Evolution of *cis*-regulatory sequences in *Drosophila*: a systematic approach

by

Daniel Avery Pollard

B.A. (Bowdoin College) 1998

Doctor of Philosophy

in

BIOPHYSICS

and the Designated Emphasis in

COMPUTATIONAL AND GENOMIC BIOLOGY

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Michael B. Eisen, Chair

Professor Ian Holmes

Professor Montgomery Slatkin

Fall 2007

Evolution of *cis*-regulatory sequences in *Drosophila*: a systematic approach

© 2007

by Daniel Avery Pollard

Abstract

Evolution of *cis*-regulatory sequences in *Drosophila*: a systematic approach

Daniel Avery Pollard

Doctor of Philosophy in Biophysics

and the Designated Emphasis in Computational and Genomic Biology

University of California, Berkeley

Michael B. Eisen, Chair

Alterations in gene expression regulatory networks during animal development have been associated with morphological innovation and debilitating disease. Such alterations potentially result from variation in *cis*-regulatory sequences controlling gene expression, yet little is known about the evolutionary forces shaping *cis*-regulatory variation or the specific consequences of such variation. To address two fundamental questions about the selective constraints on and evolutionary dynamics of *cis*-regulatory sequences, I developed a comparative genomics infrastructure in *Drosophila*. This infrastructure included genome-wide binding locations for early embryonic transcription factors in *D. melanogaster*, protein-coding annotations and synteny maps across 11 fully sequenced Drosophila, alignment accuracies for non-coding and *cis*-regulatory sequences, a phylogenetic tree and its variation across chromosomes, and probabilistic methods for identifying conserved and non-conserved functional transcription-factor binding-sites.

Selective constraints on *cis*-regulatory sequences were found to be very strong, with more than six out of ten mutations removed by selection, however, these constraints did not stand out against ubiquitous non-coding constraints and were found to be less than constraints on most ncRNAs and protein-coding sequences. Using estimates of constraint, I find that *cis*-regulatory sequences cover up to 90% of the non-coding genome and binding-sites cover up to 84% of regulatory regions. Despite evidence of strong purifying selection, I infer that binding-sites are gained and lost for some factors on the order of a few percent per million years. Although no evidence was found for compensatory gains and losses, to maintain regulatory function, binding-sites were found to accumulate and dissipate in a lineage-specific fashion, implying either functional variation or rapid restructuring of regulatory regions within loci.

Chair Date

To my family and friends

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would like to thank Mike Eisen for providing an intellectual home for me during graduate school. His shared interest in computational approaches to studying genetics was an obvious fit for me and I am quite grateful that he accepted me into his group at a time when competition to get into the group was fierce. His honesty and enthusiasm helped me develop into an independent scientist.

I would like to thank Casey Bergman for overseeing my early development during graduate school. His technical and sociological mentorship was invaluable.

I would like to thank Venky Iyer and Alan Moses for being outstanding colleagues, collaborators and friends during graduate school. Certainly the most rewarding aspect of science for me is a good discussion or debate, and I can think of nobody who engaged me more in scientific discourse than these two fine people.

I would like to thank Angela DePace for her friendship and boundless support. Angela's clarity of thought and keen sense for priorities helped me resolve countless challenges during my graduate work.

I would like to thank Audrey Gasch, Derek Chiang, Eric Weiss, Justin Fay, Lisa Simirenko, Ben Berman, Hunter Fraser, Heather McCullough, David Nix, Emily Hare, Brant Peterson, Aaron Garnett, Hanchuan Peng, Ryan Shultzaberger, Lenny Teytelman, Stewart MacArthur, Garmay Leung, Colin

Brown, Stuart Davidson, Rich Lusk, Annie Tsong, Erica Rosenblum, Aaron Hechmer, Matt Davis, Dan Richter, Devin Scannell, Genny Gould and all other Eisen Lab affiliates for support and friendship throughout my time in graduate school.

I would like to thank my committee members, Monty Slatkin and Ian Holmes, for guidance and support throughout my graduate work.

I would like to thank my collaborators Sue Celniker, Mark Biggin and Dan Halligan for making the commitment to work with me on overlapping scientific interests.

I would like to thank John Novembre for his friendship and guidance. John very much played the role of my big brother in graduate school.

I would like to thank my family, Tom, Patty and Katie, and my close friends, Suzanne Lee and José Ayerve, for love and support.

Finally, I'd like to thank everyone else who supported my graduate work.

CHAPTER 1

Introduction

My research is motivated by two synergistic questions: what molecular evolutionary processes lead to the phenotypic diversity of life on earth and how can we take advantage of our knowledge of these processes to investigate how phenotypes are encoded in the genome?

While functional evolution in protein coding sequences has been implicated in numerous cases of dramatic phenotypic variation (e.g. (Hoekstra, Hirschmann et al. 2006)), few such examples exist for functional non-coding elements, such as *cis*-regulatory sequences (reviewed in (Wray 2007)), despite the intuitive importance of the genetic control of when and where genes are expressed. This disparity is largely the result of an historic paucity of examples of regulatory sequences and techniques for studying them and a poorly resolved framework for how they mechanistically function.

In recent years, bioinformatic (e.g. (Berman, Nibu et al. 2002)) and biochemical (e.g. (Lieb, Liu et al. 2001)) techniques have been developed to take advantage of whole genome sequences to infer the location of putative regulatory sequences, increasing the number of examples to study. Biochemical and genetic approaches have also closely inspected a small but growing number of regulatory sequences, providing a few basic principles for how regulatory sequences

function (e.g. (Small, Kraut et al. 1991)). Unfortunately these methods are currently prohibitively difficult to apply systematically.

The whole genome sequencing of species, at varying levels of relatedness to model organisms, such as the eleven recently sequenced *Drosophila* species (Clark, Eisen et al. 2007), however, opens the door for an alternative approach to the study of *cis*-regulatory sequences. Examination of how *cis*-regulatory sequences vary over short and long divergence distances has the potential to reveal new details of how *cis*-regulatory sequences function as well as the constraints placed on them by evolutionary forces and the role they play in organismal diversification (Wray, Hahn et al. 2003).

The study of variation in *cis*-regulatory sequences introduces new methodological challenges. Non-coding sequences require comparison at the nucleotide level, unlike protein coding sequences, which can be compared at the slower evolving and higher information amino acid level. *Cis*-regulatory sequences are also comprised of very short, often sparsely organized, functional units, making comparisons much more difficult compared to the highly structured and constrained sequences of protein coding genes. New approaches need to be developed and tested in order for sound conclusions to be made with regard to variation in regulatory sequences.

My graduate work has focused on advancing techniques and resources for comparing *cis*-regulatory sequences across species and utilizing this infrastructure to examine the constraints and dynamics of *cis*-regulatory sequences in the genus

Drosophila. In this introductory chapter I review what *cis*-regulatory sequences are, how they function, how they are modeled and how they evolve. This chapter concludes with an overview of my approaches to my research objectives. In the next two chapters I describe studies I performed examining the quality of automated nucleotide alignments, specifically for the comparison of *cis*-regulatory sequences in *Drosophila*. In the fourth chapter I describe a study I performed examining the phylogeny of the *Drosophila melanogaster* subgroup, another fundamental resource needed for accurate species comparisons. In the fifth chapter I describe a study I performed examining selective constraints on functional non-coding elements, including *cis*-regulatory sequences. In the final chapter I summarize my accomplishments and discuss prospects for future directions. In the Appendices, I describe other studies I worked on during my graduate work, including the generation of *cis*-regulatory annotations using ChIP on chip experiments, the generation of annotations and alignments of the twelve sequenced *Drosophila* species, a method for identifying conserved transcription factor binding sites and an analysis of transcription factor binding site turnover.

What are cis-regulatory sequences and how do they function?

Cis-regulatory sequences are stretches of DNA containing binding sites for sequence-specific transcription factors that act locally to either positively (activation) or negatively (repression) influence the basal level of expression of cognate genes (Davidson 2001).

The mechanisms by which transcription factors independently and combinatorially control transcription rates are understood to fall into the areas of the recruitment of chromatin modifying activities and the recruitment of basal polymerase machinery to core promoters (Carey 1998; Merika and Thanos 2001; Alvarez, Rhodes et al. 2003).

Transcription factors can be organized into families based on their structurally homologous DNA binding domains (e.g. (Finn, Tate et al. 2007)). The span of nucleotides contacted by transcription factors, acting either as monomers or polymers, ranges from as few as three and as many as a few dozen. The overall affinity of transcription factors for their target DNA sequences has been estimated in the millimolar range (Carey and Smale 2001) and typically a subset of the positions across binding sites are involved in determining the majority of the strength of binding (Mirny and Gelfand 2002).

The functional importance of the strength, order, orientation and spacing of binding sites in regulatory regions is poorly understood, though a handful of examples of binding site "grammars" have been reported. One example is the three modes of repression observed in regulatory sequences, where (1) long-range repressors act on a locus scale, allowing their binding sites to have an unrestricted location, (2) short-range repressors act over short distances (<100bp) to "quench" locally bound activators, and (3) steric repressors physically prevent binding of activators with overlapping binding sites (Gray, Cai et al. 1995; Chatterjee, Zhou et al. 1997). Another example is the helical phasing of binding sites within

regulatory regions, indicating that transcription factors can either directly interact or avoid direct interaction by being in the same or opposite helical phase of the DNA (Chiang, Moses et al. 2003; Makeev, Lifanov et al. 2003; Hittinger and Carroll 2007). Much, however, remains unclear regarding the functional importance of regulatory sequence architectures.

How are cis-regulatory sequences modeled and predicted?

While experimental techniques for the identification of transcription factor interactions with specific genes, regulatory regions and binding sites are improving rapidly, there has been and continues to be a need for computational modeling and prediction of regulatory sequences.

Transcription factor binding sites are typically modeled as a consensus sequence (e.g. (Hertz, Hartzell et al. 1990)), a word collection (e.g. (Markstein, Markstein et al. 2002)) or a position weight matrix (PWM) (e.g. (Schneider, Stormo et al. 1986)). Consensus sequences and word collections have the advantage that they only represent known examples of binding sites, eliminating problems with inference errors. PWMs have the advantage that they can be used to infer the binding strength of any sequence based on a limited number of known binding sites.

There are two basic approaches for predicting binding sites: *ab initio* and evidence-based. *Ab initio* methods typically seek to predict binding sites based on

their over-representation in a set of sequences inferred to or known to be co-regulated by the same factor(s) (e.g. (Bussemaker, Li et al. 2001)). Evidence-based methods utilize known binding sites for a given factor and one of the above models to predict new binding sites (e.g. (Schneider, Stormo et al. 1986)). The primary challenge for *ab initio* methods is to be sensitive enough to discover weak signals (Tompa, Li et al. 2005) while evidence-based methods suffer from low specificity in predicting functionally relevant binding sites, particularly in large eukaryotic genomes (Berman, Nibu et al. 2002). For both approaches, evolutionary information has been utilized to increase signal to noise (e.g. (Blanchette and Tompa 2002; Moses, Chiang et al. 2004)).

The modeling of regulatory regions can be broken down into three problems:

(1) Which transcription factors regulate which genes?

(2) Where are the regulatory regions for each gene located?

(3) How does each regulatory region convert the expression patterns of its regulators into the expression pattern for its cognate gene?

The network of transcription factors and their targets has not proved an easy target of computational modeling, however, many successful inferences of regulatory interactions have been made using gene expression analysis, either classically, one gene at a time (e.g. (Nusslein-Volhard and Wieschaus 1980)) or recently, globally (e.g. (Brem, Yvert et al. 2002)). Large-scale efforts to collect high-resolution gene expression data (e.g. (Luengo Hendriks, Keranen et al.

2006)) together with modeling efforts (Reeves, Muratov et al. 2006) may prove a powerful complement to genetic and biochemical approaches.

The most successful techniques for predicting the location of regulatory regions take advantage of the local enrichment of predicted binding sites known to be co-regulators of target genes (e.g. (Berman, Nibu et al. 2002)). *Ab initio* approaches looking for locally enriched motifs have been implemented with less success (e.g. (Xing, Wu et al. 2004)). Evolutionary information has been used to aid in predictions, either on its own (e.g. (Boffelli, McAuliffe et al. 2003)) or in combination with motif finding approaches (e.g. (Berman, Pfeiffer et al. 2004)).

Classically the logic of how a *cis*-regulatory region converts the expression patterns of its trans-acting regulators into the entirety or a portion of the expression pattern of its cognate gene was dissected through genetic manipulation of each system and described qualitatively (e.g. (Stanojevic, Small et al. 1991)). In recent years, however, attempts have been made to infer such logic computationally using quantitative modeling (Schroeder, Pearce et al. 2004; Janssens, Hou et al. 2006; Zinzen and Papatsenko 2007). These models have been based on the relatively simple logic of antagonistic gradients of activators and repressors being integrated by regulatory regions, with some use of spacing requirements for synergistic interactions of binding sites. Physically motivated modeling, while perhaps providing a better picture of reality, suffer from poorly estimated parameters, making naïve modeling approaches a promising avenue for the near future (Reeves, Muratov et al. 2006).

How do cis-regulatory sequence evolve?

Though the study of the evolution of *cis*-regulatory sequences is in its early genesis, some principles have been established. The analysis of the evolution of *cis*-regulatory sequences can be broken down into three related problems:

(1) What is the expected variation in *cis*-regulatory sequences through evolutionary time given the conservation of function?

(2) How does variation in *cis*-regulatory sequences lead to functional variation through evolutionary time?

(3) Are *cis*-regulatory sequences subject to unique evolutionary forces, particularly with respect to protein coding genes?

The dominant principle in *cis*-regulatory sequence evolution is primary sequence constraint. Much of the motivation for whole genome sequencing of closely related species is the identification of *cis*-regulatory sequences based on their relative conservation compared to other non-coding sequences (Pribnow 1975; Hardison 2000). This approach has proven highly successful in mammalian comparisons(Prabhakar, Poulin et al. 2006), though less fruitful in clades with more compact genomes, such as *Drosophila* (Sinha, Schroeder et al. 2004). Despite the general assumption that regulatory sequences evolve under appreciable constraints, little previous work has been done to quantify these selective constraints.

In addition to primary sequence constraints, some work has been done identifying higher order constraints on regulatory sequences. One such constraint is the maintenance of binding site composition. The theory is that stabilizing selection maintains the composition of binding sites while allowing individual sites to come and go through evolutionary time (Ludwig, Bergman et al. 2000). This process is sometimes referred to as binding site turnover (Costas, Casares et al. 2003). It is not clear if this principle will prove to be sufficient to explain the observed variation in regulatory regions or if more nuanced, functionally motivated models will be required.

Variation in phenotype and/or gene expression patterns has been associated with *cis*-regulatory sequences in a small but growing number of studies (reviewed in (Wray 2007)). Phenotypes affected by *cis*-regulatory sequences include behavioral, physiological and morphological and span model organisms, though most have been identified in either fruitflies or humans. In only a few cases have specific sequence differences identified that can explain all or most of the phenotypic variation. Most of these examples have involved the lineage-specific gain or loss of one or a few transcription factor binding sites (e.g. (Tournamille, Colin et al. 1995; Hamblin and Di Rienzo 2000; Costas, Pereira et al. 2004; Gompel, Prud'homme et al. 2005; Rockman, Hahn et al. 2005; Jeong, Rokas et al. 2006)). Duplication and divergence is another mechanism by which new outputs may be achieved while maintaining ancestral functions (Hittinger and Carroll 2007). In this case, the helical phasing of binding sites appears to have

played a large role in the divergence of the duplicate *cis*-regulatory regions. Our ability to identify and understand *cis*-regulatory variation that affects phenotype will improve together with our understanding of the mechanisms of *cis*-regulatory control of gene expression.

The degree to which *cis*-regulatory sequences are subject to qualitatively different selective forces than protein coding sequences remains to be revealed, but two theories currently have some support. The first suggests that because expression is allele specific (e.g. (Wittkopp, Haerum et al. 2004)), *cis*-regulatory mutations will tend to be semi-dominant (e.g. (Ruvkun, Wightman et al. 1991)). Selection would then be expected to act on mutations immediately, unlike protein-coding mutations, which tend to be recessive and therefore must rise in frequency in the population before being subject to selection in homozygotes. The second theory is that the functional modularity of *cis*-regulatory sequences leads to lower pleiotropy than protein coding genes (e.g. (Iwamoto, Li et al. 1996; Gompel, Prud'homme et al. 2005)). Both of these theories suggest selection ought to be more efficient in *cis*-regulatory sequences compared to protein coding genes but the generality of these principles remains to be tested.

Overview of Approach

A systematic approach to studying the evolutionary properties of *cis*-regulatory sequences requires considerable infrastructure for testing hypotheses. My graduate work has attempted to put together the numerous pieces of this

infrastructure and then has utilized the infrastructure to begin what will hopefully be of the first of many global analyses of evolutionary principles for *cis*-regulatory sequences.

The infrastructure can be broken down into five basic categories (chapters where each is addressed are noted):

(1) *Cis*regulatory sequence annotations (Appendix A)

(2) Orthologous non-coding sequence mapping (Appendix B & C)

(3) Non-coding multiple alignment (Chapters 2 & 3)

(4) Phylogenetic trees (Chapter 4)

(5) Computational hypothesis testing (Appendix D & E)

Using this framework I approached two areas of evolutionary principle (chapters where each is addressed are noted):

(1) Selective constraints on functional non-coding elements (Chapter 5)

(2) Phylogenetic dynamics of transcription factor gains and losses (Appendix E)

CHAPTER 2

Variation in accuracy across tools for the pairwise alignment of functionally constrained non-coding DNA

Abstract

Numerous tools have been developed to align genomic sequences. However, their relative performance in specific applications remains poorly characterized. Alignments of protein-coding sequences typically have been benchmarked against "correct" alignments inferred from structural data. For non-coding sequences, where such independent validation is lacking, simulation provides an effective means to generate "correct" alignments with which to benchmark alignment tools. Using rates of non-coding sequence evolution estimated from the genus *Drosophila*, I simulated alignments over a range of divergence times under varying models incorporating point substitution, insertion/deletion events, and short blocks of constrained sequences such as those found in *cis*-regulatory regions. I then compared "correct" alignments generated by a modified version of the ROSE simulation platform to alignments of the simulated derived sequences produced by eight pairwise alignment tools (Avid, BlastZ, Chaos, ClustalW, DiAlign, Lagan, Needle, and WABA) to determine the off-the-shelf performance of each tool. As expected, the ability to align non-coding sequences accurately decreases with increasing divergence for all tools,

and declines faster in the presence of insertion/deletion evolution. Global alignment tools (Avid, ClustalW, Lagan, and Needle) typically have higher sensitivity over entire non-coding sequences as well as in constrained sequences. Local tools (BlastZ, Chaos, and WABA) have lower overall sensitivity as a consequence of incomplete coverage, but have high specificity to detect constrained sequences as well as high sensitivity within the subset of sequences they align. Tools such as DiAlign, which generate both local and global outputs, produce alignments of constrained sequences with both high sensitivity and specificity for divergence distances in the range of 1.25–3.0 substitutions per site. For species with genomic properties similar to *Drosophila*, I conclude that a single pair of optimally diverged species analyzed with a high performance alignment tool can yield accurate and specific alignments of functionally constrained non-coding sequences. Further algorithm development, optimization of alignment parameters, and benchmarking studies will be necessary to extract the maximal biological information from alignments of functional non-coding DNA.

Background

The increasing availability of genome sequences of related organisms offers myriad opportunities to address questions in gene function, genome organization and evolution, but also presents new challenges for sequence analysis. Many classical tools for sequence analysis are obsolete, and there has

been active effort in recent years to develop tools that work efficiently with whole genome data. Aligning long genomic sequences – the first step in many analyses – is substantially more complex and computational taxing than aligning short sequences, and many methods have been developed in recent years to address this challenge (reviewed in (Miller 2001; Frazer, Elnitski et al. 2003)). Nevertheless, comparative genomic researchers are still faced with the task of making decisions such as which alignment tools to use and which genomes to compare for their particular application. Benchmarking studies that address both the selection of alignment methods and the choice of species can provide the needed framework for informed application of genomic alignment tools and biological discovery in the field of comparative genomics.

Research in alignment benchmarking has focused on the alignment of protein-coding sequences (McClure, Vasi et al. 1994; Thompson, Plewniak et al. 1999), where independent evidence (either the three-dimensional structure of a protein sequence (Brenner, Chothia et al. 1998; Sauder, Arthur et al. 2000) or cDNA sequence (Bray, Dubchak et al. 2003; Brudno, Do et al. 2003)) is available to use as a "gold standard" to assess the relative performance of different alignment tools. In contrast, little is known about the relative performance of tools to align non-coding sequences, which comprise the vast majority of metazoan genomes and contain many functional sequences including *cis*-regulatory elements that control gene regulation. For non-coding sequences, little external evidence is available to evaluate alignment tool performance. Benchmarking,

however, can be achieved by simulating sequence divergence in silico where it is possible to generate sequences that are related by a known, "correct" alignment (Stoye, Evers et al. 1998). Simulation experiments have been used extensively to assess the performance of different methods for phylogenetic reconstruction (Hillis, Huelsenbeck et al. 1994). Yet only a few studies to date have exploited simulated data to benchmark alignment tools (Thorne, Kishino et al. 1991; Thorne, Kishino et al. 1992; Holmes and Durbin 1998; Stoye 1998; Hein, Wiuf et al. 2000; Katoh, Misawa et al. 2002; Lassmann and Sonnhammer 2002; Metzler 2003), and currently none have done so explicitly for the purposes of functional non-coding sequence alignment.

Here I present results of a simulation-based benchmarking study designed to assess the performance of eight tools (Avid (Bray, Dubchak et al. 2003), BlastZ (Schwartz, Kent et al. 2003), Chaos (Brudno, Do et al. 2003), ClustalW (Thompson, Higgins et al. 1994), DiAlign (Morgenstern 1999), Lagan (Brudno, Do et al. 2003), Needle (Rice, Longden et al. 2000), and WABA (Kent and Zahler 2000)) for the pairwise alignment of non-coding sequences. I have chosen to address the question of pairwise alignment since pairwise alignment methods often are used in the construction of multiple alignments, since the evaluation of pairwise alignment performance is more tractable than that of multiple alignment, and since pairwise alignment performance is an important part of a general assessment of non-coding alignment strategies. I have chosen to model non-coding sequence evolution in the genus *Drosophila* as a biological system for

methodological evaluation, because of the high quality sequence and annotations available for *D. melanogaster* (Celniker, Wheeler et al. 2002; Misra, Crosby et al. 2002), and the recent availability of the genome sequence for the related species, *D. pseudoobscura* (Richards, Liu et al. 2005). In addition, because of the high rate of deletion as well as the relatively low density of repetitive DNA as compared with mammalian genomes (Petrov, Lozovskaya et al. 1996; Petrov and Hartl 1998; Kaminker, Bergman et al. 2002), *Drosophila* non-coding regions are likely to be enriched for sequences under functional constraint. Previous results indicate that *Drosophila* non-coding regions contain an abundance of short blocks of highly conserved sequences, but that the detection of these sequences is dependent on the alignment method used (Bergman and Kreitman 2001). Optimizing strategies for the accurate identification of functionally constrained non-coding sequences will play a critical role in the annotation of *cis*-regulatory elements and other important non-coding sequences in *Drosophila* as well as other metazoan genomes.

In this study, I use empirically-derived estimates to parameterize simulations of non-coding sequence evolution over a range of divergences that includes those between species commonly used in comparative genomics such as *H. sapiens*-*M. musculus* (Nekrutenko, Makova et al. 2002; Waterston, Lindblad-Toh et al. 2002), *C. elegans*-*C. briggsae* (Castillo-Davis and Hartl 2002; Stein, Bao et al. 2003) and *D. melanogaster*-*D. pseudoobscura* (Zeng, Comeron et al. 1998; Bergman, Pfeiffer et al. 2002). Alignments of simulated descendent

sequences produced by the tools under consideration were compared to correct alignments and various performance measures were calculated. In general, I find that global tools (Avid, ClustalW, DiAlign-G, Lagan, and Needle), which align the entirety of input sequences, tend to have the highest accuracy over entire sequences as well as within interspersed blocks of constrained sequences, but both measures were decreasing functions of divergence. Local tools (BlastZ, Chaos, DiAlign-L, and WABA), which align subsets of input sequences, tend to have the highest accuracy for the portion of the sequences they align, but the proportion of sequences included in their alignments decreased quickly with increasing divergence distance. For intermediate to high divergences, local tools also showed a high specificity for only aligning interspersed blocks of constrained sequences. Despite these general trends, I find that some tools can systematically out-perform others over a wide range of divergence distances. These results should prove useful for comparative genomics researchers and algorithm developers alike.

Results

Properties of non-coding DNA in Drosophila

To make my simulation results as biologically meaningful as possible, I estimated properties of non-coding regions in *D. melanogaster* using Release 3 euchromatic genome sequences and annotations (Celniker, Wheeler et al. 2002; Misra, Crosby et al. 2002). As described in the methods, I masked all annotated coding exons and known transposable elements to derive a data set of unique

sequences representative of non-coding regions in the *D. melanogaster* genome. In total, I obtained 55,325 non-coding regions ranging in size from 1 to 156,299 bp with two modes at approximately 70 and 500 bp (Figure 2.1). Greater than 95% of non-coding sequences in the *D. melanogaster* genome are less than 10 Kb in length, thus 10 Kb was used as the sequence length for my simulations. Nucleotide frequencies derived from this set of non-coding regions were used to parameterize both my model of non-coding DNA as well as my substitution model used in my simulations.

Estimates of divergence between taxa used in comparative genomics

To link my simulations to species commonly used in comparative genomic analyses of non-coding DNA, I estimated silent site divergence (Ks) between *H. sapiens* vs. *M. musculus*, *C. elegans* vs. *C. briggsae*, and *D. melanogaster* vs. *D. pseudoobscura* (see methods). Since estimates of Ks are highly dependent on methodology, I sought to generate estimates between these three species pairs using a single method. I estimate the mean (and median) of Ks measured in expected number of substitutions per silent site, for these species pairs to be: *H. sapiens* vs. *M. musculus* 0.64 (0.56); *C. elegans* vs. *C. briggsae*, 1.39 (1.26); and *D. melanogaster* vs. *D. pseudoobscura*, 2.40 (2.24). I note that these divergence estimates do not underlie my simulation, but rather are intended to frame the interpretation of my simulation results in a biological context.

Table 2.1. Summary of parameters used in simulations

Parameters	Value	Source	Reference
Sequence Length	10kb	*Dmel*	-
AT : GC	60 : 40	*Dmel*	(Bergman, Pfeiffer et al. 2002) (Moriyama and Hartl 1993)
Transition / Transversion Bias	2	*Drosophila spp.*	(Bergman and Kreitman 2001) (Moriyama and Powell 1996)
Substitution Model	HKY	-	(Hasegawa, Kishino et al. 1985)
Point Substitutions : Indels	10 : 1	*Drosophila spp.*	(Bergman and Kreitman 2001) (Petrov, Lozovskaya et al. 1996) (Petrov and Hartl 1998)
Indel Spectrum	-	*Dmel*	(Comeron and Kreitman 2000)
Median Constrained Block Length	18bp	*Dmel* vs. *Dvir*	(Bergman and Kreitman 2001)
Mean Density of Constrained Blocks	0.2	*Dmel* vs. *Dvir*	(Bergman and Kreitman 2001)

Simulating non-coding sequence evolution

Using a model of non-coding DNA, parameterized with *D. melanogaster* nucleotide frequencies (see Methods for details), I generated 10 Kb sequences which were used as "ancestral" inputs to the ROSE sequence evolution simulation program (Stoye, Evers et al. 1997; Stoye, Evers et al. 1998) to create pairs of "derived" output sequences. It is important to note that ROSE provides both pairs of derived sequences and their correct alignment, and that the modifications to ROSE implemented here allow ancestral constraints to be mapped onto derived sequences. Sequence evolution in ROSE occurred under four simulation regimes: A) without insertion/ deletion (indel) evolution and without constrained blocks; B) with indel evolution and without constrained blocks; C) without indel evolution and with constrained blocks; and D) with indel evolution and with constrained blocks. Regime D is the most realistic and relevant for the interpretation of real biological data. Other regimes were used to calibrate the

outputs of my simulations and address the effects of different models of evolution on non-coding sequence alignment. Under each regime, 1,000 replicate pairs of sequences were evolved to each of eleven divergence distances ranging from 0.25 to 5.0 substitutions per site. Levels of constraint as well as relative evolutionary rates of constrained to unconstrained sites and of indels to point substitution were chosen based on previously reported estimates from the literature (see Table 2.1 and Methods).

Characterization of simulation outputs

To characterize simulation outputs, derived pairs of sequences in alignments provided by ROSE were analyzed for the following measures: estimated overall divergence, estimated divergence in constrained blocks, estimated divergence in unconstrained blocks, overall identity, identity in constrained blocks, identity in unconstrained blocks, fraction of ancestral sequence remaining, fraction of sequences constrained, and differences in length. These simulation statistics are summarized in Figure 2.2 and demonstrate that the expected outputs of my simulations are observed. In the absence of constrained blocks, estimated overall divergences correspond well with the input distance parameters up to 3.0–4.0 substitutions per site (Figure 2.2A and 2.2B, black boxes). In the presence of constrained blocks, estimated overall divergences (Figure 2.2C and 2.2D, black boxes) are less than the input distance parameters because these sequences are made up of both unconstrained sites evolving at the

rate set by the input parameter (Figure 2.2C and 2.2D, brown triangles) as well as blocks of constrained sites evolving ten times more slowly (Figure 2.2C and 2.2D, grey circles). The more pronounced deviation of the estimated overall divergences from the input distance parameters in the regime with indel evolution (Figure 2.2C vs. 2.2D) is due to preferential deletion of sequence under no constraint which enriches for constrained sites and leads to a decrease in estimated divergences.

Overall identity between derived pairs in the regimes without constrained blocks decreases to the random background of 0.26 (the sum of the squares of the mononucleotide frequencies) by 5.0 substitutions per site with and without indel evolution (Figure 2.2A and 2.2B, red crosses). In the regimes with constrained blocks, unconstrained sites have the same level of identity as entire sequences in the regimes without constrained blocks (Figure 2.2C and 2.2D, green diamonds), whereas the identity in the constrained blocks is much greater (Figure 2C and 2D, yellow x's). In the regimes with indel evolution, the fraction of the ancestral sequence remaining diminishes most quickly in the absence of constrained blocks (Figure 2.2B, green triangles). In regime C (with constrained blocks and without indel evolution), the fraction of constrained sites in derived sequences matches the input parameter of 0.2 (Figure 2.2C, blue checked-boxes). However, in regime D (with constrained blocks and indel evolution), the fraction of constrained sites in derived sequences decreases below the input parameter of 0.2 at large divergence distances (Figure 2.2D, blue checked-boxes). This is because the derived

sequences are on average longer than ancestral sequences in regime D, differing by 300–400 bp at 1 substitution per site, 400–500 bp at 2 substitutions per site and 700–800 bp at 5 substitutions per site. In my simulation there are equal input rates of insertion and deletion, however deletions are unable to extend into constrained blocks and are omitted, creating a net excess of insertions to deletions. This phenomenon was recently proposed as a possible explanation for differences in observed insertion:deletion ratios in unconstrained dead-on-arrival retrotransposon pseudogenes versus non-coding sequences flanking genes (Ptak and Petrov 2002).

Comparative analysis of genomic alignment tools

Unaligned pairs of derived sequences generated by ROSE were used as input to each of the eight genomic alignment tools (see Methods) and resulting alignments were compared to the simulated alignments produced by ROSE. My objective was to test the off-the-shelf performance of these tools over a wide range of different divergences, so each tool was run using default parameter settings. In addition, BlastZ and Chaos were run using author suggested settings (BlastZ-A and Chaos-A), as described in the Methods. I note that the output of DiAlign can be treated as both a global alignment as well as a local alignment, so I analyzed both (DiAlign-G and DiAlign-L). Alignments produced by each tool were scored for the overall coverage and overall sensitivity for all regimes (A–D), and were also scored for constraint coverage, constraint sensitivity, constraint

specificity, and local constraint sensitivity in the regimes with constrained blocks (C and D) (see Methods for details).

Coverage

Overall coverage was measured to understand the proportion of ungapped, orthologous pairs of sites in the simulated alignment that were aligned by local tools under various evolutionary scenarios. The coverage of each tool under the four simulation regimes is a decreasing function of divergence for local (but not global) tools (Figure 2.3). In the absence of constrained blocks, local tools tend to align most or all of the sequences for only small divergence distances (0.25–1.0 substitutions per site), but little or none of the sequences for intermediate to large divergence distances (Figure 3.3A and 3.3B). [For convenience, for the remainder of this report I shall refer to 0.25–1.0 substitutions per site as small distances, 1.25–3.0 substitutions per site as intermediate distances, and 4.0–5.0 substitutions per site as large distances.] One exception is Chaos, which has negligible coverage past 0.25 substitutions per site. In the presence of constrained blocks, the coverage of local tools improves substantially at all but the most extreme divergence distances. WABA, which was typical of local tools in the absence of constrained blocks, maintains high coverage out to more than twice the divergence distance of the rest of the local tools in the presence of constrained blocks. WABA also appears to be relatively unaffected by indel evolution, while

the other local tools show a reduction in coverage of about 0.5 substitutions per site in regimes with indel evolution (Figure 3.3A vs. 3.3B, 3.3C vs. 3.3D).

Sensitivity

Overall sensitivity was measured to understand the accuracy of each tool to align all orthologous nucleotide sites under various evolutionary scenarios. The sensitivity of each tool under the four simulation regimes is a decreasing function of divergence for both local and global tools (Figure 2.4). It is important to note that the maximum sensitivity a tool can attain is limited by its coverage. Thus for most divergence distances, global tools (which by definition have complete coverage) have greater potential for high sensitivity relative to local tools, which have incomplete coverage (see above, Figure 2.3). Nevertheless, with the exception of WABA, the sensitivity of local tools tends to remain very close to the maximum set by their coverage. This implies that although local tools have diminishing coverage with divergence, the portion of the sequence they do align is aligned quite accurately (see below). Despite the trend of high sensitivity in aligned regions for local tools, the sensitivity of the top global tools tends to be as good as or better than the sensitivity for the top local tools (Figure 2.4). This is particularly true for intermediate to high divergence distances in the absence of indel evolution. In each of the four regimes, at least one global tool has a higher sensitivity than the next best local tool for intermediate to high divergence distances. In the most biologically relevant regime D, the sensitivity of the highest

performing tools (such as Lagan and DiAlign) plateau over the range of 1.25–3.0 substitutions per site at higher than 0.35, implying that sites other than those in constrained blocks are being accurately aligned (Figure 2.4D). In contrast, in the absence of constraint but with indels (regime B), the sensitivity of all alignment tools is practically nil for divergences greater than 1 substitution per site (Figure 2.4B).

Coverage and sensitivity in constrained sequences

Alignment coverage and sensitivity across all orthologous sites are informative for understanding the overall performance of a tool, but, for many applications (such as aligning characterized *cis*-regulatory elements), researchers may only be interested in accurately aligning functionally constrained sites. To assess the ability of each tool to align potentially functional portions of sequences I measured the coverage and sensitivity only for orthologous nucleotide sites within constrained blocks (Figure 2.5). Constraint coverage is better than overall coverage for local tools but the degree of improvement varies considerably (Figure 2.5A and 2.5B). WABA has very similar overall and constraint coverage, suggesting little discrimination in attempting to align constrained versus unconstrained sites. In contrast, BlastZ, Blastz-A, DiAlign-L and Chaos-A have much improved constraint coverage compared with overall coverage, suggesting a preferential alignment of constrained sites.

Constraint sensitivity of all tools is much better than overall sensitivity but, as with constraint coverage, the degree of improvement varies considerably across tools (Figure 2.5C and 2.5D). Similar to overall sensitivity, global tools tend to maintain the highest sensitivity out to large divergence distances in the presence of constrained sites. It is of note that in the presence of indel evolution (Figure 2.5D), constraint sensitivity of the best performing global tools (as well as the local Dialign-L) closely parallels the decrease in identity of constrained sites (Figure 2.2D). Most tools show only moderate decreases in constraint sensitivity in the presence of indel evolution but a few, like ClustalW, Chaos-A, and BlastZ have dramatic decreases in constraint sensitivity in the presence of indel evolution.

Specificity to detect constrained sequences

Constraint coverage and constraint sensitivity reveal the ability of alignment tools to detect and align all orthologous nucleotides sites within constrained blocks, but for some purposes (like *cis*-regulatory element prediction) researchers may want to align only constrained nucleotide sites and nothing else, even at the expense of missing some functionally constrained sites. To evaluate the ability of each tool to provide high quality alignments of just potential functionally constrained sites, I measured their constraint specificity and local constraint sensitivity. As shown in Figure 2.6, constraint specificity is an increasing function of divergence for most tools because unconstrained sequences

accumulate mismatches and indels more quickly than the constrained blocks and are thus more likely to be gapped or left out of local alignments. This is particularly true for local tools where decreasing coverage can increase constraint specificity, and less so for global tools for which it is gap parameters that predominantly affect constraint specificity at different divergence distances. Most tools have higher constraint specificity in the presence of indel evolution, although this trend is less pronounced in the highest specificity tools, Chaos and DiAlign-L. All local tools except WABA increase quickly until they reach a constraint specificity of 0.8–0.9 at which point their constraint specificity plateaus. In the presence of indel evolution, near-maximal constraint specificity is achieved between 1.25 and 3.0 substitutions per site.

Local constraint sensitivity (Figure 2.6) is equivalent to constraint sensitivity (Figure 2.5) for the global tools, but for the local tools it differs in that it is a measure of their constraint sensitivity just within the subsequences they align. For BlastZ, BlastZ-A, Chaos, and DiAlign-L, local constraint sensitivity is nearly maximal (1.0) with and without indel evolution across all divergences studied. For Chaos-A and WABA, local constraint sensitivity varies with divergence distance and is less than the other local tools. Thus local tools can produce nearly perfect alignments within constraint blocks while maintaining relatively high constraint specificity, though it is important to note that this may

not be meaningful if the coverage of a tool is extremely low (e.g. BlastZ, BlastZ-A, Chaos).

Discussion

In this report I investigate the performance of eight pairwise genomic alignment tools to align functional non-coding DNA such as that found in metazoan *cis*-regulatory regions. To do so, I have used a biologically-informed simulation approach to determine off-the-shelf performance over a range of divergence distances. This study provides important information regarding the ability of genomic alignment tools to identify and align constrained sequences in non-coding regions, which would not otherwise be possible. I argue that a simulation study is necessary to achieve my goal since large datasets of functionally annotated non-coding sequences are not available to use as "gold standards" of alignment accuracy. Likewise, datasets of large orthologous genomic regions spanning a range of divergence distances are only recently becoming available (Bergman, Pfeiffer et al. 2002; Thomas, Touchman et al. 2003). As is common in alignment benchmarking (Morgenstern, Frech et al. 1998; Thompson, Plewniak et al. 1999; Lassmann and Sonnhammer 2002), I have studied performance of alignment tools using default parameters since fundamental differences in objective functions, scoring matrices, the type and values of parameters, and algorithmic design prevent a systematic exploration of parameter space.

I have attempted to construct a realistic simulation of non-coding sequence evolution and test alignment performance for species with genomic properties similar to *Drosophila*. Non-coding alignment assessment for mammalian and other species with large, repeat-rich genomes would require modifications to my current simulation, such as the inclusion of ancestral repeats and lineage-specific transposition events. Moreover, as more becomes known about the substitution process in non-coding regions (especially those under weak primary sequence constraint), it will be important to implement more realistic models such as context-dependent substitution (Averof, Rokas et al. 2000; Arndt, Burge et al. 2003; Siepel and Haussler 2004). It would be also instructive to assess alignment performance based on a simulation that decouples suppression of indel rates from substitution rates, given the possibility that the spacing (but not the primary sequence) between conserved non-coding segments may be constrained (Bergman, Pfeiffer et al. 2002). In addition, though I have attempted to be systematic in my evaluation of tools, I unfortunately cannot have included all available pairwise alignment tools. Moreover, assessment of tools which take advantage of the phylogenetic information and higher signal-to-noise inherent in multiple alignments will be an essential extension to this work to provide a more general evaluation of strategies for non-coding alignment.

From the standpoint of the most biologically relevant simulation regime studied here (D, which includes indel evolution and interspersed blocks of constrained sequences), my results indicate that global alignment tools have the

highest sensitivity in general to align orthologous sites accurately in non-coding sequences, as well as blocks of constrained sites (Figures 2.4D, 2.5D). I find that constraint sensitivity of the top global tools can be quite high (>75%) and limited only by sequence identity in constrained sites at intermediate divergence distances (1.25–3.0 substitutions per site), whereas overall sensitivity is relatively low beyond such intermediate divergence distances. The improved performance of global tools over local tools is largely a consequence of incomplete coverage of both constrained and unconstrained sites in alignments produced by local tools (Figure 2.3). The subset of sequences aligned by the highest performing local tools, however, is accurately aligned and specifically corresponds to constrained sites (Figure 2.6). In fact, most local tools can effectively discriminate between constrained and unconstrained sites to greater than 80% specificity at intermediate divergence distances while the constrained portions of their alignments are nearly perfectly aligned at large divergence distances. Finally, when compared with regime C (which excludes indel evolution but includes interspersed constrained blocks), it is clear that my model of indel evolution affects alignment coverage, sensitivity and specificity, but not enough to overturn these major trends.

These results have important implications for the analysis of functional non-coding sequences. First, if a researcher's goal is to align all constrained sites in a non-coding region, then a global tool like Lagan will reliably produce the best results, but will require post-processing to identify constrained sequences (Boffelli, McAuliffe et al. 2003; Elnitski, Hardison et al. 2003). Conversely, if

one's goal is to align only constrained blocks in a non-coding region, then a local tool like Chaos will reliably produce the best results, provided that complete recovery of all constrained sequences is not required. The distinct virtues of both global and local tools are currently incorporated in the output of only one alignment tool, DiAlign. For this reason, use of the global parse of DiAlign (DiAlign-G) can provide high coverage and sensitivity across entire non-coding regions, while use of the local parse of DiAlign (DiAlign- L) will specifically provide highly accurate alignments of blocks of constrained sites. In light of these results, I recommend the further development of global alignment tools that also output a local parse of high confidence local alignments contained within, which should be possible since local anchors are often used in the construction of the global alignment (e.g. (Bray, Dubchak et al. 2003; Brudno, Do et al. 2003)).

My results also indicate that for species with structural and evolutionary constraints on non-coding sequences such as those found in *Drosophila*, DiAlign can produce alignments with high coverage and sensitivity, as well as high specificity to detect constrained sites in the range of 1.25–3.0 substitutions per site. Since the divergence between *D. melanogaster* vs. *D. pseudoobscura* and between *C. elegans* vs. *C. briggsae* falls within this range, I suggest that the use of DiAlign for detecting functionally constrained non-coding sequences will prove successful in these taxa on a genomic scale. In contrast, my results also indicate that species pairs such as *H. sapiens* and *M. musculus* may not be sufficiently diverged for a single pairwise comparison to provide the needed

resolution to detect functionally constrained non-coding sequences, though differences in genome organization and evolution between flies and mammals require a more thorough evaluation of this claim. This conclusion, however, supports results based on Poisson modelling of point substitution that approximately 3 substitutions per site would be needed to detect functional constrained sites reliably in mammalian non-coding DNA (Cooper, Brudno et al. 2003).

Finally, the results presented here also imply that biological and technical conditions exist with which to study with accuracy the evolutionary events underlying the process of *cis*-regulatory evolution in flies and worms. Current evolutionary models of *cis*-regulatory sequence divergence posit the gain and loss of transcription factor binding sites, even under constant functional constraints (Ludwig, Bergman et al. 2000; Cuadrado, Sacristan et al. 2001). However, the absence of alignable binding sites in comparisons of divergent sequences may result from inaccuracies in alignment as well as the bona fide loss of transcription factor binding sites. I suggest that alignments of non-coding sequences using tools such as DiAlign in the range of 1.25–3.0 substitutions per site are of sufficient accuracy to measure binding site loss among divergent species pairs, such as the high levels recently reported in the genus *Drosophila* (Costas, Casares et al. 2003; Emberly, Rajewsky et al. 2003).

Conclusions

Our study demonstrates that recently developed alignment tools have the potential to produce biologically meaningful alignments of functional non-coding DNA on a genome scale. Continued development of alignment algorithms in conjunction with parameter optimization and continued benchmarking will be necessary to provide the highest quality genomic alignments under the wide diversity of genomic and evolutionary scenarios to be studied.

Methods

Modelling input sequences for the simulation of Drosophila non-coding DNA

To generate biologically relevant input sequences for my simulation, I estimated properties of non-coding sequences in the genome sequences of the fruitfly, *D. melanogaster*. First I extracted all non-coding regions from the Release 3 *D. melanogaster* genomic sequences based on annotations in the Gadfly database (Celniker, Wheeler et al. 2002; Misra, Crosby et al. 2002; Mungall, Misra et al. 2002). This was accomplished by masking all DNA corresponding to coding exons, producing inter-coding-exon intervals. Subsequent to extracting non-coding regions, transposable elements were masked using annotations in Gadfly to create "pre-integration" non-coding sequences. In my analysis, I chose to treat all non-coding sequences (intergenic, intronic, untranslated region) together since many non-coding sequences cannot be unambiguously categorized because of alternative splicing or alternative promoter usage. Moreover, previous results revealed that similar evolutionary constraints act on intergenic and intronic

sequences in *Drosophila* (Bergman and Kreitman 2001). Summary statistics of non-coding sequence lengths were calculated using the R statistical package (Figure 2.1) (Ihaka 1996).

The probabilistic dependence of adjacent bases in *D. melanogaster* non-coding sequences was assessed by Markov chain analysis in order to create an accurate model of random non-coding sequences (Weir 1996). TE-masked non-coding sequences were concatenated, and n-mers of size 1 to 10 were counted. Counts of reverse complementing n-mers were averaged, and used to estimate frequencies of each nmer (Burge, Campbell et al. 1992). Based on these counts and frequencies, I determined the likelihood of Markov chains of orders 1 through 9 describing *Drosophila* non-coding sequences, and evaluated the likelihood of each Markov chain using the Bayesian information criterion (Katz 1981; Weir 1996). This analysis revealed that *D. melanogaster* non-coding sequences are best modeled by a 7th-order Markov chain (data not shown). I therefore created the ancestral input sequences for my evolution simulations using a 7th-order Markov chain. I note that because my evolutionary simulation models bases independently (see below), the higher order structure of these ancestral input sequences was not maintained in the more divergent derived output sequences. Nevertheless, sequences generated by a 0thorder Markov chain gave qualitatively and quantitatively similar simulation and alignment results, with correlation among performance measures for the 0th-order and 7thorder generated sequences exceeding an r^2 of 0.97 (data not shown).

Divergence estimates in flies, worms and mammals

Estimates of silent site divergence (Ks) between *H. sapiens* vs. *M. musculus*, *C. elegans* vs. *C. briggsae*, and *D. melanogaster* vs. *D. pseudoobscura* were obtained using the yn00 method in PAML (version 3.13) (Yang 1997; Yang and Nielsen 2000). The mean and median of Ks were calculated for 29 fly, 193 worm, and 153 mammalian coding sequence alignments taken from references (Bergman, Pfeiffer et al. 2002), (Castillo-Davis and Hartl 2002) and (Nekrutenko, Makova et al. 2002), respectively.

Simulating non-coding sequence divergence

Non-coding sequence evolution was simulated using a modified version of the sequence simulation program ROSE (Stoye, Evers et al. 1998). In general, in the absence of large datasets of non-coding sequences from closely related *Drosophila* species, I have taken estimates of non-coding evolution from previous results reported in the literature. Beginning with ancestral sequences, evolution occurred on two descendent branches of equal length under the HKY model of point substitution (Hasegawa, Kishino et al. 1985), with a transition/transversion bias of 2 to reflect the nucleotide and transition biases observed in *Drosophila* non-coding sequences (Moriyama and Hartl 1993; Moriyama and Powell 1996; Bergman and Kreitman 2001). The substitution rate was set to 0.01 such that a branch length unit was on average 0.01 substitutions per site. Total branch lengths

spanned a range of divergence times from 0.25 to 5.0 substitutions per site. Insertion/ deletion evolution was based on the length distribution of polymorphic indels estimated in (Comeron and Kreitman 2000), and occurred at a 10-fold lower rate than point substitution, approximating relative rates estimated in (Petrov, Lozovskaya et al. 1996; Petrov and Hartl 1998).

To model the evolution of constrained blocks in non-coding sequences a modification of the ROSE sequence simulation program was developed to map constraints on ancestral sequences onto derived sequences (available in ROSE version 1.3). Constraints on non-coding sequences were modelled as short blocks of highly conserved sequences typical of *cis*-regulatory sequences, and follow a lognormal distribution with parameters estimated in (Bergman and Kreitman 2001). On average, interspersed blocks of constrained sites accounted for 20% of the sites in ancestral sequences, a conservative estimate of constraint in *Drosophila* non-coding DNA (Bergman and Kreitman 2001). Parameters used in my simulations are summarized in Table 2.1.

Estimation of evolutionary distance for simulated alignments was performed using the F84 model of sequence evolution in the DnaDist program of the PHYLIP package (Felsenstein 1989) with a transition:transversion ratio of 1.0 (note that a transition:transversion ratio of 1.0 in PHYLIP is equivalent to a transition/transversion bias of 2 in ROSE). Summary statistics for the simulations were calculated using the R statistical package (Figure 2.2) (Ihaka 1996).

Tools for aligning non-coding DNA

The alignment tools tested in this study were chosen based on the criteria that they are (1) publicly available, (2) run in batch mode from the command line and are able to produce (3) strictly co-linear, (4) error-free, pairwise genomic alignments of sequences (5) up to 10 Kb in length. Tools like BBA (Zhu, Liu et al. 1998) (5), Bl2seq (Tatusova and Madden 1999) (3), DBA (Jareborg, Birney et al. 1999) (4), MUMmer (Delcher, Phillippy et al. 2002) (3), Owen (Ogurtsov, Roytberg et al. 2002) (2) and SSEARCH (Pearson and Lipman 1988) (3) were not evaluated since they do not satisfy one of these criteria. I now briefly describe the tools that I tested.

Avid (Bray, Dubchak et al. 2003) is a pairwise global alignment tool whose general strategy for aligning two sequences is to anchor and align iteratively. A set of maximal (but not necessarily unique) matches between the sequences is constructed using a suffix tree. Dynamic programming is used to order and orient the longest matches, which are then fixed. For each subsequence remaining between the fixed matches, the process is repeated until every base is aligned. When sequences are short and the matches make up less than half of the total sequence, the program defaults to the Needleman-Wunsch algorithm (Needleman and Wunsch 1970).

The Chaos/Lagan (Brudno, Do et al. 2003) suite of tools consists of a pairwise local alignment tool, Chaos, and a global alignment tool, Lagan. Chaos starts by finding all words between the two sequences of a specified length and a

specified maximum number of mismatches. These words are then chained together if they are close together in both sequences. These maximal chains are then scored and all chains that are above a specified threshold are returned. Lagan starts by running Chaos with conservative parameter settings and then finds the optimal path through the maximal chains using dynamic programming. Lagan then recursively calls Chaos with increasingly more permissive parameters on the regions between each maximal chain in the optimal path. When the recursion has created a dense map of maximal chains that have been ordered with dynamic programming, Lagan runs the Needleman-Wunsch algorithm on the whole length of both sequences but puts close bounds around the maximal chains to provide the final global alignment. Chaos was run on default parameters as well as using parameters suggested by the authors: word length = 7, number of degeneracies = 1, score cut-off = 20 and extension mode on.

BlastZ (Schwartz, Kent et al. 2003) is a pairwise local alignment tool that is based on the gapped BLAST algorithm that has been redesigned for the alignment of long genomic sequences. BlastZ first removes lineage-specific interspersed repeats from each sequence, then searches for short near-perfect matches between the two sequences. Each match is extended first using gap-free dynamic programming and if it scores above a specified threshold it will be extended using dynamic programming with gaps; extended matches that score above a specified threshold are then kept. Part of the unique implementation of BlastZ is that it can be forced to return alignments that are both unique within

each sequence as well as collinear with respect to each other. To satisfy my strict collinear requirement, I ran BlastZ with both of these options. Blastz was also run using the author's suggestion of lowering the score cut-off (k) to 2000 (BlastZ-A).

DiAlign (v. 2.1) (Morgenstern 1999) is a segment-to-segment alignment algorithm. Like the BLAST algorithms, DiAlign looks for short ungapped segments that have a similarity that deviates from what would be expected by random chance, keeping segments with a score above a certain threshold. These high scoring segments are then aligned into a collinear global alignment using a dynamic programming algorithm. DiAlign produces a global alignment but distinguishes high confidence columns of an alignment from low confidence columns. I used DiAlign as both a global (DiAlign-G) and a local (DiAlign-L) alignment tool.

ClustalW (v. 1.8) (Thompson, Higgins et al. 1994) was used on default settings. ClustalW is a progressive multiple alignment tool that reduces to the Needleman-Wunsch algorithm in the pair-wise case with default parameters of a match score of 1.9, mismatch penalty of 0, a gap open penalty of 10 and a gap extension penalty of 0.1.

The second implementation of the Needleman-Wunsch algorithm used in this study is the needle program in the EMBOSS suite of tools (Rice, Longden et al. 2000). needle was used with default parameter settings of a match score of 5, a mismatch penalty of 4, a gap open penalty of 10 and a gap extension penalty of 0.5.

The final tool tested, WABA (Kent and Zahler 2000), is a three-tier alignment algorithm. The first tier partitions the first sequence into overlapping windows of 2 Kb and then defines a synteny map of high scoring 2 Kb windows of the first sequence onto the second sequence. The second tier then carefully aligns syntenic regions using a seven-state, pair Hidden Markov Model that includes separate query and database insertion/deletion states, high and low non-coding conservation states, as well as three coding states (one for each position in a codon). The final tier then attempts to assemble individual alignments together into a more global alignment.

Alignment performance measures

The performance of alignment tools was assessed using six basic measures: overall coverage, overall sensitivity, constraint coverage, constraint sensitivity, constraint specificity and local constraint sensitivity. Overall coverage and overall sensitivity were measured for all four evolutionary regimes (A-D) while the constraint measures were only measured in the two regimes that included constrained blocks (C, D). Alignments produced by each alignment tool were parsed to generate the statistics, which were then used to calculate each performance measure.

Each site in an alignment produced by a tool (a site being a base in one strand of a column of an alignment) can have two simulated alignment states, two constraint states, three tool alignment states, and two conditional tool alignment

states. The two simulated alignment states are "homolog" (h), ungapped sites in the simulated alignments, and "no homolog" (nh), gapped sites in the simulated alignments. Simulations without indel evolution have only homolog sites since there are no gaps in the simulated alignments. The two constraint states are "constrained" (c), sites in constraint blocks, and "unconstrained" (u), sites not in constrained blocks. The three tool alignment states are "aligned" (a), sites aligned in the tool alignment, "gapped" (g), sites gapped in the tool alignment, and "not aligned" (na), sites not included in a local tool alignment. The two conditional tool alignment states are "aligned correctly" (ac), sites aligned to the same site in both the tool and simulated alignments, and "aligned incorrectly" (ai), sites aligned to different sites in the tool and simulated alignments. There are fourteen possible combinations of these states (e.g. homolog constrained aligned correctly, h_c_ac), giving us fourteen statistics to calculate for each estimated alignment. Counts for each statistic were used to calculate the following measures:

Overall coverage is the fraction of ungapped sites in a simulated alignment that are included in a tool alignment. Overall Coverage = (h_c_ac + h_c_ai + h_c_g + h_u_ac + h_u_ai + h_u_g) / (h_c_ac + h_c_ai + h_c_g + h_c_na + h_u_ac + h_u_ai + h_u_g + h_u_na)

Overall sensitivity is the fraction of ungapped sites in a simulated alignment that are aligned to the correct base in a tool alignment. Overall Sensitivity = (h_c_ac + h_u_ac) / (h_c_ac + h_c_ai + h_c_g + h_c_na + h_u_ac + h_u_ai + h_u_g + h_u_na)

Constraint coverage is the fraction of ungapped constrained sites in a simulated alignment that are included in a tool alignment. Constraint Coverage = (h_c_ac + h_c_ai + h_c_g) / (h_c_ac + h_c_ai + h_c_g + h_c_na)

Constraint sensitivity is the fraction of ungapped constrained sites in a simulated alignment that are aligned to the correct base in a tool alignment. Constraint Sensitivity = (h_c_ac) / (h_c_ac + h_c_ai + h_c_g + h_c_na)

Constraint specificity is the fraction of unconstrained sites in a simulated alignment that are gapped or not included in a tool alignment. Constraint Specificity = (h_u_g + h_u_na + nh_u_g + nh_u_na) / (h_u_ac + h_u_ai + h_u_g + h_u_na + nh_u_a + nh_u_g + nh_u_na)

Local constraint sensitivity is the fraction of sites that are both, contained in a tool alignment and are ungapped constrained sites in a simulated alignment, that are aligned to the correct base in the tool alignment. Local Constraint Sensitivity = (h_c_ac) / (h_c_ac + h_c_ai + h_c_g)

For each of these six measures, a mean and standard error of the mean were calculated for up to 1000 replicates (local tools do not always return an alignment and replicates which produced no alignment were not counted toward the mean) using R.

CHAPTER 3

Inaccuracies in multiple alignments and the inferences we can make from them

Abstract

Molecular evolutionary studies of non-coding sequences rely on multiple alignments. Yet how multiple alignment accuracy varies across sequence types, tree topologies, divergences and tools, and further how this variation impacts specific inferences, remains unclear. Here I develop a molecular evolution simulation platform, CisEvolver, with models of background non-coding and transcription factor binding site evolution, and use simulated alignments to systematically examine multiple alignment accuracy and its impact on two key molecular evolutionary inferences: transcription factor binding site conservation and divergence estimation. I find that the accuracy of multiple alignments is determined almost exclusively by the pairwise divergence distance of the two most diverged species and that additional species have a negligible influence on alignment accuracy. Conserved transcription factor binding sites align better than surrounding non-coding DNA yet are often found to be misaligned at relatively short divergence distances, such that studies of binding site gain and loss could easily be confounded by alignment error. Divergence estimates from multiple alignments tend to be overestimated at short divergence distances but reach a tool

specific divergence at which they cease to increase, leading to underestimation at long divergences. My most striking finding was that overall alignment accuracy, binding site alignment accuracy and divergence estimation accuracy vary greatly across branches in a tree and are most accurate for terminal branches connecting sister taxa and least accurate for internal branches connecting sub-alignments. My results suggest that variation in alignment accuracy can lead to errors in molecular evolutionary inferences that could be construed as biological variation. These findings have implications for which species to choose for analyses, what kind of errors would be expected for a given set of species and how multiple alignment tools and phylogenetic inference methods might be improved to minimize or control for alignment errors.

Background

Annotation of *cis*-regulatory sequences, non-coding RNAs and other functional non-coding sequences is a major challenge in molecular genetics today. Whole genome sequences of closely related species, such as those now available in mammals, flies, worms, yeast and bacteria, provide an opportunity for evolutionary analyses to greatly aid in this effort, but also present new challenges for sequence analysis (Stone, Cooper et al. 2005).

The first step in studying the evolution of non-coding sequences is alignment. New tools have been developed for fast and accurate alignment of long stretches of genomic sequence (reviewed in (Miller 2001; Miller, Makova et al.

2004; Batzoglou 2005)) and benchmarking studies have begun to address the accuracy of these pairwise (Pollard, Bergman et al. 2004; Rosenberg 2005) and multiple (Blanchette, Kent et al. 2004; Rosenberg 2005) alignment tools under various evolutionary scenarios. Knowing the nucleotide-level accuracy of alignment tools greatly informs decisions about which tools to use and which species to compare, but the impact of alignment error on evolutionary studies of non-coding sequences is only just beginning to be explored (Rosenberg 2005; Rosenberg 2005).

Sophisticated molecular evolution models and tests have been developed over the last few decades to identify various forms of selection and sequence features, yet their application nearly always assumes a perfect alignment (Eddy 2005). It is commonly appreciated that highly diverged species align poorly and therefore are unsuitable for many alignment based evolutionary inferences. Thus cautious researchers tend to study recently diverged species that align trivially, but which have the potential to not be as informative as more diverged species. Ideally one would use the set of species that maximize information for an acceptable amount of error in an estimate.

Because of the inferential nature of evolutionary studies, no experiment in extant taxa could generate information about the true orthology of sequences, so simulations offer a tractable alternative. Molecular evolution simulations have been used to assess evolutionary analysis methods, including divergence estimation (Zharkikh 1994; Kishino, Thorne et al. 2001) and phylogeny

reconstruction methods (Felsenstein 1988; Lin and Nei 1991; Hillis, Huelsenbeck et al. 1994; Tateno, Takezaki et al. 1994), as well as protein (McClure, Vasi et al. 1994; Thompson, Plewniak et al. 1999) and non-coding alignment accuracy (Blanchette, Kent et al. 2004; Keightley and Johnson 2004; Pollard, Bergman et al. 2004; Rosenberg 2005; Rosenberg 2005; Huang, Umbach et al. 2006).

Here I present the results from a simulation-based study assessing the accuracy of multiple alignments and the effect of alignment accuracy on two fundamental evolutionary inferences: transcription factor binding site conservation and divergence distance estimation.

The most frequent non-coding targets of comparative analyses are *cis*-regulatory DNAs that contain functional binding sites for transcription factors and thereby control gene expression (Davidson 2001). Although transcription-factor binding sites are generally more conserved than surrounding sequences (Wasserman, Palumbo et al. 2000; McCue, Thompson et al. 2002; Emberly, Rajewsky et al. 2003; Johnson, Bergman et al. 2003; Wang and Stormo 2003; Berman, Pfeiffer et al. 2004; Grad, Roth et al. 2004; Moses, Chiang et al. 2004; Sinha, Schroeder et al. 2004; Bejerano, Siepel et al. 2005; Doniger, Huh et al. 2005; Gertz, Riles et al. 2005; Johnson, Zhou et al. 2005; Wang and Stormo 2005), they have also been observed to be gained and lost through evolution (Ludwig, Bergman et al. 2000; Dermitzakis and Clark 2002; Ludwig 2002; Costas, Casares et al. 2003; Dermitzakis, Bergman et al. 2003; Costas, Pereira et al. 2004; MacArthur and Brookfield 2004; Sinha and Siggia 2005). Precise

measurements of binding site conservation, therefore, are essential for studying their evolutionary dynamics as well as identifying regulatory regions.

Divergence estimates inform nearly all evolutionary analyses. Accurate measurements of non-coding divergences are used for many purposes including differentiating functional from non-functional sequences based on constraint (Hardison 2000; Chiaromonte, Weber et al. 2003; Cooper, Brudno et al. 2003; Elnitski, Hardison et al. 2003; Keightley and Gaffney 2003; Halligan, Eyre-Walker et al. 2004; Kolbe, Taylor et al. 2004; Keightley, Kryukov et al. 2005; King, Taylor et al. 2005), showing lineage specific rate changes (Sarich and Wilson 1973; Wagner, Fried et al. 2004) and as a baseline for comparing other kinds of rates, like binding site gain and loss (Costas, Casares et al. 2003).

Below I first examine multiple alignment accuracy across tools, sequence types, trees and divergences. I show that multiple alignment accuracy is primarily determined by the pairwise divergence of the two most diverged species. I next look at alignment accuracy of transcription factor binding sites. I show that although they align better than their surrounding non-coding DNA, they are misaligned at a high enough frequency such that precise studies of gain and loss events could easily be confounded by alignment errors. Finally I look at the impact multiple alignment accuracy has on divergence distance estimation. I show that divergences tend to be overestimated at short distances and cease to increase at a tool specific maximum divergence, corresponding to the point at which alignment accuracy reaches its minimum. I also show that overall alignment

accuracy, binding site alignment accuracy and divergence estimation accuracy vary across branches in a tree such that terminal branches are aligned better than internal branches. Implications for method development and evolutionary analysis are discussed.

Results

CisEvolver

For the purposes of this study I developed a molecular evolution simulator, CisEvolver, that incorporates several known characteristics of non-coding sequences. CisEvolver takes an ancestral DNA sequence and evolves it along a mutation guide tree, producing sequences for which I know the true alignment. The utility of such a simulation is that the sequences can be re-aligned using standard alignment tools and the accuracy of the tool alignment as well as the accuracy of any inference from the tool alignment can be measured by comparison with the true alignment. In cases where the error in an inference is due to both alignment error and error in the inference method itself, the contribution of alignment error to the total inference error can be directly measured by comparison of inference from the tool alignment and inference from the true alignment.

I implemented CisEvolver with two types of sequences, background genomic sequence and transcription factor binding sites. Background genomic sequences are evolved according to the Hasegawa Kashina Yano 1985 (HKY85)

substitution model (Hasegawa, Kishino et al. 1985), a Poisson insertion/deletion (indel) event model and an empirical indel length frequency distribution (Comeron and Kreitman 2000). Transcription factor binding sites are evolved according to the Halpern Bruno 1998 (HB98) model of position specific substitution rates (Halpern and Bruno 1998; Moses, Chiang et al. 2003), which requires the less degenerate positions in a transcription factor binding site to evolve more slowly and more specifically according to a position specific weight matrix (Schneider, Stormo et al. 1986) (see Methods for more details).

Simulations & Alignments

Using CisEvolver I simulated a large set of alignments on which downstream analyses were performed. Sequences were simulated over a range of total divergence distances on two, three and four species trees with fixed topologies and fixed branch length proportions as depicted in figure 3.1. The relative branch lengths in these three topologies were chosen for direct comparisons of branches within the tree, as discussed below (see *Alignment Accuracy*). Two basic classes of sequences were simulated representing either 10kb background genomic sequences or variable length enhancer sequences. Background genomic sequences were simulated with uniform substitution and indel rates. Enhancer sequences were evolved from 36 experimentally characterized regulatory regions from *Drosophila melanogaster* (Berman, Pfeiffer et al. 2004; Schroeder, Pearce et al. 2004) containing the binding sites for eight

transcription factors with known binding specificity: Bicoid, Caudal, Giant, Hunchback, Knirps, Kruppel, Tailless and Torso-Response Element (Papatsenko, Makeev et al. 2002; Schroeder, Pearce et al. 2004; Bergman, Carlson et al. 2005). Binding sites within the enhancers were evolved using CisEvolver's binding site evolution model with no gain or loss events and surrounding sequences were evolved as genomic background with substitutions and indels (see Methods for more details). One hundred replicates and 25 replicates for each divergence and tree topology were generated for background genomic sequences and each of the 36 enhancers respectively.

All alignments were performed using default parameter settings for Clustalw (Thompson, Higgins et al. 1994), Mavid (Bray and Pachter 2004), Mlagan (Brudno, Do et al. 2003) and Blastz/Tba (Schwartz, Zhang et al. 2000; Schwartz, Kent et al. 2003; Blanchette, Kent et al. 2004) (see Methods for details). These tools were chosen based on their usage, availability, speed and ability to produce collinear multiple alignments of large genomic regions and were meant to be representative of algorithms and parameter settings. I note that Blastz/Tba is a local alignment tool and therefore, unlike the global alignment tools, does not always return an alignment. Finally, although I present the relative performance of these specific tools, my focus in this study is on the relationship of their accuracy with evolutionary scenarios and the inferences that can be made from their alignments.

Alignment Accuracy

Using simulated true alignments and tool alignments I characterized the variation in alignment accuracy across alignment tools, divergences and trees. Alignment accuracy was defined as the fraction of ungapped columns in a true alignment that were aligned identically in a tool alignment (see Methods & "sensitivity" in Chapter 2). I examined many aspects of pairwise and multiple alignment accuracy and my major observations were:

i. Alignment accuracy varies across tools and divergences (figure 3.2A).

ii. The presence of transcription factor binding sites leads to higher alignment accuracy (figure 3.2B).

iii. More species results in better accuracy when comparing trees of equal total divergence but different numbers of leaves (figure 3.2C).

iv. The improvement of adding a fourth species is less than that of adding a third when comparing trees of equal total divergence but different numbers of leaves (figure 3.2C).

v. Adding in-group species or out-group species to a pair of species has an insignificant effect on the alignment accuracy of the pair (figures 3.2D, 3.2E & 3.2F).

In addition to these investigations into alignment accuracy across all species in alignments, I also examined the alignment accuracy for subsets of species within multiple alignments, attempting to relate the accuracy to the tree topology. I measured what I call leaf-to-leaf accuracy, node-to-leaf accuracy and node-to-

node accuracy (see Methods). Leaf-to-leaf accuracy refers to the accuracy of the alignment of sister taxa (i.e. seq3 to seq4 in the four species alignments in figure 3.1), conditioned on the columns being ungapped across all the sequences. Node-to-leaf accuracy refers to the accuracy of the three species alignments, conditioned on the columns containing correct alignments of seq1 to seq2. Node-to-leaf accuracy therefore only depends on the alignment accuracy of node1 to seq3. Similarly, node-to-node accuracy refers to the accuracy of the four species alignments, conditioned on the columns containing correct alignments of seq1 to seq2 and seq3 to seq4. Node-to-node accuracy therefore only depends on the alignment accuracy of node1 to node2. Using these measures I also found that:

vi. Leaf-to-leaf alignments are more accurate than node-to-leaf alignments, which are more accurate than node-to-node alignments, with the exception of highly diverged enhancers (figures 3.2E & 3.2F).

Observations i and ii were consistent with my expectations. Although all four tools in this study use some form of the Needleman-Wunsch algorithm, they each utilize unique algorithmic features and scoring schemes, leading to variation in their alignments and therefore alignment accuracy under different evolutionary conditions (figure 3.2A). Both, the decrease in alignment accuracy with greater divergence distance (figure 3.2A) as well as the increase in alignment accuracy with the addition of transcription factor binding sites (figure 3.2B), are the expected outcome of higher similarity and fewer indels leading to higher alignment accuracy (as I have previously reported for pairwise alignments

(Pollard, Bergman et al. 2004)).

Our results on the relationship of alignment accuracy to the number of species aligned (observations iii, iv and v) are consistent with the hypothesis that the pairwise distance between the two most diverged species in a tree effectively determines alignment accuracy. Across tools and divergences, adding ingroup or outgroup species to a pair of species of fixed divergence had an insignificant effect on alignment accuracy (t-test, $p>0.05$) (figure 3.2D and leaf-to-leaf accuracy in 3.2E & 3.2F). Brudno et al found Mlagan alignments of human and fugu exons were improved by 3% with the addition of mouse as an in-group (Brudno, Do et al. 2003), which is consistent with the trend I observed with Mlagan alignments improving with in-group addition, but this trend was not found to be highly significant at any divergence. Observations iii and iv, that dividing a fixed total divergence up with more species improves accuracy incrementally (figure 3.2C), may appear to be in conflict with this hypothesis but are in fact consistent. The increase in alignment accuracy with additional species dividing up a fixed total divergence is due to a decrease in the pairwise divergence between the two most diverged species, not the actual addition of species (figures 3.2D, 3.2E & 3.2F). Thus the span of the two most diverged species, not the number of species in the alignment, appears to be the primary determinant of alignment accuracy.

Finally, observation vi, that alignment accuracy varies across branches in a tree, is quite unexpected. The progressive alignment steps that these four tools use

appear to be biased toward aligning leaf sequences better than internal nodes, where sub-alignments must be aligned (figure 3.2E). This bias was found to be inconsistent for enhancer sequences, for which alignment accuracy of node-to-node and node-to-leaf branches actually were better than leaf-to-leaf branches at high divergences (figure 3.2F). This variation is surprising given that the accuracy of the alignment of a node to another node or sequence is conditioned on the sequences below that node (in the tree) having been aligned correctly (see Methods). These results suggest that the step of aligning sub-alignments is harder than aligning sequences, consistent with the idea that progressive alignment heuristics often lead to sub-optimal alignments (Keceioglou and Starrett 2004). Variation of alignment accuracy across branches in a tree has profound implications for phylogenetic analysis.

To understand the relationship of the observed variation in alignment accuracy with phylogenetic analyses performed using automated alignments, I explored the following two evolutionary inferences.

Transcription Factor Binding Site Alignment

Using simulated true alignments and tool alignments of enhancers containing conserved transcription factor binding sites I examined the accuracy of binding site alignment and its relationship with overall alignment accuracy. I used two definitions of binding site alignment. Aligned sites were classified as either perfectly aligned, meaning every base in the binding site was aligned correctly

across all species, or overlapping, meaning the binding sites across the species overlapped at at least one position (similar to definitions in (Emberly, Rajewsky et al. 2003)).

I first looked to see if binding site alignment accuracy varies across tools and divergences. Indeed, across tools binding alignment accuracy is a decreasing function of divergence distance. Figure 3.3A shows the fraction of sites overlapping in four species enhancer alignments.

I next compared my two binding site alignment scores. I was somewhat surprised to see how different the two scores are, based on the intuition that conserved binding sites should make for good anchors and large indels in flanking sequences therefore ought to be the cause of most alignment errors. Instead it appears that binding sites are often still overlapping in an alignment even if they are not perfectly aligned. Figure 3.3B shows the difference between my two scores in four species alignments. The large difference between the two scores suggests that evolved binding sites might not be strong anchors and therefore alignment errors in regulatory regions may often be subtle.

I next looked to see how binding site alignment accuracy is related to overall alignment accuracy. Across tools, divergence distances and trees, binding site alignment accuracy is highly correlated with overall alignment accuracy, however, binding site alignment accuracy is consistently higher than overall alignment accuracy. Figure 3.3C shows overlap binding site accuracy as a function of overall alignment accuracy for four species alignments. Similar to

overall alignment accuracy of enhancers (figure 3.2F), binding site alignment accuracy also varies across branches in trees (figure 3.3D).

Lastly, I looked at properties of enhancers and binding sites to see how they are related to binding site alignment accuracy. I expected that enhancers with a greater density of binding sites would align more easily. Indeed, across tools, divergence distances and trees, binding site alignment accuracy is strongly and significantly correlated with the density of binding sites in an enhancer (figure 3.3E, Spearman's *rho*=0.92 $p<10^{-10}$). I also looked at the length and average information content of binding sites to see if longer or more highly specified sites tend to align better. Across tools, divergence distances and trees, binding site alignment accuracy is correlated with binding site length (figure 3.3F, Spearman's *rho*=0.44 p<0.3) and average information content (Spearman's *rho*=0.40 p<0.35) but neither correlation is significant, likely because of the small number of factors used in this study. Thus the greater the density and the longer and more specified the sites in an enhancer, the more likely the sites will be aligned correctly.

Divergence Estimation

Using simulated true alignments and tool alignments of 10kb background non-coding sequences I investigated the effects of alignment errors on divergence estimation. Divergence distances were estimated from alignments using the Baseml program from the PAML package (Yang 1997) (see Methods for run parameters). I used divergence estimation error, instead of accuracy, so as to

capture the directionality of errors (overestimated or underestimated). I defined it as the fractional difference between the Baseml estimate and the true divergence used in the simulation: (Estimate – True) / True.

I first checked to see if divergence estimates from the simulated alignments are accurate. Indeed out to high divergence distances, Baseml estimates are very close to input divergences (figure 3.4).

I next looked to see if and how divergence estimation accuracy varies across tools and divergences. My expectation was that divergence estimation accuracy would steadily decrease with divergence distance at a tool specific rate, as alignment accuracy does. Instead I found estimated divergences tend to be mostly accurate (or somewhat overestimated) at short divergence distances but are always underestimated at long divergence distances. Figure 3.4A shows divergence estimates from four species alignments across tools and divergences. Figure 3.4B shows the same data presented as divergence estimation error, as a function of true divergence distance. Perhaps most striking is the asymptotic approach of estimates to tool specific maxima. This result is consistent with Shabalina and Kondrashov's findings that the alignment of random sequences results in a percent identity much greater than the random expectation of the sum of the squared base frequencies (Shabalina and Kondrashov 1999). If diverging sequences evolve to a lower identity than that of random sequences then alignment tools treat them like they are random and produce an alignment that has a fixed divergence. Indeed aligned random sequences produce similar divergences

as the observed maximum divergences from my simulations (data not shown). Interestingly, the two tools that have the highest maximum divergence (Clustalw and Mlagan) both overestimate divergences at short divergence distances while the two other tools do not. Finally, Tba, the only local alignment tool in my analysis, stops returning alignments before it reaches its maximum divergence, indicating that the algorithm can avoid aligning random alignments but therefore also cannot return weakly informative alignments at large divergence distances.

Because divergence estimation accuracy appears to vary in different ways than alignment accuracy, I looked directly at their relationship. Figure 3.4C shows four species divergence estimation error as a function of alignment error. With the exception of Tba, which stops returning alignments while alignment error is still small, tools reach the point at which divergence estimates cease to increase close to the point at which alignment accuracy reaches its minimum. The accuracy of divergence estimates from Mavid may be due to the fact that it requires a tree with branch lengths and I provided the true divergences. The accuracy of divergence estimates from the other three tools is remarkable given the poor quality of the alignments at long divergence distances.

I last looked to see if divergence estimation accuracy varies across branches in trees as alignment accuracy does. Across tools, divergence estimates were most accurate for leaf-to-leaf branches, less accurate for node-to-leaf branches and least accurate for node-to-node branches. Figure 3.4D shows the error in divergence estimates from Mlagan alignments of leaf-to-leaf, node-to-leaf

and node-to-node branches in two, three and four species trees. Mlagan's tendency to overestimate divergence distances at short divergence distances and to underestimate divergence distances at long divergence distances is least pronounced in leaf-to-leaf alignments and most pronounced in node-to-node alignments. The point at which divergence distances cease to increase also appears to be at a shorter divergence distance for node-to-node branches than leaf-to-leaf branches, reflecting the lower alignment accuracy of those branches. The variation in divergence estimation accuracy across branches in a tree has significant implications for phylogenetic analysis of DNA sequences.

Discussion

Molecular evolutionary studies of non-coding DNA have either relied on the intuition that closely related species can be aligned well or have ignored alignment error all together (Miller 2001; Miller, Makova et al. 2004; Batzoglou 2005; Eddy 2005; Stone, Cooper et al. 2005). To gain perspective on how alignment might impact evolutionary analysis, I investigated multiple alignment accuracy and its relationship with two fundamental evolutionary inferences: transcription factor binding site conservation and divergence estimation.

Because gold standards for base-level non-coding and regulatory DNA alignment accuracy do not exist, I developed a simulation platform called CisEvolver that can evolve background non-coding DNA, transcription factor binding site DNA or a mixture of the two (enhancers). I implemented CisEvolver

with features of background and regulatory sequence evolution that are well modeled and are present in most comparative systems. Certainly more complicated evolutionary phenomena have been observed, and in cases where modeling is successful, ought be the subject of future studies. For instance, substitution rates have been shown to vary across sequences and have been modeled in various ways, most commonly using a gamma distribution (Yang 1994). In my study I modeled both substitution and indel rate variation using interspersed transcription factor binding sites, but rates may vary for additional reasons other than regulatory constraints, in which case a gamma distribution in my background model may be appropriate. Interestingly, a recent study showed that using a gamma distribution in simulations has no effect on Clustalw alignment accuracy when comparing sequences with the same overall identity (Rosenberg 2005), suggesting that my results are likely robust to rate variation. Compensatory substitutions (like those observed in structural non-coding RNAs) (Durbin 1998; Rivas and Eddy 2001; Pedersen, Bejerano et al. 2006), ancient and lineage specific repetitive sequences (like those common in mammals), inversions and rearrangements (Coghlan, Eichler et al. 2005; Negre, Casillas et al. 2005) could all be incorporated into the CisEvolver platform for alignment analysis. As models of the *cis*-regulatory code (Markstein and Levine 2002) and binding site evolution (Costas, Casares et al. 2003; Moses, Chiang et al. 2003) are developed, they too should be tested for affects on alignment accuracy. Additionally, the trees I chose to study are idealistic, in that they are ultrametric (leaves are equidistant

from parent nodes), and they contain relatively few species compared to many real datasets. Trees with rate changes across many lineages would likely present additional problems that should be examined in future studies. A comprehensive analysis of the influence of tree shapes on alignment would be an interesting future direction (see (Rosenberg 2005) for an initial analysis). Despite the absence of these more complicated or unexplored aspects of non-coding evolution in the current study, my results suggest that even under these simple and ideal circumstances numerous issues arise from alignment error that ought to be qualitatively informative for all systems.

Using alignments generated by CisEvolver I tested the accuracy of alignments generated by four commonly used genomic alignment tools. All tools were run using their default parameter values (see Methods). It is quite possible that the accuracy of the alignments generated by some of these tools is highly sensitive to parameter settings and scoring schemes. In this study I focused on consistent behavior across tools and also how variation influenced inferences and was therefore not concerned with the relative performance of each tool. In order for users to optimize the use of current tools and also in order for designers of alignment tools to understand which algorithmic innovations actually improve alignment accuracy (beyond parameter settings), a systematic analysis of parameters is needed. In this study, using just default parameters, I found that the primary determinant of multiple alignment accuracy is the pairwise divergence distance between the two most diverged species in the alignment (figure 3.2D).

Although dividing up a given divergence distance by more species improves accuracy (figure 3.2C), this appears to be simply due to the decrease in pairwise divergence separating the two most diverged species. Although I found that adding additional species (either in-groups or out-groups) to two species of a fixed divergence distance had an insignificant and inconsistent (across tools) impact on alignment accuracy (figure 3.2D), a concurrent study found that Clustalw alignments are most improved when an additional species is added at a distance equal to one third the pairwise distance separating two other species (Rosenberg 2005) (which I note is the topology I used in this study; see figure 3.1). Brudno et al also found that adding mouse to human-fish alignments improved Mlagan alignments by 3% (Brudno, Do et al. 2003). If there is an affect of adding an in-group, my results suggest that it is weak and is not robust to alignment tool choice. Perhaps my most striking finding is that the accuracy of alignments varies across branches in a tree such that they are most accurate for alignments of sister taxa and least accurate between internal nodes that align sub-alignments. As I discuss below, this variation in accuracy causes variation in inferences across the tree, which could easily be construed as lineage specific biological variation. Future development of tools that minimize this distortion in accuracy across branches in a tree will be extremely valuable.

The first evolutionary inference I examined was the measurement of the conservation of transcription factor binding sites in regulatory regions. Studies have used conservation of binding sites as either a means of classifying functional

from spurious predictions (Wasserman, Palumbo et al. 2000; McCue, Thompson et al. 2002; Johnson, Bergman et al. 2003; Wang and Stormo 2003; Berman, Pfeiffer et al. 2004; Grad, Roth et al. 2004; Moses, Chiang et al. 2004; Sinha, Schroeder et al. 2004; Bejerano, Siepel et al. 2005; Doniger, Huh et al. 2005; Gertz, Riles et al. 2005; Johnson, Zhou et al. 2005; Wang and Stormo 2005) or for the purposes of understanding their rates of change, or turnover (Ludwig, Bergman et al. 2000; Dermitzakis and Clark 2002; Ludwig 2002; Costas, Casares et al. 2003; Dermitzakis, Bergman et al. 2003; Costas, Pereira et al. 2004; MacArthur and Brookfield 2004; Sinha and Siggia 2005). Here I wanted to understand how far out such estimates might be accurate, what approaches might be taken to improve such estimates and also which features of regulatory regions might affect such estimates. I found that binding sites are usually aligned better than their surrounding sequences (figures 3.2B & 3.3C) but are still misaligned starting at very short divergence distances (figure 3.3A). For instance, given the approximate divergence of *Drosophila pseudoobscura* from *Drosophila melanogaster*, at 1.79 substitutions per site (Richards, Liu et al. 2005), according to my results, only about 40% of truly conserved binding sites should even be overlapping in alignments. Unless the rate of binding site turnover is high enough such that the number of sites that have turned over is much larger than the 60% of truly conserved sites that have been misaligned, its unlikely that such a comparison would be useful for studying binding site evolution. If 40% binding site conservation, however, is higher than what might be expected in non-

functional regions, then comparing these species might still be useful for predicting functional regulatory regions. My finding that binding sites are often still overlapping in alignments even if they are not perfectly aligned (figure 3B) suggests that binding sites are not always strong alignment anchors, that small indels could lead to small alignment errors and that methods for identifying conserved binding sites that do not rely on perfect alignments would have greater sensitivity (Wasserman, Palumbo et al. 2000; Loots, Ovcharenko et al. 2002; Moses, Chiang et al. 2004) (the specificity of such methods, however, would need to be explored to understand their predictive power). Finally I found that the higher the density of sites in an enhancer, the higher the alignment accuracy of the binding sites within, presumably due to the overall higher constraint and suppression of indels. Bacterial and yeast regulatory regions, for instance, are not understood to contain such high-density arrays of binding sites as metazoans (Harbison, Gordon et al. 2004; Hershberg, Yeger-Lotem et al. 2005) and would therefore be expected to align more poorly, all else being equal. Although I also found that longer and more highly specified binding sites are easier to align, this requires further investigation with a larger panel of transcription factors. The variance in alignment accuracy introduced by such regulatory sequence properties is important to consider before determining the expected error from simulations or before interpreting an evolutionary comparison across regulatory regions.

The second inference I considered was divergence distance estimation. I was impressed that my estimates using PAML's Baseml program on the true

alignments generated in my simulations were highly accurate out to rather high divergences, suggesting that saturation does not lead to inaccuracies at short divergence distances, at least when the right model is used (figure 3.4A & 3.4B). Because of the accuracy of the divergence inference step, I was able to look directly at the contribution of alignment error to divergence estimation. Although the tendency of two of the tools to overestimate divergences at short divergence distances is noteworthy (as was observed for Clustalw in (Rosenberg 2005)), most striking is the behavior that all tools reach a unique divergence distance at which divergence estimates cease to increase (figures 3.4A & 3.4B) (this underestimate was also observed for Clustalw in (Rosenberg 2005)). This point of maximum divergence corresponded with the point at which tools reached their minimum alignment accuracy or where they were essentially randomly aligned (figure 3.4C). Shabalina and Kondrashov previously reported that unrelated sequences produce alignments that have a greater percent identity than would be theoretically predicted from base composition, suggesting that alignment tools add gaps to create more matches and fewer mismatches in order to maximize their scores (Shabalina and Kondrashov 1999). The “twilight zone” (the point where alignments become random) (Rost 1999) is therefore not 25% identity but instead is a much shorter divergence (or higher identity), which varies across alignment tools. For instance, pairwise alignments generated by Mavid reach the point where divergence estimates cease to increase at about 0.5 substitutions per site, which is approximately the divergence estimated for human and mouse,

suggesting that fast evolving human or mouse sequences would on average not be detected as such from Mavid alignments. It is worth noting that Tba, stops returning alignments before it reaches the point where divergence estimates cease to increase, suggesting that the scoring scheme Tba uses to filter its alignments can avoid producing random alignments but also that it might fail to return an alignment with some remaining phylogenetic signal.

Our findings that overall alignment accuracy, binding site alignment accuracy and divergence estimation accuracy vary greatly across branches in a tree have profound implications for phylogenetic research of non-coding DNA. All four of the tools I examined exhibit systematic biases toward higher accuracy along branches connecting sister taxa relative to branches connecting internal nodes (figures 3.2E, 3.2F, 3.3D & 3.4D). Consider the example of studying binding site turnover rates relative to substitution rates in human, mouse and rat alignments. Even if these rates were constant across the tree, binding site turnover might be detected as higher along the human branch because of increased alignment error along the longer node-to-leaf branch and substitution rates might be underestimated along the human branch because it is longer than an alignment tool's maximum divergence. Theses two biases combined would then cause turnover events per substitution to be even more distorted along the human branch. These results strongly suggest that either new alignment tools that minimize this bias or new phylogenetic methods that control for this bias need to be developed.

Conclusions

Errors in the alignment of non-coding DNA are systematic phenomena that affect evolutionary inferences, decreasing accuracy and biasing results. In order to use the rich diversity of variation in more diverged sequences, new alignment and phylogenetic methods need to be developed to reduce and control for errors in automated alignment.

Methods

CisEvolver

CisEvolver was written in Perl. It is available for download (http://rana.lbl.gov/CisEvolver/).

Trees

For both the divergence estimation and binding site conservation estimation simulations, each divergence distance tested was transformed into a Newick formatted tree. Figure 3.1 shows how divergences were distributed across trees.

Divergence Simulations

For the divergence estimation simulations, 100 simulations were run for each divergence distance. For each simulation, a 10kb ancestral sequence was

randomly generated from the *D. melangaster* mono-nulceotide base frequencies (60/40 AT/CG). The 10kb sequences were evolved from the root node of the tree down the branches to leaves using a substitution and indel model. Substitutions occurred according to the HKY85 substitution model (Hasegawa, Kishino et al. 1985), using the *D. melanogaster* mono-nucleotide base frequencies and kappa set to 2.0 as has been observed in *Drosophila* (Bergman and Kreitman 2001). Indel events occurred according to a Poisson indel event model:

$$p_{indel} = 1 - e^{-Rk}$$

where R is the relative rate of indels to substitutions and k is the length of the branch. In *Drosophila,* indels have been found to occur approximately 10% the rate of substitutions so I used R=0.1 (Petrov, Lozovskaya et al. 1996; Petrov and Hartl 1998). Indel lengths were determined by a frequency distribution derived from *D. melanogaster* indel polymorphisms with a maximum of 58bp (Comeron and Kreitman 2000). Insertions and deletions were treated identically.

Cis-Regulatory Sequences

Thirty-six experimentally characterized *cis*-regulatory regions that have been found to drive expression patterns in reporter assays recapitulating some or all of the expression pattern of an adjacent gene were collected from two recent papers on anterior/posterior patterning in *D. melanogaster* (Berman, Pfeiffer et al. 2004; Schroeder, Pearce et al. 2004). The sequences were mapped to release 4.0 of *D. melanogaster* using BLAT (Kent 2002).

Transcription Factor Binding Sites

The 36 *cis*-regulatory regions used in the study have been reported to be bound or predicted to be bound by some combination of the following factors: Bicoid (Papatsenko, Makeev et al. 2002), Caudal (Papatsenko, Makeev et al. 2002), Giant (Bergman, Carlson et al. 2005), Hunchback (Bergman, Carlson et al. 2005), Knirps (Bergman, Carlson et al. 2005), Krüppel (Bergman, Carlson et al. 2005), Tailless (Bergman, Carlson et al. 2005) and Torso-response element (Schroeder, Pearce et al. 2004). Position weight matrices (PWMs) were either taken from published resources (Papatsenko, Makeev et al. 2002; Schroeder, Pearce et al. 2004) or were generated from published footprints (Bergman, Carlson et al. 2005) using MEME (Bailey and Elkan 1995).

For each of the 36 *cis*-regulatory regions, PASTER (Hertz, Hartzell et al. 1990) was used to find sites with a p-value less than 10^{-3} for each of the eight PWMs. If sites were overlapping one was randomly chosen and the others were thrown out.

Transcription Factor Binding Site Conservation Simulations

For the binding site conservation simulations, 25 replicates for each of the 36 *cis*-regulatory regions were evolved to each of the divergence distances. Sequences were evolved from the root down the branches of each tree using either a background or binding site mutation model. Non-binding site sequences in the

enhancers were evolved according the HKY85 and indel models described above. Binding sites were evolved according to the HB98 substitution model (Halpern and Bruno 1998). I have previously shown that there is position-specific variation in substitution rates across functional binding sites and that the HB98 substitution model accurately predicts the relationship between the degeneracy of positions in a PWM and the position specific substitution rate across binding sites (Moses, Chiang et al. 2003; Moses, Chiang et al. 2004). The rate of change from residue a to b at position *i* in the binding site is given by:

$$R(i)_{ab} = Q_{ab} \frac{\log\left(\frac{f_{ib}Q_{ba}}{f_{ia}Q_{ab}}\right)}{1 - \frac{f_{ia}Q_{ab}}{f_{ib}Q_{ba}}},$$

where Q is the background substitution model (HKY85) and f is the PWM for the factor. Indel events were not permitted in binding sites and deletions from background sequences were not allowed to extend into binding sites.

Alignments

Alignments were performed using default parameter values for each of the following tools: Clustalw (Thompson, Higgins et al. 1994), Mavid v0.9 (Bray and Pachter 2004), Mlagan v1.2 (Brudno, Do et al. 2003) and Blastz/Tba (Schwartz, Zhang et al. 2000; Schwartz, Kent et al. 2003; Blanchette, Kent et al. 2004).

Alignment Accuracy

Alignment accuracy was defined as

$$Acc = \frac{C_{TSU}}{C_{SU}},$$

where C_{SU} is the count of the ungapped columns in the simulated alignment and C_{TSU} is the count of the ungapped columns in the simulated alignment that are aligned identically in the tool alignment. This measure is the same as "sensitivity" defined in (Pollard, Bergman et al. 2004).

Branch specific alignment accuracy was calculated similarly except that C_{SU} was the count of ungapped columns for which the alignment was joining either sequences or correctly aligned sub-alignments and C_{TSU} was the count of such columns in the simulated alignment that were aligned identically in the tool alignment. For instance, in a four species alignment, the node-to-node alignment accuracy was only based on the columns for which Seq1 and Seq2 were aligned correctly to each other and Seq3 and Seq4 were aligned correctly to each other (figure 3.1). Similarly, in a three species alignment, the node-to-leaf alignment accuracy was only based on the columns for which Seq1 and Seq2 were aligned correctly to each other. The motivation for this was to consider only the contribution to alignment accuracy a given branch contributes.

A script written in PERL that can calculate these measures is available for download at http://rana.lbl.gov/CisEvolver/.

Binding Site Alignment Measures

Binding site alignment was evaluated based on two measures. Sites that had the same start and stop position in each sequence in an alignment were considered to be perfectly aligned. Sites that were overlapping by at least one base in each of the sequence in an alignment were considered to be overlapping. The fraction of sites that were perfectly aligned and the fraction of sites overlapping in alignments across all *cis*-regulatory regions and all replicates are reported. The Pearson correlation between the density of binding sites in *cis*-regulatory regions and each measure as well as the correlation between the length of binding sites for each factor and each measure were calculated using the R statistics package (Ihaka 1996).

Divergence Estimation

Divergence estimates were calculated using the Baseml program from the PAML package v3.14 (Yang 1997). Baseml was run with the HKY85 model, estimating kappa with an initial value of 2, fixed alpha at infinity, no clock and estimating the equilibrium base frequencies from the observed averages.

CHAPTER 4

Discordance of gene trees with species tree in the *Drosophila melanogaster* species subgroup: evidence for incomplete lineage sorting

Abstract

The phylogenetic relationship of the now fully sequenced species *Drosophila erecta* and *D. yakuba* with respect to the *D. melanogaster* species complex has been a subject of controversy. All three possible groupings of the species have been reported in the past, though recent multi-gene studies suggest that *D. erecta* and *D. yakuba* are sister species. Using the whole genomes of each of these species as well as the four other fully sequenced species in the subgenus *Sophophora*, I set out to investigate the placement of *D. erecta* and *D. yakuba* in the *D. melanogaster* species group and to understand the cause of the past incongruence. Though I find that the phylogeny grouping *D. erecta* and *D. yakuba* together is the best supported, I also find widespread incongruence in nucleotide and amino acid substitutions, insertions and deletions, and gene trees. The time inferred to span the two key speciation events is short enough that under the coalescent model, the incongruence could be the result of incomplete lineage sorting. Consistent with the lineage sorting hypothesis, substitutions supporting

the same tree were spatially clustered. Support for the different trees was found to be linked to recombination such that adjacent genes support the same tree most often in regions of low recombination and substitutions supporting the same tree are most enriched roughly on the same scale as linkage disequilibrium, also consistent with lineage sorting. The incongruence was found to be statistically significant and robust to model and species choice. No systematic biases were found. I conclude that phylogenetic incongruence in the *D. melanogaster* species complex is the result, at least in part, of incomplete lineage sorting. Incomplete lineage sorting will likely cause phylogenetic incongruence in many comparative genomics data sets. Methods to infer the correct species tree, the history of every base in the genome and comparative methods that control for and/or utilize this information will be valuable advancements for the field of comparative genomics.

Background

With the sequencing of twelve species from the genus *Drosophila*, the field of comparative genomics is now presented with the opportunity and challenge of understanding the function and history of every base in the model organism *Drosophila melanogaster* (*Dmel*). This process will hopefully result in the discovery of new biological phenomena and the development of new methodologies that will eventually help with the task of annotating other clades in the tree of life, particularly the human genome. Because most analyses of multiple

genome sequences involve inferences about evolutionary history, they require an accurate description of the relationship of the species being analyzed.

The species history of the genus *Drosophila* has been the subject of numerous studies and the consensus from the literature suggests that the relationship of the twelve sequenced species is well resolved, with the exception of the species within the *Dmel* species subgroup and perhaps the placement of the Hawaiian, *D. grimshawi*, and the *virilis-repleta* species, *D. virilis* and *D. mojavenis* (Russo, Takezaki et al. 1995; Powell 1997; O'Grady and Kidwell 2002; Remsen and O'Grady 2002; Lewis, Beckenbach et al. 2005). Within the *Dmel* species group, the placement of *D. erecta* (*Dere*) and *D. yakuba* (*Dyak*) relative to the *Dmel* lineage has been the subject of numerous conflicting studies (Lemeunier and Ashburner 1976; Barnes, Webb et al. 1978; Eisses 1979; Solignac, Monnerot et al. 1986; Caccone, Amato et al. 1988; Jeffs, Holmes et al. 1994; Schlotterer, Hauser et al. 1994; Russo, Takezaki et al. 1995; Shibata and Yamazaki 1995; Powell 1997; Ko, David et al. 2003; Parsch 2003; Lewis, Beckenbach et al. 2005). Considering the placement of *Dmel*, *Dere* and *Dyak*, all three of the possible phylogenies (Figure 4.1) have received support. The topology (*Dmel*,(*Dere*,*Dyak*)), which I shall refer to as tree 1, was supported by studies of polytene chromosome banding sequences (Lemeunier and Ashburner 1976), satellite DNA (Barnes, Webb et al. 1978), the *COI* and *COII* mitochondrial genes (Lewis, Beckenbach et al. 2005), mitochondrial DNA (Nigro, Solignac et al. 1991), the *fru* gene (Gailey, Ho et al. 2000), the *Cu/Zn SOD* gene (Arhontaki,

Eliopoulos et al. 2002), the *H3* gene family (Matsuo 2000), a concatenation of mitochondrial and nuclear genes (Kopp and True 2002), a concatenation of the genes *Adh*, *Adhr*, *Gld* and *ry* (Ko, David et al. 2003) and a concatenation of the genes *Adh*, *Amyrel*, *janA*, *janB* and *Sod* (Parsch 2003). The topology ((*Dmel*,*Dere*),*Dyak*), which I shall refer to as tree 2, was supported by studies of an internal transcribed spacer region of ribosomal RNA genes (Schlotterer, Hauser et al. 1994), nucleotide sequences 5' of the *Amy* gene (Shibata and Yamazaki 1995) and the *Adh* gene (Moriyama and Gojobori 1992; Ko, David et al. 2003). The topology ((*Dmel*,*Dyak*),*Dere*), which I shall refer to as tree 3, was supported by studies of protein electrophoresis (Eisses 1979), mitochondrial DNA (Solignac, Monnerot et al. 1986), single copy nuclear and mitochondrial DNA hybridization (Caccone, Amato et al. 1988), the *Adh* gene (Jeffs, Holmes et al. 1994; Russo, Takezaki et al. 1995) and the *Amy* gene (Shibata and Yamazaki 1995). The support that each of these studies provides for the three phylogenies, however, is not uniformly strong. The most recent study by Ko et al. using the concatenation of multiple nuclear genes provides the most compelling evidence, with 100% bootstrap support, for the placement of *Dere* and *Dyak* as sister taxa relative to the *Dmel* lineage. That Ko et al. found such strong support for tree 1, despite using the *Adh* gene, which on its own has been found to support the other two trees, suggests that the past incongruence was likely the result of sampling variance (Gadagkar, Rosenberg et al. 2005; Rokas and Carroll 2005). Incongruence, however, can also be the result of numerous systematic biases

(Arbogast, Edwards et al. 2002; Sanderson and Shafer 2002; Zwickl and Hillis 2002; Felsenstein 2004; Jermiin, Poladian et al. 2005) that are not overcome by increased sampling (Phillips, Delsuc et al. 2004; Brinkmann, van der Giezen et al. 2005; Jeffroy, Brinkmann et al. 2006), as well as phylogenetically meaningful phenomena, such as lateral transfer (Andersson 2005) and incomplete linage sorting (Avise, Shapira et al. 1983; Pamilo and Nei 1988; Takahata 1989; Wu 1991; Hudson 1992; Maddison 1997; Kliman, Andolfatto et al. 2000; Chen and Li 2001; Arbogast, Edwards et al. 2002; Rosenberg 2002; Rosenberg and Nordborg 2002; Rosenberg 2003; Holland, Huber et al. 2004; Degnan and Salter 2005; Mossel and Vigoda 2005; Osada and Wu 2005; Maddison and Knowles 2006).

In this study, I set out to examine the possible causes of incongruence in this phylogeny and to investigate the placement of *Dere* and *Dyak* in the *Dmel* species subgroup, using the newly sequenced genomes in the genus *Drosophila*. Although I found that tree 1, placing *Dere* and *Dyak* as sister species, is the best-supported tree, I found genome-wide incongruence in substitutions, indels and gene trees. I show that the branch separating the split of *Dmel* from the split of *Dere* and *Dyak* is sufficiently short that incomplete lineage sorting is a plausible explanation for the incongruence. I further show that the support for the three possible trees is non-randomly distributed across the genome such that adjacent genes supporting the same tree are more likely in regions of low recombination and substitutions supporting the same tree are most enriched roughly on the same scale as estimates of linkage disequilibrium, consistent with theoretical

predictions under the coalescent (Slatkin and Pollack 2006). I tested for obvious systematic biases and found that no factor I examined could account for the incongruence. I conclude by suggesting that incongruence due to incomplete lineage sorting has important implications for comparative genomics research.

Results

Comparative Annotation of Drosophila Species

To analyze the phylogenetic history of the gene compliment of each of the seven fully sequenced species in the subgenus *Sophophora*, I mapped *Dmel* gene annotations onto each unannotated genome. 19,186 *Dmel* coding sequences were mapped to potential orthologous regions in each species using TBLASTN, and GeneWise was used to build gene models based on the *Dmel* gene in each region. These GeneWise models were matched back to *Dmel* translations using BLASTP and genes for which clear orthologs could be found were used in downstream analysis (see Methods). Peptide sequences from orthologs were aligned using TCoffee (Notredame, Higgins et al. 2000) and cDNA alignments were mapped onto the peptide alignments.

Species and Trees

Of these seven subgenus *Sophophora* species, I chose to use *Dmel*, *Dere*, *Dyak* and *D. ananassae* (*Dana*) for my initial analysis of the placement of *Dere* and *Dyak* within the *Dmel* species subgroup (I examine the effects of species

choice on my results below). *Dmel* was chosen because the annotations were mapped from *Dmel* and it is the primary model organism of the subgenus. *Dsim* and *Dsec* were excluded from initial analysis because they were assumed to provide mostly redundant information to *Dmel* and they reduced the number of clear orthologs spanning the species by 2544 genes, presumably because of lower sequence coverage and issues regarding the assembly of polymorphic reads in *Dsim*. *Dana* was chosen over *Dpse*, because it is the closest fully sequenced outgroup to the *Dmel* species subgroup. 9405 genes were found to have clear orthologs in all four of the chosen species. Figure 4.1 shows the three possible unrooted trees relating the species.

Genome Wide Incongruence

I began my analysis looking directly at the genome-wide counts of amino acid substitutions, nucleotide substitutions and insertion/deletion (indel) events that were informative with respect to each of the three possible trees (see Methods). For all three characters, tree 1, which groups *Dere* and *Dyak* together, was found to have the most support (Figure 4.2ABC). By a majority-rule consensus, tree 1 would be inferred to be the species tree, consistent with the findings of Ko et al. (Ko, David et al. 2003). The high proportion of substitutions and indels supporting the alternate trees, however, suggests a poorly resolved tree and pervasive incongruence.

What is the cause of this incongruence? The incongruent substitutions could be the product of any of a number of systematic biases but the incongruent indels are unambiguous characters that are more difficult to explain as methodological artifacts (Rokas and Holland 2000; Ogurtsov, Sunyaev et al. 2004). The population genetic theory of the coalescent states that sufficiently close speciation events will lead to incongruence due to incomplete lineage sorting (Figure 4.3) (Maddison 1997). Below I explore the compatibility of my data with the coalescent as well as test for possible systematic biases.

Maximum Likelihood Gene Trees Show Incongruence

I first repeated my analysis using maximum likelihood (ML) methods (Felsenstein 1981; Huelsenbeck 1995) to measure the informative divergence spanning the inferred speciation events and test the robustness of the incongruent substitutions using more complex models of sequence evolution. ML analysis is not currently scalable to entire genomes in a single calculation, so I partitioned the genome into individual genes. If incomplete lineage sorting is the underlying cause of the incongruence, such a partition might also reveal variation in allelic histories that multi-gene concatenations could obscure (Felsenstein 2004; Holland, Jermiin et al. 2005; Mossel and Vigoda 2005). Wanting to capture both the observed nucleotide and amino acid differences across the species (Ren, Tanaka et al. 2005), I used the F3x4 codon-based model from the PAML package (Yang 1997) to compare the likelihood of each tree given each cDNA alignment

(we test other models below). Consistent with the parsimony-based analysis, the majority of genes (57.8%) support tree 1, while a high proportion (42.2%) support the other two trees (Figure 4.2D).

The median synonymous divergence trees for the sets of genes supporting each tree are: (dmel:0.1301,(dere:0.1095,dyak:0.1201):0.0664,dana:1.3246) for tree 1, ((dmel:0.1744,dere:0.1076):0.0498,dyak:0.0757,dana:1.2871) for tree 2 and
((dmel:0.1801,dyak:0.1163):0.0454,dere:0.0719,dana:1.3147) for tree 3 (Figure 4.4). The branches between the speciation events are quite short, with the tree 1 branch being the longest at only 0.066, suggesting that these species split in rapid succession.

Incongruence Is Expected For These Species Under the Coalescent

Is the time spanning these speciation events short enough to expect the observed levels of incongruence? Using the coalescent, the probability of congruence, or monophyly, can be directly calculated for the three taxa case using the equation p(congruence)=1-2/3exp(-t), where t is the time between speciation events in units of generations/2Ne and Ne is the effective population size (Kingman 1982; Tajima 1983; Nei 1986). Figure 4.5 shows this probability graphically as a function of t. In order to go from an estimate of the informative divergence to this probability, the substitutions per site per year, the ancestral generation time and the ancestral population size must be known. Synonymous-

substitutions per site per year have been estimated to be in the range of $1\text{-}2\text{x}10^{-8}$ in *Drosophila* (Caccone, Amato et al. 1988; Sharp and Li 1989; Russo, Takezaki et al. 1995; Li, Satta et al. 1999). Generations per year for the extant taxa in the *Dmel* species subgroup is about 10 and can be used as an estimate for the ancestral generation time (Sawyer and Hartl 1992). The ancestral population size has been estimated in the range of 10^6 to 10^7 but this should be considered a poorly resolved parameter (Singh 1989). Theoretically, the median informative branch length measured above includes both divergence prior to the first speciation event and divergence between the two speciation events. If I take the informative divergence estimated from genes supporting the alternative trees to represent the expected amount of divergence prior to the first speciation event (0.05 and 0.045 for trees 2 and 3 respectively) and subtract their average (0.0475) from the tree 1 total informative divergence (0.066), I can get an estimate of the informative divergence spanning the two speciation events (0.019). This leads to an estimate of $9.5\text{x}10^5$ to $1.9\text{x}10^6$ years or $9.5\text{x}10^6$ to $1.9\text{x}10^7$ generations. The range of values for t becomes 0.48 to 9.5, which produces probabilities for congruence in the range of 0.59 to 0.99995 (Figure 4.5). Although the uncertainty in these parameter estimates does not permit us to say that incongruence would be guaranteed, they do allow us to say that incongruence due to incomplete lineage sorting is expected under plausible assumptions about these species' ancestral population and speciation events.

Spatial Structure of Tree Support

Given that I observed incongruence in individual sites as well as for whole genes, I wanted to better understand the extent to which sites supporting the same tree are spatially correlated, with a particular interest in the compatibility of this structure with the incomplete lineage sorting hypothesis. The above analysis of gene trees suggests that sites can be correlated out to the length of genes. To see if this correlation extends beyond individual genes I looked for blocks of adjacent genes supporting the same gene and tested for unusual block lengths. Using permutations of ML gene tree state to obtain significance, I found gene tree block lengths at expected frequencies with the exception of an excess of long blocks supporting tree 3, in the range of 250kb to 700kb, three of which were highly significant ($p<0.05$).

If the blocks of genes supporting the same tree were the product of incomplete lineage sorting, then regions of low recombination ought to have larger blocks (Wiuf, Zhao et al. 2004). Although the ancestral recombination rates are not known, I looked to see if block lengths are correlated with *Dmel* recombination rates (Hey and Kliman 2002). I found a weak negative correlation for all blocks (Pearson's $R=-0.13, p<0.1$) as well for blocks for each specific tree, with tree 2 blocks showing the strongest correlation (Pearson's $R=-0.30, p<0.05$). These weak correlations suggest a minor role for recombination rates in determining the spatial structure of support for different trees across the genome, however, there are many reasons for why strong correlations would not be

expected, including poorly conserved recombination rates across these species (True, Mercer et al. 1996; Takano-Shimizu 2001; Wang, Thornton et al. 2004) and gene conversion in regions of low recombination (Langley, Lazzaro et al. 2000; Andolfatto and Wall 2003; Braverman, Lazzaro et al. 2005). Nonetheless, these weak correlations establish a connection between recombination and the spatial structure of support that is at least consistent with lineage sorting. I next looked at the spatial correlation of individual sites to understand the spatial correlation at a finer scale.

Using the whole genome frequencies of informative amino acid and nucleotide substitutions supporting each tree, I looked to see if sites supporting the same tree are locally enriched across chromosomes (see Methods for more details). Figure 5.6 shows that informative amino acid and nucleotide substitutions supporting the same tree cluster together on the scale of less than 8kb for trees 1 and 2 and less than 2kb for tree 3. These local deviations in the frequencies of informative substitutions from the expected frequencies are quite highly significant (X^2 test, $p<10^{-10}$).

What forces might have shaped these clusters of informative sites supporting the same tree? Under the coalescent, linked neutrally evolving sites supporting the same tree have been proposed to be correlated at an expected distance equal to linkage disequilibrium (Slatkin and Pollack 2006). Linkage disequilibrium in *Dmel* has been estimated to extend to the length of a few kb (Zapata and Alvarez 1993), suggesting that my results are consistent with

theoretical expectations (Slatkin and Pollack 2006). Recent empirical evidence from *Dmel*, however, implies that neutral sites would not be expected to be in disequilibrium at distances greater than a few hundred base pairs (Ohta and Kimura 1971; Thornton and Andolfatto 2006), suggesting that perhaps selection has acted to increase the scale of these correlations (Wiuf, Zhao et al. 2004). Regardless of the influence of selection, the structure of the support for different trees across the genome is consistent with recombination acting within the context of incomplete lineage sorting.

Additional support for this conclusion comes from the observation that mitochondrial genes exhibit no incongruence (Montooth K. & Rand D., Personal Communication). This is expected, as recombination is not thought to occur in the mitochondrial genome. While mitochondrial evolution differs from nuclear evolution in more ways than just recombination (Bazin, Glemin et al. 2006), the complete lack of incongruence is nevertheless striking.

Thus far I have presented results suggesting that incomplete lineage sorting is a plausible explanation for the observed incongruence. I next sought to rule out alternate explanations.

Statistical Support For Incongruence

Is the incongruence in gene trees unexpected given the strength of support for each inference? To address this question, I used the bootstrap (Felsenstein 1985) value, RELL (Kishino, Thorne et al. 2001), from 10,000 replicates as an

estimate of the expected incongruence due to chance alone. Taylor and Piel have shown for a large set of yeast genes, originally reported by Rokas et al. (Rokas, Williams et al. 2003), that there is no significant difference between nonparametric bootstrap values and accuracy, as measured by congruence (Taylor and Piel 2004). Earlier work suggests that bootstrap values are conservative and likely to underestimate accuracy (Hillis and Bull 1993; Soltis and Soltis 2003). Figure 4.7A shows the proportion of genes supporting each tree in bins of bootstrap value. Unlike the yeast phylogeny, my observed incongruence consistently exceeds that expected by bootstrap values. Thus the incongruence for these four species using the F3x4 codon model appears to be statistically significant.

Incongruence Is Robust To Model Choice

I next tested if the incongruence is robust to model choice. An empirical study of model choice and accuracy by Ren et al. found that codon based models are able to recover both recent and deep divergences well, while nucleotide based models are less efficient at deep divergences and amino acid based models are less efficient at recent divergences (Ren, Tanaka et al. 2005). They also found that while more complex models fit the data better, they are not necessarily more accurate, a conclusion that has been made by other studies (Yang 1997; Sullivan and Swofford 2001). I looked at six models: nucleotide (HKY,HKY+G), codon (F3x4,F3x4+G) and amino acid (WAG+F,WAG+F+G) based models both with

and without a discrete gamma model of variable rates amongst sites (see Methods). Incongruence was found to exceed expected levels from bootstrap values across all models, suggesting that the incongruence is indeed robust to model choice.

Table 4.1. Congruence and fit to data across six models of evolution

Model	Tree 1	Tree 2	Tree 3	Nucleotide AIC	Amino Acid AIC
HKY	3,615 (62.5%)	1,284 (22.2%)	885 (15.3%)	0 (0%)	-
HKY+G	2,882 (49.8%)	1,696 (29.3%)	1,206 (20.9%)	1 (0.04%)	-
F3x4	3,215 (56.1%)	1,383 (24.1%)	1,135 (19.8%)	107 (4.6%)	-
F3x4+G	3,068 (52.5%)	1,455 (25.4%)	1,210 (21.1%)	2,225 (95.4%)	-
WAG+F	2,971 (51.9%)	1,446 (25.2%)	1,312 (22.9%)	-	618 (26.7%)
WAG+F+G	2,917 (50.9%)	1,502 (26.2%)	1,310 (22.9%)	-	1,694 (73.3%)

Incongruence is robust to model choice, and simple models of evolution result in greater congruence while more complex models fit the data better. Across the 5,784 genes for which an ortholog was found in each of the seven Sophophora subgenus species, the counts of genes supporting each of the three trees were measured using the Dmel, Dere, Dyak, and Dana species combination and each of six different models of evolution. Using Akaike's information criterion (AIC) the counts of consistent genes for which each of the six models provided the best fit to the data were measured, with nucleotide and amino acid-based models compared separately. Not all genes had a ML tree for each model of evolution.

Comparing congruence across models, simpler models seem to produce more congruence than more complex models (Table 4.1). For each of the three types of models, addition of a discrete gamma resulted in lower congruence. For the models without discrete gamma, HKY was more congruent than F3x4, which was more congruent than WAG+F, perhaps due to the relatively recent divergences in this phylogeny. Interestingly, the more complex models, F3x4+G for nucleotides and WAG+F+G for amino acids, fit the alignments better for most genes, according to Akaike's information criterion (Table 4.1) (Akaike 1974).

Thus, consistent with the finding of Ren et al. with the yeast dataset (Ren, Tanaka et al. 2005), more complex models fit the data better but produce less congruence.

Table 4.2. Congruence across 21 Species Combinations

Species Combination	Tree 1	Tree 2	Tree 3
Dmel/Dere/Dyak/Dana	3,615 (62.5%)	1,284 (22.2%)	885 (15.3%)
Dsim/Dere/Dyak/Dana	3,468 (61.0%)	1,296 (22.8%)	917 (16.1%)
Dsec/Dere/Dyak/Dana	3,490 (61.0%)	1,359 (23.8%)	869 (15.2%)
Dsim/Dere/Dyak/Dana/Dpse	3,452 (60.8%)	1,321 (23.3%)	905 (15.9%)
Dsec/Dere/Dyak/Dana/Dpse	3,447 (60.3%)	1,383 (24.2%)	884 (15.5%)
Dsim/Dsec/Dere/Dyak/Dana/Dpse	3,419 (60.2%)	1,345 (23.7%)	912 (16.1%)
Dmel/Dere/Dyak/Dana/Dpse	3,477 (60.1%)	1,371 (23.7%)	935 (16.2%)
Dsim/Dsec/Dere/Dyak/Dana	3,390 (59.7%)	1,347 (23.7%)	943 (16.6%)
Dmel/Dere/Dyak/Dpse	3,365 (58.2%)	1,403 (24.3%)	1,015 (17.6%)
Dsec/Dere/Dyak/Dpse	3,324 (58.2%)	1,403 (24.6%)	987 (17.3%)
Dsim/Dere/Dyak/Dpse	3,299 (58.1%)	1,374 (24.2%)	1,004 (17.7%)
Dsim/Dsec/Dere/Dyak/Dpse	3,249 (57.2%)	1,418 (25.0%)	1,009 (17.8%)
Dmel/Dsec/Dere/Dyak/Dana/Dpse	3,302 (57.1%)	1,504 (26.0%)	977 (16.9%)
Dmel/Dsim/Dsec/Dere/Dyak/Dana/Dpse	3,276 (56.7%)	1,506 (26.1%)	996 (17.2%)
Dmel/Dsim/Dere/Dyak/Dana/Dpse	3,277 (56.7%)	1,502 (26.0%)	1,001 (17.3%)
Dmel/Dsim/Dsec/Dere/Dyak/Dana	3,240 (56.0%)	1,519 (26.3%)	1,022 (17.7%)
Dmel/Dsim/Dere/Dyak/Dana	3,229 (55.8%)	1,521 (26.3%)	1,033 (17.9%)
Dmel/Dsec/Dere/Dyak/Dana	3,215 (55.6%)	1,531 (26.5%)	1,038 (17.9%)
Dmel/Dsim/Dsec/Dere/Dyak/Dpse	3,085 (53.4%)	1,578 (27.3%)	1,114 (19.3%)
Dmel/Dsim/Dere/Dyak/Dpse	3,084 (53.4%)	1,576 (27.3%)	1,120 (19.4%)
Dmel/Dsec/Dere/Dyak/Dpse	3,067 (53.0%)	1,591 (27.5%)	1,125 (19.5%)

Incongruence is robust to species choice. Across the 5,784 genes for which an ortholog was found in each of the seven Sophophora subgenus species, the counts of genes supporting each of the three trees were measured using the HKY model and each of the 21 informative species combinations. Not all genes had a ML tree for each species combination.

Species Choice Does Not Explain The Observed Incongruence

To evaluate the robustness of the incongruence to species choice I examined the set of 5778 genes for which a clear ortholog could be found in all seven fully sequenced species in the subgenous *Sophophora*: *Dmel*, *Dsim*, *Dsec*, *Dere*, *Dyak*, *Dana* and *Dpse*. All 21 possible species combinations that include *Dere* and *Dyak* and at least one of *Dmel*, *Dsim* and *Dsec* as well as at least one of *Dana* and *Dpse* were considered. The HKY model was used both because it was found to produce the most congruence in the original four species as well as

because it is considerably more computationally efficient than the codon models. Across all species combinations, incongruence is consistently greater than expected from bootstrap values, suggesting that incongruence is not species choice dependent.

Ranking species combinations by levels of congruence reveals that my original species choice produces the most congruence (Table 4.2), suggesting that my estimates are conservative. The relative congruence of the species combinations appears non-random, with respect to presence or absence of individual species, so I calculated the average congruence for each species across the combinations containing that species. Although the average congruence is very similar for each species, I found that *Dana* (82.4%) contributes most to congruence, while *Dsim* (80.8%), *Dsec* (80.4%) and *Dpse* (79.7%) contribute roughly equally and *Dmel* (78.9%) actually contributes least to congruence. I note that the presence of *Dmel* in the most congruent species combination goes against this general trend, perhaps reflecting further complexities in the impact of species choice on congruence.

Consistency

Although the incongruence appears to be robust to model and species choice, a much more stringent test is to look at incongruence in the partition of genes that consistently support the same tree across all models and across all species combinations (Gatesy, Milinkovitch et al. 1999). Of the 5778 genes

analyzed, 2347 are consistent across all models and of those, 1600 (68.2%) are congruent while 443 (18.9%) support tree 2 and 304 (12.9%) support tree 3. Similarly, 1918 genes are consistent across species combinations and of those, 1474 (76.8%) are congruent while 291 (15.2%) support tree 2 and 153 (8%) support tree 3. Finally, 970 genes are consistent across all models and all species combinations and of those, 804 (82.9%) are congruent while 101 (10.4%) support tree 2 and 61 (6.3%) support tree 3. This conservative partitioning reduces the amount of incongruence but does not eliminate it. I note that under the incomplete lineage sorting hypothesis, incongruent genes are expected to have accumulated fewer informative substitutions (Figure 4.4) and therefore might be expected to be less robust to such a consistency test.

To assess the statistical significance of the incongruence in the partition of genes consistent across all models and species combinations (Jeffroy, Brinkmann et al. 2006), I used the HKY model bootstrap values from the *Dmel*, *Dere*, *Dyak* and *Dana* species combination to look at congruence as a function of bootstrap value. As shown in figure 4.7B, the congruence is less than expected for the highest bootstrap values. For the 521 genes with bootstrap values between 0.9 and 1.0, which is more than half of consistent genes, the incongruence was highly significant (X^2 test, $p<10^{-3}$).

To further test whether the statistical support from the incongruent genes is the result of consistent signal, as opposed to having hidden support (Gatesy and Baker 2005) for tree 1, I concatenated the 804 consistent tree 1 genes, 101

consistent tree 2 genes and 61 tree 3 genes into three large alignments and repeated the ML analysis for the *Dmel*, *Dere*, *Dyak* and *Dana* species combination and the HKY model. Interestingly, each tree-specific concatenation supported its tree with 100% bootstrap support (Cunningham 1997). Thus the signal for incongruence appears to be consistent, highly significant and robust to model and species choice consistency partitioning.

Sequence & Evolutionary Properties

I next looked at sequence and evolutionary properties of the genes supporting each tree to see if any clear biases could explain the incongruence. The properties I examined are sequence quality, gene length (measured in ungapped codons in the alignment), base composition (GC content) across the species at each position in the codon, transition:transversion ratio (kappa), dN/dS, informative synonymous divergence (ISD), ratio of informative synonymous divergence to non-informative synonymous divergence (RINSD), and total synonymous divergence (TSD). Table 4.3 shows the correlation of bootstrap values to each of these properties for the whole set of genes and for genes supporting each tree. Distributions for properties are shown in Figure 4.8.

The strongest and most consistent correlations with bootstrap value are for ISD and RINSD (Table 4.3), which are in essence the signal and signal to noise. We've already shown that the median informative divergence in the genes supporting tree 1 is greater than that for the genes supporting trees 2 and 3 (Figure

4.4). Reflecting this, the distributions of ISD and RINSD for genes supporting trees 2 and 3 are shifted toward lower values compared to genes supporting tree 1 (figure 4.8A). Comparing consistent genes and inconsistent genes reveals that nearly all genes with ISD values close to zero are classified as inconsistent (figure 4.8B). Amongst consistent genes, those supporting trees 2 and 3 still have distributions of ISD and RINSD shifted slightly toward lower values compared to those supporting tree 1 (figures 4.8C). The fact that incongruent genes are expected to have lower ISD than congruent genes under the incomplete lineage sorting model (see above) and the fact the ISD and RINSD distributions are highly overlapping for each of the three trees suggests that a simple lack signal or low signal to noise cannot explain the observed incongruence.

The long branch out to *Dana* (figure 4.4) presents the concern that the incongruence may be due to homoplasy and perhaps long branch attraction. TSD is distributed nearly identically across all sets of genes, including consistent and inconsistent genes, with a very slight bias toward tree 2 and 3 genes and inconsistent genes having lower TSD (figure 4.8D). Although this does not rule out homoplasy as a source for noise in the inference of gene trees, it appears that regions with high mutational rates are not biased toward supporting incongruent or inconsistent genes (Wilcox, Garcia de Leon et al. 2004), making it a less likely explanatory factor. Additionally, although the trees in figure 4 are not ultrametric (leaves equidistant from internal nodes), they are biased in the opposite direction as would be expected under long branch attraction, with the shortest branch in the

Dmel species subgroup pairing with the longest branch out to *Dana*. Thus, homoplasy and long branch attraction do not appear to be responsible for the incongruence.

Another possibility is that sampling variance in short genes is leading to the incongruence (Pollock, Zwickl et al. 2002). We've already shown that a concatenation of the consistent genes supporting each tree gives 100% bootstrap support, making sampling variance an unlikely explanation. Gene length is very similar across the sets of genes supporting each tree but tree 1 genes tend to be slightly longer than genes supporting trees 2 and 3 (figure 4.8E). Gene length is also weakly correlated with bootstrap value for the whole set, consistent genes and tree 1 genes (both inconsistent and consistent) (Table 4.3). My above results on the spatial correlation of sites, however, suggest that genes that extend more than a few kb would not be expected to be enriched for sites supporting the same tree above their background frequencies. I also found that enrichment is most pronounced for tree 1 sites and less so for incongruent sites. This increased mosaic structure (Hare 2001) in incongruent genes is likely to be responsible for most of the shift to slightly larger genes in the tree 1 genes. The influence of sampling variance, however, is reflected in the shift of inconsistent genes compared to consistent genes toward shorter lengths. Thus the small decrease in long genes in the incongruent set is probably a result of the spatial clustering of sites while the small increase in short genes may be a combination of that effect

and noise from sampling variance. Regardless, gene length is so similar across trees that it is unlikely to explain the incongruence.

Table 4.3. Spearman correlation of sequence property with bootstrap value

Property	All		Tree 1		Tree 2		Tree 3	
	rho	p-value	rho	p-value	rho	p-value	rho	p-value
Dere Quality	-0.011	3.1E-01	-0.026	5.4E-02	0.003	9.0E-01	0.012	6.3E-01
Dyak Quality	-0.028	7.4E-03	-0.048	4.8E-04	-0.011	6.2E-01	0.027	2.7E-01
Dana Quality	-0.008	4.6E-01	-0.010	4.5E-01	0.001	9.6E-01	-0.001	9.8E-01
Length	0.107	<1.0E-10	0.137	<1.0E-10	-0.032	1.3E-01	-0.027	2.5E-01
1st Pos Mean GC	-0.011	2.9E-01	-0.022	1.1E-01	0.006	7.8E-01	-0.015	5.4E-01
1st Pos Std GC	-0.018	7.4E-02	-0.021	1.2E-01	0.021	3.3E-01	0.001	9.7E-01
2nd Pos Mean GC	-0.019	6.4E-02	-0.023	8.6E-02	0.024	2.6E-01	0.008	7.5E-01
2nd Pos Std GC	-0.010	3.1E-01	-0.008	5.5E-01	-0.024	2.7E-01	0.015	5.3E-01
3rd Pos Mean GC	-0.067	<1.0E-10	-0.092	<1.0E-10	-0.005	8.2E-01	0.047	5.1E-02
3rd Pos Std GC	-0.007	4.7E-01	-0.026	5.5E-02	0.015	5.0E-01	0.048	4.4E-02
Kappa	-0.023	2.9E-02	-0.036	8.6E-03	-0.007	7.3E-01	0.030	2.1E-01
dN/dS	0.039	1.7E-04	0.045	8.4E-04	-0.010	6.5E-01	-0.020	4.0E-01
ISD	0.454	<1.0E-10	0.416	<1.0E-10	0.379	<1.0E-10	0.373	<1.0E-10
RINSD	0.515	<1.0E-10	0.469	<1.0E-10	0.471	<1.0E-10	0.419	<1.0E-10
TSD	-0.004	7.2E-01	0.010	4.8E-01	-0.047	2.7E-02	-0.024	3.1E-01

GC content has been estimated to vary considerably across the species in the *Dmel* species subgroup (Akashi, Ko et al. 2006) and is therefore a major concern for systematic bias. I found that GC content is highly similar across species at 1st and 2nd codon positions but varied systematically at the 3rd codon position (figure 4.8F). *Dmel* and *Dana* have nearly identical distributions of 3rd codon position GC content, which is shifted toward lower values compared to

Dere and *Dyak*, which also have nearly identical distributions. This bias in GC content across species is very conservative with respect to the inference of incongruent genes because the incongruence would need to overcome the signal from base composition alone (Phillips, Delsuc et al. 2004). To further verify that this bias only works to decrease the incongruence, I converted the cDNA alignments into R's and Y's, for purines and pyrimadines respectively, and repeated the ML analysis using the F81 model of evolution, effectively averaging the contribution of GC and AT content and only measuring transversions (Delsuc, Phillips et al. 2003; Phillips and Penny 2003; Phillips, Delsuc et al. 2004). As expected, incongruence actually increases (45.2%) under the RY-coding and is still statistically significant (data not shown). Other methods, for example those of Galtier and Gouy (Galtier and Gouy 1995; Galtier and Gouy 1998) and Gu and Li (Gu and Li 1998), attempt to explicitly model non-stationary evolution, rather than control for it. These methods might reveal more precisely the underestimation of incongruence due to the base composition bias in these species but are not expected to provide an explanation for the observed incongruence.

Sequence qualities, transition:transversion ratios and dN/dS values were found be distributed similarly across trees, suggesting they are unlikely factors for systematic bias.

Sequence Properties Associated With Spatial Clustering

I last look to see if the spatial clustering of sites supporting the same tree could be explained by evolutionary rate or base composition variation. To examine the relationship of evolutionary rate and the clustering of sites supporting each tree, I measured total divergence and the fraction of sites supporting each tree in overlapping windows across the chromosomes. For windows of size 5kb or 1kb no correlation can be found between divergence and the fraction of sites supporting each tree, suggesting that evolutionary rate is unlikely to explain the spatial clustering. To test if changes in GC content could explain the clustering of sites I used the RY-coded alignments (described above) (Delsuc, Phillips et al. 2003; Phillips and Penny 2003; Phillips, Delsuc et al. 2004) and repeated the spatial clustering analysis. Sites are still correlated in a similar range of a few kb (data not shown), suggesting that variance in GC content is unlikely to be causing the spatial clustering of sites. Thus both the incongruence as well as the spatial clustering of sites appears to be robust to the sequence and evolutionary properties examined.

Discussion

I initially set out to confirm the placement of *Dere* and *Dyak* as sister species, relative to the *Dmel* lineage, in the *Dmel* species subgroup, using the fully sequenced genomes of seven species in the subgenus *Sophophora*. Although I did find that the best-supported phylogeny is that which places *Dere* and *Dyak* as sister species, I also found pervasive incongruence of substitutions, indels and

gene trees (figure 4.2). While incongruence in substitutions and gene trees could be the result of systematic biases, the incongruent indels, particularly unique insertions, presented strong enough evidence for unbiased incongruence that I also considered incomplete lineage sorting as a possible explanation. Assuming plausible values of substitution rate, generation time and ancestral population size, I found that the time between the split of *Dmel* and the split of *Dere* and *Dyak* is sufficiently short that incomplete lineage sorting would be expected (figures 4.3-4.5). Interestingly, I observed that the support for each of the three trees has a spatial structure across the genome, which is related to low recombination, both locally and globally (figure 4.6). This further supports the hypothesis that the observed incongruence is due, at least in part, to incomplete lineage sorting.

To test for other plausible explanations I examined model choice, species choice and variation in sequence and evolutionary properties and found no obvious candidate factors to explain the incongruence or the spatial structure of support for trees (tables 4.1, 4.2 & 4.3; figures 4.7 & 4.8). I therefore conclude that incomplete lineage sorting is the best going explanation for the lack of resolution in this phylogeny.

Nevertheless, I likely did not exhaust the possible tests for alternate hypotheses for incongruence and suspect that this dataset will prove an interesting area for systematic research, much as the Rokas et al. yeast dataset has [69]. Comparing my results to the yeast dataset reveals important differences: there is significant incongruence beyond what would be expected by chance (figure

4.7A), the level of incongruence is relatively robust to model choice (tables 4.1 & 4.2; figure 4.7B), and basic sequence properties, like GC content, vary in ways that are conservative with respect to the incongruence (figure 4.8) (Phillips, Delsuc et al. 2004). Similar to the yeast dataset, however, I find that the evolutionary model that maximizes the congruence (or accuracy as Ren et al. refer to it) is typically the simplest (HKY) while the model that fits the data best is the most complex (F3x4+G) (table 4.1) (Ren, Tanaka et al. 2005).

To further understand the extent and nature of incomplete lineage sorting in the *Dmel* species subgroup, I suggest several types of future studies. First, to further test the agreement of the observed incongruence with theoretical predictions, better estimates of the ancestral effective population size, mutation rates, time between speciation events, ancestral recombination events (Husmeier 2005) and examining the effects of selection (both directional and balancing (Charlesworth 2006)) would be of clear benefit. In addition, of great interest will be studies of lineage sorting across all taxa in the species group (especially the *Dsim* species complex (Kliman, Andolfatto et al. 2000)) and the influence of migration and gene flow on the symmetry of lineage sorting (because tree 2 is asymmetrically favored). Genome-wide population data already exists for *Dsim* and is expected for *Dmel*, which have the potential to help in the effort to understand these processes. Finally, methodological improvements might include increased large-scale taxon sampling, particularly from closely related taxa outside the species subgroup, such as the *D. suzukii* and *D. takahashii* subgroups

(Lewis, Beckenbach et al. 2005), would alleviate potential biases introduced by the long branches out to *Dana* and *Dpse*.

Although this study should prove quite valuable to the increasing numbers of comparative genomics researchers studying the genus *Drosophila*, I believe my findings have important implications for comparative genomics as a whole. The idea that speciation events have occurred in rapid bursts throughout the tree of life (Adoutte, Balavoine et al. 2000; Rokas, Kruger et al. 2005; Scannell, Byrne et al. 2006) is likely broadly understood (for example the short branch connecting the human, mouse and dog lineages (Kirkness, Bafna et al. 2003)), but the idea that genomes may be mosaics of conflicting genealogies as a result of rapid speciation is perhaps less well appreciated. As more species are sequenced, particularly the dense taxon sampling that is currently beginning in model organism clades, increasing numbers of close speciation events will likely result in many cases of incomplete lineage sorting in genome-scale data. As many methods used in comparative genomics require an accurate phylogeny, the comparative genomics community must develop methods that are robust to or take into account variation in phylogeny.

I envision three types of methods that will need to be developed to appropriately account for this kind of variation. The first are methods that can infer the most likely species tree using an entire genome in a single calculation, considering lineage sorting explicitly. The second are methods that can infer the most likely history of every base in every species, given the species tree. Lastly,

comparative genomics methods that use phylogenies would need to be altered to control for and utilize the output from the second kind of method. Progress is being made in the first two categories (Maddison 1997; Nielsen 1998; Edwards and Beerli 2000; Nielsen and Wakeley 2001; Beaumont, Zhang et al. 2002; Knowles and Maddison 2002; Rannala and Yang 2003; Wall 2003; Beaumont and Rannala 2004; Felsenstein 2004; Hey and Nielsen 2004; Degnan and Salter 2005; Husmeier 2005; Felsenstein 2006; Maddison and Knowles 2006) although no currently available method can deal with a whole genome dataset such as this one. Though well appreciated in the systematics and population genetics communities, the issue of incomplete lineage sorting is rarely considered in the bioinformatics and comparative genomics communities, so the third category of method is virtually non-existent. Accounting for variation in evolutionary histories will have different effects on different classes of methods, but I suggest that parsimony-based methods would be most strongly affected. An important example of such a phylogeny-based method is genome-wide multiple alignment using a guide tree (i.e. (Brudno, Do et al. 2003) & (Blanchette, Kent et al. 2004)), which is the first step in nearly all comparative genomic analyses. The availability of genome-scale datasets such as the one analyzed here should allow rapid progress in all three of these types of methods; I suggest that their development will be of great benefit to the evolutionary and comparative genomics community in the near future.

Methods

Assemblies

Dmel release 4.2 genome, cDNA and translation sequences were downloaded from Flybase (http://www.flybase.net). Pre-publication assemblies for *Dere* and *Dana* (dated August 01st, 2005), sequenced and assembled by Agencourt Bioscience, and for *Dsec* (dated October 28th, 2005), assembled and sequenced by the Broad Institute were downloaded from the Berkeley AAA website (http://rana.lbl.gov/*Drosophila*/). The pre-publication assemblies for *Dyak* (dated July 4th, 2004) and *Dsim* (dated June 2nd 2005) were downloaded from the Washington University School of Medicine Genome Sequencing Center's website (ftp://genome.wustl.edu/pub/seqmgr/yakuba/). The *Dpse* v1.04 assembly was downloaded from Flybase. Sequencing traces corresponding to these genomes are in the NCBI trace archive (http://ncbi.nlm.nih.gov/Traces/trace.cgi, *species_code*, '*DROSOPHILA* ERECTA', '*DROSOPHILA* YAKUBA', '*DROSOPHILA* ANANASSAE', '*DROSOPHILA* SIMULANS', '*DROSOPHILA* SECHELLIA', '*DROSOPHILA PSEUDOOBSCURA*').

Comparative Annotation

Each of the sequence assemblies were annotated separately by mapping *Dmel* gene models onto the unannotated genome in a pairwise fashion using a modified reciprocal-BLAST approach (Wall, Fraser et al. 2003) to assign orthology/paralogy relationships, and a comparative gene finder, GeneWise

(Birney and Durbin 2000; Birney, Clamp et al. 2004), to build gene models. The annotation pipeline consisted of three steps: (I) For each *Dmel* translation, I used the protein sequence as a NCBI TBLASTN (Wheeler, Barrett et al. 2005) query (e-value threshold 1e-3) against the scaffolds of the target assembly. (II) The scaffolds were ordered by the hit e-value reported by TBLASTN and up to two regions were selected from the two best scaffolds and used as input to construct gene models using GeneWise. To improve the chance of constructing a complete gene model using GeneWise, the regions were selected by clustering HSPs on the scaffold such that every HSP within 100kb of another HSP was included in the same region, and a buffer of 10kb was included at the ends of the regions. (III) The predicted translations of the models reported by GeneWise were then used as BLASTP queries against a database of *Dmel* translations, with an e-value threshold of 1e-3.

I then assigned orthology/paralogy relationships using a heuristic algorithm that takes into account (a) the rank of the starting *Dmel* translation in the BLASTP results, (b) the rank of alternative translations from the gene corresponding to the starting *Dmel* translation, and (c) whether or not there were highly ranked hits to genes other than the gene corresponding to the starting *Dmel* translation. One-to-one orthology was assigned when the only top-ranked hits in the BLASTP results were translations from the gene corresponding to the starting *Dmel* translation. Hits having e-values within one order of magnitude were considered to be equivalently ranked. For genes with more than one translation

with clear orthologs in each species, the first historically annotated (translation with the lowest letter ID) was used to represent the gene.

Informative Substitutions & Indels

Informative substitutions supporting each tree were counted across all cDNA and peptide alignments. Only single substitutions that split the four species into two groups of two were considered. Informative substitutions for tree 1 grouped *Dmel* and *Dana* together and *Dere* and *Dyak* together. Likewise tree 2 grouped *Dmel* and *Dere* together and tree 3 grouped *Dmel* and *Dyak* together.

Informative indels supporting each tree were counted across all peptide alignments. Indels were classified as informative in the same way substitutions were. Indels were further filtered to avoid artifacts from alignment errors. Only indels with five amino acids of perfect identity in flanking sequences, with no mono-, di- or tri-amino acid repeats, were included. Insertions were inferred based on an absence in *Dana* and one of the ingroup species. Such insertions, where the inserted sequence is the same in the two species containing it, provided strong, unambiguous characters.

ML Gene Trees

The Codeml program of the PAML package (version 3.14) (Yang 1997; Yang and Nielsen 2000) was run on each gene using the following three unrooted trees: Tree1 - ((*Dmel*,(*Dere*,*Dyak*),*Dana*), Tree2 - ((*Dmel*,*Dere*),*Dyak*,*Dana*) &

Tree3 - ((*Dmel*,*Dyak*),*Dere*,*Dana*) (see figure 4.1). Codeml was run using the F3x4 model, such that equilibrium codon frequencies were calculated from the average nucleotide frequencies at the three codon positions (CodonFreq = 2), amino amino acid distances were equal (aaDist = 0), one dN/dS value was estimated for all lineages using an initial value of 0.4 (model = 0, fix_omega = 0, omega = 0.4), the transition:transversion ratio was estimated with an initial value of 2 (fix_kappa = 0, kappa = 2), substitution rates across sites were set to be equal (fix_alpha = 1, alpha = 0), substitution rates were allowed to vary freely across lineages (clock = 0) and codons with ambiguous positions (gaps or Ns) were ignored (cleandata = 1).

Spatial Analysis

Based on the maximum likelihood tree for each gene, the genome was divided up into blocks supporting each tree. A ten-gene sliding-window was used to calculate a running average of the support for each tree along each chromosome. Each window was assigned a tree based on the most frequent genealogy in the window. Each gene was then reassigned a tree based on the most frequent tree of all the windows that contained it. This effectively allows the neighbors of a gene to influence its assignment, and near neighbors have more influence than far neighbors. Adjacent genes which support the same tree were combined together into blocks. To measure the significance of the size of the blocks, the labels for each gene in the genome were randomized 1000 times and

the blocks were recalculated for each replicate, using the windowing method described above. Recombination rates for a subset of genes in *Dmel,* calculated by Hey and Kliman (Hey and Kliman 2002) using the *R* statistic, were downloaded. The average *R* in each block was calculated where a gene could be found in their set. The Pearson correlation of the average *R* within blocks and the length of blocks was calculated using the R statistics package.

Informative substitutions in genes were used to look at the structure of support for the different trees across the genome independent of the likelihood inference. The counts of each type of informative substitution were calculated in 60 non-overlapping 1kb windows surrounding each informative substitution across all chromosomes. The frequency of each kind of informative substitution across the whole genome was used to calculate an expected count for each 1kb window. In each window, the enrichment of informative substitutions supporting the same tree was calculated. The X^2 significance of windows was calculated by comparing the observed frequencies of informative mutations supporting each tree with the genome averages of those frequencies.

Bootstrap Values

RELL bootstrap values (Kishino, Thorne et al. 2001) from 10,000 replicates were taken from the Codeml output.

PAML Models

All models were run using the same settings as described above for F3x4 except where HKY (model = 4) or WAG+F (model = 3) was specified and where the gamma function was used (fix_alpha = 0, alpha = 1.0, ncatg = 8).

AIC

AIC was calculated as AIC = -2 ln L + 2 N, where L is the likelihood of the model given the data and N is the degrees of freedom (Akaike 1974). Only consistent genes were used in this analysis so the tree was the same across all models. The likelihood and degrees of freedom were taken directly from PAML output. HKY, HKY+G, F3x4 and F3x4+G were compared and WAG+F and WAG+F+G were compared.

Sequence & Evolutionary Properties Analysis

The sequence quality in each species was calculated as the mean sequence quality score of the coding bases. Bootstrap value, length, GC content, transition:transversion ratio, dN/dS, informative synonymous divergence, non-informative synonymous divergence and total synonymous divergence were taken directly from the PAML output for the ML tree from the original analysis using the F3x4 model and the *Dmel*, *Dere*, *Dyak* & *Dana* species combination. The Spearman rank correlations were calculated using the R statistics package (Ihaka 1996).

Divergence Windows

To examine the correlation of divergence with the proportion of sites supporting each tree in local areas across the genome I used 5kb and 1kb windows, overlapping by 2.5kb and 0.5 kb respectively. Using the synonymous site divergences reported by Codeml from the original analysis, I calculated the synonymous divergence per coding site in each window. I also calculated the proportion of sites supporting each tree in each window. Windows with no synonymous coding sites were excluded.

CHAPTER 5

Selective constraints on functional non-coding elements in *Drosophila*

Abstract

Non-coding sequences represent a majority of DNA sequences in metazoan genomes, yet little is know about their composition and how selection acts on non-coding sequences. Selective constraint, the fraction of mutations in a sequence removed by purifying selection, has been estimated to be at high levels throughout the non-coding genome of *Drosophila melanogaster*. Here I examine selective constraints in a large panel of known functional non-coding elements to better understand if they can explain the observed wide-spread selective constraint. *Cis*-regulatory modules, transcription factor binding sites and ncRNAs all appear to be evolving under strong purifying selection. Using these estimates I formulate a relationship between the proportion of non-coding sequences covered by *cis*-regulatory modules and ncRNAs, which suggests that up to 65% of non-coding sequences are covered by ncRNAs and up to 90% of non-coding sequences are covered by *cis*-regulatory modules. Finally, I estimate that between 70-80% of bases in *cis*-regulatory modules are transcription factor binding sites.

Background

Genome sequences of closely related species provide the opportunity to improve our understanding of how selection acts on genetic diversity as well as our understanding of the molecular function of genome sequences. Protein-coding sequences have long been the target of evolutionary analyses, yet little is known about the selective processes that act on non-coding DNA nor the functional role most non-coding DNA performs.

Selection acting on non-coding sequences likely includes many diverse mechanisms (e.g. (Andolfatto 2005)), but purifying selection is perhaps both the most common and the most useful for annotating molecular functions (Li 1997). Purifying selection acts to stabilize phenotypes and molecular functions, leading to a decrease in genetic diversity within and between populations through the removal of disfavored alleles. The signal of purifying selection at the sequence level, therefore, is lowered levels of interspecific diversity. This so-called "selective constraint" has been effectively used to identify portions of non-coding sequences with molecular functions actively maintained between species, such as *cis*-regulatory modules (CRMs) (e.g. (Boffelli, McAuliffe et al. 2003)) and non-coding RNAs (ncRNAs) ((Rivas and Eddy 2001)). Despite the effectiveness of using selective constraint to help with annotation, selective constraints on non-coding sequences have only recently been the subject of quantitative characterized.

Analyses of selective constraints on non-coding sequences have revealed a few basic properties of selective processes. First, most of the constrained DNA in

the genome falls outside of protein coding genes (Keightley and Gaffney 2003; Halligan, Eyre-Walker et al. 2004; Keightley, Kryukov et al. 2005; Halligan and Keightley 2006). Second, constrained nucleotides tend to be clustered together, suggesting a blocky distribution of functional elements, consistent with our understanding of modular *cis*-regulatory elements and ncRNAs (e.g. (Bergman, Pfeiffer et al. 2002; Boffelli, McAuliffe et al. 2003; Halligan and Keightley 2006)). Third, compact genomes, such as those found in fruitflies, compared to large genomes, such as those found in mammals, tend to have ubiquitous non-coding constraints with little unconstrained sequences separating clusters of constrained nucleotides (Nobrega, Ovcharenko et al. 2003; Singh and Petrov 2004; Halligan and Keightley 2006). Fourth, the fastest evolving sequences in compact genomes are bases in short introns that are not part of the splicing machinery (Halligan and Keightley 2006).

These observations have lead to the hypothesis that the constraint in non-coding sequences may be explainable by broad distributions of CRMs and ncRNAs across the genome (Halligan and Keightley 2006). This hypothesis rests on three assumptions. The first assumption is that CRMs and ncRNAs tend to be subject to strong enough levels of purifying selection to explain non-coding constraints. The second assumption is that CRMs and ncRNAs are broadly distributed enough across the genome to explain non-coding constraints. The third assumption is that there are not additional classes of functional non-coding elements that have not yet been identified that could also partially explain the

non-coding constraints. Here I address the first assumption of the strength of selective constraints on CRMs and ncRNAs and discuss the implications for the second and third assumptions.

Results

Definition of constraint

Constraint, *C*, is defined as the fraction of mutation removed due to purifying selection. Using the approach of Halligan & Keightley (Halligan and Keightley 2006), constraint for a region R is estimated as: $C_R = 1 - d_R / d_{SI}$, where d_R is the interspecific divergence for regions R and d_{SI} is the mean interspecific divergence for short introns in the megabase chromosomal window in which region R is found. Thus *C* varies from zero to one and is normalized for megabase-scale variation in mutation rate.

Species choice & alignments

In this study I calculate constraint using pairwise alignments of *D. melanogaster* with either *D. simulans* or *D. yakuba*. Alignments with the more closely related *D. simulans* ought to have fewer errors but also smaller divergence differences between constrained and unconstrained sequences compared to alignments with the more distantly related *D. yakuba*. Simulations indicate that the error in pairwise divergence estimates for these species is expected to be at most a few percent (see chapter 3).

Orthologous sequences and alignments were generated using synteny maps and the alignment tool MLAGAN, as described in Moses et al. (Moses, Pollard et al. 2006).

Cis-regulatory and ncRNA annotations

Decades of published experimental annotations of *cis*-regulatory elements in *Drosophila* have recently been compiled into two databases. Experimentally identified enhancers and promoters have been collected into the REDfly database (Gallo, Li et al. 2006). After filtering out large encompassing constructs and collapsing overlapping features, a set of 322 non-redundant CRMs was produced. 879 DNAse 1 footprinted binding sites for 79 transcription factors at 101 loci have been compiled into the FlyReg database (Bergman, Carlson et al. 2005).

Non-coding RNAs of different classes have undergone rapid identification in recent years. Here I use a large collection of validated tRNAs, miRNAs, snRNAs and snoRNAs compiled by Bergman and colleagues (Clark, Eisen et al. 2007).

Selective constraints in protein-coding & non-coding sequences

I first examined selective constraints in protein coding and non-coding sequences as points of comparison for the functional non-coding elements. I find that mean selective constraint is high for non-synonymous sites, and low but significantly positive for four-fold degenerate synonymous sites (figure 5.1) in

both pairwise comparisons (*D. simulans* data not shown). Although significantly different, my estimates for the *D. melanogaster* / *D. simulans* comparison are comparable with those from a previous study (Halligan and Keightley 2006), which used the same pairwise comparison but a subset of the genes used here. They are also consistent with many previous studies that have demonstrated that amino-acid changing mutations are, for most the part deleterious (e.g. (Kimura and Takahata 1983)) and that there is weak selection on synonymous sites in *Drosophila*, potentially due to selection for translation efficiency, resulting in codon-usage bias. Furthermore, I find that selective constraint, per base pair, in non-coding DNA is approximately 0.56 (figure 5.1), confirming previous results (Halligan and Keightley 2006).

Interestingly, estimates of selective constraints appear to be higher when using the *D. melanogaster* / *D. yakuba* comparison relative to the *D. melanogaster* / *D. simulans* comparison. Four-fold degenerate sites in particular show a discrepancy (0.200 vs 0.123), but the difference is significant for all sequences classes (data not shown). It has been suggested that selection has been less effective along the *D. melanogaster* lineage (Akashi 1996), which is consistent with these results due to the relatively greater proportion of total branch length the *D. melanogaster* lineage makes in the comparison with *D. simulans*. This discrepancy may also be the result of greater power to detect constraint in the more divergent comparison with *D. yakuba*. The effect, however, appears to be greatest for four-fold degenerate sites (the ratio of constraint in *D. melanogaster* /

D. simulans to *D. melanogaster* / *D. yakuba* is 0.617, 0.995 and 0.897 for four-fold, non-degenerate and noncoding sites respectively), which is more consistent with weakened selection along the *D. melanogaster* lineage, because power issues would influence more slowly evolving sites while the efficacy of selection would influence more weakly constrained sites. To further investigate this, I calculated lineage specific estimates of constraint in *D. melanogaster* and *D. simulans* separately, assigning mutations to each lineage by parsimony using *D. yakuba* as an outgroup. Lineage specific constraint is not significantly lower in the *D. melanogaster* lineage than the *D. simulans* lineage for non-degenerate sites (0.882 95% C.I. [0.878,0.885] and 0.882 [0.879,0.886] for *D. melanogaster* and *D. simulans* respectively), but is significantly lower for four-fold degenerate sites (0.0635 [0.0554,0.0715] and 0.184 [0.176,0.193] respectively). These results suggest that the higher constraints in the *D. melanogaster* / *D. yakuba* comparison are the result of both relaxed constraints along the *D. melanogaster* lineage as well as the increased power from the longer total branch length. Thus for the remainder of this study I shall focus on comparisons using *D. yakuba*.

Selective constraints on functional elements are higher than non-coding averages

Both the CRMs in the REDfly database and the DNAse I footprints in the FlyReg database show significantly higher selective constraints than average non-coding sequences (figure 5.1). While the constraints on these regulatory sequences are high (more than 6 out of ten mutations removed by selection), they

are only modestly greater than non-coding averages, consistent with previous findings that conservation alone does not discriminate functional *cis*-regulatory sequences from other sequences in *Drosophila* (Berman, Pfeiffer et al. 2004; Sinha, Schroeder et al. 2004). Because DNAse I footprinting is an imprecise method for identifying transcription factor binding sites, I searched for the most likely binding sequence overlapping the footprints using position weight matrices (PWMs) for the 30 factors for which good quality matrices had been constructed (Down, Bergman et al. 2007). These binding sites, which I shall refer to as PWMFPs, have significantly higher constraints compared to the footprints from which they are derived, yet are still less than non-degenerate coding sites (figure 5.1). Thus, *cis*-regulatory sequences are under high selective constraints and may be able to explain some portion of the ubiquitous non-coding constraints in *Drosophila*.

Non-coding RNAs, of all the classes examined here, are under considerable selective constraints (figure 5.1). For snRNAs and snoRNAs, constraints are similar to those of *cis*-regulatory sequences, however, for miRNAs and tRNAs, constraints are even higher than for non-degenerate coding sites. These results are consistent with results suggesting that evolutionary information is greatly beneficial for the identification of functional ncRNAs (e.g. (Rivas and Eddy 2001)). The high constraints in ncRNAs suggest that they too would be able to explain some portion of the ubiquitous non-coding constraints in *Drosophila*.

Variation in selective constraints across functional non-coding elements

While functional non-coding elements are subject to very high selective constraints, there is considerable variation in constraint within each kind of element. In an attempt to explain the variation across elements, I first examined the relationship of annotation length with constraint. None of the functional element classes showed a significant correlation of length with constraint with the exception of REDfly CRMs (Pearson's R = -0.2919767 95% C.I. [-0.4037162,-0.1798992]) and FlyReg footprints (R = -0.2120202 [-0.2738236,-0.1508729]), which both showed weak negative correlations, presumably due to variation in the methods used to define the boundaries of the elements.

I next focused on variation in constraint across the FlyReg footprints and the PWMFPs. According to both sets, there is significant variation in constraint with respect to the transcription factors that bind the footprints (Kruskal-Wallis: FlyReg p-value<0.016; PWMFP p-value<0.0022), suggesting either actual selective differences between factors or experimental differences between sources for different factors. FlyReg footprints also showed significant variation in constraint across the loci in which they are located (Kruskal-Wallis p-value<7.222e-15), suggesting again either variation in the strength of purifying selection acting on the regulatory sequences for different genes or, again, experimental differences between sources for different loci. Finally, I was interested in seeing if the predicted affinity of the PWMFPs (see methods) is related to constraint. My expectation was that mutations in stronger sites would be

more likely to have a large impact on gene expression and would therefore be more strongly selected against. In the PWMFPs I found no correlation of predicted affinity with constraint (-0.01884575 95% C.I. [-0.08538009,0.04888508]), suggesting either constraint is unrelated to affinity or that affinity estimations using PWMs are too imprecise to detect a relationship (Maerkl and Quake 2007).

Implications of constraints on functional elements for genome architecture

As I stated above, our ability to explain the ubiquitous non-coding constraints in *Drosophila* depends on the strength of selection acting on known functional non-coding elements, their distribution across non-coding sequences and the strength of selection on and distribution of yet-to-be-identified functional non-coding elements. This relationship can be formally defined as follows:

$C_{NC} = f_{CRM} C_{CRM} + f_{ncRNA} C_{ncRNA} + f_{Unknown} C_{Unknown}$

where C_{NC} is the mean constraint in non-coding sequences, f_{CRM} is the proportion of non-coding sequences covered by CRMs, C_{CRM} is the mean constraint in CRMs, f_{ncRNA} is the proportion of non-coding sequences covered by ncRNAs, C_{ncRNA} is the mean constraint in ncRNAs, $f_{Unknown}$ is the proportion of non-coding sequences covered by unknown functional elements and $C_{Unknown}$ is the mean constraint in these unknown functional elements. We can use our estimates of constraint in known CRMs and ncRNAs for the genome averages and we already know the mean non-coding constraint. That leaves four unknowns. For

simplification to better understand the explanatory power of just CRMs and ncRNAs, I dropped the unknown elements from the equation. This leaves two unknowns. Figure 5.2A shows the relationship of f_{CRM} and f_{ncRNA} if they are the only explanatory factors for the ubiquitous non-coding constraints. The maximum proportion of non-coding sequences covered by CRMs would be 90% and the maximum for ncRNA would be 65%, with the proportion of unconstrained sequence ranging from 10% to 35%. While the actual distributions of CRMs and ncRNAs across the genome are not well understood, the *even-skipped* locus, which is arguably the best annotated locus in *Drosophila*, has CRMs for the *eve* gene covering 73% of non-coding sequences. Note that this proportion does not include regulatory sequences for adjacent genes. The corresponding ncRNA proportion is 12%, leaving 15% unconstrained non-coding sequences (figure 5.2A).

A similar analysis was performed examining the proportion of CRMs covered by transcription factor binding sites. The total selective constraint in CRMs can be written as:

$$C_{CRM} = f_{BS}\, C_{BS} + (1 - f_{BS})\, C_{SP}$$

where C_{CRM} is the mean constraint in CRMs, f_{BS} is the proportion of CRMs covered by transcription factor binding sites, C_{BS} is the mean constraint in binding sites and C_{SP} is the mean constraint on non-binding site sequences, or spacers. This equation leaves two unknowns, f_{BS} and C_{SP}, if we use our estimates from above for C_{CRM} (0.617) and C_{BS} (0.735). Figure 5.2B shows the relationship of f_{BS}

and C_{SP}. The maximum estimate for f_{BS} is 84% and, trivially, the maximum estimate for C_{SP} is 0.617. The shape of the curve suggests that considerable constraints in spacer sequences are necessary in order for the estimate of f_{BS} to be much less than its maximum.

Discussion

In this study I examined the selective constraints acting on functional non-coding elements in order to better understand the wide-spread constraints (Halligan and Keightley 2006) acting on non-coding sequences in the compact genome of *Drosophila melanogaster* (Petrov and Hartl 1997).

I used divergences estimated from short introns as a neutral standard to calculate selective constraints on other sequences (Halligan and Keightley 2006). While this approach handles mega-base scale variation in mutations rates, smaller-scale variation may cause some error in specific constraint estimates, but is unlikely to affect the general results of this analysis (Halligan, Eyre-Walker et al. 2004). Also, if positive selection is acting frequently on non-coding sequences (Andolfatto 2005), constraint may be underestimated for those sequences. The short range of linkage disequilibrium in *Drosophila* (ref), however, suggests that positive selection ought to have a minimal impact on general estimates of constraint. Thus, my estimates of selective constraint and the conclusions I draw from them ought to be largely accurate.

I find that selective constraints on experimentally annotated *cis*-regulatory

elements and ncRNA elements are very high, with between 6 and 9 out of ten mutations are removed due to purifying selection (figure 5.1). Regulatory sequences are generally less constrained than ncRNAs, while many ncRNAs are as constrained or more than non-degenerate protein coding sites (figure 5.1). Variation in selective constraints across functional elements could be partially explained by length (CRMs & footprints), locus (footprints), and cognate transcription factor (footprints & PWMFPs).

Using my accurate measurements of selective constraint in CRMs and ncRNAs, I addressed the fundamental question of the functional composition of the non-coding genome (Bird, Stranger et al. 2006). Making the necessary simplifying assumption that CRMs and ncRNAs are the only functionally constrained non-coding elements, I find CRMs are expected to cover between 0-90% of non-coding sequences while ncRNAs are expected to cover between 0-65% of non-coding sequences. Although current knowledge of the distribution of these non-coding elements is limited, the example of the eve locus with 73% of surrounding non-coding sequence covered by CRMs for eve alone, as well as the intuition that most genes require some minimal *cis*-regulatory machinery to orchestrate expression, suggest that the actual distribution of CRMs in the genome may be closer to my estimated maximum of 90% than ncRNAs to their maximum.

My estimates of constraint in CRMs and binding sites allowed me to investigate a similar question of how much of CRM sequences are covered by transcription factor binding sites. Even for the best studied CRMs, there is always

lingering doubt as to whether or not all the regulators and all of the binding sites have been identified. Allowing both the proportion of CRMs covered by binding sites and the strength of constraint in non-binding site spaces sequence to be unknowns, I found that the maximum binding site proportion of CRMs is 84% and that constraints in spacer sequences would need to be very high in order for this proportion to drop much below 70%. This result combined with the previous result, suggesting that much of the genome is covered by CRMs, suggests that there may be millions of transcription factor binding sites functioning in the *D. melanogaster* genome.

Methods

Short introns

Short introns were defined as introns 65bp or less in length. To avoid sequences involved in splicing I only used positions 8-35 (5'-3'). Mean divergence was calculated across non-overlapping megabase sized windows. See Halligan and Keightley for more details (Halligan and Keightley 2006).

Position weight matrix predictions

Position weight matrix (PWM) predictions were done using PATSER (Hertz, Hartzell et al. 1990) with a background model parameterized with the *Drosophila melanogaster* genome average base composition of 60:40, AT:GC.

CHAPTER 6

Conclusion

In this work I have developed and tested an infrastructure for studying *cis*-regulatory evolution on a systematic level and utilized this framework to measure selective constraints on, and the evolutionary dynamics of, *cis*-regulatory sequences in *Drosophila*.

Taking the process an evolutionary biologist might go through "by hand" to find, align and analyze sequences from divergent species to an automated level, where quality is kept high, requires the development of substantial computational infrastructure. Much of my thesis work concentrated on asking if the methods available for automated evolutionary analysis are sufficiently accurate. In cases where it was clear how I could improve the process, I developed new approaches. With this detailed understanding of the strengths and weaknesses of the methods used in automated comparisons, I was able to clearly interpret my results with minimal ambiguity as to how methodological errors may impact my inferences.

My graduate work made substantive advancements in our understanding of *cis*-regulatory sequences, yet much remains unclear about the genome distribution, organizational structure, and mechanism of function of *cis*-regulatory sequences as well as the evolutionary processes that act on *cis*-regulatory sequences to maintain and diversify function. Experimental and evolutionary

results suggest that each of the few hundred transcription factors encoded in the *Drosophila melanogaster* genome may regulate hundreds of genes through thousands of binding sites in the genome. Each of these regulatory regions appear to be dense clutters of binding sites for different factors, with little non-functional sequence within or between elements. Purifying selection, presumably to maintain protein binding and therefore regulatory functions, acts strongly to maintain regulatory sequences. Compared to neutral expectations, fewer than three out of ten mutations are tolerated, comparable to constraints on protein-coding sites and non-coding RNAs. The little variation in regulatory sequences that is tolerated, however, results in the loss and gain of binding sites at measurable rates over the course of just a few million years. The birth and death of these sites appear to not strictly maintain the same concentration of binding in the same location but rather lead to local fluctuations in a lineage-specific manner.

While my work answered many questions, it also produced many new ones. Perhaps the most fundamental questions pertain to how gene expression and phenotype are maintained in the face of this diversity of regulatory sequences and further, the role that these constantly shifting regulatory sequences play in the generation of phenotypic diversity. Both of these questions step out from the framework of my graduate work, which was to condition on the known regulatory function of a sequence in one species and to then ask questions about how that sequence varies across species. These questions require more functional knowledge in other species. To ask how phenotype and expression is maintained

between species requires that you can assert that function is actually maintained. To ask how new phenotypes and expression patterns are generated across species requires that you can assert that function has significantly changed and would benefit from a quantification of such changes. Single locus work in the field has contributed greatly to our understanding of these relationships, but addressing these questions on a systematic level is the challenge for tomorrow.

APPENDICES

Appendix A

Systematic identification of cis-regulatory sequences active in Drosophila early embryonic development

Efforts to understand how *cis*-regulatory sequences function and evolve have been limited by the number of known *cis*-regulatory elements with known regulators. So-called "promoter bashing" molecular genetics approaches to identifying *cis*-regulatory elements require enormous resources to perform on a large scale. New genome-scale approaches, such as ChIP-chip (Lieb, Liu et al. 2001), have been developed that have the potential to greatly accelerate the annotation of *cis*-regulatory sequences.

I participated in a large group study (http://bdtnp.lbl.gov) in which we performed chromatin immunoprecipitations (ChIP) of maternally deposited and gap transcription factors on early staged whole *Drosophila melanogaster* embryos and then hybridized the enriched DNA onto high-density tiling arrays (chips) (Lia, MacArthura et al. In Press). The results of these experiments were that transcription factors known to act in the early stages of the developmental network bind an overlapping set of hundreds to thousands of sequences across the genome with quantitatively different specificities. A manuscript describing this work is currently in press at PLoS Biology (Lia, MacArthura et al. In Press).

I helped identify that these bound regions are typically enriched for

predicted binding sites (Hertz, Hartzell et al. 1990) using position weight matrices derived from *in vitro* binding experiments (Bergman, Carlson et al. 2005). I also identified that bound sequences are typically biased in their base composition relative to surrounding non-coding sequences, such that the peak of binding intensity corresponds with a local peak in GC bases. I found that the enrichment of binding sites is independent of the GC enrichment by demonstrating that permuted PWMs do not show enrichment in bound regions. The GC bias may be a general feature of *cis*-regulatory sequences as it has been observed in regulatory sequence identified by traditional methods (Li, Zhu et al. 2007). I helped show that bound sequences tend to have lower divergences than surrounding non-coding sequences and that bound regions are enriched for conserved binding sites. I also examined the phylogenetic patterns of binding site gains and losses in bound regions, as will be discussed in Appendix E.

We found that the enrichment of cognate binding sites is insufficient to fully classify these bound regions from the rest of the non-coding genome, suggesting that yet unidentified properties of functional *cis*-regulatory sequences are determining binding. In support of the quality of our experiments, we found that all known targets of the studied factors (Gallo, Li et al. 2006) were recovered as highly bound. Further hundreds of previously unknown binding regions were identified near transcription factors known to participate in the embryonic network (Nusslein-Volhard and Wieschaus 1980) as well as genes previously identified to have patterned embryonic expression (Tomancak, Berman et al.

2007). Interestingly, the new targets included most of the miRNAs transcribed during embryogenesis (Aboobaker, Tomancak et al. 2005; Biemar, Zinzen et al. 2005) as well as the dorsal-ventral patterning genes. Finally, we found thousands of reproducibly bound regions that tend to be more weakly bound, tend to be distant from genes active during early embryogenesis, tend to bind protein-coding sequences and tend to be relatively poorly conserved, suggesting they are bound but not functional during early development.

In conclusion, this study vastly expands the number of *cis*-regulatory sequences with known regulators for future studies of *cis*-regulatory function and evolution.

Appendix B

Annotating & aligning protein coding genes in 12 fully sequenced Drosophila species

High-throughput sequencing technologies are making comparative genomics a standard component of biological research, shifting the rate limiting step to annotation and analysis of newly sequence genomes. In addition to the previously sequenced genomes of *Drosophila melanogaster* and *D. pseudoobscura*, the genomes of ten other *Drosophila* species were recently sequenced and assembled. To facilitate studying gene evolution as well as identifying orthologous non-coding sequences (see Appendix C), I participated in the comparative annotation and alignment of protein coding sequences across

these twelve *Drosophila* species. The results of these efforts were recently published in the community-wide effort to analyze the 12 species (Clark, Eisen et al. 2007).

In brief, we utilized the high quality gene annotations in *Drosophila melanogaster* to identify putatively homologous sequences in the other eleven species and ran several forms of gene mapping software (EXONERATE (Slater and Birney 2005), GENEMAPPER (Chatterji and Pachter 2006) & GENEWISE (Birney, Clamp et al. 2004)) on these putatively homologous sequences to build gene models. We then used the consensus gene building software, GLEAN (Elsik, Mackey et al. 2007), to produce a single annotation homology-based annotation set. Other research groups produced non-homology-based gene model predictions which we combined with the homology-based gene models in another run of GLEAN to produce an additional set of consensus gene models. This second set included exons that had been missed by the homology-based annotation but also included cases of incorrectly fused or separated exons and genes. To address this problem, we replaced models in the combined set with models from the homology-based set when incorrect fusions and separations were detected. This final resolved set was used by the community for downstream analyses.

It is worth noting that this project revealed two major problems with the current state of the art for genome annotation. The first issue is that GLEAN, which is the only current option for consensus gene building, is not able to handle alternative splice forms of a given gene. In cases where alternate exons are used

in different splice forms, GLEAN either picked one or created a path through the two exons that does not exist as far as we know. Development of alternative splice-aware consensus gene annotation software is an essential next step for genome annotation. The second issue is the difficulty in resolving homology-based and non-homology-based annotations at the stage of making consensus gene annotations. While the methodology we used allowed us to resolve some of the conflicts, it would greatly benefit consensus annotation if software like GLEAN allowed different classes of input annotations to be treated in appropriately different ways.

The homology relationships of these gene models across the 12 species were established using a similar method to the reciprocal blast approach (Wall, Fraser et al. 2003). Homology groups were built using all-by-all BLASTP (Altschul, Gish et al. 1990) searches where connections between models were conditioned on reciprocal-best-hits (allowing non-top hits to fall within 2 logs of the top E-value). These homology groups identified both one-to-one orthologs as well as closely related paralogous relationships.

Alignments of different partitions of the homology groups were performed at the amino-acid level using TCOFFEE (Notredame, Higgins et al. 2000). Low quality alignment segments were masked by identifying species-specific alignment properties (divergence & gap coverage) that were significantly associated with bad alignment segments flagged by human inspection of a subset

of the total alignments. High-quality alignments were thus generated for community analyses of gene evolution (e.g. (Sackton, Lazzaro et al. 2007)).

Appendix C

Using synteny maps and similarity to map and align orthologous non-coding regions across 11 fully sequenced Drosophila species

Finding and aligning orthologous non-coding sequences in divergent speciecs is an essential step in the process of studying non-coding and *cis*-regulatory evolution. While local alignments methods can recover large highly similar nucleotide sequences (see chapter 2), additional information about orthology can be used to improve inferences. Here I describe a method I developed that uses gene orthology to aid in the identification of orthologous non-coding sequences that has been used in a number of studies (chapter 5, appendices A & E).

The first step is to build a synteny map of orthologous chromosomal regions across species. This is accomplished by starting with a set of orthology assignments of gene models across species and grouping orthologous genes into synteny blocks where the order and orientation of the genes is maintained across species.

The second step is to determine the appropriate set of sequences to consider as putative orthologous non-coding sequences. If sequences are contained within synteny blocks then the appropriate sequences are the intergenic

sequences spanning adjacent gene models. If the sequences are not in synteny blocks then multiple intergenic sequences must be considered. The nearest adjacent genes with orthologs in the other species are considered potential anchors. To avoid problems with inversions switching the relative location of the non-coding segment and the gene, both the upstream and downstream intergenic sequences are returned.

The third step is to perform a BLASTN (Altschul, Gish et al. 1990) search with the query sequence, as a backup in case this method fails to improve the orthology mapping. The top HSP, if significant (Evalue<10^{-3}), is extended to attempt to recover the whole orthologous region.

The fourth step is to perform a global pairwise alignment of each of the putatively orthologous non-coding regions with the query non-coding sequence. Based on my analysis of pairwise alignment tools in chapter 2, I have been using LAGAN (Brudno, Do et al. 2003). To catch small inversions, both strands of the putatively orthologous non-coding sequence is aligned. In each alignment the segment of the alignment overlapping the query sequences is extracted for the next step.

The fifth step is to measure alignment quality for all the putative alignments, filter out clearly non-orthologous sequences and determine the best match. Alignment quality is based on the percent identify and the percent ungapped of the columns in the alignment. Using LAGAN alignments of random non-coding sequences, I have fit a line to the percent identity and percent

ungapped expected for alignments non-orthologous sequences. Using this filter, clearly non-orthologous sequences are removed. Finally the statistic of the sum of the percent identify and percent ungapped is used to rank the remaining candidate alignments.

While no systematic testing of this methodology has been performed, empirically, the synteny-based sequences are determined as the best orthologous sequences over the BLASTN-based method in more than 99% of cases. Further, local inversions are successfully handled some fraction of the time by simply considering the opposite strand of a sequence. Thus the method is an improvement over simply running BLASTN but clearly greater improvements could be made. Obvious improvements could include using an additional local alignment step prior to the global alignment to attempt to identify local inversions that might not be properly handled in the global alignment and using a statistical approach to the filtering step to better handle highly divergent species as well as clear outliers in less divergent species. A simulation-based evaluation of the method would also be helpful for identifying the strengths and weaknesses of the approach.

Appendix D

Finding functional transcription factor binding sites using a factor-specific evolutionary model

The prediction of transcription factor binding sites using affinity estimating position weight matrix (PWM) models is challenged by low specificity yet conditioning on functional conservation across species promises to improve predictive power. Using a factor-specific evolutionary model, originally developed for modeling position-specific evolutionary rates for codons (Halpern and Bruno 1998), I helped develop software, called MONKEY, that can evaluate the probability of a sequence having been functionally conserved through a phylogeny using a multiple alignment and PWM describing the binding site (Moses, Chiang et al. 2004).

To evaluate MONKEY we compared its ability to assign a higher probability to known functional vs. non-functional binding sites in yeast to two alternate approaches. The first alternate approach was the single genome probability using a PWM (Hertz, Hartzell et al. 1990). The second alternate approach was to consider the average of the single genome probabilities across a set of species. The single genome method consistently assigned lower probabilities than the averaging method, which consistently assigned lower probabilities than MONKEY to functional sites. For the non-functional sites, MONKEY assigned a lower probability than the averaging method 70-75% of the time, while the averaging method assigned a lower probability than the single genome method 52-70% of the time. Thus MONKEY appears to be superior to current methodologies for identifying functional binding sites.

Lastly, we tested the dependence of the power of the method on the divergence of the species analyzed. As expected, as the divergence spanning the species in the alignment increases, the ability of the method to distinguish functional from non-functional sites increases and is well modeled by the p-values assigned to the probabilities by the method.

The development of this method facilitates more accurate predictions of functional transcription factor binding sites using comparative sequences data. The specificity of the method is still not particularly high, suggesting that further improvements in modeling functional binding sites are possible and needed.

Appendix E

Systematic analysis of transcription factor binding site gains and losses in Drosophila

Gains and losses of transcription factor binding sites potentially underlie a large proportion of gene expression and phenotypic evolution between species (Wray 2007). Single locus studies (e.g. (Gompel, Prud'homme et al. 2005; Ludwig, Palsson et al. 2005)), suggest that such gains and losses can happen over the span of a few million years and that these changes can occur in a compensatory fashion that does not change expression, in which losses of sites for a factor are matched with gains along the same lineage, or a lineage-specific fashion that does change expression, in which new sites accumulate or are lost along a given lineage. To systematically assess the tempo and mode by which

sites are gained and lost across phylogenetic trees, I helped develop a method for probabilistically identifying non-conserved binding sites and applied it the ChIP-chip for early embryonic transcription factors (see appendix A) (Moses, Pollard et al. 2006).

We found that over the span of the few million years separating *D. melanogaster* from the other four sequenced species in the *D. melanogaster* subgroup, between 15-48% of binding sites in *D. melanogaster* bound regions have been gained or lost across the tree. Concerned that these estimates were inflated, we examined if alignment errors or spurious non-functional binding site prediction might be causing an over-estimation of turnover. Simulations revealed that alignment errors are expected to contribute on the order of 0.1% to our estimates of turnover within the *D. melanogaster* subgroup even though inclusion of the next most diverged species, D. anannase, would produce high levels of error (~15%). To correct for background levels total sites and particularly non-conserved sites in the ChIP-chip bound regions we measured the density of total sites and non-conserved sites in flanking non-coding regions and subtracted this expected density from the density of sites in the bound regions. This correction reduced the proportion of sites that have been gained or lost over the tree to between 0-23%, with Krüppel and Zeste showing the highest levels of turnover (13-23%), Bicoid and Hunchback intermediate levels (4-6%) and Caudal and Medea levels indistinguishable from zero. Thus we can say that the tempo of binding site turnover is idiosyncratic but can be as high as a few percent per

million years, expanding our understanding of this process from previous more heuristic basic estimates (Costas, Casares et al. 2003; Dermitzakis, Bergman et al. 2003).

To evaluate the mode by which sites are gained and lost, we looked for sites in any species in alignments of the ChIP-chip bound regions, and then classified these sites as non-conserved or not, and then further classified them as lineage-specific gains and losses. Compensatory evolution ought to show a signal of gains and losses in a given region falling on the same branch of the tree. Across the factors we found many examples of matched gains and losses, but permutation tests suggested that the observed levels of matched gains and losses would be expected by chance. Lineage-specific gains and losses that alter function ought to show the signal of an increased proportion of gains or losses along that lineage compared to the expected proportion based on divergence. Across factors, the *D. melanogaster* lineage consistently showed the greatest deviation from expectation, with increased proportions of gains and decreased proportions of losses. This net flux of more sites along the *D. melanogaster* lineage suggests that our experiments identified bound regions that are not bound with the same strength in the other species, perhaps because these are new functions or because the binding regions are moving within their loci.

This work establishes a framework for analyzing binding site gains and losses but much work remains for deriving a deep understanding of this process. The ChIP-chip data used in this study has a low enough resolution that many of

the predicted sites are expected to be non-functional, effectively lowering our signal to noise. Newer genome-wide binding methods, such as ChIP-seq (Schmid and Bucher 2007), ought to help greatly with reducing this noise. Secondly, while our method uses a probabilistic model to identify non-conserved sites, our method for identifying the lineage on which gains and losses occur is heuristic. Probabilistic methods that infer specific lineage gains and losses (e.g. (Wagner, Otto et al. 2007)) would likely provide more accurate estimates of the tempo and mode of changes.

REFERENCES

Aboobaker, A. A., P. Tomancak, et al. (2005). "*Drosophila* microRNAs exhibit diverse spatial expression patterns during embryonic development." Proc Natl Acad Sci U S A **102**(50): 18017-22.

Adoutte, A., G. Balavoine, et al. (2000). "The new animal phylogeny: reliability and implications." Proc Natl Acad Sci U S A **97**(9): 4453-6.

Akaike, H. (1974). "A new look at the statistical model identification." IEEE Trans Autom Contr **19**: 716-723.

Akashi, H. (1996). "Molecular evolution between *Drosophila melanogaster* and *D. simulans*: reduced codon bias, faster rates of amino acid substitution, and larger proteins in D. *melanogaster*." Genetics **144**(3): 1297-307.

Akashi, H., W. Y. Ko, et al. (2006). "Molecular evolution in the *Drosophila melanogaster* species subgroup: frequent parameter fluctuations on the timescale of molecular divergence." Genetics **172**(3): 1711-26.

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

Alvarez, M., S. J. Rhodes, et al. (2003). "Context-dependent transcription: all politics is local." Gene **313**: 43-57.

Andersson, J. O. (2005). "Lateral gene transfer in eukaryotes." Cell Mol Life Sci **62**(11): 1182-97.

Andolfatto, P. (2005). "Adaptive evolution of non-coding DNA in *Drosophila*." Nature **437**(7062): 1149-52.

Andolfatto, P. and J. D. Wall (2003). "Linkage disequilibrium patterns across a recombination gradient in African *Drosophila melanogaster*." Genetics **165**(3): 1289-305.

Arbogast, B. S., S. V. Edwards, et al. (2002). "Estimating Divergence Times From Molecular Data On Phylogenetic And Population Genetic Timescales." Annu Rev Ecol Syst **33**: 707-740.

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

Arndt, P. F., C. B. Burge, et al. (2003). "DNA sequence evolution with neighbor-dependent mutation." J Comput Biol **10**(3-4): 313-22.

Averof, M., A. Rokas, et al. (2000). "Evidence for a high frequency of simultaneous double-nucleotide substitutions." Science **287**(5456): 1283-6.

Avise, J. C., J. F. Shapira, et al. (1983). "Mitochondrial DNA differentiation during the speciation process in Peromyscus." Mol Biol Evol **1**(1): 38-56.

Bailey, T. L. and C. Elkan (1995). "The value of prior knowledge in discovering motifs with MEME." Proc Int Conf Intell Syst Mol Biol **3**: 21-9.

Barnes, S. R., D. A. Webb, et al. (1978). "The distribution of satellite and main-band DNA components in the *melanogaster* species subgroup of *Drosophila*. I. Fractionation of DNA in actinomycin D and distamycin A density gradients." Chromosoma **67**(4): 341-63.

Batzoglou, S. (2005). "The many faces of sequence alignment." Brief Bioinform **6**(1): 6-22.

Bazin, E., S. Glemin, et al. (2006). "Population size does not influence mitochondrial genetic diversity in animals." Science **312**(5773): 570-2.

Beaumont, M. A. and B. Rannala (2004). "The Bayesian revolution in genetics." Nat Rev Genet **5**(4): 251-61.

Beaumont, M. A., W. Zhang, et al. (2002). "Approximate Bayesian computation in population genetics." Genetics **162**(4): 2025-35.

Bejerano, G., A. C. Siepel, et al. (2005). "Computational screening of conserved genomic DNA in search of functional noncoding elements." Nat Methods **2**(7): 535-45.

Bergman, C. M., J. W. Carlson, et al. (2005). "*Drosophila* DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, *Drosophila melanogaster*." Bioinformatics **21**(8): 1747-9.

Bergman, C. M. and M. Kreitman (2001). "Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences." Genome Res **11**(8): 1335-45.

Bergman, C. M., B. D. Pfeiffer, et al. (2002). "Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome." Genome Biol **3**(12).

Berman, B. P., Y. Nibu, et al. (2002). "Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome." Proc Natl Acad Sci U S A **99**(2): 757-62.

Berman, B. P., B. D. Pfeiffer, et al. (2004). "Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila* pseudoobscura." Genome Biol **5**(9): R61.

Biemar, F., R. Zinzen, et al. (2005). "Spatial regulation of microRNA gene expression in the *Drosophila* embryo." Proc Natl Acad Sci U S A **102**(44): 15907-11.

Bird, C. P., B. E. Stranger, et al. (2006). "Functional variation and evolution of non-coding DNA." Curr Opin Genet Dev **16**(6): 559-64.

Birney, E., M. Clamp, et al. (2004). "GeneWise and Genomewise." Genome Res **14**(5): 988-95.

Birney, E. and R. Durbin (2000). "Using GeneWise in the *Drosophila* annotation experiment." Genome Res **10**(4): 547-8.

Blanchette, M., W. J. Kent, et al. (2004). "Aligning multiple genomic sequences with the threaded blockset aligner." Genome Res **14**(4): 708-15.

Blanchette, M. and M. Tompa (2002). "Discovery of regulatory elements by a computational method for phylogenetic footprinting." Genome Res **12**(5): 739-48.
Boffelli, D., J. McAuliffe, et al. (2003). "Phylogenetic shadowing of primate sequences to find functional regions of the human genome." Science **299**(5611): 1391-4.
Braverman, J. M., B. P. Lazzaro, et al. (2005). "DNA sequence polymorphism and divergence at the erect wing and suppressor of sable loci of *Drosophila melanogaster* and *D. simulans*." Genetics **170**(3): 1153-65.
Bray, N., I. Dubchak, et al. (2003). "AVID: A global alignment program." Genome Res **13**(1): 97-102.
Bray, N. and L. Pachter (2004). "MAVID: constrained ancestral alignment of multiple sequences." Genome Res **14**(4): 693-9.
Brem, R., G. Yvert, et al. (2002). "Genetic Dissection of Transcriptional Regulation in Budding Yeast." Science **296**(5568): 752-755.
Brenner, S. E., C. Chothia, et al. (1998). "Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships." Proc Natl Acad Sci U S A **95**(11): 6073-8.
Brinkmann, H., M. van der Giezen, et al. (2005). "An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics." Syst Biol **54**(5): 743-57.
Brudno, M., C. B. Do, et al. (2003). "LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA." Genome Res **13**(4): 721-31.
Burge, C., A. M. Campbell, et al. (1992). "Over- and under-representation of short oligonucleotides in DNA sequences." Proc Natl Acad Sci U S A **89**(4): 1358-62.
Bussemaker, H. J., H. Li, et al. (2001). "Regulatory element detection using correlation with expression." Nat Genet **27**(2): 167-171.
Caccone, A., G. D. Amato, et al. (1988). "Rates and patterns of scnDNA and mtDNA divergence within the *Drosophila melanogaster* subgroup." Genetics **118**(4): 671-83.
Carey, M. (1998). "The enhanceosome and transcriptional synergy." Cell **92**(1): 5-8.
Carey, M. and S. T. Smale (2001). Transcriptional Regulation in Eukaryotes: Concepts, Stategies, and Techniques, Cold Spring Harbor Laboratory Press.
Castillo-Davis, C. I. and D. L. Hartl (2002). "Genome evolution and developmental constraint in Caenorhabditis elegans." Mol Biol Evol **19**(5): 728-35.
Celniker, S. E., D. A. Wheeler, et al. (2002). "Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence." Genome Biol **3**(12): RESEARCH0079.

Charlesworth, D. (2006). "Balancing selection and its effects on sequences in nearby genome regions." PLoS Gen **2**(4): 379-384.
Chatterjee, S., Y. N. Zhou, et al. (1997). "Interaction of Gal repressor with inducer and operator: induction of gal transcription from repressor-bound DNA." Proc Natl Acad Sci U S A **94**(7): 2957-2962.
Chatterji, S. and L. Pachter (2006). "Reference based annotation with GeneMapper." Genome Biol **7**(4): R29.
Chen, F. C. and W. H. Li (2001). "Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees." Am J Hum Genet **68**(2): 444-56.
Chiang, D. Y., A. M. Moses, et al. (2003). "Phylogenetically and spatially conserved word pairs associated with gene-expression changes in yeasts." Genome Biol **4**(7): R43.
Chiaromonte, F., R. J. Weber, et al. (2003). "The share of human genomic DNA under selection estimated from human-mouse genomic alignments." Cold Spring Harb Symp Quant Biol **68**: 245-54.
Clark, A. G., M. B. Eisen, et al. (2007). "Evolution of genes and genomes on the *Drosophila* phylogeny." Nature **450**(7167): 203-18.
Coghlan, A., E. E. Eichler, et al. (2005). "Chromosome evolution in eukaryotes: a multi-kingdom perspective." Trends Genet **21**(12): 673-82.
Comeron, J. M. and M. Kreitman (2000). "The correlation between intron length and recombination in *Drosophila*. Dynamic equilibrium between mutational and selective forces." Genetics **156**(3): 1175-90.
Cooper, G. M., M. Brudno, et al. (2003). "Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes." Genome Res **13**(5): 813-20.
Costas, J., F. Casares, et al. (2003). "Turnover of binding sites for transcription factors involved in early *Drosophila* development." Gene **310**: 215-20.
Costas, J., P. S. Pereira, et al. (2004). "Dynamics and function of intron sequences of the wingless gene during the evolution of the *Drosophila* genus." Evol Dev **6**(5): 325-35.
Cuadrado, M., M. Sacristan, et al. (2001). "Species-specific organization of CpG island promoters at mammalian homologous genes." EMBO Rep **2**(7): 586-92.
Cunningham, C. W. (1997). "Can Three Incongruence Tests Predict When Data Should be Combined?" Mol Biol Evol **14**(7): 733-740.
Davidson, E. H. (2001). Genomic Regulatory Systems. San Diego, CA, Academic Press.
Degnan, J. H. and L. A. Salter (2005). "Gene tree distributions under the coalescent process." Evolution Int J Org Evolution **59**(1): 24-37.
Delcher, A. L., A. Phillippy, et al. (2002). "Fast algorithms for large-scale genome alignment and comparison." Nucleic Acids Res **30**(11): 2478-83.

Delsuc, F., M. J. Phillips, et al. (2003). "Comment on "Hexapod origins: monophyletic or paraphyletic?"" Science **301**(5639): 1482; author reply 1482.
Dermitzakis, E. T., C. M. Bergman, et al. (2003). "Tracing the evolutionary history of *Drosophila* regulatory regions with models that identify transcription factor binding sites." Mol Biol Evol **20**(5): 703-14.
Dermitzakis, E. T. and A. G. Clark (2002). "Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover." Mol Biol Evol **19**(7): 1114-21.
Doniger, S. W., J. Huh, et al. (2005). "Identification of functional transcription factor binding sites using closely related Saccharomyces species." Genome Res **15**(5): 701-9.
Down, T., C. Bergman, et al. (2007). "Large-Scale Discovery of Promoter Motifs in *Drosophila melanogaster*." PLoS Computational Biology **3**(1): e7.
Durbin, R. (1998). Biological sequence analysis: Probabilistic models of proteins and nucleic acids, Cambridge University Press.
Eddy, S. R. (2005). "A model of the statistical power of comparative genome sequence analysis." PLoS Biol **3**(1): e10.
Edwards, S. V. and P. Beerli (2000). "Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies." Evolution Int J Org Evolution **54**(6): 1839-54.
Eisses, K. (1979). "Genetic differntiation within *melanogaster* species group of the genus *Drosophila* (*Sophophora*)." Evolution **33**: 1063-1068.
Elnitski, L., R. C. Hardison, et al. (2003). "Distinguishing regulatory DNA from neutral sites." Genome Res **13**(1): 64-72.
Elsik, C. G., A. J. Mackey, et al. (2007). "Creating a honey bee consensus gene set." Genome Biol **8**(1): R13.
Emberly, E., N. Rajewsky, et al. (2003). "Conservation of regulatory elements between two species of *Drosophila*." BMC Bioinformatics **4**(1): 57.
Felsenstein, J. (1981). "Evolutionary trees from DNA sequences: a maximum likelihood approach." J Mol Evol **17**(6): 368-76.
Felsenstein, J. (1985). "Confidence limits on phylogenies: An approach using the bootstrap." Evolution **39**: 783-791.
Felsenstein, J. (1988). "Phylogenies from molecular sequences: inference and reliability." Annu Rev Genet **22**: 521-65.
Felsenstein, J. (1989). "PHYLIP - Phylogeny Inference Package (Version 3.2)." Cladistics **5**: 164-166.
Felsenstein, J. (2004). Inferring Phylogenies. Sunderland, MA, Sinauer Associates.
Felsenstein, J. (2006). "Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more Loci?" Mol Biol Evol **23**(3): 691-700.
Finn, R. D., J. Tate, et al. (2007). "The Pfam protein families database." Nucleic Acids Res.

Frazer, K. A., L. Elnitski, et al. (2003). "Cross-species sequence comparisons: a review of methods and available resources." Genome Res **13**(1): 1-12.
Gadagkar, S. R., M. S. Rosenberg, et al. (2005). "Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree." J Exp Zoolog B Mol Dev Evol **304**(1): 64-74.
Gailey, D. A., S. K. Ho, et al. (2000). "A phylogeny of the Drosophilidae using the sex-behaviour gene fruitless." Hereditas **133**(1): 81-3.
Gallo, S. M., L. Li, et al. (2006). "REDfly: a Regulatory Element Database for *Drosophila*." Bioinformatics **22**(3): 381-3.
Galtier, N. and M. Gouy (1995). "Inferring phylogenies from DNA sequences of unequal base compositions." Proc Natl Acad Sci U S A **92**(24): 11317-21.
Galtier, N. and M. Gouy (1998). "Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis." Mol Biol Evol **15**(7): 871-9.
Gatesy, J. and R. H. Baker (2005). "Hidden likelihood support in genomic data: can forty-five wrongs make a right?" Syst Biol **54**(3): 483-92.
Gatesy, J., M. Milinkovitch, et al. (1999). "Stability of cladistic relationships between Cetacea and higher-level artiodactyl taxa." Syst Biol **48**(1): 6-20.
Gertz, J., L. Riles, et al. (2005). "Discovery, validation, and genetic dissection of transcription factor binding sites by comparative and functional genomics." Genome Res **15**(8): 1145-52.
Gompel, N., B. Prud'homme, et al. (2005). "Chance caught on the wing: *cis*-regulatory evolution and the origin of pigment patterns in *Drosophila*." Nature **433**(7025): 481-487.
Grad, Y. H., F. P. Roth, et al. (2004). "Prediction of similarly acting *cis*-regulatory modules by subsequence profiling and comparative genomics in *Drosophila melanogaster* and D.pseudoobscura." Bioinformatics **20**(16): 2738-50.
Gray, S., H. Cai, et al. (1995). "Transcriptional repression in the *Drosophila* embryo." Philos Trans R Soc Lond B Biol Sci **349**(1329): 257-262.
Gu, X. and W. H. Li (1998). "Estimation of evolutionary distances under stationary and nonstationary models of nucleotide substitution." Proc Natl Acad Sci U S A **95**(11): 5899-905.
Halligan, D. L., A. Eyre-Walker, et al. (2004). "Patterns of evolutionary constraints in intronic and intergenic DNA of *Drosophila*." Genome Res **14**(2): 273-9.
Halligan, D. L. and P. D. Keightley (2006). "Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison." Genome Res **16**(7): 875-84.
Halpern, A. L. and W. J. Bruno (1998). "Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies." Mol Biol Evol **15**(7): 910-7.

Hamblin, M. T. and A. Di Rienzo (2000). "Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus." Am J Hum Genet **66**(5): 1669-1679.
Harbison, C. T., D. B. Gordon, et al. (2004). "Transcriptional regulatory code of a eukaryotic genome." Nature **431**(7004): 99-104.
Hardison, R. C. (2000). "Conserved noncoding sequences are reliable guides to regulatory elements." Trends Genet **16**(9): 369-72.
Hare, M. (2001). "Prospects for nuclear gene phylogeography." Trends Ecol Evol **16**: 700-706.
Hasegawa, M., H. Kishino, et al. (1985). "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA." J Mol Evol **22**(2): 160-74.
Hein, J., C. Wiuf, et al. (2000). "Statistical alignment: computational properties, homology testing and goodness-of-fit." J Mol Biol **302**(1): 265-79.
Hershberg, R., E. Yeger-Lotem, et al. (2005). "Chromosomal organization is shaped by the transcription regulatory network." Trends Genet **21**(3): 138-42.
Hertz, G. Z., G. W. Hartzell, 3rd, et al. (1990). "Identification of consensus patterns in unaligned DNA sequences known to be functionally related." Comput Appl Biosci **6**(2): 81-92.
Hey, J. and R. M. Kliman (2002). "Interactions between natural selection, recombination and gene density in the genes of *Drosophila*." Genetics **160**(2): 595-608.
Hey, J. and R. Nielsen (2004). "Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila* pseudoobscura and D. persimilis." Genetics **167**(2): 747-60.
Hillis, D. M. and J. J. Bull (1993). "An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis." Syst Biol **42**: 182-192.
Hillis, D. M., J. P. Huelsenbeck, et al. (1994). "Application and accuracy of molecular phylogenies." Science **264**(5159): 671-7.
Hittinger, C. and S. Carroll (2007). "Gene duplication and the adaptive evolution of a classic genetic switch." Nature **449**(7163): 677-681.
Hoekstra, H. E., R. J. Hirschmann, et al. (2006). "A single amino acid mutation contributes to adaptive beach mouse color pattern." Science **313**(5783): 101-4.
Holland, B. R., K. T. Huber, et al. (2004). "Using consensus networks to visualize contradictory evidence for species phylogeny." Mol Biol Evol **21**(7): 1459-61.
Holland, B. R., L. S. Jermiin, et al. (2005). "Improved Consensus Network Techniques for Genome-Scale Phylogeny." Mol Biol Evol.
Holmes, I. and R. Durbin (1998). "Dynamic programming alignment accuracy." J Comput Biol **5**(3): 493-504.

Huang, W., D. M. Umbach, et al. (2006). "Accurate anchoring alignment of divergent sequences." Bioinformatics **22**(1): 29-34.
Hudson, R. R. (1992). "Gene trees, species trees and the segregation of ancestral alleles." Genetics **131**(2): 509-13.
Huelsenbeck, J. P. (1995). "The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining." Mol Biol Evol **12**(5): 843-9.
Husmeier, D. (2005). "Discriminating between rate heterogeneity and interspecific recombination in DNA sequence alignments with phylogenetic factorial hidden Markov models." Bioinformatics **21 Suppl 2**: ii166-ii172.
Ihaka, R., & Gentleman, R. (1996). "R: a language for data analysis and graphics." Journal of Computational and Graphical Statistics **5**: 299-314.
Iwamoto, S., J. Li, et al. (1996). "Characterization of the Duffy gene promoter: evidence for tissue-specific abolishment of expression in Fy(a-b-) of black individuals." Biochem Biophys Res Commun **222**(3): 852-859.
Janssens, H., S. Hou, et al. (2006). "Quantitative and predictive model of transcriptional control of the *Drosophila melanogaster* even skipped gene." Nat Genet.
Jareborg, N., E. Birney, et al. (1999). "Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs." Genome Res **9**(9): 815-24.
Jeffroy, O., H. Brinkmann, et al. (2006). "Phylogenomics: the beginning of incongruence?" Trends Genet **22**(4): 225-31.
Jeffs, P. S., E. C. Holmes, et al. (1994). "The molecular evolution of the alcohol dehydrogenase and alcohol dehydrogenase-related genes in the *Drosophila melanogaster* species subgroup." Mol Biol Evol **11**(2): 287-304.
Jeong, S., A. Rokas, et al. (2006). "Regulation of Body Pigmentation by the Abdominal-B Hox Protein and Its Gain and Loss in *Drosophila* Evolution." Cell **125**(7): 1387-1399.
Jermiin, L. S., L. Poladian, et al. (2005). "Evolution. Is the "Big Bang" in animal evolution real?" Science **310**(5756): 1910-1.
Johnson, A. N., C. M. Bergman, et al. (2003). "Embryonic enhancers in the dpp disk region regulate a second round of Dpp signaling from the dorsal ectoderm to the mesoderm that represses Zfh-1 expression in a subset of pericardial cells." Dev Biol **262**(1): 137-51.
Johnson, D. S., Q. Zhou, et al. (2005). "De novo discovery of a tissue-specific gene regulatory module in a chordate." Genome Res **15**(10): 1315-24.
Kaminker, J. S., C. M. Bergman, et al. (2002). "The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective." Genome Biol **3**(12): RESEARCH0084.

Katoh, K., K. Misawa, et al. (2002). "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform." Nucleic Acids Res **30**(14): 3059-66.
Katz, R. W. (1981). "On Some Criteria for Estimating the Order of a Markov Chain." Technometrics **23**(3): 243-249.
Keceioglou, J. and D. Starrett (2004). "Aligning Alignments Exactly." RECOMB: 85-96.
Keightley, P. D. and D. J. Gaffney (2003). "Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents." Proc Natl Acad Sci U S A **100**(23): 13402-6.
Keightley, P. D. and T. Johnson (2004). "MCALIGN: stochastic alignment of noncoding DNA sequences based on an evolutionary model of sequence evolution." Genome Res **14**(3): 442-50.
Keightley, P. D., G. V. Kryukov, et al. (2005). "Evolutionary constraints in conserved nongenic sequences of mammals." Genome Res **15**(10): 1373-8.
Kent, W. J. (2002). "BLAT--the BLAST-like alignment tool." Genome Res **12**(4): 656-64.
Kent, W. J. and A. M. Zahler (2000). "Conservation, regulation, synteny, and introns in a large-scale C. briggsae-C. elegans genomic alignment." Genome Res **10**(8): 1115-25.
Kimura, M. and N. Takahata (1983). "Selective constraint in protein polymorphism: study of the effectively neutral mutation model by using an improved pseudosampling method." Proc Natl Acad Sci U S A **80**(4): 1048-52.
King, D. C., J. Taylor, et al. (2005). "Evaluation of regulatory potential and conservation scores for detecting *cis*-regulatory modules in aligned mammalian genome sequences." Genome Res **15**(8): 1051-60.
Kingman, J. F. C. (1982). "The coalescent." Stoch Proc Appl **13**: 235-248.
Kirkness, E. F., V. Bafna, et al. (2003). "The dog genome: survey sequencing and comparative analysis." Science **301**(5641): 1898-903.
Kishino, H., J. L. Thorne, et al. (2001). "Performance of a divergence time estimation method under a probabilistic model of rate evolution." Mol Biol Evol **18**(3): 352-61.
Kliman, R. M., P. Andolfatto, et al. (2000). "The population genetics of the origin and divergence of the *Drosophila* simulans complex species." Genetics **156**(4): 1913-31.
Knowles, L. L. and W. P. Maddison (2002). "Statistical phylogeography." Mol Ecol **11**(12): 2623-35.
Ko, W. Y., R. M. David, et al. (2003). "Molecular phylogeny of the *Drosophila melanogaster* species subgroup." J Mol Evol **57**(5): 562-73.
Kolbe, D., J. Taylor, et al. (2004). "Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat." Genome Res **14**(4): 700-7.

Kopp, A. and J. R. True (2002). "Phylogeny of the Oriental *Drosophila melanogaster* species group: a multilocus reconstruction." Syst Biol **51**(5): 786-805.
Langley, C. H., B. P. Lazzaro, et al. (2000). "Linkage disequilibria and the site frequency spectra in the su(s) and su(w(a)) regions of the *Drosophila melanogaster* X chromosome." Genetics **156**(4): 1837-52.
Lassmann, T. and E. L. Sonnhammer (2002). "Quality assessment of multiple alignment programs." FEBS Lett **529**(1): 126-30.
Lemeunier, F. and M. A. Ashburner (1976). "Relationships within the *melanogaster* species subgroup of the genus *Drosophila* (Sophophora). II. Phylogenetic relationships between six species based upon polytene chromosome banding sequences." Proc R Soc Lond B Biol Sci **193**(1112): 275-94.
Lewis, R. L., A. T. Beckenbach, et al. (2005). "The phylogeny of the subgroups within the *melanogaster* species group: likelihood tests on COI and COII sequences and a Bayesian estimate of phylogeny." Mol Phylogenet Evol **37**(1): 15-24.
Li, L., Q. Zhu, et al. (2007). "Large-scale analysis of transcriptional *cis*-regulatory modules reveals both common features and distinct subclasses." Genome Biology **8**: R101.
Li, W.-H. (1997). Molecular Evolution. Massachusetts, Sinauer Associates.
Li, Y. J., Y. Satta, et al. (1999). "Paleo-demography of the *Drosophila melanogaster* subgroup: application of the maximum likelihood method." Genes Genet Syst **74**(4): 117-27.
Lia, X., S. MacArthura, et al. (In Press). "Transcription Factors Bind Thousands of Active and Inactive Regions in the *Drosophila* Blastoderm." PLoS Biol.
Lieb, J. D., X. Liu, et al. (2001). "Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association." Nat Genet **28**(4): 327-34.
Lin, J. and M. Nei (1991). "Relative efficiencies of the maximum-parsimony and distance-matrix methods of phylogeny construction for restriction data." Mol Biol Evol **8**(3): 356-65.
Loots, G. G., I. Ovcharenko, et al. (2002). "rVista for comparative sequence-based discovery of functional transcription factor binding sites." Genome Res **12**(5): 832-9.
Ludwig, M. Z. (2002). "Functional evolution of noncoding DNA." Curr Opin Genet Dev **12**(6): 634-9.
Ludwig, M. Z., C. Bergman, et al. (2000). "Evidence for stabilizing selection in a eukaryotic enhancer element." Nature **403**(6769): 564-7.
Ludwig, M. Z., A. Palsson, et al. (2005). "Functional evolution of a *cis*-regulatory module." PLoS Biol **3**(4): e93.
Luengo Hendriks, C. L., S. V. Keranen, et al. (2006). "Three-dimensional morphology and gene expression in the *Drosophila* blastoderm at cellular resolution I: data acquisition pipeline." Genome Biol **7**(12): R123.

MacArthur, S. and J. F. Brookfield (2004). "Expected Rates and Modes of Evolution of Enhancer Sequences." Mol Biol Evol.
Maddison, W. P. (1997). "Gene Trees in Species Trees." Syst. Biol. **46**(3): 523-536.
Maddison, W. P. and L. L. Knowles (2006). "Inferring phylogeny despite incomplete lineage sorting." Syst Biol **55**(1): 21-30.
Maerkl, S. and S. Quake (2007). "A Systems Approach to Measuring the Binding Energy Landscapes of Transcription Factors." Science **315**(5809): 233-237.
Makeev, V. J., A. P. Lifanov, et al. (2003). "Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information." Nucleic Acids Res **31**(20): 6016-26.
Markstein, M. and M. Levine (2002). "Decoding *cis*-regulatory DNAs in the *Drosophila* genome." Curr Opin Genet Dev **12**(5): 601-6.
Markstein, M., P. Markstein, et al. (2002). "Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo." Proc Natl Acad Sci U S A **99**(2): 763-8.
Matsuo, Y. (2000). "Molecular evolution of the histone 3 multigene family in the *Drosophila melanogaster* species subgroup." Mol Phylogenet Evol **16**(3): 339-43.
McClure, M. A., T. K. Vasi, et al. (1994). "Comparative analysis of multiple protein-sequence alignment methods." Mol Biol Evol **11**(4): 571-92.
McCue, L. A., W. Thompson, et al. (2002). "Factors influencing the identification of transcription factor binding sites by cross-species comparison." Genome Res **12**(10): 1523-32.
Merika, M. and D. Thanos (2001). "Enhanceosomes." Curr Opin Genet Dev **11**(2): 205-8.
Metzler, D. (2003). "Statistical alignment based on fragment insertion and deletion models." Bioinformatics **19**(4): 490-9.
Miller, W. (2001). "Comparison of genomic DNA sequences: solved and unsolved problems." Bioinformatics **17**(5): 391-7.
Miller, W., K. D. Makova, et al. (2004). "Comparative genomics." Annu Rev Genomics Hum Genet **5**: 15-56.
Mirny, L. A. and M. S. Gelfand (2002). "Structural analysis of conserved base pairs in protein-DNA complexes." Nucleic Acids Res **30**(7): 1704-11.
Misra, S., M. A. Crosby, et al. (2002). "Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review." Genome Biol **3**(12): RESEARCH0083.
Morgenstern, B. (1999). "DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment." Bioinformatics **15**(3): 211-8.
Morgenstern, B., K. Frech, et al. (1998). "DIALIGN: finding local similarities by multiple sequence alignment." Bioinformatics **14**(3): 290-4.

Moriyama, E. N. and T. Gojobori (1992). "Rates of synonymous substitution and base composition of nuclear genes in *Drosophila*." Genetics **130**(4): 855-64.
Moriyama, E. N. and D. L. Hartl (1993). "Codon usage bias and base composition of nuclear genes in *Drosophila*." Genetics **134**(3): 847-58.
Moriyama, E. N. and J. R. Powell (1996). "Intraspecific nuclear DNA variation in *Drosophila*." Mol Biol Evol **13**(1): 261-77.
Moses, A. M., D. Y. Chiang, et al. (2003). "Position specific variation in the rate of evolution in transcription factor binding sites." BMC Evol Biol **3**(1): 19.
Moses, A. M., D. Y. Chiang, et al. (2004). "MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model." Genome Biol **5**(12): R98.
Moses, A. M., D. A. Pollard, et al. (2006). "Large-scale turnover of functional transcription factor binding sites in *Drosophila*." PLoS Comput Biol **2**(10): e130.
Mossel, E. and E. Vigoda (2005). "Phylogenetic MCMC algorithms are misleading on mixtures of trees." Science **309**(5744): 2207-9.
Mungall, C. J., S. Misra, et al. (2002). "An integrated computational pipeline and database to support whole-genome sequence annotation." Genome Biol **3**(12): RESEARCH0081.
Needleman, S. B. and C. D. Wunsch (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins." J Mol Biol **48**(3): 443-53.
Negre, B., S. Casillas, et al. (2005). "Conservation of regulatory sequences and gene expression patterns in the disintegrating *Drosophila* Hox gene complex." Genome Res **15**(5): 692-700.
Nei, M. (1986). "Stochastic errors in DNA evolution and molecular phylogeny." Prog Clin Biol Res **218**: 133-47.
Nekrutenko, A., K. D. Makova, et al. (2002). "The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study." Genome Res **12**(1): 198-202.
Nielsen, R. (1998). "Maximum likelihood estimation of population divergence times and population phylogenies under the infinite sites model." Theor Popul Biol **53**(2): 143-51.
Nielsen, R. and J. Wakeley (2001). "Distinguishing migration from isolation: a Markov chain Monte Carlo approach." Genetics **158**(2): 885-96.
Nigro, L., M. Solignac, et al. (1991). "Mitochondrial DNA sequence divergence in the *Melanogaster* and oriental species subgroups of *Drosophila*." J Mol Evol **33**(2): 156-62.
Nobrega, M. A., I. Ovcharenko, et al. (2003). "Scanning human gene deserts for long-range enhancers." Science **302**(5644): 413.
Notredame, C., D. G. Higgins, et al. (2000). "T-Coffee: A novel method for fast and accurate multiple sequence alignment." J Mol Biol **302**(1): 205-17.

Nusslein-Volhard, C. and E. Wieschaus (1980). "Mutations affecting segment number and polarity in *Drosophila*." Nature **287**(5785): 795-801.
O'Grady, P. M. and M. G. Kidwell (2002). "Phylogeny of the subgenus sophophora (Diptera: drosophilidae) based on combined analysis of nuclear and mitochondrial sequences." Mol Phylogenet Evol **22**(3): 442-53.
Ogurtsov, A. Y., M. A. Roytberg, et al. (2002). "OWEN: aligning long collinear regions of genomes." Bioinformatics **18**(12): 1703-4.
Ogurtsov, A. Y., S. Sunyaev, et al. (2004). "Indel-based evolutionary distance and mouse-human divergence." Genome Res **14**(8): 1610-6.
Ohta, T. and M. Kimura (1971). "Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population." Genetics **68**: 571-580.
Osada, N. and C. I. Wu (2005). "Inferring the mode of speciation from genomic data: a study of the great apes." Genetics **169**(1): 259-64.
Pamilo, P. and M. Nei (1988). "Relationships between gene trees and species trees." Mol Biol Evol **5**(5): 568-83.
Papatsenko, D. A., V. J. Makeev, et al. (2002). "Extraction of functional binding sites from unique regulatory regions: the *Drosophila* early developmental enhancers." Genome Res **12**(3): 470-81.
Parsch, J. (2003). "Selective constraints on intron evolution in *Drosophila*." Genetics **165**(4): 1843-51.
Pearson, W. R. and D. J. Lipman (1988). "Improved tools for biological sequence comparison." Proc Natl Acad Sci U S A **85**(8): 2444-8.
Pedersen, J. S., G. Bejerano, et al. (2006). "Identification and classification of conserved RNA secondary structures in the human genome." PLoS Comput Biol **2**(4): e33.
Petrov, D. A. and D. L. Hartl (1997). "Trash DNA is what gets thrown away: high rate of DNA loss in *Drosophila*." Gene **205**(1-2): 279-289.
Petrov, D. A. and D. L. Hartl (1998). "High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila* virilis species groups." Mol Biol Evol **15**(3): 293-302.
Petrov, D. A., E. R. Lozovskaya, et al. (1996). "High intrinsic rate of DNA loss in *Drosophila*." Nature **384**(6607): 346-9.
Phillips, M. J., F. Delsuc, et al. (2004). "Genome-scale phylogeny and the detection of systematic biases." Mol Biol Evol **21**(7): 1455-8.
Phillips, M. J. and D. Penny (2003). "The root of the mammalian tree inferred from whole mitochondrial genomes." Mol Phylogenet Evol **28**(2): 171-85.
Pollard, D. A., C. M. Bergman, et al. (2004). "Benchmarking tools for the alignment of functional noncoding DNA." BMC Bioinformatics **5**(1): 6.
Pollock, D. D., D. J. Zwickl, et al. (2002). "Increased taxon sampling is advantageous for phylogenetic inference." Syst Biol **51**(4): 664-71.
Powell, J. R. (1997). Progress and prospects in evolutionary biology: The *Drosophila* model, Oxford University Press.

Prabhakar, S., F. Poulin, et al. (2006). "Close sequence comparisons are sufficient to identify human *cis*-regulatory elements." Genome Res.
Pribnow, D. (1975). "Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter." Proc Natl Acad Sci U S A **72**(3): 784-788.
Ptak, S. E. and D. A. Petrov (2002). "How intron splicing affects the deletion and insertion profile in *Drosophila melanogaster*." Genetics **162**(3): 1233-44.
Rannala, B. and Z. Yang (2003). "Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci." Genetics **164**(4): 1645-56.
Reeves, G. T., C. B. Muratov, et al. (2006). "Quantitative models of developmental pattern formation." Dev Cell **11**(3): 289-300.
Remsen, J. and P. O'Grady (2002). "Phylogeny of Drosophilinae (Diptera: Drosophilidae), with comments on combined analysis and character support." Mol Phylogenet Evol **24**(2): 249-64.
Ren, F., H. Tanaka, et al. (2005). "An empirical examination of the utility of codon-substitution models in phylogeny reconstruction." Syst Biol **54**(5): 808-18.
Rice, P., I. Longden, et al. (2000). "EMBOSS: the European Molecular Biology Open Software Suite." Trends Genet **16**(6): 276-7.
Richards, S., Y. Liu, et al. (2005). "Comparative genome sequencing of *Drosophila* pseudoobscura: chromosomal, gene, and *cis*-element evolution." Genome Res **15**(1): 1-18.
Rivas, E. and S. R. Eddy (2001). "Noncoding RNA gene detection using comparative sequence analysis." BMC Bioinformatics **2**: 8.
Rockman, M. V., M. W. Hahn, et al. (2005). "Ancient and Recent Positive Selection Transformed Opioid *cis*-Regulation in Humans." PLoS Biol **3**(12).
Rokas, A. and S. B. Carroll (2005). "More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy." Mol Biol Evol **22**(5): 1337-44.
Rokas, A. and P. W. Holland (2000). "Rare genomic changes as a tool for phylogenetics." Trends in Ecology and Evolution **15**(11): 454-459.
Rokas, A., D. Kruger, et al. (2005). "Animal evolution and the molecular signature of radiations compressed in time." Science **310**(5756): 1933-8.
Rokas, A., B. L. Williams, et al. (2003). "Genome-scale approaches to resolving incongruence in molecular phylogenies." Nature **425**(6960): 798-804.
Rosenberg, M. S. (2005). "Evolutionary distance estimation and fidelity of pair wise sequence alignment." BMC Bioinformatics **6**(1): 102.
Rosenberg, M. S. (2005). "Multiple sequence alignment accuracy and evolutionary distance estimation." BMC Bioinformatics **6**(1): 278.
Rosenberg, N. A. (2002). "The probability of topological concordance of gene trees and species trees." Theor Popul Biol **61**(2): 225-47.

Rosenberg, N. A. (2003). "The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model." Evolution Int J Org Evolution **57**(7): 1465-77.
Rosenberg, N. A. and M. Nordborg (2002). "Genealogical trees, coalescent theory and the analysis of genetic polymorphisms." Nat Rev Genet **3**(5): 380-90.
Rost, B. (1999). "Twilight zone of protein sequence alignments." Protein Eng **12**(2): 85-94.
Russo, C. A., N. Takezaki, et al. (1995). "Molecular phylogeny and divergence times of drosophilid species." Mol Biol Evol **12**(3): 391-404.
Ruvkun, G., B. Wightman, et al. (1991). "Dominant gain-of-function mutations that lead to misregulation of the C. elegans heterochronic gene lin-14, and the evolutionary implications of dominant mutations in pattern-formation genes." Dev Suppl **1**: 47-54.
Sackton, T. B., B. P. Lazzaro, et al. (2007). "Dynamic evolution of the innate immune system in *Drosophila*." Nat Genet **39**(12): 1461-8.
Sanderson, M. J. and H. B. Shafer (2002). "Troubleshooting molecular phylogenetic analyses." Annu Rev Ecol Syst **33**: 49-72.
Sarich, V. M. and A. C. Wilson (1973). "Generation time and genomic evolution in primates." Science **179**(78): 1144-7.
Sauder, J. M., J. W. Arthur, et al. (2000). "Large-scale comparison of protein sequence alignment algorithms with structure alignments." Proteins **40**(1): 6-22.
Sawyer, S. A. and D. L. Hartl (1992). "Population genetics of polymorphism and divergence." Genetics **132**(4): 1161-76.
Scannell, D. R., K. P. Byrne, et al. (2006). "Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts." Nature **440**(7082): 341-5.
Schlotterer, C., M. T. Hauser, et al. (1994). "Comparative evolutionary analysis of rDNA ITS regions in *Drosophila*." Mol Biol Evol **11**(3): 513-22.
Schmid, C. D. and P. Bucher (2007). "ChIP-Seq Data Reveal Nucleosome Architecture of Human Promoters." Cell **131**(5): 831-2.
Schneider, T. D., G. D. Stormo, et al. (1986). "Information content of binding sites on nucleotide sequences." J Mol Biol **188**(3): 415-31.
Schroeder, M., M. Pearce, et al. (2004). "Transcriptional Control in the Segmentation Gene Network of *Drosophila*." PLoS Biology **2**(9): e271.
Schroeder, M. D., M. Pearce, et al. (2004). "Transcriptional control in the segmentation gene network of *Drosophila*." PLoS Biol **2**(9): E271.
Schwartz, S., W. J. Kent, et al. (2003). "Human-mouse alignments with BLASTZ." Genome Res **13**(1): 103-7.
Schwartz, S., Z. Zhang, et al. (2000). "PipMaker--a web server for aligning two genomic DNA sequences." Genome Res **10**(4): 577-86.
Shabalina, S. A. and A. S. Kondrashov (1999). "Pattern of selective constraint in C. elegans and C. briggsae genomes." Genet Res **74**(1): 23-30.

Sharp, P. M. and W. H. Li (1989). "On the rate of DNA sequence evolution in *Drosophila*." J Mol Evol **28**(5): 398-402.
Shibata, H. and T. Yamazaki (1995). "Molecular evolution of the duplicated Amy locus in the *Drosophila melanogaster* species subgroup: concerted evolution only in the coding region and an excess of nonsynonymous substitutions in speciation." Genetics **141**(1): 223-36.
Siepel, A. and D. Haussler (2004). "Phylogenetic estimation of context-dependent substitution rates by maximum likelihood." Mol Biol Evol **21**(3): 468-88.
Singh, N. D. and D. A. Petrov (2004). "Rapid sequence turnover at an intergenic locus in *Drosophila*." Mol Biol Evol **21**(4): 670-680.
Singh, R. S. (1989). "Population genetics and evolution of species related to *Drosophila melanogaster*." Annu Rev Genet **23**: 425-53.
Sinha, S., M. D. Schroeder, et al. (2004). "Cross-species comparison significantly improves genome-wide prediction of *cis*-regulatory modules in *Drosophila*." BMC Bioinformatics **5**: 129.
Sinha, S. and E. D. Siggia (2005). "Sequence turnover and tandem repeats in *cis*-regulatory modules in *Drosophila*." Mol Biol Evol **22**(4): 874-85.
Slater, G. S. and E. Birney (2005). "Automated generation of heuristics for biological sequence comparison." BMC Bioinformatics **6**: 31.
Slatkin, M. and J. L. Pollack (2006). "The concordance of gene trees and species trees at two linked loci." Genetics **172**(3): 1979-84.
Small, S., R. Kraut, et al. (1991). "Transcriptional regulation of a pair-rule stripe in *Drosophila*." Genes Dev **5**(5): 827-39.
Solignac, M., M. Monnerot, et al. (1986). "Mitochondrial DNA evolution in the *melanogaster* species subgroup of *Drosophila*." J Mol Evol **23**(1): 31-40.
Soltis, P. S. and D. E. Soltis (2003). "Applying the bootstrap in phylogeny reconstruction." Stat Sci **18**: 256-267.
Stanojevic, D., S. Small, et al. (1991). "Regulation of a segmentation stripe by overlapping activators and repressors in the *Drosophila* embryo." Science **254**(5036): 1385-7.
Stein, L. D., Z. Bao, et al. (2003). "The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics." PLoS Biol **1**(2): E45.
Stone, E. A., G. M. Cooper, et al. (2005). "Trade-offs in detecting evolutionarily constrained sequence by comparative genomics." Annu Rev Genomics Hum Genet **6**: 143-64.
Stoye, J. (1998). "Multiple sequence alignment with the Divide-and-Conquer method." Gene **211**(2): GC45-56.
Stoye, J., D. Evers, et al. (1997). "Generating benchmarks for multiple sequence alignments and phylogenetic reconstructions." Proc Int Conf Intell Syst Mol Biol **5**: 303-6.
Stoye, J., D. Evers, et al. (1998). "Rose: generating sequence families." Bioinformatics **14**(2): 157-63.
Sullivan, J. and D. L. Swofford (2001). "Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site

rate variation and nucleotide substitution pattern are violated?" Syst Biol **50**(5): 723-9.
Tajima, F. (1983). "Evolutionary relationship of DNA sequences in finite populations." Genetics **105**(2): 437-60.
Takahata, N. (1989). "Gene genealogy in three related populations: consistency probability between gene and population trees." Genetics **122**(4): 957-66.
Takano-Shimizu, T. (2001). "Local changes in GC/AT substitution biases and in crossover frequencies on *Drosophila* chromosomes." Mol Biol Evol **18**(4): 606-19.
Tateno, Y., N. Takezaki, et al. (1994). "Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site." Mol Biol Evol **11**(2): 261-77.
Tatusova, T. A. and T. L. Madden (1999). "BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences." FEMS Microbiol Lett **174**(2): 247-50.
Taylor, D. J. and W. H. Piel (2004). "An assessment of accuracy, error, and conflict with support values from genome-scale phylogenetic data." Mol Biol Evol **21**(8): 1534-7.
Thomas, J. W., J. W. Touchman, et al. (2003). "Comparative analyses of multi-species sequences from targeted genomic regions." Nature **424**(6950): 788-93.
Thompson, J. D., D. G. Higgins, et al. (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic Acids Res **22**(22): 4673-80.
Thompson, J. D., F. Plewniak, et al. (1999). "A comprehensive comparison of multiple sequence alignment programs." Nucleic Acids Res **27**(13): 2682-90.
Thorne, J. L., H. Kishino, et al. (1991). "An evolutionary model for maximum likelihood alignment of DNA sequences." J Mol Evol **33**(2): 114-24.
Thorne, J. L., H. Kishino, et al. (1992). "Inching toward reality: an improved likelihood model of sequence evolution." J Mol Evol **34**(1): 3-16.
Thornton, K. and P. Andolfatto (2006). "Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*." Genetics **172**(3): 1607-19.
Tomancak, P., B. Berman, et al. (2007). "Global analysis of patterns of gene expression during *Drosophila* embryogenesis." Genome Biology **8**: R145.
Tompa, M., N. Li, et al. (2005). "Assessing computational tools for the discovery of transcription factor binding sites." Nat Biotechnol **23**(1): 137-44.
Tournamille, C., Y. Colin, et al. (1995). "Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals." Nat Genet **10**(2): 224-228.

True, J. R., J. M. Mercer, et al. (1996). "Differences in crossover frequency and distribution among three sibling species of *Drosophila*." Genetics **142**(2): 507-23.
Wagner, G. n. P., W. Otto, et al. (2007). "A stochastic model for the evolution of transcription factor binding site abundance." J Theor Biol.
Wagner, G. P., C. Fried, et al. (2004). "Divergence of conserved non-coding sequences: rate estimates and relative rate tests." Mol Biol Evol **21**(11): 2116-21.
Wall, D. P., H. B. Fraser, et al. (2003). "Detecting putative orthologs." Bioinformatics **19**(13): 1710-1.
Wall, J. D. (2003). "Estimating ancestral population sizes and divergence times." Genetics **163**(1): 395-404.
Wang, T. and G. D. Stormo (2003). "Combining phylogenetic data with co-regulated genes to identify regulatory motifs." Bioinformatics **19**(18): 2369-80.
Wang, T. and G. D. Stormo (2005). "Identifying the conserved network of *cis*-regulatory sites of a eukaryotic genome." Proc Natl Acad Sci U S A **102**(48): 17400-5.
Wang, W., K. Thornton, et al. (2004). "Nucleotide variation and recombination along the fourth chromosome in *Drosophila* simulans." Genetics **166**(4): 1783-94.
Wasserman, W. W., M. Palumbo, et al. (2000). "Human-mouse genome comparisons to locate regulatory sites." Nat Genet **26**(2): 225-8.
Waterston, R. H., K. Lindblad-Toh, et al. (2002). "Initial sequencing and comparative analysis of the mouse genome." Nature **420**(6915): 520-62.
Weir, B. S. (1996). Genetic Data Analysis II, Sinauer.
Wheeler, D. L., T. Barrett, et al. (2005). "Database resources of the National Center for Biotechnology Information." Nucleic Acids Res **33**(Database issue): D39-45.
Wilcox, T. P., F. J. Garcia de Leon, et al. (2004). "Convergence among cave catfishes: long-branch attraction and a Bayesian relative rates test." Mol Phylogenet Evol **31**(3): 1101-13.
Wittkopp, P., B. Haerum, et al. (2004). "Evolutionary changes in cis and trans gene regulation." Nature **430**(6995): 85-88.
Wiuf, C., K. Zhao, et al. (2004). "The probability and chromosomal extent of trans-specific polymorphism." Genetics **168**(4): 2363-72.
Wray, G. A. (2007). "The evolutionary significance of *cis*-regulatory mutations." Nat Rev Genet **8**(3): 206-16.
Wray, G. A., M. W. Hahn, et al. (2003). "The evolution of transcriptional regulation in eukaryotes." Mol Biol Evol **20**(9): 1377-419.
Wu, C. I. (1991). "Inferences of species phylogeny in relation to segregation of ancient polymorphisms." Genetics **127**(2): 429-35.
Xing, E. P., W. Wu, et al. (2004). "Logos: a modular bayesian model for de novo motif detection." J Bioinform Comput Biol **2**(1): 127-54.

Yang, Z. (1994). "Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods." J Mol Evol **39**(3): 306-14.
Yang, Z. (1997). "How often do wrong models produce better phylogenies?" Mol Biol Evol **14**(1): 105-8.
Yang, Z. (1997). "PAML: a program package for phylogenetic analysis by maximum likelihood." Comput Appl Biosci **13**(5): 555-6.
Yang, Z. and R. Nielsen (2000). "Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models." Mol Biol Evol **17**(1): 32-43.
Zapata, C. and G. Alvarez (1993). "On the detection of nonrandom associations between DNA polymorphisms in natural populations of *Drosophila*." Mol Biol Evol **10**(4): 823-41.
Zeng, L. W., J. M. Comeron, et al. (1998). "The molecular clock revisited: the rate of synonymous vs. replacement change in *Drosophila*." Genetica **102-103**(1-6): 369-82.
Zharkikh, A. (1994). "Estimation of evolutionary distances between nucleotide sequences." J Mol Evol **39**(3): 315-29.
Zhu, J., J. S. Liu, et al. (1998). "Bayesian adaptive sequence alignment algorithms." Bioinformatics **14**(1): 25-39.
Zinzen, R. P. and D. Papatsenko (2007). "Enhancer responses to similarly distributed antagonistic gradients in development." PLoS Comput Biol **3**(5).
Zwickl, D. J. and D. M. Hillis (2002). "Increased taxon sampling greatly reduces phylogenetic error." Syst Biol **51**(4): 588-98.

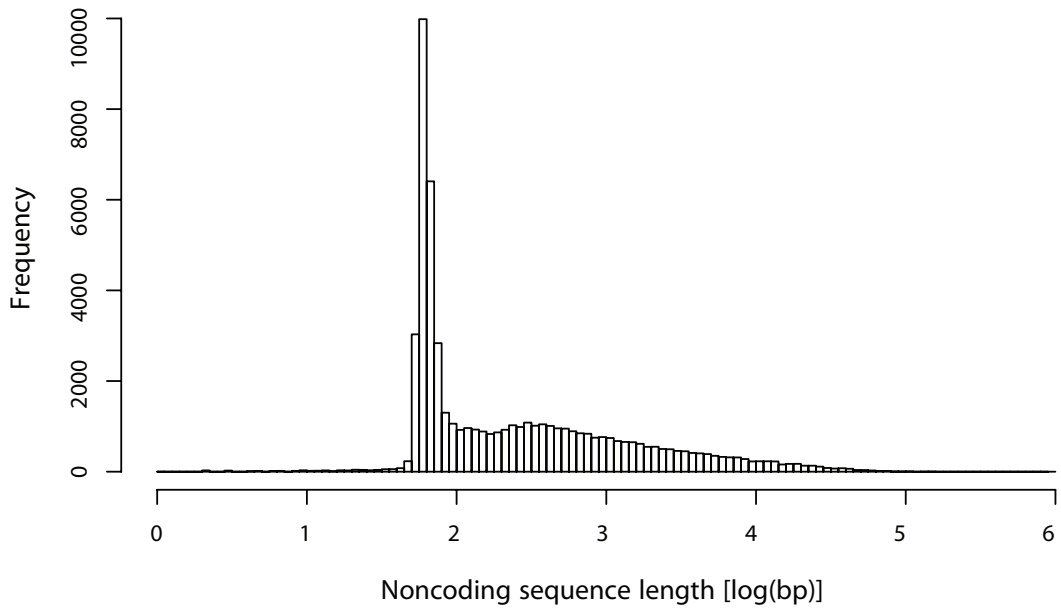

Figure 2.1
Distribution of non-coding sequence lengths in the D. melanogaster Release 3 genome sequence. Sequences between coding exons were extracted from the D. melanogaster Release 3 euchromatic genome sequence and annotations, and transposable element sequences were subsequently subtracted to produce the "pre-integration" distribution of non-coding sequence lengths.

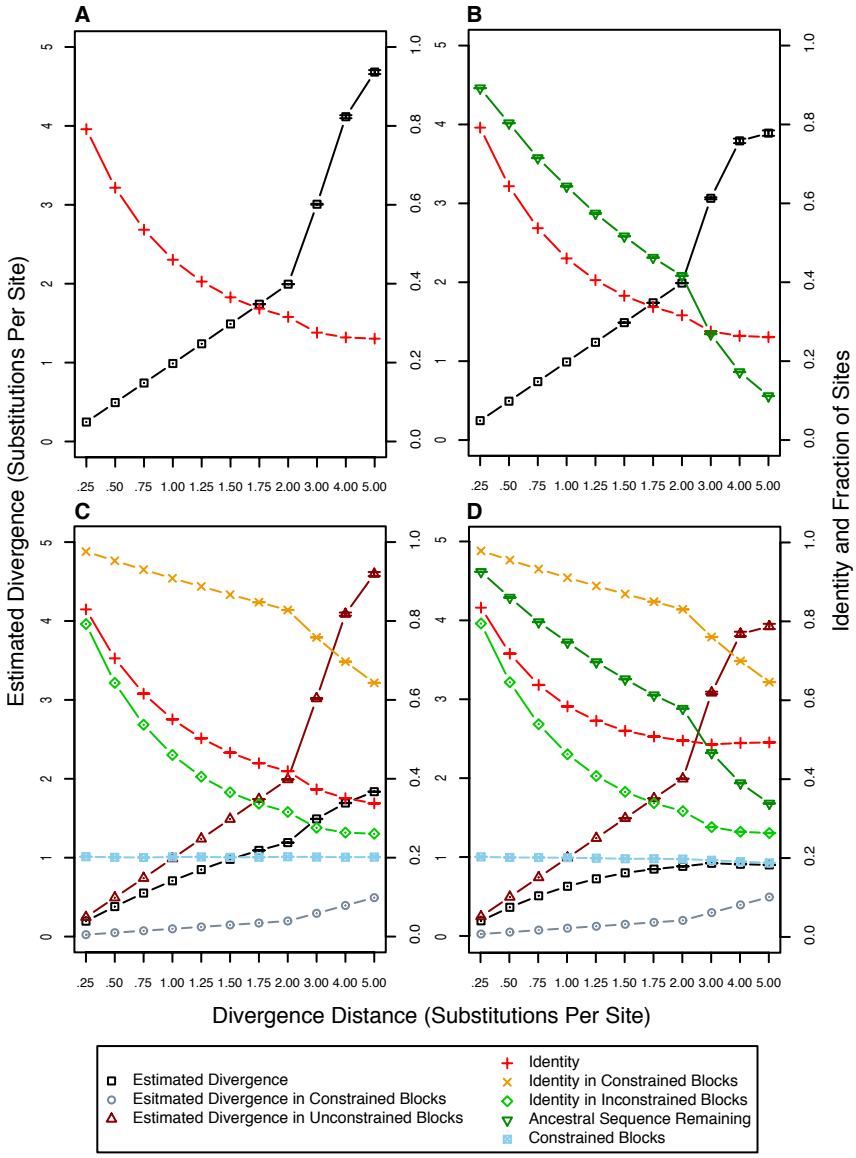

Figure 2.2

Simulation statistics. Pairwise alignments were simulated for a range of divergence distances, using a modified version of the ROSE simulation platform under four different regimes: A) w/o constrained blocks and w/o indels; B) w/o constrained blocks and w/ indels; C) w/ constrained blocks and w/o indels; D) w/ constrained blocks and w/ indels. For each divergence distance, 1,000 replicates were used to calculate the mean and standard error for the following statistics: estimated overall divergence (black boxes), estimated divergence in constrained blocks of sites (grey circles), estimated divergence in unconstrained blocks of sites (brown triangles), identity (red crosses), identity in constrained blocks (yellow x's), identity in unconstrained blocks (green diamonds), fraction of ancestral sequence remaining in derived sequences (green triangle), and fraction of constraint (light blue checked boxes). Note that the divergence scale in this and following figures is discontinuous.

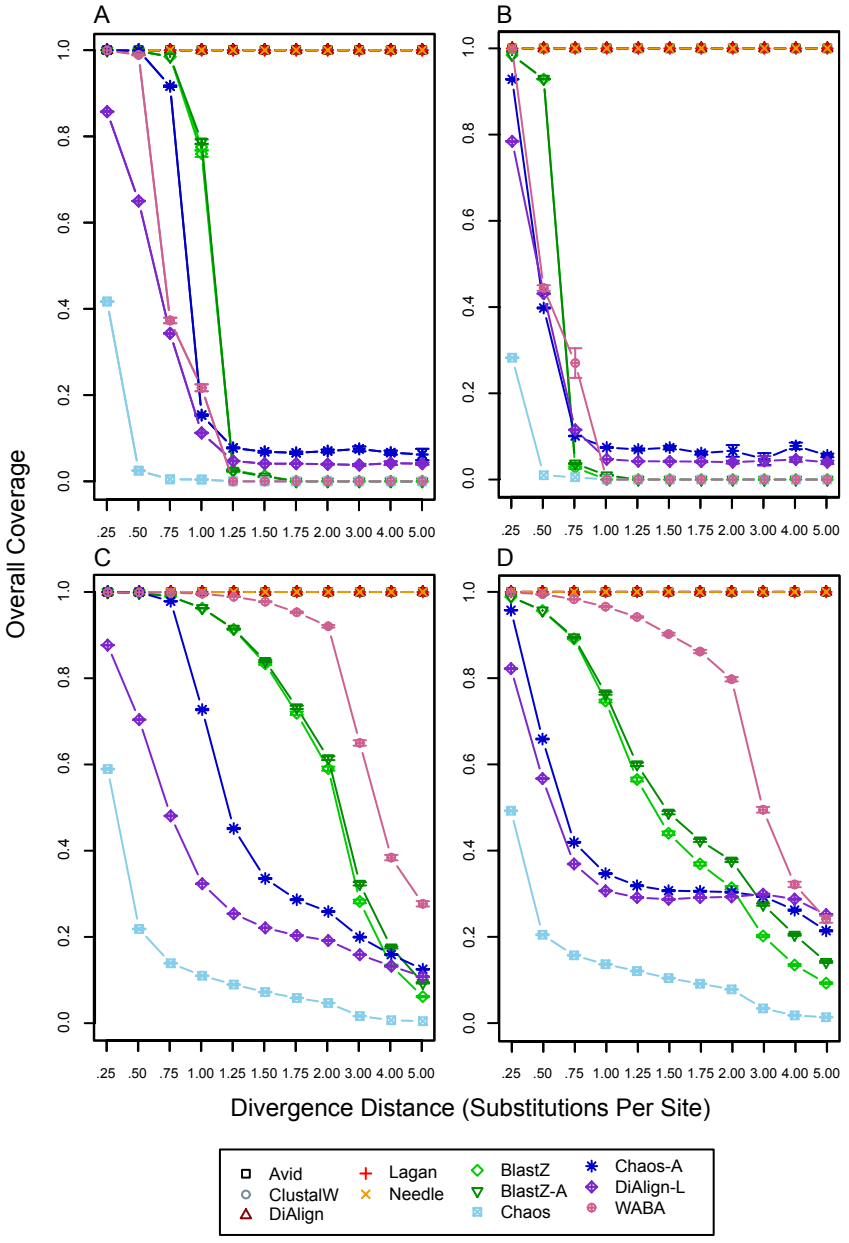

Figure 2.3
Overall alignment coverage For each divergence distance and each tool, 1,000 replicates were used to calculate the mean and standard error of overall alignment coverage, which was defined as the fraction of ungapped, orthologous pairs of sites in the simulated alignment that were included in an alignment produced by a tool (see Methods for details). A) w/o constrained blocks and w/o indels; B) w/o constrained blocks and w/ indels; C) w/ constrained blocks and w/o indels; D) w/ constrained blocks and w/ indels.

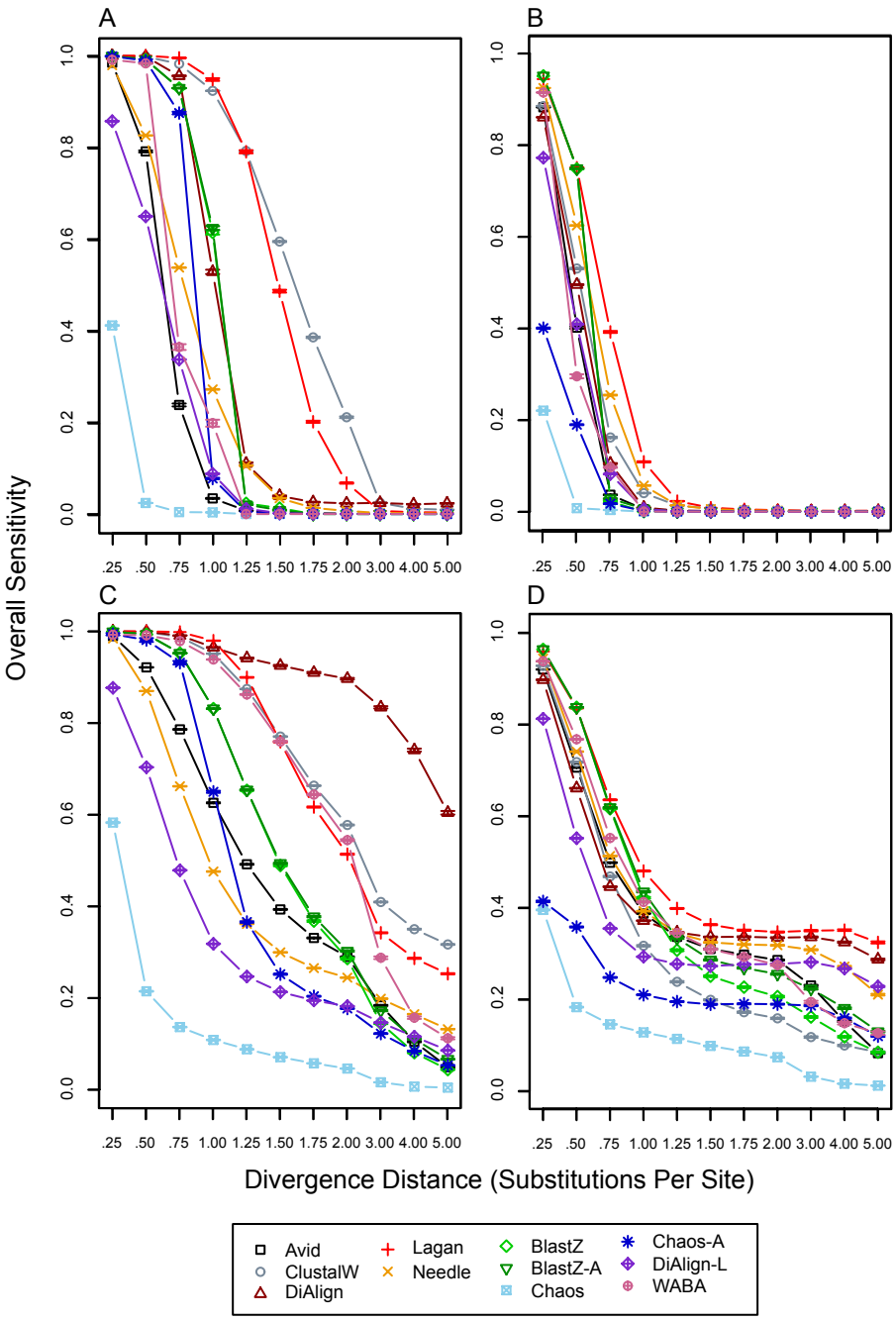

Figure 2.4

Overall alignment sensitivity. Overall alignment sensitivity For each divergence distance and each tool, 1,000 replicates were used to calculate the mean and standard error of overall alignment sensitivity, which was defined as the fraction of ungapped, orthologous pairs of sites in the simulated alignment that were aligned correctly in an alignment produced by a tool (see Methods for details). A) w/o constrained blocks and w/o indels; B) w/o constrained blocks and w/ indels; C) w/ constrained blocks and w/o indels; D) w/ constrained blocks and w/ indels.

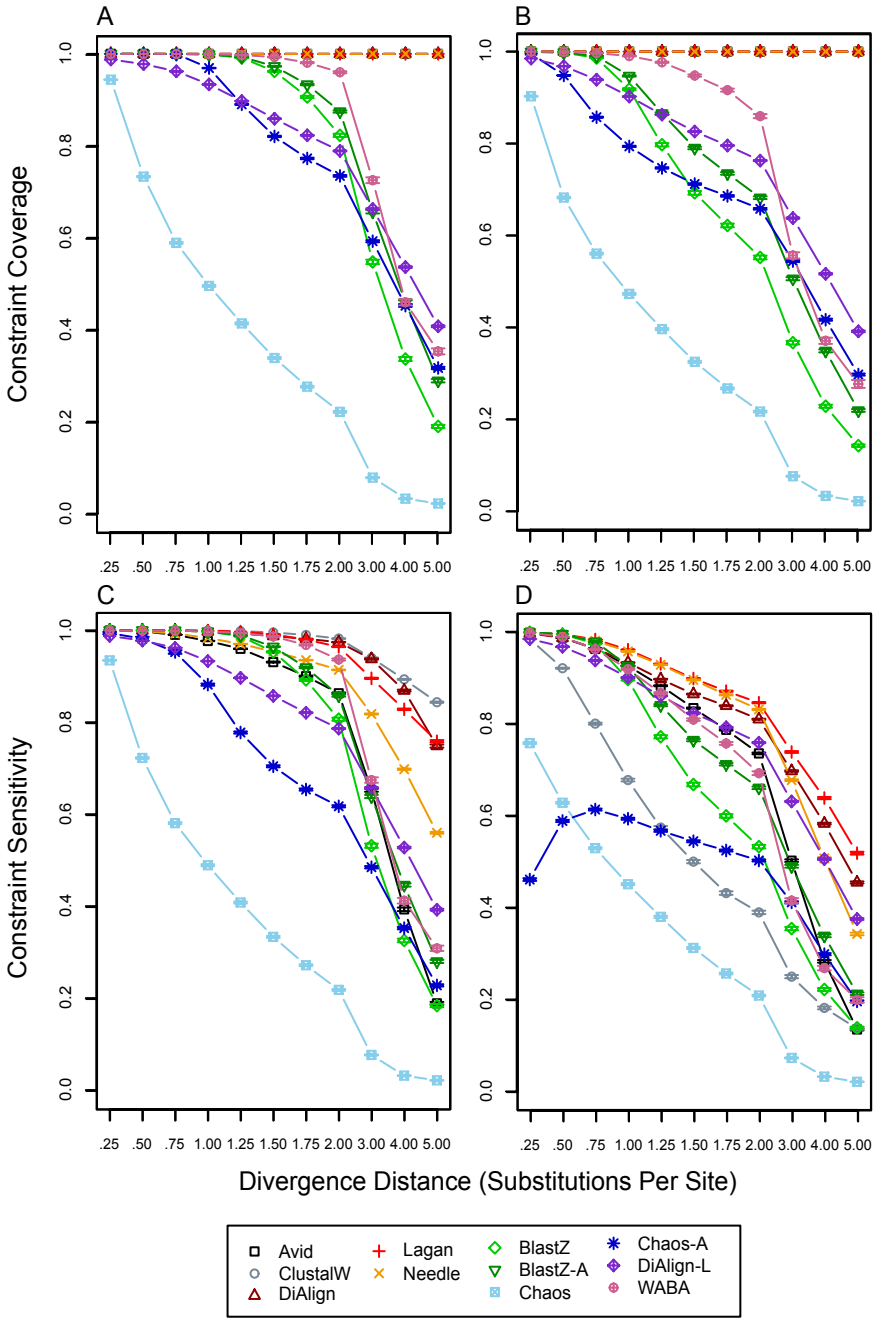

Figure 2.5
Constraint coverage and sensitivity For each divergence distance and each tool, 1,000 replicates were used to calculate the mean and standard error of constraint coverage and constraint sensitivity, which were defined as the coverage and sensitivity within interspersed constrained blocks (see Methods for details). A) constraint coverage w/o indels; B) constraint coverage w/ indels; C) constraint sensitivity w/o indels; D) constraint sensitivity w/ indels.

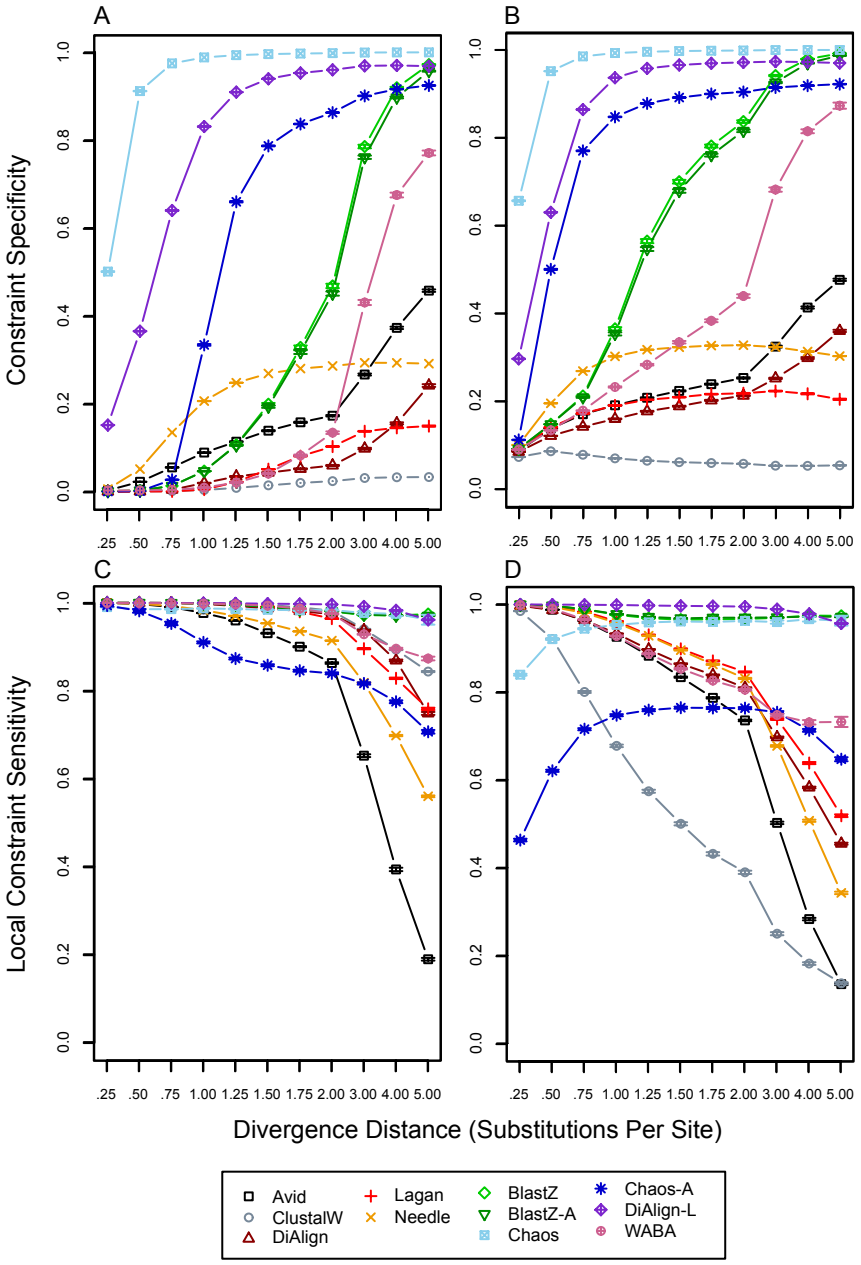

Figure 2.6

Constraint specificity and local constraint sensitivity For each divergence distance and each tool, 1,000 replicates were used to calculate a mean and standard error of constraint specificity and local constraint sensitivity. Constraint specificity was defined as the fraction of unconstrained sites in the simulated alignment that were unaligned or gapped in an alignment produced by a tool. Local constraint specificity was defined as the constraint sensitivity for just the sites contained in an alignment produced by a tool (see Methods for details). A) constraint specificity w/o indels; B) constraint specificity w/ indels; C) local constraint sensitivity w/o indels; D) local constraint sensitivity w/ indels.

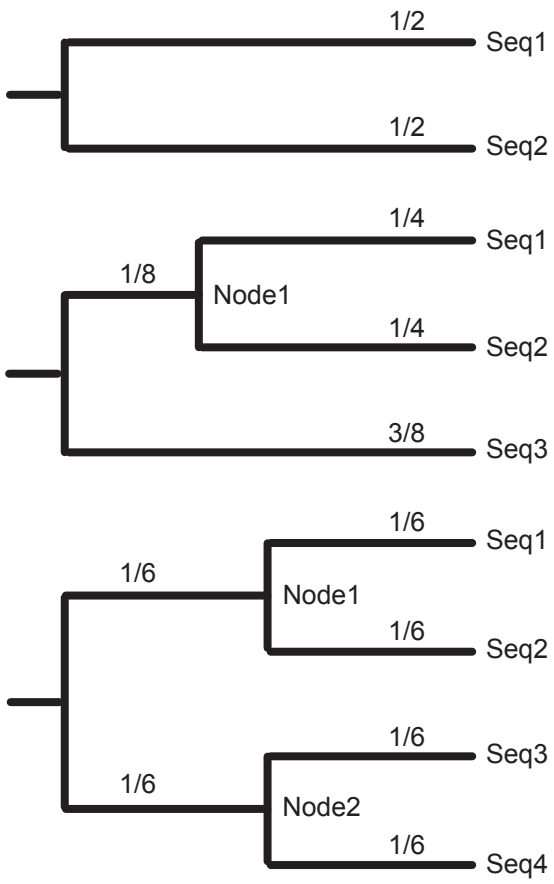

Figure 3.1
Mutation Guide Trees. Simulations were performed on two, three and four species trees. Numbers on the branches indicate the fraction of the total tree divergence distance on each branch.

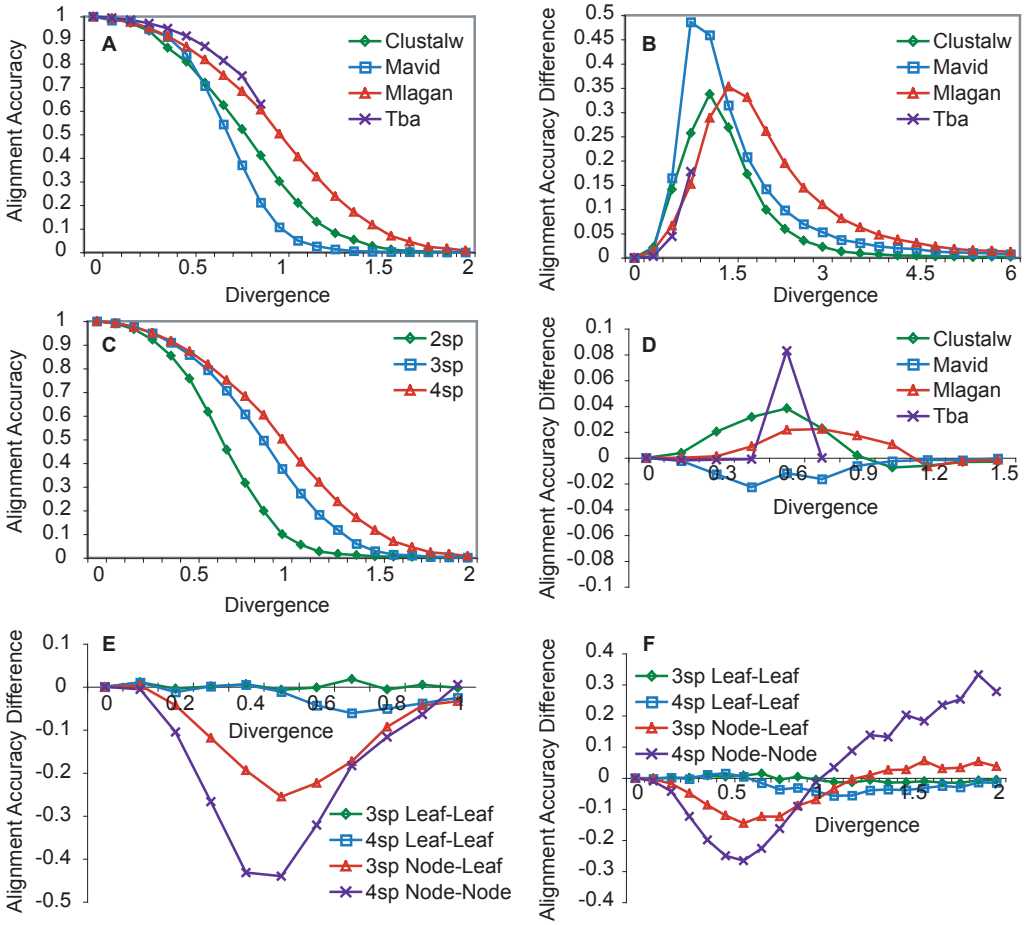

Figure 3.2

Multiple Alignment Accuracy. A: Alignment accuracy varies across tools and divergences. Mean four species alignment accuracy for each tool was measured as a function of total divergence distance. B: Alignment accuracy improves with the presence of transcription factor binding sites. Mean improved alignment accuracy of enhancers over background sequences for four species alignments was measured as a function of total divergence distance. C: Dividing a fixed total divergence up with more species improves alignment accuracy. Mean Mlagan alignment accuracy for two, three and four species trees was measured as a function of total divergence distance. D: Adding in-group species to a pair of species has no effect on the alignment accuracy of the pair. Mean improved alignment accuracy of three species alignments over two species alignments, where the divergence distance between Seq1 and Seq3 in the three species alignment was the same as the divergence distances between Seq1 and Seq2 in the two species alignment, was measured as a function of divergence distance. E & F: Alignment accuracy varies across branches in a tree and is best for leaf-to-leaf alignments and worst for node-to-node alignments, with the exception of highly diverged enhancers. Mean Clustalw alignment accuracy along branches in three and four species trees subtracted from mean two species alignment accuracy, where divergence along each branch is the same as the two species divergence, was measuredin background sequences (E) and enhancers (F) as a function of divergence distance.

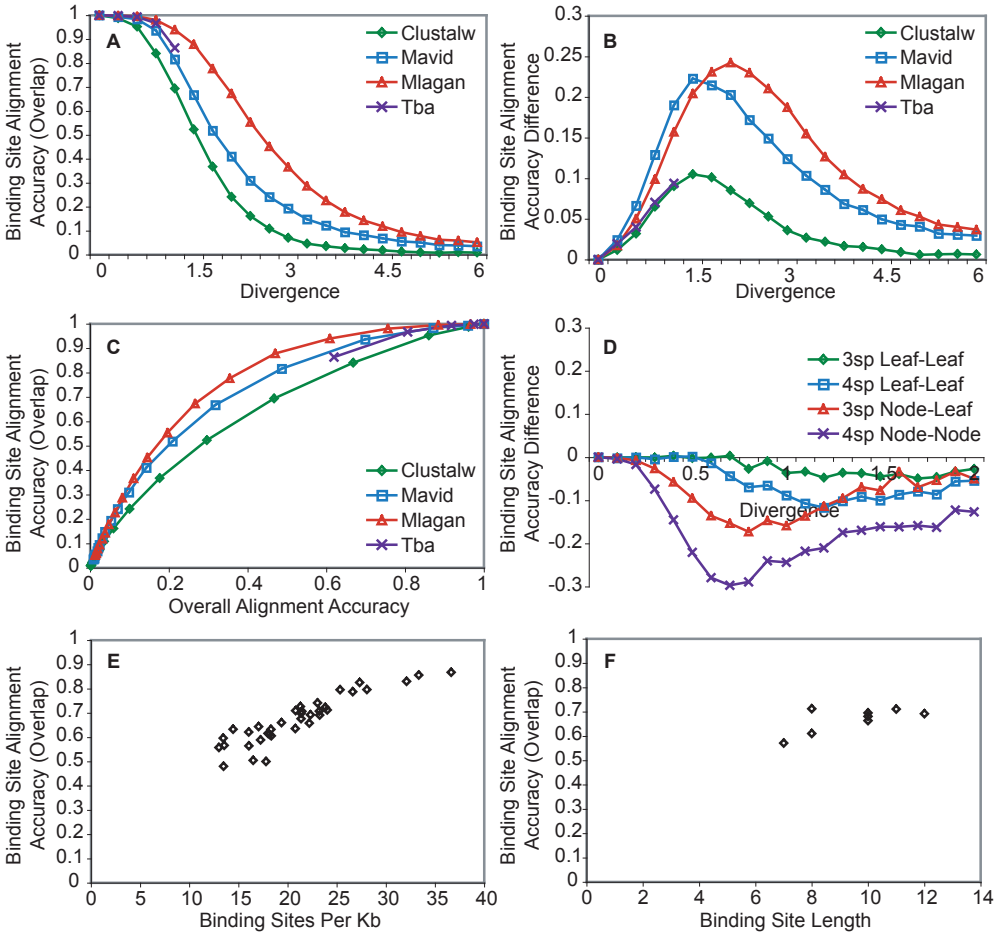

Figure 3.3

Transcription Factor Binding Site Alignment Accuracy. A: Binding site alignment accuracy varies across tools and divergences. Fraction of binding sites overlapping in four species alignments was measured as a function of total divergence distance. B: Binding sites are often still overlapping in alignments even when they are not perfectly aligned. Fraction of binding sites perfectly aligned in four species alignments subtracted from the fraction of binding sites overlapping in four species alignments was measured as a function of total divergence distance. C: Binding site alignment accuracy is highly correlated with overall alignment accuracy and is consistently higher. Fraction of binding sites overlapping in four species alignments was measured as a function of overall alignment accuracy. D: Binding site alignment accuracy varies across branches in a tree and is best for leaf-to-leaf alignments and worst for node-to-node alignments. Fraction of binding sites overlapping along branches in three and four species trees subtracted from the fraction of binding sites overlapping in two species Clustalw alignments, where the divergence along each branch is the same, was measured as a function of divergence distance. E: Binding site alignment accuracy is positively correlated with binding site density in an enhancer. Fraction of binding sites overlapping in replicate four species Mlagan alignments of each of the 36 enhancers was measured as a function of the density of binding sites in the enhancer. F: Binding site alignment accuracy is positively correlated with binding site length. Fraction of binding sites overlapping in four species Mlagan alignments for each of the eight transcription factors was measured as a function of the length of the transcription factors' binding sites.

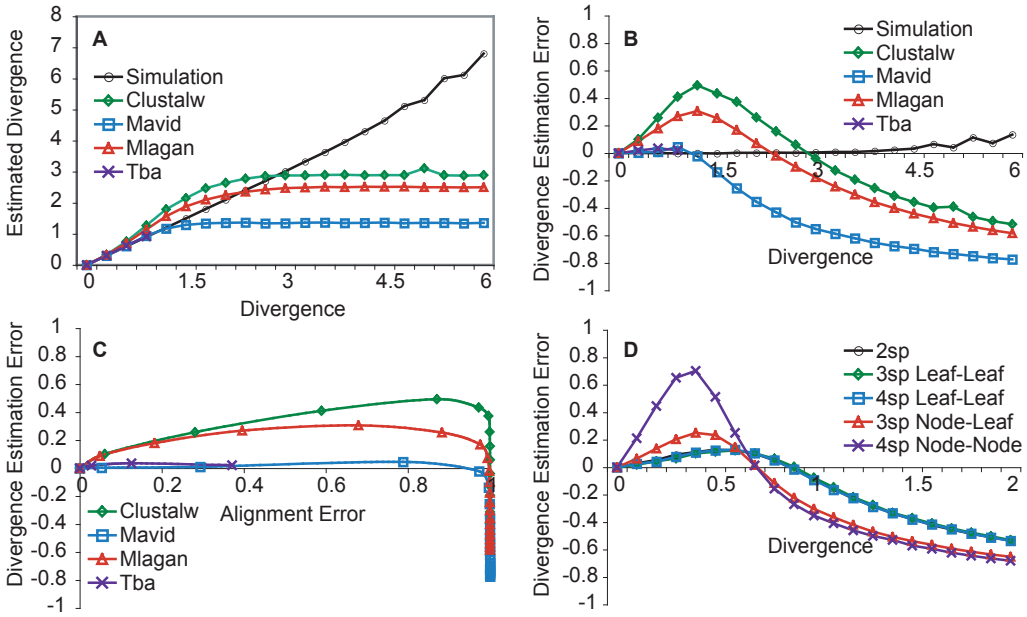

Figure 3.4

Divergence Distance Estimation. Divergences estimated from tool alignments are overestimated at short divergence distances and underestimated at large divergence distances while divergences estimated from true simulated alignments are accurate to large divergence distances. A: Mean divergence distance estimated from simulated alignments and tool alignments for four species trees was measured as a function of total true divergence distance. B: Mean divergence estimation error (Estimate – True/True) for four species trees was measured as a function of total true divergence distance. C: Divergence estimation error from tool alignments is not correlated with alignment error. Mean divergence estimation error for four species trees was measured as a function of mean alignment error. D: Divergence estimation error varies across branches in a tree and is best for leaf-to-leaf alignments and worst for node-to-node alignments. Mean divergence estimation error along branches of equal true divergence from two, three and four species Mlagan alignments was measured as a function of true divergence distance.

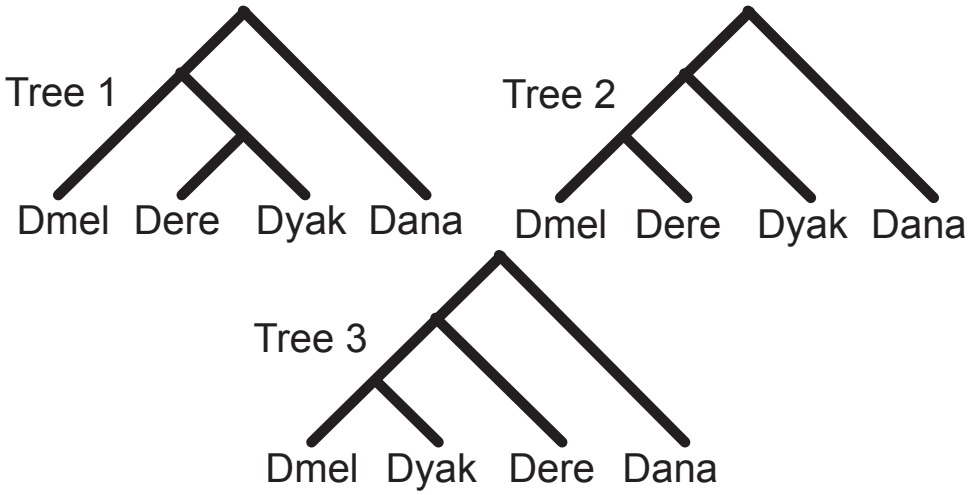

Figure 4.1
Phylogenies. The three possible phylogenies for Dmel, Dere, and Dyak, with Dana as an outgroup.

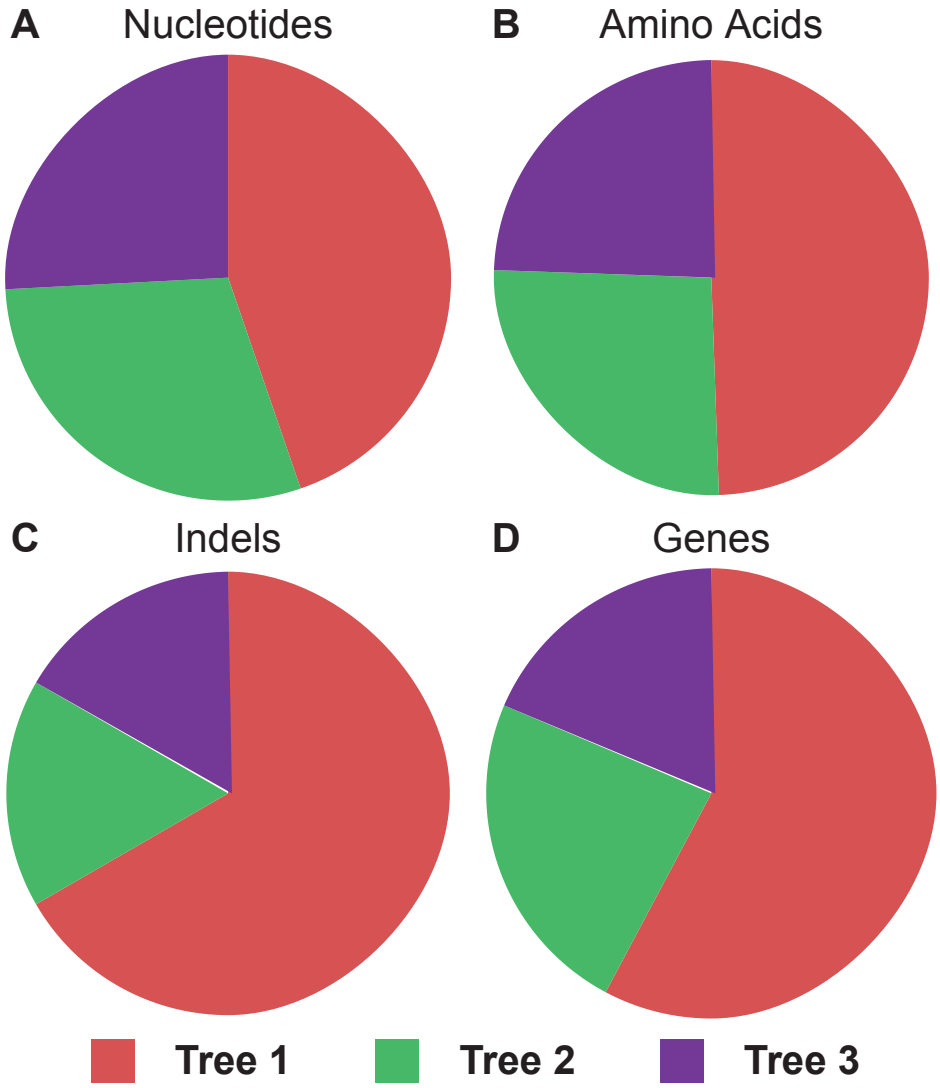

Figure 4.2
Widespread Incongruence of Substitutions, Indels, and Gene Trees. (A) The proportion of informative nucleotide substitutions in 9,405 genes supporting each of the three trees. Tree 1 (red) is supported by 170,002 (44.7%) nucleotide changes; tree 2 (green), 112,278 (29.5%) nucleotide changes; and tree 3 (purple), 98,117 (25.8%) nucleotide changes. (B) The proportion of informative amino acid substitutions in 9,405 genes supporting each of the three trees. Tree 1 (red) is supported by 28,628 (49.3%) amino acid changes; tree 2 (green), 15,182 (26.2%) amino acid changes; and tree 3 (purple), 14,203 (24.5%) amino acid changes. (C) The proportion of informative insertions or deletions (indels) in 9,405 genes supporting each of the three genes. Indels were filtered, requiring five flanking amino acids of perfect identity and no repetitive sequence. Tree 1 (red) is supported by 2 deletions and 6 insertions (66.7%); tree 2 (green), 1 deletion and 1 insertion (16.7%); and tree 3 (purple), 2 insertions (16.7%). Similar proportions but much larger counts are found when the indels are not filtered. (D) The proportion of 9,315 genes with ML support for each of the three trees. Tree 1 (red) has ML support for 5,381 (57.8%); tree 2 (green), 2,188 (23.5%); and tree 3 (purple), 1,746 (18.7%).

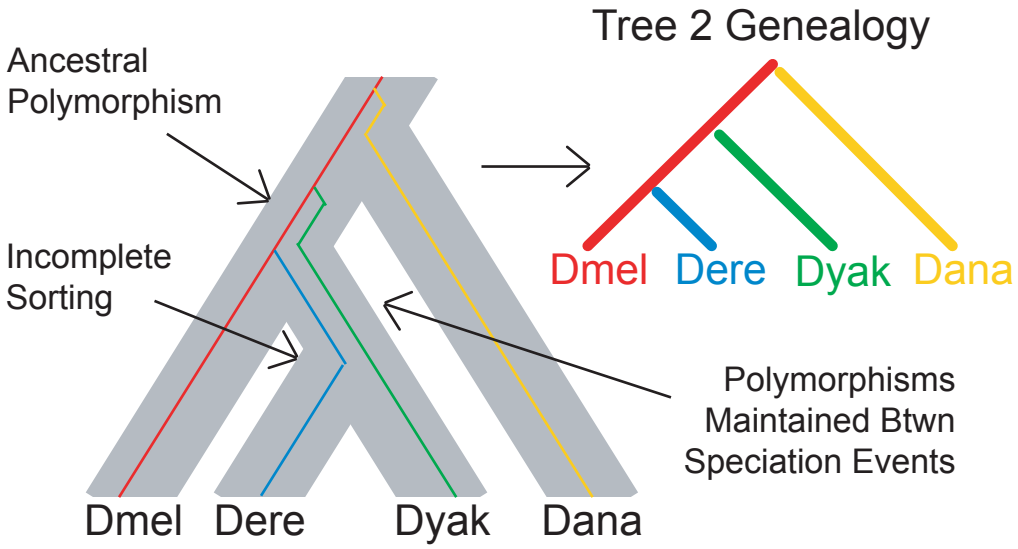

Figure 4.3
Incomplete Lineage Sorting. The history of a gene (colored lines) is drawn in the context of a species tree (gray bars). New lineages arising from new polymorphisms in the gene are drawn in different colors. In this case, the two alleles in the population prior to the split of Dmel are maintained through to the split of Dere and Dyak, leading to incomplete lineage sorting and an incongruent genealogy (tree 2). The greater the diversity in the ancestral population and the shorter the time between speciation events, the more likely nonspecies genealogies are.

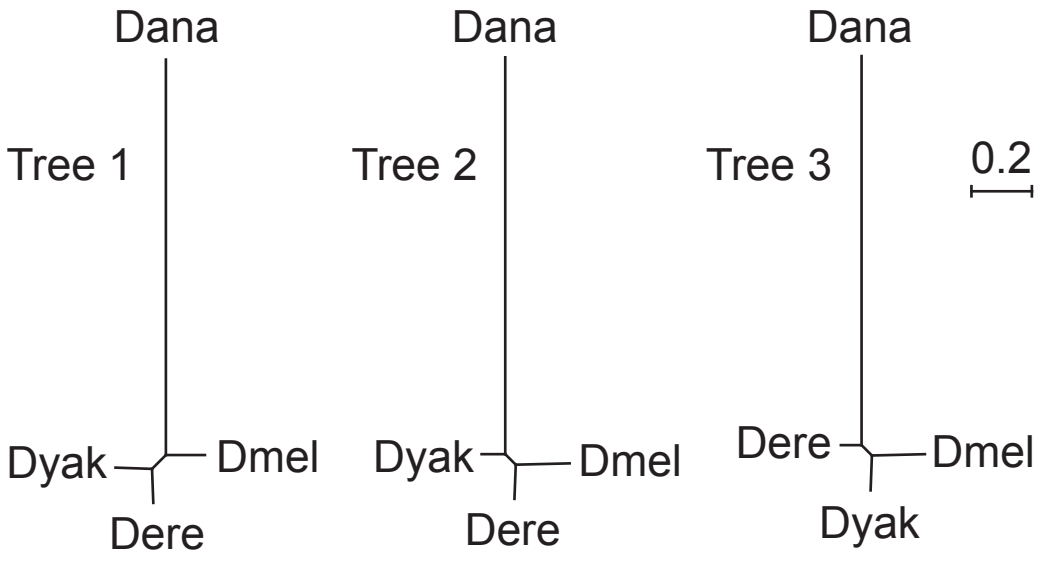

Figure 4.4
Median Synonymous Trees. Median synonymous branch length trees derived from the genes supporting each of the three trees are drawn to the same scale. The branch spanning the two speciation events is quite short for all trees.

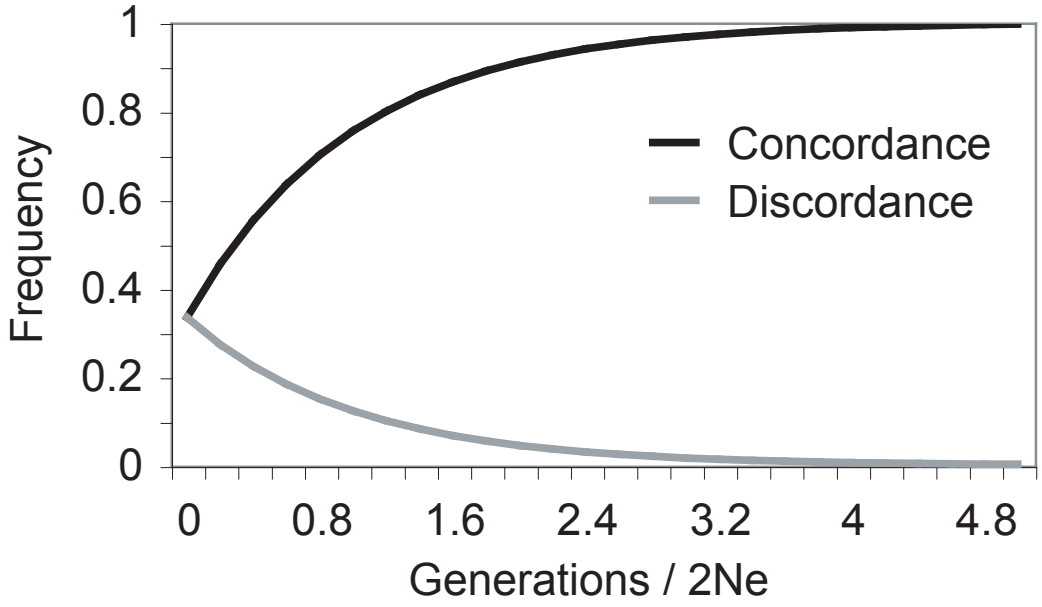

Figure 4.5
Coalescence Probabilities for Each Tree. Using the formula p(congruence)=1-2/3exp(-t), where t=generations / 2Ne, the probability of the species tree (black) and the probability of one of the two alternate trees (gray) was plotted as a function of t.

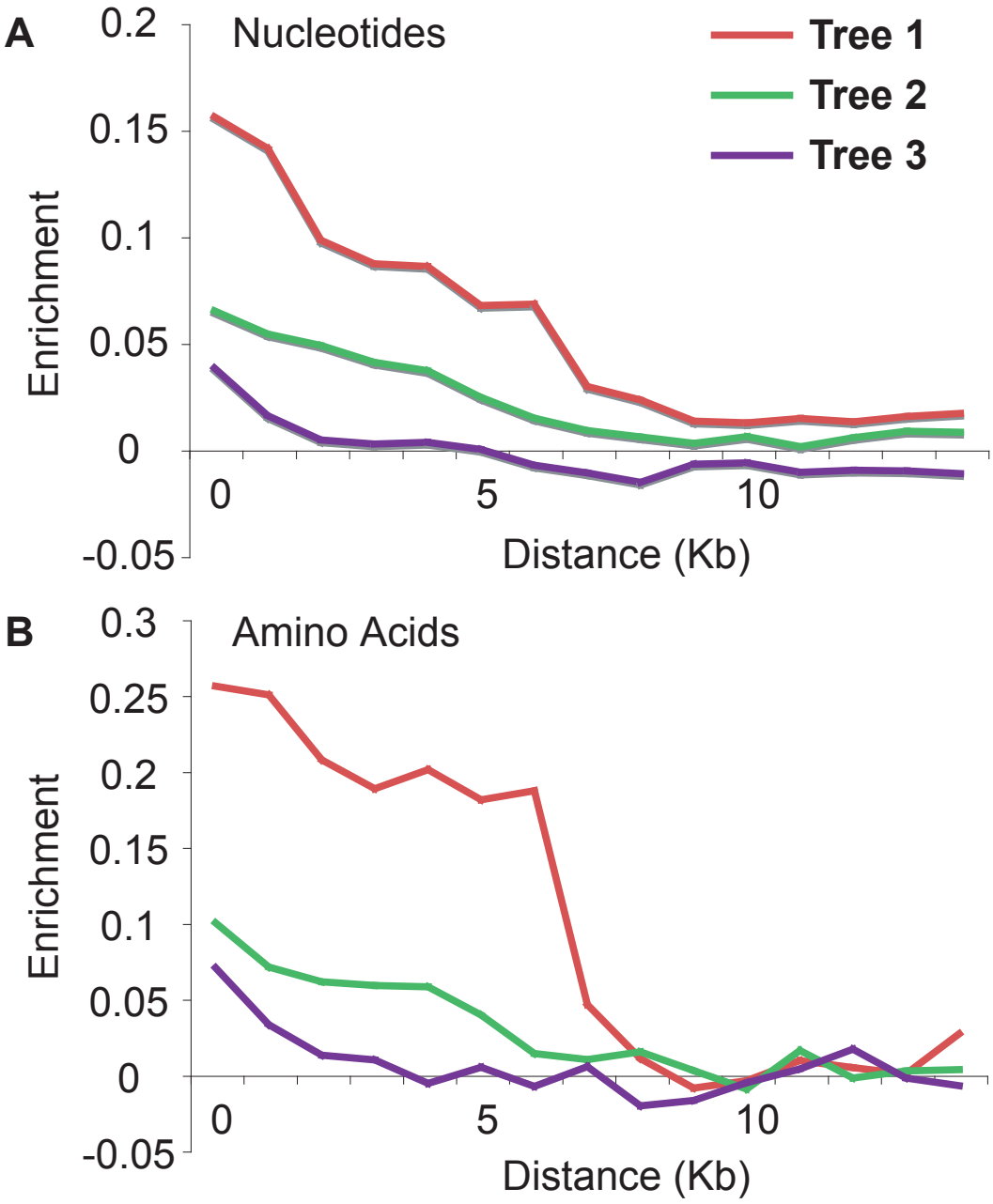

Figure 4.6
Clustering of Informative Sites. The enrichment of informative nucleotide (A) and amino acid (B) substitutions near other substitutions that support the same phylogeny was found for all three trees and is on a scale roughly similar to estimates of linkage disequilibrium. At each informative site in the genome, the counts of informative sites supporting each of the three trees in 1-kb windows extending 30 kb up- and downstream were measured. For each type of informative site, the enrichment of the same type of informative site in each 1-kb window was calculated using the observed counts and the expected number of sites based on their genome-wide frequency. Enrichment is log10(observed / expected).

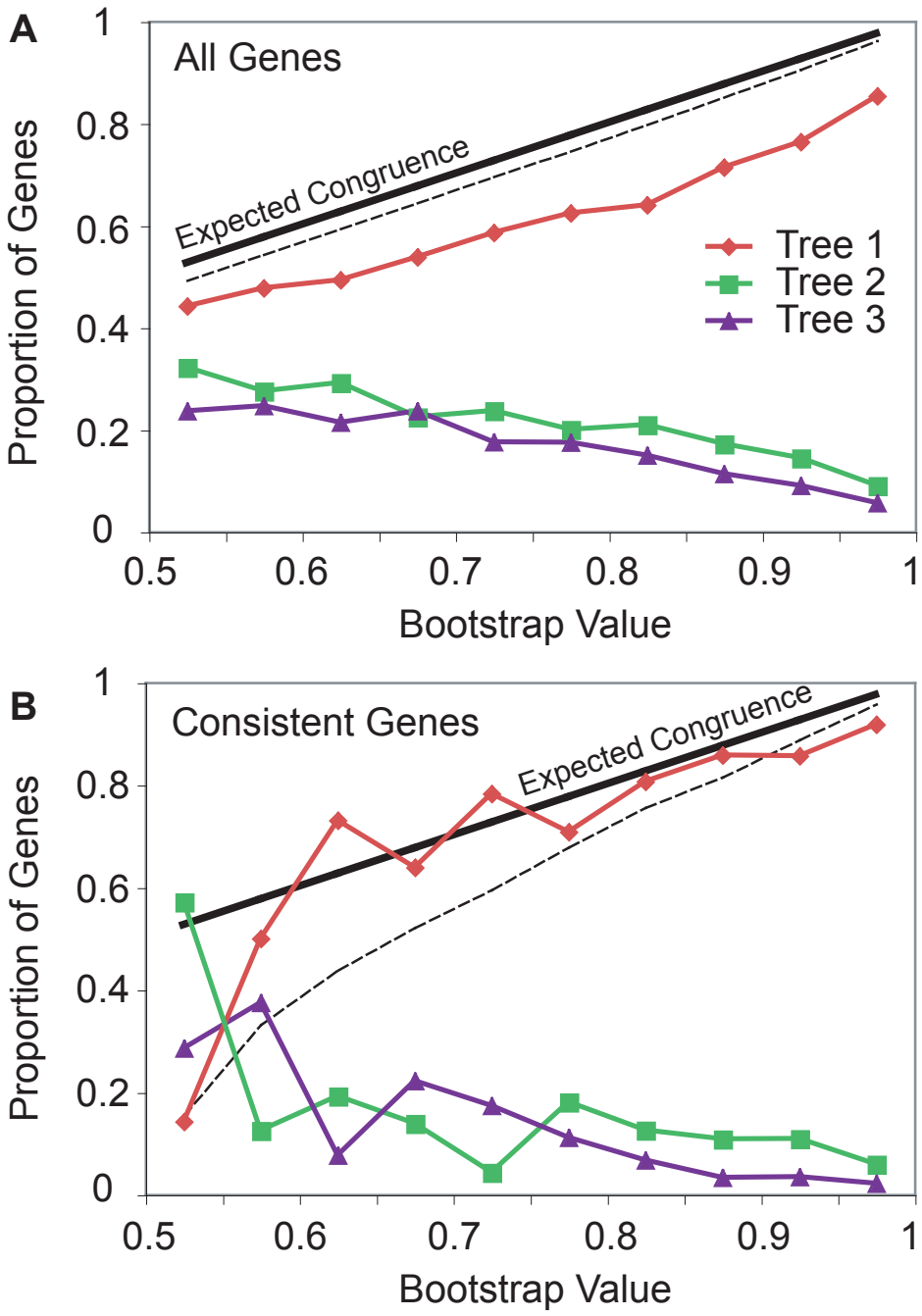

Figure 4.7
Significance of Incongruence. An excess of incongruence above what is expected by chance was observed for the set of all genes (A) as well as the set of genes that consistently supported the same tree across models and species combinations (B). Genes were binned by bootstrap value, and the proportion of genes supporting tree 1 (red line), tree 2 (green line), and tree 3 (purple line) were plotted. The expected congruence based on the bootstrap value in each bin (black solid line) and the 95% confidence interval based on a $X2$ distribution (black dash line) demonstrate the excess incongruence.

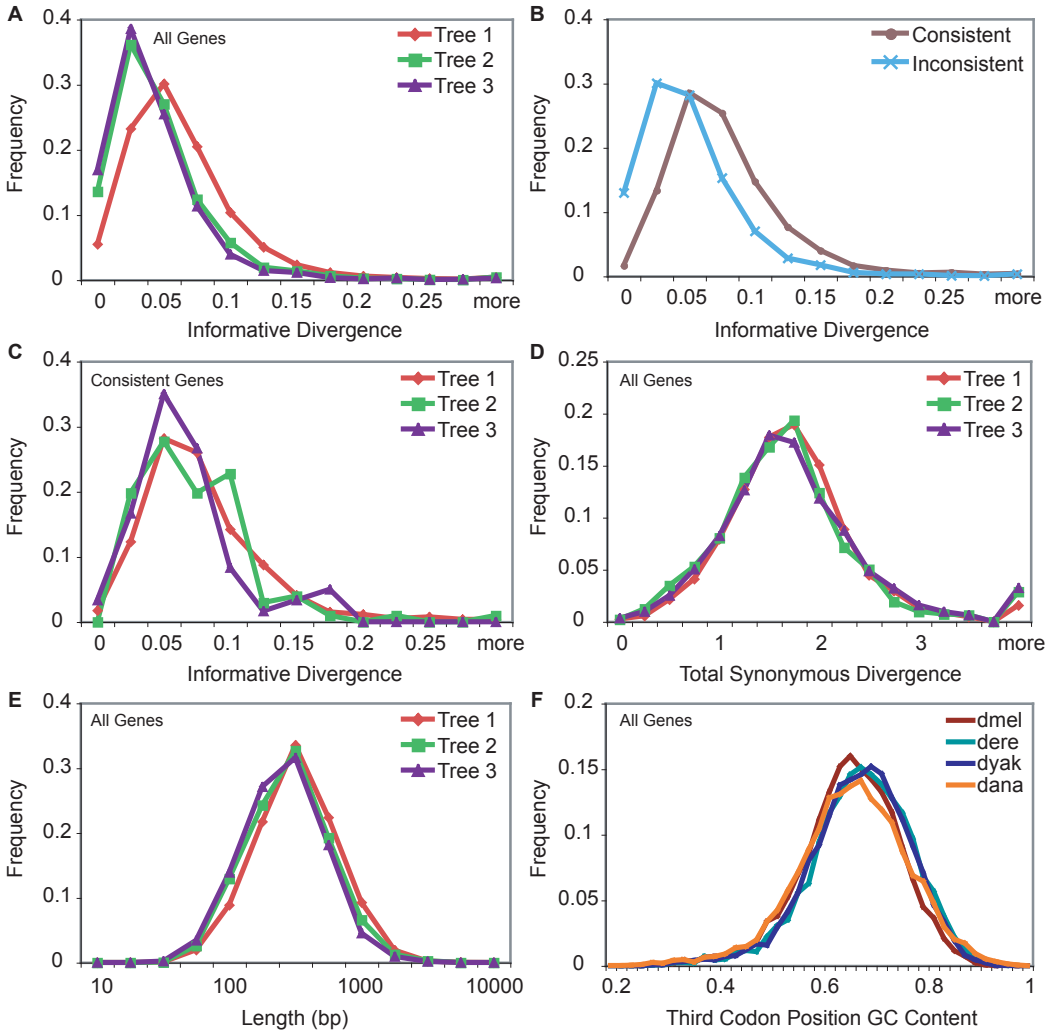

Figure 4.8

Sequence and Evolutionary Gene Properties Sequence and evolutionary properties of the genes are unable to explain the incongruence. Distributions are calculated using results from the original ML analysis using the F334 model and the Dmel, Dere, Dyak, and Dana species combination. The distributions of informative synonymous divergences in genes supporting each tree reveal a bias toward lower values for the incongruent genes (A). Nearly all genes with little or no informative synonymous divergence, however, are classified as inconsistent (B). Therefore, consistent genes have very similar distributions of ISD across trees (C). TSD is distributed similarly across trees, suggesting homoplasy due to increased mutation rates is not causing the incongruence (D). Gene length is slightly higher in tree 1 genes but overall is very similar across trees (E). Third codon position GC content is slightly biased toward lower values for Dmel and Dana and higher values for Dere and Dyak, creating a conservative bias for the incongruence (F).

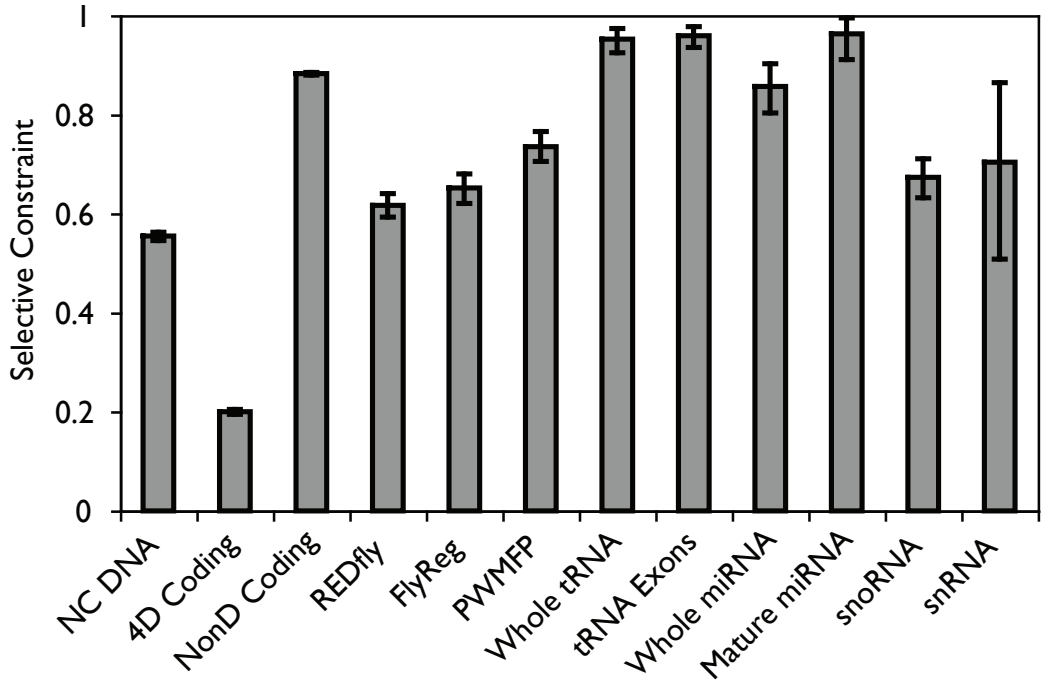

Figure 5.1
Selective constraints on genomic annotation classes. Mean selective constraint across elements in each annotation class was calculated using divergence in short introns as a neutral standard. 95% confidence intervals were calculated using 1000 bootstrap resamplings.

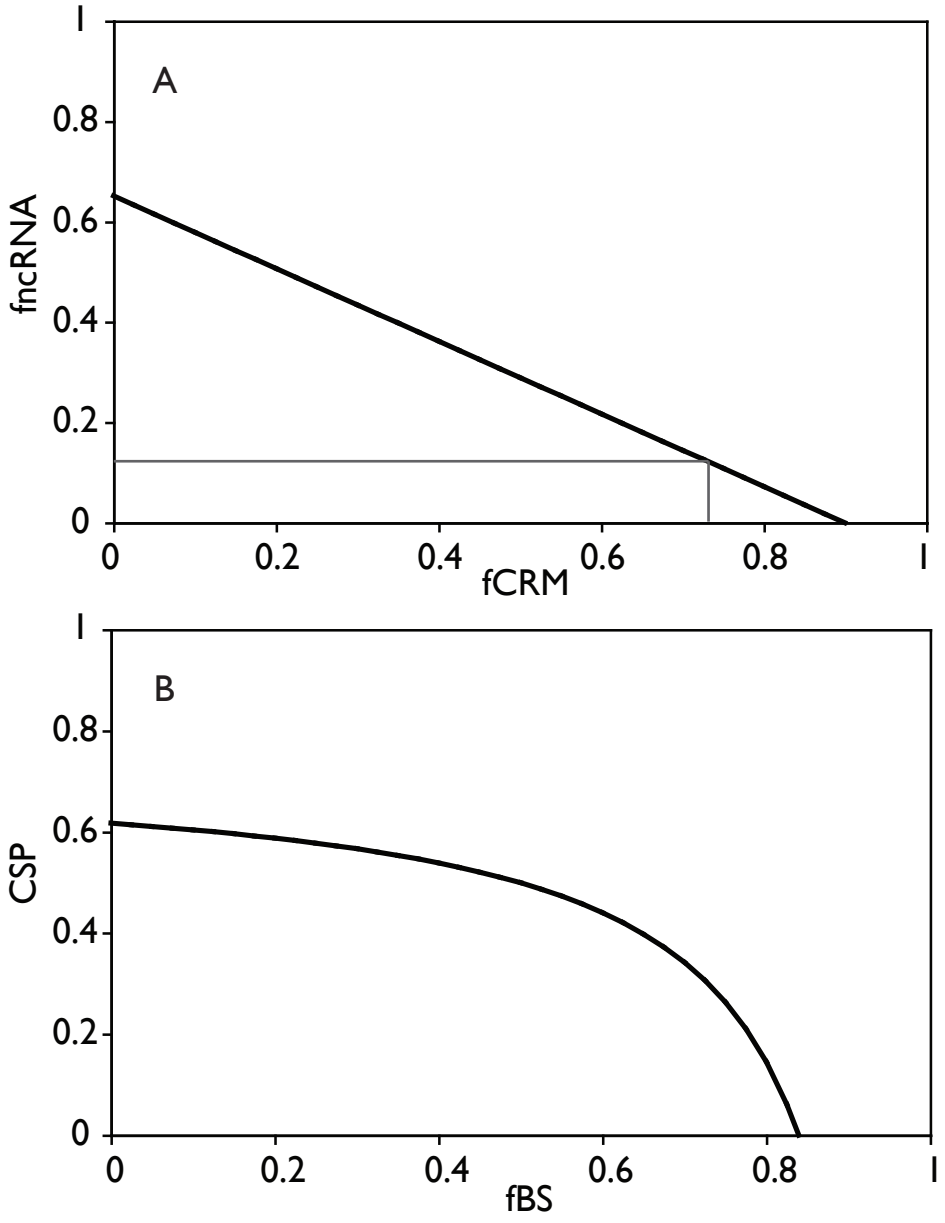

Figure 5.2
Genome and CRM content. (A) Predicted proportion of non-coding genome covered by cis-regulatory modules (fCRM) and non-coding RNAs (fncRNA). Assuming CRMs and ncRNAs are the only functionally constrained non-coding elements, a line describes the relationship of the proportion of the non-coding genome that each would need to cover (black line). The maximum CRM proportion is 90% and the maximum ncRNA proportion is 65%. The best annotated Drosophila locus contains the even-skipped gene, which has CRMs covering 73% of adjacent non-coding sequences (grey verticle line). The corresponding ncRNA proportion is 12% (grey horizontal line), leaving 15% unconstrained sequence. (B) Predicted proportion of of CRMs covered by transcription factor binding sites (fBS) and predicted non-binding site spacer sequence mean constraint (CSP). A curve describes relationship of fBS and CSP (black line). The maximum binding site proportion is 84% and the maximum spacer constraint is 0.617.