

Spring 2009

Genetic variation within the *Daphnia pulex* genome

Abraham Eaton Tucker
University of New Hampshire, Durham

Follow this and additional works at: <https://scholars.unh.edu/dissertation>

Recommended Citation

Tucker, Abraham Eaton, "Genetic variation within the *Daphnia pulex* genome" (2009). *Doctoral Dissertations*. 490.
<https://scholars.unh.edu/dissertation/490>

This Dissertation is brought to you for free and open access by the Student Scholarship at University of New Hampshire Scholars' Repository. It has been accepted for inclusion in Doctoral Dissertations by an authorized administrator of University of New Hampshire Scholars' Repository. For more information, please contact nicole.hentz@unh.edu.

GENETIC VARIATION WITHIN THE *DAPHNIA PULEX* GENOME

BY

ABRAHAM EATON TUCKER

B.S., University of Southern Maine, 2003

DISSERTATION

Submitted to the University of New Hampshire
in Partial Fulfillment of
the Requirements for the Degree of

Doctor of Philosophy

In

Genetics

May, 2009

UMI Number: 3363734

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

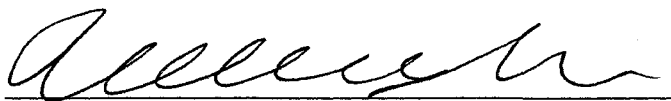
In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI[®]

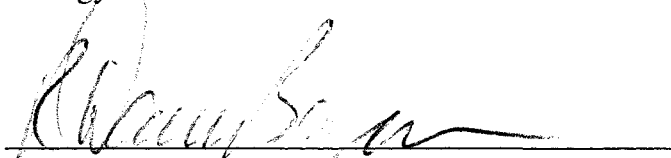
UMI Microform 3363734
Copyright 2009 by ProQuest LLC
All rights reserved. This microform edition is protected against
unauthorized copying under Title 17, United States Code.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

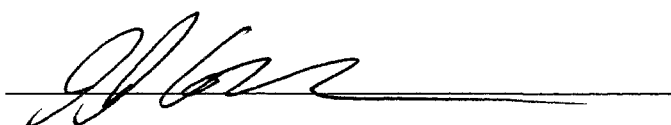
This dissertation has been examined and approved.



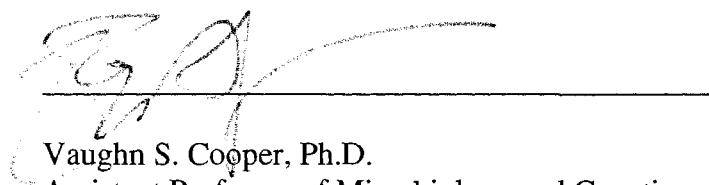
Dissertation Director, W. Kelley Thomas, Ph.D.
Director, Hubbard Center for Genome Studies
Professor of Biochemistry and Molecular
Biology and Genetics



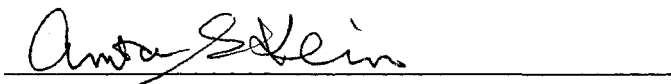
R. Daniel Bergeron, Ph.D.
Professor of Computer Science



John J. Collins, Ph.D.
Associate Professor of Biochemistry and Molecular
Biology and Genetics



Vaughn S. Cooper, Ph.D.
Assistant Professor of Microbiology and Genetics



Anita S. Klein, Ph.D.
Associate Professor of Biochemistry and Molecular
Biology and Genetics

2/13/09
Date

DEDICATION

To my grandparents, Rev. Ralph Lewis and Mildred Moore Tucker, who have always encouraged, supported and shown interest in my education.

ACKNOWLEDGEMENTS

Funding Sources:

Sequencing Facility Technician (September 2003-December 2004)

Genetics Department Teaching Assistantship (Spring 2005)

NSF GK-12 PROBE Fellowship (May 2005-May 2007)

NSF Research Assistantship (May 2007-December 2008)

People:

Special thanks to my advisor Kelley Thomas for his mentorship and support.

Dissertation Committee: Profs. R. Daniel Bergeron, John Collins, Vaughn Cooper and Anita Klein

This research project was aided directly and indirectly by the collaborations and contributions of the following people:

Way Sung, Morel Henley, Sanjuro Jogdeo, Phil Hatcher, Darren J. Bauer, Shilpa Kulkarni, Karen Carleton, Wenli Li, John Colbourne, Don Gilbert, Mike Lynch, Jeong-Hyeon Choi.

Genome center comrades: Feseha, Joe, Janet, Kazu, Krys, Weston, Will

Most importantly, I acknowledge the friendship and love of my wife, Antoinette.

TABLE OF CONTENTS

DEDICATION	iii
ACKNOWLEDGEMENTS	iv
LIST OF TABLES	viii
LIST OF FIGURES	ix
ABSTRACT	xii
CHAPTER	PAGE
INTRODUCTION	1
Current state of genome biology	1
Population genomics and evolution	4
<i>Daphnia pulex</i> genome	6
I. SMALL NUCLEOTIDE POLYMORPHISMS	11
Background	11
SNPs and evolution	11
Neutral theory and population genetics	13
SNP studies	18
<i>Daphnia</i> variation	22

Methods	24
Comparative Assembly (TCO)	24
Comparative Assembly (TRO)	28
Results and Discussion	29
Pre-assembly quality control	29
Comparative Assembly (TCO vs. TCO)	30
Scope of study	31
Magnitude of variation	32
SNP types	33
Functional distribution	41
Physical distribution	42
Recombination	45
Windows analysis	47
Windows analysis	49
Comparative Assembly (TCO vs. TRO)	55
Maximum likelihood analysis	58
Conclusion	59
Acknowledgements	59

II. PATTERNS OF VARIATION IN RECENTLY DIVERGED

MITOCHONDRIAL GENOMES OF *DAPHNIA PULEX* 60

Background 60

Mitochondrial genomes and evolution 60

Methods	63
Results and discussion	64
III. INTRON GAIN/LOSS POLYMORPHISMS IN <i>DAPHNIA PULEX</i>	69
Background	69
Intron evolution	69
Methods	71
Results and discussion	72
Acknowledgements	78
IV. GENE DUPLICATION IN <i>DAPHNIA PULEX</i>	79
Background	79
Methods	83
Results and discussion	85
Acknowledgements	99
LIST OF REFERENCES	103
APPENDICES	122

LIST OF TABLES

<u>Table</u>	<u>Title</u>	<u>Page</u>
Table 1-1	Pre and post LUCY-treated data	29
Table 1-2	Types of the transversions	40
Table 2-1	Observed and expected substitutions in mtDNA	66
Table 3-1	Intron absences in TRO	77
Table 4-1	Gene content of fully sequenced invertebrates	82
Table 4-2	Birth rates of gene duplicates	86
Table 4-3	Ks values for all, single and pseudogene pairs	96
Table 4-4	Spatial categories among all gene duplicates	99

LIST OF FIGURES

<u>Table</u>	<u>Title</u>	<u>Page</u>
Figure I-1	Phylogeny of completed arthropod genome	6
Figure I-2	Reproductive mode in <i>Daphnia</i>	8
Figure I-3	Distribution of N50 scaffold sizes	10
Figure 1-1	SNP Pipeline flowchart	26
Figure 1-2	TRO reads mapped to the TCO assembly	29
Figure 1-3	Average coverage of assemblies	31
Figure 1-4	Frequency of coverage	31
Figure 1-5	N50 sites under criteria	32
Figure 1-6	Categories of polymorphic sites	33
Figure 1-7	Base substitutional matrix	34
Figure 1-8	Frequency of SNP types by scaffold	35
Figure 1-9	Distribution of sequential indels found within TCO	36
Figure 1-10	Distribution of sequential base substitutions within TCO	37
Figure 1-11	Frequency of SNP types in N50 scaffolds	38
Figure 1-12	Ts/Tv by scaffold	39
Figure 1-13	Ts/Tv and SNP frequency	40
Figure 1-14	SNPs in exons, intron and intergenic sequence	41

Figure 1-15	Distribution of base substitutions in exons	42
Figure 1-16	Scaffold-wide base substitution frequency	44
Figure 1-17	Polymorphism on mapped scaffolds	45
Figure 1-18	SNP frequency and recombination rate	47
Figure 1-19	Windows analysis of SNP frequency across scaffold 1	48
Figure 1-20	Runs of Zero-SNP windows	49
Figures 1-21	Polymorphism across scaffolds	50-55
Figure 1-31	Frequency of coverage for TRO	56
Figure 1-32	Polymorphism and divergence between TRO and TCO	57
Figure 1-33	Nucleotide divergence between TRO and the TCO	58
Figure 2-1	TCO-TRO-Crease tree	64
Figure 2-2	Dilatory mutation-selection	68
Figure 3-1	Duplications at intron/exon boundaries in 13/30 TCO introns	76
Figure 4-1	Homology among predicted genes in <i>D. pulex</i>	88
Figure 4-2	Age distribution of gene duplicates at >60% AA identity	89
Figure 4-3	Age distribution of gene duplicates at >40% AA identity	90
Figure 4-4	Distribution of single copy gene pairs	91
Figure 4-5	Single gene duplicates using 40%	92
Figure 4-6	Selection intensity (Ka/Ks) and age of single gene pairs	94
Figure 4-7	Selection intensity (Ka/Ks) and age of all gene pairs	94
Figure 4-8	Selection intensity (Ka/Ks) and age of pseudogene-gene pairs	95
Figure 4-9	Distribution of Ks for gene pairs	96
Figure 4-10	KOG classes of gene duplicates	98

Figure 4-11	All gene pairs along Ks in four spatial categories	100
Figure 4-12	N50 gene pairs along Ks in four spatial categories	101
Figure 4-13	Number of <i>cis</i> and <i>trans</i> duplicates over time.	102

ABSTRACT

GENETIC VARIATION WITHIN THE *DAPHNIA PULEX* GENOME

by

Abraham Eaton Tucker

University of New Hampshire, May, 2009

Genetic variation within the diploid *Daphnia pulex* genome was examined using a high quality *de novo* assembly and shotgun reads from two distinct *D. pulex* clones. Patterns of variation and divergence at single nucleotides were examined in physical and functional regions of the genome using comparative assembly output and available annotations. Additionally, mitochondrial genomes of the same *D. pulex* clones were assembled and compared for patterns of divergence, and substitutional biases. Intron presence/absence polymorphisms were identified computationally and verified experimentally. Finally, gene duplicate demographics were examined for patterns of divergence and estimates of gene birth rates.

INTRODUCTION

Current state of genome biology

The scientific study of inheritance is entering the post-genomic phase. Over the past decade, whole genome sequences, from dozens of mammals, reptiles and fish to hundreds of viruses and bacteria have been published (www.genomesonline.org). Among ~100 eukaryotes published to date, chordates, fungi, nematodes and arthropods are among the most studied (Appendix A). The rapidly accelerating pace of genome sequencing, assembly and annotation has moved the field of genomics past a mentality of a “canonical” genome sequence for each species to a recognition that substantial genomic variation underlies the diversity within individuals and populations. When funding major genome projects, an early emphasis on the macroevolutionary trends of genome evolution led to the prioritization of phylogenetic breadth over population depth when choosing taxa. This may have led to a misconception of static genome structure within species and an underemphasis on intraspecific variation in genome analysis. From a purely computational point of view, genetic variation has been considered a problem and not an opportunity (Green 1997, Vinson et al. 2005). This has meant that many organisms are chosen for genome sequencing specifically because they lack variation (through artificial inbreeding and/or recently bottlenecked populations). The avoidance of genome projects with natural levels of genetic variation has exacerbated the all-to-common view of

canonical genomes, although the increasing affordability of whole genome sequencing is changing this.

The commonly quoted statistic that humans are “99.9% similar” belies the fact that much of the variation within species is not at single nucleotides, but at larger segments and sections that are modified, lost and gained (Kidd et al. 2008; Feuk et al. 2006; Redon et al. 2006; Khaja et al. 2006). Genetic variation comes in many forms, from single nucleotide polymorphisms to gene duplications and large-scale karyotype-level changes. In fact, there is evidence that genomes of individual humans can differ by hundreds of active gene copies (Nozawa et al. 2007; Zhang 2007; Young et al. 2008) and substantially more other segmental variants (Jakobsson et al. 2008; Tuzun et al. 2005; Nguyen et al. 2008). Similarly, the oft-quoted statistic that humans and chimps are 98.8% similar (The Chimpanzee Sequencing and Analysis Consortium 2005; Wildman 2003; Kumar and Hedges 1998; Eichler et al. 2004; Tishkoff and Kidd 2004) gives an overly conservative estimate of genetic divergence. Larger scale structural divergence is estimated to be many times that estimated for nucleotide substitution (Kerher-Sawatzki and Cooper 2007; Newman et al. 2005; Nguyen et al. 2008; Shianna and Willard 2006). With a greater appreciation for the many scales of genomic variation and the development of high throughput sequencing technologies (Mardis 2008; Wang et al. 2008; Wheeler et al. 2008), a new era of individual genome sequencing has begun. As genomes from related individual organisms are sequenced, it is clear that we have only started to understand the many forms of genetic variation and their consequences for genome evolution and biology.

Characterizing the mechanisms and forces driving genome evolution is a fundamental challenge for the field of biology. The first decade of whole-genome sequencing (~1995-2005) gave a glimpse of macroevolutionary trends in genome structure (Lynch 2007; Gregory 2005). In the context of a phylogenetic framework, the broad comparative genome approach has proved to be an informative and powerful strategy for cataloging genomic differences and similarities. However, the signatures of evolutionary events are quickly masked by subsequent divergence. For example, an extreme bias towards transition substitutions in animal mitochondrial DNA becomes less apparent as comparative distances increase, due to saturation. To preempt the erosion of signal in newly arisen genetic novelties, highly related genomes must be compared. For instance, newly arisen mutant alleles such as an intron loss (Llopert et al. 2002) or gain (Omilian et al. 2008) have more information about their origin and fate when discovered and described in the context of population genetic data. The mutational processes responsible for generating new variants (e.g. point mutation, micro insertion-deletion, duplication, recombination, transposon activity, segmental duplication and deletion), and the microevolutionary forces responsible for maintaining them (i.e. drift, selection) can best be described through the examination of genomic variation within individuals and populations. As we enter an era of population genomics, the microevolutionary perspective will help describe genomic variation soon after it originates. With a solid grounding in the principles of population genetics and with genomic data from closely related alleles in populations, the genome biologist can begin to more fully and specifically describe the mechanisms and forces underlying microevolution.

Population genomics and evolution

Theodosius Dobzhansky famously stated “Nothing in biology makes sense except in the light of evolution.” (Dobzhansky 1973). Recently, Lynch (2007) extended Dobzhansky’s statement to reflect the microevolutionary perspective when he wrote “Nothing in evolution makes sense except in the light of population genetics.” As more organisms are fully sequenced, analyzed and compared, the forces affecting genome content and structure can be detected, leading to an improved understanding of the processes of genome evolution. Principles of population genetics and molecular evolution contribute to predictions about how variation, from single nucleotides to genes and large segmental variants, are proliferated, maintained and purged from genomes. Understanding how forces of mutation, recombination, drift and selection act to shape the genome in the process of biological evolution requires a look at how variation, in its many forms, originates within a genome. From this perspective, comparing two genomes from the same or highly related species will be more informative than comparing phylogenetically distant taxa.

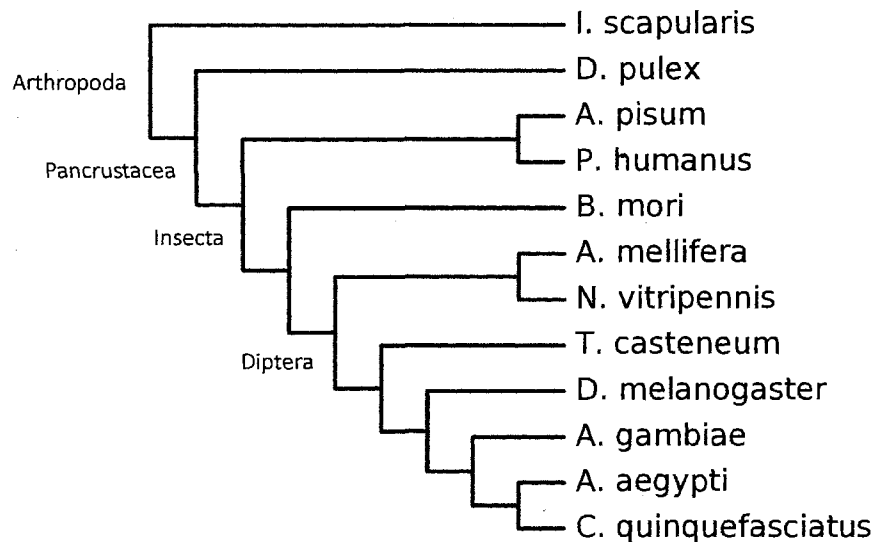
Population genomics is an emerging field that promises to deliver direct and practical insight into personal medicine and modern evolution (Jorde et al. 2001; Butlin 2008; Li et al. 2008; Begun et al. 2007; Tsai et al. 2008; Stranger et al. 2007). Association studies using genomic variation such as SNPs, copy number variants (CNVs), and structural polymorphisms are becoming increasingly common (McCarroll and Altshuler 2007; Iuliana Ionita-Laza et al. 2008; Hiroaki and Sato 2008; McCarroll 2008). Similarly, insights into recent evolutionary pressures on populations are being accelerated by comparative population genomic approaches that examine polymorphism on a genomic

scale (Begun et al. 2007; Hawks et al. 2007; Wang et al. 2006; Zhang et al. 2006; Sabeti et al. 2007; Tang et al. 2007; Williamson et al. 2007; Anisamova and Liberles 2007; Cutter and Payseur 2003). As biologists strive to describe and understand how the genome builds, develops and manages individual organisms, the differences between individual genome sequences can tell us much more than simple phylogeny, but can serve as a historical record of the evolutionary forces that act on organisms, helping us understand how and why genomes change over time. The population genomics paradigm promises a renaissance of environmental genomic research where the genetic basis of ecological specialization and adaptation can be elucidated through the examination of genomes from natural isolates as well as traditional ecological models (Stinchcombe and Hoekstra 2008; McKay and Stinchcombe 2008; Ungerer et al. 2008; Cooper and Lenski 2000).

The series of projects described below examine genetic variation across the *Daphnia pulex* genome from four different scales, but all from the perspective of population-level molecular evolution. In Chapter 1, SNPs in the nuclear genome are characterized and examined on a regional and functional basis. Chapter 2 outlines an analysis of patterns of substitution among recently diverged mitochondrial genomes from three *D. pulex* clones. Chapter 4 describes how intron turnover in genes was assayed using the assembly and shotgun reads of two related *D. pulex* clones. Finally, in Chapter 4, patterns of divergence between gene duplicates are quantified using the latest *D. pulex* gene predictions. These four aspects of variation arise from different mutational processes that contribute to the ongoing molecular evolution of the genome and give important clues to the microevolutionary processes that direct the evolution of the *D. pulex* genome.

Daphnia pulex genome

Arthropods are one of the most diverse and successful animal phyla with millions of species (Ruppert et al. 2003). Whole genome projects from Arthropoda, however, have been heavily skewed towards Insecta, an overwhelmingly terrestrial class that includes many disease model organisms as well as traditional genetics workhorse species (Figure I-1). Ongoing arthropod genome projects continue to emphasize the insects (Appendix B). *Daphnia pulex* was recently tapped as the first member of Crustacea, sister taxon to Insecta (Dunn et al. 2008), to be fully sequenced. *Daphnia* is also the first aquatic arthropod genome sequence. As an outgroup to the many complete insect genomes, *Daphnia pulex* will serve an important role in clarifying lineage-specific genetic novelties (Colbourne et al. 2007) and provide unprecedented opportunities for linking evolutionary ecology and genomics (<http://daphnia.cgb.indiana.edu/files/papers/WhitePaper.pdf>).



I-1: Phylogeny of completed arthropod genome projects as of 2008. Common names (from top to bottom) are tick, waterflea, aphid, louse, silkworm, honeybee, wasp. beetle, fruitfly, mosquito (last three). Many species of *Drosophila* have

been fully sequenced and not included here (Drosophila 12 Genomes Consortium 2007).

Commonly known as the “waterflea”, *Daphnia* are globally distributed microcrustaceans generally found in freshwater lakes and ponds and serve as keystone species in aquatic food chains as foragers of algae and bacteria and prey for carnivorous zooplankton and fish. *Daphnia* have a long history as an ecological model and are one of the most widely studied model organisms (Peters and de Bernardi 1987; Banta 1939).

The *Daphnia* system is unique among genomically characterized model systems for its combination of ecological tractability and vast history of ancient and recent evolutionary radiation (Colbourne and Hebert 1996). Through an examination of various scales of genomic variation, the *Daphnia* molecular toolbox will be expanded to prepare for the coming age of population genomics where natural genetic variation will be used to understand the basis of phenotypic evolution (Mitchell-Olds et al. 2007; Benfey and Mitchell-Olds 2008; Colbourne et al. 2000). Among animal models, *Daphnia* is quickly being established as a premier model system for evolutionary and ecological genomics (Feder and Mitchell-Olds 2003) and promises a unique chance to tie natural genomic variation to local ecological adaptation (Lynch 1983; Eads et al. 2007). Heavily studied by limnologists, ecotoxicologists and other ecologists, *Daphnia* are known to inhabit a wide variety of aquatic environments, from freshwater to saline, coastal to alpine, eutrophic to oligotrophic and temperate to arctic. Evolutionary studies have shown that *Daphnia* provide a rich model for understanding physiological and morphological diversification, convergence and adaptation (Colbourne et al. 2000). For instance, tolerance to toxic cyanobacteria (Hairston et al. 2001), hypersaline conditions (Hebert et

al. 2002), predation (Cousyn et al. 2001), acidification and metal contamination (Pollard et al. 2003) and other anthropomorphic disturbances (Weider et al. 1997) have been investigated using *Daphnia*.

Additionally, most lineages of *D. pulex* are cyclic parthenogens that alternate between asexual and sexual reproduction (Figure I-2, left). However, some lineages of *D. pulex* have evolved obligate asexuality, where parthenogenesis is the sole mode of reproduction (Figure I-2, right). The divergent reproductive modes of *Daphnia* make them useful models for studying the genomic consequences of recombination (Paland and Lynch 2006; Lynch et al. 2008).

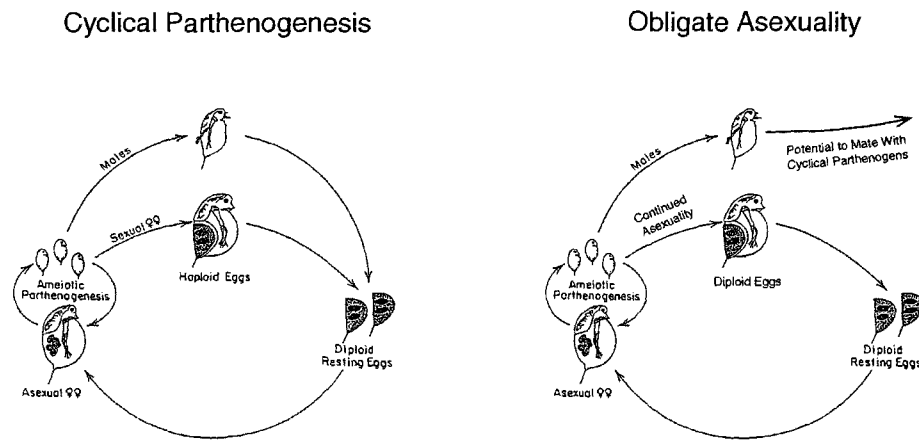


Figure I-2: Reproductive mode in *Daphnia*

Genomic data from *Daphnia* populations will enable genomicists and ecologists to combine forces to infer the genetic consequences of environmental disturbance, life-history (Dudycha and Tessier 1999) and other ecological forces that affect population parameters. Conversely, with genome in hand, the geneticist can provide candidate loci that may be evolutionarily and ecologically important (Li et al. 2008).

This study examines standing variation across the diploid genome of an individual microcrustacean, providing a platform to infer the nature of recent mutation, the fate of which is determined by immediate evolutionary forces. The evolutionary scale of this population genomics study provides the power to describe recent, ecologically relevant variation, a stated goal of the *Daphnia* Genome Consortium.

The 200 Mb *Daphnia* genome (clone TCO) was sequenced at 9X coverage using the whole genome shotgun approach (WGS) and Sanger sequencing using libraries of 3, 8 and 40 Kb. The first draft assembly contains 100 N50 scaffolds (the largest scaffolds that make up half the genome), with scaffold 100 containing 0.5 Mb of sequence (Figure I-3). The entire assembly contains 30,104 gene predictions, with $\approx 20\%$ of the genome coding for predicted proteins. An additional 1X coverage of another *D. pulex* clone (TRO) was sequenced, providing additional opportunities for comparative genome analysis. *Daphnia pulex* was chosen as the first crustacean genome to be fully sequenced, a decision motivated by the deep ecological understanding of the *Daphnia* system and its modest genome size. The TCO clone was isolated from an ephemeral pond along the Oregon coast and was chosen for its naturally low heterozygosity.

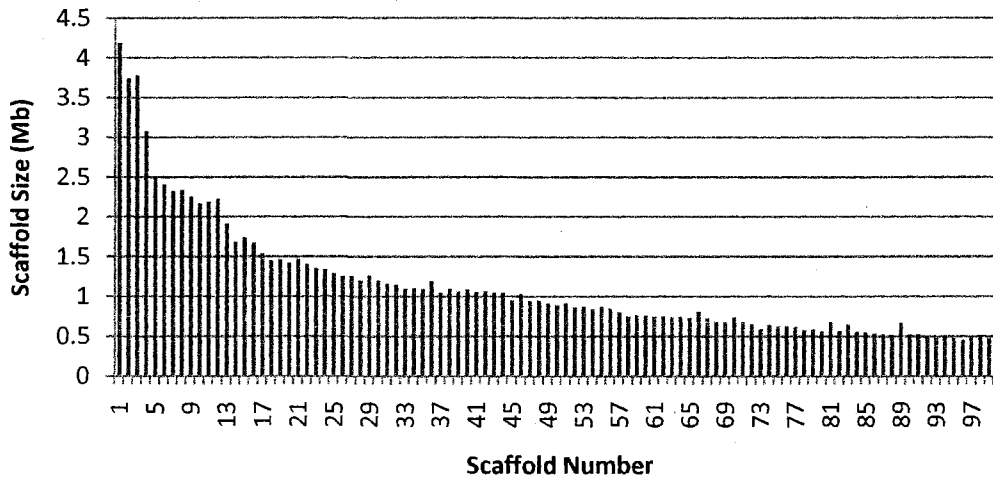


Figure I-3: Distribution of N50 scaffold sizes from JAZZ Assembly of *D. pulex*.

The *Daphnia* Genome Project is a collaborative effort of the *Daphnia* Genome Consortium (DGC) and the Joint Genome Institute (JGI). The Thomas Lab at University of New Hampshire is a founding member of the DGC and, along with Indiana University, Utah State University and Dartmouth College, has been a primary contributor to the development and analysis of the *Daphnia* genome sequence.

The following chapters describe a series of projects undertaken to describe genetic variation across the first draft *Daphnia pulex* genome. With many more *Daphnia* genomes to be sequenced in the near future, this research provides a springboard for further analysis of *Daphnia* genome evolution. A series of pipelines to systematically scan the genome for a number of variant types are described. The computational infrastructure developed for this project will be used for future analysis of *D. pulex* genomic data.

CHAPTER I

SMALL NUCLEOTIDE POLYMORPHISMS

Background

SNPs and evolution

Small nucleotide polymorphisms (SNPs) are a fundamental form of genetic variation within populations and are indispensable tools for genetic research. Allelic variation associated with phenotypic characters can be used to physically map loci responsible for traits of interest using SNPs as genetic markers (Lynch and Walsh 1998). SNPs are widely used as molecular markers in association studies used for positional cloning and medical diagnosis (Altshuler et al. 2008; Donnelly 2008; Hindorff et al. 2008). Additionally, SNPs are important for population studies and evolutionary research, as SNP patterns (haplotypes) are the basis of tracking gene flow, population structure and biogeography of populations (Tishkoff and Kidd 2004; Novembre et al. 2008; Gilbert et al. 2007; Lynch and Ritland 1999; Anderson and Weir 2007; Nei 1987; Weir 1996). For the purposes of this study, patterns of SNPs across the genome are used to consider recent evolutionary forces acting on the genome.

While the ultimate source of genetic variation is mutation, the maintenance and sorting of variation in a population involves the combined processes of genetic drift,

natural selection and recombination. The patterns of nucleotide polymorphism over a genome help to identify the magnitude and localization of these important evolutionary processes. For instance, regions of relatively low SNP density along a chromosome (“SNP deserts”) may result from recent selective sweeps that reduce variation at a locus (Cai et al. 2009; Andolfatto 2007). High rates of recombination, however, can lead to regions of relatively high SNP density and low linkage disequilibrium within a genome as higher crossover rates break up associations between alleles. Additionally, SNP patterns may provide insights into underlying substitution biases and may, ultimately, reflect mutational trends across a genome.

Observed levels of nucleotide polymorphism within species represent the fairly recent mutational events that have persevered through evolutionary filters of drift and selection. Their overall spectrum across the genome includes mutational events that range from highly deleterious to beneficial, with most substitutions being of the nearly neutral variety (Kimura 1983; Ohta 1992; Hughes 2008). Most of the polymorphisms detected in modest population samples involve alleles of intermediate frequency since rare alleles (< 1%) are difficult to sample. While the present analysis includes alleles that range the entire allele frequency spectrum of the population, it is not possible to infer the allele frequencies by sampling a single or few diploid genomes (i.e. all SNP frequencies are 50% in this analysis). However, by considering SNPs in different categories of genome function, we can begin to infer the recent evolutionary pressures acting on regions of the genome.

Neutral theory and population genetics

Neutral theory posits that most polymorphisms within species and fixed substitutions between species are the result of the random drift of nearly neutral mutations, rather than of natural selection (Kimura 1983). This view of the predominance of drift has persevered in the age of genomics, and underlies the statistical approaches to discovering rare instances of recent selection in the genome (Nielsen 2005; Biswas and Akey 2006; Tang et al. 2007). The relative role of drift and selection in shaping genome sequences continues to be an important and dynamic debate that will continue into the age of population genomics (Hahn 2008; Wagner 2008; Hughes 2008; Lynch 2007; Nei 2005; Bernardi 2007).

While neutral models of genetic evolution have been around since the Modern Synthesis (Wright 1931), a formalized neutral theory for molecular evolution was developed by Motoo Kimura, who modeled the dynamics of neutral mutations in finite populations using math from diffusion theory (Kimura 1955, 1964). The Neutral Theory emphasizes the effects of genetic drift over natural selection when considering the turnover of neutral and nearly-neutral mutations in populations (Kimura 1983; Ohta 1992). The formalized theory, applied to molecular evolution by Kimura in the late 1960s with elaborate mathematical justification (Kimura 1968; King and Jukes 1969; Kimura and Ohta 1971), gained legitimacy after the empirical study of gene products showed unexpectedly high levels of molecular variation within and between populations (Zuckerkandl and Pauling 1965; Lewontin and Hubby 1966; Harris 1966). The traditional school of thought had attributed intraspecific genetic diversity to balancing selection and had assumed that natural selection maintained the bulk of measurable variation (Ford

1965; Mayr 1963; Dobzhansky 1955; Lerner 1954). Kimura promoted an alternative view where mutations are rarely selected for, but instead are selected out (purifying selection) or are selectively neutral. With many nearly-neutral mutations being lost or fixed randomly, the continuous turnover of many mutations over time would be the basis for most of the genetic diversity in populations (Kimura 1983; Ohta 1992). With the proposition that nearly-neutral mutations might explain much of molecular evolution, neutral theory was tagged as “non-darwinian”, since Darwin had emphasized natural selection as the main force responsible for evolutionary change. But Kimura recognized natural selection as the main force for adaptation, stipulating that instances of positive selection at the molecular level are rare compared to allele fluctuations due to drift. In this sense, neutral theory is consistent with and fully integrates with neo-darwinism (Nei 2005).

Neutral theory does not suggest that most variants have equal fitness, but predicts that the fixation of allele variants is largely determined by drift and not by weak selection. For instance, an allele with mildly deleterious functional *effects* may still evolve neutrally. The neutral theory emphasizes the role of drift over selection more than neutrality of function. In fact, Kimura himself wrote that the theory may be better served being called the “mutation-random drift theory” since functional neutrality is not a prerequisite for fixation through drift (Kimura 1983, pg. xii). However, a fundamental implication of neutral theory is the primacy of population size in establishing the relative importance of drift and selection in evolving populations.

Eventual fixation of mutations by drift can occur even when there is a small selective force (where fitness is increased or decreased by proportion, s) acting on an

allele as long as $|s| < 1/2N_e$. Since the number of new mutants per generation is $2 N_e\mu$ and the rate of fixation is $1/2N_e$, it follows that the rate of substitution of a neutral allele (k) is $2 N_e\mu (1/2N_e)$, or $k=\mu$. In other words, the rate of substitution is independent of population size and equal to the neutral mutation rate (Kimura 1983). This makes intuitive sense when you consider that in small populations there are fewer mutations that are each more likely to be fixed. In larger populations, there are more mutations, but each is less likely to be fixed. In other words, the probability of fixation in a small population increases with the same magnitude that the number of mutations is reduced in small population. Even assuming mutation-selection balance, we should not expect the same magnitude of genetic variation in small and large populations, however. Since time to fixation is $4N_e$ generations (Kimura and Ohta 1969), larger populations will, all things being equal, have more standing variation at any given point in time. In a nutshell, these theoretical expectations provide the null hypothesis from which we begin to study variation in natural populations.

Because sequence diversity (π) is proportional to population size (N_e) and the mutation rate (μ), a population-mutation parameter that describes neutral sequence diversity, θ , can be used to estimate effective population size (N_e) and mutation rate (μ), since $\theta = 2N_e\mu$. θ must be inferred from observed levels of heterozygosity within populations where theoretically ideal conditions are rarely met, so a number of estimators have been proposed. For instance, silent site diversity (π_s), a measure of per site heterozygosity, is calculated from synonymous sites of protein-coding genes to minimize any purifying or adaptive selection that may interfere with a measure of θ . Recent population bottlenecks will depress all measures of θ across the genome, while variance

in θ across loci within a genome is attributable to local changes in mutation rate and/or recent localized selective pressures. Using these principles, the well-known HKA test (Hudson et al. 1987) uses within and between species variation to test for selection across multiple loci. If assuming a constant N_e among populations, loci that are outliers to the correlation of divergence and polymorphism are candidate targets for positive (lower polymorphism) or balancing (higher polymorphism) selection. In this way, the local measures of polymorphism and divergence can be used to scan the genome for candidate selective targets.

However unrealistic many of the assumptions of the model may be, population genetic analysis begins with the standard neutral model. This model assumes a randomly mating, demographically stable population where mutations are neutral across infinite sites. These ideal conditions serve as a null model from which violations are detected and other forces, such as local selection, mutation heterogeneity and recombination, are proposed.

Most mutations are quickly lost in large and small populations, where the probability of fixation of a neutral allele is equal to its initial frequency ($1/2N$ in diploids). While positive selection ($s > 0$) reduces the probability of a rapid initial exit of a beneficial allele, most new mutations, neutral or beneficial, are eventually lost (~30% chance of being lost in first generation!). Tightly linked nucleotide sites are transmitted across generations as a unit depending on rate of recombination (c). Linkage causes fixation rates of beneficial mutations to be lower and of deleterious alleles to be higher, since beneficial alleles will spread slowly when inhibited by the baggage of their genetic backgrounds, a phenomenon known as selective interference (Hill and Robertson 1966;

Birky and Walsh 1988; Comeron et al. 2008; Gordo and Campos 2006). Because drift dominates selection as effective populations decrease in size, the range of mutations that are effectively neutral is inversely proportional to N_e . In other words, selection can “see” a greater range of s (hence a larger proportion of mutations) in larger populations (Lynch 2006; Yi 2006; Lynch 2007).

The frequency of meiotic recombination (c) within evolving populations controls long-term genetic opportunities by modifying the effective number of independent selective targets in the genome. For instance, the maintenance of neutral and/or beneficial genetic combinations may be decoupled from the elimination of deleterious factors when alleles are shuffled during meiosis. By constantly trading alleles, a sexual population of individuals ensures that, over time, targets of negative selection are pressured independently from targets of positive selection. Genetic hitchhiking of neutral and deleterious alleles during selective sweeps increases the overall magnitude of drift within infrequently recombining populations (i.e. asexual or self-fertilizing lineages). Even in low recombining regions of fully sexual genomes, the increased role of hitch-hiking (genetic “draft”) reduces the efficiency of natural selection (Gillespie 2000; Gillespie 2004). In fact, *Drosophila* genomic regions with low levels of recombination show elevated levels of replacement substitution and intron divergence (Haddrill et al. 2007) as well as elevated gene expression levels, possibly due to the reduced efficacy of purifying selection, leading to looser regulatory control (Haddrill et al. 2008).

D. pulex is an ideal system for testing neutral expectations in natural populations because the *D. pulex* species complex contains well-studied populations distributed globally in semi-isolated freshwater lakes and pond systems. These populations range in

effective size from tiny (where a handful of diapausing eggs found new populations every season) to extremely large. The extensive history of ecological research on *Daphnia* populations means that demographic and biological characteristics can be brought to bear when analyzing population-genetic parameters. For instance, obligate asexual lineages of *D. pulex* are expected to have increased deleterious mutation accumulation due to lack of recombination and reduced N_e (Lynch 2008; Paland and Lynch 2006; Paland et al. 2005).

SNP studies

Over the past century, the neutralist-selectionist debate in molecular evolution has swung back and forth. The relative influences of natural selection and genetic drift on transient genomic features such as nucleotide diversity and gene duplication as well as ancient, enduring products of evolution such as the genetic code (Koonin and Novozhilov 2008; Massey 2008; Sella and Ardell 2006; Freeland et al. 2000) and the molecular clock (Wilson and Sarich 1969; Wilson et al. 1987; Hedges et al. 2003; Hedges and Kumar 2003; Ho and Larson 2006), have not been resolved with much certainty. Many of the predictions of population genetic models have been supported by empirical data. For instance, large populations tend to have more sequence diversity (Lynch 2006, Supplemental Table 3; Tishkoff and Williams 2002; Wilhelm et al. 2007; Sauvage et al. 2007) and while relatively few allelic variants have conclusively been shown to have beneficial (Tishkoff et al. 2007; Hoekstra et al. 2006) or deleterious effects (Palti et al. 2000), small populations tend to have more deleterious variation (Lohmueller et al. 2008; Cruz et al. 2008), consistent with theory. Neutral theory has been, and remains, an effective null hypothesis for studying molecular evolution. However, population genomic

analysis of polymorphism across multiple, related genomes will be the ultimate test of the validity of many neutralist claims. The first population genomic studies have called into question the assumption that most mutations are nearly neutral (Hahn 2007; Begun et al. 2007; Orr 2009; Cai et al. 2009). While the modern form of the neutralist-selectionist debate is nuanced and a disagreement over degree rather than wholesale worldview, it may be that “rampant nonneutrality”, like that found in *Drosophila* (Fay et al. 2002; Smith and Eyre-Walker 2002; Sawyer et al. 2003; Bierne and Eyre-Walker 2004; Shapiro et al. 2007) makes the current neutral model unrealistic. However, some have questioned the validity of popular comparative methods for detecting selection (Hughes 2007; Hughes 2008). It remains to be seen what inferences will be made as variation data accumulates for other genomes.

Many of the earliest eukaryotic genome projects were carried out on highly inbred lab organisms (e.g. *C. elegans*, *D. melanogaster*, *M. musculus*) and species with naturally low levels of polymorphism (e.g. *H. sapiens*). Because even moderate levels of heterozygosity can confound *de novo* assembly, individuals chosen for genome sequencing are often intentionally inbred. Even in cases where natural isolates are used for genome projects, low polymorphism individuals are preferred. Genome projects that possess natural levels of heterozygosity often produce lower quality assemblies (Holt et al. 2002; Vinson et al. 2005; The French-Italian Public Consortium for Grapevine Genome Characterization 2007), a significant problem for genomes from populations with high N_e . Polymorphism data are often collected from skim sequencing of diverged lineages after an initial high quality assembly is produced (Kasahara et al. 2007; The Honeybee Genome Sequencing Consortium 2006). However, as the desire for broader

sampling of natural populations increase, more attention has been paid to detecting variation inherent in the diploid genome projects themselves (Levy et al. 2007; Kim et al. 2007; Wheeler et al. 2008; Wang et al. 2008; Holt et al. 2002; Lynch 2008).

Not surprisingly, the deepest sampling of genome-wide diversity has been collected from *H. sapiens*, where the International HapMap Consortium has generated a database of over 3.9 million mapped SNPs from hundreds of individuals from geographically diverse populations (Frazer et al. 2007). While the central objective for the HapMap project is to develop SNP markers for biomedical studies, these data are a boon to the evolutionary biologist (Manolio et al. 2008). Using the SNPs generated by the HapMap project and other efforts (e.g. Perlegen, Hinds et al. 2005), studies have begun to identify genomic regions under selective pressures by using a variety of newly developed computational approaches (Voight et al. 2006; Williamson et al. 2007; Sabeti et al. 2006; Tang et al. 2007; Sabeti et al. 2007; Wang et al. 2006; Cai et al. 2009; Hawks et al. 2007; Wright and Gaut 2005; for reviews on genomic approaches for identifying selection, see Nielsen 2005; Biswas and Akey 2006; Anisimova and Liberles 2007; Jensen et al. 2007; Thornton et al. 2007; Pavlidis et al. 2008). These and other recent studies have pioneered a set of computational approaches that use the SNP and linkage disequilibrium data from HapMap to model the nature of recent molecular evolution at sites across the entire human genome. When variation and recombination rates are considered, haplotype sizes and frequencies can be used to test for the signature of recent positive selection since targets of a recent selective sweep will show up as large haplotype blocks that rise to high frequency (Sabeti et al. 2002).

SNPs in protein-coding genes have been used to infer selective pressures across the genome through an examination of the relative levels of silent and replacement polymorphisms and rates of silent and replacement substitution between lineages (Liu et al. 2008; Ellegren 2008). Relative levels of nuclear diversity at silent and replacement site in protein-coding genes reflect the relative power of selection (Graur and Li 2000; Nielsen 2005; Begun et al. 2007). Recent studies have attempted to use human polymorphism data to estimate the distribution of fitness effects of new mutations (Boyko et al. 2008).

While it takes substantial resources to apply population genetic tests on genomic data sets, the promise of locating genomic regions under recent selection is an exciting prospect to the evolutionary biologist. While much of the population genomic analysis is being developed with the massive and well-curated effort of human SNP discovery, other organisms traditionally favored by molecular evolutionary biologists have seen genomic surveys of variation put to use for understanding recent evolution. Much of the pioneering work on detecting natural selection using population genetic data was developed on the *Drosophila* model system (Hudson et al. 1987; McDonald and Kreitman 1991; Tajima 1989; Akashi 1995; Kreitman and Akashi 1995). Recently, genome-wide SNP distributions were used to detect selection in *Drosophila* (Begun et al. 2007) and *C. elegans* (Cutter and Payseur 2003; Cutter et al. 2006).

Since the advent of whole genome sequencing, SNPs have been detected within diploid genome assemblies (Levy et al. 2007; Wang et al. 2008), between lineages (The Honeybee Genome Consortium; Lindblad-Toh et al. 2005; Kasahara et al. 2007; Cruz et al. 2008; Wayne et al. 2007) and from ESTs (Cheng et al. 2004). New approaches for

estimating levels of polymorphism from ESTs (Long et al. 2007) and genome assemblies (Hellmann et al. 2008; Lynch 2008) are being developed to cope with the influx of large genomic data sets.

The increasing affordability of whole genome sequencing has expanded the taxonomic sampling of established and non-traditional evolutionary model organisms. As serious population genomic studies become possible, there is hope that the power to relate population-level evolution to ecological circumstances has arrived. With a long history of ecological research, the microcrustacean *Daphnia pulex* genome is a unique resource for discovering ecologically relevant variation.

Daphnia variation

Daphnia pulex was chosen for whole genome sequencing based on its proven utility as an ecological model organism. The potential to decipher ecologically relevant genetic variation has been a selling point of the *Daphnia* model. Here, we outline a series of steps used to detect SNPs in *Daphnia pulex* by generating conservative estimates of variable sites on a scaffold-by-scaffold basis. Because the genomic DNA for the *Daphnia* Genome Project was prepared from a clonal population started from a single, low-heterozygosity individual, this study is equivalent to an assay of heterozygosity within an individual daphniid. However, heterozygous sites within a diploid individual represent segregating alleles of the larger population, and thus, with this first pass of SNP detection, we are able to describe some patterns of genetic variation across the whole genome of the species *Daphnia pulex*.

The source of the *Daphnia* genome sequence used in this study (clone TCO, “The Chosen One”) comes from a geographically isolated sexual, diploid population with reduced long term effective population size relative to other populations (Omilian et al. 2008; M. Lynch, personal comm.) The TCO clone was chosen among natural isolates for its relatively low heterozygosity, possibly attributable to its history of population bottleneck. The population-genetic implications for a genome from a clade with a considerably smaller long term effective population size invites future genome comparisons with other *D. pulex* lineages. Using another natural isolate, TRO (“The Rejected One”), we were able to compare relative levels of variation across the genome. TRO was isolated from the core *D. pulex* group and is substantially diverged from TCO.

In order to identify small nucleotide polymorphisms (SNPs) within the *Daphnia pulex* genome, a pipeline of analyses that uses the comparative assembly of whole genome shotgun reads against reference scaffolds was developed to conservatively estimate sites of true polymorphism within TCO, the clone of the *Daphnia* Genome Project. This study offers a first pass of the genome-wide level of polymorphism and identifies a large number of variable sites in the ecological model organism, *Daphnia pulex*.

Methods

Comparative Assembly (TCO)

The *Daphnia* Genome Project produced 2.7 million reads with an average length of 1011 base pairs (bp). The trace files containing the raw sequence reads were downloaded from NCBI (<http://www.ncbi.nih.gov/Traces/trace.cgi?>) and trimmed for vector and quality using the LUCY program (Chou and Holmes 2001). The output from LUCY was trimmed using the Perl scripts *lucyTrim* and *lucyTrimQual* (Appendix C). The LUCY program purges vector sequence and identifies optimal trimming points for the 5' and 3' ends of each raw read based on quality score information. We used the default LUCY parameters and the vector sequence of pUC19 as input to LUCY.

Quality-trimmed TCO shotgun reads (~9X) were assembled against the 100 largest scaffolds (N50~100) of the latest *Daphnia* JAZZ assembly (Draft 1.0) using AMOS Comparative Assembler (Pop et al. 2004) on Fedora 9 using a Dell Dimension 9200 with a 2.40 GHz Intel Core 2 Duo processor and 3 Gb RAM. After assigning each shotgun read to a scaffold using a best blast hit filter, a 98% minimum pair-wise identity cutoff was applied for the AMOS comparative assembly. Assembly information used to detect SNPs was generated by AMOScmp output files for each scaffold assembly.

Steps were taken to minimize the contribution of highly paralogous regions in the SNP analysis (through a coverage filter and best-blast assignments) and to include sites with a low probability of sequencing error (quality trimming and double evidence criteria). For these and other reasons, this analysis excludes up to a third of sites in the N50. While exclusions were mostly due to undetermined sequence in the reference scaffolds (Ns), the effects of sequencing error and poor sequence quality were minimized

with initial quality trimming and a rejection of ambiguous sites. Therefore, the SNP calls reported here have a high probability of being true sites of allelic variation, but underestimate the magnitude of genomic variation. A correction for undersampling is discussed later.

In order to identify SNPs in the *Daphnia pulex* genome, a stringent set of criteria for defining variable sites as true SNPs was implemented. The criteria are outlined below:

1. Trimming of raw sequence reads to improve average quality scores and purge ambiguous data and vector sequence.
2. Best blast hit filter to assign reads uniquely to scaffolds.
3. Assembly of trimmed reads to reference scaffolds with a minimum 98 percent pair-wise match requirement.
4. Rejection of sites with excessive coverage to minimize identification of variable sites due to paralogous misassembly.
5. Exclusion of variable sites that contain more than two types of nucleotides.
6. Counting only SNPs with at least two reads of each variant type of nucleotide.

Since the accuracy of the subsequent SNP analysis depends on the average quality of the input sequence information, the reduction of information due to trimming gave us more confidence for all SNP calls by minimizing low-quality base calls.

Based on the AMOS assembly and the delta file, the number of reads that occurred at each base was calculated. The percent coverage across each scaffold of the reference sequence was calculated using the Perl script *avgCoverage*.

A binomial probability distribution can be approximated by Poisson when $N(x)$ is large and $p(x)$ is small. Assuming the absence of a strong cloning bias for a given genomic region, the sequence coverage of reads in a shotgun genome project follows a

Poisson distribution. The probability (P) of a given nucleotide being sequenced x times based on an average genome coverage of λ is given by :

$$P(x, \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}$$

Using the estimated average coverage, we calculated the expected coverage probability as a function of x , the number of sequence reads at any base. Using this distribution, we determined the value of x at which 99% of the genome would be covered at least once. This defines an upper bound to x , x_{max} . Regions where the coverage exceeds this value are more likely to have extra coverage due to the alignment of reads from paralogous regions from elsewhere in the genome. These regions were therefore excluded from further analysis. Simple sequence repeats (homopolymers and microsatellites with 8 repeats or greater) were excluded from our SNP analysis since variation at simple sequence repeats are a unique category of polymorphism that are being studied in a separate analysis (Sung et al. in prep). Additionally, regions in the scaffold where blocks of undetermined sequence are located (represented by Ns) were also excluded.

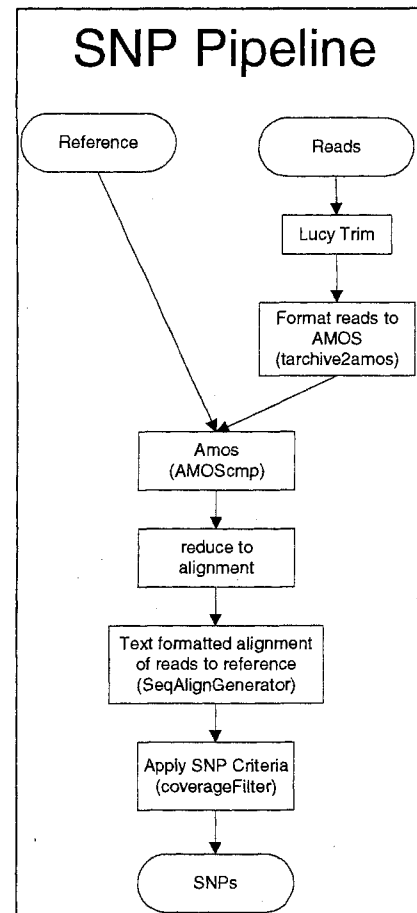


Figure 1-1: Flowchart of SNP pipeline

A site-by-site list of the nucleotides that assembled to each location in each reference scaffold was generated using the Java program, *SNPfinder* was generated. The output includes the reference base and a list of nucleotides that aligned to the reference at each site.

With the goal of identifying true SNPs, stringent criteria that minimized false assignment due to paralogous assembly or sequencing error were implemented. The Perl script *SNPFilter* generates data from loci with maximum and minimum numbers of aligned nucleotides with user specified SNP criteria. For instance, additional criteria also required that a SNP have at least two nucleotides of exactly two variant nucleotide types in order to reduce SNP calls due to sequencing error. SNP determination was completed by counting the number of base substitutions or indels per site. The output of the program *SNPFilter* contains the SNP locations from our scaffold, the reference base call, and the bases within the trimmed reads at each reference site.

Perl scripts to analyze SNP variation within the data were created. In order to view SNP variation across the scaffolds a Perl script *SNPvariationWindow* was used, which enables us to view variation at different window sizes. We analyzed our data using a 100-100,000 bp window sizes. High, moderate and low SNP density regions we examined to test for paralogy using BLAST against all the reference scaffolds.

The Perl script *kindOfSNP* outputted the totals of all SNP types (i.e. transitions, transversions, and indels). Sequential SNPs were found using the script, *clusterSNP*. We produced an output file that determined how many SNPs were in clusters of two, three, four, etc.

Based on the window analysis, there were two classes of SNP bins, one with no SNPs and one with SNPs. To test whether the two classes of SNP bins were randomly distributed we tested for significance using a normal approximation for the number of runs (r) where μ_r is the mean and σ_r is the standard deviation. Sokal and Rohlf (1981, pgs. 782-787) proposed a runs test which calculates the standard deviation of r :

$$t_s = \frac{r - \mu_r}{\sigma_r} = \frac{r - [2n_1n_2/(n_1 + n_2)] - 1}{\sqrt{[2n_1n_2(2n_1n_2 - n_1 - n_2)] / [(n_1 + n_2)^2(n_1 + n_2 - 1)]}}$$

Here n_1 is the number of bins with SNPs and n_2 is the number of bins with no SNPs. If $t_s > 1.96$, then the distribution differs statistically from a random assortment of the two bin classes.

Comparative Assembly (TRO)

Shotgun reads from *D. pulex* clone TRO (1X) were quality-trimmed and assembled to the TCO draft assembly N50 scaffolds. Based on the preliminary distribution of blast hits (Figure 1-2), average divergence of TRO reads was estimated at 4.5%. Using the expected distribution around this average, the minimum pair-wise identity for comparative assembly was set at 90%. TRO reads were assigned to the N50 scaffolds using a best-blast test for unique placement and assembled using the AMOScmp (described above).

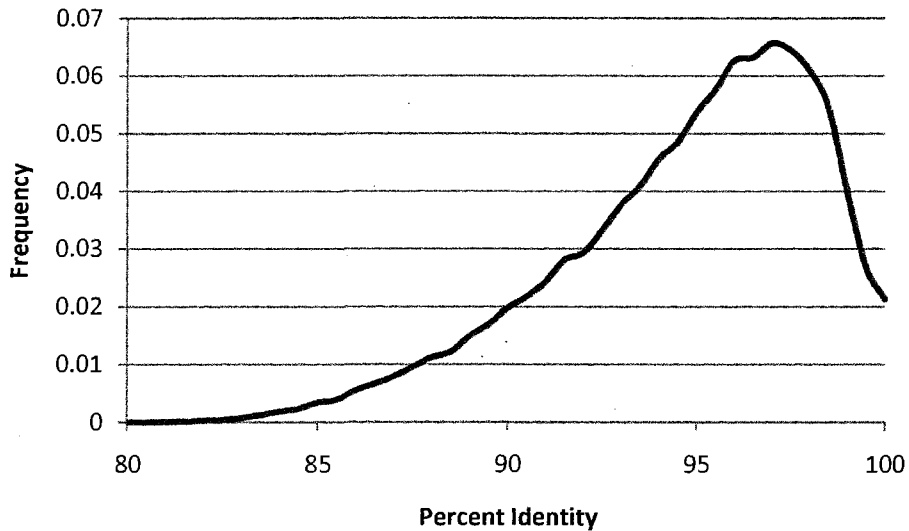


Figure 1-2: Distribution of genetic identity of TRO reads mapped to the TCO assembly.

Analysis of the TRO-TCO comparative assembly was carried out using a combination of custom scripts. Because of the low coverage, we considered sites containing 2-4 X coverage to examine heterozygosity (within TRO) and 1-4X to measure divergence (between TRO and TCO). Results are described below.

Results and Discussion

Pre-assembly quality control

LUCY trimming reduced the number of raw TCO shotgun reads from 2.7 to 2.5 million and cut our average sequence length to 774 bp (Table 1-1). The reduction of total sequence information (28.6%) after vector and quality trimming is proportional to that of other large data sets (unpublished data). Trimming the raw reads increased average quality score to 40, reducing the expected probability of sequencing error to 0.0001 per site.

	# reads	avg. length	QS	Total sites	% reduction
Raw shotgun reads	2,724,771	1,012	31	2,756,774,088	n/a
LUCY-trimmed reads	2,542,760	774	40	1,968,938,346	28.6%

Table 1-1. Pre and post LUCY-treated data. Average quality scores (QS) improved nearly 10X.

Comparative Assembly (TCO vs. TCO)

Comparative assembly of 9X shotgun reads against N50 scaffolds of the *Daphnia* JAZZ Assembly 1.0 produced 100 separate assemblies ranging from 6.6X to 9.9X average coverage (Figure 1-3). Overall, the comparative assemblies produced an average coverage of 8.8X (Figure 1-4), slightly less than the predicted sequencing coverage of the raw reads (9X). This is reasonable considering the error associated with estimating genome size and other factors such as non-assembled reads reducing actual coverage, contamination sequence (non-*D. pulex* DNA), and an edge effect which depresses comparative assembly at the ends of scaffolds and near gaps between contigs. Additionally, highly diverged alleles (>2% different) will fail to assemble. These factors all contribute to depressing coverage in the comparative assembly.

General platykurtosis of the actual coverage distribution is mostly due to the enrichment of low coverage sites (Figure 1-4, left tail), which may be due to single allele assembly in some regions and possible paralagous assembly in others, although to a lesser degree. While most N50 scaffolds assembled with normal average coverage, 4 scaffolds had particularly low coverage (scaffolds 30, 71, 93 and 98). These scaffolds

also have relatively high polymorphism levels (Figure 1-8). Our criteria may have excluded a higher proportion of reads from these assemblies where allelic divergence was $\gg 2\%$.

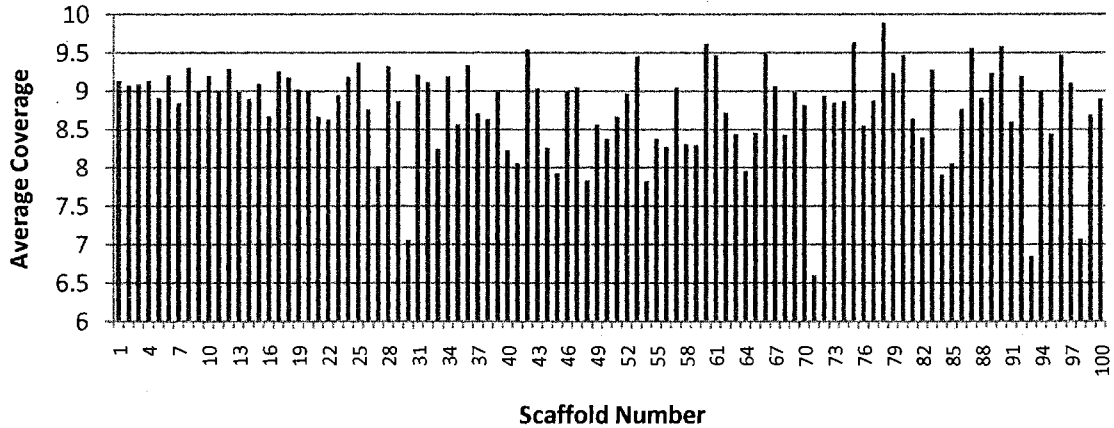


Figure 1-3: Average coverage of assemblies ranged from $\sim 6.5X$ – $\sim 9.9X$

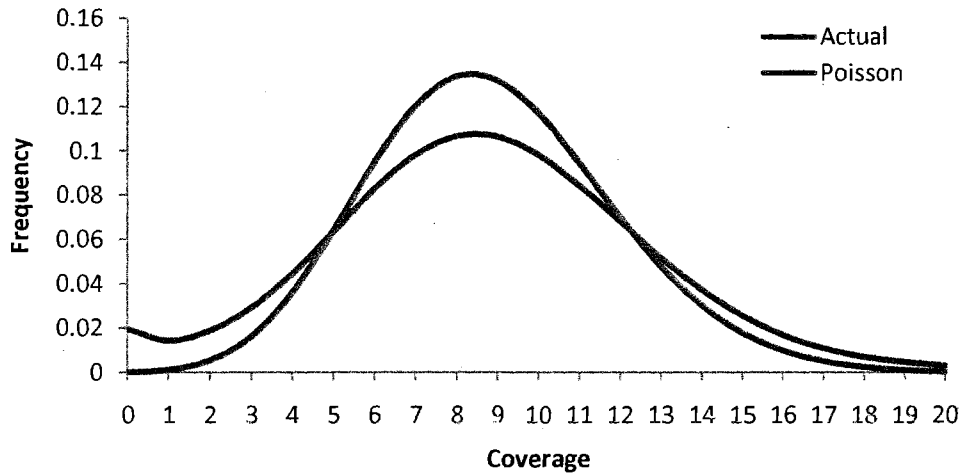


Figure 1-4. Frequency of coverage for all sites of N50 comparative assembly of TCO. (Average = 8.8X)

Scope of study

This analysis sampled roughly 40% of the 200 Mb *Daphnia pulex* genome after discarding sites that failed to satisfy our criteria (Figure 1-5). Because we began the

analysis using the best assembled half of the genome (N50 scaffolds), there is reason to believe that we sampled a biased set of sites, avoiding highly repetitive regions or areas that failed to produce continuous stretches of unique sequence in the *de novo* assembly. However, just over 50% of predicted genes are in the N50, indicating that the best half of the assembly is not biased with respect to gene density.

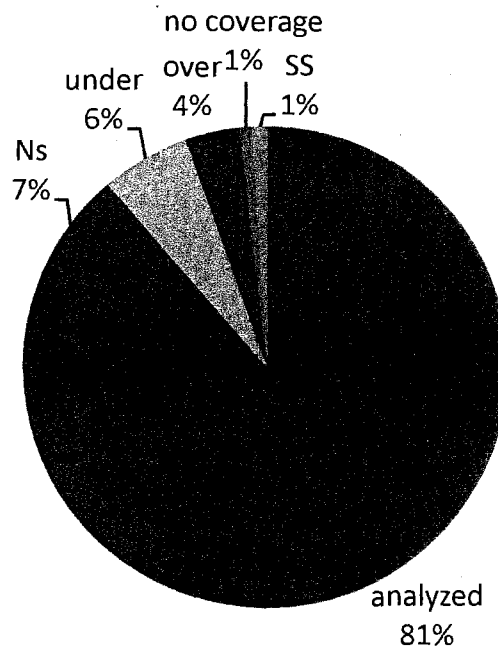


Figure 1-5. Breakdown of N50 sites under our criteria from the comparative assembly. We analyzed roughly 40% of the *D. pulex* genome ((200 Mb x 0.81 x 0.5)/200Mb). “Ns” refers to undetermined sequence in scaffolds. “Under” includes sites with 1-3X coverage, “Over” with >16X. “SS” refers to simple sequences with >8 repeats.

Magnitude of variation

The observed average heterozygosity for single base substitutions in TCO was 0.00101 per site across the genome. The average heterozygosity of TRO is much higher at 0.0144 per site. This enormous difference in nucleotide diversity shows that intraspecific lineages can range in natural levels of polymorphism by over an order of

magnitude. Average divergence between TCO and TRO at single nucleotide sites was 0.02499 per site. Single insertion-deletion polymorphisms contribute significant variation in both clones, adding 0.000268 in TCO and 0.0155 in TRO. For all observed levels of heterozygosity within TCO, the estimates are downwardly biased by at least 12.8% (based on expected coverage) and 17.9% (adjusted for actual coverage), due to binomial undersampling alone. For TRO, where only sites between 2-4X were used to estimate heterozygosity, 11.4% undersampling is estimated.

SNP types

A majority of the polymorphisms detected in this analysis are single site differences (68%, Figure 1-6). Base substitutions were classified into six types, according to the base substitutional matrix (Figure 1-7). Most indels were part of larger insertion-deletion events. For the range of sequential indels detectable in this analysis, average indel size was 3.5 bp.

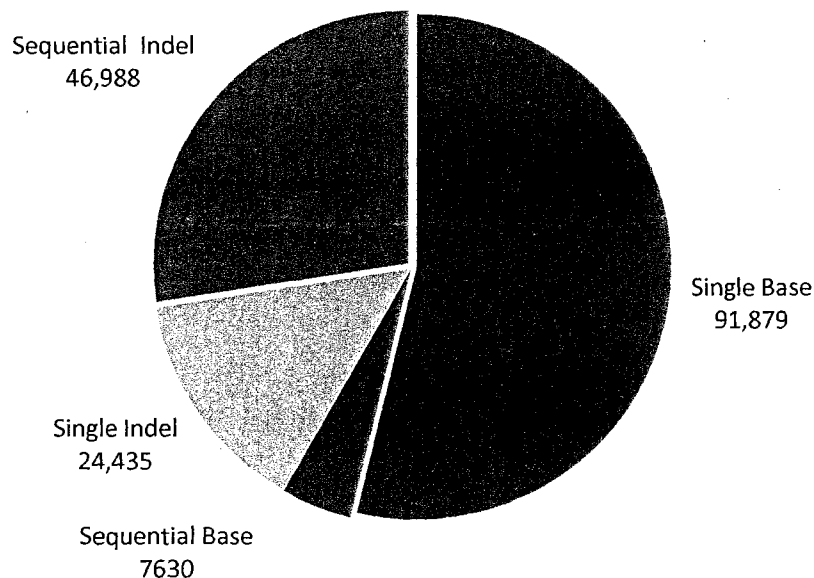


Figure 1-6. 170,932 polymorphic sites fall into 4 major categories.

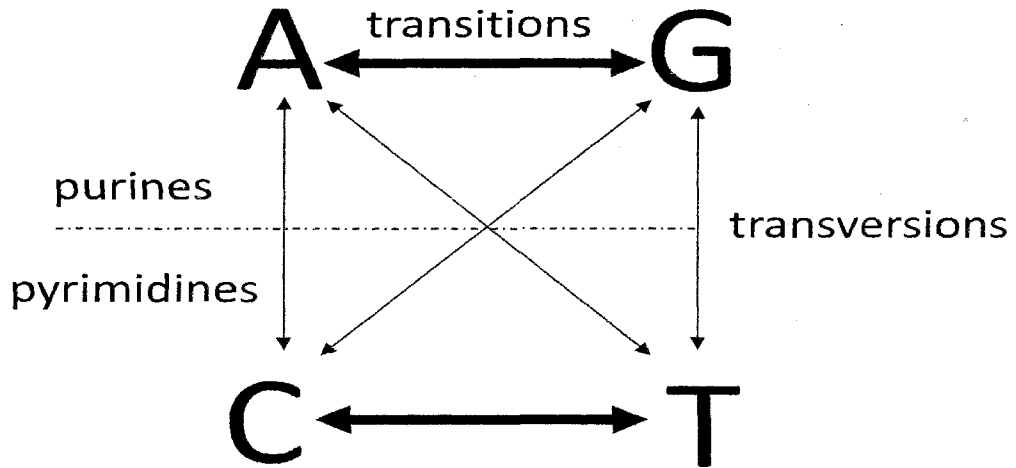


Figure 1-7: Base substitution matrix.

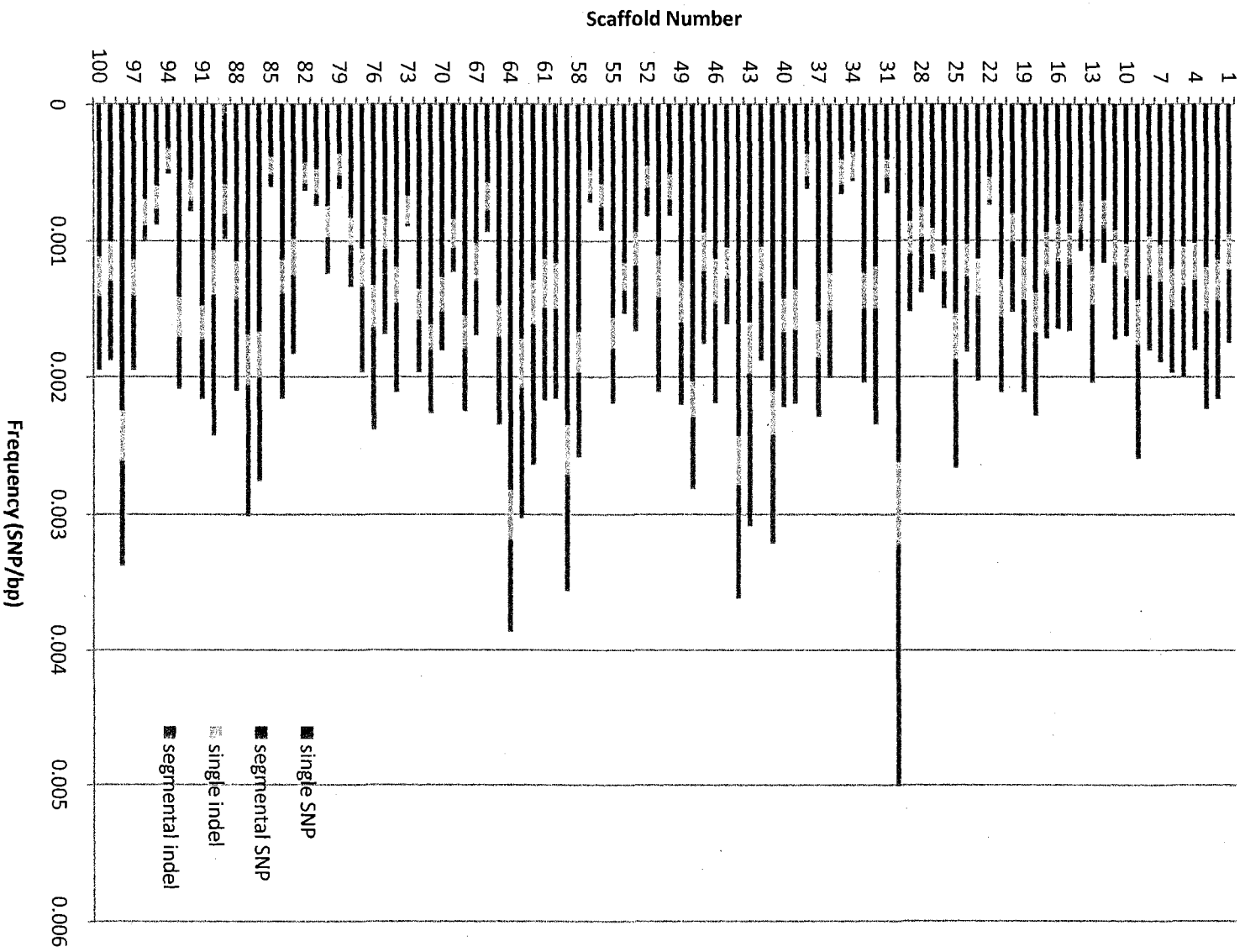


Figure 1-8: Frequency of SNP types by scaffold.

The frequency of small sequential indels (2-20 bps) across the genome follow a negative exponential distribution (Figure 1-9). However, sequential indels are more common than would be expected if each indel substitution occurred independently. For instance, using the observed single indel frequency of 0.000268, the expected number of sequential indels of size 2 would be 7. For sequential indel size 3, the expected number of observation would be 0. It is therefore likely that most sequential indels arose from single mutational events. The distribution of sequential indel substitutions observed in *D. pulex* is similar to that observed in *C. elegans* (Solorzano et al. in prep). Sequential base substitutions (Figure 1-10) are likewise thought to be part of larger mutational events. There may be an ascertainment bias in the observed distribution of sequential SNPs since the likelihood of detection is expected to decrease as larger segments fail to assemble under our strict criteria. However, it is clear that sequential SNPs, especially indels, are a prevalent class of polymorphism in the *D. pulex* genome (Figure 1-6).

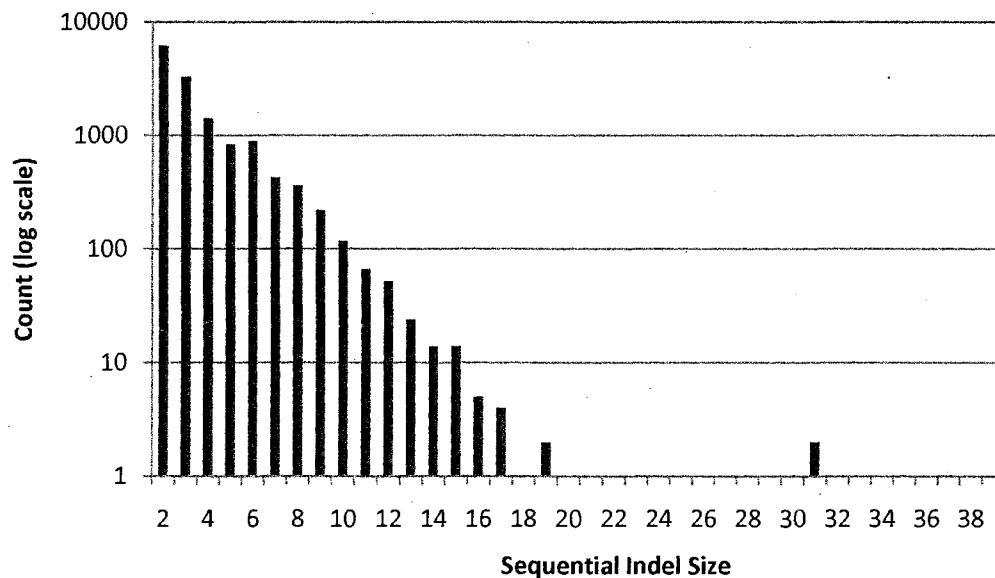


Figure 1-9: Distribution of sequential indels found within TCO.

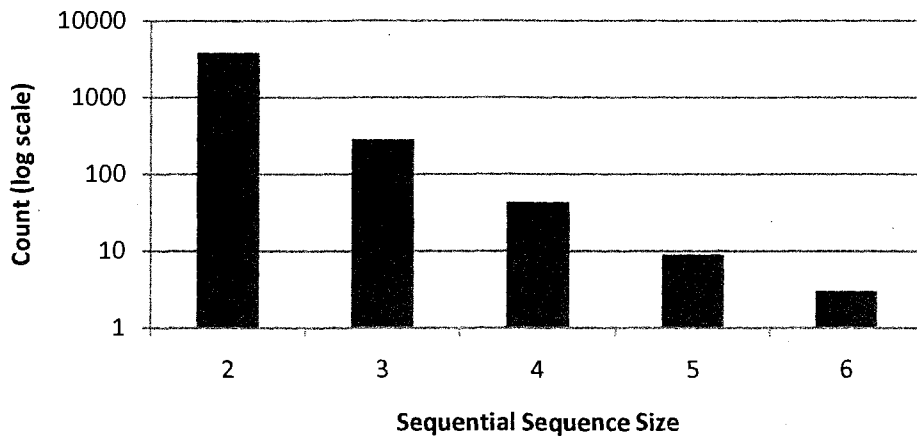


Figure 1-10: Distribution of sequential base substitutions within TCO.

From an analysis of base substitutions, we found that roughly half are transitions (Figure 1-11). Overall transition/transversion ratio (Ts/Tv) ranges from 0.468 to 1.196 across the scaffolds, for an average of 0.923 (Figure 1-12). If all base substitution types occur at equal rates, the expected Ts/Tv should be 0.5 (2 types of transitions/4 types of transversions, Figure 1-7). Transition bias (Ts/Tv >0.5) is widely observed among metazoan nuclear DNA comparisons (Jiang and Zhao 2006; Cargill et al. 1999; Gojobori et al. 1982; Collins and Jukes 1994; Lindblad-Toh et al. 2000; Rosenberg et al. 2003) and thought to be driven by underlying chemical properties of DNA that favor transition mutations, specifically the deamination of cytosine (Wakeley 1996). Zhao et al. (2006) found Ts/Tv to be related to GC. A few studies have found no transition bias, depending on the type of sequences examined (Keller et al. 2007; Moriyama and Powell 1996).

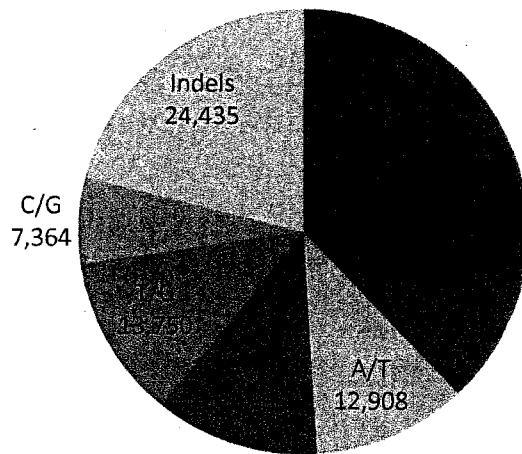


Figure 1-11. Frequency of SNP types in N50 scaffolds. Ts/Tv=0.923

Among *D. pulex* scaffolds, transition bias positively correlates with SNP frequency (Figure 1-13, $r^2=0.35$, $p < 10^{-11}$), but not GC content. If Ts/Tv purely reflects a mutational bias, variation in Ts/Tv may reflect the mosaic nature of mutation across the genome. However, if selection drives observable substitution patterns, fluctuations in Ts/Tv may reflect heterogeneity in selective regimes among scaffolds. Even with an extreme bias towards transition substitutions, the observed Ts/Tv value for two or more DNA sequences may be affected by a saturation phenomenon. For instance, the signature of past and future transitions are erased by occasional transversion substitutions, eroding the detectable transition bias over time. For this reason, the true transition bias can only be measured by looking at highly related sequences where few sites have multiple hits (i.e. intraspecific). This study provides a genome-wide view of substitution patterns, but may not fully minimize the effects of repeat substitution. Other SNP studies have found Ts/Tv to be *negatively correlated* with polymorphism (Solorzano et al. in prep),

suggesting that high SNP regions may be older and influenced by the transversion-saturation phenomenon. The positive correlation found here suggests other factors may be at play (e.g. mutation heterogeneity, wide selection on substitution type).

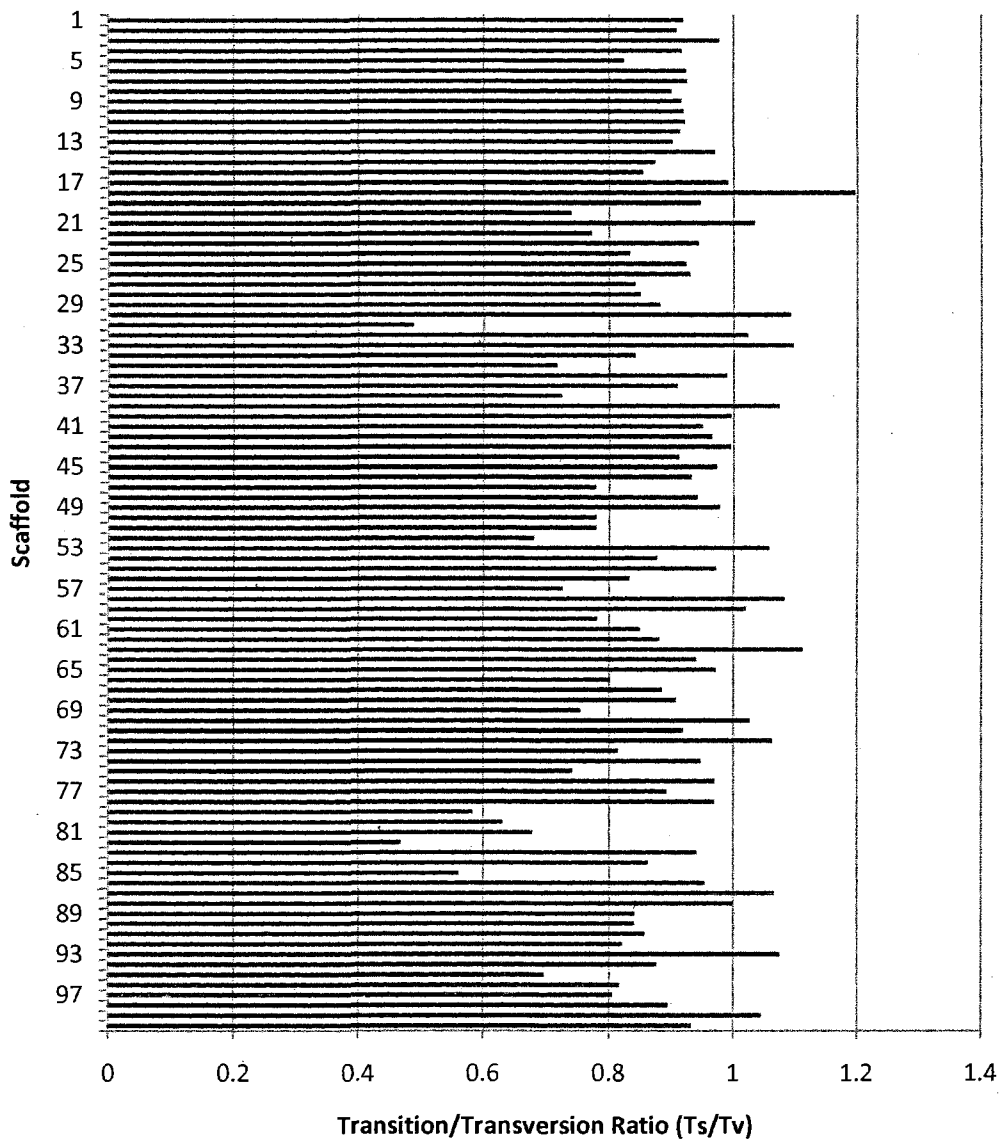


Figure 1-12: Transition/transversion ratio (Ts/Tv) for TCO polymorphisms by scaffold.

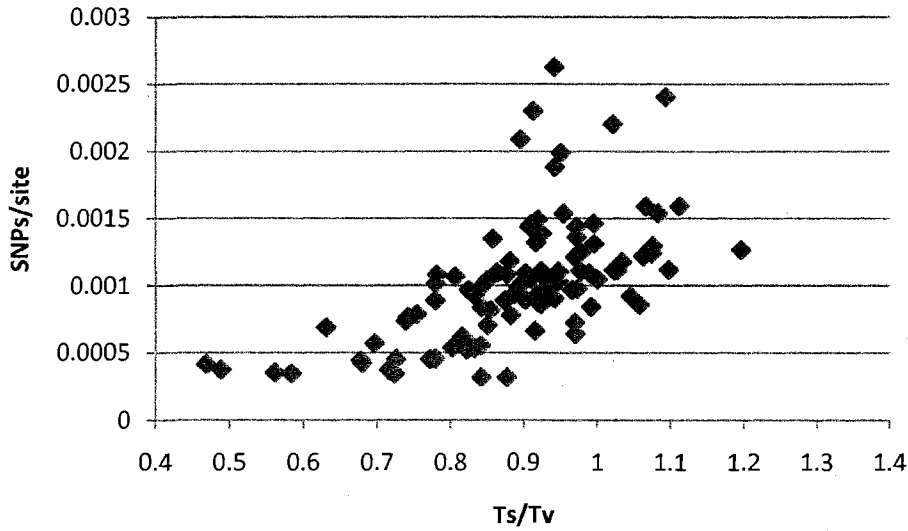


Figure 1-13: Polymorphism and Ts/Tv ratio. ($r^2 > 0.36$, $p < 1 \times 10^{-11}$).

The distribution of specific transversion types follows expectations from complementary base pairing rules and genomic base composition (Table 1-2).

Tv type	Expected	Observed
A/T	0.3502	0.2702
A/C	0.2415	0.2879
T/G	0.2416	0.2878
C/G	0.1666	0.1541

Table 1-2: Types of the transversions (Tv) over 100 scaffolds meet expectations based on base composition (A=0.2959, T=0.2959, C=0.2041, G=0.2041) in *Daphnia pulex*. (chi sq. test, $p < 0.002$)

Functional distribution

SNP types were examined separately in exons, introns and intergenic sequences across the genome. Base substitutions and indels were more frequent in non-coding sequence and transition bias was more pronounced in exons (Figure 1-14), a signature of selection against replacement substitutions.

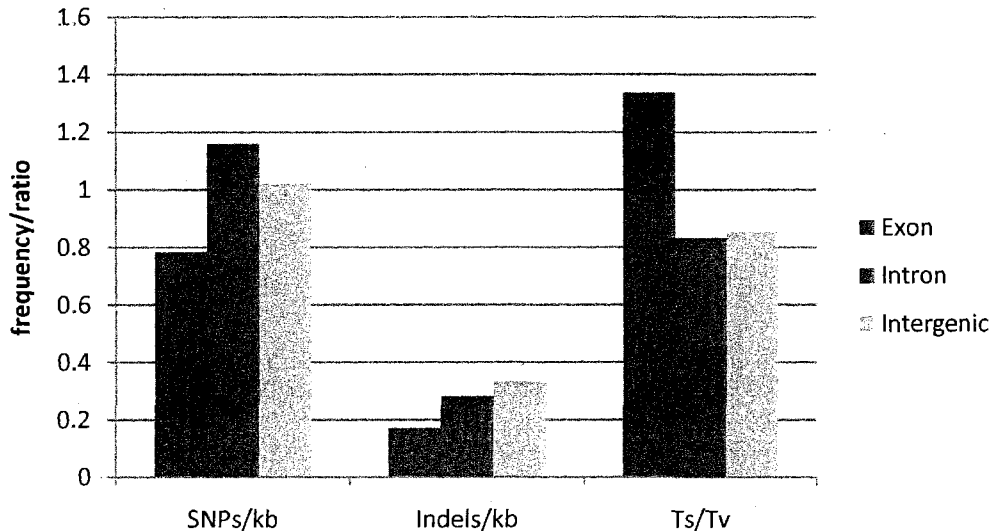


Figure 1-14. Frequency of base substitutions and indels in three major functional categories of sequence type (base substitutions/1000bp) and transition bias (Ts/Tv).

Overabundance of base substitutions in third positions (Figure 1-15) may result from selection against replacement substitutions (Kimura 1977), which mostly occur at second and first positions of codons. Replacement to silent substitution ratio (R/S) was 1.2, well below a neutral expectation of ~ 3 , suggesting overall purifying selection in protein-coding sequence among segregating alleles. Humans, *C. elegans* and cichlid fish are estimated to have genome wide R/S of 0.8 (Liu et al. 2008), 1.3 (Solorzano et al. in prep) and 1.54 (Loh et al. 2008), respectively. Evidence for purifying selection on synonymous sites is mounting (Chamary et al. 2006; Parmley et al. 2006; Resch et al.

2007). Reis and Wernisch (2009) relate levels of codon usage bias (translational selection) to expression levels in a pan-eukaryotic study.

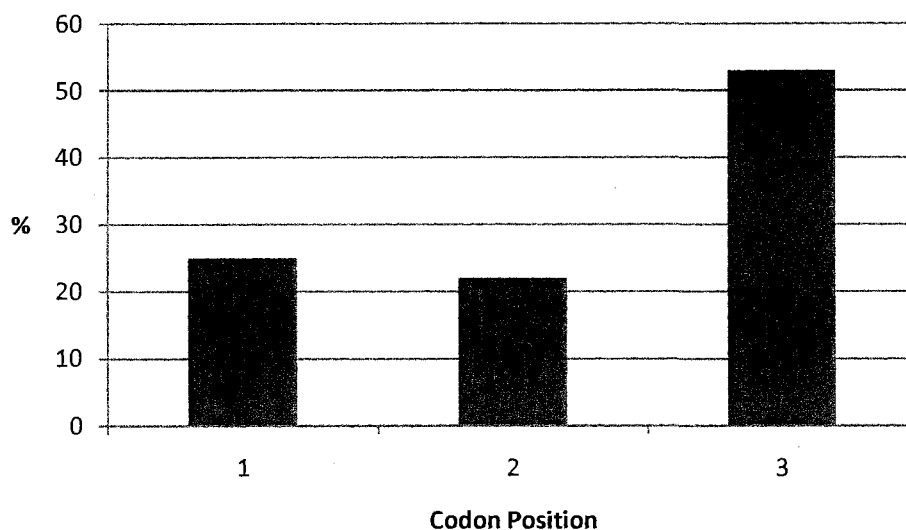


Figure 1-15. Distribution of base substitutions in exons.

Physical distribution

In order to understand how polymorphism varies across regions of the genome, data were analyzed on a scaffold-by-scaffold basis. Scaffolds varied in overall SNP frequency from 0.32 SNPs/1000 bp to 2.63 SNPs/1000 bp (Figure 1-16). While no scaffolds showed significantly low overall polymorphism levels, 5 scaffolds had particularly high SNP rates (scaffolds 30, 44, 59, 64 and 98, Figure 1-16). Scaffold 30 and scaffold 64 were mapped to chromosomes 2 and 3, respectively, suggesting that high SNP regions do not necessarily map to the same chromosomes.

The source of high relative rates of polymorphism within a genome can be boiled down to few general possibilities. Mutation rates may be heterogeneous (Baer et al. 2007; Fox et al. 2008; Gaffney and Keightley 2005; Malcolm et al. 2003; Wolfe et al. 1989).

Exceptionally high recombination rates would minimize polymorphism-clearing effects of hitch-hiking (i.e. low linkage disequilibrium). Recent introgression of diverged lineages may also leave a signature of high polymorphism in regions with the acquired alleles (Castric et al. 2008). Additionally, balancing selection in low-recombining regions would maintain high levels of polymorphism in a population (Charlesworth 2006; Hedrick 2007). Interestingly, Lawniczak et al. (2008) recently reported a positive relationship between polymorphism and expression variation. Further investigation of high polymorphism regions of the *D. pulex* genome will examine these and other contributing forces.

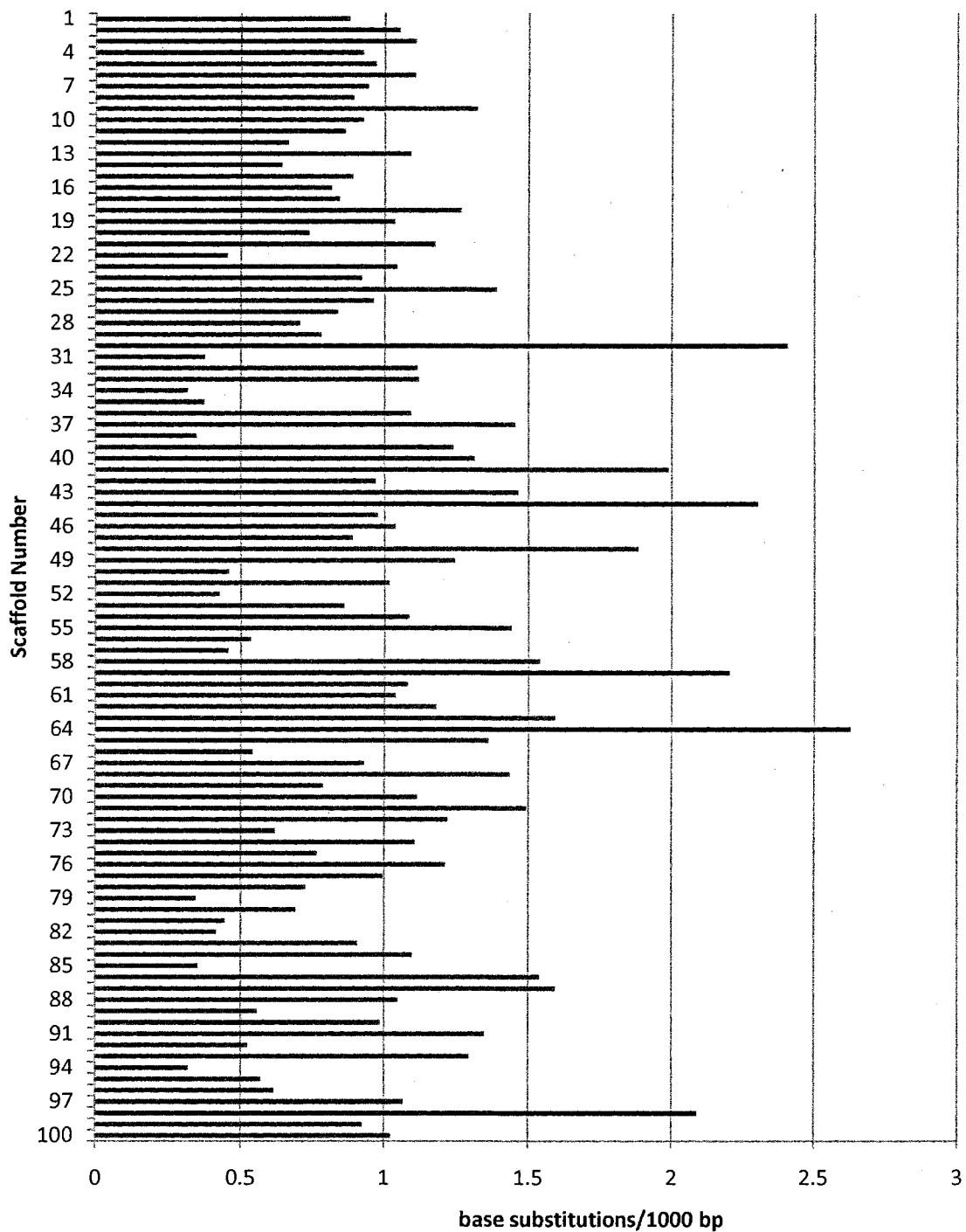


Figure 1-16. Scaffold-wide base substitution frequency. Min = 0.32/1000, Max = 2.63/1000 bp. Average frequency across N50 = 1.01/1000 bp. High polymorphism scaffolds in red.

Over half (59/100) of the largest scaffolds could be assigned to chromosomes based on a combination of genetic map and paired-end data (Figure 1-17). While scaffolds on chromosomes 2 and 3 have a higher average SNP frequency, the differences are insignificant when variance is considered.

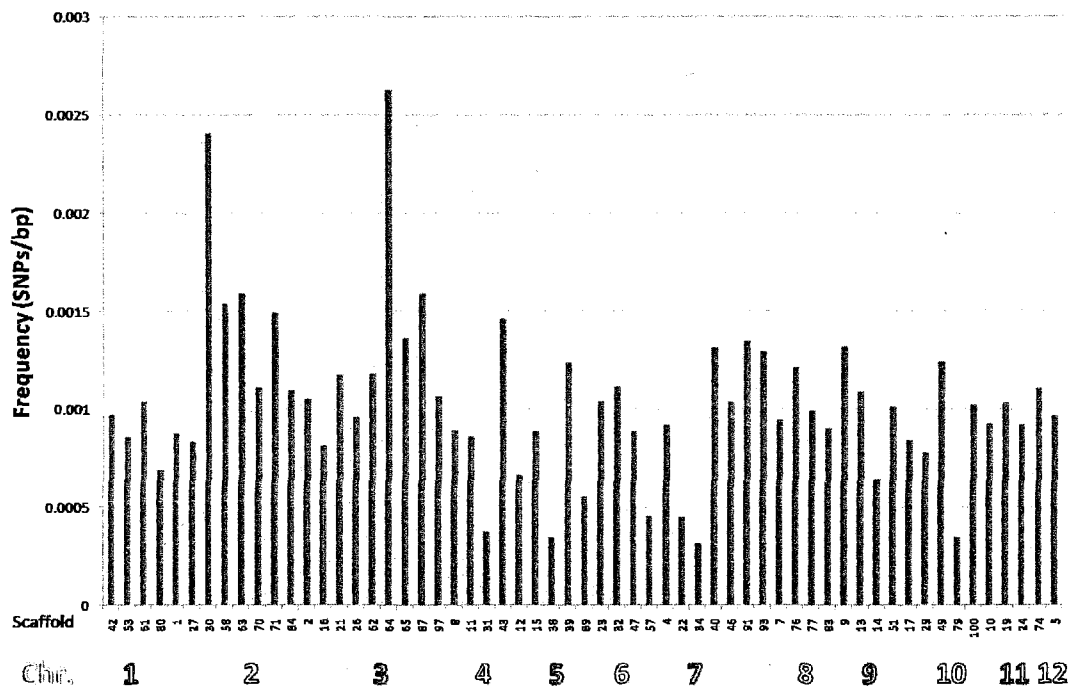


Figure 1-17: Polymorphism on mapped scaffolds.

Recombination

Theory suggests that nucleotide diversity will positively correlate with recombination rate as genetic hitch-hiking of neutral alleles (which reduces diversity) is reduced by higher crossover frequencies (Smith and Haigh 1974; Begun and Aquadro 1992; Kaplan et al. 1989). Both purifying and adaptive selection will clear neutral variation in low-recombining regions (Charlesworth et al. 1993; Hudson and Kaplan

1995). For those genomes with high quality genetic maps, evidence from many genomes appear to support this (Beye et al. 2006; Payseur and Nachman 2002; Betancourt and Presgraves 2002; Kulathinal et al. 2008). Some have suggested that the correlation of diversity and recombination is due other factors such as the mutagenic effects of the recombination process itself (Spencer et al. 2006; Bussell et al. 2005; Hellmann et al. 2003) or the co-variation of diversity and recombination with other variables. Even in genomes where diversity correlates with recombination rate, divergence may not (Begun et al. 2007). Recombination landscapes may evolve rapidly (Winckler et al. 2005; Crawford et al. 2004; Ptak et al. 2005), therefore current recombination rates may be a superior predictor of diversity, but not divergence. This would explain why both diversity and divergence correlate with recombination when mapped at a fine scale (Kulathinal et al. 2008; Noor 2008).

Cristescu et al. (2006) published the first genetic linkage map of *D. pulex*, describing the segregation of 185 polymorphic microsatellite markers in 129 F₂ progeny of two divergent lineages. While their genetic map is not dense, their map assigned markers to 12 linkage groups, presumably corresponding to the 12 chromosomes, providing the platform to compare physical and genetic maps using the genome assembly.

For the purposes of this study, scaffolds were assigned to linkage groups from Cristescu et al. (2006) and, where two or more markers could be mapped to the same scaffold, genetic distances (cM) were divided by physical distances (Mb) for an estimate of recombination rate (cM/Mb) (see Appendix E). Estimates of average recombination rate were extremely high compared to other mapped invertebrates (Beye et al. 2006;

Severson et al. 2002; Wicks et al. 2001; Yasukochi 1998), however an improved genetic map for *D. pulex* is necessary for a robust analysis. The low density of markers (185) and modest sampling of the F₂ generation (129 individuals), combined with some problematic genotyping issues (Cristescu, personal comm.), the first generation linkage map for *D. pulex* has limited power to precisely measure recombination rates. For the regions of the genome where recombination rates could be estimated, SNP frequencies were compared. No correlation between recombination rate and polymorphism was detected (Figure 1-18).

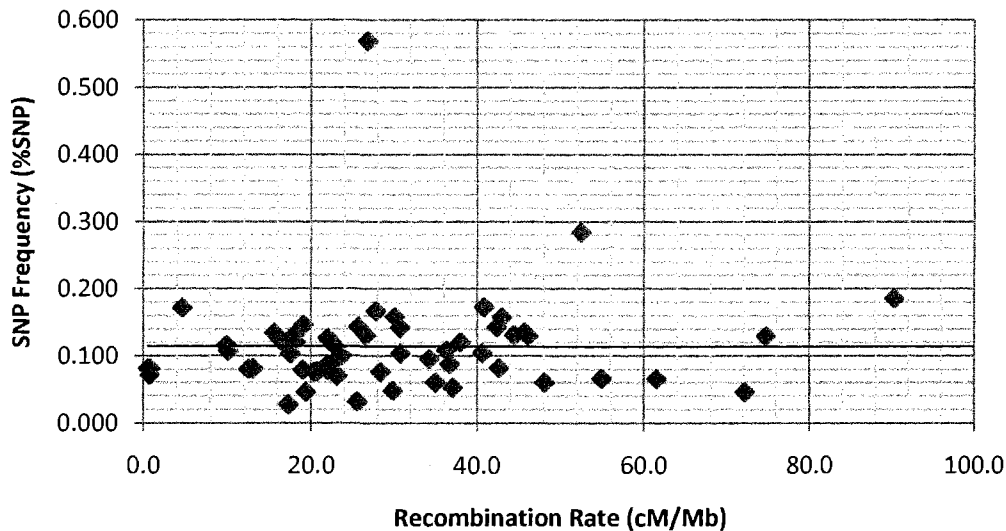


Figure 1-18: SNP frequency does not correlate with estimates of recombination rate in *D. pulex*. ($r^2 < 10^{-5}$, $p > 0.94$)

Windows Analysis

Unitary single base substitutions and single indels within TCO and TRO were examined across all scaffolds using a windows analysis of varying size (e.g. Figure 1-19). Windows with and without SNPs were not randomly distributed within scaffolds based

on a runs test of the 1000 bp window output. Using 1000 bp windows, over 250 regions of the genome were found to have unusually high numbers of consecutive windows (>10kb) without SNPs (Appendix F). Low SNP density regions of the genome may result from selective sweeps (Cai et al. 2009; Sabeti et al. 2002; Diller et al. 2002). Rampant gene conversion among alleles may also lead to increased homozygosity, however the scale of conversion is generally quite small (i.e. 1-500 bp, Chen et al. 2007; Xu et al. 2008). The low polymorphism regions detected in the *D. pulex* genome are >10kb (Figure 1-20), have typical gene density and are not enriched for recent gene duplicates or large gene families, which are thought to have increased rates of gene conversion (Teshima and Innan 2003; Nei and Rooney 2005; Pan and Zhang 2007).

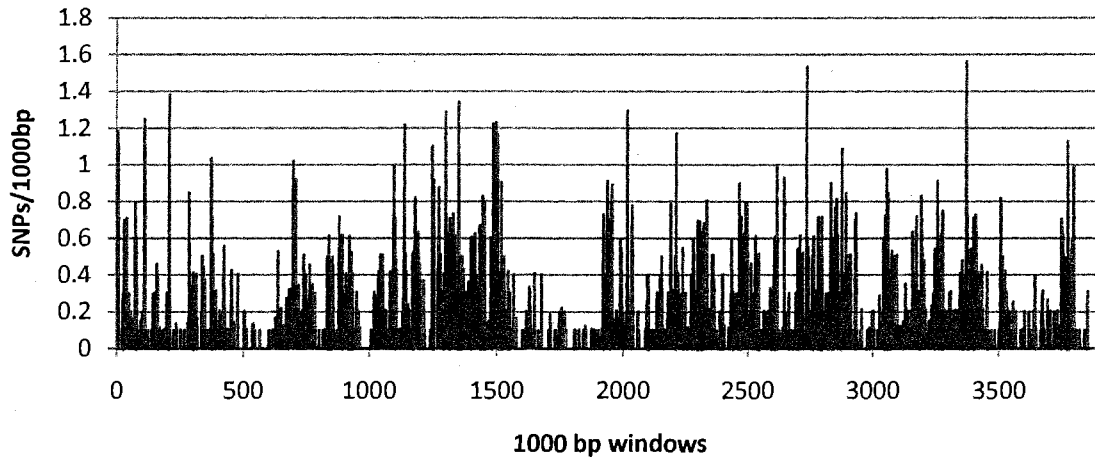


Figure 1-19: Fine-scale windows analysis of SNP frequency across scaffold 1.

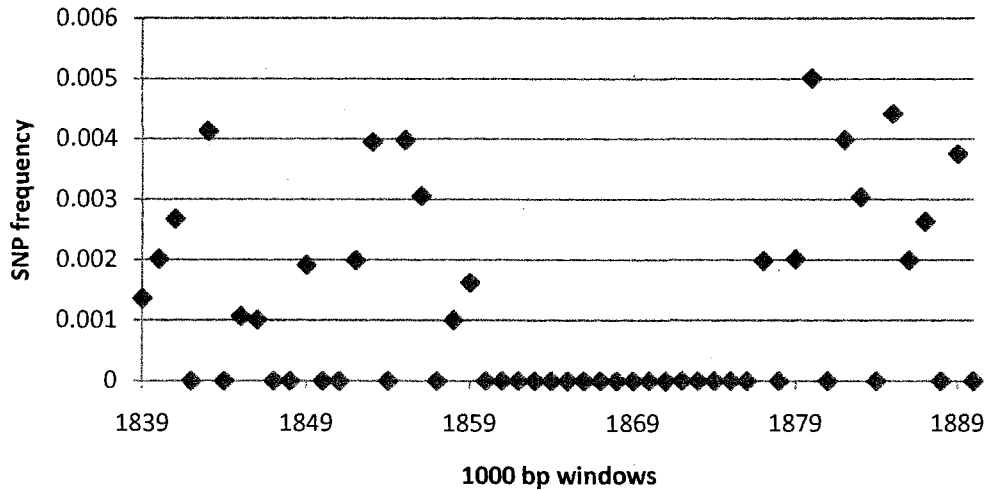


Figure 1-20: A run of 1000 bp windows with 0 SNPs along scaffold 1.

Using the windows analysis, polymorphism levels in TCO and TRO were compared. The following figures depict the difference in relative rates of single indels vs. base substitutions in TCO vs. TRO (Figures 1-21 through 1-30). For example, the indel rate is generally equal to base substitution rate in TRO, but relatively lower in TCO. TCO polymorphism levels are consistently lower than TRO, but more highly variable across scaffolds. Within each clone, rates of indel polymorphism often track with base substitutions, as would be expected in a neutral model where polymorphism levels are affected by evolutionary forces acting locally on genomic regions. However, many regions exist where relative levels of indels and SNPs diverge significantly (see TCO, scaffold 1, 6×10^{-5} bp). Both selective and mutational forces may explain such phenomena. However, when indels and base substitutions diverge in magnitude, it is unlikely to be due to sweeps in regions of high linkage disequilibrium, since all variant types would be affected.

Polymorphism levels in TCO and TRO do not correlate globally, although some local correlations exist. Considering the substantial time of divergence between TCO and TRO, it may be unlikely that selective forces acting on a common ancestor would be detectable. Therefore, while long-term, wide-net background (purifying) selection in both lineages would be detectable as low-polymorphism regions in both, there should be little expectation that polymorphism levels between the two lineages will track.

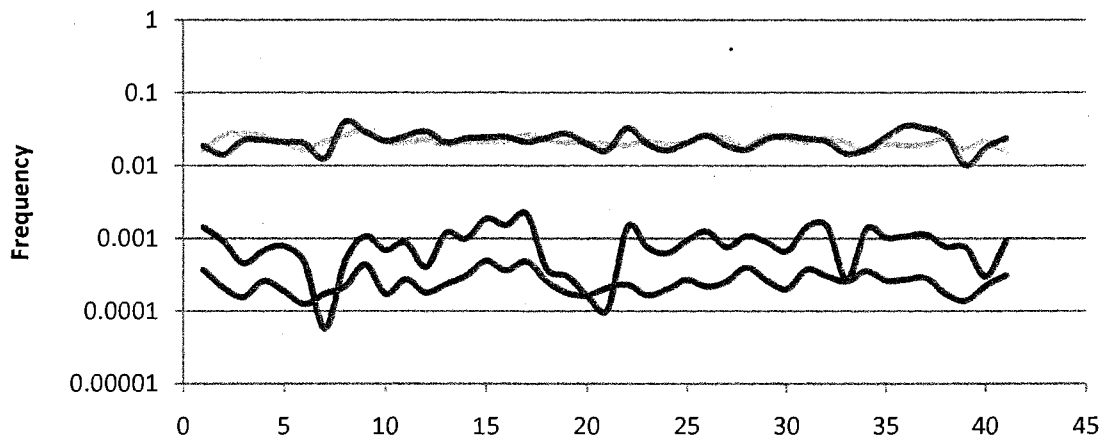


Figure 1-21: Distribution of polymorphism across scaffold 1 in TRO (Base substitutions=Green, Indels=Purple) and TCO (Base substitutions=Blue, Indels=Red)

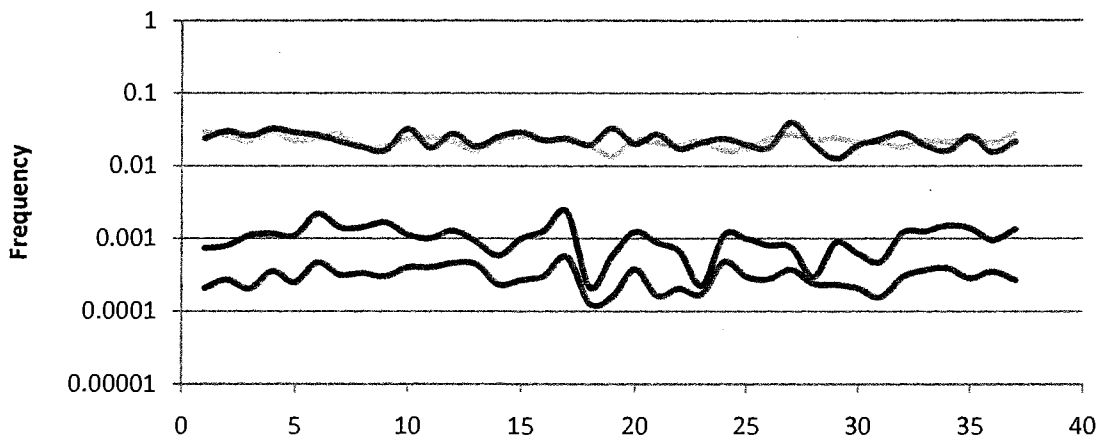


Figure 1-22: Distribution of polymorphism across scaffold 2 in TRO (Base substitutions=Green, Indels=Purple) and TCO (Base substitutions=Blue, Indels=Red)

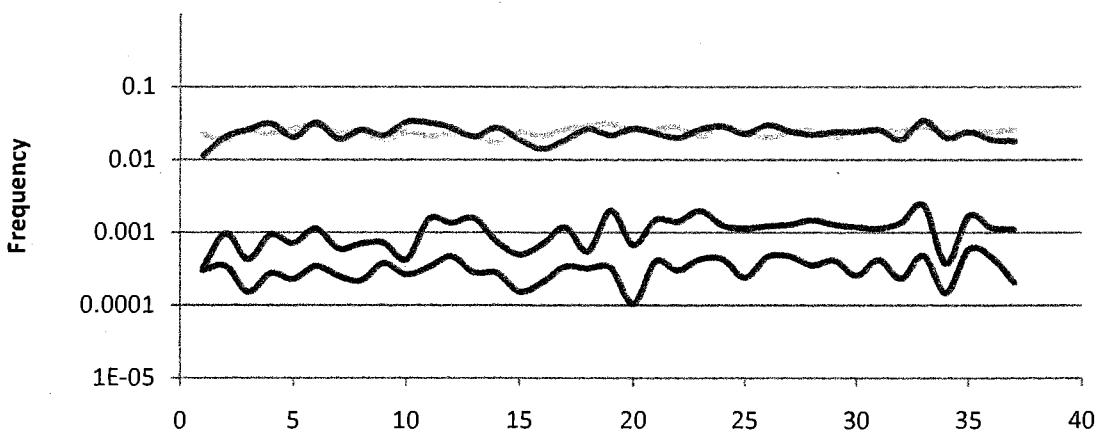


Figure 1-23: Distribution of polymorphism across scaffold 3 in TRO (Base substitutions=Green, Indels=Purple) and TCO (Base substitutions=Blue, Indels=Red)

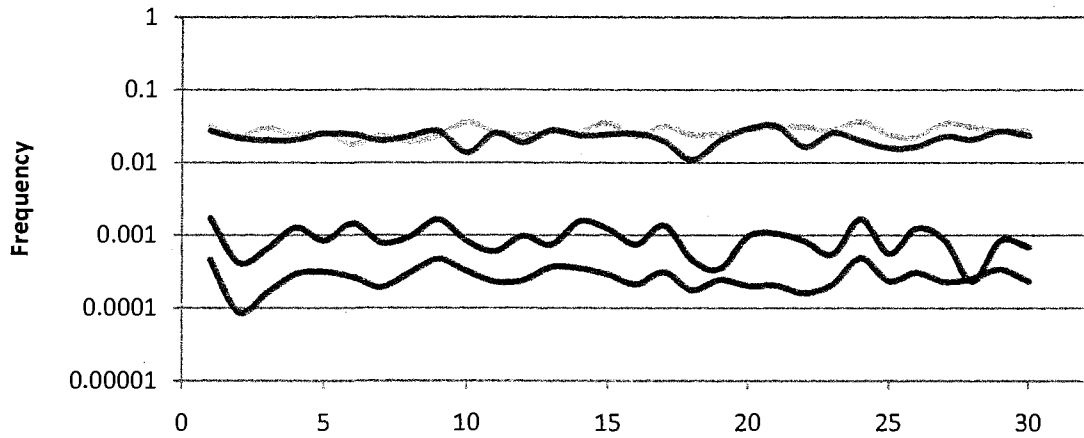


Figure 1-24: Distribution of polymorphism across scaffold 4 in TRO (Base substitutions=Green, Indels=Purple) and TCO (Base substitutions=Blue, Indels=Red)

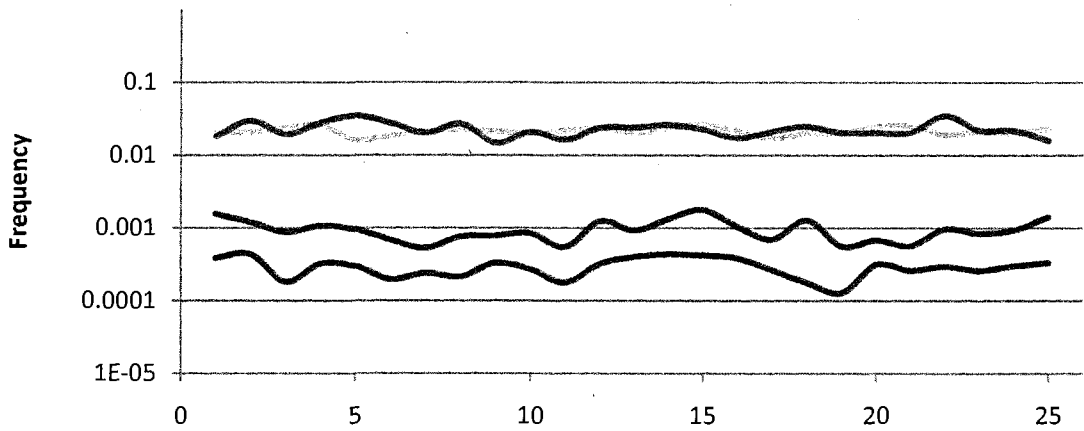


Figure 1-25: Distribution of polymorphism across scaffold 5 in TRO (Base substitutions=Green, Indels=Purple) and TCO (Base substitutions=Blue, Indels=Red)

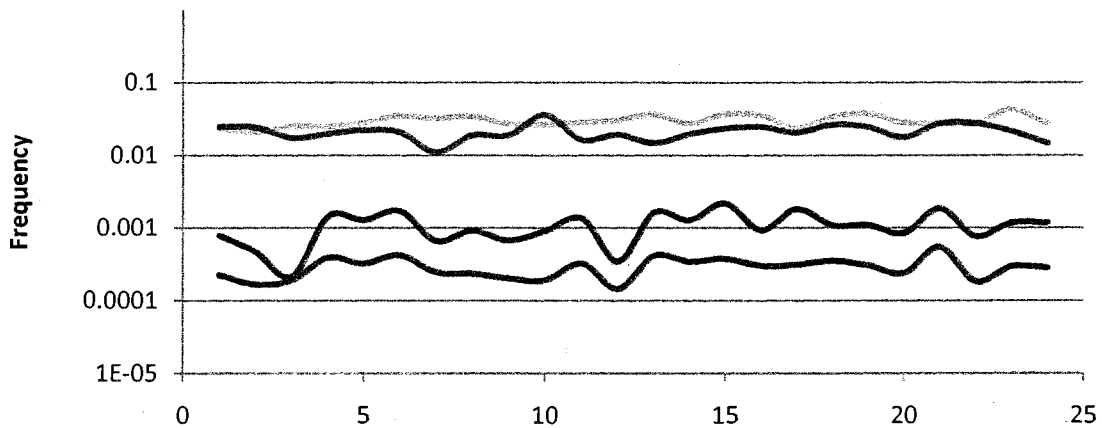


Figure 1-26: Distribution of polymorphism across scaffold 6 in TRO (Base substitutions=Green, Indels=Purple) and TCO (Base substitutions=Blue, Indels=Red)

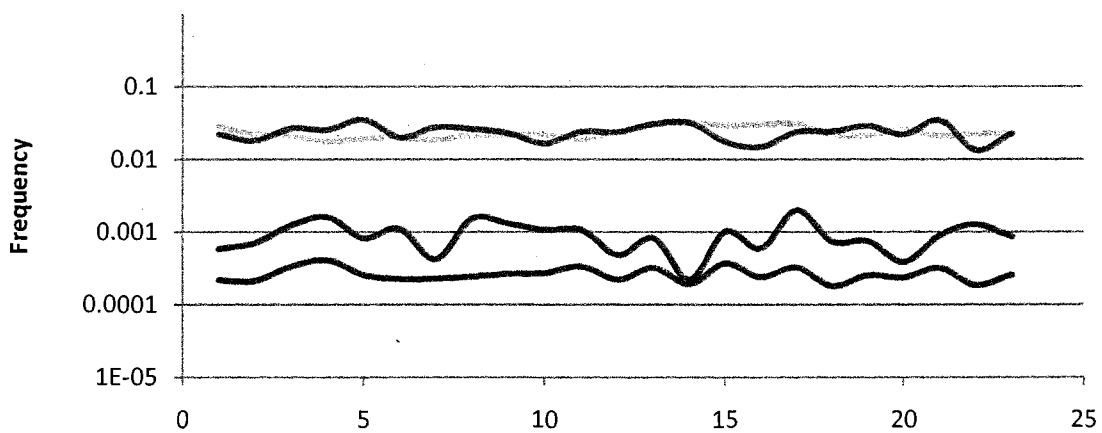


Figure 1-27: Distribution of polymorphism across scaffold 7 in TRO (Base substitutions=Green, Indels=Purple) and TCO (Base substitutions=Blue, Indels=Red)

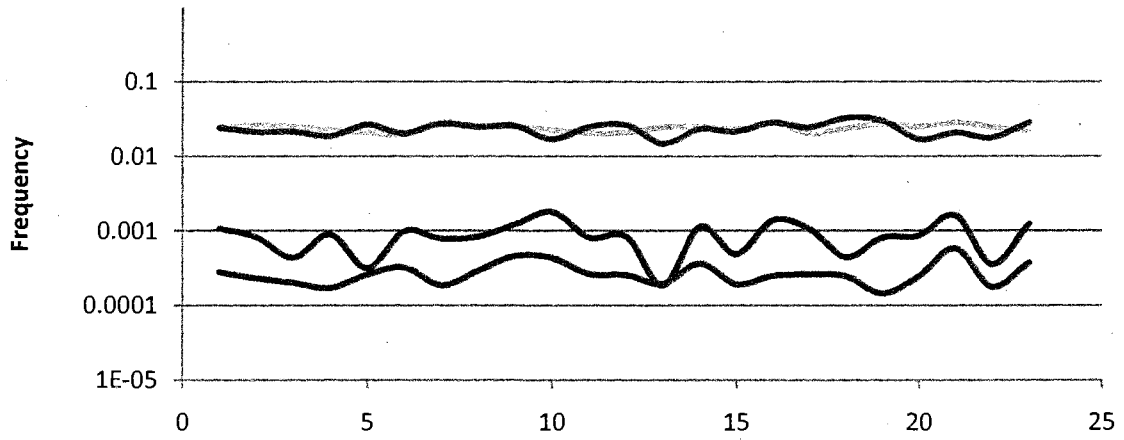


Figure 1-28: Distribution of polymorphism across scaffold 8 in TRO (Base substitutions=Green, Indels=Purple) and TCO (Base substitutions=Blue, Indels=Red)

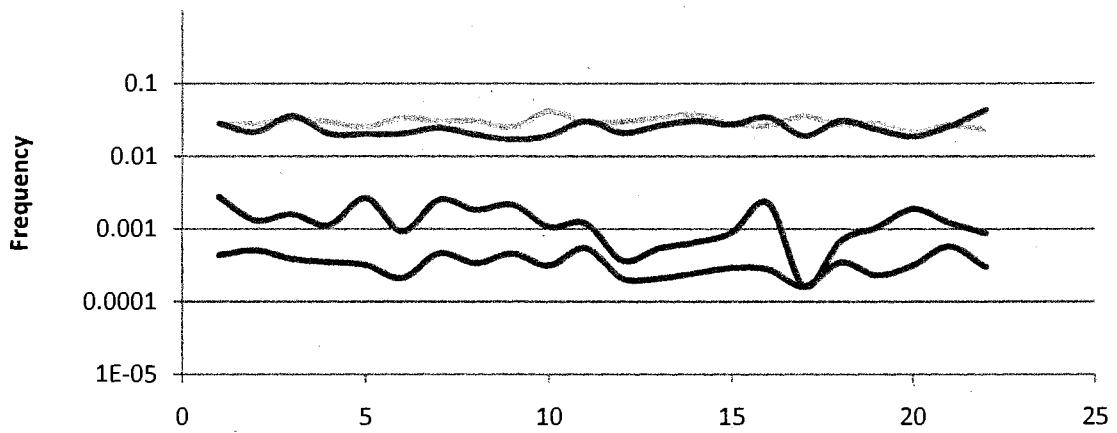


Figure 1-29: Distribution of polymorphism across scaffold 9 in TRO (Base substitutions=Green, Indels=Purple) and TCO (Base substitutions=Blue, Indels=Red)

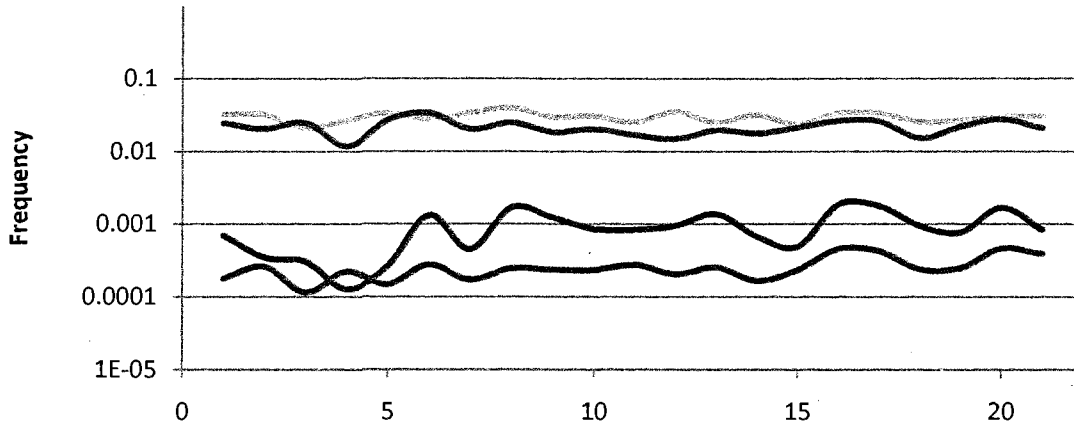


Figure 1-30: Distribution of polymorphism across scaffold 10 in TRO (Base substitutions=Green, Indels=Purple) and TCO (Base substitutions=Blue, Indels=Red)

Comparative assembly (TCO vs TRO)

The comparative assembly of TRO produced a distribution with low average coverage (Figure 1-31). Scaffold-by-scaffold analysis of TRO-TCO divergence at single nucleotides ranged from ~ 2-4% (Figure 1-32). Scaffold-by-scaffold polymorphism levels in TRO correlated with divergence from TCO (Figure 1-33), consistent with an overall neutral mode of molecular evolution. However, TCO polymorphism levels did not. The relatively low polymorphism and high variance in TCO may reduce the power to detect a correlation, even if one exists. Additionally, reduced capacity for recombination can magnify the diversity-cleansing power of genetic draft, making polymorphism levels independent of divergence at the megabase scale. Polymorphism levels across the genomes of TRO and TCO also failed to correspond in any way, suggesting that the clones may have divergent recombinatorial landscapes and/or exposure to mutational and selective pressures at the genomic scale. The low diversity and high variance of TCO polymorphism values within scaffolds may cause scaffold-wide values to be unreliable.

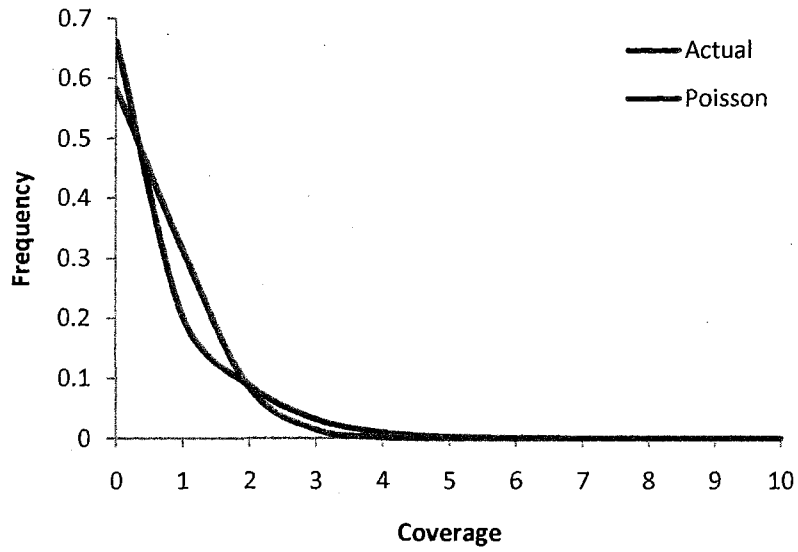


Figure 1-31. Frequency of coverage for all sites of N50 comparative assembly of TRO. (Average = 0.53X)

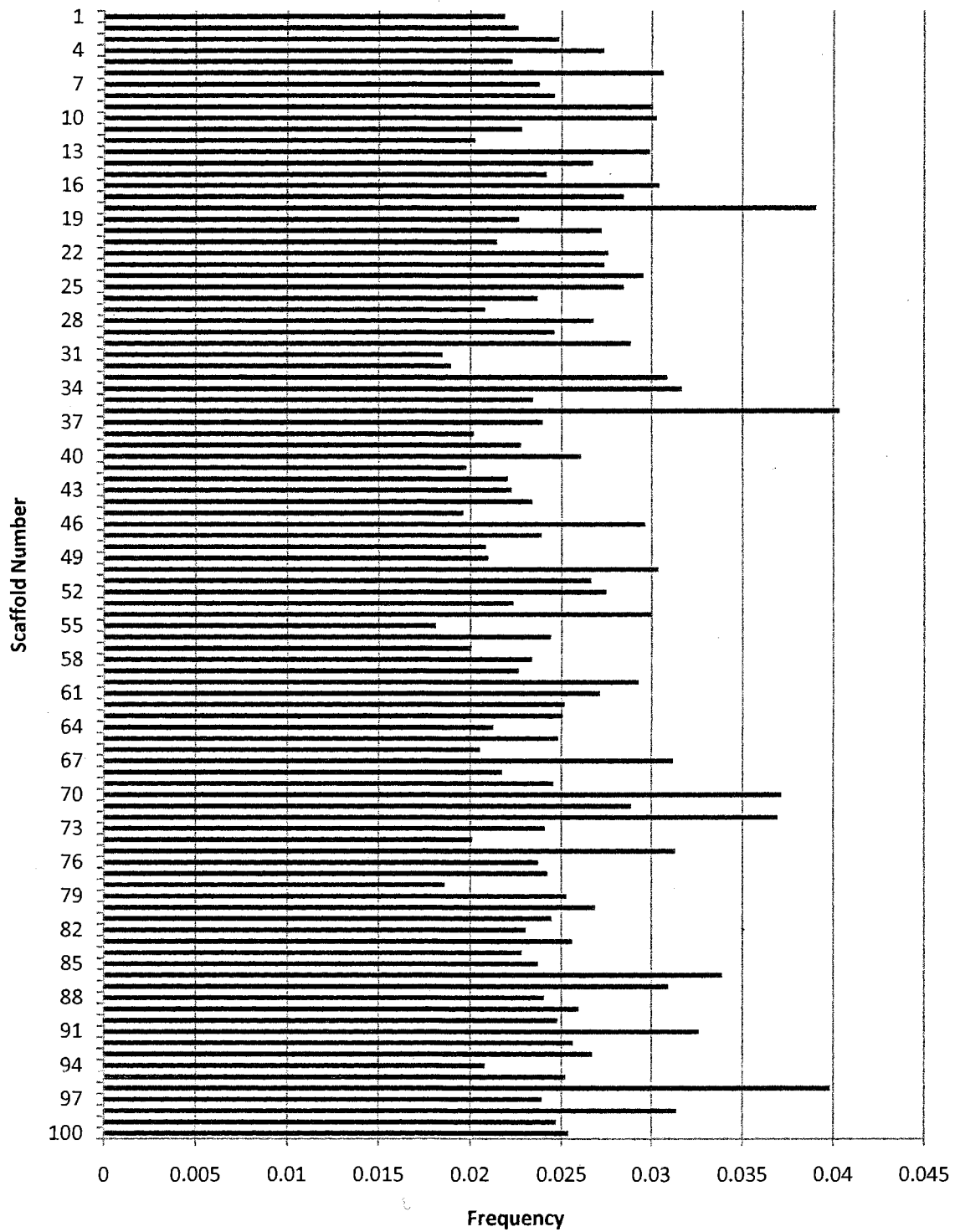


Figure 1-32: Estimates of nucleotide divergence between TRO and the TCO assembly.

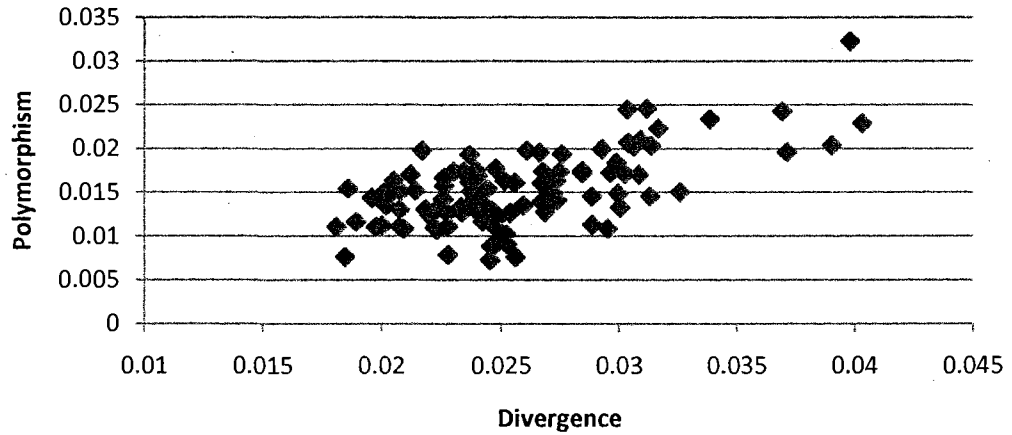


Figure 1-33: Polymorphism in TRO correlates with divergence between TRO and TCO. ($r^2 > 0.40$, $p < 10^{-12}$)

Maximum Likelihood Analysis

To evaluate the accuracy of the overall estimate of TCO heterozygosity, the results were compared to a maximum likelihood (ML) approach. Lynch (2008) developed an analytical approach using ML to factor out problems with sequencing error and under-sampling of alleles to estimate population-wide nucleotide diversity from base calls and coverage patterns across the assembly of a diploid genome. Lynch used the SNP data generated from this project as a test set for his approach. While the observed base substitution frequency from this study (corrected for undersampling and including sequential base substitutions) averaged 0.001290 substitutions/site, the ML corrected estimate of average heterozygosity was 0.001237 with 95% support interval of (0.001225, 0.001249).

Conclusion

Natural levels of genetic variation in evolving populations are traditionally measured through population sampling at relatively few loci. With the advent of whole genome sequencing of natural isolates, it has become feasible to quantify the distribution and magnitude of genetic variation across the entire genome to infer evolutionary forces acting on the population at large. This project uses the comparative assembly of shotgun reads to a high quality reference genome to detect polymorphic sites in the *Daphnia pulex* genome. Analysis of roughly 100 million sites across the genome has not only provided a wide glimpse of small genetic variation across the genomes of two divergent *D. pulex* clones, but has suggested loci that have undergone recent evolutionary pressures. The catalogue of variation reported here, including genomic regions with high and low SNP levels, is now part of the genome annotation, which will aid in the evaluation of coding function and contribution of allelic variation to *Daphnia* biology. These insights contribute valuable information to the ecological and evolutionary study of the first crustacean genome sequence, *Daphnia pulex*.

Acknowledgements

Morel Henley wrote many of the Perl scripts (Appendix C) used for the analysis. Shilpa Kulkarni wrote the Java program *SNPfinder*. Way Sung wrote Perl scripts and provided simple sequence positions. Sanjuro Jogdeo provided scaffold bridging data from paired end reads and helped with the silent vs. replacement analysis. Karen Carleton suggested the runs test. The *Daphnia* SNP project began as a class project under the guidance of Profs. W. Kelley Thomas and R. Daniel Bergeron.

CHAPTER 2

PATTERNS OF VARIATION IN RECENTLY DIVERGED MITOCHONDRIAL GENOMES OF *DAPHNIA PULEX*

Background

Mitochondrial genomes and evolution

Mitochondrial genomes are long established models of molecular evolution (Brown et al. 1979, 1982; Ferris et al. 1981). Most animal mitochondrial genomes are simple and streamlined, contain diverse functional domains, providing plentiful data in easily characterized, circular genomes (Chen and Butow 2005; Boore 1999). The technical simplicity of mitochondrial DNA analysis, minimal recombination and relative ease of mitochondrial DNA isolation led to an early explosion of comparative sequence data for the mitochondrial genomes of animals in the 1980s and 90s (Brown 1985; Thomas and Wilson 1991; Kocher et al. 1989; Thomas et al. 1989; Thomas and Kocher 1993). These genomes were popular sources of data for phylogenetic and population analysis (Avisé et al. 1987; Irwin et al. 1991; DeSalle et al. 1987). Phylogenetically broad analyses of mitochondrial DNA revealed a rapid rate of base substitution relative to nuclear sequences and an extreme bias toward transition substitutions (Brown et al. 1979; Denver et al. 2000). The legitimacy of using mitochondrial sequences for deep phylogenetic

comparisons has been questioned, in part due to the problematic qualities of mitochondrial molecular evolution, such as substitution bias and rapid saturation (Cuore and Kocher 1999; Hassanin et al. 2005; Blouin et al. 1998). Progress toward a mechanistic explanation of mitochondrial genome evolution requires an understanding of population genetic factors such as mutation, selection, demography and population size, all forces that ultimately shape the DNA sequences from which many of our evolutionary inferences originate.

Animal mitochondrial evolution is characterized by a high rate of transition substitutions (Belle et al. 2005; Aquadro and Greenberg 1983; Tamura and Nei 1993). In addition, many mitochondrial genomes have base compositional skew between the DNA strands (i.e. $G \neq C$ and/or $A \neq T$ on the same strand, Perna and Kocher 1995; Frank and Lobry; Asakawa et al. 1991; Andersson and Kurland 1991). If similar substitutional processes are occurring on both strands, the Parity Rule 2 states that AT and GC skew will be 0 (Sueoka 1995). While skew is observed locally in nuclear genomes, it is a global feature of some mitochondrial sequences, especially mammals (Saccone et al. 2002). Recent analysis of animal mitochondrial DNA has focused on understanding the mechanisms responsible for the high rate of substitution and base compositional bias between the two DNA strands (Niu et al. 2003). Current theory favors a strand-specific mutation-driven model for base substitution, which proposes that directional mutation drives the biases in stand composition and codon usage (Reyes et al. 1998; Tanaka and Ozawa 1994). Because deamination of cytosines occurs asymmetrically on the two strands, it is thought that the strand that spends more time single-stranded during replication suffers more C→T mutations (Bielawski and Gold 2002; Faith and Pollock

2003), although the single-strand exposure mechanism has been questioned (Yang et al. 2002; Rocha et al. 2006). This C→T transition bias on one strand reduces the frequency of C on one strand and G on the other. Replication is asymmetric in most vertebrate mitochondrial DNA (Shadel and Clayton 1997; Clayton 2000, for exceptions see Reyes et al. 2005), however in many invertebrate genomes (e.g. *C. elegans*, *D. melanogaster*), both strands are replicated asymmetrically. These genomes tend to be extremely A+T rich (Thomas and Wilson 1991; Clary and Wolstenholme 1987) and have less extreme GC skew. However, largely ignored is the alternative explanation that *selection* contributes to these patterns, a scenario consistent with the first direct analysis of mutation in invertebrate mitochondrial genomes (Denver et al. 2000).

The mitochondrial genomes of *Daphnia* represent a significant opportunity to expand our understanding of mitochondrial genome evolution. First, from the perspective of ecology and population biology, *Daphnia* is one of the best characterized genera in biology, providing both a context of population size and large numbers of lineages from which to carry out polarized genome comparisons. Additionally, the complete sequence of the *Daphnia* nuclear genome provides a catalog of relevant nuclear genes, such as genes involved in mitochondrial function, repair, and inheritance. Ultimately, a rich understanding of the role of these genes in mitochondrial evolution will shed light on lineage specific rates and patterns of change.

The observed patterns of substitution between genomes are a product of mutational forces and subsequent filtering by natural selection and drift. Many attempts have been made to capture the underlying mode and tempo of mutation by measuring molecular change at sites that are thought to undergo minimal selection (silent sites or non-coding

DNA). However, it is increasingly apparent that previously unimagined selective forces are at play in surprising locations of genomes (Chamary et al. 2006; Svensson et al. 2006; Chen and Blanchette 2007; Vavouri et al. 2007; Andolfatto 2005). Inconveniently, mitochondrial genomes are famously devoid of many of the “non-functional” domains that occur in nuclear genomes. But because of their relatively rapid level of nucleotide substitution, a sufficient number of base changes can be observed between moderately diverged genomes for an analysis of evolutionary divergence between closely-related genomes.

In this study, the rate and pattern of nucleotide substitution in *D. pulex* mitochondrial genomes are described and hypotheses are developed about relative roles of selective and mutational determinants of observed substitution patterns.

Methods

The *Daphnia* Genome Project has provided deep sequence coverage of mitochondrial DNA in two isolates of *Daphnia pulex*, TCO and TRO. For each data set, genomic clones were aligned and assembled against a reference *Daphnia pulex* mitochondrial (Crease 1999) genome sequence using AMOS Comparative Assembler (Pop et al. 2004), providing 40-60X coverage of the mitochondrial genome for each strain. The mitochondrial genome sequences, derived from the consensus of the clones, were then aligned in ClustalW (Larkin et al. 2007) and compared, allowing for analysis of substitution patterns among the three strains using MEGA 3.1 (Tamura et al. 2007).

Using TCO as an outgroup, the direction of substitutions in TRO and Crease was inferred, allowing for polarization of a subset of the substitutions that occurred since the

et al. 2005). The proportion of substitutions in protein-coding sites (175/233) is roughly equivalent to the proportion of protein-coding sites in the mitochondrial genome (72.2%). Of the 58 substitutions not in protein-coding genes 11 were located in tRNA genes, 17 in rRNA genes and 32 in putative non-coding DNA. Of the 175 substitutions in protein-coding regions, 143 were synonymous and 32 nonsynonymous (4.5:1), indicative of the strong purifying selection on mitochondrial protein-coding sequences. 57.1% were in major strand genes, while 42.9% of protein-coding substitutions were in minor strand encoded genes. A single codon among all protein-coding sites had a detectable double substitution, evidence that the multiple hits can influence data from recently diverged sequences.

To investigate the specific pattern of nucleotide change we identified the subset of substitutions that could be polarized based on the relationship of the three genomes compared. The matrix of 53 substitutions at 2- and 4- fold degenerate sites is consistent in pattern and distribution with the overall dataset but reveals a significant bias toward substitutions from G to A (Table 2-1). Our analysis of 2 and 4-fold degenerate sites in *Daphnia* is consistent with that observed in *Drosophila* (Haag-Liautard et al. 2008) vertebrates as well as with a mutational mechanism driving strand specific nucleotide composition. G to A substitutions are observed twice as often as expected from a strand-specific nucleotide composition model in both the major and minor protein-coding regions (Table 2-1, chi square test, $p=0.0017$). This observation suggests that the probability of any G on the major coding strand changing to an A is higher than the probability of an A changing to a G.

Base from↓	<u>To</u>					Exp	Obs
	A	T	C	G	NC		
A	-	1	2	16	.328	17.4	20
T	0	-	10	0	.309	16.4	10
C	0	6	-	1	.221	11.7	6
G	15	0	0	-	.144	7.6	17

Table 2-1: Ratio of Observed to Expected Substitutions from each nucleotide based on nucleotide composition (NC) of 2- and 4-fold sites on the major strand.

The observed bias towards G→A substitutions is consistent with repeated claims that mutational and/or selective forces are driving base compositional differences between the major and minor strands. However, observed substitution patterns at degenerate sites are a proxy for underlying mutations and there is no direct evidence that directional substitutions or strand bias are the result of an underlying mutational bias rather than a selection. In fact, the mutation accumulation experiments in *C. elegans* suggest that selection, rather than mutation, drives substitutional bias in the mitochondria of nematodes (Denver et al. 2000).

Unlike the nuclear chromosomes, the replication of mitochondrial genomes within a cell is not controlled such that all chromosomes replicate to completion before cell division. While the dynamics of mitochondrial DNA inheritance remains a mystery, it is clear that there exist serious bottlenecks, resulting in rapid fixation of new mutations. Consequently, it must be assumed that there exists a replication race where the molecules that are slower to replicate necessarily contribute less to the daughter population after cell division. The repair of deaminated cytosines will necessarily delay replication of molecules, selecting against molecules having deaminated sites compared to other

molecules with less damage. Therefore, the consequence of deamination of cytosine may not be directional mutation but selection against molecules with sites prone to deamination (Figure 2-2). To further evaluate the potential role of repair enzymes involved, we have identified the key gene in the mitochondrial deamination repair pathway, uracil-DNA glycosylase (UNG), in the *D. pulex* genome (Nilsen et al. 1997).

The directional mutation model suggests that GC skew is driven by a high frequency of deamination on one strand. However, most animal genomes have a host of DNA repair genes that operate in the mitochondria. It may be more likely that high rates of deamination on one strand slow replication for strands with higher cytosine content. An ongoing replication race would lead to selection for lower C on one strand. While the current model is not mutually exclusive with the model proposed here, the distinction is important for making predictions about the underlying mutational spectrum and the evolutionary forces shaping the mitochondrial genome. Since *Daphnia* mutation accumulation lines will be sequenced soon, a key prediction from our model is that a bias towards G/C to A/T substitution, as reported here, will not be observed.

We show that substitution type is not proportional to nucleotide frequency when comparing closely-related *Daphnia pulex* mitochondrial genomes. If mutation accumulation experiments show that substitution patterns are proportional to base frequency, it is likely that 2- and 4-fold sites in *Daphnia pulex* mitochondrial DNA undergo substantial selective pressure and are not reliable measures of the underlying baseline mutational spectrum.

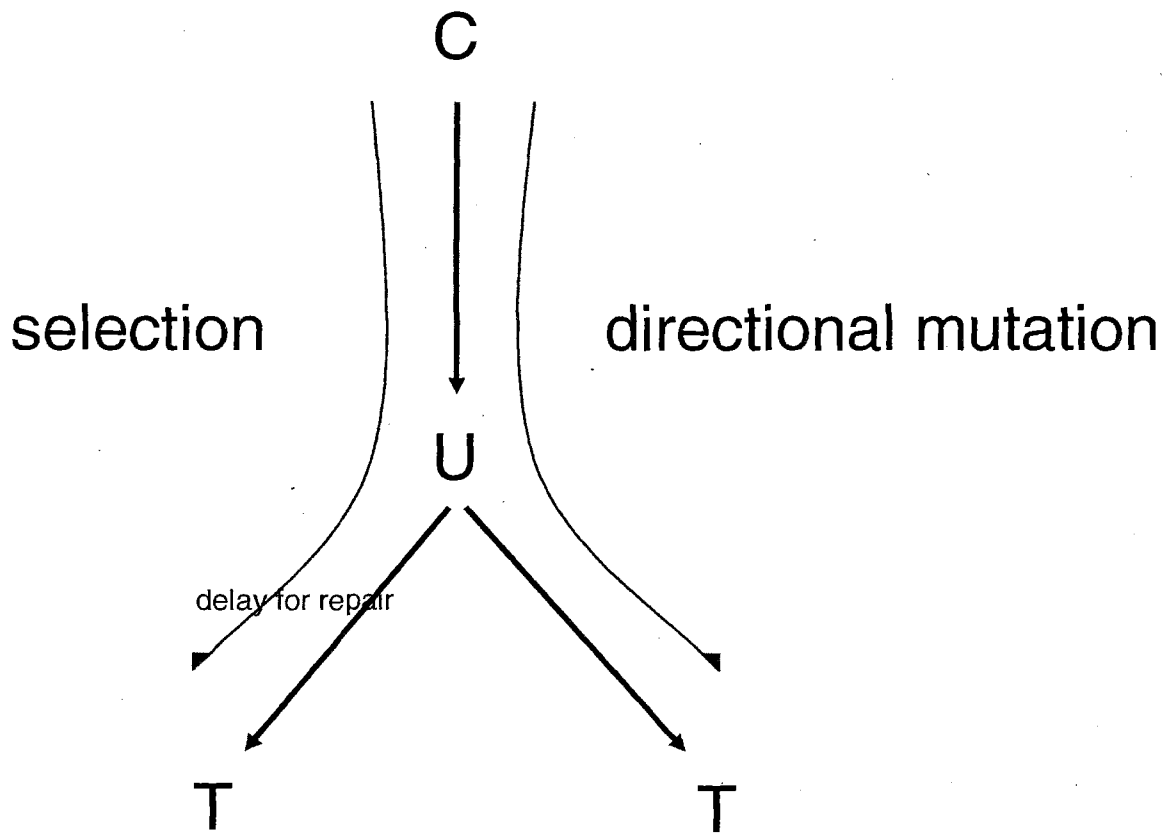


Figure 2-2. Dilatory mutation-selection model predicts biased substitution pattern (e.g. G/C \rightarrow A/T) due to natural selection against G/C regardless of underlying mutation pattern.

CHAPTER 3

INTRON GAIN/LOSS POLYMORPHISMS IN *DAPHNIA PULEX*

Background

Intron evolution

Spliceosomal introns are a defining characteristic of eukaryotic genomes and have taxon-specific patterns of proliferation, extinction and structural evolution. However, little is known about the evolutionary dynamics of introns in populations. The study of intron evolution is central to understanding gene structure evolution and the origin of genome and organismal complexity. Nevertheless, over thirty years after their discovery (Berget et al. 1977; Chow et al. 1977; Gilbert 1978), and in spite of being widely studied (Koonin 2006; Lynch 2002; Lynch 2007), the origin, function and evolutionary consequences of introns are open questions. Analysis of intron gain and loss between phylogenetically distant taxa has revealed long term trends in taxon-specific intron proliferation and suggested general patterns of intron gain/loss across eukaryotes (Cho et al. 2004; Carmel et al. 2007; Roy and Gilbert 2005; Belshaw and Bensasson 2006; Coulombe-Huntington and Majewski 2007). These patterns include differences in rates and mechanisms of turnover as well as spatial biases of gain/loss. For instance, some studies report more intron loss towards the 3' end of genes and preferential loss of introns between codons (phase 0 introns). Variation in gene family rates of gain/loss have also

been reported (Roy and Gilbert 2006; Jeffares et al. 2005). While these studies are rigorous and expansive, they are limited to uncovering general trends between relatively few, distant genomes. While some of the trends uncovered in these large analyses of divergent taxa point to particular mechanisms of intron loss (e.g. reverse transcription/gene conversion vs. genomic deletion) and gain (e.g. ectopic seeding, local duplication), they may not be the most powerful or informative approaches for testing specific hypotheses regarding the molecular mechanisms of intron loss and gain. For instance, introns that are ectopically seeded will lose detectable homology rapidly after speciation, making them impossible to detect. Additionally, broad phylogenetic comparisons of intron positions rely on assumptions of character irreversibility (Dollo parsimony, Farris 1977), a conservative view of intron evolution that may be unrealistic, especially if cryptic, unoccupied “proto” splice sites play a large role in intron gain (Sadusky et al. 2004; Lynch 2007). However, recent studies have indicated that intron gain and loss may be studied at the population level, a prospect that will contribute power for inferring gain/loss mechanisms. In one of the first examples of intraspecific intron gain/loss polymorphisms, Llopert et al. (2002) uncovered a standing polymorphism for intron loss/gain in a natural isolate of *Drosophila teisseri*, with evidence of loss through genomic deletion with possible selective forces acting on the deletion allele. More recently, Omilian et al. (2008) reported two novel intron gains segregating in *Daphnia pulex*. The Llopert and Omilian studies suggest it is possible to discover intron gain/loss soon after introduction of new alleles in a population, improving our ability to understand the process of intron turnover in eukaryotes. In light of the Omilian study, this study used a combination of genomic and population genetic resources to assay *D. pulex* for intron

turnover. The data reported here suggests that genome-wide population level studies may be essential to understanding intron evolution. The results further underscore the importance of comparing closely related genomes to understand the origin and evolution of genomic variation.

In this study, the predicted intron-exon boundaries in the *Daphnia pulex* genome (clone TCO) were used to detect the absence of introns in a second clone (TRO). The loci with intron absences identified in this comparison were assayed for polymorphisms across a panel of *D. pulex* populations at Indiana University. In all, 22 cases of putative intron gain and 2 cases of intron loss are reported. The results further indicate that intron turnover can be studied at the population level, at least in some taxa, and that a thorough understanding of the evolutionary dynamics of introns awaits population genomic level investigation.

Methods

Intron polymorphism was examined in *Daphnia pulex* using two genomic data sets, TCO (9X), and TRO (1X), both from the *Daphnia* Genome Project. A Perl script *getIntronJunction.pl* was written to extract and concatenate 50 bp before and after every predicted intron of the TCO Assembly using the FrozenGeneCatalog_2007_07_03.gff and *Daphnia_pulex.fasta* (<http://genome.jgi-psf.org/Dappu1/Dappu1.download.ftp.html>). These 85,353 100 bp exon-exon hybrid sequences were queried against the TRO 1X shotgun reads using BLASTn $< e^{-35}$, a threshold allowing alignment of 100 bp exon-exon hybrid sequence from TCO against an “intronless” genomic clone from TRO with a handful of mismatches. Putative intronless TRO clones were then aligned with the TCO

exon-exon hybrid, TCO parent gene sequence (containing intron) and other TRO clones that hit the gene elsewhere, all using Sequencher. Alignments were examined and adjusted by eye.

The alignments were used to design primers for amplification of the intron absence regions in both TCO and TRO to confirm the *in silico* analysis. Primers were designed in regions of perfect conservation between TRO and TCO, upstream and downstream from the intron site. The primers were also used for the population assays carried out at Indiana University, where these loci (TCO introns absent in TRO) were assayed in 96 *Daphnia* (mostly *D. pulex*) isolates from across North America.

To test the hypothesis of intron origin through ectopic seeding, BLASTn was used to search the TCO genome assembly for sequence homology to all introns involved in gain/loss (n=34).

Results and Discussion

The recently assembled *D. pulex* genome (TCO) was used to survey intron presence/absence in another *D. pulex* clone (TRO), for which there is substantial genomic data (1X shotgun sequence). 34 genes with instances of intron absence were observed in TRO. These putative intron gain/loss loci were then amplified in TCO, TRO as well as in dozens of other *D. pulex* lineages. 4 were found to have upstream and downstream intron absences (Table 3-1) and to rule out the processed pseudogenes, all cases with adjacent losses were eliminated from further analysis. 24 loci are confirmed by direct PCR analysis to be polymorphic for intron presence/absence within the *D. pulex* species.

An analysis of the DNA sequences flanking the TCO introns lost in TRO reveal small, direct sequence duplications for 13 TCO introns (Figure 3-2). Although the exact mechanism generating these duplications is unclear, they strongly suggest that these 13 TRO introns are recent gains.

To put the confirmed intron polymorphisms in a greater phylogenetic context, primers were sent to IU to amplify and sequence putative gain/loss loci in a diverse panel of *Daphnia* lineages (The “Big 96”). Through direct sequencing, the polymorphisms were phylogenetically polarized using outgroups to the TRO/TCO clade. Additionally, these loci were assayed across clones inside the TCO/TRO clade. 12/24 intron polymorphisms were confirmed to be putative intron gains, while 2/24 cases are putative losses based on a phylogenetic analysis of character state in the “Big 96”. Most of the gains have EST evidence from TCO, suggesting that the introns are actively spliced (wleabase.org/release1/current_release/est). The introns appeared in conserved regions of the alignment, and flanking sequences do not appear to contribute to the gained introns, arguing for gain by the insertion of exogenous sequences. Interestingly, 5 loci show independent intron insertions at the exact same sites (parallel gain), supporting the proto-splice site theory of origin. 1 intron has significant homology to introns of other genes, an observation consistent with an ectopic seeding model of intron gain (Roy and Gilbert 2006).

Over all, the intron presence/absence polymorphisms are in a functionally diverse set of protein coding genes and do not appear to be biased towards intron phase or gene location. However, it is notable that all genes found to have an intron polymorphism in *D. pulex* have significant homology with other sequenced animals and appear to be highly

conserved genes (Table 3-1). A random set of genes from the *D. pulex* gene catalogue would include many (>30%) genes without homology outside *Daphnia*. While there may be an ascertainment bias contributing to this result, further investigation of these intron polymorphisms will shed light on the distribution of intron turnover within the *D. pulex* gene set.

The results of this study indicate that intron turnover in *D. pulex* is rapid and that *Daphnia* may be a useful model for understanding intron evolution. Intron turnover at the same site may be high even within genomes of the same species, making inferences based on Dollo parsimony unreliable.

Although numerous broad phylogenetic comparisons have been employed to measure taxon-specific rates of intron gain and loss in highly conserved genes, it is possible that the assumptions of parsimony underlying these measurements are often violated and/or not applicable when sampling a small biased set of genes (i.e. highly conserved genes that can be aligned between different phyla or classes). This study is a genome-wide assay of intron-exon boundaries in at the population scale. If extensive intron gain/loss combined with rapid divergence of introns limits inference of intron phylogeny between even moderately distant taxa, phylogenetically broad inferences of intron turnover may be misled by a saturation phenomenon. Additionally, an understanding of the mechanisms of intron gain and loss are improved by detecting intron birth and death soon after the introduction of new alleles. The signature of gain or loss may still be detectable when new alleles aren't fixed in the population. For instance, the presence of small duplications flanking 13 of the intron positions discovered in this study may indicate a mechanism of insertion (Figure 3-1). For instance, a staggered DNA break

followed by synthesis of a new strand leads to small target site duplications. The observation of such features at the intron presence/absence sites suggests that some of the polymorphisms may be result of ectopic or *de novo* intron gain.

This study is not a comprehensive analysis of intron turnover, but a conservative genome-wide look at intron evolution in *D. pulex*. Our study is restricted to absences of TCO introns in TRO, which were then investigated in a diverse panel of *D. pulex* populations. Since a TRO assembly with gene predictions does not exist, exact intron positions in TRO are unavailable for the reverse assay of TCO absences. Additionally, only “perfect” absences are detected in TRO. For instance, a genomic deletion resulting in 3n leftover bases in an intron leading to extension of an exon, like the Llopart (2002) deletion, would have been identified in our data set. No such cases were discovered in this analysis.

In addition, we are aware of the possibility of false positive intron absences in TRO due to the recent insertion of a whole or partial pseudogene in TRO. However, none of the genes involved in TRO intron absence (n=34) have evidence of a pseudogene copy in TCO, meaning a pseudogene would have had to insert in TRO or be lost in TCO after the divergence of the two strains. Although this cannot be ruled out, 30 of the 34 genes with intron losses in TRO have up/down stream introns. Additionally, PCR amplification should yield two different size products if both intron-containing and intron-lost paralogs existed. However, for the genes that show evidence of heterozygosity for intron gain/loss in TRO, it may be hard to distinguish between allelic vs. paralogous variation.

ACCAGGT.....ACCAGGT
CAG.....CAG
GGTACT.....GGTACT
CAAATGAATGAAGGT.....CTAAATACTGAAAGT
GGTAAGAA.....GGTAAGAA
TAG.....TAG
AGGTAAC.....AGTAAC
TCAG.....TCAG
ACCCACAAGG.....ACCCACAGG
ATCATAG.....ATCATAGG
AAAAACAGG.....AAAAACAGG
ACA ACTTACAGT.....ACA ACTTACAGGT
CAAGG.....CAAGG

Figure 3-1. Duplications at intron/exon boundaries in 13/30 TCO introns that are absent in TRO. Exonic sequence is black, intronic is red.

	Gene	Location (scaffold:start-stop)	Intron size, # (TCO)	Function	Notes
1	299809	1:642148-644866	69, 3	Isocitrate dehydrogenase	0, P ¹ , S
2	300064	1:1735823-1738795	60, 1	Phosphodiesterase/endonuclease	1, P ² , S
3	300174	1:3884606-3897221	74, 7	Chromatin structure	0, P ³ , S
4	210176	15:391013-394586	69, 2	Polyamine transporter	1, P ⁴ , H
5	49696	19:1131355-1132925	62, 6	Bestrophin (macular dystrophy)	1, P ⁵
6	220747	3:829775-831403	90, 6	Proteasome activator activity	2, P ⁶
7	192333	5:1937808-1940803	63, 3	NADH-quinone oxidoreductase	1, P ⁷
8	308710	83:571471-573372	84, 4	Poorly characterized	2, P ⁸ *
9	60686	92:448210-451261	63, 9	Aldehyde dehydrogenase	2, P ⁹
10	320441	35:552343-560119	62, 14	Glutaminase	1, S
11	42116	4:2813476-2817541	76, 2	Histidine ammonia-lyase	0, H
12	305612	4:813051-817358	85, 11	Peptidase M1	1
13	220780	3:1275256-1279669	88, 9	Poorly characterized	1
14	306461	5:2453607-2454266	119, 2	Complex 1 lyr	0, H
15	312878	7:1370402-1380160	61, 3	Na ⁺ /solute symporter	2
16	310999	3:2381836-2386328	94, 5	Transcription factor	1
17	305001	32:423509-424982	85, 5	Citrate lyase, beta subunit	0
18	324526	62:653445-655802	102, 4	<u>Cytoskeleton</u>	1
19	30917	10:1512564-1514700	65, 2	H ⁺ /oligopeptide symporter	1
20	309681	1:1652388-1656457	61, 5	Puromycin-sensitive aminopeptidase	0
21	211626	23:1173577-1177875	91, 5	STE20-like serine/threonine kinase MST	1
22	318553	25:645321-648617	70, 9	Peroxidase/oxygenase	0
23	304027	25:810711-812098	63, 5	Thioredoxin/protein disulfide isomerase	2
24	54063	39:521867-526821	202, 5	FOG: Immunoglobulin C-2 Type/fibronectin type III domains	1
25	305669	4:1820026-1822728	65, 3	Predicted polypeptide N-acetylgalactosaminyltransferase	1
26	305741	4:3053366-3058687	106, 12	Uncharacterized conserved protein	1
27	323635	55:335765-344494	177, 3	Collagens (type IV and type XIII), and related proteins	1
28	58732	75:537505-538847	67, 1	Zn finger proteins	2
29	317361	20:1017643-1021231	62, 5	Predicted E3 ubiquitin ligase	1
30	313288	8:1690769-1695101	67, 12	Glutamate-gated kainate-type ion channel receptor subunit GluR5	2
31	111110	81:63490-64695	276, 1	Protein kinase PKN/PRK1, effector,	0, H
32	230511	8:2032486-2034185	233/59, 2/3	Ribosomal protein S6e	H, psi
33	109145	63:230454-232102	117/70, 5/6	Tyrosine phosphatase	Psi
34	197668	31:989787-993690	63/66/64, 11/12/13	Kinesin	H, psi

Table 3-1. Intron absences in TRO relative to the same genes in TCO. Notes: 0, 1, 2 refer to intron phase; P-PCR confirmed; H-evidence of heterozygosity within TRO; psi- adjacent intron lost, possible processed pseudogene. *PCR result inconclusive. S- intron absence confirmed with sequencing. Pⁿ refers to PCR products in Figure 1. The last four rows (bold) include genes with multiple intron absences in which we haven't ruled out processed pseudogene origination.

For the intron gains, most occur within an isolated clade (Oregon) of *D. pulex*, a population possibly susceptible to mildly deleterious mutation accumulation due to a prolonged period of bottleneck that magnified the power of genetic drift.

Acknowledgements

Mike Lynch, Kelley Thomas and Way Sung contributed to the conception of the project (DGC meeting, July 2007). Wenli Li (Indiana University) sequenced the polymorphic loci in the “Big 96” *Daphnia* collection. Way Sung wrote the Perl script *getIntronJuction.pl*.

CHAPTER 4

GENE DUPLICATION IN *DAPHNIA PULEX*

Background

Gene duplication is an important source of genomic variation within eukaryotic lineages (Ohno 1970; Graur and Li 2000; Lynch 2007). Segmental duplications that include partial and entire protein-coding genes have been observed on the microevolutionary scale (Redon et al. 2006; Zhang et al. 2005; Khaja et al. 2006). From broader comparative analyses, it is clear that gene gain and loss cause fluctuations in gene family sizes (Demuth et al. 2006; Hahn et al. 2007). The relative roles of positive selection, purifying selection and drift on the retention and removal of new duplicates remain in dispute and may vary among taxa. While examples of gene duplicates contributing to adaptive evolution have been proposed (Nei and Rooney 2005; Irish and Litt 2005; Beisswanger and Stephan 2008), the process of gene duplication and loss have been treated like other stochastic mutational events and can be modeled as a neutral, random process, with a rate estimated to be roughly equivalent to the probability of a single nucleotide mutation (Lynch and Conery 2000). Under this view, the gene content of a genome is the outcome of a long-term equilibrium of gene gain and loss, with positive and negative selection affecting the retention of new duplicates at the margins, depending on the magnitude of beneficial or deleterious effects. Assuming a steady-state equilibrium of birth and death rates, the demography of duplicate genes can be inferred

from the contemporary gene catalogue. For example, using synonymous substitution rate (Ks) between duplicates as an estimate for age, all eukaryotic genomes studied to date show an exponential decay curve of retained duplications over time (Lynch and Conery 2002). However, whole and partial genome duplication events in the evolutionary past may appear as large cohorts containing significant duplication peaks like those found in vertebrates and *Arabidopsis* (Vandepoele et al. 2004; Zhang et al. 2004; Lynch and Conery 2003; Maere et al. 2005).

Because all genes are thought to come from other genes, the mutational processes leading to gene duplication are important for understanding evolution. Rates of unequal crossing over, transposable element-mediated transfer and whole/partial genome duplication are important factors determining the potential for gene duplication (Lynch 2007). As an ongoing, stochastic process, gene duplication seeds the genome with new sequence whose fate is determined by evolutionary pressures of drift and selection. While some gene duplicates are retained as functional copies, most duplications are lost through drift, deletion and/or silencing via deleterious mutation accumulation (Lynch et al. 2001; Lynch 2007). The fate of newly arisen gene duplicates has been given hefty theoretical and empirical consideration (Lynch and Force 1999; Katju and Lynch 2003; Rastoni and Liberles 2005; Moore and Purugganan 2003, 2005; Lynch and Katju 2004; Kondrashov et al. 2002). Some evidence suggests that gene duplication can serve as a buffer for deleterious mutation and contribute to genetic robustness (Hsiao and Vitkup 2008; Nowak 1997; Wagner 1999; Gu et al. 2003), however most duplications are not retained.

If a new duplication allele arises in a population, current models entertain three potential fates (Hurles 2004; Lynch 2007). Unlike a single gene whose function may be

essential, a new duplicate may initially escape the constraints of purifying selection. Degenerative mutation may silence the duplicate, leading to *nonfunctionalization*. However, freedom from intense purifying selection can, in rare instances, lead to new, advantageous alleles, a process termed *neofunctionization*. In this way, new duplicates can become test beds for evolutionary novelty. Numerous examples of neofunctionalization have been reported (Zhang et al. 1998; Lynch 2007a; Escriva et al. 2006; Beisswanger and Stephan 2008). A third evolutionary fate of a new duplicate gene, *subfunctionalization*, has been proposed. Subfunctionalization occurs when both the parent and child gene undergo compromising mutations that split the functions of the parent gene between the relatives. This Duplication-Degeneration-Complementation model suggests a mechanism by which new evolutionary opportunities may arise even in the presence of purifying selection (Force et al. 1999; Lynch and Force 2000).

Since the dawn of eukaryotic genomics, the simple observation has been made that gene content does not correlate with organismal complexity. That a nematode and a human both have roughly 20,000 protein-coding genes, begs an explanation for how the chasm of organismal complexity is achieved. The study of recent gene duplications is a tractable phenomenon for testing the evolutionary potential of new mutations. The generation and fate of gene duplicates is certainly not deterministic, but depends on the local and long-term population-genetic environment of populations. Here, we compare the overall demography of the *Daphnia pulex* gene duplicate catalogue to other taxa and attempt to summarize some general patterns of gene copy evolution in the recently sequenced microcrustacean. We calculate estimates of non-synonymous substitution rate (K_a) and synonymous substitution rate (K_s) for each gene pair. K_a and K_s calculations

(often referred to as dN and dS) are commonly used to infer a variety of evolutionary phenomena such as substitution rate heterogeneity, magnitude of purifying or positive selection and rapid gene evolution.

The *Daphnia pulex* genome appears to have an expanded number of genes compared to other fully sequenced invertebrates (Table 4-1). This phenomenon may be attributed to genome-wide duplication event(s) or from a higher rate of gene duplication relative to loss. These hypotheses were tested by examining the distribution of gene duplications over time using Ks as a proxy for time. Also, by examining gene duplicates that appear to evolve rapidly and/or under positive selective pressure, candidate loci were identified for further study of evolutionary significance. In addition, the demography of gene duplicates was examined by testing two identity cutoffs (40 and 60%) and parsing the data between an all-inclusive gene duplicate set (with large families) and single pair gene duplicates (family size=2). Patterns of evolution between dispersed and tandem duplicates and between *cis* and *trans* tandem duplicates are reported.

Taxon	Gene number	Reference
<i>D. pulex</i>	>32,000	Colbourne et al. In prep
<i>C. elegans</i>	~20,000	Stein et al. 2003
<i>A. mellifera</i>	17,000	Honey Bee Cons. 2006
<i>Drosophila</i> *	13-16,000	Drosophila 12 Genomes 2007
<i>A. gambiae</i>	~19,000	Holt et al. 2002

Table 4-1: Estimated gene content of fully sequenced invertebrates. *Multiple *Drosophila* species have been fully sequenced.

Methods

To characterize the gene duplicate catalogue of *Daphnia*, we conducted a “genome history” analysis, with a focus on highly related genes. In order to decipher patterns of molecular evolution among these gene duplicates, we compared all protein coding gene models (Frozen Gene Set v1.1, n=30,940) to each other using a modified installation of Genome History (Conant and Wagner 2002). By analyzing substitution patterns between gene copies and in the context of gene family assignments, we can better understand the process of gene copy evolution in *D. pulex*. This study includes other genomes for comparative insights. The entire gene catalogue from *C. elegans*, *A. thaliana* and *H. sapiens* were downloaded from Ensemble (www.ensembl.org). For genes with multiple splice variants, the largest gene was chosen.

Genome History (GH) detects and compares gene duplicates within a genome using a set of user-specified parameters and input. Here is an outline of the process as it was carried out on the *Daphnia pulex* v1.1 gene set:

1. All predicted protein sequences were WU-gapped-BLASTPed against each other. Self hits were thrown out. Hits $> e^{-10}$ proceed to next step.
2. Gene matches were aligned (ClustalW, Larkin et al. 2007) and minimum alignment length (100 amino acids) and percent identity (40 or 60%) cutoffs are applied. These strict settings minimized false relationships due to highly conserved motifs and narrowed the focus of this study to recent gene duplicates ($K_s < 1$).
3. Each aligned gene pair was then backtranslated using the nucleotide gene file. For each pair, K_a and K_s are calculated using the maximum likelihood, codon-

based model similar to Yang and Nielsen (2000). (For clarity, Ks means "Substitutions / Silent Site" and Ka means "Substitutions / Replacement Site".)

Zhang et al. (2004) argue that a genomic analysis of gene duplicates should include pairs with Ks values between 0.005 and 1 to avoid mistaking independently assembled alleles of the same gene ($Ks < 0.005$) and because accurate estimates of Ks are increasingly difficult at higher Ks values. In fact, our analysis of multiple gene duplicate pairs using 11 different analytical methods (Zhang et al. 2006) showed higher variance on estimates as distances surpassed $Ks \gg 1$. It should also be pointed out that the number of duplicate pair comparisons within a family is often higher than the actual number of duplication events since for any combination of genes there are $n(n-1)/2$ pairwise comparisons.

For this analysis we removed splice variants and transposable element genes. We also ran the same Ka and Ks calculations on a set of predicted pseudogenes generated using PseudoPipe (Zhang et al. 2006).

Depending on the specific analysis, we chose to include or exclude exact copy duplicates ($Ks=0$) and gene families >2 . It was important not to take all estimates of Ka and Ks at face value, but to consider the appropriateness of each estimate based on the assumptions underlying each test and the information content of each measurement. For example, a gene pair with one single non-synonymous substitution would not have a meaningful Ka/Ks value. Likewise, a balance exists where more information about the average mode of evolution can be gathered from diverged sequences up to a point at which saturation makes substitution rate estimates unreliable.

Results and Discussion

After 30,940 predicted *Daphnia pulex* genes (v1.1 frozen set) were run through the pipeline, GH output 36,186 gene pairs from 11,862 different genes. This proportion is higher than for many other vertebrate or invertebrate genomes, with a pronounced overabundance of very similar gene pairs ($K_s < 0.1$) (Figure 4-2).

Gene conversion is known to reduce variation in some large gene families (Liao 1999) such as rRNAs (Arnheim et al. 1980;), RNU2 (Paveltiz et al. 1995), histones (Coen et al. 1982), ubiquitin (Nenoi et al. 1998) as well as in non-coding repeat sequences (Elder and Turner 1995). To test the possibility of sequence homogenization among large families in *D. pulex*, average K_s for families of size 2, 3, 4-5, 6-99 and 100 were compared and not found to be significantly different (ANOVA, $p=0.265$). The abundance of very similar gene pairs in *D. pulex* ($K_s < 0.1$) appears to be a consequence of recent gene duplication rather than gene conversion *in large gene families* given the lack of detectable correlation between gene family size and K_s .

Birth rates of gene duplicates were calculated using the number of single-pair duplicates in the youngest cohort ($K_s < 0.01$), the baseline number of single copy genes and the synonymous substitution rate (K_s), giving units of duplications/gene/ K_s . Birth rates of nematodes and humans were comparable to those found in earlier studies (Lynch and Conery 2000, 2003). *D. pulex* appears to have a higher rate of gene duplication than other animals studied to date (Lynch 2007, Table 8.1).

While the observed number of new duplicates can be used to estimate a birth rate, it should be considered a downwardly biased estimate, since observed duplications may

represent a subset of events that rose to high frequency in the population, and were not purged by selection.

	Tucker	Lynch 2007
<i>D. pulex</i>	0.0085	N/A
<i>C. elegans</i>	0.0030	0.0028
<i>H. sapiens</i>	0.0055	0.0049

Table 4-2: Estimated birth rates for gene duplicates. Units are duplications/gene/Ks.

To test for the existence of gene duplicates where at least one member is evolving under overall positive selection, we compared the synonymous and nonsynonymous substitution rates between gene duplicates using the Ka/Ks test (Hurst 2002; Yang and Bielawski 2000). Based on this analysis, we were also able to identify a subset of recently duplicated genes that appear to be evolving in a positive mode ($Ka/Ks > 2$). We found 175 gene pairs with both a $Ka/Ks > 2$ and at least ten nonsynonymous substitutions. Functional analysis of positively evolving genes showed most to be of unknown function and without a homolog in Genbank (67%).

This analysis takes a conservative approach (minimum 100 AA alignment and $> 60\%$ identity) and is not meant to be an all inclusive analysis of all detectable paralogs. The power of the substitution analysis is in informing young gene pairs ($Ks \ll 0.7$). For instance, ancient gene duplicates are highly saturated with synonymous substitutions and are not as likely to play a role in recent genome evolution. Unlike other gene duplication analyses (Lynch and Conery 2000; Zhang et al. 2004), we have not removed larger gene families ($n > 5$) since the larger families, especially in *Daphnia*, are as much a part of the

recent evolutionary story as other pairs. However, we include an analysis in single pair families to gauge the effect of large families. It has been argued that the stochastic process of gene conversion biases larger families to smaller K_s values (Pan and Zhang 2007). This does not appear to be a detectable problem in the *Daphnia* gene set. However, it cannot be ruled out that many recent or exact gene copies are in fact the result of recent gene conversion and not recent gene duplication events. In fact, there is some evidence that *D. pulex* may undergo biased gene conversion during mitotic recombination as loss of heterozygosity (LOH) was observed on a short time scale in asexually propagated mapping lines (Omilian et al. 2006). Recent analysis of the *D. pulex* genes has also suggested high gene conversion rates (J. Colbourne, personal comm.) With the recent revelation that the aphid genome has many recent duplicates as well, it has been speculated that an expanded gene set may be related to an asexual reproductive mode. However, more data are needed.

Because the *D. pulex* gene set was generated from a combination of automated gene prediction algorithms and has not been manually and experimentally overhauled to the degree of older genome projects, there may be some gene predictions that are not actually protein coding genes. It has been estimated that over 20% of current human gene predictions may in fact not be protein coding genes (Clamp et al. 2007). This analysis takes the current predicted gene set at face value. However, there is reason to believe that the current gene number (~30,000) for *Daphnia* is an underestimate. Considering the relative phylogenetic isolation of *Daphnia* compared to other genome projects, it is not surprising that many genes without homology exist. In addition, expression analysis has recently been shown to support many *ab initio* models not yet included in the gene set.

Since *D. magna* (a distant relative of *D. pulex*) is in the process of being sequenced, all *D. pulex* models could be compared to a draft assembly yielding valuable information about gene model legitimacy (Figure 4-1). The current *Daphnia pulex* gene catalogue with previously excluded genes (green and purple) is summarized below (Figure 4-1).

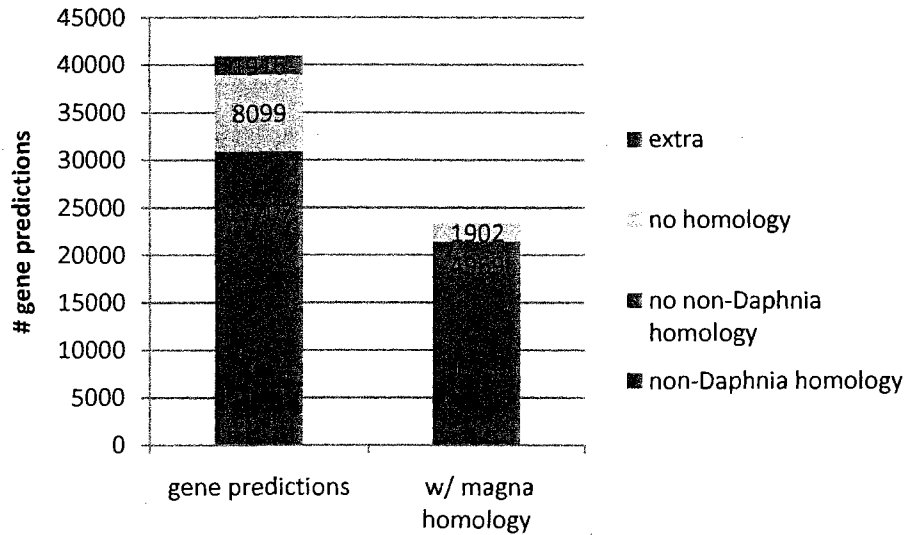


Figure 4-1: Homology among predicted genes in *D. pulex*. This analysis uses the Frozen Gene Set for *D. pulex* Draft 1.0 (left, bar, blue and red only).

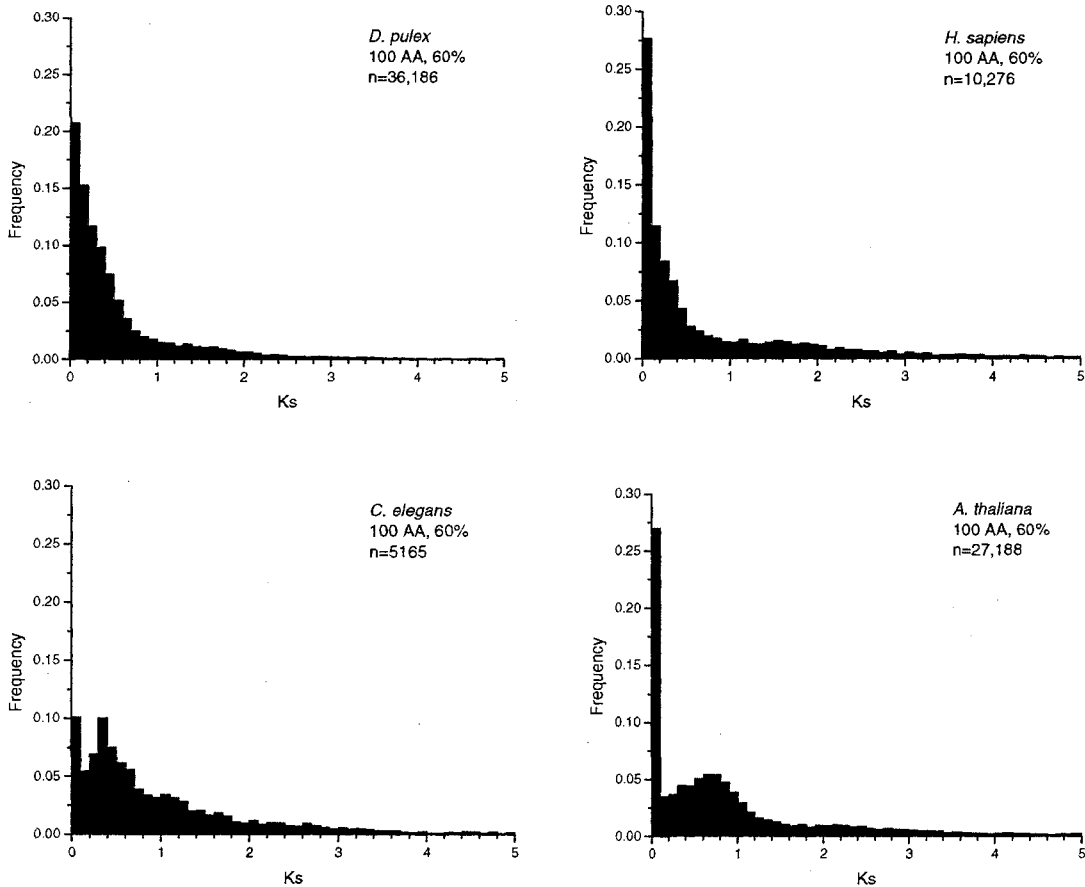


Figure 4-2. Age distribution of gene duplicates at >60% AA identity.

Figures 4-2 and 4-3 depict the frequency distribution of all gene pairs for four taxa plotted against their Ks values. The first panel (Figure 4-2) shows the distribution using the 60%, 100 amino acid minimum, while the second (Figure 4-3) shows the frequency distribution using the 40%, 100 amino acid identity. As would be expected, all taxa show an enrichment of older duplicates in the 40% panel (Figure 4-3). However, *D. pulex* (upper left quadrant of both panels) shows the smoothest decay curve using both cutoffs. This suggests a steady birth/death turnover over time. *H. sapiens*, a vertebrate, shows signs of ancient duplication activity when enriched for older duplicates (Figure 4-3, upper right). *C. elegans* (lower left in both panels), shows a younger explosion of duplication (Ks ~ 0.4), magnified in the first panel (Figure 4-2). *C. elegans* also appears to maintain many ancient duplicates (Figure 4-3, lower left).

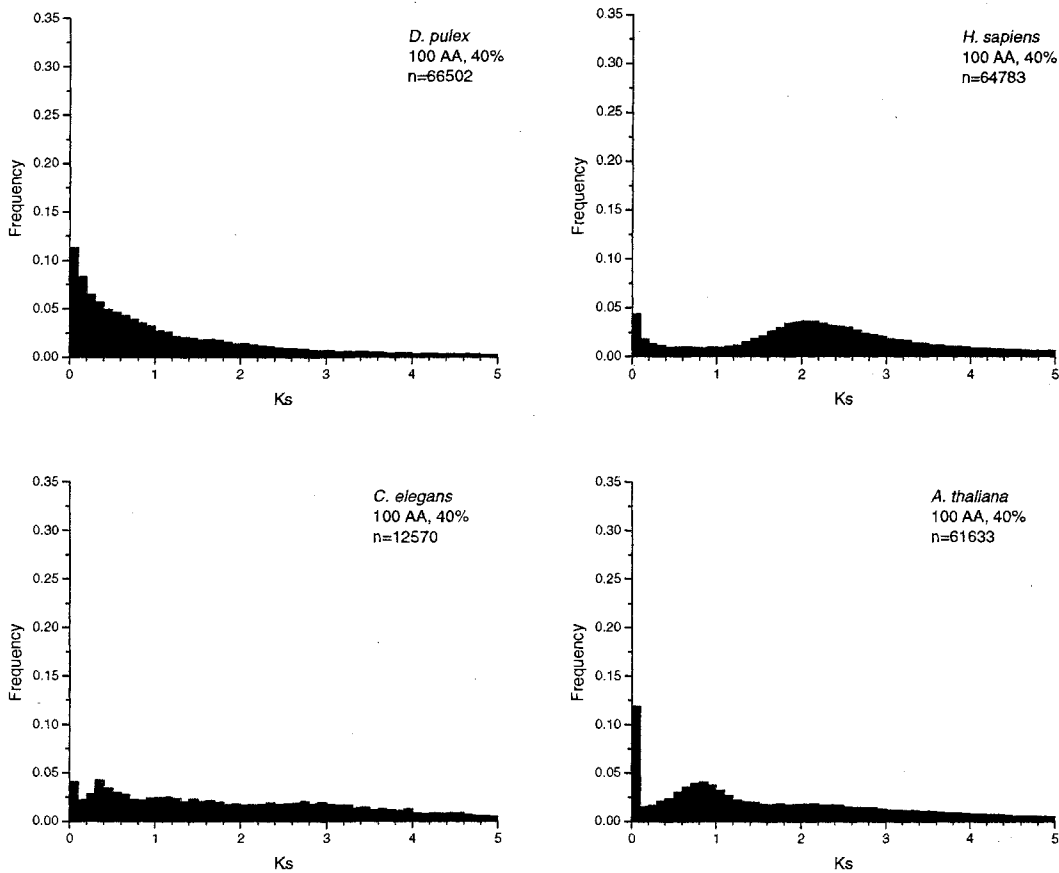


Figure 4-3: Age distribution of gene duplicates at >40% AA identity.

A. thaliana clearly shows a spike in gene duplication at $K_s \sim 0.75$, most likely due to an ancient polyploidization event (Maere et al. 2005). Overall, when comparing all gene duplicate pairs, *D. pulex* shows a high rate of birth with the most steady decay of duplicates, both in the panel enriched for recent duplicates (Figure 4-2) and with older duplicates (Figure 4-3). Because the number of duplicate comparisons overestimates the number of duplication events, single duplicate pairs (i.e. family size =2) were used to estimate birth rates for the four taxa. Figures 4-4 and 4-5 show the age distribution of single pair duplicates at the conservative (60% amino acid identity) and relaxed (40%) cutoffs, respectively.

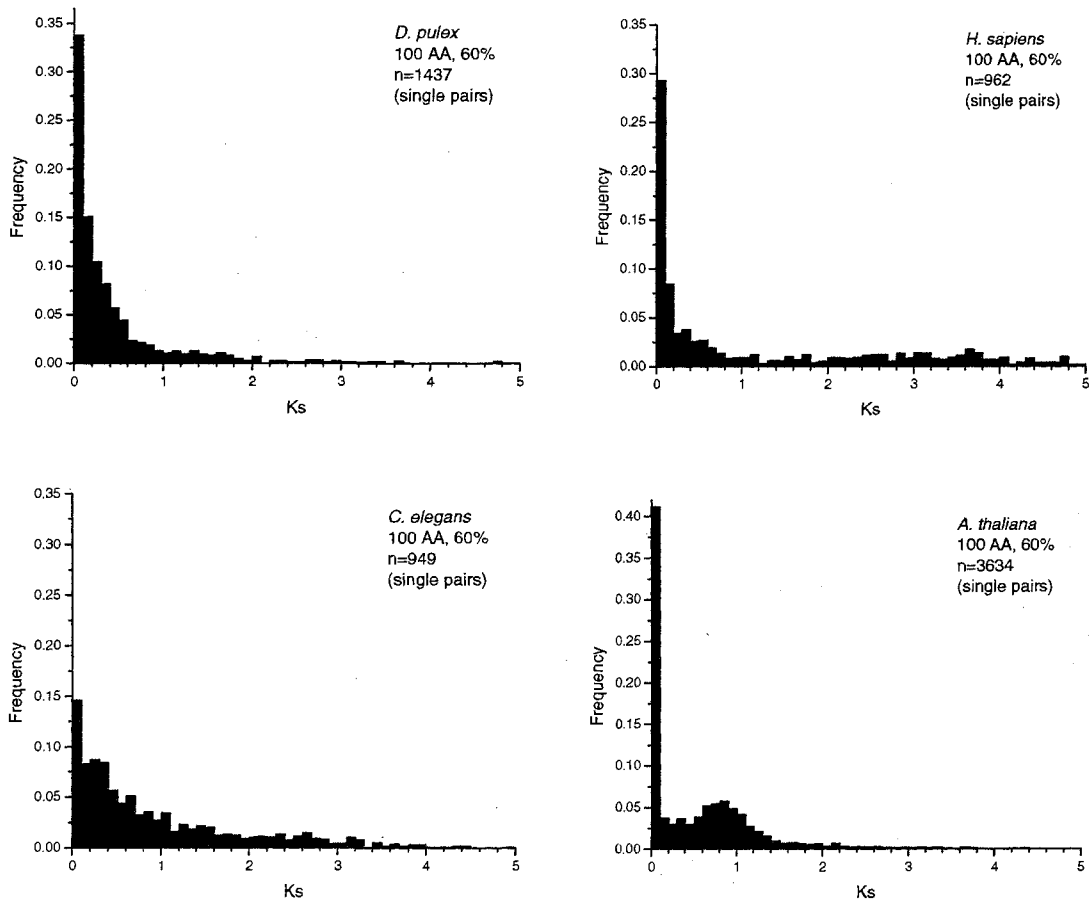


Figure 4-4: Distribution of single copy gene pairs. Average and median K_s values for single copy gene pairs are lower than for all gene pairs mostly due to a surplus of exact copy gene pairs.

Interestingly, using only single pair duplicates, the frequency distributions do not change significantly at the two cutoffs. Additionally, duplicate explosions described above are only apparent in *A. thaliana* when looking at single gene pairs. *A. thaliana* shows an extreme surplus of near-exact duplicates in all panels.

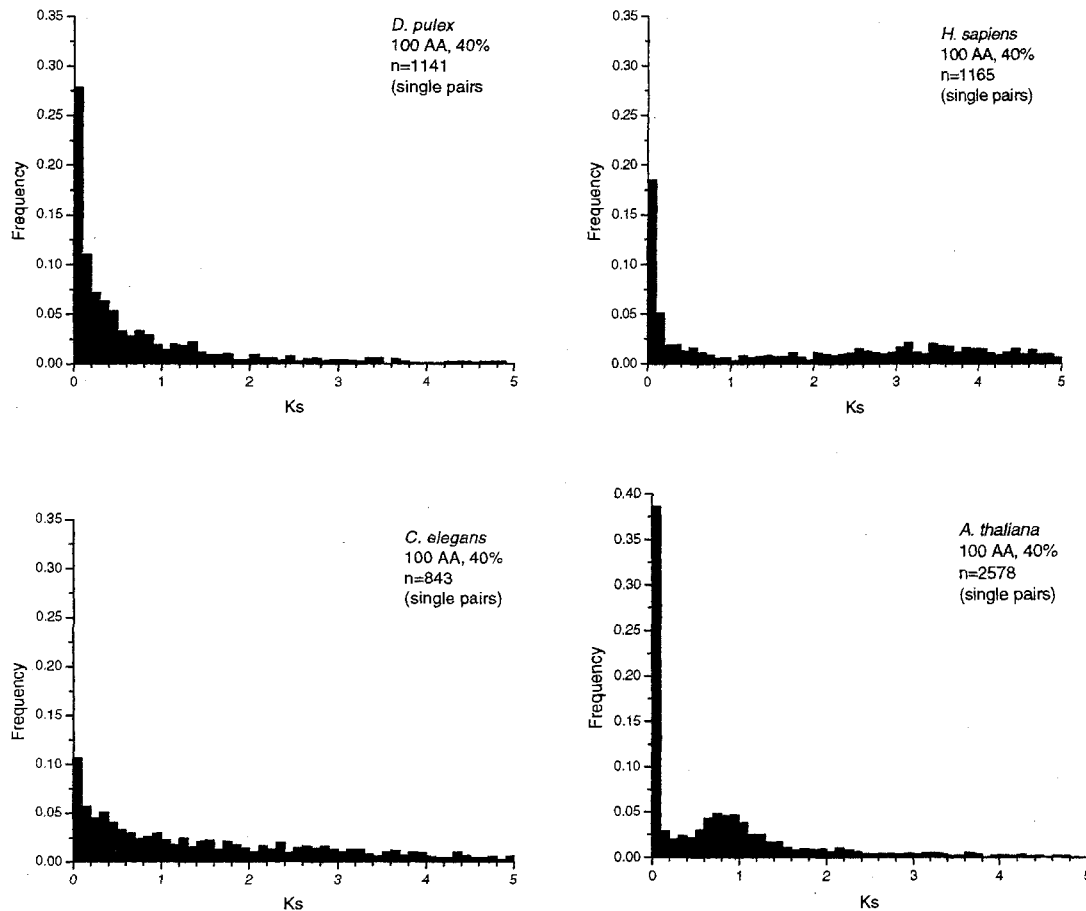


Figure 4-5: Frequency distribution of single gene duplicates using 40%, 100 amino acid identity cutoff.

In order to evaluate the magnitude of selective pressure on gene duplicate cohorts, the ratio of replacement substitution to silent substitution rate (Ka/Ks) was plotted against age (Ks) for each gene duplicate pair. Without purifying or positive selection, the rate of silent and replacement substitution are expected to be the same. Therefore, under a neutral model of sequence evolution, Ka/Ks is expected to be around 1 (Nei and Kumar 2000; Yang and Bielawski 2000).

The Ka/Ks statistic employed here is an average value for an aligned portion of a gene (>100 AA). Extreme recent selection on a small portion of sites in a gene would be undetectable. Also, for any given gene pair, it is not possible to identify the gene under

selection without polarizing the analysis. For instance, if a single member of a gene family is undergoing positive selection, it will show a high K_a/K_s when paired with all members of its family.

Theories regarding the fate of gene duplicates predict that, on average, younger gene duplicate pairs are expected to evolve with reduced selective pressure due to redundancy (Wagner 2002). In fact, most genome-wide studies to date support this generalization (Lynch and Conery 2003; Zhang et al. 2004), although some have questioned the extent of relaxation (Kondroshov et al. 2002). *D. pulex* is no exception. The majority of gene duplicates evolve with intense purifying selection ($K_a/K_s \ll 1$). Single pair gene duplicates in the *D. pulex* genome were found to have larger ranges of K_a/K_s at lower K_s values (Figure 4-6).

General patterns of evolution between gene pairs are depicted in Figures 4-6 – 4-8. Single gene pairs (family size=2, Figures 4-6), all gene pairs (family sizes >1, Figure 4-7) and predicted pseudogenes (Figures 4-8) are plotted separately. Once disabled, pseudogenes are expected to evolve neutrally. However, the signature of purifying selection can be detected in young pseudogenes since they may have been functional for a period of time after duplication (Figures 4-8). However, on average pseudogene pairs have an elevated K_a/K_s values (Table 4-2). Additionally, when comparing pseudogenes to real genes, a significant portion of the substitutions observed *occurred* in the real gene, therefore giving the pattern of divergence between the pairs the signature of purifying selection. Relative levels of purifying selection between gene-gene and pseudogene-gene comparisons are therefore more informative. Dead-on-arrival duplicates are expected to neutrally evolve, even at a young age.

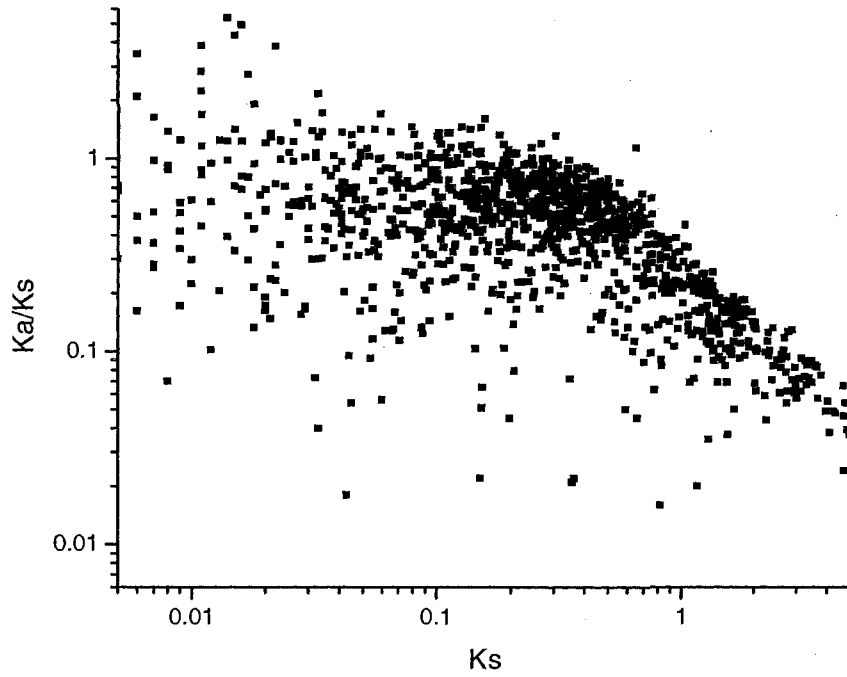


Figure 4-6: Selection intensity (Ka/Ks) and age of single gene pairs.

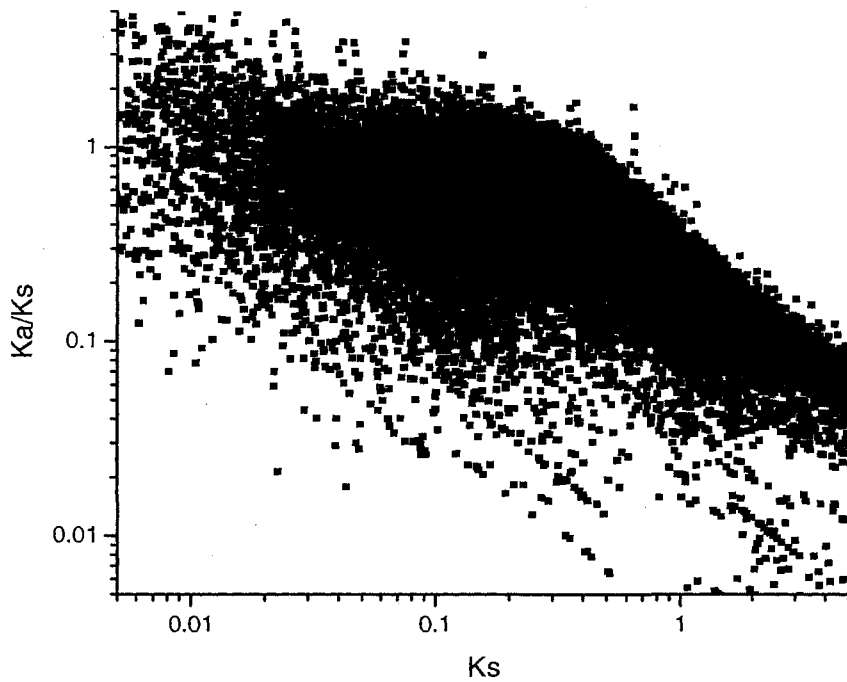


Figure 4-7: Selection intensity (Ka/Ks) and age of all gene pairs

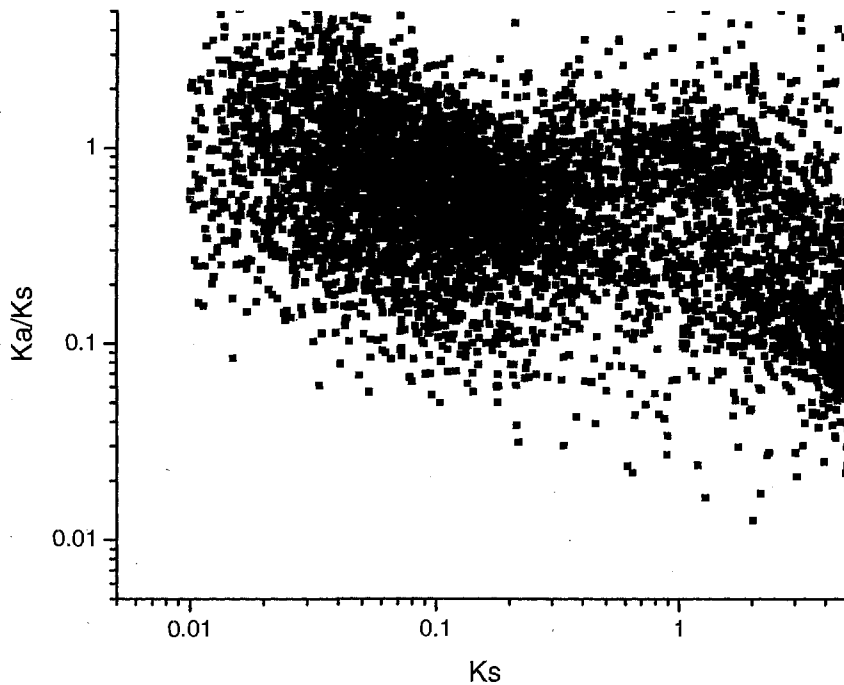


Figure 4-8: Selection intensity (Ka/Ks) and age of pseudogene-gene pairs.

Pseudogenes were compared to their closest living relative in the predicted *D. pulex* gene catalogue. Pseudogene pairs are younger on average and show a steeper decay in frequency when plotted along Ks (Figure 4-9). While this would be expected if pseudogenes accumulate deleterious mutations rapidly (i.e. higher death rate), there may be an ascertainment bias towards discovering younger pseudogenes.

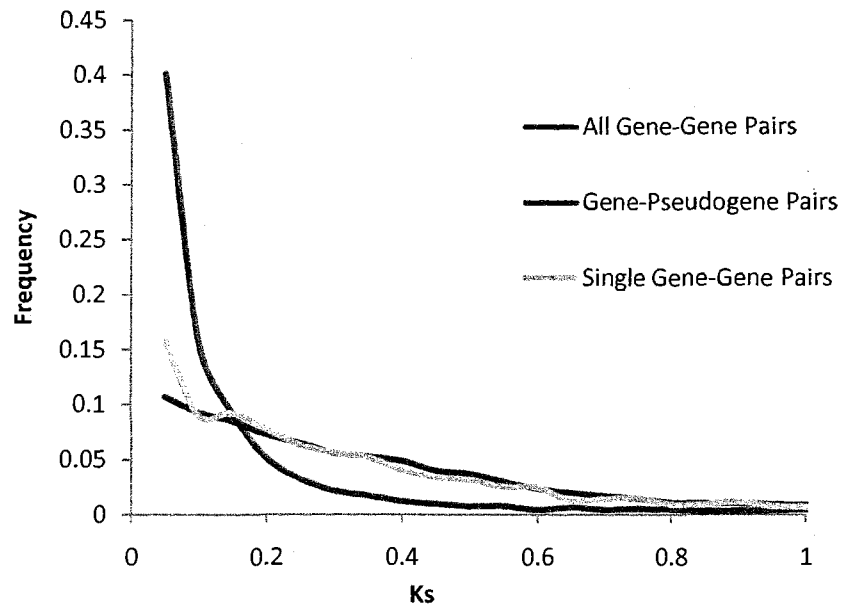


Figure 4-9: Distribution of Ks for gene pairs.

By comparing different ranges of Ks, comparative estimates were made between all, single and pseudogene pairs (Table 4-3).

	Ks		Ka		Ka/Ks		N pairs
	Avg	median	avg	median	avg	median	
0 < Ks < 5							
all pairs	0.607	0.312	0.153	0.143	0.252	0.458	35520
single pairs	0.469	0.206	0.137	0.127	0.292	0.617	1423
psi pairs	0.485	0.078	0.649	0.057	1.338	0.731	9274
.01 < Ks < 5							
all pairs	0.632	0.331	0.159	0.151	0.252	0.456	34105
single pairs	0.554	0.278	0.148	0.139	0.267	0.500	1202
psi pairs	0.622	0.125	0.497	0.0807	0.799	0.646	7234
.01 < Ks < 1							
all pairs	0.305	0.248	0.144	0.124	0.472	0.500	27594
single pairs	0.278	0.21	0.133	0.111	0.478	0.529	1005
psi pairs	0.165	0.091	0.211	0.058	1.279	0.637	5857

Table 4-3: Average and median values for substitution analysis of all, single and pseudogene pairs and varying ranges of Ks.

For K_s values 0-5 (Table 4-2, top), pseudogenes have a relatively high rate of replacement substitution (K_a). Single pairs and pseudogenes tend to be younger than all gene pairs.

K_s values that exclude exact copy duplicates ($K_s > 0.01$, middle rows, Table 4-2) show similar comparative values. However, when examining younger pairs ($K_s < 1$, bottom rows, Figure 4-2), values are considered more reliable. Median and average values converge between values for all taxa at $K_s < 1$, suggesting more normal distributions of K_a and K_s .

Recent gene duplicates ($K < 0.1$) were functionally annotated using the JGI-generated KOG report for the predicted gene set. When compared to all genes, recent duplicates were enriched for post-translational modification and chromatin structure and underrepresented in the general function category (Figure 4-10, chi sq. $p = .023$).

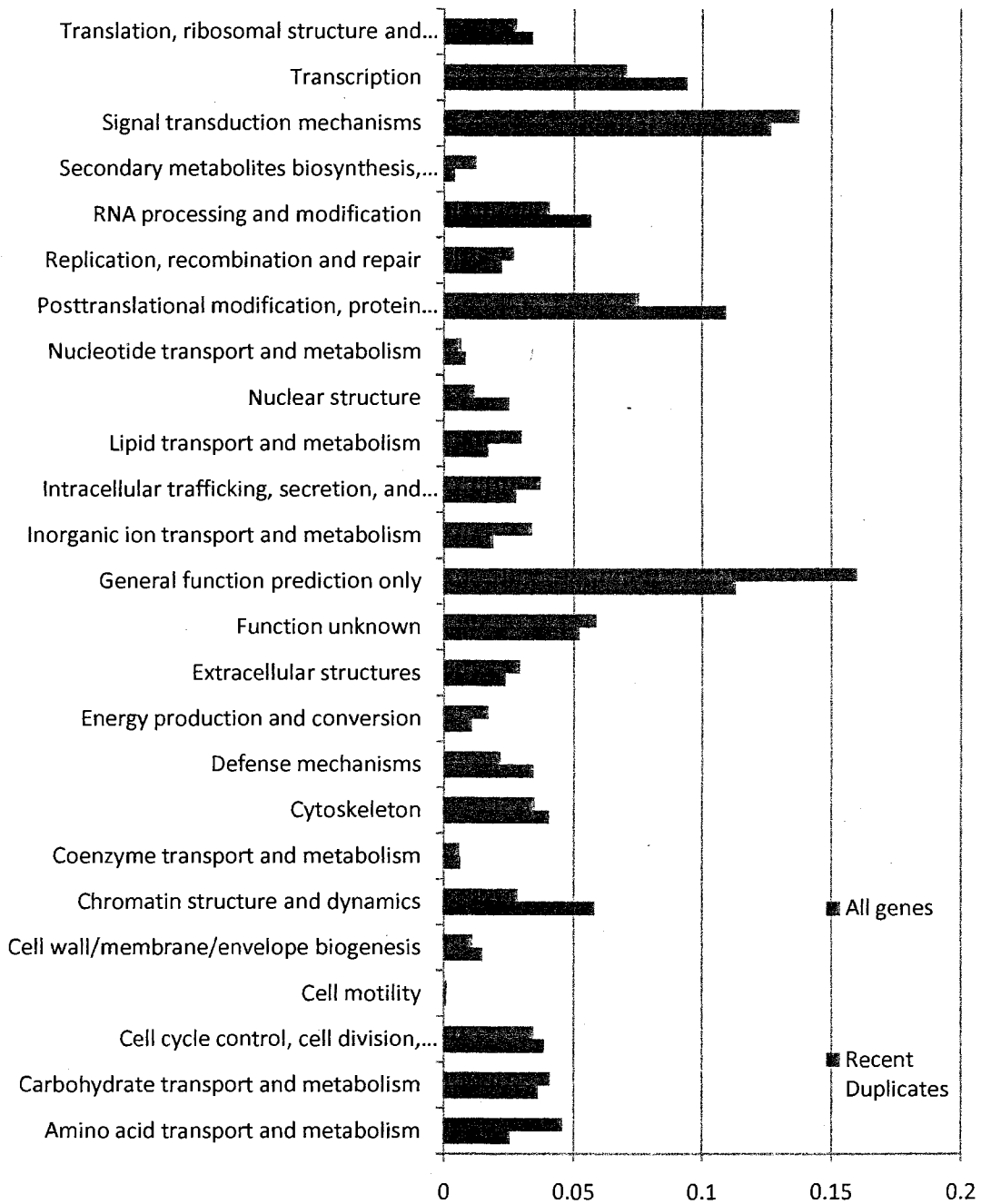


Figure 4-10: Comparison of KOG classes all genes vs. recent duplicates ($K_s < 0.1$).

Tandem vs. dispersed gene duplicates in *D. pulex* genome

In order to gauge the effect of physical and spatial orientation of duplicate pairs on patterns of evolution, duplicate pairs were classified into four groups: Tandem (< 20 kb apart) in *cis* (same coding strand), tandem in *trans* (on opposite strands), dispersed on the same scaffold and dispersed on different scaffolds. A general pattern of decay in the number of functional gene duplicates over time is apparent in all categories (Figure 4-11). However, the tandem duplicates appear older on average ($K_s > 0.750$), with *trans* duplicates showing the largest signature of purifying selection ($K_a/K_s = 0.170$, Table 4-4). The relative youth of dispersed duplicates suggests that there may be more gene conversion with these families or that the general mechanism of gene duplication is dispersive and tandem duplicates are special cases of co-regulated gene families.

Spatial relationship	Ks	Ka	Ka/Ks	# gene pairs
Tandem-cis <20kb	0.770	0.166	0.216	1704
Tandem-trans<20kb	0.750	0.127	0.170	363
Dispersed >20kb	0.597	0.136	0.228	2767
Dispersed	0.597	0.154	0.258	30663
N50 Tandem-cis <20kb	0.932	0.175	0.188	1462
N50 Tandem-trans<20kb	0.961	0.128	0.133	281
N50 Dispersed >20kb	0.826	0.142	0.172	2622
N50 Dispersed	0.909	0.172	0.189	19454

Table 4-4: Substitution analysis between gene duplicate pairs in four spatial categories among all gene pairs (top) and those in the N50 (bottom).

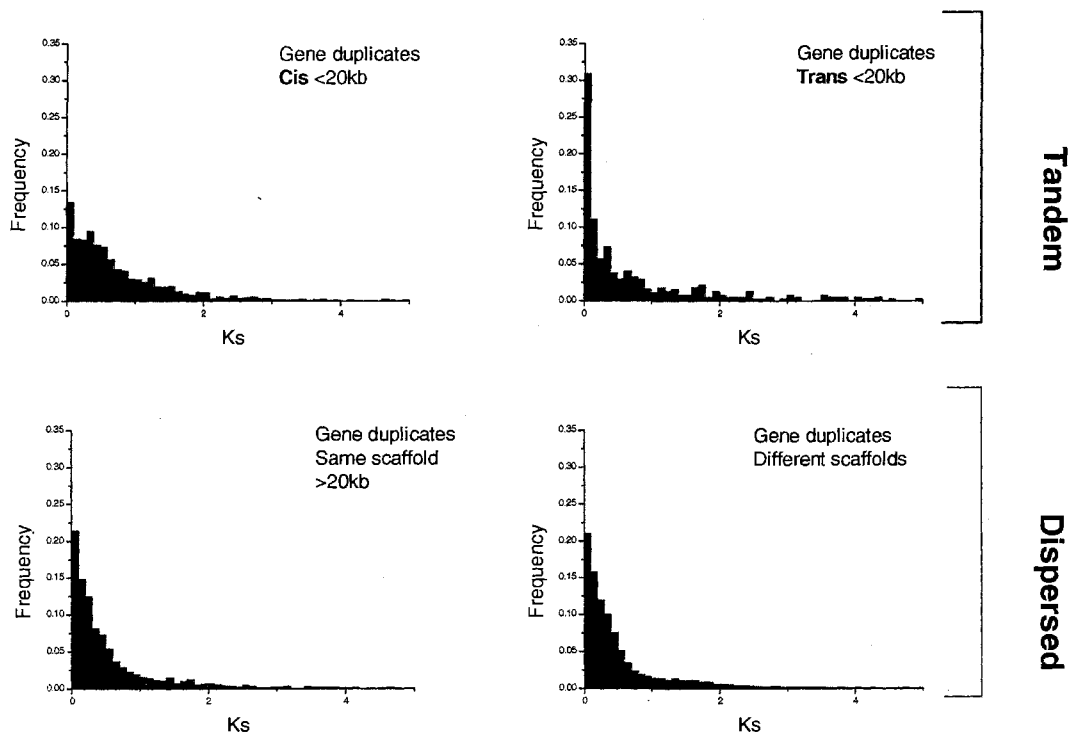


Figure 4-11: Frequency distribution of all gene pairs along Ks in four spatial categories.

However, restricting the dispersed class to those from the best-assembled half of the genome ($N_{50}=100$ scaffolds) (Figure 4-12) brings average age (Ks) from 0.597 to 0.909 from 7793 dispersed, different scaffold gene pairs (Table 4-4). This suggests that many pairs involving orphans (gene on micro-scaffolds) may be young. Although many scaffolds outside the N_{50} are quite large (e.g. scaffold 200 = 184,404 bp), there are hundreds of scaffolds that contain single or few protein-coding genes. Pan and Zheng (2008) estimate that 10-20% of genes in most eukaryotes are in tandem. *Daphnia* meets this expectation. However, the observation that tandem duplicates are older on average suggests that duplicate birth and/or gene conversion, forces that would lead to pairs with low Ks, are not necessarily biased towards tandem genes.

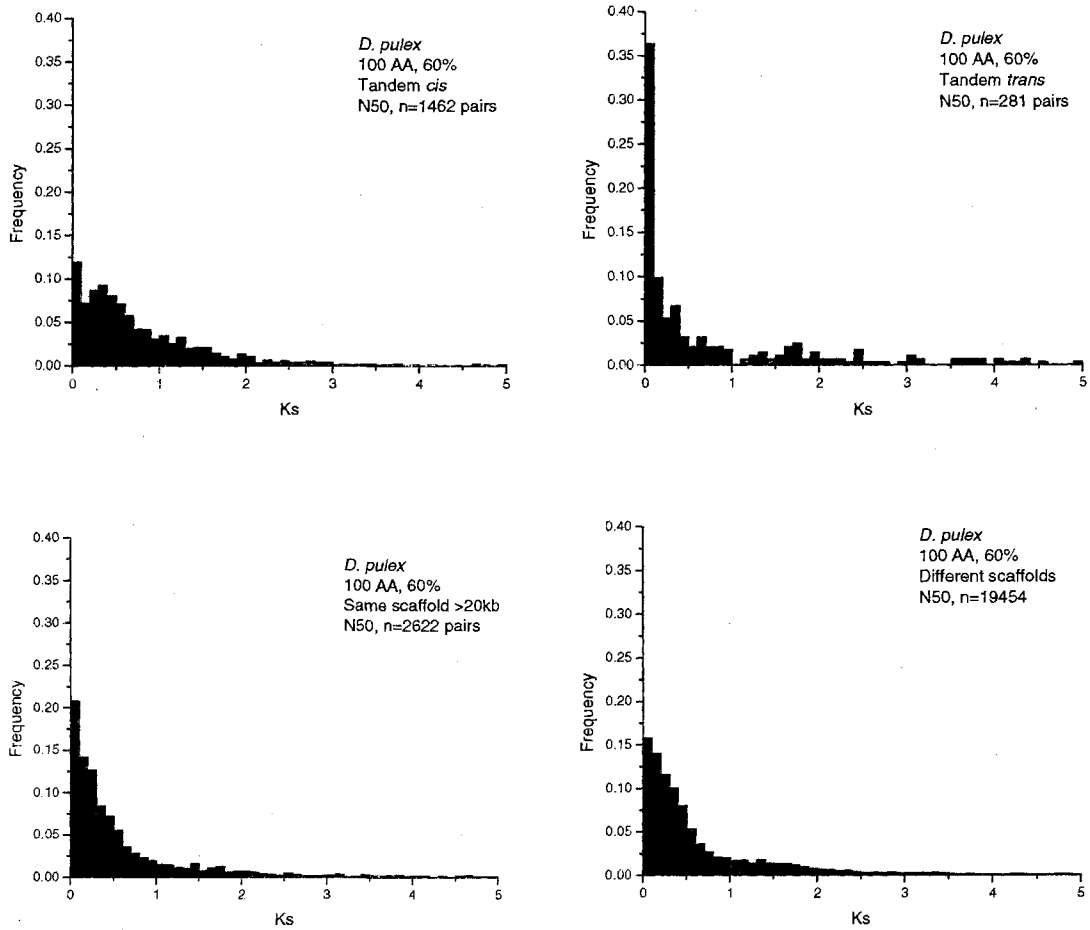


Figure 4-12: Frequency distribution of N50 gene pairs along Ks in four spatial categories.

While *cis* and *trans* tandem duplicates have similar birth rates (0.00349 and 0.00391 duplications/gene/0.01 Ks, respectively), *trans* duplicates have higher initial death rates. The retention of *cis* duplicates (Figure 4-13) is responsible for the expanded number (> 5X) of *cis* vs. *trans* duplicates.

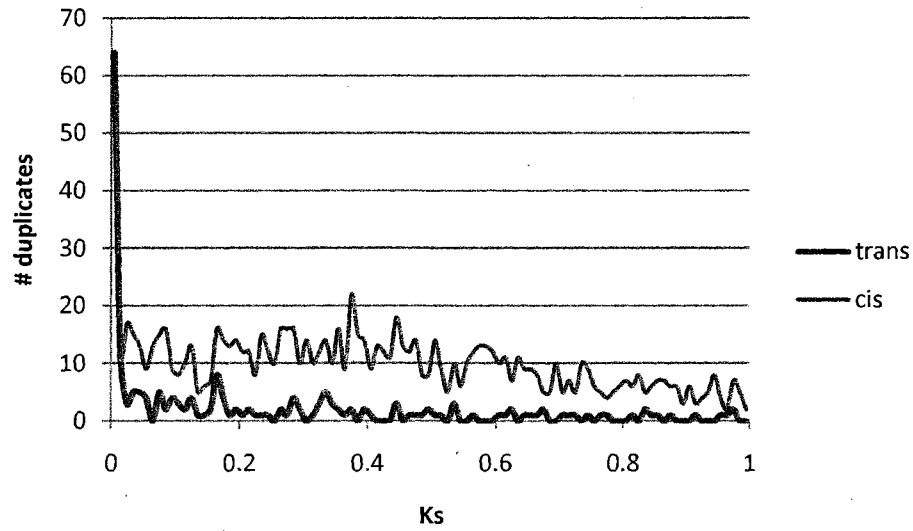


Figure 4-13: Number of *cis* and *trans* duplicates over time.

Acknowledgements

Phil Hatcher set up a modified installation of Genome History, ran the gene sets. Phil Hatcher and Kelley Thomas contributed to the experimental design. Jeong-Hyeon Choi (IU) generated the PseudoPipe data.

LIST OF REFERENCES

1. Akashi, H. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. *Genetics* **139**, 1067-76(1995).
2. Altshuler, D., Daly, M.J. & Lander, E.S. Genetic Mapping in Human Disease. *Science* **322**, 881-888(2008).
3. Anderson, A.D. & Weir, B.S. A maximum-likelihood method for the estimation of pairwise relatedness in structured populations. *Genetics* **176**, 421-40(2007).
4. Andersson, G.E. & Kurland, C.G. An extreme codon preference strategy: codon reassignment. *Mol Biol Evol* **8**, 530-44(1991).
5. Andolfatto, P. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* **437**, 1149-52(2005).
6. Andolfatto, P. Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melanogaster* genome. *Genome Res* **17**, 1755-62(2007).
7. Anisimova, M. & Liberles, D.A. The quest for natural selection in the age of comparative genomics. *Heredity* **99**, 567-79(2007).
8. Aquadro, C.F. & Greenberg, B.D. Human mitochondrial DNA variation and evolution: analysis of nucleotide sequences from seven individuals. *Genetics* **103**, 287-312(1983).
9. Arnheim, N. et al. Molecular evidence for genetic exchanges among ribosomal genes on nonhomologous chromosomes in man and apes. *Proc Natl Acad Sci U S A* **77**, 7323-7(1980).
10. Asakawa, S. et al. Strand-specific nucleotide composition bias in echinoderm and vertebrate mitochondrial genomes. *J Mol Evol* **32**, 511-20(1991).
11. Baer, C.F., Miyamoto, M.M. & Denver, D.R. Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nat Rev Genet* **8**, 619-31(2007).
12. Begun, D.J. & Aquadro, C.F. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356**, 519-20(1992).
13. Begun, D.J. et al. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol* **5**, e310(2007).
14. Beisswanger, S. & Stephan, W. Evidence that strong positive selection drives neofunctionalization in the tandemly duplicated polyhomeotic genes

- in *Drosophila*. *Proc Natl Acad Sci U S A* **105**, 5447-52(2008).
15. Belle, E.M. et al. An investigation of the variation in the transition bias among various animal mitochondrial DNA. *Gene* **355**, 58-66(2005).
 16. Belshaw, R. & Bensasson, D. The rise and falls of introns. *Heredity* **96**, 208-13(2006).
 17. Benfey, P.N. & Mitchell-Olds, T. From genotype to phenotype: systems biology meets natural variation. *Science* **320**, 495-7(2008).
 18. Berget, S.M., Moore, C. & Sharp, P.A. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci U S A* **74**, 3171-5(1977).
 19. Bernardi, G. The neoselectionist theory of genome evolution. *Proc Natl Acad Sci U S A* **104**, 8385-90(2007).
 20. Betancourt, A.J. & Presgraves, D.C. Linkage limits the power of natural selection in *Drosophila*. *Proc Natl Acad Sci U S A* **99**, 13616-20(2002).
 21. Beye, M. et al. Exceptionally high levels of recombination across the honey bee genome. *Genome Res* **16**, 1339-44(2006).
 22. Bielawski, J.P. & Gold, J.R. Mutation patterns of mitochondrial H- and L-strand DNA in closely related Cyprinid fishes. *Genetics* **161**, 1589-97(2002).
 23. Bierne, N. & Eyre-Walker, A. The genomic rate of adaptive amino acid substitution in *Drosophila*. *Mol Biol Evol* **21**, 1350-60(2004).
 24. Birky, C.W. & Walsh, J.B. Effects of linkage on rates of molecular evolution. *Proc Natl Acad Sci U S A* **85**, 6414-8(1988).
 25. Biswas, S. & Akey, J.M. Genomic insights into positive selection. *Trends Genet* **22**, 437-46(2006).
 26. Björnerfeldt, S., Webster, M.T. & Vilà, C. Relaxation of selective constraint on dog mitochondrial DNA following domestication. *Genome Res* **16**, 990-4(2006).
 27. Blouin, M.S. et al. Substitution bias, rapid saturation, and the use of mtDNA for nematode systematics. *Mol Biol Evol* **15**, 1719-27(1998).
 28. Boore, J.L. Animal mitochondrial genomes. *Nucleic Acids Res* **27**, 1767-80(1999).
 29. Boyko, A.R. et al. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet* **4**, e1000083(2008).
 30. Brown, W.M. The mitochondrial genome of animals. *Molecular Evolutionary Genetics* 95-130
 31. Brown, W.M., George, M. & Wilson, A.C. Rapid evolution of animal mitochondrial DNA. *Proc Natl Acad Sci U S A* **76**, 1967-71(1979).
 32. Brown, W.M. et al. Mitochondrial DNA sequences of primates: tempo and mode of evolution. *J Mol Evol* **18**, 225-39(1982).
 33. Bussell, J.J. et al. Human polymorphism and human-chimpanzee divergence in pseudoautosomal region correlate with local recombination rate. *Gene* **368**, 94-100(2006).
 34. Bustamante, C.D. et al. Natural selection on protein-coding genes in the human genome. *Nature* **437**, 1153-7(2005).
 35. Butlin, R. Population genomics and speciation. *Genetica*

(2008).doi:10.1007/s10709-008-9321-3

36. Cai, J.J. et al. Pervasive Hitchhiking at Coding and Regulatory Sites in Humans. *PLoS Genet* **5**, e1000336(2009).
37. Cargill, M. et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* **22**, 231-238(1999).
38. Castric, V. et al. Repeated adaptive introgression at a gene under multiallelic balancing selection. *PLoS Genet* **4**, e1000168(2008).
39. Chamary, J.V., Parmley, J.L. & Hurst, L.D. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet* **7**, 98-108(2006).
40. Charlesworth, D. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet* **2**, e64(2006).
41. Charlesworth, J. & Eyre-Walker, A. The McDonald-Kreitman test and slightly deleterious mutations. *Mol Biol Evol* **25**, 1007-15(2008).
42. Charlesworth, B., Morgan, M.T. & Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289-303(1993).
43. Chen, H. & Blanchette, M. Detecting non-coding selective pressure in coding regions. *BMC Evol Biol* **7 Suppl 1**, S9(2007).
44. Chen, X.J. & Butow, R.A. The organization and inheritance of the mitochondrial genome. *Nat Rev Genet* **6**, 815-825(2005).
45. Chen, J. et al. Gene conversion: mechanisms, evolution and human disease. *Nat Rev Genet* **8**, 762-75(2007).
46. Cheng, T. et al. Mining single nucleotide polymorphisms from EST data of silkworm, *Bombyx mori*, inbred strain Dazao. *Insect Biochem Mol Biol* **34**, 523-30(2004).
47. Chou, H.H. & Holmes, M.H. DNA sequence quality trimming and vector removal. *Bioinformatics* **17**, 1093-104(2001).
48. Chow, L.T. et al. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* **12**, 1-8(1977).
49. Clamp, M. et al. Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci U S A* **104**, 19428-33(2007).
50. Clary, D.O. & Wolstenholme, D.R. *Drosophila* mitochondrial DNA: conserved sequences in the A + T-rich region and supporting evidence for a secondary structure model of the small ribosomal RNA. *J Mol Evol* **25**, 116-25(1987).
51. Clayton, D.A. Transcription and replication of mitochondrial DNA. *Hum Reprod* **15 Suppl 2**, 11-7(2000).
52. Coates, B.S. et al. Partial mitochondrial genome sequences of *Ostrinia nubilalis* and *Ostrinia furnicalis*. *Int J Biol Sci.* **1**, 13-18(2005).
53. Coen, E., Strachan, T. & Dover, G. Dynamics of concerted evolution of ribosomal DNA and histone gene families in the melanogaster species subgroup of *Drosophila*. *J Mol Biol* **158**, 17-35(1982).
54. Colbourne, J.K. & Hebert, P.D. The systematics of North American *Daphnia* (Crustacea: Anomopoda): a molecular phylogenetic approach. *Philos Trans*

R Soc Lond B Biol Sci **351**, 349-60(1996).

55. Colbourne, J.K., Hebert, P.D. & Taylor, D.J. Evolutionary Origin of Phenotypic Diversity in *Daphnia*. *Molecular Evolution and Adaptive Radiation* 163-188(2000).
56. Colbourne, J.K. et al. Sampling *Daphnia*'s expressed genes: preservation, expansion and invention of crustacean genes with reference to insect genomes. *BMC Genomics* **8**, 217(2007).
57. Collins, D.W. & Jukes, T.H. Rates of transition and transversion in coding sequences since the human-rodent divergence. *Genomics* **20**, 386-96(1994).
58. Comeron, J.M., Williford, A. & Kliman, R.M. The Hill-Robertson effect: evolutionary consequences of weak selection and linkage in finite populations. *Heredity* **100**, 19-31(2008).
59. Conant, G.C. & Wagner, A. GenomeHistory: a software tool and its application to fully sequenced genomes. *Nucleic Acids Res* **30**, 3378-86(2002).
60. Conant, G.C. & Wagner, A. Asymmetric sequence divergence of duplicate genes. *Genome Res* **13**, 2052-8(2003).
61. Coulombe-Huntington, J. & Majewski, J. Characterization of intron loss events in mammals. *Genome Res* **17**, 23-32(2007).
62. Coulombe-Huntington, J. & Majewski, J. Intron loss and gain in *Drosophila*. *Mol Biol Evol* **24**, 2842-50(2007).
63. Cousyn, C. et al. Rapid, local adaptation of zooplankton behavior to changes in predation pressure in the absence of neutral genetic changes. *Proc Natl Acad Sci U S A* **98**, 6256-60(2001).
64. Crawford, D.C. et al. Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nat Genet* **36**, 700-6(2004).
65. Crease, T.J. The complete sequence of the mitochondrial genome of *Daphnia pulex* (Cladocera: Crustacea). *Gene* **233**, 89-99(1999).
66. Cristescu, M.E.A. et al. A microsatellite-based genetic linkage map of the waterflea, *Daphnia pulex*: On the prospect of crustacean genomics. *Genomics* **88**, 415-30(2006).
67. Cruz, F., Vilà, C. & Webster, M.T. The legacy of domestication: accumulation of deleterious mutations in the dog genome. *Mol Biol Evol* **25**, 2331-6(2008).
68. Curole & Kocher Mitogenomics: digging deeper with complete mitochondrial genomes. *Trends Ecol Evol* **14**, 394-398(1999).
69. Cutter, A.D. & Payseur, B.A. Selection at linked sites in the partial selfer *Caenorhabditis elegans*. *Mol Biol Evol* **20**, 665-73(2003).
70. Cutter, A.D., Baird, S.E. & Charlesworth, D. High nucleotide polymorphism and rapid decay of linkage disequilibrium in wild populations of *Caenorhabditis remanei*. *Genetics* **174**, 901-13(2006).
71. Cutter, A.D., Wasmuth, J.D. & Washington, N.L. Patterns of molecular evolution in *Caenorhabditis* preclude ancient origins of selfing. *Genetics* **178**, 2093-104(2008).

72. Dehal, P. et al. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* **298**, 2157-67(2002).
73. Demuth, J.P. et al. The evolution of mammalian gene families. *PLoS ONE* **1**, e85(2006).
74. Denver, D.R. et al. High direct estimate of the mutation rate in the mitochondrial genome of *Caenorhabditis elegans*. *Science* **289**, 2342-4(2000).
75. DeSalle, R. et al. Tempo and mode of sequence evolution in mitochondrial DNA of Hawaiian *Drosophila*. *J Mol Evol* **26**, 157-64(1987).
76. Diller, K.C., Gilbert, W.A. & Kocher, T.D. Selective sweeps in the human genome: a starting point for identifying genetic differences between modern humans and chimpanzees. *Mol Biol Evol* **19**, 2342-5(2002).
77. Dobzhansky, T. A review of some fundamental concepts and problems of population genetics. *Cold Spring Harb Symp Quant Biol* **20**, 1-15(1955).
78. Dobzhansky, T. Nothing in Biology Makes Sense Except in the Light of Evolution. *The American Biology Teacher* **35**, 125-129(1973).
79. Donnelly, P. Progress and challenges in genome-wide association studies in humans. *Nature* **456**, 728-731(2008).
80. Drosophila 12 Genomes Consortium Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450**, 203-218(2007).
81. Dudycha, J.L. & Tessier, A.J. Natural Genetic Variation of Life Span, Reproduction, and Juvenile Growth in *Daphnia*. *Evolution* **53**, 1744-1756(1999).
82. Dunn, C.W. et al. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* **452**, 745-9(2008).
83. Eads, B.D., Andrews, J. & Colbourne, J.K. Ecological genomics in *Daphnia*: stress responses and environmental sex determination. *Heredity* **100**, 184-190(2007).
84. Elder, J.F. & Turner, B.J. Concerted evolution of repetitive DNA sequences in eukaryotes. *Q Rev Biol* **70**, 297-320(1995).
85. Ellegren, H. Comparative genomics and the study of evolution by natural selection. *Molecular Ecology* **17**, 4586-4596(2008).
86. Ellegren, H. & Sheldon, B.C. Genetic basis of fitness differences in natural populations. *Nature* **452**, 169-75(2008).
87. Emerson, J.J. et al. Natural Selection Shapes Genome-Wide Patterns of Copy-Number Polymorphism in *Drosophila melanogaster*. *Science* **320**, 1629-1631(2008).
88. Escriva, H. et al. Neofunctionalization in Vertebrates: The Example of Retinoic Acid Receptors. *PLoS Genet.* **2**, e102(2006).
89. Eyre-Walker, A. The genomic rate of adaptive evolution. *Trends in Ecology & Evolution* **21**, 569-575(2006).
90. Faith, J.J. & Pollock, D.D. Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. *Genetics* **165**, 735-45(2003).

91. Farris, J. Phylogenetic analysis under Dollo's law. *Syst. Zool.* **26**, 220-223(1977).
92. Fay, J.C., Wyckoff, G.J. & Wu, C. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature* **415**, 1024-6(2002).
93. Feder, M.E. & Mitchell-Olds, T. Evolutionary and ecological functional genomics. *Nat Rev Genet* **4**, 651-7(2003).
94. Fedorov, A. et al. Mystery of Intron Gain. *Genome Res.* **13**, 2236-2241(2003).
95. Ferris, S.D., Wilson, A.C. & Brown, W.M. Evolutionary tree for apes and humans based on cleavage maps of mitochondrial DNA. *Proc Natl Acad Sci U S A* **78**, 2432-6(1981).
96. Force, A. et al. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531-45(1999).
97. Ford, E. *Genetic Polymorphism*. (Faber and Faber: London, 1965).
98. Fox, A.K., Tuch, B.B. & Chuang, J.H. Measuring the prevalence of regional mutation rates: an analysis of silent substitutions in mammals, fungi, and insects. *BMC Evol Biol.* **8**, 186(2008).
99. Frank, A.C. & Lobry, J.R. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene* **238**, 65-77(1999).
100. Frazer, K.A. et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-61(2007).
101. Freeland, S.J. et al. Early fixation of an optimal genetic code. *Mol Biol Evol* **17**, 511-8(2000).
102. Fumagalli, M. et al. Widespread balancing selection and pathogen-driven selection at blood group antigen genes. *Genome Res* (2008).doi:gr.082768.108
103. Gaffney, D.J. & Keightley, P.D. The scale of mutational variation in the murid genome. *Genome Res* **15**, 1086-94(2005).
104. Gilbert, W. Why genes in pieces? *Nature* **271**, 501(1978).
105. Gilbert, M.T.P. et al. The emergence of HIV/AIDS in the Americas and beyond. *Proc Natl Acad Sci U S A* **104**, 18566-70(2007).
106. Gillespie, J.H. Genetic drift in an infinite population. The pseudohitchhiking model. *Genetics* **155**, 909-19(2000).
107. Gillespie, J.H. *Population genetics: A Concise Guide*. (Johns Hopkins University Press: Baltimore, MD, 2004).
108. Gladyshev, E. & Meselson, M. Extreme resistance of bdelloid rotifers to ionizing radiation. *Proc Natl Acad Sci U S A* **105**, 5139-44(2008).
109. Gladyshev, E.A., Meselson, M. & Arkhipova, I.R. Massive Horizontal Gene Transfer in Bdelloid Rotifers. *Science* **320**, 1210-1213(2008).
110. Gojobori, T., Li, W. & Graur, D. Patterns of nucleotide substitution in pseudogenes and functional genes. *Journal of Molecular Evolution* **18**, 360-369(1982).
111. Gordo, I. & Campos, P. Adaptive evolution in a spatially structured asexual

- population. *Genetica* **127**, 217-229(2006).
112. Graur, D. & Li, W. *Fundamentals of Molecular Evolution*. (Sinauer: Sunderland, MA, 2000).
 113. Green, P. Against a whole-genome shotgun. *Genome Res* **7**, 410-7(1997).
 114. Gregory, T.R. *The Evolution of the Genome*. (Elsevier Academic Press: Boston, MA, 2005).
 115. Gu, Z. et al. Role of duplicate genes in genetic robustness against null mutations. *Nature* **421**, 63-6(2003).
 116. Haag-Liautard, C. et al. Direct estimation of the mitochondrial DNA mutation rate in *Drosophila melanogaster*. *PLoS Biol* **6**, e204(2008).
 117. Haddrill, P.R. et al. Reduced efficacy of selection in regions of the *Drosophila* genome that lack crossing over. *Genome Biol.* **8**, R18(2007).
 118. Haddrill, P.R., Waldron, F.M. & Charlesworth, B. Elevated levels of expression associated with regions of the *Drosophila* genome that lack crossing over. *Biol Lett* (2008).doi:10.1098/rsbl.2008.0376
 119. Hahn, M.W. Toward a selection theory of molecular evolution. *Evolution* **62**, 255-65(2008).
 120. Hahn, M.W., Han, M.V. & Han, S. Gene family evolution across 12 *Drosophila* genomes. *PLoS Genet* **3**, e197(2007).
 121. Hairston, N.G. et al. Natural selection for grazer resistance to toxic cyanobacteria: evolution of phenotypic plasticity? *Evolution* **55**, 2203-14(2001).
 122. Harris, H. Enzyme polymorphisms in man. *Proc R Soc Lond B Biol Sci* **164**, 298-310(1966).
 123. Hassanin, A., Léger, N. & Deutsch, J. Evidence for multiple reversals of asymmetric mutational constraints during the evolution of the mitochondrial genome of metazoa, and consequences for phylogenetic inferences. *Syst Biol* **54**, 277-98(2005).
 124. Hawks, J. et al. Recent acceleration of human adaptive evolution. *Proc Natl Acad Sci U S A.* **104**, 20753–20758(2007).
 125. Hebert, P.D.N. et al. Accelerated molecular evolution in halophilic crustaceans. *Evolution* **56**, 909-26(2002).
 126. Hedges, S.B. et al. A genomic timescale for the origin of eukaryotes. *BMC Evol Biol* **1**, 4(2001).
 127. Hedrick, P.W. Balancing selection. *Curr Biol* **17**, R230-1(2007).
 128. Hellmann, I. et al. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am J Hum Genet* **72**, 1527-35(2003).
 129. Hellmann, I. et al. Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. *Genome Res* **18**, 1020-9(2008).
 130. Hill, W.G. & Robertson, A. The effect of linkage on limits to artificial selection. *Genet Res* **8**, 269-94(1966).
 131. Hill, W.G. & Robertson, A. Linkage disequilibrium in finite populations. *TAG Theoretical and Applied Genetics* **38**, 226-231(1968).

132. Hillier, L.W. et al. Whole-genome sequencing and variant discovery in *C. elegans*. *Nat Meth* **5**, 183-188(2008).
133. Hindorff, L. et al. A Catalog of Published Genome-Wide Association Studies. www.genome.gov/26525384 (2008).
134. Hinds, D.A. et al. Whole-Genome Patterns of Common DNA Variation in Three Human Populations. *Science* **307**, 1072-1079(2005).
135. Hiroaki Kakinuma & Hitoshi Sato Copy-number variations associated with autism spectrum disorder. (2008).at
<<http://www.futuremedicine.com/doi/abs/10.2217/14622416.9.8.1143>>
136. Ho, S.Y.W. & Larson, G. Molecular clocks: when times are a-changin'. *Trends Genet* **22**, 79-83(2006).
137. Hoekstra, H.E. et al. A single amino acid mutation contributes to adaptive beach mouse color pattern. *Science* **313**, 101-4(2006).
138. Holt, R.A. et al. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298**, 129-49(2002).
139. Hsiao, T. & Vitkup, D. Role of duplicate genes in robustness against deleterious human mutations. *PLoS Genet* **4**, e1000014(2008).
140. Hudson, R.R. & Kaplan, N.L. Deleterious background selection with recombination. *Genetics* **141**, 1605-17(1995).
141. Hudson, R.R., Kreitman, M. & Aguade, M. A Test of Neutral Molecular Evolution Based on Nucleotide Data. *Genetics* **116**, 153-159(1987).
142. Hughes, A.L. Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity* **99**, 364-73(2007).
143. Hughes, A.L. Near neutrality: leading edge of the neutral theory of molecular evolution. *Ann N Y Acad Sci* **1133**, 162-79(2008).
144. Hurles, M. Gene duplication: the genomic trade in spare parts. *PLoS Biol* **2**, E206(2004).
145. Hurst, L.D. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet* **18**, 486(2002).
146. Ionita-Laza, I. et al. On the frequency of copy number variants. *Bioinformatics* **24**, 2350-2355(2008).
147. Irish, V.F. & Litt, A. Flower development and evolution: gene duplication, diversification and redeployment. *Curr Opin Genet Dev* **15**, 454-60(2005).
148. J C Avise et al. Intraspecific Phylogeography: The Mitochondrial DNA Bridge Between Population Genetics and Systematics. *Annual Review of Ecology & Systematics* **18**, 489-522(1987).
149. Jakobsson, M. et al. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**, 998-1003(2008).
150. Jeffares, D.C., Mourier, T. & Penny, D. The biology of intron gain and loss. *Trends Genet* **22**, 16-22(2006).
151. Jensen, J.D., Wong, A. & Aquadro, C.F. Approaches for identifying targets of positive selection. *Trends Genet* **23**, 568-77(2007).
152. Jiang, C. & Zhao, Z. Directionality of point mutation and 5-methylcytosine

- deamination rates in the chimpanzee genome. *BMC Genomics* **7**, 316(2006).
153. Jiang, C. & Zhao, Z. Mutational spectrum in the recent human genome inferred by single nucleotide polymorphisms. *Genomics* **88**, 527-34(2006).
 154. Jorde, L.B., Watkins, W.S. & Bamshad, M.J. Population genomics: a bridge from evolutionary history to genetic medicine. *Hum Mol Genet* **10**, 2199-207(2001).
 155. Juárez, P. et al. Evolution of snake venom disintegrins by positive darwinian selection. *Mol Biol Evol* **25**, 2391-407(2008).
 156. Kaplan, N.L., Hudson, R.R. & Langley, C.H. The "hitchhiking effect" revisited. *Genetics* **123**, 887-99(1989).
 157. Kasahara, M. et al. The medaka draft genome and insights into vertebrate genome evolution. *Nature* **447**, 714-719(2007).
 158. Katju, V. & Lynch, M. The structure and early evolution of recently arisen gene duplicates in the *Caenorhabditis elegans* genome. *Genetics* **165**, 1793-803(2003).
 159. Katju, V. & Lynch, M. On the formation of novel genes by duplication in the *Caenorhabditis elegans* genome. *Mol Biol Evol* **23**, 1056-67(2006).
 160. Kehrer-Sawatzki, H. & Cooper, D.N. Understanding the recent evolution of the human genome: insights from human-chimpanzee genome comparisons. *Human Mutation* **28**, 99-130(2007).
 161. Keller, I., Bensasson, D. & Nichols, R.A. Transition-transversion bias is not universal: a counter example from grasshopper pseudogenes. *PLoS Genet* **3**, e22(2007).
 162. Khaja, R. et al. Genome assembly comparison identifies structural variants in the human genome. *Nat Genet* **38**, 1413-8(2006).
 163. Kidd, J.M. et al. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56-64(2008).
 164. Kim, J.H., Waterman, M.S. & Li, L.M. Diploid genome reconstruction of *Ciona intestinalis* and comparative analysis with *Ciona savignyi*. *Genome Res* **17**, 1101-10(2007).
 165. Kimura, M. Solution of a process of random genetic drift with a continuous model. *Proc Natl Acad Sci U S A* **41**, 144-50(1955).
 166. Kimura, M. Diffusion Models in Population Genetics. *Journal of Applied Probability* **1**, 177-232(1964).
 167. Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624-6(1968).
 168. Kimura, M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267**, 275-6(1977).
 169. Kimura, M. *The Neutral Theory of Molecular Evolution*. (Cambridge University Press: Cambridge, 1983).
 170. Kimura, M. & Ohta, T. The Average Number of Generations until Fixation of a Mutant Gene in a Finite Population. *Genetics* **61**, 763-771(1969).
 171. Kimura, M. & Ohta, T. Protein polymorphism as a phase of molecular evolution. *Nature* **229**, 467-9(1971).

172. King, J.L. & Jukes, T.H. Non-Darwinian evolution. *Science* **164**, 788-98(1969).
173. Kocher, T.D. et al. Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. *Proc Natl Acad Sci U S A* **86**, 6196-200(1989).
174. Kondrashov, F.A. et al. Selection in the evolution of gene duplications. *Genome Biol* **3**, RESEARCH0008(2002).
175. Koonin, E.V. The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? *Biol Direct* **1**, 22(2006).
176. Koonin, E.V. & Novozhilov, A.S. Origin and evolution of the genetic code: The universal enigma. *IUBMB Life* (2008).doi:10.1002/iub.146
177. Kreitman, M. & Akashi, H. Molecular evidence for natural selection. *Annual Review of Ecology & Systematics* **26**, 403-422(1995).
178. Kulathinal, R.J. et al. Fine-scale mapping of recombination rate in *Drosophila* refines its correlation to diversity and divergence. *Proc Natl Acad Sci U S A* **105**, 10051-6(2008).
179. Kumar, S. & Hedges, S.B. A molecular timescale for vertebrate evolution. *Nature* **392**, 917-20(1998).
180. Larkin, M.A. et al. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-8(2007).
181. Lawniczak, M.K.N. et al. Genomic analysis of the relationship between gene expression variation and DNA polymorphism in *Drosophila simulans*. *Genome Biol* **9**, R125(2008).
182. Lerner, I. *Genetic Homeostasis*. (Oliver and Boyd: Edinburgh, 1954).
183. Levy, S. et al. The diploid genome sequence of an individual human. *PLoS Biol* **5**, e254(2007).
184. Lewontin, R.C. & Hubby, J.L. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics* **54**, 595-609(1966).
185. Li, W., Yang, J. & Gu, X. Expression divergence between duplicate genes. *Trends Genet* **21**, 602-7(2005).
186. Li, Y.F. et al. "Reverse ecology" and the power of population genomics. *Evolution* **62**, 2984-94(2008).
187. Liao, D. Concerted evolution: molecular mechanism and biological implications. *Am J Hum Genet* **64**, 24-30(1999).
188. Lindblad-Toh, K. et al. Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nat Genet* **24**, 381-386(2000).
189. Lindblad-Toh, K. et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803-819(2005).
190. Liu, J. et al. Natural selection of protein structural and functional properties: a single nucleotide polymorphism perspective. *Genome Biol* **9**, R69(2008).
191. Llopart, A. et al. Intron presence-absence polymorphism in *Drosophila*

- driven by positive Darwinian selection. *Proc Natl Acad Sci U S A* **99**, 8121-6(2002).
192. Loh, Y.E. et al. Comparative analysis reveals signatures of differentiation amid genomic polymorphism in Lake Malawi cichlids. *Genome Biol* **9**, R113(2008).
 193. Lohmueller, K.E. et al. Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451**, 994-7(2008).
 194. Long, A.D., Beldade, P. & Macdonald, S.J. Estimation of population heterozygosity and library construction-induced mutation rate from expressed sequence tag collections. *Genetics* **176**, 711-4(2007).
 195. Lynch, M. Ecological Genetics of *Daphnia pulex*. *Evolution* **37**, 358-374(1983).
 196. Lynch, M. The origins of eukaryotic gene structure. *Mol Biol Evol* **23**, 450-68(2006).
 197. Lynch, M. *The Origins of Genome Architecture*. (Sinauer: Sunderland, MA, 2007).
 198. Lynch, V.J. Inventing an arsenal: adaptive evolution and neofunctionalization of snake venom phospholipase A2 genes. *BMC Evol Biol.* **7**, 2(2007).
 199. Lynch, M. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc Natl Acad Sci U S A* **104 Suppl 1**, 8597-604(2007).
 200. Lynch, M. Estimation of Nucleotide Diversity, Disequilibrium Coefficients, and Mutation Rates from High-Coverage Genome-Sequencing Projects. *Mol Biol Evol* **25**, 2409-2419(2008).
 201. Lynch, M. & Conery, J.S. The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151-5(2000).
 202. Lynch, M. & Conery, J.S. The evolutionary demography of duplicate genes. *J Struct Funct Genomics* **3**, 35-44(2003).
 203. Lynch, M. & Force, A. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**, 459-73(2000).
 204. Lynch, M. & Katju, V. The altered evolutionary trajectories of gene duplicates. *Trends Genet* **20**, 544-9(2004).
 205. Lynch, M. & Richardson, A.O. The evolution of spliceosomal introns. *Curr Opin Genet Dev* **12**, 701-10(2002).
 206. Lynch, M. & Ritland, K. Estimation of pairwise relatedness with molecular markers. *Genetics* **152**, 1753-66(1999).
 207. Lynch, M. & Walsh, J.B. *Genetics and Analysis of Quantitative Traits*. (Sinauer: Sunderland, MA, 1998).
 208. Lynch, M. et al. The probability of preservation of a newly arisen gene duplicate. *Genetics* **159**, 1789-804(2001).
 209. Lynch, M. et al. Localization of the genetic determinants of meiosis suppression in *Daphnia pulex*. *Genetics* **180**, 317-27(2008).
 210. Maere, S. et al. Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci U S A* **102**, 5454-9(2005).

211. Malaria Genomic Epidemiology Network A global network for investigating the genomic epidemiology of malaria. *Nature* **456**, 732-737(2008).
212. Malcom, C.M., Wyckoff, G.J. & Lahn, B.T. Genic mutation rates in mammals: local similarity, chromosomal heterogeneity, and X-versus-autosome disparity. *Mol Biol Evol* **20**, 1633-41(2003).
213. Manolio, T.A., Brooks, L.D. & Collins, F.S. A HapMap harvest of insights into the genetics of common disease. *J Clin Invest* **118**, 1590-605(2008).
214. Mardis, E.R. The impact of next-generation sequencing technology on genetics. *Trends Genet* **24**, 133-41(2008).
215. Mark Welch, D.B., Mark Welch, J.L. & Meselson, M. Evidence for degenerate tetraploidy in bdelloid rotifers. *Proc Natl Acad Sci U S A* **105**, 5145-9(2008).
216. Massey, S.E. A neutral origin for error minimization in the genetic code. *J Mol Evol* **67**, 510-6(2008).
217. Mayr, E. *Animal Species and Evolution*. (Harvard University Press: 1963).
218. McCarroll, S.A. Extending genome-wide association studies to copy-number variation. *Hum. Mol. Genet.* **17**, R135-142(2008).
219. McCarroll, S.A. & Altshuler, D.M. Copy-number variation and association studies of human disease. *Nat Genet* (2007).
220. McDonald, J.H. & Kreitman, M. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* **351**, 652-4(1991).
221. McKay, J.K. & Stinchcombe, J.R. Ecological genomics of model eukaryotes. *Evolution* **62**, 2953-7(2008).
222. Mitchell-Olds, T., Willis, J.H. & Goldstein, D.B. Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nat Rev Genet* **8**, 845-56(2007).
223. Moore, R.C. & Purugganan, M.D. The early stages of duplicate gene evolution. *Proc Natl Acad Sci U S A* **100**, 15682-7(2003).
224. Moore, R.C. & Purugganan, M.D. The evolutionary dynamics of plant duplicate genes. *Curr Opin Plant Biol* **8**, 122-8(2005).
225. Moriyama, E.N. & Powell, J.R. Intraspecific nuclear DNA variation in *Drosophila*. *Mol Biol Evol* **13**, 261-77(1996).
226. Nardi, F. et al. The mitochondrial genome of the olive fly *Bactrocera oleae*: two haplotypes from distant geographical locations. *Insect Molecular Biology* **12**, 605-611(2003).
227. Nei, M. *Molecular Evolutionary Genetics*. (Columbia University Press: New York, 1987).
228. Nei, M. Selectionism and neutralism in molecular evolution. *Mol Biol Evol* **22**, 2318-42(2005).
229. Nei, M. & Kumar, S. *Molecular Evolution and Phylogenetics*. (Oxford University Press: New York, 2000).
230. Nei, M. & Rooney, A.P. Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet* **39**, 121-52(2005).
231. Neno, M. et al. Higher frequency of concerted evolutionary events in

- rodents than in man at the polyubiquitin gene VNTR locus. *Genetics* **148**, 867-76(1998).
232. Newman, T.L. et al. A genome-wide survey of structural variation between human and chimpanzee. *Genome Res.* **15**, 1344–1356(2005).
233. Nguyen, D. et al. Reduced purifying selection prevails over positive selection in human copy number variant evolution. *Genome Res* **18**, 1711-23(2008).
234. Nilsen, H. et al. Nuclear and mitochondrial uracil-DNA glycosylases are generated by alternative splicing and transcription from different positions in the UNG gene. *Nucleic Acids Res* **25**, 750-5(1997).
235. Niu, D., Lin, K. & Zhang, D. Strand Compositional Asymmetries of Nuclear DNA in Eukaryotes. *Journal of Molecular Evolution* **57**, 325-334(2003).
236. Noor, M. Connecting recombination, nucleotide diversity and species divergence in *Drosophila*. *Fly (Austin)* **2**, (2008).
237. Novembre, J. et al. Genes mirror geography within Europe. *Nature* **456**, 98-101(2008).
238. Nozawa, M., Kawahara, Y. & Nei, M. Genomic drift and copy number variation of sensory receptor genes in humans. *Proceedings of the National Academy of Sciences* **104**, 20421-20426(2007).
239. Ohno, S. *Evolution by Gene Duplication*. (Springer-Verlag: Berlin, 1970).
240. Ohta, T. The Nearly Neutral Theory of Molecular Evolution. *Annual Review of Ecology and Systematics* **23**, 263-286(1992).
241. Omilian, A.R. et al. Ameiotic recombination in asexual lineages of *Daphnia*. *Proc Natl Acad Sci U S A.* **103**, 18638–18643(2006).
242. Omilian, A.R., Scofield, D.G. & Lynch, M. Intron Presence-Absence Polymorphisms in *Daphnia*. *Mol Biol Evol* **25**, 2129-2139(2008).
243. Orr, H.A. Testing Natural Selection. *Scientific American* **300**, 44-51(2009).
244. Paland, S. & Lynch, M. Transitions to asexuality result in excess amino acid substitutions. *Science* **311**, 990-2(2006).
245. Paland, S., Colbourne, J.K. & Lynch, M. Evolutionary history of contagious asexuality in *Daphnia pulex*. *Evolution* **59**, 800-13(2005).
246. Palti, Y. et al. Detection of genes with deleterious alleles in an inbred line of tilapia (*Oreochromis aureus*). *Aquaculture* **206**, 151-164(2002).
247. Pan, D. & Zhang, L. Quantifying the major mechanisms of recent gene duplications in the human and mouse genomes: a novel strategy to estimate gene duplication rates. *Genome Biol* **8**, R158(2007).
248. Pan, D. & Zhang, L. Tandemly arrayed genes in vertebrate genomes. *Comp Funct Genomics* 545269(2008).doi:10.1155/2008/545269
249. Parmley, J.L., Chamary, J.V. & Hurst, L.D. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol Biol Evol* **23**, 301-9(2006).
250. Pavelitz, T. et al. Concerted evolution of the tandem array encoding primate U2 snRNA occurs in situ, without changing the cytological context of the RNU2 locus. *EMBO J* **14**, 169-77(1995).

251. Pavlidis, P., Hutter, S. & Stephan, W. A population genomic approach to map recent positive selection in model species. *Molecular Ecology* **17**, 3585-3598(2008).
252. Payseur, B.A. & Nachman, M.W. Gene density and human nucleotide polymorphism. *Mol Biol Evol* **19**, 336-40(2002).
253. Perna, N.T. & Kocher, T.D. Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. *J Mol Evol* **41**, 353-8(1995).
254. Phillips, N. et al. Spontaneous Mutational and Standing Genetic (Co)Variation at Dinucleotide Microsatellites in *Caenorhabditis briggsae* and *C. elegans*. *Mol Biol Evol* (2008).doi:10.1093/molbev/msn287
255. Pollard, H.G., Colbourne, J.K. & Keller, W. Reconstruction of centuries-old *Daphnia* communities in a lake recovering from acidification and metal contamination. *Ambio* **32**, 214-8(2003).
256. Ponjavic, J., Ponting, C.P. & Lunter, G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res* **17**, 556-65(2007).
257. Pop, M. et al. Comparative genome assembly. *Brief Bioinform* **5**, 237-48(2004).
258. Przeworski, M., Hudson, R.R. & Di Rienzo, A. Adjusting the focus on human variation. *Trends in Genetics* **16**, 296-302(2000).
259. Ptak, S.E. et al. Fine-scale recombination patterns differ between chimpanzees and humans. *Nat Genet* **37**, 429-34(2005).
260. Rastogi, S. & Liberles, D.A. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol Biol* **5**, 28(2005).
261. Redon, R. et al. Global variation in copy number in the human genome. *Nature* **444**, 444-454(2006).
262. Reis, M.D. & Wernisch, L. Estimating translational selection in Eukaryotic genomes. *Mol Biol Evol* msn272(2009).doi:10.1093/molbev/msn272
263. Resch, A.M. et al. Widespread Positive Selection in Synonymous Sites of Mammalian Genes. *Mol Biol Evol* **24**, 1821-1831(2007).
264. Reyes, A. et al. Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Mol Biol Evol* **15**, 957-66(1998).
265. Reyes, A. et al. Bidirectional Replication Initiates at Sites Throughout the Mitochondrial Genome of Birds. *J. Biol. Chem.* **280**, 3242-3250(2005).
266. Rocha, E.P.C., Touchon, M. & Feil, E.J. Similar compositional biases are caused by very different mutational effects. *Genome Res* **16**, 1537-47(2006).
267. Rosenberg, M.S., Subramanian, S. & Kumar, S. Patterns of Transitional Mutation Biases Within and Among Mammalian Genomes. *Mol Biol Evol* **20**, 988-993(2003).
268. Roy, S.W. & Gilbert, W. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat Rev Genet* **7**, 211-21(2006).
269. Ruppert, E.E., Barnes, R.D. & Fox, R.S. *Invertebrate Zoology: A Functional Evolutionary Approach*. (Brooks Cole: 2003).

270. Sabeti, P.C. et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832-7(2002).
271. Sabeti, P.C. et al. Positive natural selection in the human lineage. *Science* **312**, 1614-20(2006).
272. Sabeti, P.C. et al. Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913-8(2007).
273. Saccone, C. et al. Mitochondrial DNA in metazoa: degree of freedom in a frozen event. *Gene* **286**, 3-12(2002).
274. Sadusky, T., Newman, A.J. & Dibb, N.J. Exon junction sequences as cryptic splice sites: implications for intron origin. *Curr Biol* **14**, 505-9(2004).
275. Sauvage, C. et al. Single Nucleotide polymorphisms and their relationship to codon usage bias in the Pacific oyster *Crassostrea gigas*. *Gene* **406**, 13-22(2007).
276. Sawyer, S.A. et al. Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. *J Mol Evol* **57 Suppl 1**, S154-64(2003).
277. Sea Urchin Genome Sequencing Consortium et al. The Genome of the Sea Urchin *Strongylocentrotus purpuratus*. *Science* **314**, 941-952(2006).
278. Sella, G. & Ardell, D.H. The coevolution of genes and genetic codes: Crick's frozen accident revisited. *J Mol Evol* **63**, 297-313(2006).
279. Severson, D.W. et al. Linkage map organization of expressed sequence tags and sequence tagged sites in the mosquito, *Aedes aegypti*. *Insect Mol Biol* **11**, 371-8(2002).
280. Shadel, G.S. & Clayton, D.A. Mitochondrial DNA maintenance in vertebrates. *Annu Rev Biochem* **66**, 409-35(1997).
281. Shapiro, J.A. et al. Adaptive genic evolution in the *Drosophila* genomes. *Proc Natl Acad Sci U S A* **104**, 2271-6(2007).
282. Shianna, K.V. & Willard, H.F. Human genomics: In search of normality. *Nature* **444**, 428-429(2006).
283. Small, K.S. et al. A haplome alignment and reference sequence of the highly polymorphic *Ciona savignyi* genome. *Genome Biol* **8**, R41(2007).
284. Small, K.S. et al. Extreme genomic variation in a natural population. *Proc Natl Acad Sci U S A* **104**, 5698-703(2007).
285. Smith, N.G.C. & Eyre-Walker, A. Adaptive protein evolution in *Drosophila*. *Nature* **415**, 1022-4(2002).
286. Smith, J.M. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genet Res* **23**, 23-35(1974).
287. Sokal, R. & Rohlf, F. *Biometry*. (W.H. Freeman and Company: San Francisco, 1981).
288. Solorzano, E. et al. Patterns of Variation in the Nuclear Genome of *Caenorhabditis elegans*. *in prep*
289. Spencer, C.C.A. et al. The influence of recombination on human genetic diversity. *PLoS Genet* **2**, e148(2006).
290. Stein, L.D. et al. The genome sequence of *Caenorhabditis briggsae*: a

- platform for comparative genomics. *PLoS Biol* **1**, E45(2003).
291. Stinchcombe, J.R. & Hoekstra, H.E. Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity* **100**, 158-70(2008).
 292. Stranger, B.E. et al. Population genomics of human gene expression. *Nat Genet* **39**, 1217-24(2007).
 293. Sueoka, N. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *J Mol Evol* **40**, 318-25(1995).
 294. Sung, W. et al. Simple sequence repeats in the *Daphnia pulex* genome. *in prep*
 295. Svensson, O., Arvestad, L. & Lagergren, J. Genome-wide survey for biologically functional pseudogenes. *PLoS Comput Biol* **2**, e46(2006).
 296. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585-95(1989).
 297. Tamura, K. & Nei, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* **10**, 512-26(1993).
 298. Tamura, K. et al. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**, 1596-9(2007).
 299. Tanaka, M. & Ozawa, T. Strand asymmetry in human mitochondrial DNA mutations. *Genomics* **22**, 327-35(1994).
 300. Tang, K., Thornton, K.R. & Stoneking, M. A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome. *PLoS Biol* **5**, e171(2007).
 301. Tenesa, A. et al. Recent human effective population size estimated from linkage disequilibrium. *Genome Res* **17**, 520-6(2007).
 302. Teshima, K.M. & Innan, H. The effect of gene conversion on the divergence between duplicated genes. *Genetics* **166**, 1553-60(2004).
 303. The French-Italian Public Consortium for Grapevine Characterization The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463-467(2007).
 304. The Honeybee Genome Sequencing Consortium Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* **443**, 931-949(2006).
 305. Thomas, W.K. & Kocher, T.D. Sequencing of polymerase chain reaction-amplified DNAs. *Methods Enzymol* **224**, 391-9(1993).
 306. Thomas, W.K. & Wilson, A.C. Mode and tempo of molecular evolution in the nematode *Caenorhabditis*: cytochrome oxidase II and calmodulin sequences. *Genetics* **128**, 269-79(1991).
 307. Thomas, W.K., Maa, J. & Wilson, A.C. Shifting constraints on tRNA genes during mitochondrial DNA evolution in animals. *New Biol* **1**, 93-100(1989).
 308. Thornton, K.R. et al. Progress and prospects in mapping recent selection in the genome. *Heredity* **98**, 340-348(2007).
 309. Tishkoff, S.A. & Kidd, K.K. Implications of biogeography of human

- populations for 'race' and medicine. *Nat Genet* **36**, S21-7(2004).
310. Tishkoff, S.A. & Williams, S.M. Genetic analysis of African populations: human evolution and complex disease. *Nat Rev Genet* **3**, 611-621(2002).
 311. Tishkoff, S.A. et al. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet* **39**, 31-40(2007).
 312. Tribolium Genome Sequencing Consortium The genome of the model beetle and pest *Tribolium castaneum*. *Nature* **452**, 949-955(2008).
 313. Tsai, I.J. et al. Population genomics of the wild yeast *Saccharomyces paradoxus*: Quantifying the life cycle. *Proc Natl Acad Sci U S A* **105**, 4957-62(2008).
 314. Tuskan, G.A. et al. The Genome of Black Cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**, 1596-1604(2006).
 315. Tuzun, E. et al. Fine-scale structural variation of the human genome. *Nat Genet* **37**, 727-32(2005).
 316. Ungerer, M.C., Johnson, L.C. & Herman, M.A. Ecological genomics: understanding gene and genome function in the natural environment. *Heredity* **100**, 178-83(2008).
 317. Vandepoele, K. et al. Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc Natl Acad Sci U S A* **101**, 1638-43(2004).
 318. Vavouri, T. et al. Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. *Genome Biol* **8**, R15(2007).
 319. Vinson, J.P. et al. Assembly of polymorphic genomes: algorithms and application to *Ciona savignyi*. *Genome Res* **15**, 1127-35(2005).
 320. Voight, B.F. et al. A map of recent positive selection in the human genome. *PLoS Biol* **4**, e72(2006).
 321. Wagner, A. Neutralism and selectionism: a network-based reconciliation. *Nat Rev Genet* **9**, 965-74(2008).
 322. Wakeley, J. The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. *Trends in Ecology & Evolution* **11**, 158-162(1996).
 323. Wang, E.T. et al. Global landscape of recent inferred Darwinian selection for *Homo sapiens*. *Proc Natl Acad Sci U S A* **103**, 135-40(2006).
 324. Wang, J. et al. The diploid genome sequence of an Asian individual. *Nature* **456**, 60-5(2008).
 325. Ward, R. & Durrett, R. Subfunctionalization: How often does it occur? How long does it take? *Theor Popul Biol* **66**, 93-100(2004).
 326. Wayne, R.K. & Ostrander, E.A. Lessons learned from the dog genome. *Trends Genet* **23**, 557-67(2007).
 327. Weider, L.J. et al. Long-term genetic shifts in a microcrustacean egg bank associated with anthropogenic changes in the Lake Constance ecosystem. *Proc Biol Sci*. **264**, 1613-1618(1997).
 328. Wheeler, D.A. et al. The complete genome of an individual by massively

- parallel DNA sequencing. *Nature* **452**, 872-6(2008).
329. Wicks, S.R. et al. Rapid gene mapping in *Caenorhabditis elegans* using a high density polymorphism map. *Nat Genet* **28**, 160-4(2001).
330. Wildman, D.E. et al. Implications of natural selection in shaping 99.4% nonsynonymous DNA identity between humans and chimpanzees: enlarging genus *Homo*. *Proc Natl Acad Sci U S A* **100**, 7181-8(2003).
331. Wilhelm, L. et al. Natural variation in SAR11 marine bacterioplankton genomes inferred from metagenomic data. *Biology Direct* **2**, 27(2007).
332. Williamson, S.H. et al. Localizing recent adaptive evolution in the human genome. *PLoS Genet* **3**, e90(2007).
333. Wilson, A.C. & Sarich, V.M. A molecular timescale for human evolution. *Proceedings of the National Academy of Sciences of the United States of America* **63**, 1088-1093(1969).
334. Wilson, A.C., Ochman, H. & Prager, E.M. Molecular time scale for evolution. *Trends in Genetics* **3**, 241-247
335. Winckler, W. et al. Comparison of fine-scale recombination rates in humans and chimpanzees. *Science* **308**, 107-11(2005).
336. Wolfe, K.H., Sharp, P.M. & Li, W.H. Mutation rates differ among regions of the mammalian genome. *Nature* **337**, 283-5(1989).
337. Wright, S. Evolution in Mendelian populations. 1931. *Bull Math Biol* **52**, 241-95; discussion 201-7(1990).
338. Wright, S.I. & Gaut, B.S. Molecular population genetics and the search for adaptive evolution in plants. *Mol Biol Evol* **22**, 506-19(2005).
339. Xu, S. et al. Gene conversion in the rice genome. *BMC Genomics* **9**, 93(2008).
340. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol* **24**, 1586-1591(2007).
341. Yang & Bielawski Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* **15**, 496-503(2000).
342. Yang, Z. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* **17**, 32-43(2000).
343. Yang, M.Y. et al. Biased incorporation of ribonucleotides on the mitochondrial L-strand accounts for apparent strand-asymmetric DNA replication. *Cell* **111**, 495-505(2002).
344. Yasukochi, Y. A Dense Genetic Map of the Silkworm, *Bombyx mori*, Covering All Chromosomes Based on 1018 Molecular Markers. *Genetics* **150**, 1513-1525(1998).
345. Yi, S.V. Non-adaptive evolution of genome complexity. *Bioessays* **28**, 979-82(2006).
346. Young, J.M. et al. Extensive copy-number variation of the human olfactory receptor gene family. *Am J Hum Genet* **83**, 228-42(2008).
347. Zhang, J. The drifting human genome. *Proc Natl Acad Sci U S A* **104**, 20147-8(2007).

348. Zhang, Z. & Kishino, H. Genomic background predicts the fate of duplicated genes: evidence from the yeast genome. *Genetics* **166**, 1995-9(2004).
349. Zhang, P., Min, W. & Li, W. Different age distribution patterns of human, nematode, and Arabidopsis duplicate genes. *Gene* **342**, 263-8(2004).
350. Zhang, L. et al. Patterns of segmental duplication in the human genome. *Mol Biol Evol* **22**, 135-41(2005).
351. Zhang, Z. et al. KaKs_Calculator: Calculating Ka and Ks Through Model Selection and Model Averaging. *Genomics, Proteomics & Bioinformatics* **4**, 259-263(2006).
352. Zhang, Z. et al. PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics* **22**, 1437-9(2006).
353. Zhang, C. et al. A whole genome long-range haplotype (WGLRH) test for detecting imprints of positive selection in human populations. *Bioinformatics* **22**, 2122-8(2006).
354. Zhao, H. et al. The study of neighboring nucleotide composition and transition/transversion bias. *Science in China Series C: Life Sciences* **49**, 395-402(2006).
355. Zuckerkandl, E. & Pauling, L. Evolutionary divergence and convergence in proteins. *Evolving Genes and Proteins* 97-166(1965).

APPENDIX A: Ongoing Arthropod Genome Projects as of December 2008 (source:
www.genomesonline.org)

Class	Order	Species
Arachnida	Ixodida	<i>Ixodes scapularis</i>
Arachnida		<i>Tetranychus urticae</i>
Branchiopoda	Diplostraca	<i>Daphnia magna</i>
Branchiopoda	Diplostraca	<i>Daphnia pulex</i>
Chilopoda	Geophilomorpha	<i>Strigamia maritima</i>
Insecta	Diptera	<i>Aedes albopictus</i>
Insecta	Diptera	<i>Aedes triseriatus</i>
Insecta	Diptera	<i>Anopheles gambiae*</i>
Insecta	Diptera	<i>Cochliomyia hominivorax</i>
Insecta	Diptera	<i>Culex pipiens</i>
Insecta	Diptera	<i>Culex pipiens quinquefasciatus</i>
Insecta	Diptera	<i>Drosophila americana</i>
Insecta	Diptera	<i>Drosophila auraria</i>
Insecta	Diptera	<i>Drosophila equinoxialis</i>
Insecta	Diptera	<i>Drosophila hydei</i>
Insecta	Diptera	<i>Drosophila littoralis</i>
Insecta	Diptera	<i>Drosophila mauritiana</i>
Insecta	Diptera	<i>Drosophila mercatorum</i>
Insecta	Diptera	<i>Drosophila mimica</i>
Insecta	Diptera	<i>Drosophila miranda</i>
Insecta	Diptera	<i>Drosophila mojavensis</i>
Insecta	Diptera	<i>Drosophila novamexicana</i>
Insecta	Diptera	<i>Drosophila repleta</i>
Insecta	Diptera	<i>Drosophila silvestris</i>
Insecta	Diptera	<i>Drosophila simulans*</i>
Insecta	Diptera	<i>Glossina morsitans</i>
Insecta	Diptera	<i>Haematobia irritans</i>
Insecta	Diptera	<i>Lutzomyia longipalpis</i>
Insecta	Diptera	<i>Phlebotomus papatasi</i>
Insecta	Hemiptera	<i>Acyrtosiphon pisum</i>
Insecta	Hemiptera	<i>Acyrtosiphon pisum</i>
Insecta	Hemiptera	<i>Diaphorina citri</i>
Insecta	Hemiptera	<i>Rhodnius prolixus</i>
Insecta	Hymenoptera	<i>Apis mellifera capensis</i>

Insecta	Hymenoptera	<i>Apis mellifera scutellata</i>
Insecta	Hymenoptera	<i>Nasonia giraulti</i>
Insecta	Hymenoptera	<i>Nasonia longicornis</i>
Insecta	Hymenoptera	<i>Nasonia vitripennis</i>
Insecta	Lepidoptera	<i>Bicyclus anynana</i>
Insecta	Lepidoptera	<i>Helicoverpa armigera</i>
Insecta	Lepidoptera	<i>Heliopsis virescens</i>
Insecta	Lepidoptera	<i>Spodoptera frugiperda</i>
Insecta	Phthiraptera	<i>Pediculus humanus corporis</i>
Malacostraca	Amphipoda	<i>Jassa slatteryi</i>
Merostomata	Xiphosura	<i>Limulus polyphemus</i>

APPENDIX B: Fully sequenced eukaryotic genomes (as of December 2008)

Phylum	Species	Sequencing	Coverage
Apicomplexa	<i>Babesia bovis</i>	Sanger	11x
Apicomplexa	<i>Theileria annulata</i>	Sanger	10x
Apicomplexa	<i>Theileria parva</i>		
Apicomplexa	<i>Cryptosporidium hominis</i>	Sanger	
Apicomplexa	<i>Cryptosporidium parvum</i>		
Apicomplexa	<i>Plasmodium yoelii yoelii</i>	Sanger	5x
Apicomplexa	<i>Plasmodium falciparum</i>	Sanger	
Arthropoda	<i>Drosophila pseudoobscura</i>	Sanger	9.1x
Arthropoda	<i>Bombyx mori</i>	Sanger	5.9x
Arthropoda	<i>Aedes aegypti</i>	Sanger	8x
Arthropoda	<i>Apis mellifera</i>	Sanger	7-8x
Arthropoda	<i>Bombyx mori</i>	Sanger	3x
Arthropoda	<i>Anopheles gambiae</i>	Sanger	10x
Arthropoda	<i>Drosophila melanogaster</i>	Sanger	
Arthropoda	<i>Drosophila ananassae</i>	Sanger	8.9x
Arthropoda	<i>Drosophila erecta</i>	Sanger	10.6x
Arthropoda	<i>Drosophila grimshawi</i>	Sanger	7.9x
Arthropoda	<i>Drosophila persimilis</i>	Sanger	4.1x
Arthropoda	<i>Drosophila sechellia</i>	Sanger	4.9x
Arthropoda	<i>Drosophila virilis</i>	Sanger	8x
Arthropoda	<i>Drosophila willistoni</i>	Sanger	8.4x
Arthropoda	<i>Drosophila yakuba</i>	Sanger	9.1x
Arthropoda	<i>Tribolium castaneum</i>	Sanger	7.3x
Ascomycota	<i>Candida albicans</i>	Sanger	10.9x
Ascomycota	<i>Kluyveromyces waltii</i>	Sanger	8x
Ascomycota	<i>Fusarium (Gibberella) graminearum</i> (zeae)	Sanger	10x
Ascomycota	<i>Vanderwaltozyma polyspora</i>	Sanger	7.8x
Ascomycota	<i>Magnaporthe grisea</i>	Sanger	7x
Ascomycota	<i>Pichia stipitis</i>	Sanger	
Ascomycota	<i>Aspergillus niger</i>	Sanger	7.5x
Ascomycota	<i>Aspergillus (Emericella) nidulans</i>	Sanger	13x
Ascomycota	<i>Aspergillus oryzae</i>		9x
Ascomycota	<i>Aspergillus fumigatus</i>	Sanger	10.5x
Ascomycota	<i>Candida glabrata</i>	Sanger	8x
Ascomycota	<i>Debaryomyces hansenii var. hansenii</i>	Sanger	9.7x

Ascomycota	<i>Kluyveromyces lactis</i>	Sanger	11.4x
Ascomycota	<i>Yarrowia lipolytica</i>	Sanger	10x
Ascomycota	<i>Ashbya (Eremothecium) gossypii</i>	Sanger	4.2x
Ascomycota	<i>Neurospora crassa</i>	Sanger	10x
Ascomycota	<i>Schizosaccharomyces pombe</i>		8x
Ascomycota	<i>Saccharomyces cerevisiae</i>	Sanger	
Ascomycota	<i>Podospora anserina</i>	Sanger	10x
Ascomycota	<i>Trichoderma (Hypocrea) virens</i>	Sanger	
Bacillariophyta	<i>Thalassiosira pseudonana</i>	Sanger	14x
Bacillariophyta	<i>Phaeodactylum tricornutum</i>	Sanger	9.6x
Basidiomycota	<i>Ustilago maydis</i>	Sanger	10x
Basidiomycota	<i>Cryptococcus neoformans</i>	Sanger	12.5x
Basidiomycota	<i>Phanerochaete chrysosporium</i>	Sanger	10.5x
Basidiomycota	<i>Laccaria bicolor</i>		9.9x
Basidiomycota	<i>Malassezia globosa</i>	Sanger	7x
Chlorophyta	<i>Chlamydomonas reinhardtii</i>	Sanger	10x
Chlorophyta	<i>Ostreococcus lucimarinus</i>	Sanger	
Chlorophyta	<i>Ostreococcus tauri</i>	Sanger	
Chordata	<i>Homo sapiens</i>	Illumina	
Chordata	<i>Homo sapiens</i>	Illumina	34x
Chordata	<i>Homo sapiens</i>	Sanger	7.5x
Chordata	<i>Pan troglodytes</i>	Sanger	3.6x
Chordata	<i>Monodelphis domestica</i>	Sanger	
Chordata	<i>Macaca mulatta</i>	Sanger	6x
Chordata	<i>Gallus gallus</i>	Sanger	6.6x
Chordata	<i>Canis lupus familiaris</i>	Sanger	7.6x
Chordata	<i>Danio rerio</i>	Sanger	
Chordata	<i>Tetraodon nigroviridis</i>	Sanger	6X
Chordata	<i>Rattus norvegicus</i>	Sanger	
Chordata	<i>Ciona intestinalis</i>	Sanger	8.2x
Chordata	<i>Mus musculus</i>		5x
Chordata	<i>Takifugu rubripes</i>		
Chordata	<i>Homo sapiens</i>		
Chordata	<i>Ornithorhynchus anatinus</i>	Sanger	6x
Chordata	<i>Homo sapiens</i>	454	
Cnidaria	<i>Nematostella vectensis</i>	Sanger	7.8x
Microsporidia	<i>Encephalitozoon cuniculi</i>	Sanger	
Nematoda	<i>Brugia malayi</i>	Sanger	
Nematoda	<i>Caenorhabditis briggsae</i>	Sanger	10x
Nematoda	<i>Caenorhabditis elegans</i>	Sanger	
Nematoda	<i>Meloidogyne hapla</i>	Sanger	10.4x

Placozoa	<i>Trichoplax adhaerens.</i>	Sanger	8x
Streptophyta	<i>Vitis vinifera</i>	Sanger	8.4x
Streptophyta	<i>Populus balsamifera trichocarpa</i>	Sanger	7.5x
Streptophyta	<i>Oryza sativa ssp. japonica</i>	Sanger	10x
Streptophyta	<i>Oryza sativa L. ssp. indica</i>	Sanger	6x
Streptophyta	<i>Oryza sativa ssp. japonica</i>		
Streptophyta	<i>Arabidopsis thaliana</i>		
Streptophyta	<i>Physcomitrella. patens patens</i>		
Streptophyta	<i>Vitis vinifera L.</i>	454, Sanger	11x
	<i>Hemiselmis andersenii</i>		
	<i>Giardia lamblia (intestinalis)</i>	Sanger	
	<i>Phytophthora sojae</i>	Sanger	
	<i>Phytophthora ramorum</i>	Sanger	7x
	<i>Paramecium tetraurelia</i>	Sanger	13x
	<i>Leishmania infantum</i>	Sanger	5x
	<i>Tetrahymena thermophila</i>	Sanger	
	<i>Leishmania major</i>		
	<i>Trypanosoma cruzi</i>		
	<i>Trypanosoma brucei</i>		
	<i>Dictyostelium discoideum</i>		
	<i>Entamoeba histolytica</i>	Sanger	12.5x
	<i>Cyanidioschyzon merolae</i>	Sanger	
	<i>Guillardia theta</i>		
	<i>Monosiga brevicollis</i>	Sanger	8.1x

APPENDIX C: Selected list of scripts

AMOScmpScript: A shell script that we wrote in order to change the parameters in AMOScmp from command line options instead of going into AMOS scripts to change the parameters. The output produced is congruent to the AMOScmp output. This script uses a Perl script called *changeAMOScmp* which actually does the revision to the AMOScmp script.

avgCoverage: Determines the number of sites that have a certain coverage. The script looks at every site in the output of deltaOut and determines how many nucleotide reads are at that site. If the number of nucleotides is less than the input number the list keeping track of that number of nucleotides is incremented.

coverageFilter : Determines the SNPs in the data. The coverageMin number is the minimum number of nucleotide reads that you require at a certain location while the coverageMax is the maximum number that you allow. It will also take into consideration a SNP has to have at least two nucleotides of each change in base and only two at a particular location.

kindOfSNP : Determines if the SNP is a transition, transversion, or indel. Then the amount of each is output.

reduceAlignment : A shell script reduces the Alignment in the delta file to an assembly delta file. The process was accomplished by a series of Perl scripts we combined to call *reduceAlignment*. First, we pulled a list of the contig read ids from the “contig file”. Then we retrieved the sequence id number from the fasta file. We could then obtain a list of the sequence numbers that were in the contig file. The sequence numbers includes the entire header of the fasta file. We only want the internal ids from that file, so we pulled those from the file. Our next step was to copy the sequences from the delta file that have the same internal ids as we found from the contig file. This allowed us to focus only on the data that was used in the assembly.

runTurboShilpa Runs the programs and scripts involved in producing the site by site analysis of the scaffold. Requires *SeqAlignGenerator*, *align-summary*, *getRefPosnBase*, and *combineRefDelta*

SNPcluster: The SNPcluster script determines how many SNPs are next to each other. The script will output the number of SNPs that are in pairs, three of a kind and so on. These are broken into base substitutions and indels as well.

SNPvariationWindow: Given the particular window size specified in the input, it will divide the scaffold into sections of that window size and determine how many SNPs are in each section.

APPENDIX D: Scaffold statistics from TCO comparative assembly

Scaffold	scaffold size	Analyzed Sites	Over16	Under4	Ns	single base subs	segmental base subs	single indels	segmental indels	# genes
1	4193030	3686992	120481	166558	134351	3232	271	966	1981	622
2	3740169	3304942	109786	180232	67038	3476	284	1003	2370	681
3	3777634	3272538	104178	160113	166658	3630	275	1064	2330	714
4	3075709	2594574	106714	142788	164678	2398	238	707	1332	623
5	2511979	2143091	77416	126828	107558	2074	163	627	1415	386
6	2406117	2011168	100416	108857	139657	2228	204	591	936	643
7	2324446	1982890	69169	118615	100981	1870	177	527	1172	363
8	2335496	2027666	59329	87794	116993	1810	150	578	1118	408
9	2251199	1878670	73292	113438	138752	2482	205	629	1547	397
10	2169786	1903562	63411	97277	61190	1763	184	487	802	521
11	2187803	1906664	59557	99768	71371	1644	116	481	1045	402
12	2218424	1688893	111437	118871	260747	1122	62	360	414	444
13	1912767	1541253	69229	102128	153469	1682	143	442	875	425
14	1680605	1437347	64743	91095	52857	924	89	307	220	370
15	1740524	1419358	65650	78484	138583	1261	76	329	694	319
16	1679881	1230636	54023	112613	243591	1003	77	337	609	306
17	1546548	1353977	46705	76937	40002	1141	128	419	638	425
18	1446981	1271770	55128	80005	19691	1610	146	376	764	439
19	1465945	1265241	35872	71014	70669	1310	104	398	855	201
20	1424641	1208836	37282	61755	88173	892	73	253	620	279
21	1469790	1114730	47965	101328	167706	1312	118	311	612	278
22	1403856	1113202	38181	81149	131383	505	88	190	34	242
23	1354378	1103788	41831	76977	99036	1151	98	302	686	230
24	1335609	1143516	43944	61061	64346	1054	118	278	627	248
25	1283025	1078929	47291	66620	70202	1499	156	364	849	283
26	1248150	984029	49651	72203	107553	946	74	191	261	207
27	1252696	977220	36230	96317	107814	818	67	189	179	179
28	1197646	1055397	40688	44473	36468	744	49	236	428	217
29	1263923	962985	42814	67350	157881	751	73	233	406	243
30	1199123	843259	16945	153502	132450	2028	175	515	1503	197
31	1161896	826921	80761	54947	157227	311	24	110	92	172
32	1144766	957687	39349	56273	65501	1068	70	300	804	153
33	1099917	825750	42094	93903	104234	925	95	220	443	195
34	1101474	880715	61750	63876	74795	280	24	169	18	196

35	1097241	904048	20751	62577	87649	338	28	166	60	165
36	1192092	857438	63516	71774	175037	937	125	236	421	325
37	1048454	842200	39440	84452	49731	1224	116	227	357	156
38	1091505	899433	25936	61462	84018	312	14	148	81	159
39	1058928	917167	29606	53952	38045	1139	107	276	489	167
40	1086000	753848	44463	100715	152112	990	84	188	409	196
41	1051352	825328	26189	94974	76672	1642	90	262	655	177
42	1066612	938275	37412	41472	31504	910	71	238	546	155
43	1050224	929886	27149	44955	31346	1361	130	349	1027	155
44	1043815	756163	57771	99379	84363	1740	94	271	630	196
45	945371	672128	24793	88507	121190	657	48	155	223	169
46	1027350	773699	43406	55576	129910	804	72	259	557	168
47	944016	822221	24564	40312	36280	731	38	236	441	156
48	945889	743661	26764	89931	58795	1400	112	191	387	164
49	913200	716210	53319	65895	55337	892	36	221	428	130
50	887758	703997	30760	67122	63167	322	38	131	82	135
51	914909	728310	18789	54487	95042	740	67	225	502	156
52	860121	714797	33215	47735	47823	304	16	118	145	157
53	870110	753715	32227	35657	35046	648	57	185	368	131
54	841328	616648	29306	86886	83259	670	44	129	103	161
55	869239	635821	32097	77279	99693	915	80	146	250	109
56	847024	697989	21518	62387	43160	372	31	121	118	103
57	797877	621483	36256	49092	72667	283	14	107	40	129
58	747206	537429	34110	56910	92850	827	70	160	328	122
59	761483	530254	48364	101888	61701	1168	77	192	451	93
60	757983	623917	55069	39563	25004	675	50	212	409	142
61	749149	687439	21679	27339	2302	714	63	252	460	180
62	748003	657645	18308	40950	20060	777	71	215	668	125
63	734086	579675	28470	48399	58057	923	75	210	545	127
64	750253	548938	23232	76492	71440	1442	105	200	373	120
65	729016	518273	50143	64260	66042	706	59	122	328	137
66	809393	608088	35390	35104	120475	328	19	125	95	136
67	727487	547407	29108	40537	85179	509	47	155	217	174
68	677430	376119	45891	57516	164151	540	42	93	170	86
69	678846	461089	44619	48311	111613	363	24	99	79	95
70	733211	548958	18134	49488	100863	612	83	143	155	191
71	677205	381899	25607	117355	117650	570	48	72	174	90
72	652215	562459	20522	36454	21666	687	76	128	214	174
73	586677	428353	34165	38011	68479	265	20	89	8	88
74	640712	570197	15839	29163	12690	631	47	154	370	61
75	626918	509653	37145	31955	35512	390	22	130	318	121

76	624867	542033	12687	40838	15020	658	61	171	401	86
77	611871	514226	22083	35473	23111	511	32	144	324	108
78	576004	445137	55889	32688	34662	323	45	90	136	105
79	589484	492148	25561	28498	30026	171	6	80	45	134
80	560210	479410	29002	29518	13718	331	24	111	127	92
81	679555	455608	19756	34696	157638	203	14	82	39	91
82	564057	444288	11316	31963	64677	185	4	69	22	64
83	642089	476155	15559	22553	119919	431	39	126	276	105
84	554210	410834	10337	41817	74171	451	16	105	314	47
85	551187	362597	16661	40014	103385	128	10	49	31	84
86	533890	415471	25880	34452	44407	639	54	140	311	128
87	522205	451500	22756	26366	10560	719	43	169	428	139
88	518914	456354	14617	24186	8073	478	46	130	304	96
89	666227	460195	15148	23055	157689	256	12	100	85	71
90	518052	446539	19883	26369	21533	440	36	148	458	96
91	522894	396583	20906	31637	60031	535	50	102	169	104
92	500392	410032	28643	37212	17319	215	12	63	30	75
93	480070	317305	11682	71643	56662	411	38	93	120	68
94	501917	383115	25938	22383	56671	122	2	58	12	95
95	476616	371763	13757	36904	35136	212	9	63	43	80
96	444560	274339	42732	31202	74959	169	21	54	31	95
97	482554	417908	12289	19543	23920	446	30	112	228	95
98	511364	311261	6661	68609	108706	650	49	114	238	90
99	475394	390091	26075	43508	6026	360	33	112	228	91
100	471848	423196	10702	23554	8392	433	40	125	228	102

APPENDIX E: Genetic and physical mapping data

<u>scaffold</u>	<u>start</u>	<u>stop</u>	<u>Distance(Mb)</u>	<u>cM/Mb</u>	<u>Chr</u>	<u>map</u> <u>ID</u>	<u>cM</u>	<u>cM/Mb</u>	<u>%SNP</u>
scaffold_1	1892387	2183854	291467	37.1	2	d162	10.8	37.1	0.053
scaffold_1	2184113	2459188	275075	28.4	2	d124	7.8	28.4	0.076
scaffold_1	2459642	2755478	295836	30.8	2	d079	9.1	30.8	0.104
scaffold_1	2755698	3257753	502055	23.7	2	d069	11.9	23.7	0.100
scaffold_1	3258007	3532515	274508	34.2	2	d183	9.4	34.2	0.095
scaffold_1	3510933	4020601	509668	12.6	2	d015	6.4	12.6	0.080
scaffold_1	3532645	4020601	487956	13.1	2	d047	6.4	13.1	0.081
scaffold_2	465115	1247096	781981	25.7	3	d147	20.1	25.7	0.143
scaffold_2	1247523	1834050	586527	17.6	3	d075	10.3	17.6	0.103
scaffold_2	1834370	2983993	1149623	20.4	3	d136	23.5	20.4	0.076
scaffold_2	1851638	2983993	1132355	20.8	3	d108	23.5	20.8	0.076
scaffold_3	739030	1383978	644948	40.6	1	d007	26.2	40.6	0.104
scaffold_3	1384162	1923837	539675	23.3	1	d091	12.6	23.3	0.097
scaffold_3	1923837	2275958	352121	15.6	1	d130	5.5	15.6	0.134
scaffold_3	1923990	2275958	351968	15.6	1	d001	5.5	15.6	0.134
scaffold_3	2276376	2318013	41637	40.8	1	d163	1.7	40.8	0.173
scaffold_3	2318270	2483839	165569	26.6	1	d174	4.4	26.6	0.130
scaffold_3	2484047	3015192	531145	22.0	1	d103	11.7	22.0	0.126
scaffold_3	3015577	3212697	197120	45.7	1	d053	9	45.7	0.135
scaffold_5	109480	117840	8360	705.7	12	d184	5.9	705.7	0.104
scaffold_5	118469	126598	8129	61.5	12	d182	0.5	61.5	0.066
scaffold_8	214644	1410937	1196293	36.7	4	d106	43.9	36.7	0.088
scaffold_8	1411116	1927328	516212	0.6	4	d105	0.3	0.6	0.081
scaffold_8	1927865	2259957	332092	9.9	4	d155	3.3	9.9	0.116
scaffold_9	848851	1082325	233474	22.7	9	d145	5.3	22.7	0.116
scaffold_9	1082581	1369547	286966	48.1	9	d118	13.8	48.1	0.060
scaffold_9	1369704	2143217	773513	36.3	9	d043	28.1	36.3	0.108
scaffold_11	203644	343688	140044	30.7	4	d139	4.3	30.7	0.142
scaffold_11	343786	725719	381933	19.1	4	d116	7.3	19.1	0.146
scaffold_12	368224	1024971	656747	0.8	5	d087	0.5	0.8	0.071
scaffold_13	1074377	1333322	258945	27.8	9	d011	7.2	27.8	0.167
scaffold_13	1333322	1397633	64311	4.7	9	d149	0.3	4.7	0.172
scaffold_21	598442	813165	214723	18.2	3	d082	3.9	18.2	0.121
scaffold_21	598442	835104	236662	16.5	3	d177	3.9	16.5	0.122
scaffold_21	835598	1051139	215541	18.1	3	d003	3.9	18.1	0.132
scaffold_21	1051430	1248646	197216	10.1	3	d122	2	10.1	0.107

scaffold_24	142351	272150	129799	42.4	11	d173	5.5	42.4	0.143
scaffold_29	225287	424770	199483	42.6	10	d186	8.5	42.6	0.082
scaffold_30	741780	883779	141999	26.8	2	d004	3.8	26.8	0.569
scaffold_32	244697	253489	8792	136.5	6	d014	1.2	136.5	0.011
scaffold_32	253637	295131	41494	74.7	6	d035	3.1	74.7	0.129
scaffold_32	295288	376888	81600	38.0	6	d142	3.1	38.0	0.120
scaffold_32	377193	471100	93907	25.6	6	d135	2.4	25.6	0.032
scaffold_39	5470	284755	279285	30.1	5	d093	8.4	30.1	0.158
scaffold_39	284948	379327	94379	35.0	5	d030	3.3	35.0	0.060
scaffold_39	379458	427971	48513	72.1	5	d042	3.5	72.1	0.046
scaffold_39	428255	589170	160915	29.8	5	d024	4.8	29.8	0.048
scaffold_39	589286	1016558	427272	46.1	5	d031	19.7	46.1	0.130
scaffold_42	112556	132315	19759	106.3	1	d170	2.1	106.3	0.232
scaffold_42	132609	674164	541555	22.0	1	d098	11.9	22.0	0.079
scaffold_43	120207	254164	133957	23.1	4	d172	3.1	23.1	0.070
scaffold_43	254353	379974	125621	43.0	4	d180	5.4	43.0	0.157
scaffold_46	456798	483793	26995	44.5	7	d107	1.2	44.5	0.131
scaffold_49	90586	580271	489685	0.8	10	d034	0.4	0.8	0.081
scaffold_49	580950	611970	31020	90.3	10	d005	2.8	90.3	0.185
scaffold_53	193436	289000	95564	22.0	1	d181	2.1	22.0	0.086
scaffold_53	289336	515330	225994	19.0	1	d114	4.3	19.0	0.079
scaffold_53	515514	685669	170155	19.4	1	d096	3.3	19.4	0.046
scaffold_62	51836	124649	72813	54.9	3	d054	4	54.9	0.066
scaffold_63	450291	480790	30499	52.5	2	d074	1.6	52.5	0.284
scaffold_89	311220	542357	231137	17.3	5	d164	4	17.3	0.027
scaffold_91	391760	401836	10076	158.8	7	d191	1.6	158.8	0.095

APPENDIX F: Windows analysis: Regions with no SNPs.

scaffold	start	Stop	Length
66	732001	761000	28999
17	184001	207000	22999
18	648001	669000	20999
34	1005001	1026000	20999
39	544001	565000	20999
4	1281001	1301000	19999
10	1349001	1369000	19999
16	1307001	1327000	19999
81	322001	342000	19999
1	4144001	4163000	18999
2	3534001	3553000	18999
3	3578001	3597000	18999
7	656001	675000	18999
8	134001	153000	18999
11	1527001	1546000	18999
14	1492001	1511000	18999
51	79001	98000	18999
63	68001	87000	18999
94	380001	399000	18999
1	2271001	2289000	17999
10	296001	314000	17999
10	400001	418000	17999
26	261001	279000	17999
29	8001	26000	17999
45	261001	279000	17999
70	19001	37000	17999
1	1860001	1877000	16999
5	710001	727000	16999
8	481001	498000	16999
9	226001	243000	16999
9	921001	938000	16999
9	1383001	1400000	16999
14	1088001	1105000	16999
21	363001	380000	16999
27	970001	987000	16999

scaffold	start	stop	length
47	641001	658000	16999
57	17001	34000	16999
61	42001	59000	16999
88	381001	398000	16999
5	2394001	2410000	15999
6	1935001	1951000	15999
7	1539001	1555000	15999
8	1235001	1251000	15999
11	1637001	1653000	15999
15	1133001	1149000	15999
26	141001	157000	15999
33	962001	978000	15999
49	309001	325000	15999
56	240001	256000	15999
65	170001	186000	15999
70	38001	54000	15999
1	3914001	3929000	14999
8	1772001	1787000	14999
8	2138001	2153000	14999
9	1099001	1114000	14999
9	1251001	1266000	14999
10	317001	332000	14999
13	446001	461000	14999
13	1492001	1507000	14999
14	446001	461000	14999
15	340001	355000	14999
15	1275001	1290000	14999
22	565001	580000	14999
31	363001	378000	14999
34	577001	592000	14999
35	765001	780000	14999
35	881001	896000	14999
36	505001	520000	14999
39	405001	420000	14999

1	1407001	1421000	13999
1	1829001	1843000	13999
3	1413001	1427000	13999
4	2188001	2202000	13999
5	345001	359000	13999
6	753001	767000	13999
6	1370001	1384000	13999
8	289001	303000	13999
10	89001	103000	13999
11	1786001	1800000	13999
12	945001	959000	13999
13	831001	845000	13999
20	395001	409000	13999
22	1040001	1054000	13999
27	1126001	1140000	13999
27	1141001	1155000	13999
28	149001	163000	13999
29	661001	675000	13999
35	470001	484000	13999
38	293001	307000	13999
45	309001	323000	13999
50	863001	877000	13999
52	322001	336000	13999
56	499001	513000	13999
57	161001	175000	13999
61	680001	694000	13999
65	64001	78000	13999
71	241001	255000	13999
79	211001	225000	13999
85	464001	478000	13999
90	257001	271000	13999
92	402001	416000	13999
95	148001	162000	13999
95	198001	212000	13999
96	326001	340000	13999
2	2841001	2854000	12999
3	2472001	2485000	12999
4	120001	133000	12999
4	2481001	2494000	12999
5	846001	859000	12999
5	1919001	1932000	12999

5	2329001	2342000	12999
6	288001	301000	12999
8	244001	257000	12999
9	1119001	1132000	12999
10	1963001	1976000	12999
11	1719001	1732000	12999
11	1953001	1966000	12999
14	93001	106000	12999
14	588001	601000	12999
14	864001	877000	12999
16	661001	674000	12999
17	1061001	1074000	12999
19	444001	457000	12999
19	648001	661000	12999
19	922001	935000	12999
21	1037001	1050000	12999
22	269001	282000	12999
26	396001	409000	12999
27	703001	716000	12999
29	546001	559000	12999
31	311001	324000	12999
34	380001	393000	12999
37	727001	740000	12999
38	731001	744000	12999
38	1005001	1018000	12999
45	527001	540000	12999
50	719001	732000	12999
53	266001	279000	12999
62	94001	107000	12999
68	187001	200000	12999
70	55001	68000	12999
72	334001	347000	12999
87	136001	149000	12999
87	496001	509000	12999
96	169001	182000	12999
1	2330001	2342000	11999
2	720001	732000	11999
2	1574001	1586000	11999
3	1964001	1976000	11999
4	1003001	1015000	11999
7	1324001	1336000	11999

10	906001	918000	11999
10	1136001	1148000	11999
11	2135001	2147000	11999
12	799001	811000	11999
13	1051001	1063000	11999
14	610001	622000	11999
14	726001	738000	11999
14	1063001	1075000	11999
15	1472001	1484000	11999
16	1241001	1253000	11999
17	1381001	1393000	11999
18	1054001	1066000	11999
19	311001	323000	11999
22	433001	445000	11999
22	504001	516000	11999
22	520001	532000	11999
23	1061001	1073000	11999
24	975001	987000	11999
26	585001	597000	11999
27	1092001	1104000	11999
28	27001	39000	11999
30	1028001	1040000	11999
31	169001	181000	11999
31	968001	980000	11999
32	1034001	1046000	11999
34	1078001	1090000	11999
35	230001	242000	11999
35	436001	448000	11999
35	856001	868000	11999
35	934001	946000	11999
36	264001	276000	11999
38	277001	289000	11999
38	874001	886000	11999
38	983001	995000	11999
39	739001	751000	11999
40	270001	282000	11999
41	695001	707000	11999
50	141001	153000	11999
50	704001	716000	11999
52	367001	379000	11999
53	639001	651000	11999

57	276001	288000	11999
68	174001	186000	11999
70	547001	559000	11999
73	265001	277000	11999
78	98001	110000	11999
79	147001	159000	11999
80	114001	126000	11999
87	475001	487000	11999
90	189001	201000	11999
94	458001	470000	11999
1	1959001	1970000	10999
1	2073001	2084000	10999
2	2192001	2203000	10999
2	3226001	3237000	10999
3	206001	217000	10999
3	232001	243000	10999
3	1268001	1279000	10999
3	3665001	3676000	10999
4	867001	878000	10999
4	2712001	2723000	10999
4	3016001	3027000	10999
5	973001	984000	10999
5	1044001	1055000	10999
5	1843001	1854000	10999
5	2255001	2266000	10999
6	217001	228000	10999
6	1110001	1121000	10999
6	1257001	1268000	10999
6	2205001	2216000	10999
9	566001	577000	10999
10	881001	892000	10999
10	1024001	1035000	10999
10	1763001	1774000	10999
10	1823001	1834000	10999
12	1822001	1833000	10999
14	1001001	1012000	10999
14	1545001	1556000	10999
16	1187001	1198000	10999
16	1223001	1234000	10999
17	846001	857000	10999
17	1131001	1142000	10999

18	37001	48000	10999
18	1165001	1176000	10999
21	1281001	1292000	10999
23	419001	430000	10999
25	714001	725000	10999
26	379001	390000	10999
26	422001	433000	10999
26	539001	550000	10999
31	266001	277000	10999
31	449001	460000	10999
34	269001	280000	10999
34	857001	868000	10999
35	410001	421000	10999
35	732001	743000	10999
36	1052001	1063000	10999
37	199001	210000	10999
37	750001	761000	10999
38	567001	578000	10999
38	609001	620000	10999
47	481001	492000	10999
47	494001	505000	10999
50	308001	319000	10999
51	99001	110000	10999
51	418001	429000	10999
53	545001	556000	10999
56	257001	268000	10999
58	672001	683000	10999
67	55001	66000	10999
70	394001	405000	10999
71	550001	561000	10999
72	322001	333000	10999
75	115001	126000	10999
76	222001	233000	10999
83	199001	210000	10999
85	178001	189000	10999
86	508001	519000	10999
88	360001	371000	10999
89	12001	23000	10999
97	313001	324000	10999
100	85001	96000	10999

234	15332	A	A	T	Ts
235	15333	-	A	T	INDEL
236	15334	-	C	T	INDEL
237	15335	C	C	T	Ts
238	15336	A	A	T	Tv
239	15337	A	A	T	Tv
240	15338	-	-	A	INDEL
241	15339	-	-	T	INDEL
242	15340	-	-	T	INDEL
243	15341	-	-	A	INDEL
244	15342	-	-	C	INDEL
245	15343	-	-	C	INDEL
246	15344	-	-	A	INDEL
247	15345	-	-	A	INDEL

APPENDIX H: KOG annotations of *D. pulex* gene duplicates with Ks/Ks > 2

#N/A

Beta-transducin family (WD-40 repeat) protein

Beta-tubulin folding cofactor D

C-type lectin

Carbonic anhydrase

Chromatin assembly factor-I

Cytoplasmic exosomal RNA helicase SKI2, DEAD-box superfamily

E3 ubiquitin ligase

Enolase-phosphatase E-1

Focal adhesion tyrosine kinase FAK, contains FERM domain

Nucleolar GTPase/ATPase p130

Predicted esterase of the alpha-beta hydrolase superfamily (Neuropathy target esterase), contains cAMP-binding domains

RNA polymerase II, large subunit

Serine carboxypeptidases

Serine proteinase inhibitor (KU family) with thrombospondin repeats

Traf2- and Nck-interacting kinase and related germinal center kinase (GCK) family protein kinases

Translation initiation factor 4F, ribosome/mRNA-bridging subunit (eIF-4G)

Trypsin

Uncharacterized conserved protein

Uncharacterized conserved protein H4

von Willebrand factor and related coagulation proteins

APPENDIX I: Intron absences in TRO

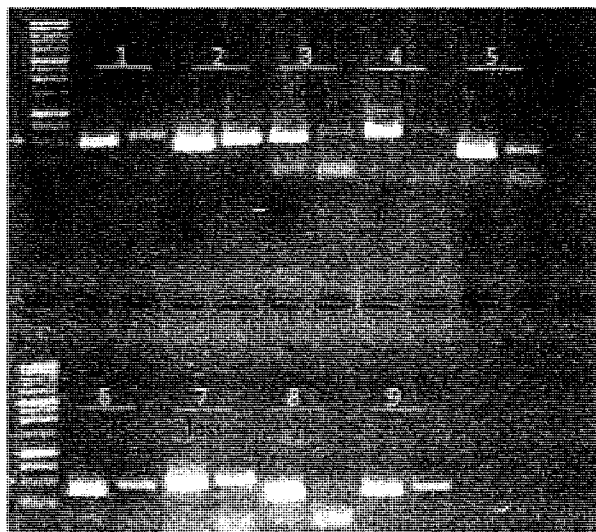


Figure 34: PCR amplification of putative intron polymorphisms in TRO and TCO, Absences are observed by smaller PCR products in TRO (left in each pair).