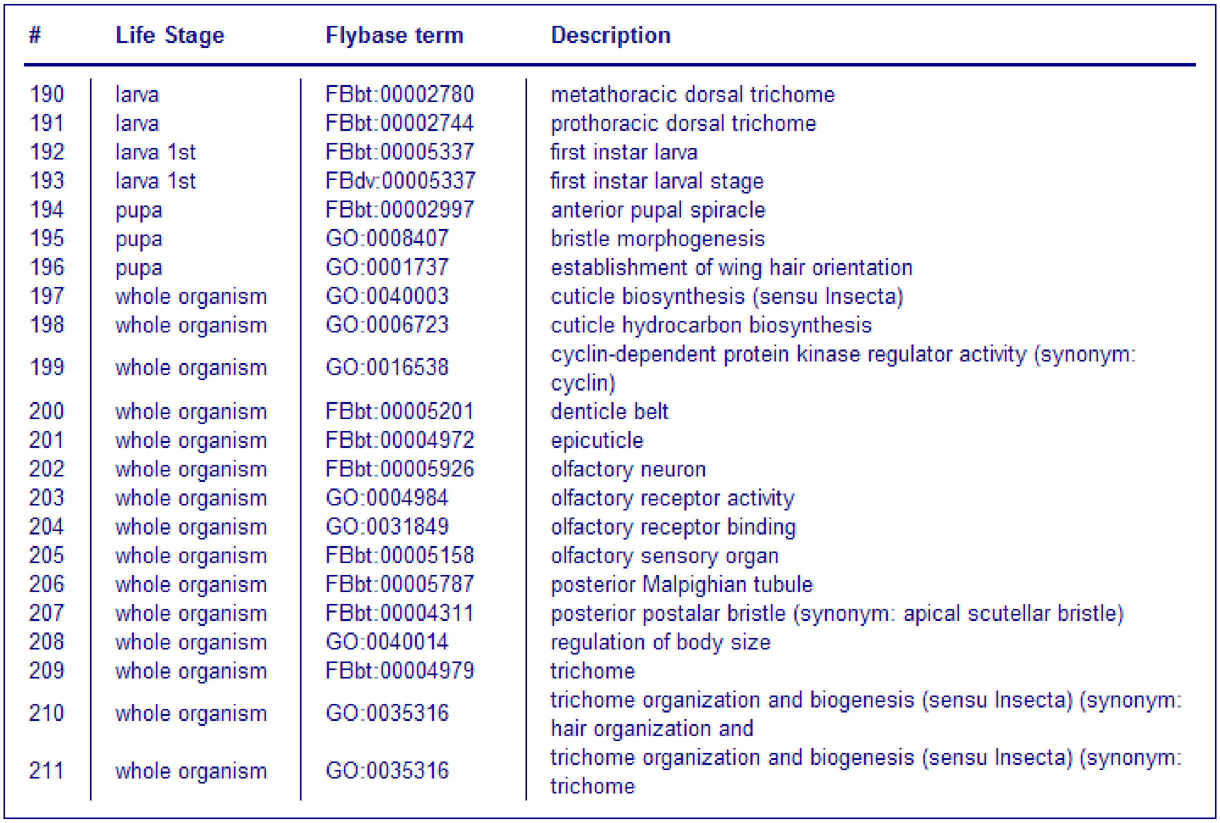
#	Life Stage	Flybase term	Description
81	adult female	FBbt:00004911	oviduct
82	adult female	GO:0018991	oviposition
83	adult female	GO:0048091	positive regulation of female pigmentation
84	adult female	GO:0048089	regulation of female pigmentation
85	adult female	GO:0046662	regulation of oviposition
86	adult female	FBbt:00004922	seminal receptacle
87	adult female	GO:0046692	sperm competition
88	adult female	GO:0046693	sperm storage
89	adult female	FBbt:00004923	spermathecal duct
90	adult female	FBbt:00004921	spermathecum
91	adult male	FBbt:00004850	aedeagus
92	adult male	FBbt:00004843	clasper
93	adult male	FBbt:00004471	clasper long bristle
94	adult male	FBbt:00004470	clasper tooth
95	adult male	FBbt:00004854	dorsal paramere
96	adult male	FBbt:00004962	ejaculatory bulb (synonym: sperm pump)
97	adult male	FBbt:00004962	ejaculatory bulb (synonym: sperm pump)
98	adult male	FBbt:00004839	genital arch (synonym: abdominal tergite 9)
99	adult male	FBbt:00004847	hypandrium
100	adult male	FBbt:00004959	male accessory gland (synonym: paragonium)
101	adult male	GO:0045433	male courtship behavior (sensu Insecta), song production
102	adult male	FBbt:00004828	male genitalia
103	adult male	GO:0030539	male genitalia development
104	adult male	GO:0007485	male genitalia development (sensu Endopterygota)
105	adult male	GO:0048808	male genitalia morphogenesis
106	adult male	GO:0048803	male genitalia morphogenesis (sensu Endopterygota)
107	adult male	GO:0048094	male pigmentation
108	adult male	GO:0048092	negative regulation of male pigmentation
109	adult male	GO:0048093	positive regulation of male pigmentation
110	adult male	GO:0048088	regulation of male pigmentation
111	adult male	FBbt:00004296	sex comb
112	adult male	GO:0045498	sex comb development
113	adult male	FBbt:01004296	sex comb tooth
114	adult male	FBcv:0000402	song defective
115	adult male	GO:0007288	sperm axoneme assembly
116	adult male	FBbt:00005809	sperm derived structure
117	adult male	GO:0042713	sperm ejaculation
118	adult male	FBbt:00004954	spermatozoon (synonym: sperm)
119	adult male	FBbt:00004928	testis
120	adult male	FBbt:00004956	testis pigment cell



#	Life Stage	Flybase term	Description
121	adult male	FBbt:00004955	testis sheath
122	adult male	FBbt:00003558	testis sheath muscle
123	cell	GO:0003990	acetylcholinesterase activity
124	cell	GO:0015629	actin cytoskeleton
125	cell	GO:0002032	arrestin mediated desensitization of G-protein coupled receptor protein signaling pathway
126	cell	GO:0050839	cell adhesion molecule binding (synonym: CAM binding)
127	cell	GO:0030131	clathrin adaptor complex
128	cell	GO:0003677	DNA binding
129	cell	GO:0017033	DNA topoisomerase I binding
130	cell	GO:0005006	epidermal growth factor receptor activity (synonym: EGF receptor activity)
131	cell	GO:0004373	glycogen (starch) synthase activity
132	cell	GO:0005978	glycogen biosynthesis
133	cell	GO:0005977	glycogen metabolism
134	cell	GO:0004696	glycogen synthase kinase 3 activity
135	cell	GO:0042393	histone binding
136	cell	GO:0005871	kinesin complex
137	cell	GO:0043236	laminin binding
138	cell	GO:0045863	negative regulation of pteridine metabolism
139	cell	GO:0043474	pigment metabolism during pigmentation
140	cell	GO:0045864	positive regulation of pteridine metabolism
141	cell	GO:0042026	protein refolding (synonym: heat shock protein activity)
142	cell	GO:0016567	protein ubiquitination (synonym: ubiquitin)
143	cell	GO:0004713	protein-tyrosine kinase activity (synonym: protein tyrosine kinase activity
144	cell	GO:0019889	pteridine metabolism
145	cell	GO:0005979	regulation of glycogen biosynthesis
146	cell	GO:0042068	regulation of pteridine metabolism
147	cell	GO:0009408	response to heat (synonym: response to heat shock)
148	cell	GO:0006986	response to unfolded protein (synonym: heat shock protein activity)
149	cell	GO:0030529	ribonucleoprotein complex
150	cell	GO:0030880	RNA polymerase complex
151	cell	SO:0000252	rRNA
152	cell	GO:0005272	sodium channel activity
153	cell	GO:0003735	structural constituent of ribosome (synonym: ribosomal protein)



#	Life Stage	Flybase term	Description
154	cell	GO:0006931	substrate-bound cell migration, cell attachment to substrate
155	cell	GO:0008023	transcription elongation factor complex
156	cell	GO:0005025	transforming growth factor beta receptor activity, type I (synonym: transforming growth factor beta ligand binding to type I receptor)
157	cell	GO:0005026	transforming growth factor beta receptor activity, type II (synonym: transforming growth factor beta ligand binding to type II receptor)
158	cell	GO:0045298	tubulin
159	cell	GO:0043130	ubiquitin binding
160	cell	GO:0000151	ubiquitin ligase complex
161	cell	GO:0030104	water homeostasis
162	egg	GO:0007339	binding of sperm to zona pellucida
163	egg	GO:0021835	chemoattraction involved in embryonic olfactory bulb interneuron migration
164	egg	FBbt:00000038	chorion
165	egg	GO:0042600	chorion
166	egg	GO:0007307	chorion gene amplification
167	egg	FBbt:00000046	dorsal appendage
168	egg	GO:0007342	fusion of sperm to egg plasma membrane
169	egg	GO:0007306	insect chorion formation
170	egg	GO:0005213	structural constituent of chorion (sensu Insecta)
171	embryo	FBbt:00001663	female genital disc primordium
172	embryo	FBbt:00001662	male genital disc primordium
173	larva	FBbt:00002913	abdominal 1 dorsal trichome
174	larva	FBbt:00002914	abdominal 2 dorsal trichome
175	larva	FBbt:00002915	abdominal 3 dorsal trichome
176	larva	FBbt:00002916	abdominal 4 dorsal trichome
177	larva	FBbt:00002917	abdominal 5 dorsal trichome
178	larva	FBbt:00002918	abdominal 6 dorsal trichome
179	larva	FBbt:00002919	abdominal 7 dorsal trichome
180	larva	FBbt:00002940	abdominal 8 dorsal trichome
181	larva	FBbt:00002912	abdominal dorsal trichome
182	larva	FBbt:00004982	dorsal fine hair (synonym: quaternary dorsal hair)
183	larva	FBbt:00004980	dorsal hair
184	larva	FBbt:00004981	dorsal thick hair (synonym: tertiary dorsal hair)
185	larva	FBbt:00001787	female genital disc
186	larva	FBbt:00004983	larval cuticle
187	larva	FBbt:00005717	larval salivary gland
188	larva	FBbt:00001785	male genital disc
189	larva	FBbt:00002762	mesothoracic dorsal trichome



#	Life Stage	Flybase term	Description
190	larva	FBbt:00002780	metathoracic dorsal trichome
191	larva	FBbt:00002744	prothoracic dorsal trichome
192	larva 1st	FBbt:00005337	first instar larva
193	larva 1st	FBdv:00005337	first instar larval stage
194	pupa	FBbt:00002997	anterior pupal spiracle
195	pupa	GO:0008407	bristle morphogenesis
196	pupa	GO:0001737	establishment of wing hair orientation
197	whole organism	GO:0040003	cuticle biosynthesis (sensu Insecta)
198	whole organism	GO:0006723	cuticle hydrocarbon biosynthesis
199	whole organism	GO:0016538	cyclin-dependent protein kinase regulator activity (synonym: cyclin)
200	whole organism	FBbt:00005201	denticle belt
201	whole organism	FBbt:00004972	epicuticle
202	whole organism	FBbt:00005926	olfactory neuron
203	whole organism	GO:0004984	olfactory receptor activity
204	whole organism	GO:0031849	olfactory receptor binding
205	whole organism	FBbt:00005158	olfactory sensory organ
206	whole organism	FBbt:00005787	posterior Malpighian tubule
207	whole organism	FBbt:00004311	posterior postalar bristle (synonym: apical scutellar bristle)
208	whole organism	GO:0040014	regulation of body size
209	whole organism	FBbt:00004979	trichome
210	whole organism	GO:0035316	trichome organization and biogenesis (sensu Insecta) (synonym: hair organization and
211	whole organism	GO:0035316	trichome organization and biogenesis (sensu Insecta) (synonym: trichome

TABLE 13. Life stage, FlyBase terms and description of features available to diagnose *Drosophila* species. Anatomical features (FBbt) and physiological processes (gene ontology (GO) categories) available through FlyBase to diagnose species within the genus *Drosophila* (Grimaldi, 1990, Markow T, & O'Grady, 2005, Wilson *et al.*, 2008).

APPENDIX VII. Phylogenies of Taxa Surveyed for Conservation of MicroRNAs and Targets to *Drosophila melanogaster*

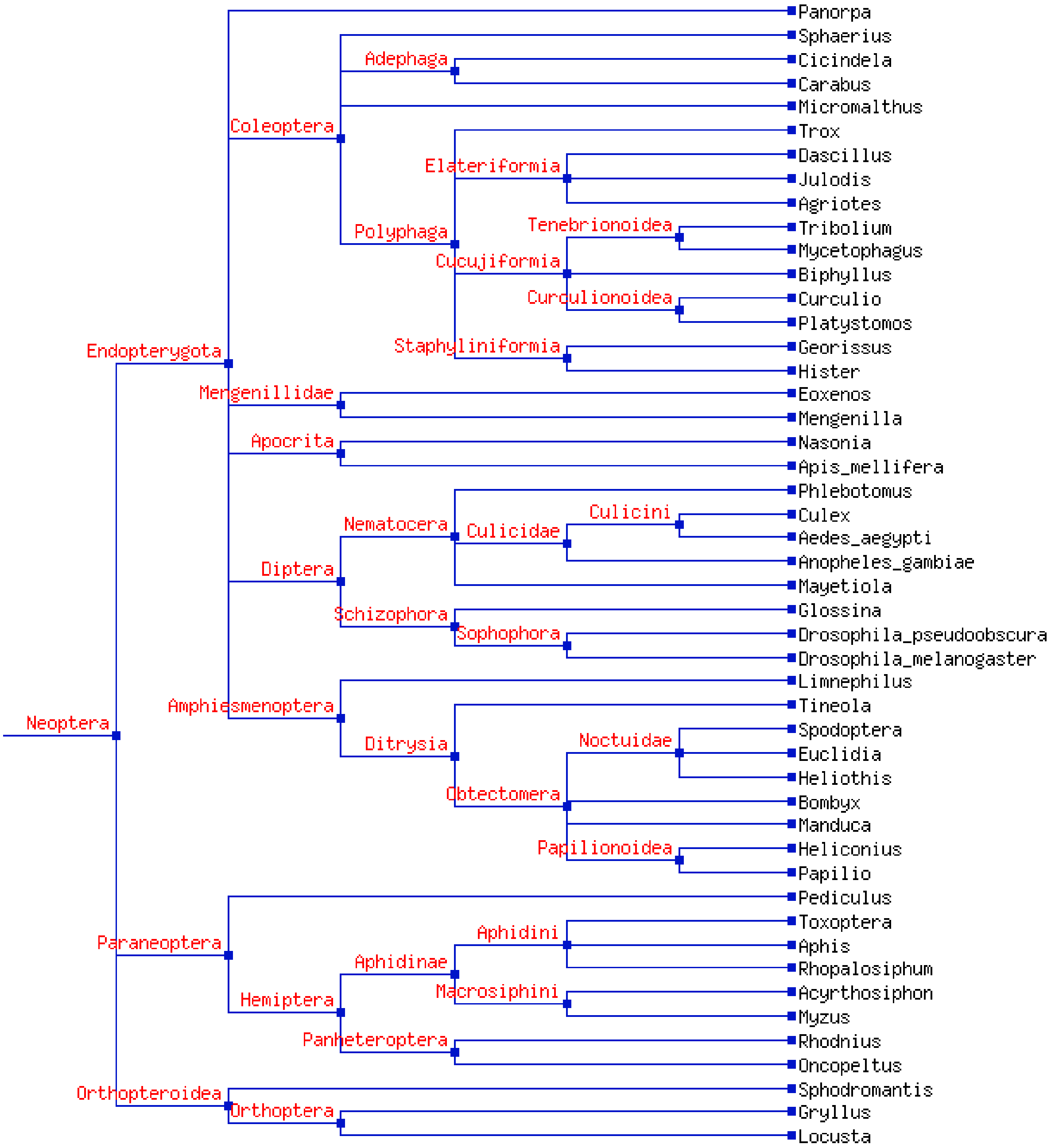


FIGURE 41. Phylogeny of insect taxa surveyed for conservation of microRNAs and targets to *Drosophila melanogaster* (Li *et al*., 2006; Ruan *et al*., 2008; Sharp, 1898).

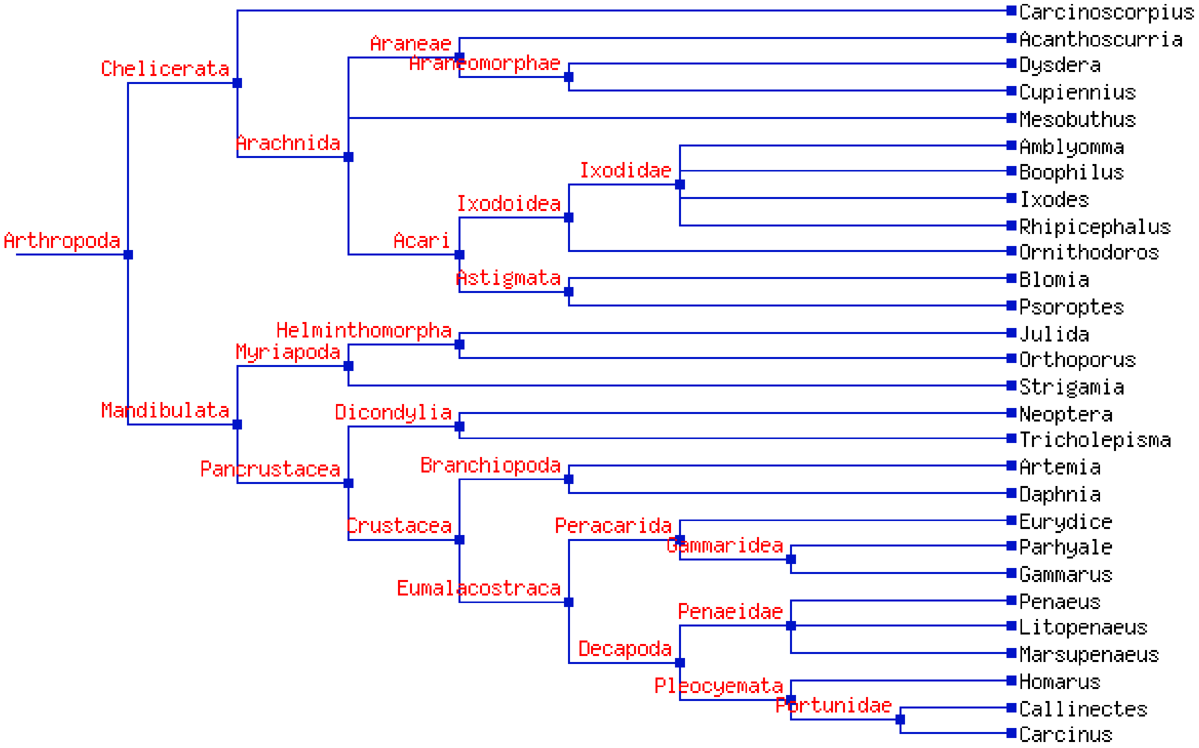


FIGURE 42. Phylogeny of arthropod taxa surveyed for conservation of microRNAs and targets to *Drosophila melanogaster* (Latreille, 1829; Li *et al.*, 2006; Ruan *et al.*, 2008).

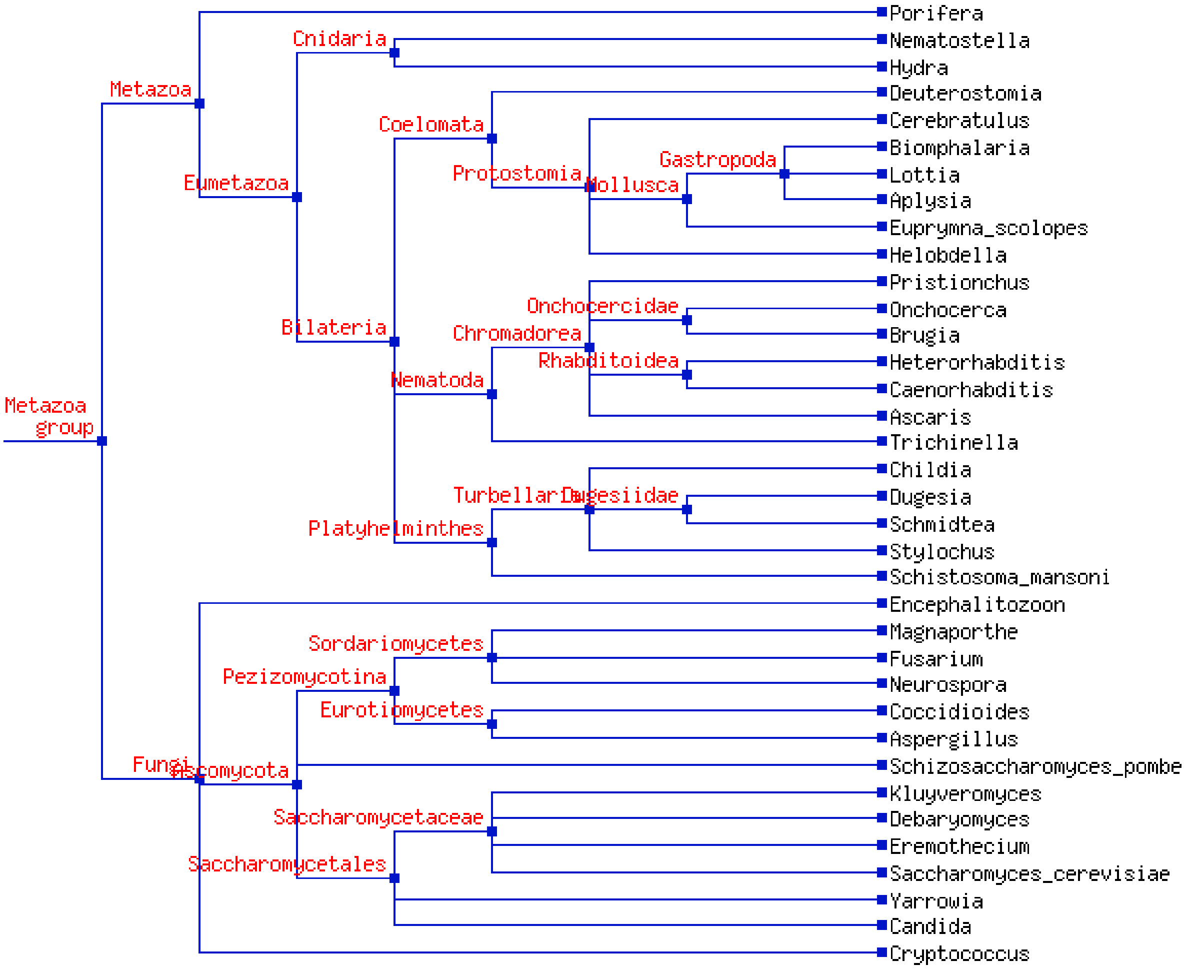


FIGURE 43. Phylogeny of opisthokont taxa surveyed for conservation of microRNAs and targets to *Drosophila melanogaster* (Li *et al.*, 2006; Ruan *et al.*, 2008). This phylogeny traces conservation to *Drosophila* through the ranks of Opisthokonta, Metazoa, Eumetazoa, Bilateria in the traditional sense, Coelomata, and Protostomia (Brands, 2005; Cavalier-Smith, 1987; Grobben, 1908; Haeckel, 1896; Hatschek, 1888; Hyman, 1951).

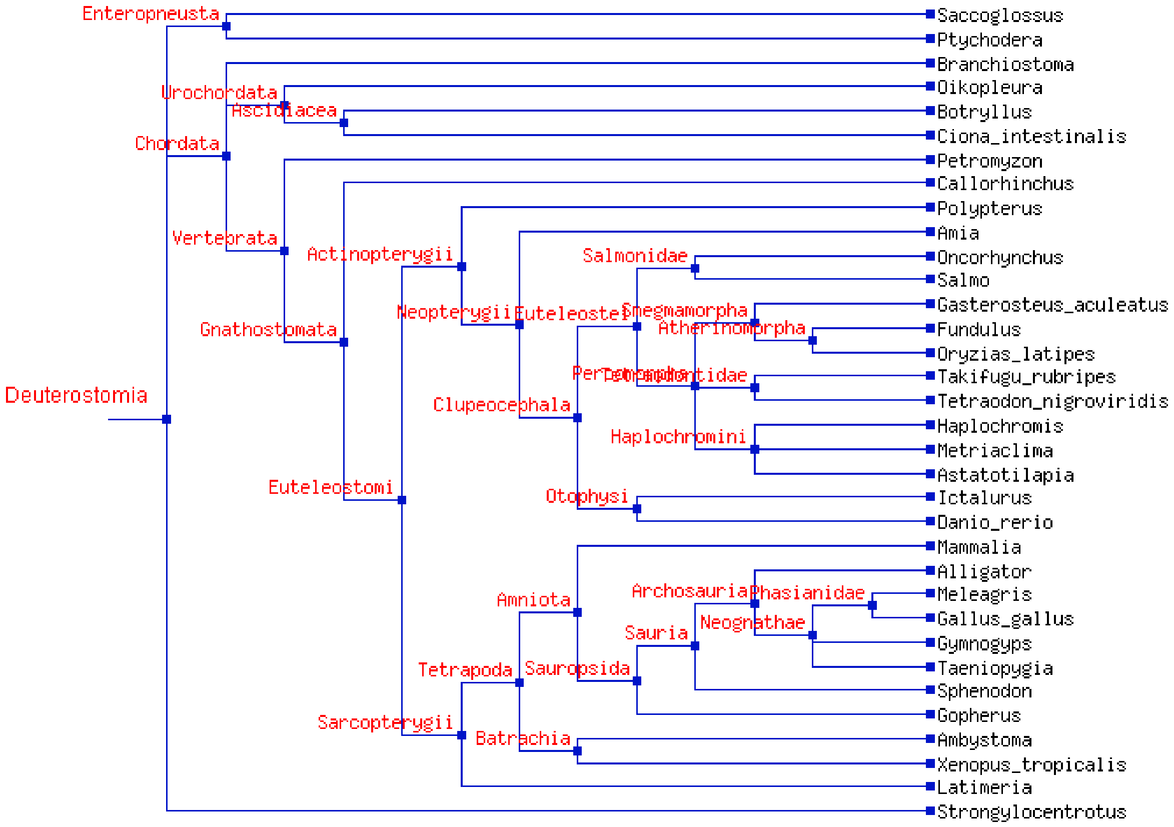


FIGURE 44. Phylogeny of deuterostome taxa surveyed for conservation of microRNAs and targets to *Drosophila melanogaster* (Li *et al*., 2006; Ruan *et al*., 2008). Deuterostomia unites to *Drosophila* at the systematic rank of Coelomata in the traditional sense (Grobben, 1908; Hyman 1951).

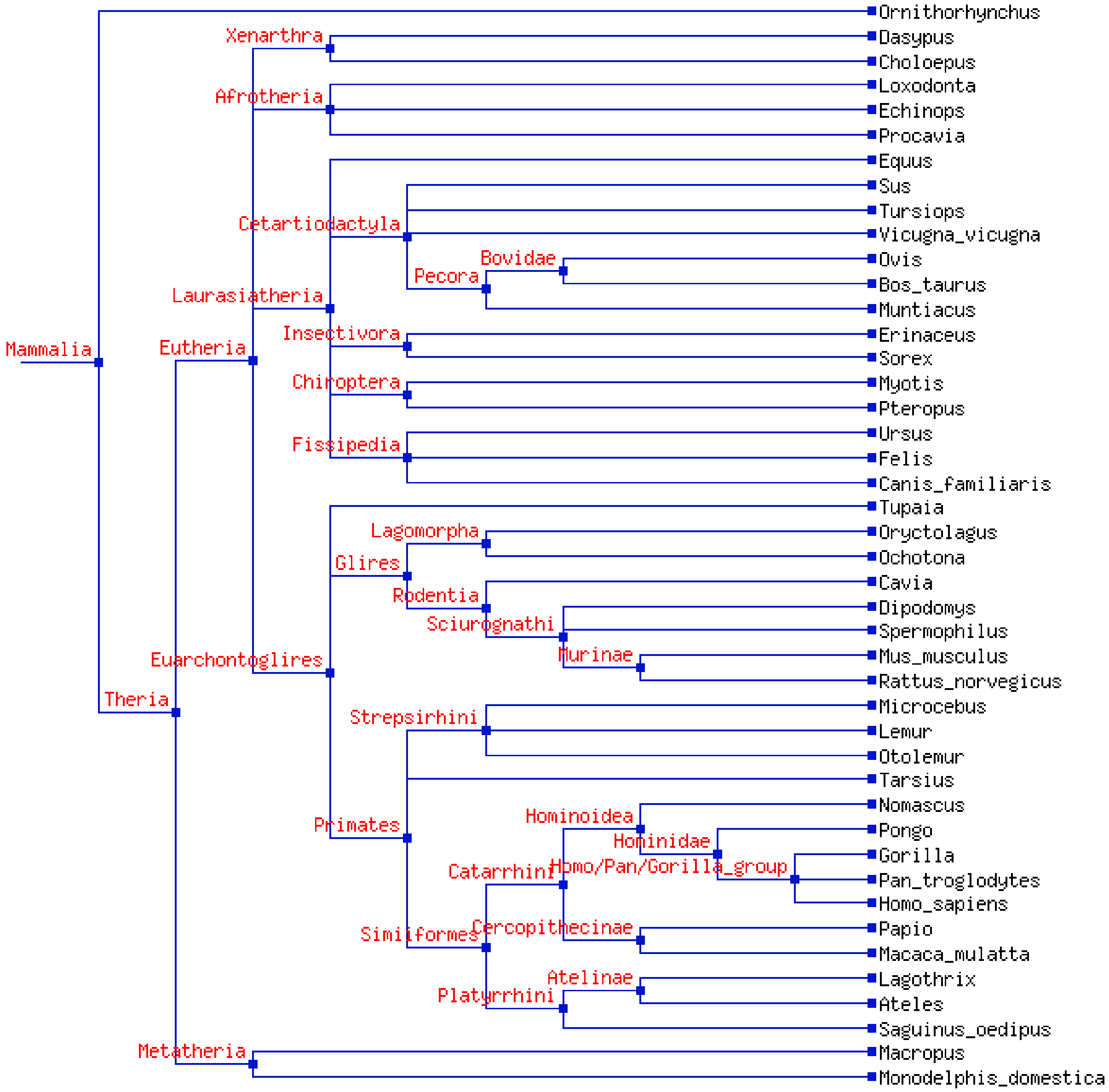


FIGURE 45. Phylogeny of mammalian taxa surveyed for conservation of microRNAs and targets to *Drosophila melanogaster* (Li *et al.*, 2006; Ruan *et al.*, 2008). Mammalian taxa are members of the Dueterostomia (FIGURE 44) and unite to *Drosophila* within the Coelomata in the traditional sense (Grobben, 1908; Hyman 1951).

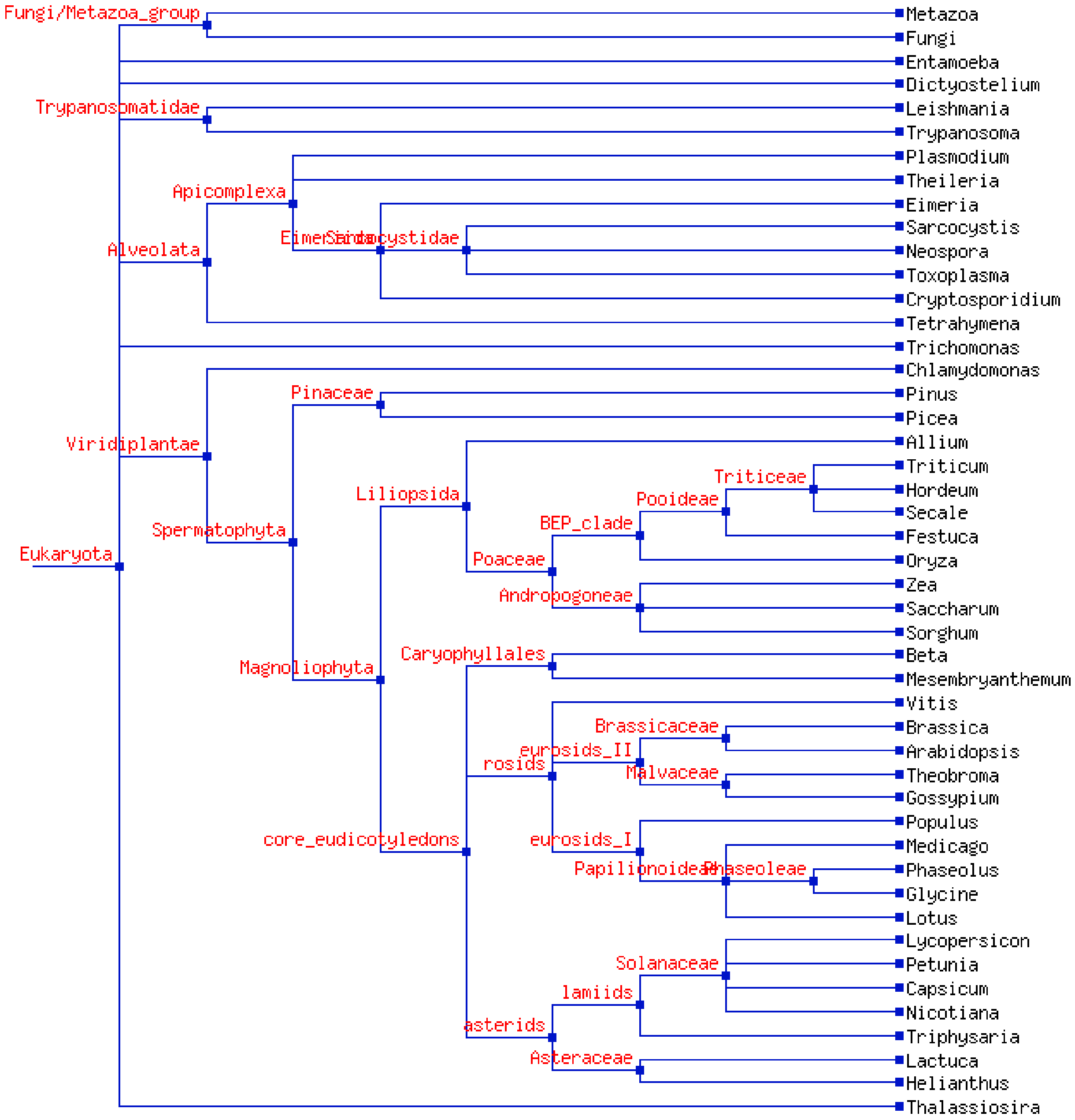


FIGURE 46. Phylogeny of eukaryote taxa surveyed for conservation of microRNAs and targets to *Drosophila melanogaster* (Chatton, 1925; Li *et al.*, 2006; Ruan *et al.*, 2008).

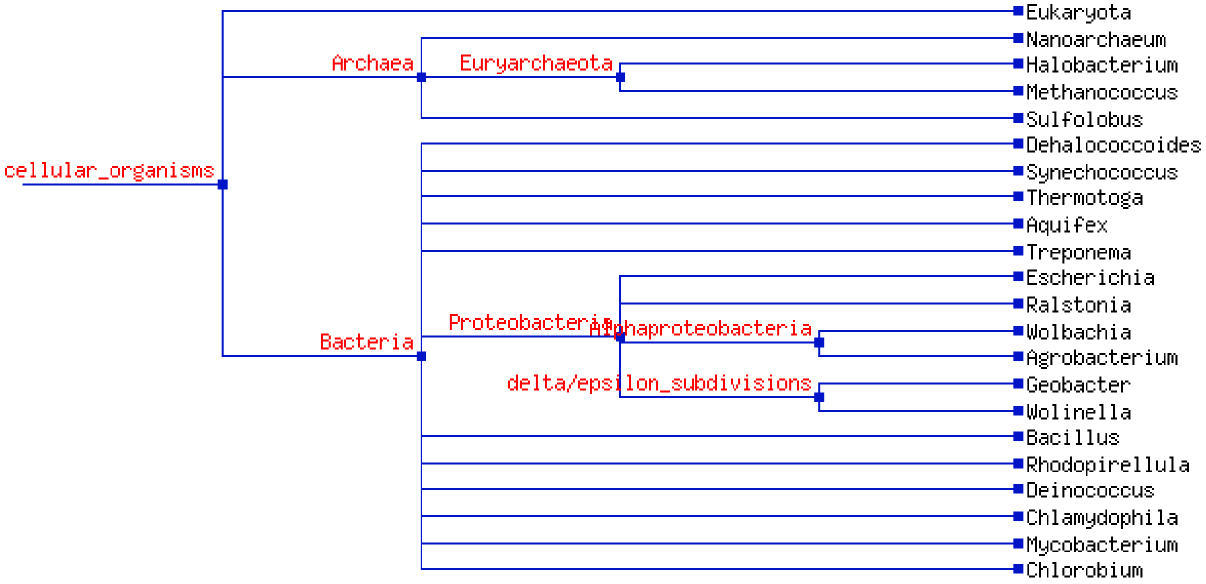


FIGURE 47. Phylogeny of biota surveyed for conservation of microRNAs and targets to *Drosophila melanogaster* (Brands, 2005; Li *et al.*, 2006; Ruan *et al.*, 2008).