The Pennsylvania State University

The Graduate School

Department of Biology

EVOLUTIONARY DYNAMICS OF THE U12-TYPE SPLICEOSOMAL INTRONS IN MULTICELLULAR ORGANISMS

A Dissertation in

Biology

by

Chiao-Feng Lin

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2008

The Dissertation of Chiao-Feng Lin was reviewed and approved* by the following:

Wojciech Makalowski
Professor and Director of Institute of Bioinformatics
University of Muenster, Germany
Dissertation Advisor
Co-chair of Committee

Edward Holmes
Professor of Biology
Co-chair of Committee

Webb Miller
Professor of Biology and Computer Science and Engineering

Ross C. Hardison
T. Ming Chu Professor of Biochemistry and Molecular Biology

Stephen M. Mount
Associate Professor of
Department of Cell Biology and Molecular Genetics,
University of Maryland, College Park, Maryland
Special Member

Douglas R. Cavener
Professor of Biology
Head of the Department of Biology

*Signatures are on file in the Graduate School

ABSTRACT

Many multicellular eukaryotes have two types of spliceosome for the removal of introns from messenger RNA precursors. The major (U2) -type spliceosome processes the vast majority of introns, while the minor (U12) -type spliceosome removes the small fraction (less than 0.5%) of introns referred to as U12-type introns. U12-type introns have distinct sequence elements, and mostly occur in genes together with U2-type introns. A phylogeny of species with and without U12-type introns shows that the minor splicing pathway has been lost repeatedly in evolution. With four parallel and one shared snRNPs, the two spliceosome function in a similar, cooperative, and competitive manner. The possibility of U12-introns converting to U2-type introns make the evolutionary dynamics of U12-type introns more complicated than just gain and loss. The studies in this thesis include computational identification and characterization of U12-type introns, and the evolutionary dynamics of U12-type introns.

I have investigated the evolution of U12-type introns among 18 metazoan genomes by analyzing orthologous U12-type intron clusters. Examination of gain, loss and type switching shows that intron type is remarkably conserved among vertebrates. Among 180 intron clusters, only eight show intron loss in any species and only five show U12 to U2 conversion. In contrast to the other insect species investigated, Dipteran genomes are characterized by a rapid evolution (or loss) of components of the U12 spliceosome and a striking loss of U12-type introns. Nevertheless, we find one case of U2 to U12 conversion, apparently mediated by activation of a cryptic U12 splice site, in Diptera. Overall, loss of U12-type introns is more common than conversion to U2-type.

U12 to U2 conversion occurs more frequently among introns of the GT-AG subtype. I also found support for natural U12-type introns with non-canonical terminal dinucleotides (CT-AC, GG-AG, and GA-AG) that have not been reported previously.

My most salient finding is that U12 introns are extremely stable in some taxa, including eutheria. U12-type intron loss is more frequent than conversion to the major type. The degeneracy of U12-type terminal dinucleotides among natural U12-type introns is higher than previously thought.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

Chiao-Feng Lin's contributions to the study in CHAPTER 4

The piece of work presented in Chapter 4 is collaboration between Mount's lab (University of Maryland, College Park) and Makalowski's lab. For that project, I computationally identified U12-type introns in 16 metazoan genomes (*Anopheles gambiae, Apis mellifera, Bos taurus, Canis familiaris, Ciona intestinalis, Drosophila melanogaster, Danio rerio, Fugu rubripes, Gallus gallus, Homo sapiens, Monodelphis domestica, Mus musculus, Pan troglodytes, Rattus norvegicus, Tetraodon nigroviridis, and Xenopus tropicalis*). The methods used are based on Ensemble annotations. They are described in that chapter and differ from those in CHAPTER 2.

Previously published middle-author papers are in Appendices A, B and C.

ACKNOWLEDGEMENTS

I acknowledge my thesis advisor, Dr. Wojciech Makalowski, for providing the opportunity to make all this possible. I am deeply grateful for his enormous patience with his headstrong student who resists moving on, and the United-Nations-like lab that he maintained. Vamsi Veeramachaneni not only helped me to pick up perl scripting quickly but also showed me how a computer scientist can be an idealistic but realistic philanthropist. Valer Gotea, a science-enthusiast, was always encouraging and never ceased nudging me into starting perl scripting, in which he succeeded. Dimitra Chalkia, a wildish but melancholy girl, from the land of myths and philosophers. It is a great honor and fulfilling experience to work with all these good people from different corners of the planet. I also acknowledge Dr. Izabela Makalowska for advising on the research of the overlapping gene.

I thank my committee members, Drs. Eddie Holmes, Webb Miller, and Ross Hardison for giving insightful guidance. The splicing expertise of my special committee member, Dr. Steve Mount (University of Maryland) made me see the big picture of the splicing world and where my little puzzle piece of U12-intron fits in the big puzzle. Working with Dr. Hiroshi Akashi ushered me to the wonderland of DNAs and evolution. I truly appreciate Wen-Ya Ko's enlightenments on so many things in so many ways.

My one year lost in translation in Muenster Germany would not have been possible if it were not for the two secretaries, Kathryn McClintock of the Biology Department and Wolfgang Garbers of the Institute of Bioinformatics at University of Muenster, who took care of complex administration issues on both sides of the Atlantic.

I was partially supported by the Center for Comparative Genomics and Bioinformatics, part of Penn State's Huck Institutes of the Life Sciences while working on the research included in this thesis.

Finally, I thank my mother for taking care of herself and my sister for looking after my mother so that I could selfishly pursue my dream

CHAPTER 1

INTRODUCTION

Sequences within precursors of the messenger RNAs of many eukaryotic protein-coding genes are split by introns. Before the RNA molecule can be translated into amino acids, introns have to be removed and exons connected. This process is known as pre-mRNA splicing and is performed by a machinery called spliceosome, which consists of U1, U2, U4, U5, U6 snRNPs (small nuclear ribonucleoproteins) and more than 200 non-snRNP proteins (Sharp, 1994; Patel and Steitz, 2003; Will and Luhrmann, 2005). Each snRNP has a small nuclear RNA associated with several proteins including a common set of core Sm proteins. These snRNAs (specifically U2 and U6 snRNAs) catalyze the major reactions of splicing. However, many proteins, including both snRNP proteins and non-snRNP proteins such as SR proteins, are also required. The spliceosome does not come and function as one entity. Instead, snRNPs associate with the pre-mRNA and each other in a complex and dynamic manner.

Spliceosomes at different stages of a splicing event are called the E, A, B, and C complexes. A spliceosomal cycle has been proposed to describe the order in which these complexes form. The sequence is generally as follows, but may be somewhat flexible (Sharp 2005). 1) When U1 binds the 5' splice site, this is E complex. 2) U2AF (U2 auxiliary factor) binds the poly-pyrimidine tract and then recruits U2snRNP to bind the branch point adenosine in an ATP-dependent manner. Thus, the A complex contains U1 and U2 snRNPs. A U4/U6.U5 tri-snRNP joins on the A complex to form the B complex.

After U1 and U4 dissociate, the B complex is catalytically active and ready for the first phosphotransfer reaction between the 5' end of the intron and the branch site. The first step of splicing generates a lariat intermediate in the C complex. Then the second phosphotransfer reaction between the 3'end of the upstream exon and the 5' end of the downstream exon takes place, and results in ligated exons and an intron lariat.

The terminal dinucleotide GT and AG at the 5' and 3' end of introns are nearly universal (Breathnach and Chambon, 1981; Sharp, 2005; references therein). However, compilation of introns with non-consensus splice sites (Jackson 1991; Hall and Padgett 1994) led to the discovery of a rarer class of introns. Such introns have an extended and nearly invariant 5' splice site (ATATCCTT at +1 to +8 positions starting from the 5' junction) and a more conserved Branch Point Site (TCCTTAAC near the 3' end), and an AC acceptor site. Possible base-pairing between the 5' splice site and U11 snRNAs, and between the BPS and U12 snRNAs, prompted Hall and Padgett (1994) to propose that the low-abundance U11 and U12 snRNAs are involved in the splicing of this new class of introns. These snRNAs had been described earlier (Montaka and Steitz, 1988), but their function had remained unknown. Later, attempts by Hall and Padgett (1996) and Tarn and Steitz (1996a; 1996b) pieced together the major pieces of a jigsaw puzzle for a novel spliceosome, which consists of U11, U12, U4atac, U5, U6atac snRNPs, and non-snRNP proteins. The rare introns were called AT-AC (pronounced as "attack") introns for their unusual terminal dinucleotides. However, Dietrich *et al.* (1997) later changed the view that introns with GT-AG termini are always processed by the major type spliceosome while those with AT-AC are always processed by the minor type. The authors demonstrated that an AT-AC intron with double mutations at the first and last

nucleotides, which turned the termini to GT-AG, could be properly removed by the minor-type spliceosome. On the other hand, splicing of intron 21 of the human *SCN4A* gene, which has AT-AC termini, requires the major-type spliceosome (Dietrich *et al.*, 1997). Independently, Wu and Krainer (1997) also reached the same conclusion on this AT-AT U2-type intron, or what they termed "AT-AC II intron". Since terminal dinucleotides do not necessarily indicate the splicing pathway that an intron undergoes, Dietrich *et al.* (1997) coined the terms "U12-dependent" and "U2-dependent" to denote the introns and spliceosome involved in the minor and the major splicing pathway, respectively. This nomenclature has been widely accepted; frequently, a shorter version, "U12-type" and "U2-type", is used.

The two types of spliceosomes consist of shared and distinct components (Will *et al.*, 1999; Will and Luhrmann 2005). For instance, in addition to U5 snRNPs, Sm proteins, which bind snRNAs, and SF3b are found in both spliceosomes. The same molecular mechanism, including the formation of an intron lariat, two nucleophilic attacks, and similar pre-mRNA-snRNA and snRNA-snRNA interactions, are observed in two spliceosomes (Frilander and Steitz 2001). The snRNAs performing parallel roles have similar secondary structure, namely, U1, U2, U4, and U6 correspond to U11, U12, U4atc, and U6atac, respectively. In the U2-type splicing, U1 snRNA binds the 5' junction and U2 snRNA binds the branch point adenine of introns subsequently. This corresponds to the U11-5' splice site and U12-BPS base-pairing in the U12-type splicing except that U11 and U12 snRNPs form a di-snRNP and assemble on the intron as a unit (Frilander and Steitz, 1999). The U4/U6.U5 tri-snRNP complex in the U2-type spliceosome has its counterpart (U4atac/U6atac.U5) in the U12-type.

In the major-type splicing, it has been shown that the binding of U1 snRNP at the downstream 5' splice site noticeably increases the splicing efficiency (Berget, 1995). The enhancement is likely facilitated by non-snRNP proteins, such as SR proteins, which bridge over exons and interact with the U2AF35 at the 3'splice site of the upstream intron as well as the U1 at the downstream 5'splice site (Berget, 1995; Reed, 1996). It is particularly interesting that this so-called exon-definition interaction was observed in the minor type splicing as well (Wu and Krainer, 1996). Nearly all U12-type introns reside with U2-type introns in the same genes. In this context, exon-definition interaction indicates that the minor and the major type spliceosomes function cooperatively. SR proteins are characterized by their RRM (RNA Recognition Motif) domains and an arginin-serine rich domain at the C-terminal. In addition to facilitating U1-mediated exon-definition interaction, SR proteins also so promote splicing in a U1-independent manner by cross-linking to purine-rich enhancers in exons (ESEs; Exonic Sequence Enhancers) in both splicing pathways (Reed, 1996; Wu and Krainer, 1998; Wu and Krainer, 1999).

U12- and U2-type introns not only link, through exon-definition, in the physical space but also in (evolutionary) time. In the first study that investigated the evolution of U12-type introns, Burge *et al.* (1998) proposed that U12-type introns can be "converted" into U2-type introns. They described eight sets of U12-type introns having U2-type orthologs (occurring at orthologous codon position and the same intron phase) as the outcome of U12-U2 conversion. Given that U12-type splicing signals are more conserved they noted that such type conversion is likely to occur only from U12- to U2-type but not vice versa. They also found orthologous U12-type introns having both GT-AG (including

GC-AG) and AT-AC (with slight variation at the last nucleotide) subtypes, depending on their terminal dinucleotides. They called this mixture of subtypes "subtype switch". They argued that such U12-U2 conversion and subtype switch are consistent with the experimental results. As Dietrich *et al.* (1997) demonstrated, an AT-AC U12-type intron with mutations at the first and the last nucleotides that turn the termini to GT-AG is still processed by the U12-type spliceosome, a subtype switch. An additional mutation (C to G at +5 position) converted the AT-AC U12-type intron into a GT-AG U2-type. This also suggests that a subtype switch from AT-AC to GT-AG can serve as an intermediate stage for a type conversion from U12 to U2.

Burge *et al.* (1998) identified U12-type introns using position specific weight matrix approach. They were found in vertebrates, insects, cnidarians, and plant *Arabidopsis thaliana* but not in *Caenorhabditis elegans*, two yeasts, and protists. BLASTN search for U6atac snRNA found putative orthologs in *A. thaliana* and *Drosophila melanogaster* but, again, not in the near complete *C. elegans* genome. The presence of U12-type introns and snRNAs in both plants and animals (but not all) suggests that the U12-type splicing pathway existed in the common ancestor, and it has been lost in nematodes and probably in other species, such as yeasts, too. Therefore, the later identification of traces of the minor-type splicing pathway (snRNAs, proteins, and/or introns) in a fungus (*Rhizopus oryzae*) and several protists (Russel *et al.*, 2006) is not completely surprising, but provides evidence for an early origin of the U12-type splicing. The view on the wide spread of the U12-type spliceosome in eukaryotes and its independent loss in distantly related species gained further support in a recent

computational scanning for spliceosomal snRNA genes in 149 eukaryotic genomes (Lopez *et al.*, 2008).

The question of whether these two seemingly parallel splicing machineries descended from the same ancestor is intriguing but remains unresolved. Burge *et al.* (1998) proposed the fission-fusion hypothesis; two splicing pathways diverged from one in the course of speciation (fission) and later merged (fusion) in one organism. They reasoned that the similar secondary structure and interaction between the two sets of spliceosomal snRNAs are unlikely to be the results of convergent evolution despite the lack of detectable sequence similarities between U1 and U11 or between U2 and U12 snRNAs. Since that work, this argument has been strengthened by obvious homology between protein components of the U11/U12 snRNP and proteins in the U1 and U2 snRNPs (Will *et al.*, 2004). However, Lynch and Richardson (2002) do not agree with this view. They argued that two splicing systems originated from two group II self-splicing introns, and that the similarity between functionally analogous snRNAs was shaped by the proteins involved in splicing. In this model, homologous proteins in the two spliceosomes can be explained by duplication and specialization of genes for proteins that were originally shared.

Before the discovery of the minor-type splicing pathway, the evolution of spliceosomal intron has been of great interests because of its implication in the evolution of gene structure and, in turn, evolution of genome complexity. While group I and group II self-splicing introns are found in prokaryotic genomes, spliceosomal introns are found exclusively in the nuclear genomes of all eukaryotes. The origin of introns has long been debated with competing hypotheses being “introns- early” and “introns-late”. The dispute

mainly lies whether the genome of the common ancestor of prokaryotes (intron-free) and eukaryotes (intron-containing) harbour introns or not. In other words, the absence of introns in prokaryotes is because introns arose in eukaryotes or they were lost in prokaryotes. The introns-early school proposes that introns were present and facilitated exon shuffling during formation of gene in the early life form. The striking similarity in splicing mechanism between group II and spliceosomal introns suggests that they share common ancestry; the introns-late school postulates that spliceosomal introns originated through invasion and subsequent fragmentation of group II introns. With the lack of trace of introns in prokaryotes and accumulating evidence pointing toward homology between group II and spliceosomal introns, the debate between two hypotheses has subdued. Furthermore, the presence of a second splicing pathway fits in the introns-late view better.

Lately, many attempts have been made to understand major evolutionary mechanism that drives the evolution of spliceosomal introns in eukaryotes by investigating correspondence of intron position with respect to the resultant protein sequence. The datasets that were investigated ranges from a broader spectrum of genomes or a certain phylognentic group, such as mammals or drosophila. Varying approaches (parsimony- or likelihood-based) have been applied to infer conservation, gain, and loss of introns. Even though results are not always consistent, several trends have started to emerge. First, basal metazoa are relatively intron-rich. Second, intron gain is more active in early eukaryotes and tends to occur in a surgical manner. The heterogeneity in intron abundance among extant eukaryotes is mainly due to variance in retention rate. While intron-poor organisms have experienced elevated intron loss rate,

gain and loss rate remains somewhat balanced in other lineages. However, these balancing gain and loss rates inferred from likelihood approach based on large-scale datasets appear to be puzzling because only few unambiguous cases of intron gain have been observed in lineage-based analyses (Iwamoto *et al.*, 1998; 1999; Hankeln *et al.*, 1997; O'Neill *et al.*, 1998).

Since the publication of Burge and colleagues' evolutionary study of U12-type introns, many complete genomes have become available. Computational identification and characterization of U12-type introns have been done on several genomes, such as *Homo sapiens* (Levine and Durbin, 2001), *D. melanogaster* (Schneider *et al.*, 2004), and *A. thaliana* (Zhu and Brendel, 2003). However, those studies did not fully address evolutionary aspects, which are the subject of this thesis.

The chapters of this thesis are organized as follows. Chapter 2 presents the most up-to-date analysis of U12-type intron distribution and evolutionary dynamics. For completeness, my initial method, and the results obtained with that method, are presented in Chapter 3. In that work, I used sequence and annotation data downloaded from the ensembl and TIGR databases, and a position-specific weight matrix with likelihood ratio approach, to identify U12-type introns in five animal and two plant genomes. Only a few AT-AC U12-type introns were detected, primarily because introns with AT-AC terminal dinucleotides were often misannotated. Widespread under-detection of U12-type introns and a bias towards GT-AG U12-type introns became clear with the release of U12DB (Alioto, 2007), a database of orthologous U12-type intron clusters. However, the initial results, as presented in Chapter 3, showed that intron phase and intron position are

conserved among mammalian (mostly GT-AG) U12-type introns and that U12-type introns are not randomly distributed among genes.

More U12-type introns, particularly those with AT-AC termini, were reported in the U12DB because the database incorporated introns predicted by mapping EST data in addition to genomewide annotation. I retrieved the whole database and investigated the evolution of U12-type introns in sets of genomes. These are 1) eutheria (*Homo sapiens, Mus musculus, Rattus norvegicus, Canis familiaris*), 2) vertebrates (with addition of *Monodelphis domestica, Gallus gallus, Fugu rubripes, Tetraodon nigroviridis, Danio rerio*), 3) chordates (with addition of *Ciona intestinalis*), 4) *H. sapiens vs. D. melanogaster*, 5) *H. sapiens vs. A. thaliana*. As presented in chapter 2, U12-type introns are remarkably conserved among vertebrates. There has been massive loss of U12-type introns in multiple invertebrate lineages. Interestingly, one of the very few (18) U12-type introns that were detected in the *D. melanogaster* genome, was an extremely rare case of U2 to U12-type conversion. This *de novo* U12-type intron appears likely to have arisen from cryptic splice sites near an existing U2-type intron. In general, loss of U12-type introns is more common than conversion to a U2-type. U12 to U2 conversion occurs more frequently among introns of the GT-AG subtype. I also found support for natural U12-type introns with non-canonical terminal dinucleotides (CT-AC, GG-AG, and GA-AG) that have not been reported previously.

Chapter 4 presents the application of this analysis to insect genomes as part of an analysis of the evolution of spliceosomal small nuclear RNA genes. In contrast to the other insect species investigated, Dipteran genomes are characterized both by a rapid

evolution (or loss) of components of the U12 spliceosome and by a striking loss of U12-type introns.

CHAPTER 2

EVOLUTIONARY DYNAMICS OF U12-TYPE SPLICEOSOMAL INTRONS

Abstract

Many multicellular eukaryotes have two types of spliceosome for the removal of introns from messenger RNA precursors. The major (U2) -type spliceosome processes the vast majority of introns, while the minor (U12) -type spliceosome removes a small fraction (less than 0.5%) of introns referred to as U12-type introns. U12-type introns have distinct sequence elements, and usually occur in genes together with U2-type introns. A phylogenetic distribution of U12-type introns shows that the minor splicing pathway appeared very early in eukaryotic evolution and has been lost repeatedly in evolution.

I have investigated the evolution of U12-type introns among 18 metazoan genomes by analyzing orthologous U12-type intron clusters. Examination of gain, loss and type switching shows that intron type is remarkably conserved among vertebrates. Among 180 intron clusters, only eight show intron loss in any vertebrate species and only five show U12 to U2 conversion. There are very few U12-type introns in Diptera (18 in Drosophila). Nevertheless, we find one case of U2 to U12 conversion, apparently mediated by activation of a cryptic U12 splice site, in Diptera. Overall, loss of U12-type introns is more common than conversion to U2-type. U12 to U2 conversion occurs more frequently among introns of the GT-AG subtype. I also found support for natural U12-

type introns with non-canonical terminal dinucleotides (CT-AC, GG-AG, and GA-AG) that have not been reported previously.

Although complete loss of the U12-type spliceosome has occurred repeatedly, U12 introns are extremely stable in some taxa, including eutheria. The degeneracy of U12-type terminal dinucleotides among natural U12-type introns is higher than previously thought.

Introduction

Burge *et al.* (1998) were the first to investigate evolutionary fate of U12-type introns. In addition to intron gain and loss that was normally examined to understand evolution of introns, they observed intron conversion from U12 to U2-type, and subtype switching among U12-type introns. Given that splicing signals in U2-type introns are more degenerate than that those of U12-type introns, they proposed that the conversion is likely to be unidirectional (U12 to U2-type only). Since the publication of their study, many complete genomes have become available. Computational identification and characterization of U12-type introns have been done on several genomes, such as *Homo sapiens* (Levine and Durbin, 2001), *Drosophila melanogaster* (Schneider *et al.*, 2004), and *Aarbidopsis thaliana* (Zhu and Brendel 2003). However, evolutionary aspect was not fully addressed in those studies. Two databases provide access to U12-type introns that were identified in multiple genomes: SpliceRack (Sheth *et al.*, 2006) and U12DB (Alioto, 2007). The U12DB not only has a broader range of genomes (20 eukaryotes); it also provides orthology information (based on Ensembl and Inparanoid databases) among

introns by organizing U12-type introns and their orthologous introns into intron clusters. Such a dataset is ideal for studying the evolution of U12-type introns.

I retrieved the whole database and investigated conservation and change in intron status (U12, U2, and absence) in sets of genomes. These are 1) eutheria (*Homo sapiens, Mus musculus, Rattus norvegicus, Canis familiaris*), 2) vertebrates (with addition of *Monodelphis domestica, Gallus gallus, Fugu rubripes, Tetraodon nigroviridis, Danio rerio*), 3) chordates (with addition of *Ciona intestinalis*), 4) *H. sapiens* and *D. melanogaster*, 5) *H. sapiens* and *A. thaliana*. I found that U12-type introns are remarkably conserved among vertebrates. Even though only 18 U12-type introns were detected in the *D. melanogaster* genome, only of them actually arose from activating cryptic sites of an existent U2-type intron. In general, loss of U12-type introns is more common than conversion to U2. U12 to U2 conversion occurs more frequently among introns of the GT-AG subtype. I also found support for natural U12-type introns with non-canonical terminal dinucleotides (CT-AC, GG-AG, and GA-AG) that have not been reported previously.

Results and Discussion

To study the evolution of U12-type introns, we investigated conservation and changes of intron state among U12-type introns and their orthologS identified in 20 sequenced eukaryotic genomes (Alioto, 2006) - *Homo sapiens, Pan troglodytes, Macaca mulatta, Mus musculus, Rattus norvegicus, Canis familiaris, Bos taurus, Monodelphis domestica, Gallus gallus, Xenopus laevis, Fugu rubripes, Tetraodon nigroviridis, Danio*

rerio, Ciona intestinalis, Apis mellifera, Drosophila melanogaster, Anopheles gambiae, Caenorhabditis elegans, Saccharomyces cerevisiae, Arabidopsis thaliana. We utilized orthologous U12-type intron clusters and related data available in the U12DB. Members of an orthologous intron cluster are those of orthologous genes that are flanked by orthologous exons (alignable exonic sequences). The status of each intron in each species was designated as U12, U2, ambiguous or absent. Changes in intron state between orthologous introns were noted and manually curated.

The U12-type introns in the U12DB were not directly scanned from 20 individual genomes. Instead, U12DB uses four reference genomes (*H. sapiens*, *C. intestinalis*, *A. thaliana*, and *D. melanogaster*). Their source of introns includes extraction from complete genome annotation (excluding *Ciona*) and mapping ESTs to genomic sequences. The resultant U12-type introns from their classification joining those previously identified from the human genome (Levine and Durbin, 2001) served as seed introns to form orthologous intron clusters. Exonic sequences flanking a seed intron (100 nucleotides on each side) were concatenated and then used to seek the orthologous region in the genomic sequence of the orthologous gene. The orthologous intron, if present, was then scored and determined intron type. Ensembl ortholog prediction and Inparanoid were sources of ortholog information underlying the U12DB. Consequences of this approach are: 1) for non-reference genomes, the number of reported U12-type introns is expected to be underestimated since the genome is not fully searched. 2) The results are subject to uncertainties or errors in ortholog identifiction. Particularly, for those genomes that have less than ideal genome assembly, such as *Ciona*, orthology information can be missing or incorrect due to the lack of gene annotation. Other problems include that

ancient paralogs resulting from gene duplication before speciation and recent paralogs, if present, were not distinguished in the U12DB, and that the number of orthologous introns appears to be underestimated for extremely distant species..

An overview of the entire dataset is provided by a pairwise comparison to the human genome as a reference (Table 2-1). Humans were used as the reference species because the greatest number of U12-type introns has been identified in the human genome. In Table 1 the first column shows the number of U12-type introns that were included in the analysis from each genome. Not surprisingly, both the number of intron pairs compared and the number of U12-type introns conserved between the human and the compared genomes decrease with the phylogenetic distance. Between the human and each of the other vertebrate genomes, U12-type introns are conserved for at least 90% of the pair of introns compared. Overall, there are a lot more cases of U12-absent (humans-compared species) than U12-U2. Nonetheless, these initial results also show that evidently, quality of genome assembly and the abundance of mRNA and EST data affect the number of U12-type introns identified in each genome. For instance, the number of U12-type introns identified in humans is outstandingly high even among primates or mammals. This most likely reflects the unrivalled amount of expression data, the quality of the genome assembly and quality of gene annotation. Furthermore, the mouse genome is better studied than chimp and macaca, and thus, the number of compared and conserved intron pairs in the human-mouse comparison is higher than in the human-chimp or human-macaca comparisons. With this automated result serving as a broader picture, we set out to conduct analyses that are more refined.

Table 2-1: Intron state and counts of 621 human U12-type intron orthologs in each compared species

Compared species	comparisons	absent	U2	ambiguous	U12	Proportion of conserved U12 to comparisons	# of U12 in compared species*	Proportion of conserved U12 to # of U12 in compared species
Pan troglodytes	486	6	0	7	473	0.973	474	0.998
Macaca mulatta	452	14	1	7	430	0.951	434	0.991
Mus musculus	501	11	0	7	483	0.964	490	0.986
Rattus norvegicus	483	15	1	5	462	0.957	469	0.985
Canis familiaris	516	7	0	11	498	0.965	503	0.990
Bos taurus	469	9	2	9	449	0.957	455	0.987
Monodelphis domestica	440	12	1	5	422	0.959	428	0.986
Gallus gallus	391	13	0	9	369	0.944	373	0.989
Fugu rubripes	333	21	6	5	301	0.904	308	0.977
Danio rerio	301	11	6	10	274	0.910	282	0.972
Ciona intestinalis	154	83	29	1	41	0.266	113	0.363
Apis mellifera	161	122	6	2	31	0.193	34	0.912
Drosophila melanogaster	198	169	18	5	6	0.030	16	0.375
Caenorhabditis elegans	96	80	15	1	0	0.000	0	0.000
Arabidopsis thaliana	2	0	0	0	2	1.000	216	0.009

*number of U12-type introns from the compared species included in the analysis. Results are automated inference from the intron state matrix before manual inspection.

Evolution of U12-type introns in eukaryotic taxa

Four-eutherium comparison: only one intron loss

To gain insight into the evolution of U12-type introns through closely related species and to reduce the effect of data quality, we analyzed four eutherian genomes that are of relatively good quality. According to the initial automated results, there are 442 clusters where at least one of human, mouse, rat, and dog orthologous introns has a "U12" status (Table 2-2). In 407 of these 442 clusters, introns of all four species are of "U12" state. Of the remaining 35 clusters, 17 have "U12" and "absent", 17 have "U12" and "ambiguous", and a single cluster contains one "U12", one "U2", and two "ambiguous" introns. After manual inspection of the 17 intron clusters having putative U12-type intron losses ("U12" to "absent"), only one case appears to be a true intron loss by deletion, the mouse elongation factor 1 (*Elof1*) coding gene. The other "absent" cases (24 introns in 17 clusters) are artifacts. In 14 cases, non-orthologs were compared. In the other three cases, non-cognate transcripts were compared, or genes that are present but not or partially annotated in the comparison genome.

Table 2-2: Intron state composition of eutherian orthologous U12-type intron clusters

H. sapiens	*M. musculus*	*R. norvegicus*	*C. familiaris*	# of clusters
U12	Lost[1]	U12	U12	1
U12	ambi	U2[2]	ambi	1
U12	U12	U12	U12	423
U12	U12	U12	ambi	4
U12	U12	ambi	U12	2
U12	ambi	U12	U12	3
U12	ambi	ambi	U12	1
U12	ambi	ambi	ambi	1
ambi	U12	U12	U12	2
ambi	U12	U12	ambi	1
ambi	U12	ambi	ambi	1
ambi	ambi	U12	ambi	1
ambi	ambi	ambi	U12	1

Each orthologous intron cluster represents a unique intron position.
Ambi: ambiguous (denoted as U12/U2 in the U12DB)
[1] The mouse orthologous intron of human *ELOF1* (Elongation Factor 1) gene intron 2 is lost.
[2] Intron 2 of rat *Syt16* (synaptotagmin XIV-like) gene.

Eight-vertebrate comparison: eight losses and five type conversions in fish

We then added *M. domestica, G. gallus, D. rerio, F. rubrip*, and *T. nigroviridis* to the four eutheria to compare U12-type introns within vertebrates. In this analysis, the two puffer fishes (*F. rubrip* and *T. nigroviridis*) were represented as a puffer lineage. As more species were added to the analysis, the amount of data available decreased. There are 180 orthologous intron clusters where orthologous genes from four eutheria, opossums, chickens, zebrafish, and puffers are present, and there is at least one "U12" intron. This number is considerably lower than the 442 clusters of the eutherian analysis. U12-type introns are conserved in all eight lineages for 144 (80%) of the 180 intron clusters. Manual inspection of the remaining 36 clusters indicates an even higher level of

conservation. As in the eutherian comparison, several apparent cases of change were found to be due to ambiguous orthology or incompleteness in genome assemblies and/or gene annotation. For all 180 clusters except two, U12-type introns are conserved among eutheria, opossums, and chickens (column 1, 2, and 3 in Table 2-3). Only one cluster involves a U12-U2 conversion (the *CSNK1G1* gene, case 13 in Table 2-4). This is the only cluster among the 180 where U12-type introns are present in eutheria only. For the remaining 179 clusters, U12-type introns are present in at least one eutherium and one fish (column 5 in Table 3-3). We were able to confirm intron losses in eight clusters (two clusters are within the same gene), which were the results of five intron deletions, two presumptive gene retropositions (one in intronless *DCP2* gene and two in intronless *VAC14* gene). We also observed U12-U2 changes in five clusters (two in *AOX1* and one in each of *DCTN6, CSNK1G1*, and *EXOSC2* genes). The remaining five clusters have ambiguous introns from either fish lineage.

When this analysis was extended to include the *Ciona* genome, the number of clusters available for analysis shrank to 77 from 180. The *Ciona* intron state in these 77 clusters are 25 “U12”, one “ambiguous”, 17 “U2”, and 34 “absent”. This degree of conservation (32.5% of 77) is considerably lower than the near 90% among vertebrates. This low conservation is likely to be exaggerated by under-detection of tunicate U12-type introns. As mentioned earlier, the relatively poor quality of the draft genome assembly (Dehal *et al.*, 2002) and less abundant EST and/or cDNA data prevent some genes from being annotated and U12-introns from being detected. This can be seen in the pairwise comparison where 198 pairs were compared between human and fruitfly genomes and only 154 between human and *Ciona*. Nonetheless, the low number of genes compared

and the high proportion of absence cases agree with the findings that the *Ciona* genome has undergone excessive loss of ancestral genes (Hughes and Friedman, 2005) and introns (Raible *et al.*, 2005) that were present in the common ancestor of vertebrates and *Ciona*.

Table 2-3: Intron state composition of vertebrate orthologous U12-type intron clusters

eutheria	***M.domestica***	***G.gallus***	***D.rerio***	**puffers**	**# of clusters**
U12	U12	U12	U12	U12	156
U12	U12	U12	U12	Lost	6[1-6]
U12	U12	U12	U12	U2	1[7]
U12	U12	U12	Lost	U12	2[8,9]
U12	U12	U12	U2	U12	3[10-12]
U12	U12	U12	ambi	U12	4
U12	U12	U12	U12	ambi	1
U12, ambi	U2	U2	U2	U2	1[13]
U12, ambi	ambi	ambi	U12	U12	1
U12, ambi	U12	U12	U12	U12	5

Eutheria include *H. sapiens, M. musculus, R. norvegicus, and C. familiaris;* puffers include *F. rubripes* and *T. nigroviridis*. Footnotes (superscripted numbers) refer to Table 2-4.

Table 2-4: Details of cases of losses and type conversion listed in Table 2-3

	change	**species**	**gene**	**intron number**	**termini**	**Ensembl ID**
1	Lost	*T. nigroviridis*	*DCP2* (mRNA decapping enzyme 2)	2	AT-AC	GSTENG00007196001 (intronless)
2	Lost	*T. nigroviridis*	*C1orf164* (chromosome 1 open reading frame 164)	9	GT-AG	GSTENG00015184001
3	Lost	*T. nigroviridis*	*XAB1* (XPA-binding protein 1)	3	GT-AG	GSTENG00003557001
4	Lost	*T. nigroviridis*	*C21orf33* (chromosome 21 open reading frame 33)	6	AT-AC	GSTENG00027938001
5	Lost	*T. nigroviridis*	*VAC14*(Vac14 homolog)	8	GT-AG	GSTENG00034358001 (intronless)
6	Lost	*T. nigroviridis*	*VAC14*(Vac14 homolog)	16	GT-AG	GSTENG00034358001 (intronless)
7	U12 -> U2	*T. nigroviridis*	*EXOSC2* (Exosome complex exonuclease RRP4 (Ribosomal RNA- processing protein 4))	5	GT-AG	GSTENG00017201001
8	Lost	*D. rerio*	*VPS16* (vacuolar protein sorting 16)	9	GT-AG	ENSDARG00000033588
9	Lost	*D. rerio*	*VPS16* (vacuolar protein sorting 16)	13	GT-AG	ENSDARG00000033588
10	U12 -> U2	*D. rerio*	*DCTN6* (Dynactin subunit 6)	3	GT-AG	ENSDARG00000046084
11	U12 -> U2	*D. rerio*	*AOX1*(Aldehyde oxidase 1)	2	GT-AG	ENSDARG00000020054
12	U12 -> U2	*D. rerio*	*AOX1*(Aldehyde oxidase 1)	3	GT-AG	ENSDARG00000020054
13	U12 -> U2	*D. rerio*	*CSNK1G1*(casein kinase 1, gamma 1)	7	GT-AG	ENSDARG00000033872

*for loss cases, intron number refers to that in the human orthologs

Comparison between Homo sapiens and Drosophila melanogaster

The extremely limited conservation of U12-type introns between humans and fruitflies (6 of 198 comparisons in Table 2-1) relative to that among vertebrates is not surprising given the phylogenetic distance and the small number of U12-type introns identified in the fruitfly genome. It is, however, somewhat surprising that only six of the 18 fruitfly U12-type introns are conserved in humans. These 18 introns include two that have been reported earlier (Schneider *et al.*, 2004) but are missing from U12DB.

We manually examined those cases where fruitfly U12-containing genes are not clustered with human genes in ensembl prediction. The protein alignment between the fruitfly sequence and the human sequence with the highest BLASTP score was used to compare intron position. Table 2-5 lists adjusted status of the 18 *D. melanogaster* U12-type introns. Contrary to the numbers in the U12DB and Table 2-1 we were able to identify orthologous U12-type introns in the human genome for 15 of the 18 introns. The remaining three cases are one U2 to U12 conversion (twintron) in early diptera (*URP*), one intron loss in early vertebrates (*SF3A1*; see Figure 2-4) and one presumptive gene retroposition in early mammals (*CCDC16*; see Figure 2-5). The adjusted figures for human-fruitfly comparison are therefore as follows. In the 198 comparisons where humans have U12-type introns, fruitflies have 160 absent, 18 U2, 5 ambiguous, and 15 U12.

Unlike in vertebrates where a lot more U12-type introns were detected and they are predominantly GT-AG subtype, there are only nine (out of 18) GT-AG U12-type

introns in fruitflies with the other eight being AT-AC and one GC-AG. This 1:1 ratio is clearly different from that 2:1 (GT-AG vs. AT-AC) in the human genome, where the most (208) AT-AC U12-type introns were identified. This seems to suggest that AT-AC U12-type introns are more resistant to loss. All the 15 human, and other mammalian, orthologous introns have the same terminal dinucleotides with the fruitfly orthologs except the one with GC-AG, which is a nucleotide substitution that occurred only in fruitflies (see below). Fish orthologs have an additional subtype switch from AT-AC to GT-AG (in *Syntaxin-6* gene). We also examined orthologs of *D. melanogaster* U12-type introns in 12 *Drosophila* and one mosquito genomes whenever data are available. Interestingly, while the orthologs of the U12 intron in *D.melanogaster ZDHHC8* are U12-type in most vertebrates, they are U2-type in *D. ananassae* and *Anopheles gambiae*.

Table 2-5: Conservation of *Drosophila melanogaster* U12-type introns in humans

	Annotation symbol	***Termini***	***Gene symbol (name) in fruitflies and/or humans***	***Intron***	***of***	***Fruitflies***	***Humans***
1	CG6323	GT-AG	*Tsp97E* (Tetraspanin 97E); *TSPAN13* (tetraspanin 13)	2	4	U12	U12
2	CG8408	GT-AG	*TMEM41B* (transmembrane protein 41B)	3	3	U12	U12
3	CG17912	GT-AG	*ZNF207* (Zinc finger protein 207)	1	3	U12	U12
4	CG32705	GT-AG	*ZDHHC8* (zinc finger, DHHC-type containing 8)	4	11	U12	U12
5	CG33108	GT-AG	*C19orf54* (chromosome 19 open reading frame 54)	2	3	U12	U12
6	CG4894	GT-AG	*Ca-α1D* (Ca2+-channel protein α1 subunit D); *CACNA1D* (calcium channel, voltage-dependent, L type, alpha 1D subunit)	3	31	U12	U12
7	CG7892	GT-AG	*nmo* (nemo); *NLK* (nemo-like kinase)	6	10	U12	U12
8	CG15735	GT-AG	LSM12	1	4	U12	U12
9	CG3294	GT-AG	*ZRSR2* (zinc finger (CCCH type), RNA-binding motif and serine/arginine rich 2), *URP*; U2AF small subunit related protein ;	3	4	Twintron (alternative U2 and U12)	U2 only
10	CG16941	GC-AG	*SF3A1* (splicing factor 3a, subunit 1, 120kDa)	1	6	U12	absent
11	CG11839	AT-AC	*CCDC16* (coiled-coil domain containing 16); ZNF830 (zinc finger protein 830)	1	1	U12	absent
12	CG11328	AT-AC	*Nhe3*; *SLC9A7* (solute carrier family 9 (sodium/hydrogen exchanger), member 7)	5	12	U12	U12
13	CG18177	AT-AC	FLJ14154 hypothetical protein	4	4	U12	U12
14	CG7736	AT-AC	*Syx6* (Syntaxin 6); *STX6* (syntaxin 6)	1	2	U12	U12
15	CG17228	AT-AC	*pros* (prospero); *PROX1* (prospero homeobox 1)	2	4	Twintron (alternative U12 and U2)	U12 only
16	CG15081	AT-AC	*l(2)03709* (lethal (2) 03709); *PHB2* (prohibitin 2)	3	5	U12	U12
17	CG3427	AT-AC	*Epac*; *RAPGEF3* (Rap guanine nucleotide exchange factor (GEF)3)	11	16	U12	U12
18	CG11984	AT-AC	*KCMF1* (potassium channel modulatory factor 1)	3	6	U12	U12

Comparison between Homo sapiens and Arabidopsis thaliana

Even though more than 200 U12-type introns were identified in *A. thaliana*, only two of them are reporte by U12DB to have orthologs in humans. These two introns are in genes encoding exosome component 5 (*EXOSC5*; Exosome complex exonuclease *RRP46*; Ribosomal RNA- processing protein 46) and BRCA1/BRCA2-containing complex, subunit 3 (*BRCC3*). Twenty five *A. thaliana* U12-type introns are absent in the human orthologous genes, and another 28 have orthologous genes but not introns in non-human animals. If these numbers are correct, the number of *A. thaliana* U12-containing genes that have human orthologous genes is underrepresented compared to the number of *A. thaliana* genes that have human orthologous genes (p-value 7e-17), according to a hypergeometric distribution test we conducted using the clustering data from InParanoid (a database of Eukaryotic Ortholog Groups). However, my earlier analysis indicated 20 U12-type introns shared between humans and Arabidopsis and a recent analysis by Basu *et al.* (2008) reached a similar conclusion. This is therefore likely to be a more extreme case of what was observed for *Drosophila*; U12DB fails to track orthologs over greater evolutionary distances.

Summary of comparisons within various clades

The above analyses on various subsets of genomes were based on relatively conservative datasets; many data were not considered. We meticulously documented cases of changes with details. In summary, for 423 (95.7%) of the 442 clusters (unique intron positions) where four eutherian orthologous genes are present, introns are

classified as U12-type in all four eutheria. There is only one unambiguous case of U12 intron deletion (*Elof1* in the mouse genome) and one potential U12-U2 transition (*synaptotagmin XIV* in rats). The remaining 17 clusters contain ambiguous introns, indicating potential conversion. Among the eight vertebrates, in 156 (86.7%) of the 180 clusters U12-type introns are conserved in all eight lineages. There are intron losses in eight clusters and U12-U2 conversion in five clusters. Eleven clusters contain ambiguous introns. No gain of U12-intron was observed. These results indicate that U12-type introns have been tremendously stable in vertebrates during the last 550 million years after the divergence of *Ciona*. There were only a few losses and U12-U2 conversions. Conservation between *Ciona* and vertebrates is certainly underestimated mainly due to the relatively poor quality of the genome draft, which leads to underdetection of U12-type introns and difficulties in ortholog prediction. The conservation of U12-type introns between humans and fruitflies is very low (7.5% of 200) since there are only 18 U12-type introns in *D. melanogaster*, 15 of which have orthologs in humans, one is specific to diptera and two are lost in humans (mammals).

Massive loss of U12-type introns in invertebrates

Depending on availability of data and phylogenetic distance among genomes examined, the number of intron clusters in my analysis varied from 77 to 442, which accounts for at most two thirds of the near 700 metazoan U12-intron clusters (unique intron positions) in the U12DB. Since only 47 intron clusters contains non-chordate U12-type introns, obviously, the modern-day U12-type introns identified in extant animals are

mostly present in vertebrates. To learn how these orthologous vertebrate U12-type introns (those that are present in the human and at least one non-primate genomes) relate to other lineages, we investigated their conservation pattern from a broader and more inclusive perspective.

Of the 267 orthologous intron clusters where orthology between vertebrate and non-chordate genes is available, the majority (194) are cases where a vertebrate U12-type intron corresponds to "absence" state in non-chordate. Although the data do not provide direct evidence addressing whether these are gains in vertebrates or losses in non-chordates, we feel that they are more likely to be the latter case, based on the following observations. First, our four-eutherium and eight-vertebrate analyses did not uncover a single case of U12-type intron insertion, suggesting that insertion of U12-type introns is rare in vertebrates. Second, as mentioned above, the U12-type splicing system independently went extinct in some organisms (such as *C. elegans*) via intron deletion or conversion to U2-type. The intermingling of "U12-free" genomes with "U12-bearing" genomes indicates that that loss rate of U12-type introns is greater than gain rate. Third, by studying 30 gene loci from the marine annelid (*Platynereis dumerilii*), *C. elegans*, insects, *Ciona*, and humans, Raible *et al.* (2005) reported that genes in Urbilateria, the common ancestor of worms, insects, tunicates, and humans, are "vertebrate-type" (intron-rich), and that the disparity in intron abundance between invertebrates and vertebrates is mainly due to intron loss in invertebrates. Fourth, with apparent lower intron abundance and density than other higher eukaryotes, the *D. melanogaster* genome has been appears to have undergone extensive intron loss (Nguyen *et al.*, 2005; Raible *et al.*, 2005; Roy and Guilbert, 2005; Carmel *et al.*, 2007). A study using maximum likelihood approach

infers that the ratio of intron loss rate over gain rate are is 1.25/0.01 =~ 125 in *D. melanogaster*, 1.04/0.02 =~ 50 in *C. elegans*, 0.77/0.04 =~ 20 in *Ciona* , and 0.35/0.21 =~ 2 in humans (Carmel *et al.*, 2007).

The loss rate of U2-type introns in various species agrees with the prediction from population genetics perspective. Lynch and Richardson (2002) argued that harboring introns impose deleterious effect to the organisms because errors in splicing can lead to failure in producing functional proteins. The larger the effective population size of a species, the greater the selection pressure to remove introns. This explains the intron-rich vertebrates and intron-poor invertebrates and microbes. The loss rate may be even higher for U12-type introns because hosting U12-type introns can be more deleterious than U2-type introns. The primary argument is that the requirement for extended recognition motif makes U12-type more susceptible to mutations. Other factors that have been proposed to have negative impact include the slower processing of U12-type introns (Patel *et al.*, 2002) and higher splicing error rates (Levin and Durbin, 2001; Hasting *et al.*, 2005). Therefore, the selection pressure against U12-type introns is greater than U2-type introns, which resulted in predominant presence of U2-type introns. The disparity in number of U12- and U2-type introns can be greater in big populations than in small populations. While U12-type introns were steadily being lost in nematodes and insects, the loss process has been slowed down by the radiation of vertebrates. As new species arose, the small population size made the purging of U12-type intron less efficient, and thus led to higher retention rates.

Nearly half of U12-containing genes are vertebrate-specific

Of the 549 intron clusters where U12-type introns are present in humans and at least one non-primate vertebrate, invertebrate orthologous genes are absent in 254 of them. This indicates that not only that the majority of metazoan U12-type introns are present in vertebrates, but also that their harboring genes are often vertebrate specific. While some of the 254 intron clusters might be attributed to extensive gene losses in invertebrates (Kortschak et al., 2003; Hughes and Friedman, 2005), some of these vertebrate-specific genes might have been resulted from the large scale segmental duplication or whole genome duplication (WGD) event(s) that was/were proposed to take place in early vertebrates (McLysaght *et al.*, 2002; Dehal and Boore, 2005). The duplicated U12-containing genes provided material for "new" genes through subfunctionalization and/or neofunctionalization, and meanwhile, proliferated U12-type introns in vertebrate genomes. This is best exemplified by the two gene families that encode the alpha subunit (the pore-forming unit) of voltage-gated sodium and calcium channels, *SCN* and *CACN1* respectively. Our discussion of sodium and calcium channel genes is restricted to the one encoding the alpha subunit.

Mammals have 11 sodium (*SCN1A* through *SCN11A*) and 10 calcium (*CACN1A* through *CACN1I*, and *CACN1S*) channel alpha subunit genes while fruitflies have two and three respectively (Peixoto et el., 1996; Plummer and Heisler, 1999; Anderson and Greenberg, 2001). Most mammalian *SCN* and *CACN1* genes have two U12-type introns (one AT-AC and one GT-AG subtype); some *SCN* genes contain a rare AT-AC U2-type intron (Dietrich et al., 1997; Wu and Krainer, 1997; Wu and Krainer, 1999). Fruitflies

preserve only the GT-AG U12-type intron in one of the calcium channel gene. That the human and mouse *SCN* genes are located in four paralogous chromosomal regions, which also harbor the *HOX* genes, provides evidence that the large-scale segmental duplication or WGD and subsequent tandem duplications lead to the expansion of *SCN* genes in mammals (Plummer and Heisler, 1999). Similarities in protein sequence, protein structure, and functional characteristics have prompted the proposal that *SCN* and *CACNL* genes are both derived from potassium channel genes (Catterall 1988; Anderson and Greenberg, 2001; Yu et al., 2005). This is further supported by the shared position of the AT-AC U12-type intron, though not the GT-AG one, in a number of *SCN* and *CACN* genes (Wu and Krainer, 1999).

More deletion of U12-type intron than U12-U2 conversion in metazoa

A unique aspect of evolution of U12-type introns is the conversion from U12- to U2-type, which was first documented and proposed by Burge *et al.*, (1998). Both intron deletion and U12 to U2 conversion reduce the abundance of U12-type introns in a genome. Thus, an interesting question is how these two mechanisms contribute to the low abundance or extinction of U12-type introns in invertebrates. Assuming that there is no parallel intron gain and that type conversion can only be from U12- to U2-type introns (Burge *et al.*, 1998), those invertebrate U2-type introns orthologous to the vertebrate U12-type introns represent such U12 to U2 conversions. Even though the two assumptions are unrealistic (we have evidence of *de novo* U12-type intron creation by U2 to U12 conversion; see below), parallel insertion and U2 to U12 conversion are so rare

that the impact on the overall pattern is limited. The ratios of absence vs. U2-type introns are 2:1, 9:1, and 5:1 for *Ciona*, fruitflies, and nematodes (Table 2-1 with adjustment for fruitflies) respectively. This suggests that deletion contributes more than U12-U2 conversion to the evolution of U12-type introns in metazoa although the numbers of cases observed do not directly reflect the rates of two processes (some U12-turned-U2-type introns might have been lost and are not counted.) This might be because U12 to U2 conversion requires multiple fortuitous mutations or circumstances. It has been shown that with the presence of cryptic splice sites in the vicinity, nucleotide substitutions at positions +4 to +7, particularly a C to G change at +5, of a U12-type 5' splice site activate these cryptic sites, which lead to splice product through U2-type pathway, or leave the intron not removed (Kolossova and Padgett 1997; Dietrich *et al.*, 1997). Such mutations are more likely to result in non-functional proteins, and in turn, the mutant alleles are quickly removed from the population. Moreover, if the initial mutation does not disrupt the reading frame, the intron is likely to become an ambiguous intron, which was processed inefficiently by either spliceosome. Hence, subsequent mutations might be required to improve the splice site of the U12-turned U2-type introns. During the course of "perfection", the suboptimal U2-type introns slowly drift in the population. On the other hand, reverse transcription-mediated intron deletion is less likely to disrupt the reading frame of the mature mRNAs, and thus has less negative effect than nucleotide mutations in splice sites. In turn, the mutant alleles are more likely to be retained or go fixation more quickly.

Gene retrotransposition and intron loss

In vertebrates where U12-type introns are tremendously conserved, gene retroposition appear to account for near one half of the lost U12-type introns. Three of the eight mammalian U12-introns that are lost in fish are due to putative gene retroposition. Interestingly, the vertebrate ortholog of the *D. melanogaster* U12-containing gene (*CCDC16*) have undergone two independent retropositions leading to a functional allele, one in mammals and one in fish. On the other hand, 13 of the 160 *D. melanogaster* orthologs of the vertebrate U12-containing genes are intronless.

U12-U2 conversion occurs more frequently in GT-AG than AT-AC subtype

Of the near one thousand U12-type intron clusters in the U12DB, 87 clusters have both U12- and U2-type introns. Presumably, the 87 non-redundant U12-type introns have undergone type-conversion at least once during the course of evolution. Eight of them also have both subtypes. Of the remaining 79 clusters of U12-type introns 74 are GT-AG and only five AT-AC subtypes. The ratio of GT-AG to AT-AC subtypes in various genomes catalogued in the U12DB (Alioto, 2007) ranges between 3:1 and 1:1. Conversion of AT-AC subtype is clearly underrepresented among those undergoing type conversion ($\chi 2$ value 19.33; $P<0.00001$). Moreover, all 87 U12-turned-U2-type introns have GT-AG termini. This seems to suggest that GT-AG U12-type introns are more likely, if not exclusively, to be converted to U2-type. This is probably because a guanosine at the first position of an intron (denoted as G1) is better base-pairing with U1 snRNAs, an interaction that is crucial in the recognition of U2-type introns, which

commits the intron into the major-type splicing pathway. At least nine (five plus four of the eight) clusters might have undergone both "subtype switching" (from AT-AC to GT-AG) and "type conversion" (U12 to U2) - a U12-U2 conversion pathway Burger *et al.* (1998) proposed. Another 36 clusters have U12-introns of both subtypes.

To emphasize that terminal dinucleotides of an intron do not determine which splicing pathway it undergoes, Dietrich *et al.* (1997) not only demonstrated that introns with GT-AG termini can be processed by the minor-type spliceosome, but also that splicing of intron 21, which has AT-AC termini, of the gene encoding human SCN4A (voltage-gated sodium channel α subunit) requires the major-type spliceosome. While the majority of U12-type introns turned out to be GT-AG subtype, AT-AC U2-type introns remain very rare, if not exceptional. A large-scale splice site analysis on five genomes identified only 15, six, and three such introns in the human, mouse, and fruitfly genomes respectively (Sheth *et al.*, 2006), and most of them are ancient paralogs of the one in the human *SCN4A* gene. This AT-AC U2-type intron is present in most vertebrates as well as several other homologous sodium channel α subunit genes, such as *SCN5A*, *SCN8A*, and *SCN10A*, which are believed to arise from gene duplication (see below). Furthermore, none of the U12-turned-U2-type introns in the aforementioned 87 clusters has AT-AC termini.

A number of observations provide further support for the view that U2-type introns have a stronger bias against AT-AC than do U12-type introns. First, the study reporting the evidence for a non-Waston-Crick interaction between the first and last nucleotides of an intron showed that a G1 to A1 mutation (G to A mutation at position +1) considerably reduces the cleavage at the 5' end (the first step of splicing). The double

mutant with coupling mutations at the two terminal nucleotides of an intron (GT-AG to AT-AC) only restores 10% of the successful rate in the wild type (Parker and Siliciano, 1993). The nearly exclusive G1 in U2-type introns is mirrored by the nearly invariant 5' end of U1 snRNAs in many eukaryotes, including several fungi (Schwartz *et al.*, 2008).

A rare pathway from AT-AC U12 to AT-AC U2-type

It is likely that AT-AC II introns are the outcome of a much rarer U12 to U2 conversion pathway. Those AT-AC II introns in the sodium channel genes all have a G at position +5, a feature that has been shown to be important for recognition of the 5' splice site by the U2-spliceosome (Dietrich and Padgett, 1997). A possible scenario is that a C to G mutation at position +5 turns an AT-AC U12-type intron into a suboptimal U2-type. Subsequent mutations, particularly to an A at position +4 or a T at position +6, would then improve the recognition of this site. The resulting AT-AC II intron may then be "trapped" because, unlike in the U12-type splicing, the U2-type splicing allows less flexibility in the non-Watson-Crick interaction between the first and the last nucleotides. It would require simultaneous mutations to switch an AT-AC II intron to either GT-AG U2-type or AT-AC U12-type. This difficulty leaves it to remain as a suboptimal U2-type intron.

U12-type splicing signals and U12-U2 conversion

It seems paradoxical that there is stronger constraint on the first nucleotide of U2-type introns, while more flexibility (G1 or T1, or even C1; see below) is allowed for the minor-type introns and the 5' end of U11 snRNAs are less conserved among species than U1 snRNAs. It is, after all, the extended nearly invariant 5' splice site that led to the discovery of the minor-type splicing pathway. This in fact reflects the major difference between the two splicing pathways at the early stage of splicing. Whereas the nucleotides at positions -3 to +6 of U2-type introns can interact with U1 snRNA, only +4 to +7 of U12-introns are recognized by U11 snRNAs (Figure 2-1; Hall and Padgett, 1996; Patel and Steitz, 2003). While U1 and U2 snRNPs assemble on a U2-type intron independently, U11 and U12 snRNPs form di-snRNPs and then interact with the 5' splice site and the branch point site simultaneously (Frilander and Stetz, 1999). This difference eventually facilitates the conversion of U12- to U2-type introns. A possible pathway for the type conversion might start with a mutation at the last nucleotide, which has been shown that it can be any nucleotide if the first nucleotide is an A (Dietrich *et al.*, 2001; 2005). When the last nucleotide is a G, the subsequent mutation from A to G at the first nucleotide switches the U12-type from AT-AC to GT-AG subtype. A mutation to G at position +5 or gradual and/or multiple mutations at position +4 to +7 followed by "improving" mutations then convert a U12- to a U2-type intron (Dietrich *et al.*, 1997). The conservation of nucleotides at +2 and +3 of U12-type introns might have been shaped by interaction with other factors. Indeed, during preparation of this manuscript, the recently published study demonstrated a direct contact between the nucleotide at +2

and the U11-48K protein (Turunen *et al.*, 2008). Unlike most protein components of the U11/U12 snRNP, this protein has no homolog among U1 or U2 snRNP proteins.

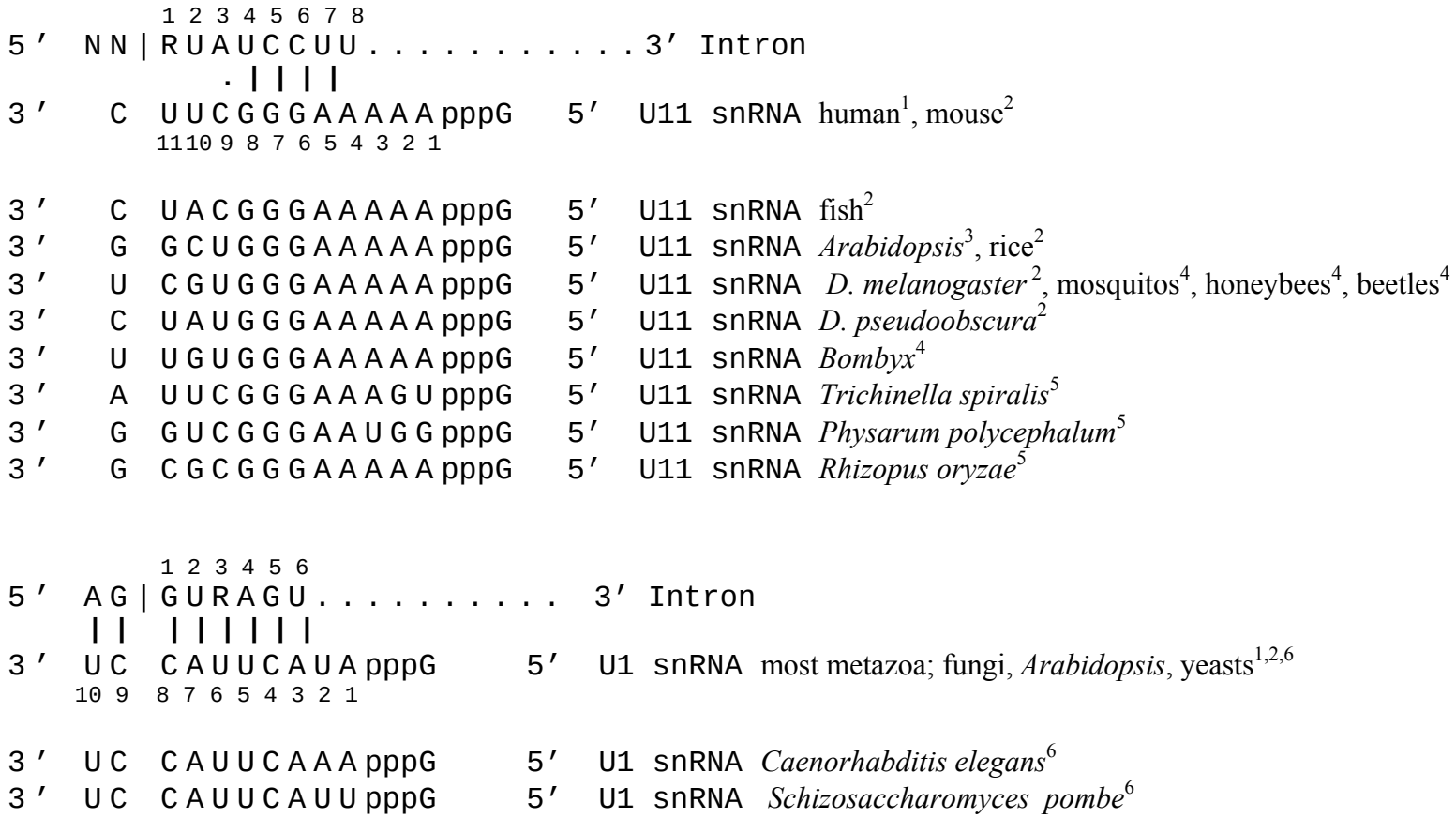

Figure 2-1:Base pairing of 5' splice sites and snRNAs in the U12-type (top) and U2-type (bottom) splicing

U12-type introns are known for their extended nearly invariant 5's splice sites. Only nucleotides at positions 4 to 8 are base-pairing with U11 snRNAs. The nucleotides immediate downstream of the interacting site in the U11 snRNA are variable while the counterpart in U1 snRNAs are nearly invariant in all eukaryotes.

Modified from Hall and Padgett, 1996, and Kolossova and Padgett, 1997. Sources of U11 and U1 snRNA sequences:[1] Hall and Padgett, 1996; [2] Rfam; [3]Schneider *et al.*, 2004; [4] Mount *et al.*, 2007; [5] Lopez *et al.*, 2008; [6] Schwartz *et al.*, 2008

Non-canonical U12-type introns

Given that the 5' end dinucleotides of U12-type introns seem to be under less constraint, we investigated non-canonical 5' dinucleotides present in U12-type introns. In the U12DB (Alioto, 2007), 19 U12-type introns have non- canonical donor sites (with canonical donor sites being GT, GC, or AT). Manual examination revealed that six are annotation errors (four have AT-AC termini and two have GT-AG). Six of the remaining 13 are unambiguous U12 non-canonical 5' dinucleotides. They are two CT-AC introns in the mouse and rat genes encoding NIK and IKK(beta) binding protein, three GG-AG ones in the macaca and cow *SLC12A4* genes and the rat *SLC25A30* gene, and one GA-AG intron in the zebrafish *actr10* gene. Naturally occurring U12-type introns with these three sets of termini (CT-AC, GG-AG, GA-AG) have never been reported before. In a mutational study of terminal dinucleotide in U12-type introns, all 16 possible combinations of first and last nucleotides were tested in the context of NU-AN. Indeed, the combination CU-AC yields correct splicing products, albeit with a lower efficiency than the wild type. Claiming that the U at the second and the A at the penultimate position of an intron (NU-AN; denoted as +2 at the 5' end and the -2 at the 3' end respectively) are highly conserved according to database search and mutant analyses, the authors also tested the functionality of 12 combinations (excluding G2) of nucleotides at these two positions in the context of AN-NC. The results suggested that the nucleotides at +2 and -2 do not interact in the same way as those at +1 and -1. Therefore, no evidence appears to challenge the functionality of GG-AG and GA-AG termini.

The discovery of these non-canonical U12-type introns has two important implications. First, greater variation has been observed in the 3' termini in natural human U12-type introns. This finding shows that the variation in the 5' termini is greater than previously thought, which confirms the results of that mutational study (Dietrich *et al.*, 2005). Second, it reveals the conundrum faced by investigation of non-canonical introns, especially at a large-scale and/or that of U2-type introns, whose 5'splice site have much lower specificity than that of U12-type introns. On the one hand, the algorithm used in predicting splice sites dictates what splice sites can/cannot be found. For instance, early prediction programs are tuned for assigning GT and AG as donor and acceptor sites. Consequently, AT-AC introns are often misannotated. Later gene prediction or cDNA mapping programs were improved but there is still a tendency to predict RT-AR introns (R stands for A or G). On the other hand, designing of algorithm is based on what is known, which can be biased. For instance, in intron studies, especially large-scale studies, there is normally some filtering based on terminal dinucleotides and/or intron length (Schwartz *et al.*, 2008), or only certain termini were examined (Levin and Durbin, 2001; Dietrich *et al.*, 2005). Such filtering is often necessary but it also somewhat skews our understanding of spliceosomal introns to what is known. Furthermore, when a non-canonical splice site is detected and there is slim or none cDNA support, issues might arise over whether it is legitimate, splicing error, or sequencing errors.

Illustrative examples of the evolution of U12-type intron

A U12-type intron in Drosophila CG3294 (URP) maybe autoregulated

Two splicing variants (CG3294-RA and CG3294-RB) are annotated for the D. melanogaster CG3294 locus based on a combination of EST data and sequence similarity to the human U2AF35-related protein (URP) sequence, which contains a divergent CCCH Zinc finger RNA-binding motif and RS (aRginine/Serine) domain. The HGNC gene symbol for this human *URP* gene is *ZRSR2*. In this paper, we denote the Drosophila and human *URP* as *dURP* and *hURP* respectively. CG3294-RA and CG3294-RB are annotated as having almost identical gene structures with five exons, the only difference being at the third intron, where alternative U2 and U12 5' splice sites are joined to a common 3' splice site. However, it is more likely that different 3' splice sites are also used, and the two alternative introns are staggered (Figure 2-2) (see below). Thus, CG3294-RA is the U2-isoform and CG3294-RB is the U12-isoform with the U12-type intron. The 5' end of the U12-type intron is 29 nucleotides downstream of its U2-type counterpart. This extended 29-nucleotide exonic sequence introduces a pre-mature stop codon, which would truncate the protein such that the RS domain in the C-terminal is missing if the processed mRNA is translated. The majority of the mature mRNA is predicted to be targeted for degradation by an mRNA surveillance system, Nonsense-mediated mRNA Decay (NMD). Such alternative splicing pattern has been observed in many SR protein genes (Rehwinkel *et al.*, 2007; Lareau *et al.*, 2007).

Characterized by their RS domain and RNA recognition motifs (RRMs), SR family and SR-related proteins are required in both constitutive and alternative splicing of

pre-mRNAs (reviewed in Blencowe, 2000). Because they interact with other proteins mainly through their C-terminal RS domains, truncated SR proteins that lack the C-terminal RS domain via alternative splicing are speculated to be involved in negative feedback loops to regulate their own expression (Screaton *et al.*, 1995; Sureau *et al.*, 2001). Later, several studies showed that this auto-regulation is achieved at the transcript level, a type of post-transcriptional expression regulation through alternative splicing coupled to NMD. In fact, a study (Rehwinkel *et al.*, 2007) where major components (UPF1, UPF2, UPF3, SMG1, SMG5, SMG6) of NMD machinery were depleted individually to identify targets of NMD genome-wide in drosophila, shows that transcripts of CG3294 were indeed upregulated for all the six components, indicating that CG3294-RB is subject to NMD. In each of the six individual knockdowns (*UPF1, UPF2, UPF3, SMG1, SMG5, SMG6*), 14.3%, 11.1%, 4.5%, 5.2%, 4.8%, and 9.2%, respectively, of detected mRNAs were upregulated, and 10.4%, 6%, 1.5%, 1.6%, 1.5%, and 3.9% respectively downregulated. 184 transcripts were at least 1.5-fold over-represented in at least 10 (of 12) profiles.

▼ (birth of) U12-type introns

▽ U2-type introns

▽ U2-type introns without EST/mRNA support

* Stop codon

== UnTranslated Region

x intron loss

Figure 2-2: Phylogenetic distribution of the GT-AG U12-type intron of *Drosophila* CG3294 (*URP*) gene and the protein alignment of the flanking regions in the orthologs.
The *URP* gene has one splice variant (orthologous to the isoform A) in most lineages (only humans and zebrafish are shown to represent vertebrates). in fruitflies and mosquitoes (both are diptera) it produces a distinct isoform (isoform B) resulted from removing a U12-type intron. Excision of the U12-type intron introduces a premature stop codon and leads to a 3' UnTranslatedRegion, which is shown in this figure to demonstrate that two alternative spliced introns are 29 nucleotides apart in the chromosome, as opposed to one amino acid in the protein sequence. The premature stop codon subjects the mature mRNA to be targeted by NMD surveillance system, or leads to a truncated version of the protein. Both dipteran U12-type introns (solid triangles) are supported by ESTs/mRNAs. Both dipteran U2-type introns (open triangles with dotted lines) have NO EST/mRNA support. *Drosophila* isoform A is likely to be a mistake in predicting 3' splice site of the U2-type intron. We proposed a "corrected" version as isoform A' in Figure 2-3. (Branch lengths are not drawn to scale.)

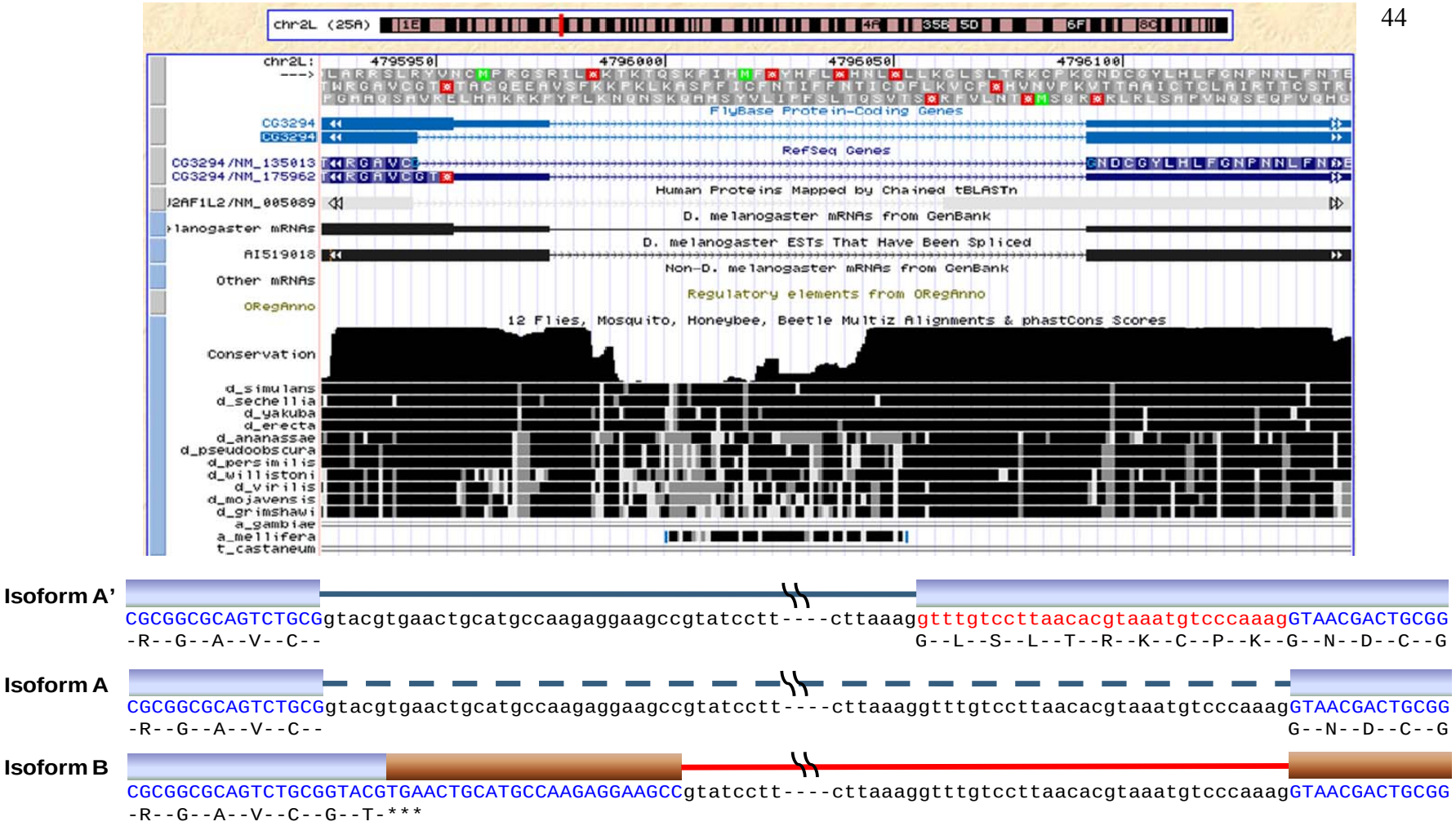

Figure 2-3: The genomic region of the GT-AG U12-type intron and its flanking exons in the *D. melanogaster* CG3294 gene

It is shown in a snapshot from the UCSC Genome Browser (Kent *et al.*, 2002) on *D. melanogaster* Apr. 2004 Assembly. Only the U12-type intron is supported by mRNA and EST data. Sequence conservation is detectable between the derived human protein sequence and that of the intronic 30 nucleotides upstream of the annotated 3' splice site. These 30 nucleotides make up 10 aa, which fills up the gap in the sequence alignment between the human and the annotated *D. melanogaster* protein sequences. Both vicinities of the 5' exon-intron and the 3' intron-exon junctions are highly conserved among 12 *Drosophila*. Nucleotide sequences and derived protein sequences surrounding the two junctions. The isoform A and B are as annotated in the flybase as as well as the RefSeq. The isoform A' is the corrected version we propose (see text). Light blue boxes represent coding exons and brown UTRs. Blue lines represent U2-type intron and red U12-type.

As described in details in the next section this U12-intron-mediated-NMD-autoregulation is present only in fruitflies and mosquitoes. Thus, it appears that the absence of an orthologous U12-type intron in hURP is a gain in *D. melanaogaster* by activating cryptic splice sites and thus a new isoform.

A new twintron in dipterans

As mentioned above, *D. melanaogaster URP* is annotated with two transcript variants and both are labeled as "reviewed" in the RefSeq database (NM_135013, NM_175962 for CG3294-RA and CG3294-RB respectively). While CG3294-RB (the U12-isoform with a pre-mature stop codon) is supported by two cDNA sequences, prediction of complete coding sequence of CG3294-RA (the U2-isoform) is based on similarity to the human protein sequence. Because the 3' end of the third intron is 30 nucleotides upstream (suggested by sequence similarity but not supported by ESTs/mRNAs) to that of the U12-type intron of RB, which is supported by ESTs/mRNAs, the U12-type intron's acceptor site is assigned to the U2-type intron. We find it likely that the U2-type intron (for which there is no published EST or cDNA support) actually lies 30 nucleotides upstream of this site. By adjusting the acceptor site the 10-amino-acid gap in the alignment of drosophila and human protein sequences is filled. We looked for evidence for presence of its protein product. A study mapping protein-interaction network of the *D. melanaogaster* proteome (Giot *et al.*, 2003) through a yeast two-hybrid system reports physical interaction between CG3294-PA (protein product of CG3294-RA) and 23 proteins, including SLU7 (a catalytically active

spliceosome splicing factor), RNA-binding protein 1 (RBP1), SAP90 (a component in the U4/U6/U5 tri-snRNP), and Mediator of RNA polymerase II transcription subunit 19 (MED19). *D. melanaogaster* RBP1 is another SR protein, orthologous to human 9G8 and SRp20. Its interaction with proteins related in splicing and transcription suggests that not only *dURP* codes for a functional protein but also its protein product is involved in pre-mRNA splicing.

To learn evolution of *dURP* U12-type intron, we looked for orthologous gene and the transcript (only the regions flanking the U12-type was considered) that is cognate to the *dURP* U12-isoform in other organisms. While the *URP* genes in12 *Drosophila* and two mosquitoes contain the orthologous U12-type intron and presumably they could produce express both the U2- and the U12-isoform, those in vertebrates, except mouse, seem to express only the U2-isoform (see Figure 2-3). Mouse *URP* expresses the U2-isoform and a second splicing variant that would be translated into truncated proteins, but this is via retention of an intron that is not related to the *dURP* U12-type intron. similarly, all non-dipetera invertebrates examined express only the U2-isoform. Thus, it is clear that the U2-isoform is the ancestral state and the U12-isoform (with the U12-type intron) was newly acquired in the common ancestor of diptera.

Furthermore, there must be an original, intact protein (product of the U2-isoform) and then a "truncated" version (product of the U12-isoform). The fact that dURP, by acquiring a new U12-intron, activated a novel isoform that allows it to be post-transcriptionally regulated by NMD mechanism is significant in several aspects. First, no other case of U12-type intron gain has been documented so far. Second, the dipteran U12-type intron within the URP gene demonstrates that U12-type introns can arise via

alternative splicing, as opposed to intron insertion. Second, the evolutionary trend of U12-type splicing system is believed to be loss of U12-type introns due to unsymmetrical stringency of cis-elements in the two types of intron sequences. It has been thought that U12-type introns can be converted to U2-type but probably not vice versa. Nonetheless, dipteran *URP* U12-type intron reveals that in genomes where U12-type introns are so scarce (fewer than 20 in diptera) new U12-type introns can still arise. Third, *Drosophila URP* gene demonstrates a "reverse" pattern of alternative splicing to that of the well studied *Drosophila prospero* gene (Scamborava *et al.*, 2004) where the alternatively spliced U2-type intron is embedded in the U12-type counter (a so-called twintron arrangement). In the case of *prospero* gene, the isoform with the U12-type intron is the ancestral state because orthologous U12-type introns are present in other species (humans for instance) while the U2-type is only present in *Drosophila*. This has been proposed to be one of the pathways whereby U12-type introns were lost in *Drosophila* (Mount *et al.*, 2007). The pattern that an ancestral U12-type intron alternatively splices with a new U2-type intron in prospero is opposite to that in *URP* that an ancestral U2-type intron alternatively splices with a new U12-type intron.

Interestingly, similar to the *dURP* U2-isoform, the mosquito U2-isoform is also predicted based on sequence similarity to *hURP* without EST/mRNA support, while the U12-isoform is not annotated even though it is supported by ESTs. Therefore, numbers of EST/mRNA sequences detected for the *Drosophila* and mosquito *URP* genes seem to suggest that the URP U12-isoform is more abundant than the U2-isoform in both diptera.

That AS-NMD was newly acquired by the *dURP* but not *hURP* gene is consistent with the finding that the majority of NMD targets in *Drosophila* are not orthologous to

those in humans (Rehwinkel *et al.*, 2007). Thus, even though NMD surveillance has been found in many eukaryotes, regulating 10 - 20 % of the transcript in addition to preventing the translation of mRNAs that contain non-sense mutation (Conti and Izaurralde, 2005), genes become target of NMD as a means of regulating expression independently in different species.

Evidence that URP functions in the splicing of U12 introns suggests a possible "U12AF."

U2AF (auxiliary factor) is a heterodimer consisting of a small and a large subunit, which are designated as U2AF35 (35kDa) and U2AF65 (65kDa) in mammals, and U2AF38 (38kDa) and U2AF50 (50kDa) in *Drosophila*. It is a splicing factor that is essential in the early stage of pre-mRNA splicing. U2AF35 has been demonstrated to play two roles. First, through the interaction between the two subunits, U2AF35 facilitates U2AF65's binding to the Poly Pyrimidine Tract (PPT), which in turn, recruits U2snRNP and helps stabilize its interaction with the branch point site of the intron. Second, U2AF35 recognizes and associates with the 3' splice site of an intron (reviewed in Blencowe 2000). U2AF35 Related Protein (*URP* or *ZRSR2*) is named after U2AF35 because it has the essential domains (CCCH type Zinc finger, RNA-binding motif and RS domain, after which the HGNC gene symbol was named) that make up U2AF35. However, two proteins have distinct lengths (438 aa and 240 aa) and low sequence identity (34%). The mouse *URP* gene was first documented for its genomic imprinting role (Yamaoka *et al.*, 1995). Later, Tronchere and colleagues (1997) reported that while human URP interacts with U2AF65 and SR proteins as U2AF35 does. The two proteins,

however, do not function interchangeably. Despite that *Drosophila* URP has not been functionally studied, as mentioned earlier, a genomewide yeast two-hybrid assay on *Drosophila* protein-protein interaction reported interactions between URP and SR proteins and snRNP proteins.

While U2AF35 has been consistently demonstrated to recognize 3' splice site, no counterpart in the minor spliceosome has been described. Shen and Green (2007) have excluded U2AF35's role in U12-type splicing with their experiment, which revealed that U2AF is not required in the U12 splicing pathway. This makes URP a perfect candidate for "U12AF35", the splicing factor that recognizes 3' splice site in the U12-type splicing. It is possible that URP acts as a U12AF without a large subunit, because U12 introns lack the pyrimidine tract to which the U2AF large subunit binds. Alternatively, another protein may play this role. The analogy (same protein structures but distinct lengths and low similarity in protein sequence) between URP and U2AF35 agrees with those demonstrated by other spliceosomal components, including snRNAs (e.g. U2 vs. U12) and snRNP proteins (U1-70K vs. U11/U12-35K), and the two splicing pathway per se. Two additional observations make the hypothesis that URP recognizes 3' splice site in the U12-type splicing even more attractive. First, URP is present in the human 18S U11/U12 di-snRNP (Will *et al.*, 2004). Second, URP gene is preserved in many eukaryotic genomes with minor introns, but is lost in *C. elegans*, an organism that has lost the U12-type splicing machinery and U12-type introns.

Thus, like those components that are shared between two spliceosomes, such as SF3b (a component in U2 snRNP and U11/U12 di-snRNP) and U5 snRNP, URP not only functions in the U2-type splicing but also in the U12-type.

An insect-specific U12-type intron in the SF3A1 coding gene

There are six introns in *D. melanogaster*'s *SF3A1* (splicing factor 3a subunit 1; 120kDa) gene. The first one is a non-consensus U12-type intron, which has GC-AG termini, rather than GT-AG. This GC-AG U12-type intron is present in the 12 *Drosophila* species. Surprisingly, while the orthologous intron remains U12-type in honey bees (*Apis mellifera*) and beetles (*Tribolium castaneum*) (both, however, have GT-AG termini), it is U2-type in mosquitoes. As diptera, mosquitoes are closer to fruitflies than honeybees and beetles, which are hymenopteran and Coleopteran, respectively. The orthologous intron is U2-type in the *C. elegans* genome, and appears to be absent in Arabidopsis thaliana and all the vertebrate genomes we examined. Based on the coelomata hypothesis, a reconstructed evolutionary history inferred from states of orthologous introns, as shown in Figure 2-4, is that the ancestral metazoan U12-type intron was converted to U2-type in nematodes (non-coelomata), and was lost in deuterostomes (represented by four vertebrates here) after speciation of insects (arthropods). The U12-type intron in early insects was later converted to U2-type in one diptera lineage (mosquitoes). Independently, GT to GC substitution occurred in the 5' dinucleotide in another dipteran lineage. Based on the alternative phylogeny where nematodes and arthropods form ecdysozoa and are joined by dueterostomes, however, the scenario, which is more favorable from a parsimony point of view, would be that the U12-type intron arose after divergence of dueterostomes. The succeeding events were the same as those inferred according to the coelomata hypothesis except that there was no loss event in deuterostomes. Despite that the fates of this group of orthologous introns

varied, the coding regions flanking them are remarkably conserved across all organisms we examined.

One of the five major components of the U2-type spliceosome is U2snRNP, which is a dimer formed by SF3A and SF3B. Two monomers consist of three (120, 66, and 62 kDa) and seven (155, 145, 130, 49, p14, 14, and 10 kDa) subunits respectively. While SF3B is also present in U12 snRNPs, which is a part of the U12-type spliceosome, SF3A is specific to U2 snRNP. Thus, in fruitflies, there is an interesting relationship between two splicing systems. That is, formation of a U2-type spliceosome requires protein products involving the U12-dependent splicing pathway (removal of the U12-type intron from SF3A1 pre-mRNAs), the SF3A 120 kDa (SF3A1).

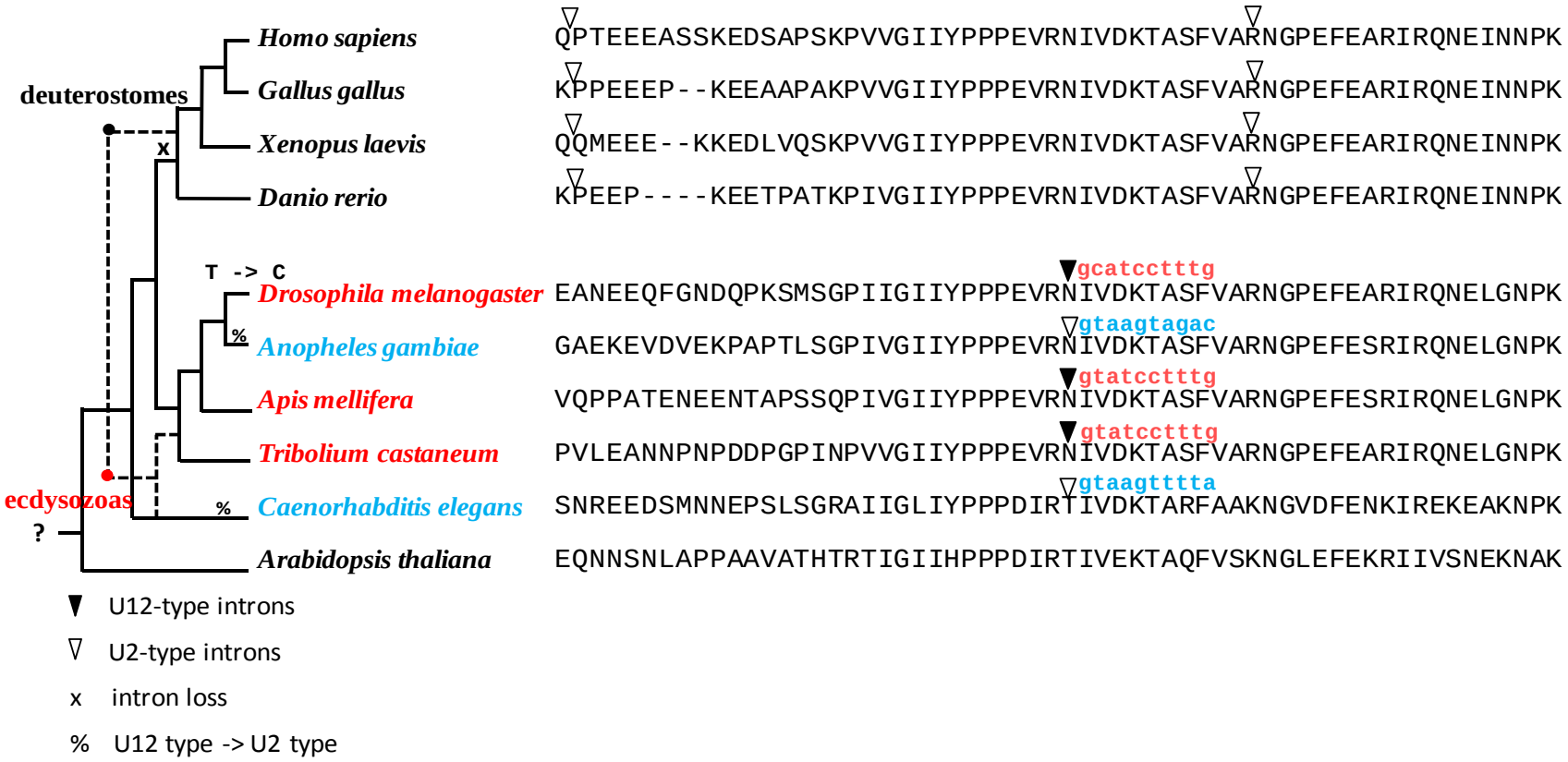

Figure 2-4: The GC-AG U12-type intron in the *D. melanogaster* splicing factor 3a subunit 1 (*SF3A1*) coding gene

The reconstructed evolutionary history of this intron based on ceolomata hypothesis. Species/taxon names are color-coded for status of the intron of interest (red: U12-type; blue: U2-type; black: absent). The dotted lines show the branching pattern according to ecdysozoa hypothesis. Note that in this alternative topology, the loss event in deuterostomes is unnecessary. Positions of introns are indicated by downward pointing triangles (followed by intron phase and 10 nucleotides, which are color coded in the way taxon names are, from the intron's 5' end) with respect to the protein alignment of the flanking regions, which are remarkably conserved across all organisms.

U12-type intron loss via retroposition

D. melanogaster coiled-coil domain containing 16 (*CCDC16*) gene contains one single intron, which is U12-type with AT-AC termini. The gene entries in the flybase and the RefSeq databases are incorrectly annotated with a U2-type 5' splice site that is 26 nucleotides upstream of the correct one. This leads to an unalignable region between the protein sequences derived from the RefSeq nucleotide entry and a cDNA sequence of this gene. The *C. elegans* ortholog has one intron that is not orthologous to the Drosophila U12-type intron. The orthologous introns in *A. thaliana* and *Oryza sativa* orthologs are U2-type. In vertebrates, mammalian orthologous genes are mostly intronless. Fish genes are less clear mainly due to incompleteness of fish genomes. For instance, the transcript of the putative zebrafish ortholog cannot be mapped to assembled contigs. Yet the best scoring sequence from a BLAST (Altschul *et al.*, 1990) search in tetradon whole genome shotgun sequences using human CCDC 16 protein sequence as a query suggests that the tetradon orthologs is intronless. However, we found that the AT-AC U12-type intron along with eight U2-type introns is present in the chicken and frog orthologs although like the one in the *D. melanogaster* genome, two AT-AC U12-type introns are misannotated.

The reconstructed evolutionary history of this intron based on parsimony principle (shown in Figure 2-5) is that it was U12-type before animals and plants split, given that the possibility of a change from U12- to U2-type is higher than the other way around or parallel insertion of introns. This ancient U12-type intron was converted to U2-type in

plants, and was lost, like all the other U12-type introns, in *C. elegans*. After insects diverged, *CCDC16* underwent at least one, or very likely two independent retroposition events in vertebrates (one in early mammals and one in early fish), which ridded the gene of all introns, including the U12-type intron. We found dense presence of SINEs (such as MIR, *Alu* families) at the vicinity of this gene in mammalian genomes. In the case of humans, there are additional LINEs (L2 mostly) and LTRs. These repeat elements might have facilitated insertion of the retrotranscribed version of the gene. Interestingly, in those vertebrate genomes that are better assembled and annotated, such as human and mouse, we were unable to identify additional copies of the gene. There seems to be some form of pressure that keeps the gene's copy number at one. While the original, intron-containing copy is retained in the *Gallus gallus* and *Xenopus tropicalis* genomes, mammals and fishes preserve the retroposed one. This case also demonstrates that annotation of non-consensus introns is prone to errors.

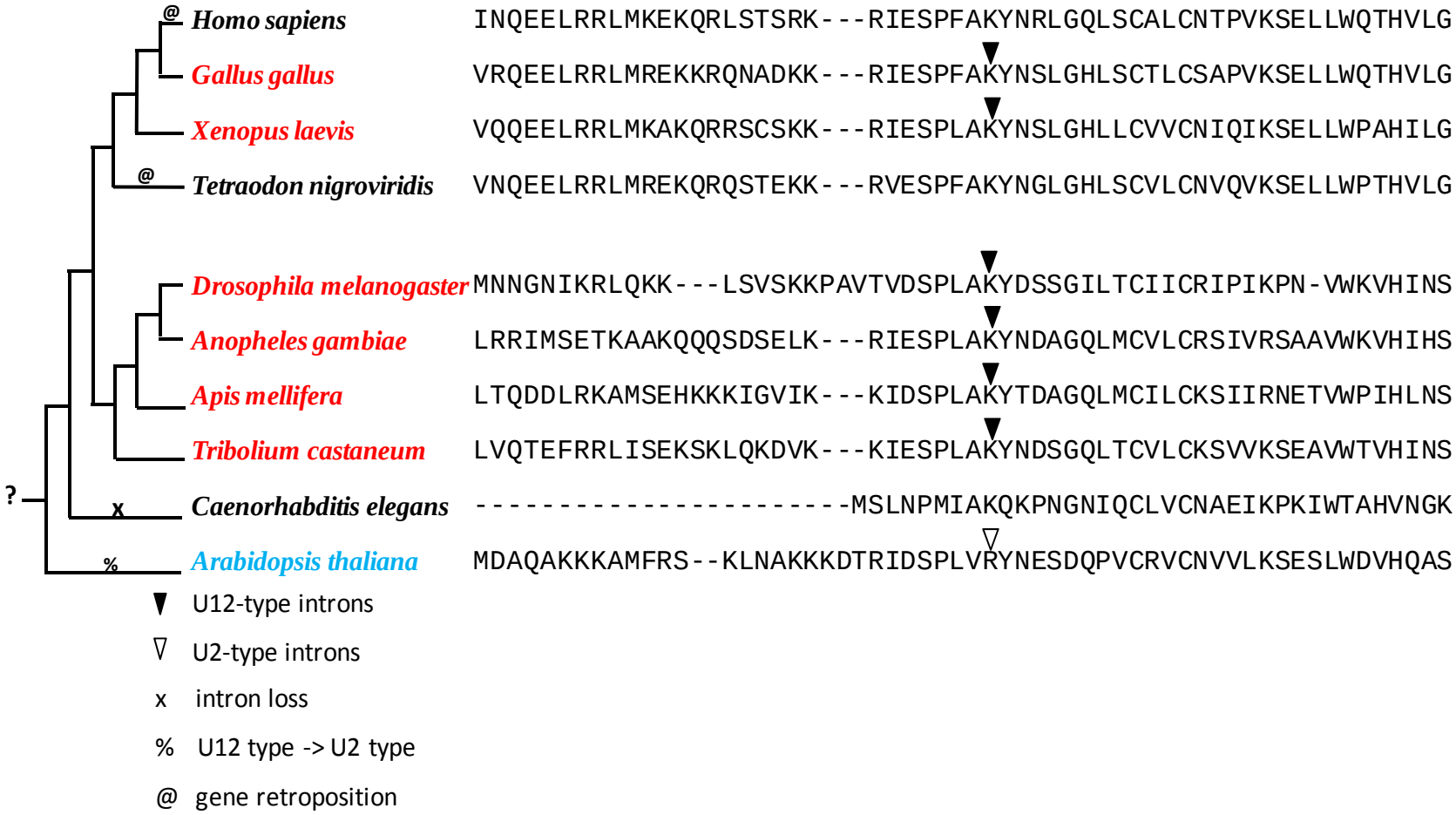

Figure 2-5: The AT-AC U12-type intron in the *D. melanogaster* coiled-coil domain containing 16 (*CCDC16*) gene and its orthologous intron in several representative organisms.

CCDC16 genes are intronless in mammals and at least one pufferfish but not chickens and frogs, suggesting that two gene retropostion events might have independently occurred in early mammals and fish.

U12-U2 conversion and neofunctionalization in XDH-AOX genes

An interesting case of U12- to U2-type conversion is the two consecutive U12-type introns in the gene encoding xanthine dehydrogenases (XDH). The second and the third of a total of 34 introns in the human XDH gene are U12-type. The gene encoding Aldehyde oxidase (AOX1) has similar gene structure (35 introns). Sequence identity between two proteins is ~50%. These proteins have similar enzymatic activities; while XDH catalyzes purine catabolism through urate, AOX1 does aldehyde substrates. They reside in peroxisome and cytoplasm respectively. The flanking exons of the two U12-type introns encode part of the iron/sulfer domain. Interestingly, while the second intron of *AOX1* remains U12-type, the third intron is U2-type. This pattern is, in fact, the circumstance for all mammals studied.

The phylogenetic tree topology of putative homologous protein sequences suggests that *XDH* gene is the ancestral copy that gave rise to AOX via a gene duplication event that occurred after the divergence of *C. intestinalis*. Although two copies of *XDH*-like genes are predicted in the *C. intestinalis* genome they are more likely resulted from another duplication event. This suggests that two genes remain *XDH* and one of them acquired new function after *C. intestinalis* diverged. While the overall gene structure varies from nematodes (15 introns) to insects (three introns), and to vertebrates (35 introns), orthologous introns of the two human *XDH* U12-type intros remain present, with the exception in fruitflies, across all *XDH* and *AOX* genes. They are both U2-type in *C. elegans* and *Ciona* but reveal an interesting pattern (Figure 2-6) among vertebrate

XDH and *AOX1* genes. The second and the third intron of *XDH* and *AOX1* genes are U12-then-U12 and U12-then-U2 respectively in all mammals, and ambiguous-then-U12 and U12-then-U12 respectively in chickens. Given that *XDH* gene is the ancestral state, this indicates that there was a U12- to U2-type switch of *AOX1*'s third intron in the common ancestor of mammals. The switch seemed to have been triggered by deletion of a thymine at position +4 of the 5' splice site, and was followed by substitution and/or deletion at the BPS. Zebrafish *XDH* and *AOX1* genes appear to have undergone a unique intron evolutionary dynamics while the two pufferfishes and chickens have more in common. Both intron 2 and 3 of *AOX1* gene are U12-type in two puffer fishes and chickens, but are U2-type in zebrafish. This suggests that the type switch in zebrafish is independent from the one mentioned above in early mammals. Interestingly, the downstream intron (intron 5) in zebrafish *AOX1* retains a perfect U12-type 5' splice site while the other vertebrates do not. Similarly, intron 2 of *XDH* is U2-type in two pufferfishes and ambiguous in frogs and chickens, but U12-type in zebrafishes. Intron type switch in *XDH* and *AOX1* genes does not seem to have posed great impact on the protein products. Most introns are lost in insect *XDH* genes. No evidence directly links the intron type change to neofunctionalization of *AOX1* gene.

Species		Intron 2		Intron 3
AOX1	**type**	**5′ and 3′ junctions**	**type**	**5′ and 3′ junctions**
H. sapiens	U12	AGGAAGAAGC|GTATCCTTT..GATCTTTAACTATACCTCTTCCAG|TTCGAC	U2	AGAGGATAAG|GTACCGTGC..GCTTTTCTTTTGCATTCTGAAG|GCATCA
M. musculus	U12	AGGAAGAATC|GTATCCTTT..TCTCTTTAACAGAGCTCCTTCCAG|TCCGAC	U2	AGGCGATCAG|GTAGGTGCA..GCTTCTCTTTTGCATTCTGAAG|GCATCA
R. norvegicus	U12	AGGAAGAACC|ATATCCTTT..TCTCTTTGACCGAGCTCCTTCCAG|TCCGAC	U2	AGAGCATCAG|GTGGGTGCA..GTTTCTCTTTTGCATTCTGAAG|GCATCA
C. familiaris	U12	AGGAAGAAGT|GTATCCTTT..TTTCCTTGACTCCCCTCCTTCCAG|TCTGTC	U2	CGAAGATAAG|GTATCCTAC..TCTTTGTGTTCCTCGTATGAAG|ACACTT
M. domestica	U12	AGGAAGAAAC|GTATCCTTT..TTTCTTTGACTACTGCCCTTCAAG|TCCATC	ambi	AAAAAATAAG|GTATCCTTC..ATTTCTTATTTTTGTTCTGAAG|ACATTA
G. gallus	U12	CGAAAGAGAC|GTATCCTTT..TTTTCCTTAATAGCGGAATTTCAG|TCCGTC	U12	AGAAGATACG|GTATCCTTA..TATTCACCTTAATGTTCCAAAG|ACACTA
D. rerio	U2	AGGAAAAAAT|GTATATTTT..GTATATGCATGTTTTTGTATGTAG|TGCGTT	U2	AGAGCATCAG|GTATTACAG..AGCATTTTTTTTCTCTGTTTAG|TCATTT
T. nigroviridis	U12	AGAGACAGGC|GTATCCTTC..GCCTTTCCTTCACCGCCGCTTCAG|TGAGGC	U12	GGAGCATCAC|GTATCCTTC..GTTTGTCTTCACTCAGGACCAG|ACATTC
F. rubripes	U12	AGACAGAAAC|GTATCCTTC..GGCTTTCCTTCACTCTGACTTCAG|TGAGGC	U12	AAACCATCAC|GTATCCTTG..ACCGACCTTCACTCAGCATTAG|ACACCT
XDH	**type**	**5′ and 3′ junctions**	**type**	**5′ and 3′ junctions**
H. sapiens	U12	AGAAGAAAGT|GTATCCTGA..TCTCCTTAACTCTTGACCACCCAG|TGGGGC	U12	ACAAGATCGT|GTATCCTTT..TGACCTTAATCTGGGGTTCTAG|CCACTT
M. musculus	U12	AGAAGAAAGT|GTATCCTGA..TGTCCTTAACAAGAGGCTGCTCAG|TGGGGC	U12	ACAAGATCGT|GTATCCTTT..TGGCCTTAATCTGTGGTTCTAG|TCATTT
R. norvegicus	U12	AGAAGAAAGT|GTATCCTGA..TGTCCTTAACAAGTGGTTGTTCAG|TGGGGC	U12	ACAAGATTGT|GTATCCTTT..CAACCTTAATCTGTGGTTCTAG|TCATTT
C. familiaris	U12	AGAAGAAAAT|GTATCCTGA..ATTCCTTAACTCTCGACCATCCAG|TGCGGC	U12	ACAAGATCGT|GTATCCTTT..TGACCTTAATCCAGGGCCCCAG|CCACTT
M. domestica	U12	AGAAGAAAAT|GTATCCCTG..TTTTCCTTAACTCTTGCACTTTCA|GTGGGG	U12	AAAAAATTGT|GTATCCTTT..TATCCTTTACTTGAGTTTCCAG|CCACTT
G. gallus	ambi	CGAAGAAAAC|GTATCTGGT..TTTCCTTGATTCCTGTACTCTCAG|TGGGCC	U12	AGAAAATCCT|GTATCCTTT..TGACCTTAATCTATCATTTTAG|CCACCA
X. tropicalis	ambi	CGGAGAAAAT|GTACGGTTT..TTCCCTTAACCTCTTTCCATTCAG|TGGGAT	U12	ACAGAATACT|GTATCTTTT..CTTCTTAAACGCATAATTCCAG|ACACTA
D. rerio	U12	AGAAGAAGCT|GTATCCTTT..TTTCCTTAATGTTGTCACCTGCAG|TGGGTC	U12	ACCGGATTAT|GTATCTTTT..TATCCTTGATTCTCTCTTGAAG|TCACTA
F. rubripes	U2	AGGAGAAAAT|GTGAGCTCA..GTGTGTGTGTGTGTGTGTGTTCAG|TGGGAT	U12	AACAGCTACT|GTATCCTTC..TGGTCTTTAACTCCGGTTCCAG|TCACTA
T. nigroviridis	U2	AGGAGGAAAT|GTAAGCTCC..AACCATCTTTTCTTTAGCCTCCAG|TGAGGA	U12	AACAGCTGCT|GTATCCGTG..AGGTCTTTAACTGCGGTTTCAG|TCACTA
C. intestinalis	U2	AGAACAAAAC|GTGAGTGAC..ATTATGTGACTTTTTTTGTTTTAG|TTCGTT	U2	ACCGCATCGT|GTGAGTTTG..CGTTGTTTTTTTGTTTTTTCAG|ACATTT
A. mellifera	U12	AGAAATAAGT|GTATCCTTT..AAGTTTTCCTAACTAGAATAATAG|TACGTT	U12	GAATTATTAC|GTATCCTTT..TTTAATTTTTTAATTTAAATAG|ACATCT
A. gambiae	lost		lost	
D. melanogaster	lost		lost	
C. elegans	U2	AGAGATAAAT|GTGAGTTCT..TTTACGAGTTTATTTTATTTTAAG|TGAAGC	U2	GTGAAATCAA|GTGAGTTGG..TTTTAAAGCCTAATTGTTCCAG|ACATTT

Figure 2-6: The two U12-type introns in *AOX1* and *XDH* genes

Two U12-type introns are present in VPS16 in most genomes but both are missing in zebrafish

As mentioned above, three of the eight U12-type intron losses in fish lineage are likely due to gene retroposition (Table 2-4); they were lost as the rest of introns within the gene. For the remaining five losses, three occurred in three pufferfish genes while two in a single zebrafish gene, which encodes VPS16 (vacuolar protein sorting 16).

Homologous to the yeast *VPS16*, the human *VPS16* gene is believed to encode proteins that may play a role in lysosomal delivery of vesicle-mediated proteins. There are 23 introns in the human *VPS16* gene. Introns 9 and 13 are both U12-type. One of the isoforms shows that the donor site of intron 9 couples with the acceptor of intron 13 such that four exons and five introns are removed as a single intron (Figure 2-7). This leads to a deletion of 144 aa in the protein. It is not clear, though, whether this isoform is functional or not. Several independent studies, including one cloning and characterizing four human *VPS* genes (Huizing *et al.*, 2001) and two large scale cDNA projects (the German cDNA consortium, 2002; the MGC Project Team, 2004), found ubiquitous expression (at least all eight tissues examined in the *VPS* study) of the longer isoform. Another large scale study characterizing full-length human cDNAs (Ota *et al.*, 2004), however, reported expression of the shorter isoform (isoform 2) only in kidnies. This seems to suggest, while isoform 1 expresses ubiquitously, expression of the isoform 2 is tissue specific and the regulation of alternative splicing is likely to involve the minor spliceosome.

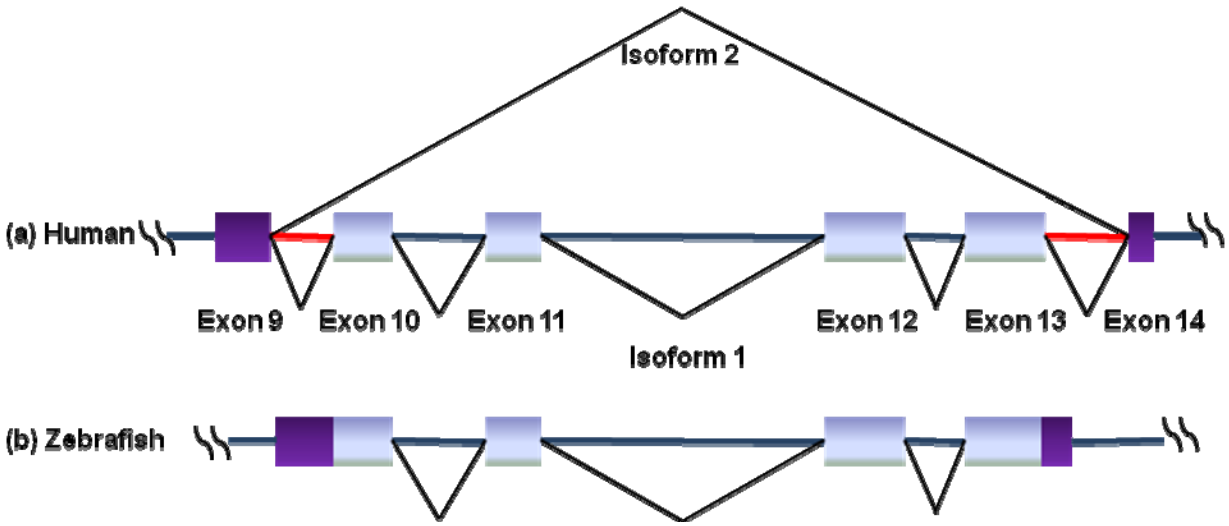

Figure 2-7: The two U12-type introns in the *VPS16* (vacuolar protein sorting 16) gene. Two introns form a hybrid one in the human gene. Double losses of zebrafish

(a) In isoform 2 of the two human splicing variants (isoform 1:NM_022575; isoform2:NM_080413), the donor site of intron 9 couples the acceptor site of intron 13, which causes four exons (exon 10, 11, 12, and 13) that encode 144 aa to be skipped. (b) Both of the two U12-type introns are missing in the zebrafish *VPS16*gene.Purple boxes represent exons. Light blue boxes are the exons that are skipped in one isoform. Red thin lines represent U12-type introns while blue lines U2-type introns.

Interestingly, orthologs of these two U12-type introns are both missing in the *Danio rerio VPS16* gene while the remaining 21 introns are all in place. Mechanisms for causing intron deletion that have been proposed/known to date (Roy and Gilbert, 2006) predict that intros are lost either as single introns (by genomic deletion) or multiple adjacent introns (mediated by reverse-transcribed-mRNA). Deleting two nonadjacent introns would require more than one deletion event. It is striking in the case of *Danio rerio VPS16* gene that the targets of two events are both U12-type introns that are among 23 and 22 introns each time. The straightforward explanation, although with a very low probability, is that two U12-type introns were lost in two independent events. However,

in light of the aforementioned human *VPS16* isoform where the donor site of intron 9 couples with the acceptor of intron 13 in lieu of its own, it might have been that a similar circumstance with a series of splicing and reinsertion led to the double losses. Among the abundant repeat elements spreading across the entire gene, there might be relics of what has facilitated the double losses. On the other hand, if the double losses are, indeed, the outcome of two independent loss events, was the gene under certain pressures to remove them or gaining advantages by doing so?

Conclusions

We investigated conservation and changes of intron state among orthologous U12-type introns identified in 18 metazoa genomes. At most, one case of U12-type intron gain was found. Intron loss contributes much greater than conversion of U12- to U2-type introns to the evolution U12-type introns. U12-U2 conversion occurred more frequently in GT-AG subtype. U2 to U12 conversion has been regarded nearly impossible we, however, discovered one such case. This dipteran U12-type intron is also a “twintron”. With the presence of the ancestral U2-type splice sites, the U12-type cis-elements arose from activation of cryptic splice site that can be dated to early diptera. We discovered six non-canonical U12-type introns with three sets of terminal dinucleotide that have never been reported (CT-TC, GG-AG, and GA-AG). This indicates that U12-type 5’ termini are more degenerative than previously thought.

The presence of orthologous U12-type introns among vertebrates has been highly conserved in the past 550 million years after *Ciona* diverged. The conservation is especially enormous among eutheria. Of the 442 unique intron positions, only one unambiguous loss in mice and one U12 to U2 conversion in rats were observed. The non-chordates investigated have considerably fewer U12-type introns than vertebrates. Interestingly, the lineage with the fewest introns, diptera, is the only one that acquired a new U12-type intron. Although there is no orthologous intron from outgroups to support that most of the absent cases in the near 200 orthologous U12-containing genes are due to introns loss, rather than intron gain in vertebrates, several facts argue for the intron loss view. Near half of vertebrate U12-containing genes are absent in the non-chordates. This is likely a combination of gene losses in the non-chordates and neofunctionalization/subfunctionalization of genes arising from large-scale segmental duplication or whole genome duplication that occurred in early vertebrates, as is exemplified by the multiple-U12-containing genes that encode ion channel alpha subunit.

To some extent, the evolution of U12-type introns resembles that of U2-type introns. The initial incorporation of introns brings slightly deleterious effect to individuals. However, those introns that survived genetic drift were given the opportunity to pick up new functions. Once they did their evolutionary fate entered a different mode. Expanding this process to many introns over long evolutionary time, we then see the beneficial and indispensable effects RNA splicing has evolved. On the other hand, U12-type splicing pathway imposes stronger deleterious effect than U2-type on individuals including higher susceptibility to mutation and lower processing rates, which invokes stronger pressure for it to be eliminated from a population. This might have eventually

led to independent extinctions of the system in many organisms as it is becoming evident that U12-type splicing pathway was once more prevalent in eukaryotes. Furthermore, the higher flexibility (or lower fidelity) in recognition of splice sites in the U12-type spliceosome makes the conversion of U12-type to U2-type intron possible. This adds more complexity to the evolution of U12-type introns. Hence, dynamics of U12-type introns can be very stochastic during a relatively short span of time, such as within metazoan, as our results have revealed.

Materials and Methods

We downloaded the MySQL version of the entire U12DB version 1.0 published by Alioto (2006; ftp://genome.imim.es/pub/software/u12/u12db_v1_0.sql.gz). U12DB, a database with a web interface, provides information for U12-type introns identified in 20 sequenced eukaryotic genomes and clusters of U12-type introns and orthologous introns. In the U12DB orthology between two introns was determined by sequence similarity of their flanking exonic sequences. Each orthologous U12-type intron cluster represents a unique intron position with respect to the protein sequence that is derived from the hosting genes. At least one of member introns is U12-type. Therefore, a set of orthologous genes can share more than one U12-type intron cluster.

Status of the introns in a cluster were encoded with “1” - U12-type, “2” - ambiguous, which is denoted as “U12/U2” in the U12DB, “3” - U2-type, and “4” - intron absent. Intron states of all intron clusters were converted into a matrix of intron status versus species. Each orthologous intron cluster is represented in the matrix as one line of

intron states for all species analyzed. If gene orthology information is lacking for a certain species, the state of this species is coded by "5" – missing data. On the other hand, if multiple genes from one species are present in one cluster, intron states of the same species were ranked in the order of U12-type, ambiguous, U2-type introns, and intron absent. The top ranking was retained to represent the species. Patterns of intron states were used to infer conservation of U12-type intron among taxa at varying phylogenetic distances and.

We conducted manual inspection extensively to correct potentially false cases of intron absence and missing orthologs in the U12DB. We resorted to a variety of sources for validation of automated results when necessary. These sources include Entrez Gene (Maglott *et al.*, 2005), evidence viewer, and Blink from NCBI, Ensembl (Birney *et al.*, 2004), the UCSC Genome Browser (Kent *et al.*, 2002), and the UniProt Knowledgebase (UniProtKB; Bairoch *et al.*, 2005).

Several scenarios can result in bogus "absent" cases. We found cases where orthologous introns were put in two clusters due to an error in alignment, which led to two false "absent" cases. If activation of cryptic splice sites by nucleotide substitution or incomplete genomic intron deletion fortuitously does not incur shifting of the reading frame, depending on the decrease/increase of the coding region, this intron would appear not alignable with the orthologous introns. These introns would be described as two intron clusters and generate two bogus absent cases.

Lack of orthologs from certain species can be caused by two artificial factors. First, heterogeneity in quality of genome assemblies and amount of EST and cDNA data. We observed many cases where U12-type introns are identifiable through sequence

conservation with other species but are not or incorrectly annotated. Second, orthology in genes and in introns are crucial in inferring conservation of introns. Clustering of orthologs of organisms across a large phylogenetic distance, however, can be challenging. Large scale and automated clustering can be further confounded by the heterogeneity in gene copy numbers (duplicated genes) and gene evolution rates. For example, there are five cases, such as *Syntaxin 6* and *prospero* genes, where human and fruitfly U12-containing genes were clustered by Ensembl as being in different orthology groups but turned out to be orthologous after manual inspection of the protein sequences. On the other hand, there are also different genes being clustered in one group. The cluster having *XDH* and *AOX* genes is one such an example (see Results and Discussion).

CHAPTER 3

PHYLOGENETIC DISTRIBUTION OF U12-TYPE INTRONS IN SEVEN EUKARYOTIC GENOMES

Abstract

In this study, I used sequence and annotation data downloaded from the ensembl and TIGR databases, and a position-specific weight matrix with likelihood ratio approach, to identify U12-type introns in seven complete genomes (*Arabidopsis thaliana*, rice, human, mouse, rat, fugu, and fruit fly). Only a few AT-AC U12-type introns were detected, primarily because introns with AT-AC terminal dinucleotides were often misannotated. Widespread under-detection of U12-type introns and a bias towards GT-AG U12-type introns became clear with the release of U12DB (Alioto, 2007), a database of orthologous U12-type intron clusters. More U12-type introns, particularly those with AT-AC termini, were reported in the U12DB because the database incorporated introns predicted by mapping EST data in addition to genomewide annotation. Phylogenetic analysis of U12-containing genes showed that in nearly two third (289) of the defined homology groups, U12 introns are present in more than one species. U12 introns are conserved across mammalian genomes investigated. Interestingly, there are 20 groups where U12 introns are conserved across animal and plant kingdoms. Examination of homologous U12 introns revealed that intron phase and position with respect to the exon-intron structure are highly conserved.

Introduction

Two genome-wide studies - in human (Levine and Durbin, 2001) and *Arabidopsis* (Zhu and Brendel, 2003) - showed that frequencies of U12 introns are different in the two species. However, those studies did not address evolutionary aspects of U12-type introns. In this study, I computationally identified U12-type introns in seven complete genomes (*Arabidopsis thaliana*, rice, human, mouse, rat, fugu, and fruit fly). I then conducted phylogenetic analysis of U12-containing genes. Conservation pattern of U12-type intron position and intron phase were also investigated.

Materials and Methods

The framework of this study consists of identification of U12-intron within species, and phylogenetic analysis of U12-introns and U12-containg-genes among different species. Sequence and annotation information related to five animal genomes were downloaded from Ensembl (http://www.ensembl.org), while plant data were downloaded from the Institute for Genomic Research (http://www.tigr.org). A full list of analyzed genomes along with the data version is presented in Table 3-6.

3-6: Genomes investigated and assembly version

Species	*Assembly version*
Arabidopsis thaliana	ATH1 version 5.0
Oryza sativa	OSA1 version 1.0
Homo sapiens	NCBI34
Mus musculus	NCBIM30
Rattus norvegicus	RGSC3_1
Fugu rubripes	FUGU2
Drosophila melanogaster	DROM3A

Identification of U12-dependent Introns

Identification of U12-dependent introns was carried out using a pipeline developed in our laboratory. The system was constructed with Perl scripts in combination of Mysql database. Ensembl Perl API and TIGR's XML parser were incorporated, too. The pipeline consists of three steps and each genome was processed separately.

First, all introns for a given genome were extracted based on a genome annotation. Because U12-dependent introns are characterized by a conserved 5' Splice Site (5' SS) and a Branch Point Site, each intron was represented by 20 and 45 nucleotides of the 5' and 3' end sequences respectively. For introns that are shorter than 70 nucleotides the whole intron sequence was stored. Both 5' and 3' end are extended by 10 nucleotides from the adjacent exon to preserves the potential non-canonical splice sites that have been mis-annotated.

To make this process of identifying U12-introns more efficient, all introns were screened by exact-letter matching for U12-intron candidates. Only these candidates were

subject to the weight-matrix (profile) based classifier. Because all known U12 introns contain an ATCC motif at position +3 (+1 being the first nucleotide of an intron) to +7 of the 5' SS, this is a criterion for a U12-intron candidate. If a more extended motif [A or G]TATCC was found in the 5' end, then the ATCC motif is not necessary at position +3. In other words, the extended motif can be shifted from the annotated exon/intron boundary. This is, as mentioned earlier, to recover non-canonical splice sites, particularly AT-AC termini, that have been falsely adjusted to agree with canonical splicing sites (see below). Although the ATCC motif is not sufficient for U12 splicing pathway, it is a required one. Introns that didn't contain the ATCC signal at their 5' end were used as a U2 training set if the first intron position is nucleotide A or G since none of experimentally characterized U2-introns so far was found to have nucleotide other than A or G at site +1. U12-intron reference set (U12 training set) consisted of 48 U12-dependent introns from various organisms that are reported in literature and compiled by Sharp and Burge (1997).

Those U12 intron candidates were then evaluated by its likelihood of being U12-type vs. that of being U2-type. We adopted the statistical model developed by Burge et al. (Burge *et al.*, 1998). A weight matrix (profile) for U12 5' SS was created from the U12 training set and was common for all genomes analyzed. However, weight matrices for U2 5' SS were generated from training sets for each genome individually. Empirical nucleotide frequencies estimated from scanned regions were added to the matrix as pseudo count to avoid zero values at any position in a matrix. Log-odd ratio between U12 and U2 signals was calculated and normalized (subtracting the mean and dividing by the standard deviation) for introns of the two training sets. Scores for a branch point site

(BPS) were calculated in a similar way. U12-intron candidates were then scanned for the profiles. A 5' SS score and a BPS score were calculated and plotted on a bivariate plane for each candidate. An outlier curve x2 + y2 = 20 was used as a determinant, which yields an error rate equivalent to p-value 0.05. In the first quadrant of the 5'SS-BPS bivariate plane, dots that are above the outlier curve represent sequences that have statistically strong 5' SS and BPS and they are classified as U12 introns. Besides U12 introns, we defined another class of intron – what we called shifted U12-intron. They are introns containing both conserved U12 motifs, but either or both are slightly shifted from the annotated 5’ boundary and/or the position where the distance between the branch point and the 3’ boundary is within eight to 35 nucleotides. As mentioned earlier, most automated annotation programs are geared toward forcing GT-AG as the termini introns. By classifying cryptic-U12-introns, misannotated U12-introns can be rescued as such misannotation can be seen by comparing a cDNA sequence with its cognate genomic location; or relics of U12 motifs can be picked up. These cryptic U12-introns were afterwards manually validated against available expression evidence.

Phylogenetic Analysis of U12-containing Genes and U12-introns

Many of U12-containing genes may be resulted from relatively recent gene duplication. On the other hand, it has been proposed that U12-introns tend to convert to U2-introns. To investigate conversation of U12-introns among species, we searched for homologs of all U12-containing genes that were detected in the previous stage, and clustered them into homology groups.

Amino acid sequences of U12-containing genes were used as queries in search against proteoms of the seven organisms using BLASTp with the default parameter except filtering option being “for lookup table only”. Proteoms of the five animals (known and novel protein sequences defined by Ensembl) and the two plants were used as a database. Stringency of criteria used in the clustering process dictates the number of clusters obtained. To determine a proper clustering scheme, various combinations of thresholds for E value and relative score (Bits score / self hit bits score) were tested. Query and BLAST hit sequences above the cutoff values were clustered into homology groups with single linkage algorithm. In the cases where there is more than one transcript for a given gene, the one that contains U12-introns was used. If none of the isoforms contains U12-introns, then the longest one was included in our dataset.

Clustalx with default parameters was then employed to create multiple alignments of amino acid sequences. Bootstrap consensus trees were obtained by Neighbor-Joining method with 1000 replicates. Resulted trees were manually inspected and divided into single-copy-gene (presumably ortholog) groups and others (with paralogs from the same species).

Position and Intron Phase of U12-introns

Introns in the same phase and at homologous positions with respect to the gene structure are likely to be derived from a common ancestor (Long et al. 1998). To investigate homology of U12-introns found within groups, positions and phases of U12-introns were analyzed for groups where at least two genes contain U12-introns. Multiple

amino acid sequence alignments were used to guide the alignment of nucleotide coding sequences and then positions of U12-introns were mapped onto these nucleotide alignments so that position and intron phase of U12-introns can be observed. A set of web-based visualization tools, such as displaying aligned gene structure and colored nucleotide alignment, were developed for this purpose.

Distribution of U12-introns within one genome

U2-introns overwhelmingly outnumber U12-introns in all genomes investigated. In the mouse genome, which has the largest proportion of U12-intron, there are 433 U12-introns, 177132 U2-introns, and near 23500 genes. With such low U12-intron density, however, in most of the genomes we analyzed, several genes are still found to harbor more than one U12-intron. U12-introns do not seem to distribute in the genome randomly. We carried out a statistical test to see whether U12-introns are randomly distributed in a genome. Under the null hypothesis that U12-type introns are randomly distributed in a genome, for a given gene, the probability that a certain number of U12-intron present in this gene follows a binomial distribution. Thus, the probability (q) that a gene contains multiple introns can be approximated by $q \approx 1- (1-f)^{d-1}$. The parameters required for the calculation – frequency of U12-intron (f), intron density (d, average number of introns per gene) were estimated from individual genome data. With q, number of U12-containning gene and number of multi-U12-containing gene detected in the genome, the probability for observing such number of multi-U12-containing gene can then be calculated.

Results

Computational identification of U12 introns in seven eukaryotic genomes

We have computationally scanned human, mouse, rat, pufferfish, fruit fly, *Arabidopsis*, and rice genomes for U12 introns. The number and frequency of detected U12 introns for each genome are shown in Table 3-7. Since the majority of U12-hosting genes contain only one U12 intron, the number of U12-hosting genes is very similar to that of U12 introns (Table 3-8). U12 introns are consistently rare in all investigated genomes. None of the U12 intron frequencies is higher than 0.24%. The frequency is the highest in mammals, then in fish and then in plants. The remarkably lower number of U12 introns detected in rice, compared to that of *Arabidopsis*, can be attributed to the relatively poor quality of annotation rather than real biological phenomena since we used the first annotation of the genome. Recently I ran a quick test to roughly estimate the quality of the rice annotation. Non-canonical introns are rarely annotated in this genome, and a few cases of introns homologous to *Arabidopsis* U12 introns that I checked manually show conservation of unannotated non-consensus introns. Many protein components of the U11 snRNP (35K, 48K, 59K) and U11/U12 di-snRNP (65K, 31K, 25K) have recently been identified in rice (Lorkovic *et al.*, 2005), indicating that U12 spliceosome is well conserved in rice. Therefore, it can be expected that the number of U12 introns will be much higher once better genome assembly is used. I have defined a different set of U12 introns, shifted U12 introns. These introns have U12-type 5' SS and BPS signals that are above the threshold to be classified as U12 introns but either the 5'SS or 3'SS, or both have to be shifted several nucleotides away from the annotated

exon intron junction. These shifted U12 introns can reflect the following possibilities. First, they are mis-annotated. This is particularly likely to be the cases for introns with non-consensus splice sites because some automatic annotating programs are biased toward assigning consensus splice sites GT-AG. Second, they are novel splicing variants. Third, they are relatively recent U12-U2 conversion event. These shifted U12 introns would be of interest for further investigation. A homology group with multiple misannotated U12 introns is demonstrated in Figure 3 and Figure 4.

Table 3-7: Numbers of U12 introns identified in seven genomes

species	Tot int	U2-int	U12 Cand.	U12-int	Shifted U12-int
Arabidopsis	117461	117289	1372 (1.2%)	164 (0.14%)	64
Rice	184624	184566	1833 (1.2%)	48 (0.03%)	47
Human	185815	185352	1829 (1.2%)	420 (0.23%)	141
Mouse	177603	177132	1955 (1.2%)	421 (0.24%)	147
Rat	153714	153331	1652 (1.2%)	344 (0.22%)	115
Pufferfish	126551	126304	1070 (1.2%)	231 (0.18%)	51
Fruit fly	46235	46220	545 (1.2%)	13 (0.03%)	31
Total	992003	990194	10256 (1.0)	1642 (0.15%)	

Tot int: numbers of total introns analyzed.

U2 int: numbers of U2-introns identified.

U12 cand.: numbers of candidate U12 introns identified based on a conserved motif (ATCC) at 5' splice site (5' SS) from nucleotide positions +3 to +6. Values in the parentheses indicate the proportions of U12 cand. to total introns (U12 cand./Tot int).

U12-int: numbers of detected U12 introns. Values in the parentheses indicate the proportions of U12 introns to total introns (U12-int/Tot int).

Shifted U12-int: shifted U12 introns have U12-type 5’ SS and BPS signals that are above the threshold to be classified as U12 introns but either the 5’SS or 3’SS or both have to be shifted several (at most ten) nucleotides away from the annotated exon intron junction

Table 3-8: Numbers of U12-hosting genes identified in seven genomes

species	Total genes	U12-hosting genes	Multiple-U12-hosting genes*
Arabidopsis	29388	158	5
Rice	57221	48	0
Human	21787	410	22
Mouse	25307	409	26
Rat	22159	325	20
Pufferfish	35180	222	10
Fruit fly	13525	14	0
total	204567	1585	83

*number of genes containing more than one U12 introns

Characterization of U12 introns

Distribution of terminal dinucleotides

Splicing of pre-mRNA is a two-step process. While the 5'SS and the BPS are critical in the initial recognition of introns, the completion of splicing requires active participation of the 3'SS. Experiments suggest non-Watson-Crick base pairing is always favored or required between the first and the last nucleotide of the intron, such as G-G base pair in GT-AG introns and A-C base pair in AT-AC introns. An experiment on AT-U12 introns shows that although the 5' terminal nucleotide A can base pair with any

nucleotide at the 3' terminal, C yielded the highest splicing efficiency, while T the least (Dietrich, 2001). Our results (Table 3-9) agree with these observations. Very few of the 5' and 3' termini observed in the detected AT- and GT-U12 introns across seven species show Watson-Crick base pairing between the first and the last nucleotide.

Table 3-9: Occurrences of different types of terminal dinucleotides in U12-dependent introns

5' termini	3' termini	count
AT	AA	5
AT	AC	57
AT	AG	13
GT	AG	1553
GT	AT	9
GT	GG	3
GT	TA	1

Comparisons of intron size and GC-content between U12 and U2-type introns

We divided introns into four categories (U12, shifted U12, U12-canidate and U2 type) to compare their size and GC-content. Even within one particular species and category, intron sizes can be extremely diverse. The size can range from a number of nucleotides to several 10,000 nucleotides. Those mini-introns (< 10 nts) are probably resulted from mis-annotation because it is questionable whether an intron with such short length is able to form lariats. Nevertheless, the trends exhibited in the intron size

distribution of U2-introns and U12-introns are not significantly different (data not shown). To have a rough but quantitative idea about intron sizes of different categories and species, mean and median values were calculated and shown in Figure 3-8 . The seven species distinctly form two groups: 1) the long-intron group, to which mammals belong, and 2) the short-intron group, which includes fish, fruitflies, and plants. In the long-intron group, U12 type has the lowest median and mean size. The variation in median size among types is not as dramatic as in mean size, which is more prone to be skewed by extremely long introns. For all species, GC-content does not differentiate U2 and U12 type introns, while GC-content is always lower in introns then in exons.

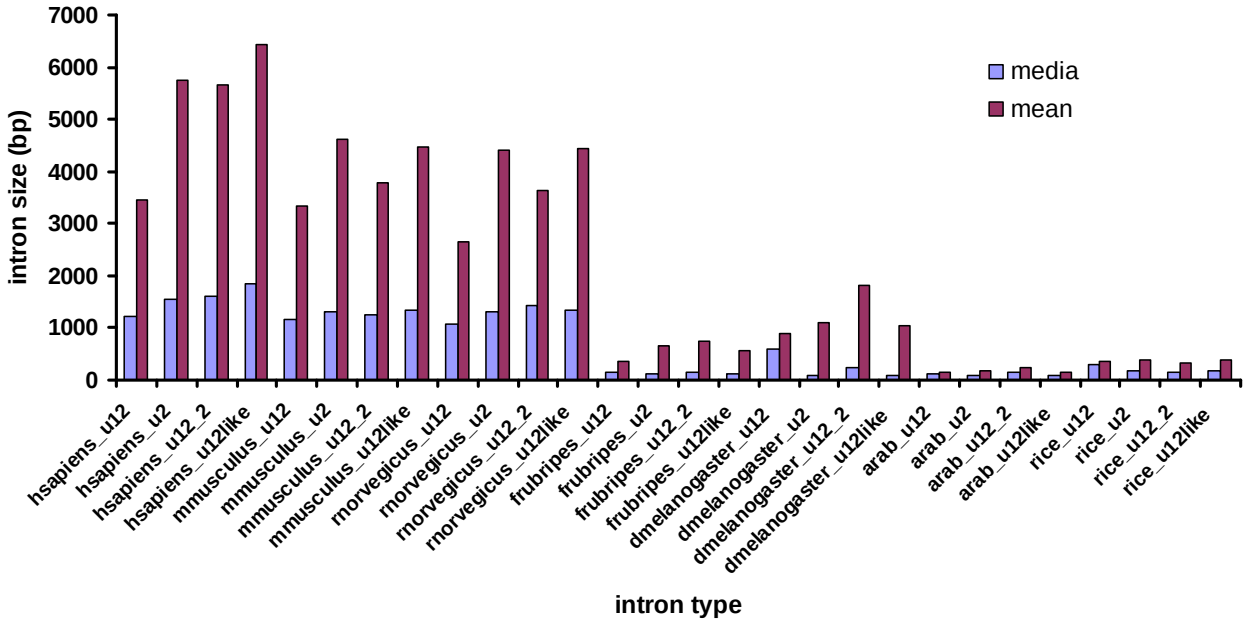

Figure 3-8: Intron size (Median and mean values) of four categories of introns in each of the seven species.

Phylogenetic analysis of genes containing U12 introns

All U12-hosting genes and their homologs obtained by BLASTp search (with thresholds E value < 10 -4 and Bits score / self hit bits score > 0.15) were clustered into homology groups using a single-linkage clustering algorithm. After three singleton groups (where only one member is in the group) were removed from the initial 490 groups, 487 homology groups were obtained. Distribution of group size is shown in Figure 3-9 We manually inspected the phylogenetic trees and divided them into two categories: multi-species group and single-species group. 298 multi-species groups (61% of 487 groups) include 1421 U12 introns (86.48% of 1642 U12 introns) whereas 189 single-species groups (39%) include 221 U12 introns (13.52%). Each species accounts for various number and percentage of the single-species groups as shown in Figure 3-10. The 1421 U12 introns are evolutionarily conserved (present in more than one homologs) thus are more likely to be genuine and presumably have stronger splicing signals.

The highest number of single-species group contributed by *Arabidopsis* does not necessarily mean that this plant has the most species-specific U12 introns. It is rather by the same token as the extraordinarily low U12 intron frequency of rice – the poor quality of the rice annotation. Improved annotation will increase the number of rice U12 introns and many of their hosting genes will very likely to be homologous to the *Arabidopsis* U12-hosting genes.

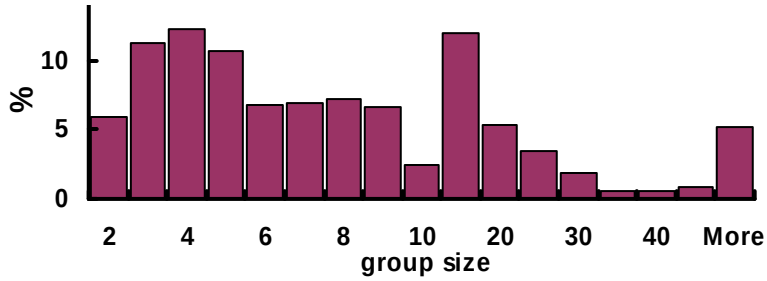

Figure 3-9: Distribution of homology group size

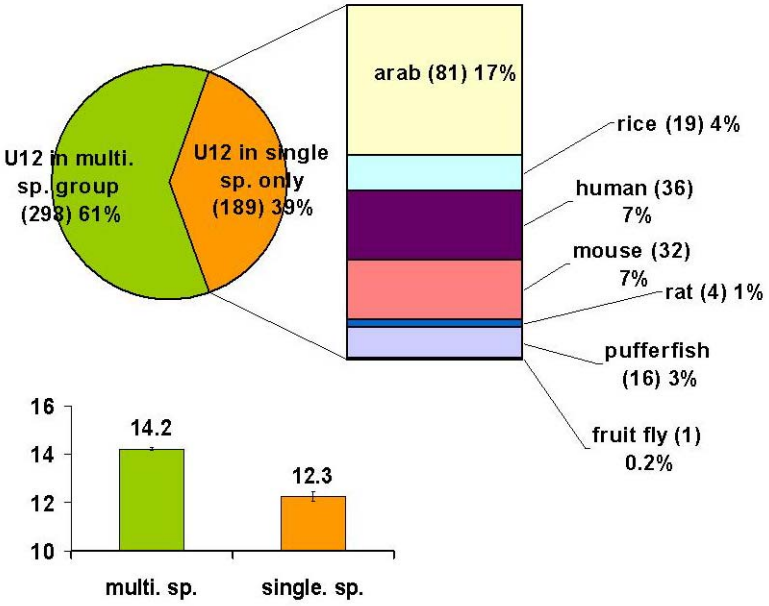

Figure 3-10: Proportion of clustered homology groups of U12-hosting genes.

A, "multi. sp." group represents identified U12-depedent introns present in multiple spices. "single sp." group represents identified U12-depedent introns present in only one spices. Proportions of "single sp" are also shown for individual species. Abbreviation of Arabidopsis is "arab" . B, Comparisons of average values of (5'SS score x BPS score) between "multi. sp" group and "single sp" group.

In most cases (219 out of 298) U12 introns are conserved across mammalian taxa, with 118 groups where the conservation even extends to fugu. The majority of U12-

containing-genes identified in the human, mouse and rat genomes turned out to be homologous to one another. There are 20 groups where U12 introns are conserved across animal and plant kingdoms. For 12 groups, U12-introns are plant-specific. Other patterns of conservation were observed but at much lower frequencies (Table 3-10).

Table 3-10: Presence/Absence of U12-dependent introns in different taxonomical lineages

Mammals*	Pufferfish	Fruit fly	Plants**	# group
+	+	+	+	1
+	+	+	-	5
+	+	-	+	10
+	+	-	-	116
+	-	+	-	3
+	-	-	+	8
+	-	-	-	134
+	+	-	+	1
-	-	-	+	11
			total	289

group, numbers of homologous gene groups that show presence (+)/absence (-) of U12-dependent introns in different organisms.
* U12-dependent introns are present in any of human, mouse, or rats.
** U12-dependent introns are present in Arabidopsis or rice.

U12 intron position and phase

Multiple nucleotide sequence alignments were manually inspected for 252 groups except 37 low-bootstrap-value groups. 221 groups (see below for the other 31 groups) include a total of 1064 U12 introns, for 1003 of which intron position and phase could be

determined unambiguously. Within each of the 221 (97.23%) groups all U12 introns reside at the same position (flanked by homologous exons). For the remaining 2.72%, U12 introns are found at different positions in the homologous genes. Regarding intron phase, for 92.63% of the groups, all U12 intron have the same phase; 5.99% and 1.38% share two and three phases respectively. Interestingly, despite the scarcity of U12 intron, 77 genes were found to harbor more than one U12 intron. Seven genes even appear to contain three U12 introns. Such multi-U12 genes are found in almost every species except rice and fruit fly (Table 2), and are distributed in 35 groups. Four groups were excluded from this analysis for at least one of the following reasons: 1) more than 100 members, 2) low (lower than 30) bootstrap values in the phylogenetic tree, 3) U12 introns exist in one species only. 31 groups include a total of 215 U12-hosting genes, for 205 of which intron position and phase of U12 introns were determined. Only 16 U12 introns are not located at the conserved position where other U12 introns reside.

Discussion

Uneven Distribution of U12-introns among Lineages

The numbers of U12-introns identified in seven organisms reveal an uneven distribution of U12-introns among lineages. There seems to be a gradient of U12-intron abundance with the highest number and frequency in mammals, decreasing in fish, and further decreasing in plants. However, there is a discontinuity in the fruitfly genome where only 16 U12-introns were detected. Not only the rarity is worth noting; what is

even more intriguing is that the proportion (~1%) of U12-candidates – those introns with at least ATCC at position +3 to +6 of the 5' ends but not necessarily with a detectable BPS near the 3' end - in fruit fly's introns is not particularly lower than that of the other species. Out of 545 U12-candidates, only 17 (15 U12-introns plus 2 cryptic-U12-introns) contain both recognizable U12-type 5' SS and BPS. It seems as if the consensus (among higher Eukaryotes) BPS is underrepresented in fruit fly compared to other U12-containing genomes. To make sure there were no systematic errors in the pipeline (more about this issue below), we compared our results with another study that has been done in Drosophila genomes. Four out of 19 U12-introns analyzed by Schneider et al. (2004) are not found in our list while one of our 16 U12-introns is not in theirs. The four that are missing in our list were indeed among our U12-candidates. However, they were not classified as U12-type because a proper BPS - another import hallmark of U12-introns – was not detected at the 3' end. The highest score obtained for the sequence resembling the consensus BPS was low. One previously reported U12-intron, which is known for being a sole intron of one gene, is also missing from our detected U12-intron set because it has been misannotated (see below).

The pattern of taxanomical distribution of U12-introns shown here matches the pattern observed in the divergence of U12-type spliceosome's major components. On the one hand, U11 and U12 snRNAs perform initial recognition of U12-introns by base-pairing with the pre-mRNA, and later on U12 snRNA interact with other snRNAs and proteins to form a catalytic core so as to remove the targeted intron. Thus, they are believed to be the essential molecules in the U12-spliceosome. Remarkably, for both snRNAs, Drosophila appears to be at an unusual phylogenetic position with plants being

closer to vertebrates (Schneider et al, 2004; Lorkovic et al, 2005). *Drosophila* U11 snRNA is so divergent that it even has been proposed to be nonhomologous to that of vertebrates and plants (Schneider *et al.*, 2004). On the other hand, from the perspective of the reacting substrate, in addition to its exceptional scarcity, *Drosophila* U12-introns were found to posses an A-rich region immediate downstream of 5' SS, which is not seen in human and *Arabidopsis* U12-introns. Given that *Drosophila* U11 snRNA is significantly different from the human and plant ones, this A-rich region is believed to assist in binding of U11 by recruiting proteins of U11 snRNP or other transacting factors. Similar mechanism has been characterized in yeast and human (Ref 25, 26 in Schneider's) U2-type splicing, where U-rich patches downstream of 5' SS facilitates the binding the U1 snRNP. Experiments on Rous sarcoma virus (RSV) negative regulator of splicing (NRS) demonstrated that, with the presence of G-rich sequences, U11 snRNP binds much more efficiently to NRS's pseudo 5' SS than to that of human's U12 introns in both P120 and SCN4A genes (McNally *et al.*, 2004). Another less obvious but note-worthy phenomenon about *Drosophila* introns is that the proportion of *Drosophila* introns having conserved U12-type 5' SS is about the same as the other species, but only very few of them (16 out of 545) have the conserved U12-type BPS as well. The conserved U12 5' SS and BPS are thought to be consensus to most U12-containing organisms because those known U12-introns (mostly from human and mouse) have such motifs. One can speculate that with the assistance of the A-rich patches, better association of U11 snRNP or even U11/U12-di-snRNP to the 5' SS relaxes the constraints on the BPS. Thus, BPSs are more degenerate in *Drosophila*, not as conserved as they are in

other species. In other words, some of the Drosophila U12-candidates that currently are not classified as U12-introns might be in fact U12-introns.

The highly divergent key snRNPs indicate that *Drosophila* splicing system might have undergone a unique evolutionary history. Our comparison of U12-intron abundance in different organisms supports this view. However, the evidence that the A-rich region downstream of 5' SS helps recruit the highly divergent U11 snRNA remains to be found.

Nonrandom Distribution of U12-introns and Models of U12-intron Origin

Given that U12-introns are so rare, one would expect that they sporadically dispersed in the genome. However, there is still a number of genes with multiple U12-introns found in Arabidopsis, human, mouse, rat and fugu genomes. For these organisms, binomial-distribution type of statistical tests for distribution of U12-introns all rejected the null hypothesis – random distribution of U12-introns. Thus, not only uneven among lineages, distribution of U12-introns within one genome appears to be nonrandom, too. With a much smaller set of U12-containing genes pooled together from various species Burge *et al.* (1998) carried out the same statistical analysis to test competing hypotheses of origin of U12-intron.

In spite that debates are not completely settled yet, the hypothesis that spliceosomal introns, which presumably were referred to U2-introns, evolved from group II introns is receiving greater and greater acceptance. Now that with the identification of the second (minor- /U12-) type spliceosomal introns, the nearly solved issue of origins of spliceosomal introns becomes even more confounded. Burge *et al.* (1998) summarized

possible evolutionary paths of two spliceosomes into three models. the first model (parasitic invasion) states that U12- and U2-spliceosomes are nonhomologous - both emerged from group II introns. The progenitor of the former entered a genome that has already been invaded by another group II intron, which has given rise to U2-spliceosome. Second, the codivergence model, two spliceosomes were resulted from genome duplication. Third, what they called fission-fusion model hypothesizes one spliceosomes speciated into two and somehow, possibly by endosymbiosis, merged in one genome later on.

Assuming that introns randomly inserted, first and second models will predict a random distribution of U12-introns. With the results of the binomial-distribution test, which rejected the null hypothesis, they maintained that the first two models were rejected, and in turn, the fission-fusion model was supported. However, such inference is not without questions. As they proposed in the same study, due to U12-U2 intron conversion is unidirectional, U12-introns are gradually lost from the genome, by converting to U2-introns. That is, the (conventionally denoted as p or q in a binomial distribution) has been decreasing. If we rolled back to ancient evolutionary time, would the frequency of U12-introns have been greater than what we observed today? Can the binomial-distribution statistical test still reject the null hypothesis? With inconstant frequency of U12- and U2-introns, binomial-distribution statistical test seems insufficient to address the issue of origins of U12-type splicing machinery. Perhaps, a computer simulation of the dynamics between two types of introns may give a better picture.

CHAPTER 4

SPLICEOSOMAL SMALL NUCLEAR RNA GENES IN 11 INSECT GENOMES

Stephen M. Mount[1,2,5], Valer Gotea[3,5], Chiao-Feng Lin[3], Kristina Hernandez[3], and Wojciech Makalowski[3,4]

[1] Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, Maryland 20742-5815, USA

[2] Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland 20742, USA

[3] Institute of Molecular Evolutionary Genetics and Department of Biology, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

[4] Department of Computer Science and Engineering, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

[5] These authors contributed equally to this work.

Previously published in *RNA* (2007), 13:5-14.

The piece of work presented in this chapter is collaboration between Mount's lab (University of Maryland, College Park) and Makalowski's lab. For the project, I computationally identified U12-type introns in 16 metazoan genomes (*Anopheles gambiae, Apis mellifera, Bos taurus, Canis familiaris, Ciona intestinalis, Drosophila*

melanogaster, Danio rerio, Fugu rubripes, Gallus gallus, Homo sapiens, Monodelphis domestica, Mus musculus, Pan troglodytes, Rattus norvegicus, Tetraodon nigroviridis, and Xenopus tropicalis). The methods used are based on Ensemble annotations and differ from those in CHAPTER 2.

Abstract

The removal of introns from the primary transcripts of protein-coding genes is accomplished by the spliceosome, a large macromolecular complex of which small nuclear RNAs (snRNAs) are crucial components. Following the recent sequencing of the honeybee (*Apis mellifera*) genome, we used various computational methods, ranging from sequence similarity search to RNA secondary structure prediction, to search for putative snRNA genes (including their promoters) and to examine their pattern of conservation among 11 available insect genomes (*A. mellifera*, *Tribolium castaneum*, *Bombyx mori*, *Anopheles gambiae*, *Aedes aegypti*, and six *Drosophila* species). We identified candidates for all nine spliceosomal snRNA genes in all the analyzed genomes. All the species contain a similar number of snRNA genes, with the exception of *A. aegypti*, whose genome contains more U1, U2, and U5 genes, and *A. mellifera*, whose genome contains fewer U2 and U5 genes. We found that snRNA genes are generally more closely related to homologs within the same genus than to those in other genera. Promoter regions for all spliceosomal snRNA genes within each insect species share similar sequence motifs that are likely to correspond to the PSEA (proximal sequence element A), the binding site for snRNA activating protein complex, but these promoter

elements vary in sequence among the five insect families surveyed here. In contrast to the other insect species investigated, Dipteran genomes are characterized by a rapid evolution (or loss) of components of the U12 spliceosome and a striking loss of U12-type introns.

Introduction

Protein-coding genes in most eukaryotes are interrupted by introns that must be removed in a process known as RNA splicing (Berget *et al.* 1977; Chow *et al.* 1977; Gilbert 1978). This complex process is carried out by the spliceosome, a large macromolecular assembly consisting of perhaps as many as 100 proteins (Jurica and Moore 2003; Hochleitner *et al.* 2005) and a set of five small nuclear RNAs (snRNAs) that are found complexed with proteins in the form of small nuclear ribonucleoproteins (snRNPs). Most introns are removed by the major spliceosome, which includes the U1, U2, U4, U5, and U6 snRNAs. Spliceosomal components, including the snRNAs, are highly conserved throughout eukaryotes (Mount and Salz 2000; Barbosa-Morais *et al.* 2006). The snRNAs are encoded by moderately repeated genes that show some variation within a species. Here we take advantage of the recently sequenced honeybee (Apis mellifera) (Honeybee Genome Sequencing Consortium 2006) and 10 other insect genomes to examine the pattern of conservation of spliceosomal snRNA genes within the class Insecta.

snRNAs and snRNA variants

Past work in *Drosophila melanogaster* has described variant U1 (Lo and Mount 1990) and U5 (Chen *et al.* 2005) snRNAs and their distinct patterns of expression throughout development. In each case, individual variants with distinct sequences showed unique patterns of developmental and tissue-specific expression. Similar observations have been made in pea (Hanley and Schuler 1991) and silk moth (Sierra-Montes *et al.* 2005). The differential expression of snRNA variants could be due to functional differences. However, differential expression could also be explained if sequence variants without functional significance are associated by chance with genes showing distinct patterns of expression driven by other factors (most likely, transcriptional regulation of the amount of snRNA produced during different developmental stages). Here we have addressed this question by examining the evolutionary stability of snRNA variants.

The minor spliceosome

In the case of ~0.2% of mammalian introns (Burge *et al.* 1998; Levine and Durbin 2001), splicing is carried out by a minor (U12) spliceosome, which contains U11, U12, U4atac, and U6atac in place of U1, U2, U4, and U6 snRNAs, respectively (Tarn and Steitz 1996). The U5 snRNP is shared between the major and minor spliceosomes (Luo *et al.* 1999; Schneider *et al.* 2002). The snRNPs that are unique to this U12 spliceosome contain, in addition to common core snRNP proteins, a number of unique proteins, most of which are homologous to proteins in the U2 spliceosome (Will *et al.* 2004). The U12 spliceosome was originally identified as required for the removal of introns carrying the

noncanonical AT-AC terminal dinucleotides in place of GT-AG (Hall and Padgett 1994; Tarn and Steitz 1996). It has since become clear that most U12 introns carry standard GT and AG dinucleotides (Dietrich *et al.* 1997; Levine and Durbin 2001), but that the U12 introns can nevertheless be recognized by distinct splice signal sequences (Sharp and Burge 1997). The U12 spliceosome appears to be ancient since U12 introns are found, together with genes for components of the minor spliceosome, in plants (Wu *et al.* 1996), insects, and chordates (Burge *et al.* 1998). However, many species (including Caenorhabditis elegans and most, if not all, unicellular eukaryotes) have lost the U12 spliceosomal machinery altogether.

The genome of *D. melanogaster* has a divergent U12 spliceosome, and was originally thought to lack U11 snRNA (Adams *et al.* 2000), but subsequent analysis including affinity purification revealed a highly divergent U11 snRNA (Schneider *et al.* 2004). We find that the divergence of the minor spliceosome appears to be limited to Diptera.

snRNA promoters

With the exception of U6 and U6atac, the spliceosomal snRNA genes are transcribed by RNA polymerase II and have a cap structure resembling that found on messenger RNAs but with a characteristic trimethylation of the cap guanosine. U6 and U6atac genes are transcribed by RNA polymerase III and have a distinct monomethyl phosphate cap. However, all snRNAs are abundant and are transcribed with the help of a

common multisubunit transcription factor known as the snRNA activating protein complex (SNAPc) (Hernandez 2001; Lai *et al.* 2005) that binds to the proximal sequence element A (PSEA). We have identified conserved sequences upstream of snRNAs in each of the insect species that are likely to correspond to the PSEA.

Results and Discussion

Putative spliceosomal snRNAs genes in the honeybee genome

We searched for snRNA genes in the honeybee genome using a combination of BLAST with customized parameters and covariance model approaches (see Materials and Methods). The honeybee genome has five putative U1 snRNA genes, three U2, two U4, three U5, three U6, and one putative gene for each of the minor spliceosomal snRNAs (Table 4-1). A putative U1 pseudogene located on Chromosome 1 (NCBI gi 63053347, position 4437-4276) carries two mutations in the essential conserved 5' terminus and lacks the conserved U1-specific start-site motif AAGC (which is present immediately upstream of all other putative U1 snRNA genes in the insect species examined here). However, this gene has the PSEA signal and relatively few other substitutions. We conclude that this gene was functional until recently and cannot exclude the possibility that it is still expressed and functional. In addition, there are four tandemly repeated U2 pseudogenes adjacent to the U2 snRNA gene on Chromosome 8. Each of them is missing 25 nucleotides (nt) from the 5' end, which affects formation of the first loop in the U2 snRNA secondary structure (see Supplemental Fig. S1 at

http://warta.bio.psu.edu/htt_doc/Projects/snRNA/). They also lack the PSEA signal upstream, which likely affects their transcription.

Table 4-1: Approximate (see Materials and Methods) gene numbers by snRNA type and species

TABLE 1. Approximate (see Materials and Methods) gene numbers by snRNA type and species

	U gene type/reference gene[a] length								
Species	U1 165	U2 192	U4 143	U5 124	U6 108	U11 275	U12 238	U4atac 160	U6atac 97
Drosophila melanogaster	5	6	3	7	3	1	1	1	1
Drosophila simulans	8	6	3	6	3	1	1	1	1
Drosophila sechellia	7	6	3	7	3	1	1	1	1
Drosophila yakuba	9	7	3	8	3	1	1	1	1
Drosophila pseudoobscura	7	8	2	7	3	1	1	1	1
Drosophila persimilis	7	7	3	7	3	1	1	1	1
Anopheles gambiae	8	7	1	7	2	2	1	1	1
Aedes aegypti	18	10	3	9	3	1	1	2	1
Bombyx mori	8	8	3	8	4	1	1	1	1
Apis mellifera	5	3	2	3	3	1	1	1	1
Tribolium castaneum	5	5	2	6	3	1	1	1	1

[a]*D. melanogaster* gene sequences were used as references and queries in BLAST searches.

The 20 snRNA genes in A. mellifera are spread across 11 chromosomes, two of them being located, however, in contigs yet unmapped. Because of this, there is little clustering of the snRNA genes, apart from two U1 genes that lie ~1 kb apart and within 100 kb of a third U1 gene on Chromosome 16. This is in contrast to *D. melanogaster*, which has four clusters of snRNA genes (Mount and Salz 2000), including one with two U2, one U4, and two U5 genes spread over 6 kb of Chromosome 2L.

A comparison of the number of snRNA genes among the 11 insect species analyzed (Table 4-1) reveals that U1, U2, and U5 genes tend to be the most abundant in each species, while U4 and U6 genes are intermediate in abundance. *A. mellifera* appears to be an exception to this rule, because it only has three copies for each of the U2 and U5 genes, just as many as there are U6 genes. The genes for snRNAs found in the minor spliceosome, U4atac, U6atac, U11, and U12, are single copy, with the exception of *Aedes*

aegypti U4atac and *Anopheles gambiae* U11, which have two copies each. Copy number polymorphism for U1 snRNA (five to seven genes) is documented in *D. melanogaster* (Lo and Mount 1990), and such minor variations in gene number are plausibly common for insect snRNA genes. These relative abundances of gene numbers reflect the overall abundance of snRNAs within cells (Mount and Steitz 1981).

Evolution of spliceosomal snRNA genes

As mentioned above, it has been shown that different U1 and U5 variants have tissue-specific expression patterns in D. melanogaster. We investigated the relationship between different variants by phylogenetic analysis to see if the pattern of conservation is consistent with expression differentiation. We found that neither the phylogenetic tree of U1 genes nor that of U5 genes clearly supports the functional differentiation of different variants. In fact, the trees of all snRNA genes reveal a pattern that is more consistent with a concerted mode of evolution or extreme purifying selection (Piontkivska *et al.* 2002; Nei and Rooney 2005). This is because different variants are more similar to variants within a genus than between genera (see a neighbor joining tree of U5 genes in Figure 4-1) suggesting that snRNA genes are being constantly homogenized by processes such as gene conversion, unequal crossover, or replacement through birth and death.

Even though the concerted mode of evolution appears as a convenient explanation at first, especially because it applies to other RNA gene families as well, it would require clustering of the genes that are being homogenized by gene conversion or unequal crossover (Nei and Rooney 2005). In the case of the spliceosomal snRNA genes, we

observe only partial clustering, mostly in the eight Dipteran species that have a small number of chromosomes as compared to the other five insect species included in this analysis. Therefore, the concerted mode of evolution can only partially explain the snRNA phylogenies. We investigated additional forces acting on these gene families by looking at their functional constraints.

The case of U6 snRNA genes is particularly interesting due to the high level of conservation observed across all species. All U6 genes are 108 nt long, of which only 13 are variable sites, the rest (95 nt; 88%) being perfectly conserved across all 33 U6 genes detected. It is remarkable that all *Drosophila* U6 genes are identical, with the exception of the *Drosophila yakuba* gene Dyak|U6|84681803|81773-81666, which differs only by 2 nt (singleton mutations) in the 5'-loop region and is otherwise perfectly conserved. Had we looked only at *Drosophila*, where the three U6 genes are found within a 1.5-kb region, the concerted evolution scenario would have seemed highly possible, in spite of protein-coding genes CG6643 and CG13624 flanking the triplet (U6:96Aa is, in fact, located in the last intron of CG6643). This scenario is contradicted, however, by low sequence conservation outside of genes and by the lack of clustering in non-Dipteran species, possibly due to the higher fragmentation of their genomes. And yet, the four *Bombyx mori* U6 genes are identical to each other, as are two of the three *Tribolium castaneum* genes, and two of the three *A. mellifera* genes (in the case of both *T. castaneum* and *A. mellifera*, only the first nucleotide is different in the third copy). Out of the 13 variable sites, four more are singletons that all belong to one *A. aegypti* gene (Aaeg|U6|78152160|82903-82796). Twelve of the 13 variable sites are concentrated in the 18-nt segment of the 5' end of the gene, the rest being almost invariant (one of the *A.*

aegypti singletons is found at position 43). This conservation pattern agrees well with expected stringent functional constraints of the U6 gene: it binds Lsm proteins and base pairs with U4 (Supplemental Fig. S2 at http://warta.bio.psu.edu/htt_doc/Projects/snRNA/), U2, and the donor splice sites of introns. Based on these facts, one can say that purifying selection rather than concerted evolution is the more likely explanation for the high conservation of U6 genes.

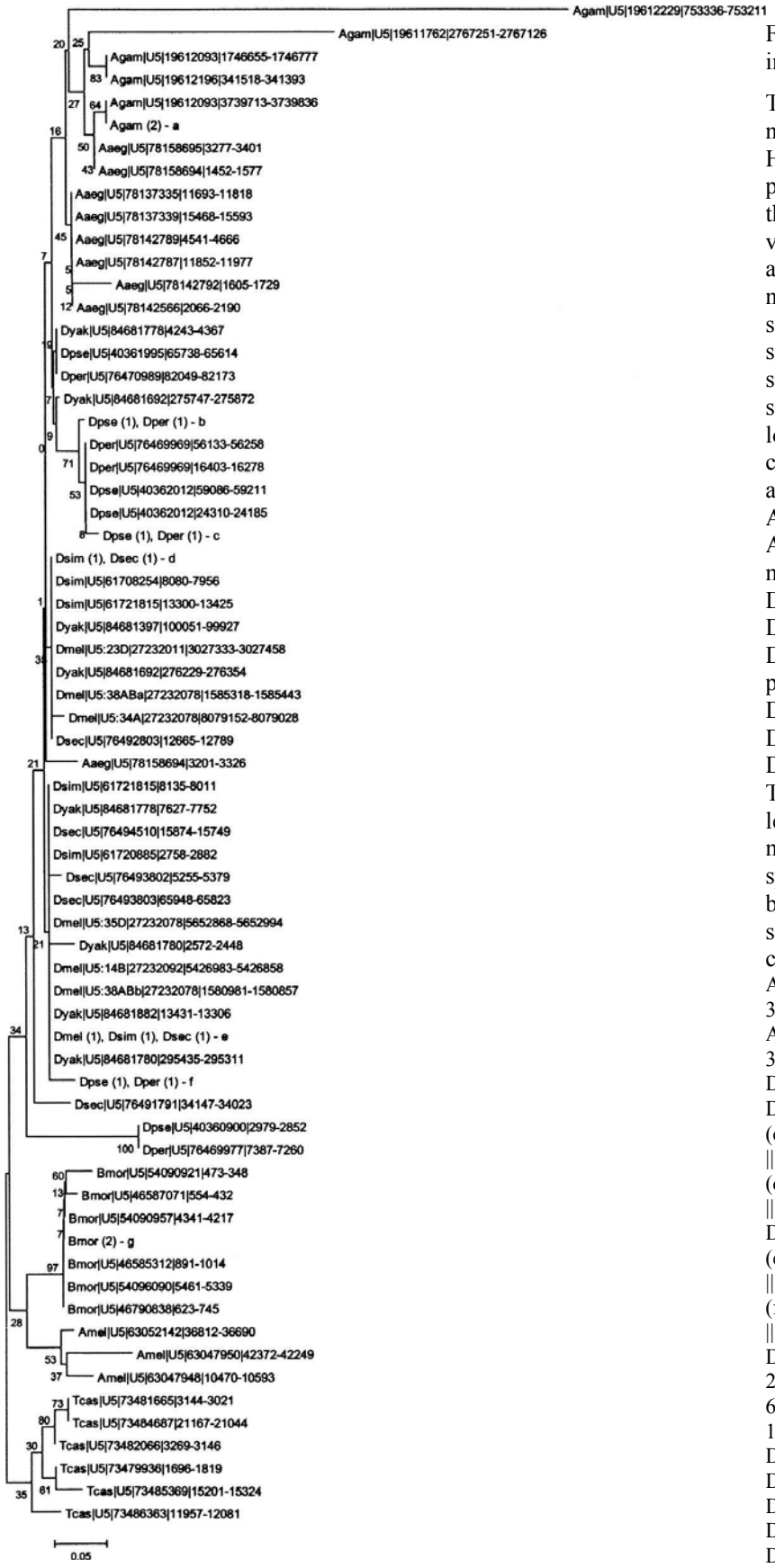

Figure 4-1: Phylogenetic tree of insect U5 snRNA molecules.

The tree was inferred using the neighbor-joining method. However, the maximum parsimony tree gave essentially the same topology. Bootstrap values based on 1000 replicas are given as percentage numbers. Leaf labels follow our standard name code: four-letter species abbreviation, type of snRNA, NCBI gi number of the sequences where the snRNA is located, and snRNA gene coordinates. Species abbreviations are as follow: Aaeg, Aedes aegypti; Agam, Anopheles gambiae; Amel, Apis mellifera; Bmor, Bombyx mori; Dmel, Drosophila melanogaster; Dper, Drosophila persimilis; Dpse, Drosophila pseudoobscura; Dsec, Drosophila sechellia; Dsim, Drosophila simulans; Dyak, Drosophila yakuba; and Tcas, Tribolium castaneum. Some leaves represent identical multiple sequences denoted by a species abbreviation, followed by the number of identical sequences and a single letter code. (a) Agam|U5|19612093|3732939-3732816 || Agam|U5|19612093|3738569-3738446; (b) Dsim|U5|61711086|937-1062 || Dsec|U5|76494084|16278-16403; (c) Dpse|U5|40362012|24310-24185 || Dpse|U5|55845750|73589-73714; (d) Dpse|U5|40362012|29122-29247 || |Dpse|U5|55845750|68777-68652 || Dper|U5|76469969|49677-49552; (e) Dpse|U5|40362012|59086-59211 || Dpse|U5|55845750|36653-36528; (f) Dpse|U5|40361995|65738-65614 || Dpse|U5|55845375|2085-2209; (g) Dmel|U5|27232088|2013728-2013604 || Dsim|U5|61738723|6567-6443 || Dsec|U5|76494655|18549-18673; (h) Dpse|U5|40362439|55864-55989 || Dpse|U5|55845417|57627-57502 || Dper|U5|76465975|1936-2061; (i) Dpse|U5|40360900|2979-2852 || Dpse|U5|55845704|65681-65554; (j) Bmor|U5|54108152|25278-25156 || Bmor|U5|54109135|218-340; (k) Bmor|U5|46790838|623-745 || Bmor|U5|54096763|3978-3856; (l) Bmor|U5|46585312|891-1014 || Bmor|U5|54046501|217-340 || Bmor|U5|54064195|209-332 || Bmor|U5|54081478|592-469.

Within the spliceosome, U6 snRNA pairs with U4 snRNA (Supplemental Fig. S1 at http://warta.bio.psu.edu/htt_doc/Projects/snRNA/), which, however, does not interact as intimately with the other components of the spliceosome as U6 does. Consequently, U4 genes are subject to more relaxed functional constraints. Interestingly, the U4 phylogeny reveals a mixed evolutionary pattern. Within the genus Drosophila, U4 genes appear to be subject to divergent evolution, as the orthology relationship between genes located in regions of conserved synteny is unambiguously solved by the phylogenetic tree. This indicates that U4 genes are not homogenized by gene conversion or unequal crossover in spite of their being located on the same chromosome (2L in the case of D. melanogaster). For the rest of the species, genes are more similar within a genus than among genera, which is more consistent with a mechanism of concerted evolution. The lack of conservation outside the genic region and the multichromosomal distribution of U4 genes (e.g., A. mellifera U4 genes are located on Chromosomes 1 and 15) indicate that gene conversion or unequal crossover is unlikely to homogenize U4 genes. Instead, it could be that a common functional constraint, such as interacting with U6 genes that are highly conserved (see above), determines all genes to evolve in a concerted manner. The same mechanism could very well apply to functional elements in the promoter region, such as the PSEA (see below).

In the case of the other major spliceosomal components, U1, U2, and U5, a complex of evolutionary forces appears to have generated the spectrum of different snRNA variants that we were able to detect (see U5 snRNA example in Figure 4-2). Higher clustering in Diptera allows gene conversion and unequal crossover to occur. For example, nine of the 18 A. aegypti U1 genes are separated by 1–3 kb, and six of the

seven U1 genes in *Drosophila pseudoobscura* form two three-gene clusters that are 17 kb apart. The potential of U1 for recombination events is demonstrated by a U1-mediated translocation event in Drosophila (Gonzalez *et al.* 2004). An increased number of U1 genes in A. aegypti (Table 4-1) and identification of pseudogenes for all snRNA types (data not shown) both agree with the death-and-birth model of evolution (Nei and Rooney 2005). So does duplication of genes, which can be observed in the case of head-to-head U2-U5 gene pairs found in the genus *Drosophila*. In the case of *D. melanogaster*, three such pairs are found on Chromosome 2L (two in region 38AB and one in region 34A) and one on Chromosome X (region 14B). Since no such pair appears to exist outside of the genus *Drosophila*, it is reasonable to believe they are the result of segmental duplication events rather than independent association of U2 and U5 genes in four genomic places. In spite of these apparent duplication events, the number of U2 and U5 genes in *Drosophila* is not significantly greater than in other species (Table 4-1), indicating that other genes were lost, in agreement with the birth-and-death model of evolution. In *Drosophila* we also observe conservation of certain variants across the genus. A good example is a U5 variant (U5:63BC in *D. melanogaster*) that has conserved the 3' end in all *Drosophila* species. It is not clear, however, if the conservation of this variant is due to functional constraints or because it is located in a region that does not favor recombination events, making it impossible for its sequence to be homogenized.

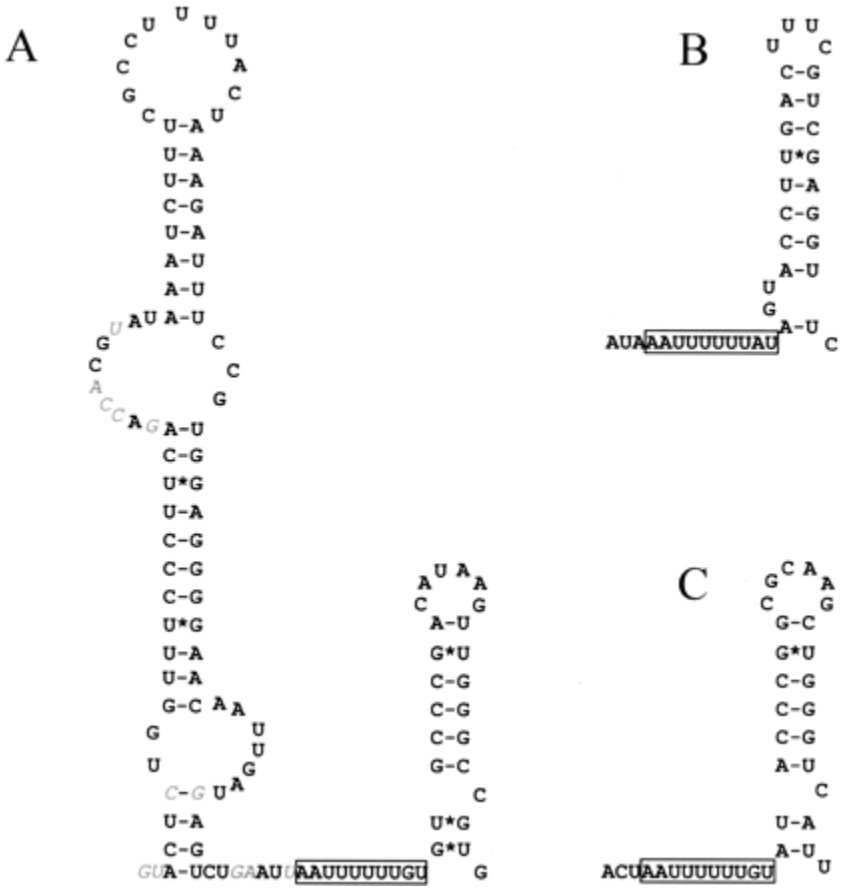

Figure 4-2: Secondary structure of A. mellifera U5 snRNA as predicted using the covariance model approach (see Materials and Methods for details). (A) snRNA structure based on the gene sequence found in Group 16, position 448405-448287; variable positions up to the Sm protein-binding site (boxed) are denoted in gray and italics. (B,C) 3' ends of two other genes found in the honeybee genome: Group 16, position 390994-390874 (B), and Group 3, position 3078080-3078200 (C).

In conclusion, we should re-emphasize that the evolution of spliceosomal snRNA is governed by several concurrent forces. Purifying selection is certainly one major force. In Dipteran species, which have a small number of chromosomes, concerted evolution by gene conversion or unequal crossover may play an important role as well. In non-Dipteran species, where snRNA genes do not form clusters (perhaps because of the higher fragmentation of genomes), concerted evolution by recombination is likely to be

extremely rare. It is thus possible that apparently concerted evolution results from coordinated changes in gene sequences as a result of selection and birth–death processes. This scenario could also explain why different species have slightly different PSEA consensi (see below), but it requires further investigations.

Small nuclear RNA promoters

The PSEA motifs upstream of snRNA genes are remarkably similar within species (Hernandez 2001). This is true for species as diverse as *Arabidopsis thaliana*, *D. melanogaster*, and human. We investigated the pattern of conservation of the PSEA elements within the 11 insect genomes available at the time of this study (see Materials and Methods and Figure 4-3). A conserved promoter motif was found in every species, and was shared among genes for all of the spliceosomal snRNAs. Sequence logos summarizing these conserved elements are presented in Figure 4-3. In each case, the motif is present at a similar distance of ~55 nt upstream of each gene. The honeybee PSEA signal can be represented by the consensus TTCTC (the underlined nucleotide denotes a variable position). All Drosophila species share a slightly longer conserved signal (ATTCCCAA), whereas the two mosquitoes share ATCGCTA (see Figure 4-3). It is important to note that in each case, the longer consensus contains a common TCNC motif. In the case of T. castaneum, however, the signal appears to be much less conserved, and MEME did not report any conserved motif at the expected position. A more careful analysis showed that Tribolium genes had the TCNC motif, but the flanking sequences are much less conserved, with the exception of U6 and U6atac genes

(Figure 4-3). Therefore, it is possible that the conservation of the short TCNC motif across all species and U gene types is due to functional constraints, being crucial for SNAPc binding, whereas the species-specific conservation of flanking residues is due to coevolution of the transcription factor with a small number of sites.

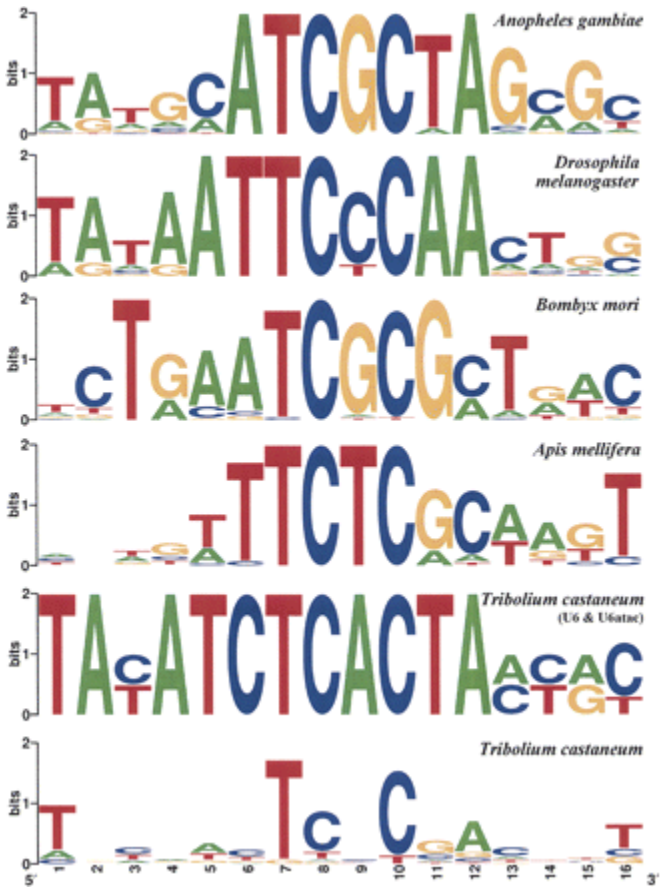

Figure 4-3: Sequence logos of insect snRNA gene promoters. They were created with WebLogo (Crooks *et al.* 2004) with the promoters corresponding to the genes presented in Table 1 for each species.

Divergence of the minor spliceosome within Diptera

The secondary structure of *Drosophila* U11 snRNA, a component of the minor spliceosome, was described as highly diverged from other known U11 sequences (Schneider *et al.* 2004), and we expected that the honeybee U11 would have a structure similar to that of Drosophila. To our surprise, we found that the structure of honeybee U11 resembles more closely that of human or plant U11 genes (Figure 4-4). In fact, the dramatic elongation of the second stem–loop of U11, which is a characteristic of *Drosophila* species, is not seen in either *Apis*, *Bombyx*, or *Tribolium*, each of which has a simple stem with 12 or fewer base pairs. Interestingly, the mosquito U11 genes contain stems of intermediate length in this position (the *A. gambiae* structure is shown in Figure 4-4). This finding suggests that the conservation of the minor spliceosome observed from plants to mammals was lost in Dipteran species.

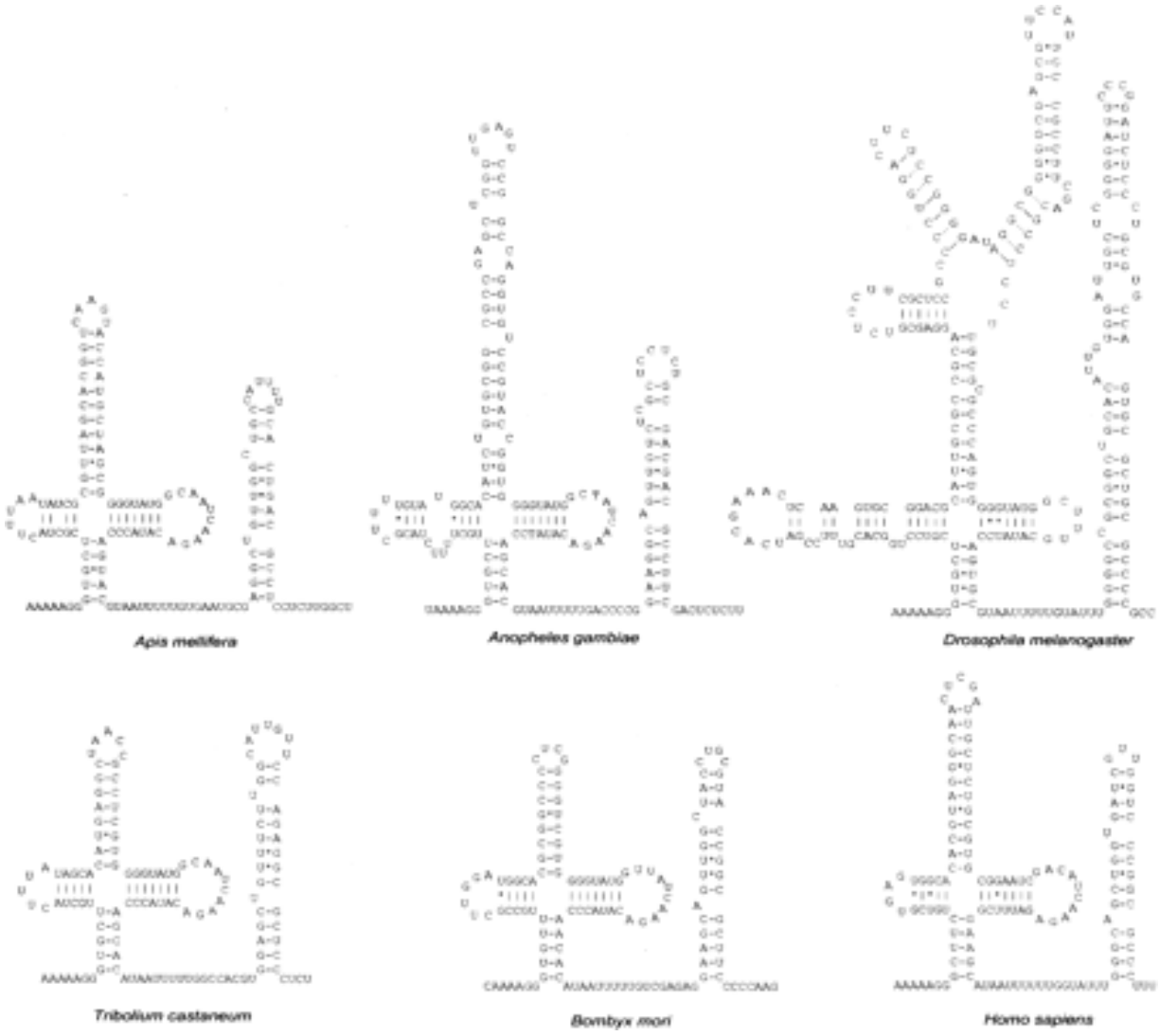

Figure 4-4: U11 snRNA secondary structures from five insect species and humans. *Drosophila* and human structures were redrawn after Schneider *et al.* (2004). All other structures were inferred based on covariance models as described in Materials and Methods.

This observation is reinforced by another feature of the divergent Drosophila U11 snRNP. Some of the U11/U12 snRNP proteins first characterized in the human U11/U12 snRNP (Will *et al.* 2004) and conserved broadly through eukaryotes (Lorkovic *et al.* 2005) are absent from Drosophila (Schneider *et al.* 2004). We searched for genes encoding these U11/U12 proteins in the 11 insect genomes using translated BLAST searches (Table 4-2). All six Drosophila species seem to lack the 31K and 35K proteins, yet these same genes are present in the other five insect species, including the two mosquito genomes. The 25K protein is also likely to be missing specifically from the dipteran species. Homologs can be identified in *Apis*, *Tribolium*, and *Bombyx* but not in *Drosophila* or mosquito species. However, because the 25K protein is not well conserved, this particular negative result is not compelling.

Table 4-2: Observation of identified U11/U12 snRNP proteins in various insect species

TABLE 2. Observation of identified U11/U12 snRNP proteins in various insect species

Protein	Query gi[a]	*Tribolium*	*Bombyx*	*Apis*	Mosquito	*Drosophlia*
20K	9506863	Yes	Yes	Yes	Yes	Yes
25K	13443018	Yes	Yes	Yes	No	No
31K	51243065	Yes	No	Yes	Yes	No
35K	5902144	Yes	Yes	Yes	Yes	No
65k	40538732	Yes	Yes	Yes	Yes	Yes

Either *D. melanogaster* (20K, 65K) or human (25K, 31K, and 35K) proteins were used as queries in a BLAST search (TBLASTN) against 11 insect genomes.
[a]NCBI's protein gi number.

U12-dependent introns

To investigate whether structural divergence of the U11/U12 snRNP is coupled with a loss of U12 introns, we examined the *A. mellifera* and 15 other metazoan genomes for the distribution of U12 introns. Using the statistical criteria described in Materials and Methods, we found that 57 out of 103,861 *A. mellifera* introns belong to the minor-type category, 49 of which are of the GT-AG type. These 57 introns are present in 56 genes, as one of the genes (ENSAPMT00000031526; a member of voltage-sensitive calcium channels) contains two U12 introns (introns 1 and 15) (Supplemental Table S1 at http://warta.bio.psu.edu/htt_doc/Projects/snRNA/). This large gene family is known to harbor many U12 introns (Wu and Krainer 1999), and our preliminary analysis suggests that the two introns present in the *Apis* gene are well conserved in different metazoan lineages (data not shown). The list of all the honeybee genes containing U12 introns is presented in Supplemental Table S2 at http://warta.bio.psu.edu/htt_doc/Projects/snRNA/. This count is fourfold higher than the number observed in Diptera (*D. melanogaster* and *A. gambiae*), but similar to that of the basal chordate *Ciona intestinalis* (Table 4-3).

Table 4-3: Number of genes and introns that have been analyzed in the genomes investigated, and results of identification of U12 introns

TABLE 3. Number of genes and introns that have been analyzed in the genomes investigated, and results of identification of U12 introns

Species name	Total number of genes	Total number of introns	Number of U12 introns	U12-intron containing genes	U12 intron frequency (%)
Anopheles gambiae	13,046	41,041	15	15	0.037
Apis mellifera	13,448	103,861	57	56	0.055
Ciona intestinalis	10,706	114,226	81	80	0.071
Drosophila melanogaster	11,399	48,251	14	14	0.029
Homo sapiens	21,190	202,715	422	395	0.208

Although the number of U12 introns in vertebrates is higher, this most likely reflects extensive gene duplications in the vertebrate lineage. The honeybee U12 introns are present in 55 unique gene families. Interestingly, while 41 of these gene families have at least one U12-intron-containing homolog in a vertebrate genome, only six of them have a homolog with a U12 intron in *D. melanogaster* (see Table 4-4). While these data are subject to minor errors due to inconsistencies in the annotation of genes between species, they indicate that >63% of the honeybee U12 introns are present in vertebrates but absent from *Drosophila*. Altogether, these data imply that most U12 introns present in the last common ancestor of flies, bees, and vertebrates have been lost in the Dipteran lineage.

One of the most interesting metazoan introns resides between exons 2 and 3 of the prospero gene of *D. melanogaster*. This U12 intron contains active splice sites for a U2 spliceosome that are alternatively used, an arrangement referred to as a "twintron." The alternative splicing of the prospero gene is temporally regulated during fly development (Scamborova *et al.* 2004). The U12 intron appears to be ancestral and is present in vertebrates, where it is not alternatively spliced (Oliver *et al.* 1993), and the U2 intron signals are not present. We examined the A. mellifera gene in order to explore the origins of the twintron arrangement. The Apis gene perfectly conserves the U12 splice sites, but not the U2 sites. Although it is possible that nonorthologous cryptic sites are used for U2 splicing in the bee (and ancestral insects), we hypothesize that the twintron arrangement is a recent development in the Dipteran lineage consistent with the switch from U12 to U2 splicing that must be occurring for other introns.

Table 4-4: Conservation pattern of homologous U12 containing genes among *A. mellifera* and other taxa

TABLE 4. Conservation pattern of homologous U12 containing genes among *A. mellifera* and other taxa

Apis mellifera	*Anopheles gambiae*	*Drosophila melanogaster*	*Ciona intestinalis*	*Homo sapiens*	Number of gene families
+	−	−	−	−	14
+	−	−	−	+	27
+	−	+	−	+	3
+	−	+	+	+	3
+	−	−	+	+	5
+	+	−	−	+	2
+	+	−	+	+	1
				Total	55

Homology information among U12-containing genes was based on each gene's gene family ID, assigned by ENSEMBL. (+/−) Presence/absence of U12-containing genes in that taxon. Each row represents a pattern where U12-containing genes are present in a certain combination of taxa. The last column is the number of gene families that shows that particular pattern.

Coordinated divergence of the U12 spliceosome and loss of U12-dependent introns

Representative Lepidopteran (*Bombyx*), Coleopteran (*Tribolium*), and Hymenopteran (*Apis*) genomes all conserve genes for components of the minor spliceosome that are missing in *Drosophila*. In addition, the honeybee genome has many more U12 introns than do Dipteran (fly or mosquito) species, indicating that divergence of the U12 spliceosome in Diptera is associated with the loss of U12 introns. These coordinated changes indicate that U12 introns are being lost from genomes that remove them inefficiently, as has been described for *Drosophila* (Patel *et al.* 2002). For a number of other genes, we have noted that bee U12 introns are either cleanly lost or replaced by

U2 introns. In the case of prospero, the twintron arrangement may represent an evolutionary intermediate where a poorly spliced U12 intron serves as an alternative to a poorly placed U2 intron. The *Drosophila* genome has a total of only 14 U12 introns, and the genes for three U11/U12 spliceosomal proteins have been lost. It is tempting to speculate that this state represents an intermediate on the path toward the complete loss of U12 introns observed for species such as *C. elegans*.

Materials and Methods

Sequence data

The genome assembly Amel_3.0 provided by the Human Genome Sequencing Center at Baylor College of Medicine (ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Amellifera/) was used for finding and annotating the honeybee snRNA genes and their promoters. For interspecies comparison, NCBI's "BLAST with arthropoda genomes" server was used (http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi?organism=insects). At the time of this analysis (January 2006), 11 insect genomes were available: *Aedes aegypti*, *Anopheles gambiae* str. PEST*, *Apis mellifera**, *Bombyx mori*, *Drosophila melanogaster**, *Drosophila persimilis*, *Drosophila pseudoobscura*, *Drosophila sechellia*, *Drosophila simulans*, *Drosophila yakuba*, and *Tribolium castaneum* (*denotes completed genomic sequence).

Annotation of the snRNA genes

The annotation of the honeybee snRNA genes involved two steps. First, sequences of all snRNAs (U1, U2, U4, U4atac, U5, U6, U6atac, U11, and U12) of D. melanogaster were used as queries against the A. mellifera genome assembly 3.0. NCBI's BLAST was used with the following parameters: -r 5 -q -4 -G 10 -E 6 -W 7 -FF -X 40 -y 20 -Z 100 -e 0.1. In the next step, the nucleotide sequences around each hit were extracted and used as input for the INFERNAL software (Eddy 2002).

The snRNA gene number in 11 insect genomes was estimated based on BLAST hits using Drosophila snRNAs as queries and NCBI's insect genomes server (http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi?organism=insects). The BLAST parameters used for finding Apis genes were used again here. In the case of very short hits or no hit at all, other insect snRNA sequences were used as queries. Functionality of the determined genes was assessed by the integrity of the gene and by the presence of the PSEA element at the expected distance from the transcription start site of a gene.

Phylogenetic analysis

We investigated the mode of evolution of spliceosomal snRNA genes by constructing a phylogenetic tree of genes specifying each of the U snRNAs found in the major spliceosome. For that purpose, we aligned all genes assessed to be functional (see above) using ClustalW (Thompson *et al.* 1994) and adjusted the alignment manually where necessary. In the case of U5 genes, we used the secondary structure of the second stem–loop as a guide for aligning the variable 3' end for Drosophila sequences. The

neighbor-joining method with 1000 bootstrap replicates was then used to construct the trees.

snRNA secondary structure prediction

snRNA secondary structures were drawn based on the INFERNAL alignment with manual adjustment when necessary. In some cases, for example, the Anopheles U11 snRNA, the structure was drawn with the aid of Mfold modeling software (http://www.bioinfo.rpi.edu/applications/mfold/old/rna/; Zuker 2003) run on fragmented sequence. Simply, the sequence was divided into known functional domains (defined by conserved segments), and domains were folded separately.

Detection of proximal sequence element A (PSEA)

One hundred nucleotides upstream of each snRNA gene were subjected to an initial motif search with MEME software (http://meme.sdsc.edu/meme/intro.html; Bailey and Elkan 1994). Next, every promoter sequence was analyzed individually to overcome the limitations of MEME software, given that the expected conserved motifs are short. Sequences of the promoters from genes that are believed to be functional (Table 1) were used to construct the PSEA profiles using WebLogo (Crooks *et al.* 2004).

U12-dependent introns

Intron position information for 16 metazoan genomes (*Anopheles gambiae, Apis mellifera*, *Bos taurus*, *Canis familiaris*, *Ciona intestinalis*, *Drosophila melanogaster*, *Danio rerio*, *Fugu rubripes*, *Gallus gallus*, *Homo sapiens*, *Monodelphis domestica*, *Mus musculus*, Pan troglodytes, Rattus norvegicus, *Tetraodon nigroviridis*, and *Xenopus tropicalis*) was downloaded from Ensembl database version 36, December 2005 (http://www.ensembl.org/index.html). Each intron was represented by 200-nt sequence windows centered at the 5'- and 3'-splice junctions, respectively. All introns containing an ATCC string at position +3 from the 5'-splice junction were selected as U12 intron candidates. Noncandidate introns with R (A or G) at +1 and G at +5 were used as a U2 intron training set. The U12 intron training set was composed of 46 U12 introns from a variety of higher eukaryotes (Burge *et al.* 1998). Weight matrices of the 5'-splice site and branch point site (BPS) for both U12 and U2 introns were created from the two training sets, and were used to calculate log-odds ratios for all candidates and introns in the two training sets. The mean and standard deviation of log-odds ratio values were calculated over all training sequences for the 5'-splice site and BPS, respectively, and then were used to normalize the log-odds ratios of all introns that have been evaluated. Normalized BPS scores of the training sequences were plotted, and since they followed a normal distribution, we chose a Z value of 2.31, which corresponds to a 99% confidence interval, as a threshold BPS score to classify U12 intron candidates. Two additional criteria were applied: at least one adenine must be present at the branch point site and the distance from the branch point to the 3' junction has to be between 9 and 36 nt. Homology

information of U12-containing genes was inferred based on the Ensembl gene family assignment, which is defined by the Markov clustering algorithm described by Enright *et al.* (2002).

Acknowledgements

C.-F.L. and V.G. were partially supported by the Center for Comparative Genomics and Bioinformatics, part of Penn State's Huck Institutes of the Life Sciences. S.M.M. was partially supported by NSF award 0544309.

BIBLIOGRAPHY

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science 287*:2185-2195.

Alioto TS. 2007. U12DB: a database of orthologous U12-type spliceosomal introns. *Nucl Acids Res 35*:D110-115.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol 215*:403-410.

Anderson PAV, Greenberg RM. 2001. Phylogeny of ion channels: clues to structure and function. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology 129*:17-28.

Bailey T, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*: AAAI Press, Menlo Park, California. pp 28-36.

Barbosa-Morais NL, Carmo-Fonseca M, Aparicio S. 2006. Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion. *Genome Res 16*:66-77.

Basu MK, Makalowski W, Rogozin IB, Koonin EV. 2008. U12 intron positions are more strongly conserved between animals and plants than U2 intron positions. *Biol Direct 3*:19.

Berget SM. 1995. Exon Recognition in Vertebrate Splicing. *J Biol Chem 270*:2411-2414.

Berget SM, Moore C, Sharp PA. 1977. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci U S A 74*:3171-3175.

Blencowe BJ. 2000. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends in Biochemical Sciences 25*:106-110.

Bonini NM, Bui QT, Gray-Board GL, Warrick JM. 1997. The *Drosophila* eyes absent gene directs ectopic eye formation in a pathway conserved between flies and vertebrates. *Development 124*:4819-4826.

Breathnach R, Chambon P. 1981. Organization and expression of eucaryotic split genes coding for proteins. *Annual Review of Biochemistry 50*:349-383.

Burge CB, Padgett RA, Sharp PA. 1998. Evolutionary fates and origins of U12-type introns. *Mol Cell 2*:773-785.

Carmel L, Wolf YI, Rogozin IB, Koonin EV. 2007. Three distinct modes of intron dynamics in the evolution of eukaryotes. *Genome Res 17*:1034-1044.

Catterall WA. 1988. Structure and function of voltage-sensitive ion channels. *Science 242*:50-61.

Chen L, Lullo DJ, Ma E, Celniker SE, Rio DC, Doudna JA. 2005. Identification and analysis of U5 snRNA variants in *Drosophila*. *RNA 11*:1473-1477.

Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol 3*:e314.

Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, *et al*. 2002. The Draft Genome of *Ciona intestinalis*: Insights into Chordate and Vertebrate Origins. *Science 298*:2157-2167.

Dietrich RC, Fuller JD, Padgett RA. 2005. A mutational analysis of U12-dependent

splice site dinucleotides. *RNA 11*:1430.

Dietrich RC, Incorvaia R, Padgett RA. 1997. Terminal intron dinucleotide sequences do not distinguish between U2- and U12-dependent introns. *Mol Cell 1*:151.

Drysdale RA, Crosby MA. 2005. FlyBase: genes and gene models. *Nucleic Acids Res 33*:D390-395.

Eddy SR. 2002. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics 3*:18.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res 30*:1575-1584.

Frilander MJ, Steitz JA. 1999. Initial recognition of U12-dependent introns requires both U11/5' splice-site and U12/branchpoint interactions. *Genes Dev 13*:851-863.

Gilbert W. 1978. Why genes in pieces? *Nature 271*:501.

Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, *et al*. 2003. A Protein Interaction Map of *Drosophila melanogaster*. *Science 302*:1727-1736.

Hall SL, Padgett RA. 1994. Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites. *J Mol Biol 239*:357-365.

Hall SL, Padgett RA. 1996. Requirement of U12 snRNA for in Vivo Splicing of a Minor Class of Eukaryotic Nuclear Pre-mRNA Introns. *Science 271*:1716-1718.

Hankeln T, Friedl H, Ebersberger I, Martin J, Schmidt ER. 1997. A variable intron distribution in globin genes of Chironomus: evidence for recent intron gain. *Gene* 205:151-160.

Hanley BA, Schuler MA. 1991. Developmental expression of plant snRNAs. *Nucleic Acids Res 19*:6319-6325.

Hastings ML, Resta N, Traum D, Stella A, Guanti G, Krainer AR. 2005. An LKB1 AT-AC intron mutation causes Peutz-Jeghers syndrome via splicing at noncanonical cryptic splice sites. *Nat Struct Mol Biol 12*:54-59.

Hernandez N. 2001. Small nuclear RNA genes: a model system to study fundamental mechanisms of transcription. *J Biol Chem 276*:26733-26736.

Hochleitner EO, Kastner B, Frohlich T, Schmidt A, Luhrmann R, Arnold G, Lottspeich F. 2005. Protein stoichiometry of a multiprotein complex, the human spliceosomal U1 small nuclear ribonucleoprotein: absolute quantification using isotope-coded tags and mass spectrometry. *J Biol Chem 280*:2536-2542.

Hughes AL, Friedman R. 2005. Loss of ancestral genes in the genomic evolution of *Ciona intestinalis*. *Evolution & Development 7*:196-200.

Huizing M, Didier A, Walenta J, Anikster Y, Gahl WA, Kramer H. 2001. Molecular cloning and characterization of human VPS18, VPS 11, VPS16, and VPS33. *Gene 264*:241-247.

Iwamoto M, Maekawa M, Saito A, Higo H, Higo K. 1998. Evolutionary relationship of plant catalase genes inferred from intron-exon structures: isozyme divergence after the separation of monocots and dicots. *Theor Appl Genet* 97:9-19.

Iwamoto M, Nagashima H, Nagamine T, Higo H, Higo K. 1999. p-SINE1-like intron of the CatA catalase homologs and phylogenetic relationships among AA-genome Oryza and related species. *Theor Appl Genet* 98:853-861.

Jackson IJ. 1991. A reappraisal of non-consensus mRNA splice sites. *Nucleic Acids Res*

19:3795-3798.

Jurica MS, Moore MJ. 2003. Pre-mRNA splicing: awash in a sea of proteins. *Mol Cell 12*:5-14.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler, David. 2002. The Human Genome Browser at UCSC. *Genome Res 12*:996-1006.

Kolossova I, Padgett RA. 1997. U11 snRNA interacts in vivo with the 5' splice site of U12-dependent (AU-AC) pre-mRNA introns. *Rna 3*:227-233.

Kortschak RD, Samuel G, Saint R, Miller DJ. 2003. EST Analysis of the Cnidarian *Acropora millepora* Reveals Extensive Gene Loss and Rapid Sequence Divergence in the Model Invertebrates. *Current Biology 13*:2190-2195.

Lai HT, Chen H, Li C, McNamara-Schroeder KJ, Stumph WE. 2005. The PSEA promoter element of the *Drosophila* U1 snRNA gene is sufficient to bring DmSNAPc into contact with 20 base pairs of downstream DNA. *Nucleic Acids Res 33*:6579-6586.

Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE. 2007. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature 446*:926-929.

Levine A, Durbin R. 2001. A computational scan for U12-dependent introns in the human genome sequence. *Nucl Acids Res 29*:4006-4013.

Lo PC, Mount SM. 1990. *Drosophila melanogaster* genes for U1 snRNA variants and their expression during development. *Nucleic Acids Res 18*:6971-6979.

Lopez MD, Alm Rosenblad M, Samuelsson T. 2008. Computational screen for spliceosomal RNA genes aids in defining the phylogenetic distribution of major

and minor spliceosomal components. *Nucl Acids Res*:gkn142.

Lorkovic ZJ, Lehner R, Forstner C, Barta A. 2005. Evolutionary conservation of minor U12-type spliceosome between plants and humans. *Rna 11*:1095-1107.

Luo HR, Moreau GA, Levin N, Moore MJ. 1999. The human Prp8 protein is a component of both U2- and U12-dependent spliceosomes. *Rna 5*:893-908.

Lynch M, Richardson AO. 2002. The evolution of spliceosomal introns. *Curr Opin Genet Dev 12*:701-710.

Maglott D, Ostell J, Pruitt KD, Tatusova T. 2005. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res 33*:D54-58.

Malicki J, Schughart K, McGinnis W. 1990. Mouse Hox-2.2 specifies thoracic segmental identity in *Drosophila* embryos and larvae. *Cell 63*:961-967.

McLysaght A, Hokamp K, Wolfe KH. 2002. Extensive genomic duplication during early chordate evolution. *Nat Genet 31*:200.

Montzka KA, Steitz JA. 1988. Additional Low-Abundance Human Small Nuclear Ribonucleoproteins: U11, U12, Etc. *PNAS 85*:8885-8889.

Mount SM, Gotea V, Lin C-F, Hernandez K, Makalowski W. 2007. Spliceosomal small nuclear RNA genes in 11 insect genomes. *RNA 13*:5-14.

Mount SM, Salz HK. 2000. Pre-messenger RNA processing factors in the *Drosophila* genome. *J Cell Biol 150*:F37-44.

Mount SM, Steitz JA. 1981. Sequence of U1 RNA from *Drosophila melanogaster*: implications for U1 secondary structure and possible involvement in splicing. *Nucleic Acids Res 9*:6351-6368.

Nei M. 1969. Gene duplication and nucleotide substitution in evolution. *Nature 221*:40-

42.

Nei M, Rooney AP. 2005. Concerted and Birth-and-Death Evolution of Multigene Families. *Annu Rev Genet*.

Nguyen HD, Yoshihama M, Kenmochi N. 2005. New maximum likelihood estimators for eukaryotic intron evolution. *PLoS Comput Biol 1*:e79.

Ohno S. 1970. *Evolution by gene duplication*. New York: Springer Verlag.

Oliver G, Sosa-Pineda B, Geisendorf S, Spana EP, Doe CQ, Gruss P. 1993. Prox 1, a prospero-related homeobox gene expressed during mouse development. *Mech Dev 44*:3-16.

O'Neill RJW, Brennan FE, Delbridge ML, Crozier RH, Graves JAM. 1998. De novo insertion of an intron into the mammalian sex determining gene, SRY. *Proc Natl Acad Sci U S A* 95:1653-1657.

Parker R, Siliciano PG. 1993. Evidence for an essential non-Watson-Crick interaction between the first and last nucleotides of a nuclear pre-mRNA intron. *Nature 361*:660-662.

Patel AA, McCarthy M, Steitz JA. 2002. The splicing of U12-type introns can be a rate-limiting step in gene expression. *EMBO J 21*:3804-3815.

Patel AA, Steitz JA. 2003. Splicing double: insights from the second spliceosome. *Nature Rev Mol Cell Biol 4*:960.

Peixoto AA, Smith LA, Hall JC. 1997. Genomic Organization and Evolution of Alternative Exons in a *Drosophila* Calcium Channel Gene. *Genetics 145*:1003-1013.

Plummer NW, Meisler MH. 1999. Evolution and Diversity of Mammalian Sodium

Channel Genes. *Genomics 57*:323-331.

Raible F, Tessmar-Raible K, Osoegawa K, Wincker P, Jubin C, Balavoine G, Ferrier D, Benes V, de Jong P, Weissenbach J, Bork P, Arendt D. 2005. Vertebrate-Type Intron-Rich Genes in the Marine Annelid *Platynereis dumerilii*. *Science 310*:1325-1326.

Reed R. 1996. Initial splice-site recognition and pairing during pre-mRNA splicing. *Curr Opin Genet Dev 6*:215-220.

Rehwinkel JAN, Letunic I, Raes J, Bork P, Izaurralde E. 2005. Nonsense-mediated mRNA decay factors act in concert to regulate common mRNA targets. *RNA 11*:1530-1544.

Rosa To, Francisco JA, Francisco Rg-T. 2008. Alternative splicing: A missing piece in the puzzle of intron gain. *Proceedings of the National Academy of Sciences* 105:7223-7228.

Roy SW, Gilbert W. 2005. Complex early genes. *Proceedings of the National Academy of Sciences 102*:1986-1991.

Russell AG, Charette JM, Spencer DF, Gray MW. 2006. An early evolutionary origin for the minor spliceosome. *Nature 443*:863-866.

Saxena A, Ma B, Schramm L, Hernandez N. 2005. Structure-function analysis of the human TFIIB-related factor II protein reveals an essential role for the C-terminal domain in RNA polymerase III transcription. *Mol Cell Biol 25*:9406-9418.

Scamborova P, Wong A, Steitz JA. 2004. An intronic enhancer regulates splicing of the twintron of *Drosophila melanogaster* prospero pre-mRNA by two different spliceosomes. *Mol Cell Biol 24*:1855-1869.

Schneider C, Will CL, Brosius J, Frilander MJ, Luhrmann R. 2004. Identification of an evolutionarily divergent U11 small nuclear ribonucleoprotein particle in *Drosophila*. *Proc Natl Acad Sci U S A 101*:9584-9589.

Schneider C, Will CL, Makarova OV, Makarov EM, Luhrmann R. 2002. Human U4/U6.U5 and U4atac/U6atac.U5 tri-snRNPs exhibit similar protein compositions. *Mol Cell Biol 22*:3219-3229.

Schwartz S, Silva J, Burstein D, Pupko T, Eyras E, Ast G. 2008. Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Res 18*:88-103.

Screaton GR, Caceres JF, Mayeda A, Bell MV, Plebanski M, Jackson DG, Bell JI, Krainer AR. 1995. Identification and characterization of three members of the human SR family of pre-mRNA splicing factors. *EMBO J 14*:4336-4349.

Sharp PA. 1994. Split genes and RNA splicing. *Cell 77*:805-815.

Sharp PA. 2005. The discovery of split genes and RNA splicing. *Trends in Biochemical Sciences 30*:279-281.

Sharp PA, Burge CB. 1997. Classification of introns: U2-type or U12-type. *Cell 91*:875-879.

Shen H, Green MR. 2007. RS domain-splicing signal interactions in splicing of U12-type and U2-type introns. *Nat Struct Mol Biol 14*:597-603.

Sheth N, Roca X, Hastings ML, Roeder T, Krainer AR, Sachidanandam R. 2006. Comprehensive splice-site analysis using comparative genomics. *Nucl Acids Res 34*:3955-3967.

Sierra-Montes JM, Pereira-Simon S, Smail SS, Herrera RJ. 2005. The silk moth *Bombyx*

mori U1 and U2 snRNA variants are differentially expressed. *Gene 352*:127-136.

Sureau A, Gattoni R, Dooghe Y, Stevenin J, Soret J. 2001. SC35 autoregulates its expression by promoting splicing events that destabilize its mRNAs. *EMBO J 20*:1785-1796.

Tarn WY, Steitz JA. 1996. Highly diverged U4 and U6 small nuclear RNAs required for splicing rare AT-AC introns. *Science 273*:1824-1832.

Tarn WY, Steitz JA. 1996. A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro. *Cell 84*:801-811.

Tronchre H, Wang J, Fu X-D. 1997. A protein related to splicing factor U2AF35 that interacts with U2AF65 and SR proteins in splicing of pre-mRNA. *Nature 388*:397.

Turunen JJ, Will CL, Grote M, Luhrmann R, Frilander MJ. 2008. The U11-48K protein contacts the 5' splice site of U12-type introns and the U11-59K protein. *Mol Cell Biol*:MCB.01928-01907.

Will CL, Luhrmann R. 2005. Splicing of a rare class of introns by the U12-dependent spliceosome. *Biol Chem 386*:713-724.

Will CL, Schneider C, Hossbach M, Urlaub H, Rauhut R, Elbashir S, Tuschl T, Luhrmann R. 2004. The human 18S U11/U12 snRNP contains a set of novel proteins not found in the U2-dependent spliceosome. *Rna 10*:929-941.

Will CL, Schneider C, Reed R, Luhrmann R. 1999. Identification of both shared and distinct proteins in the major and minor spliceosomes. *Science 284*:2003.

Wu HJ, Gaubier-Comella P, Delseny M, Grellet F, Van Montagu M, Rouze R. 1996. Non-canonical introns are at least 10(9) years old. *Nat Genet 14*:383-384.

Wu Q, Krainer AR. 1996. U1-mediated exon definition interactions between AT-AC and GT-AG introns. *Science 274*:1005-1008.

Wu Q, Krainer AR. 1997. Splicing of a divergent subclass of AT-AC introns requires the major spliceosomal snRNAs. *RNA 3*:586-601.

Wu Q, Krainer AR. 1998. Purine-rich enhancers function in the AT-AC pre-mRNA splicing pathway and do so independently of intact U1 snRNP. *Rna 4*:1664-1673.

Wu Q, Krainer AR. 1999. AT-AC pre-mRNA splicing mechanisms and conservation of minor introns in voltage-gated ion channel genes. *Mol Cell Biol 19*:3225-3236.

Yamaoka T, Hatada I, Kitagawa K, Wang X, Mukai T. 1995. Cloning and mapping of the U2af1-rs2 gene with a high transmission distortion in interspecific backcross progeny. *Genomics 27*:337-340.

Yu FH, Yarov-Yarovoy V, Gutman GA, Catterall WA. 2005. Overview of Molecular Relationships in the Voltage-Gated Ion Channel Superfamily. *Pharmacol Rev 57*:387-395.

Zamore PD, Green MR. 1991. Biochemical characterization of U2 snRNP auxiliary factor: an essential pre-mRNA splicing factor with a novel intranuclear distribution. *EMBO J 10*:207-214.

Zhu W, Brendel V. 2003. Identification, characterization and molecular phylogeny of U12-dependent introns in the *Arabidopsis thaliana* genome. *Nucl Acids Res 31*:4561-4572.

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res 31*:3406-3415.

Zuo P, Maniatis T. 1996. The splicing factor U2AF35 mediates critical protein-protein

interactions in constitutive and enhancer-dependent splicing. *Genes Dev 10*:1356-1368.

Appendix A

DETERMINANTS OF PLANT U12-DEPENDENT INTRON SPLICING EFFICIENCY

The Plant Cell, Vol. 16, 1340–1352, May 2004, www.plantcell.org © 2004 American Society of Plant Biologists

Determinants of Plant U12-Dependent Intron Splicing Efficiency

Dominika Lewandowska,[a,1,2] Craig G. Simpson,[b,1] Gillian P. Clark,[b] Nikki S. Jennings,[b] Maria Barciszewska-Pacak,[a] Chiao-Feng Lin,[c] Wojciech Makalowski,[c] John W.S. Brown,[b,3] and Artur Jarmolowski[a]

[a] Department of Gene Expression, Adam Mickiewicz University, Poznan 60-371, Poland
[b] Gene Expression Programme, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, Scotland, United Kingdom
[c] Institute of Molecular Evolutionary Genetics and Department of Biology, Pennsylvania State University, University Park, Pennsylvania 16802

Factors affecting splicing of plant U12-dependent introns have been examined by extensive mutational analyses in an in vivo tobacco (*Nicotiana tabacum*) protoplast system using introns from three different *Arabidopsis thaliana* genes: *CBP20*, *GSH2*, and *LD*. The results provide evidence that splicing efficiency of plant U12 introns depends on a combination of factors, including UA content, exon bridging interactions between the U12 intron and flanking U2-dependent introns, and exon splicing enhancer sequences (ESEs). Unexpectedly, all three plant U12 introns required an adenosine at the upstream purine position in the branchpoint consensus UCCUURAUY. The exon upstream of the *LD* U12 intron is a major determinant of its higher level of splicing efficiency and potentially contains two ESE regions. These results suggest that in plants, U12 introns represent a level at which expression of their host genes can be regulated.

INTRODUCTION

Pre-mRNA splicing in eukaryotes is a fundamental step in gene expression and represents an important level at which the expression of protein-coding genes can be regulated (Reed, 2000; Smith and Valcárcel, 2000; Graveley, 2001; Hastings and Krainer, 2001; Cartegni et al., 2002; Black, 2003). In higher eukaryotes, there are two classes of nuclear pre-mRNA introns. The most abundant class consists of U2-dependent introns (U2 introns), whereas the second rarer class (<0.4% of introns) consists of U12-dependent introns (U12 introns). U12 introns have been found in the nuclear genomes of vertebrates, plants, and insects (Hall and Padgett, 1994; Wu et al., 1996; Sharp and Burge, 1997; Tarn and Steitz, 1997; Wu and Krainer, 1998, 1999; Burge et al., 1998; Levine and Durbin, 2001; Patel and Steitz, 2003). Introns belonging to these two distinct classes are spliced by two different spliceosomes: the major U2-type spliceosome and the less abundant U12-type spliceosome (Hall and Padgett, 1996; Tarn and Steitz 1996a, 1996b, 1997). Although the first U12 introns to be described had AT-AC–terminal dinucleotides, the majority of U12-type introns contain GT-AG, and a small number contain other noncanonical terminal dinucleotides, such as AT-AA, AT-AG, AT-AT, GT-AT, or GT-GG (Jackson, 1991; Hall and Padgett, 1994; Dietrich et al., 1997, 2001a; Sharp and Burge, 1997; Burge et al., 1998; Wu and Krainer, 1999; Levine and Durbin, 2001; Zhu and Brendel, 2003). Moreover, functional analyses have shown that AT-AC–terminal dinucleotides are not a defining feature of U12 introns (Dietrich et al., 1997, 2001a). Instead, U12 introns contain highly conserved sequences in the 5′ splice site (exon:G/ATATCCTY) and branchpoint region (TCCTTRAY) (Hall and Padgett, 1994; Sharp and Burge, 1997; Burge et al., 1998), which are both required for prespliceosome complex formation (Frilander and Steitz, 1999). The 3′ splice site consensus sequence of U12 introns (YAC/G:exon) is less informative than their 5′ splice site and branchpoint sequences. U12 introns also lack a polypyrimidine tract and have a short distance between the branchpoint and 3′ splice site (between 10 and 20 nucleotides) (Hall and Padgett, 1994; Burge et al., 1998; Dietrich et al., 2001a; Tarn and Steitz, 1997). In *Arabidopsis thaliana*, a recent computational search for U12 introns predicted 165 such introns representing ~0.15% of the predicted total number of introns (Arabidopsis Genome Initiative, 2000; Zhu and Brendel, 2003). Plant U12 introns contain the same splice site and branchpoint consensus sequences as vertebrate U12 introns (Wu et al., 1996; Zhu and Brendel, 2003). Similarly, the first and last nucleotides are variable, with the most common dinucleotide combinations being GT-AG and AT-AC and only two examples that have AT-AA and GT-AT (Zhu and Brendel, 2003). In addition, 153 introns had branchpoint/3′ splice site distances of <21 nucleotides, showing that plant U12 introns in general maintain the core sequence elements required for splicing in vertebrate U12 intron splicing (Zhu and Brendel, 2003).

Both U12- and U2-type spliceosomes contain five small nuclear ribonucleoprotein particles (snRNPs) required for

[1] These authors contributed equally to this work.
[2] Current address: Gene Expression Programme, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, Scotland, UK.
[3] To whom correspondence should be addressed. E-mail jbrown@scri.sari.ac.uk; fax 44-1382-562426.
The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: Craig G. Simpson (csimps@scri.sari.ac.uk).
Article, publication date, and citation information can be found at www.plantcell.org/cgi/doi/10.1105/tpc.020743.

spliceosome assembly and function. The U5snRNP is common to both types of spliceosome (Tarn and Steitz, 1996a). In addition to U5, the U12 spliceosome contains U11, U12, U4atac, and U6atac snRNPs, which are distinct from their major spliceosome counterparts, U1, U2, U4, and U6, but have analogous roles in splice site selection and spliceosome formation (Hall and Padgett, 1996; Tarn and Steitz, 1996a, 1996b, 1997; Kolossova and Padgett, 1997; Yu and Steitz, 1997; Incorvaia and Padgett, 1998; Wu and Krainer, 1999; Patel and Steitz, 2003). In particular, the functions of analogous small nuclear RNAs (snRNAs) are conserved. For example, base-pairing interactions occur between the U11snRNA/5′ splice site and U12snRNA/branchpoint during prespliceosomal complex formation, similar to U1 and U2 in the major spliceosome (Kolossova and Padgett, 1997; Frilander and Steitz, 1999). U4atac and U6atac snRNAs show similar base-pairing interactions and functional domains as U4 and U6 (Shukla and Padgett, 1999, 2001). In addition, both the major and minor spliceosomes share many common spliceosomal proteins (Luo et al., 1999; Will et al., 1999, 2001; Schneider et al., 2002). Plant homologs of U6atac and U12 have been identified in Arabidopsis, and despite large sequence differences, their functional regions are conserved (Shukla and Padgett, 1999). U11 and U4atac remain to be identified in plants but the conservation of intron signals and U12 and U6atac suggest that these snRNAs are likely to exist as part of the plant U12 splicing system. As with U2 introns, U12 intron splicing follows a two-step reaction that involves the formation of a characteristic intron lariat intermediate at an adenosine that can be at one of two possible positions within the stringent branchpoint consensus (Tarn and Steitz, 1996a; McConnell et al., 2002). In addition, exon bridging interactions between complexes forming on splice sites of U12 introns and flanking U2 introns can enhance U12-dependent intron splicing (Wu and Krainer, 1996; Dietrich et al., 2001b; Hastings and Krainer, 2001; Kmieciak et al., 2002), and some U12 introns are involved in alternative splicing (Dietrich et al., 2001b; Zhu and Brendel, 2003).

The distinguishing feature of U2 introns in plants, compared with introns from vertebrates and yeast (*Saccharomyces cerevisiae*), is their UA richness, which is required for efficient U2 intron splicing and for 5′ and 3′ splice site selection (Goodall and Filipowicz, 1989; Simpson and Filipowicz, 1996; Brown and Simpson, 1998; Lorković et al., 2000; Reddy, 2001). The role of UA-rich elements in plant splicing is still poorly understood, but they may minimize secondary structure of plant introns or bind specific UA binding proteins to recruit splicing factors early in spliceosome formation (Simpson and Filipowicz, 1996; Brown and Simpson, 1998; Lorković et al., 2000; Reddy, 2001). Putative UA-rich binding proteins (UBP1, RBP45, and RBP47) with affinity for U-rich sequences have been isolated and characterized (Gniadkowski et al., 1996; Lambermon et al., 2000; Lorković et al., 2002). Of these, only UBP1 so far has been shown to affect splicing, where overexpression of UBP1 enhanced splicing of otherwise poorly spliced U2-type introns (Lambermon et al., 2000). Putative plant U12 introns are also UA rich in comparison with exonic sequences (Zhu and Brendel, 2003; our unpublished results), but it remains to be seen whether UA richness determines splice site selection or influences splicing efficiency.

To date, the splicing of only one plant U12-dependent intron has been experimentally analyzed in vivo (Kmieciak et al., 2002). Here, we present an analysis of sequence elements that are involved in plant U12 intron splicing efficiency using three different Arabidopsis U12-dependent introns. Differences in splicing efficiency among the introns were investigated by mutational analysis of splice site and putative branchpoint sequences, leading to identification of essential branchpoint nucleotides in vivo. We also demonstrate that UA content influences plant U12 intron splicing efficiency and that two U-rich RNA binding proteins, UBP1 and RBP45, have differential effects on splicing of U2- and U12-type introns. Finally, we show the presence of potential exon splicing enhancer sequences in the 5′ flanking exon of one U12 intron that are required for maximal splicing efficiency.

RESULTS

Isolation of Three Different Plant U12-Type Introns

Three different plant U12-dependent introns were selected from three different Arabidopsis genes on the basis of their splice site and branchpoint sequences (Figure 1). The three genes were *CBP20*, encoding the 20-kD cap binding protein (Kmieciak et al., 2002), *GSH2*, encoding glutathione synthetase (Wang and Oliver, 1996), and *LD*, encoding the LUMINIDEPENDENS protein (Lee et al., 1994). The U12 intron from the *CBP20* gene is 134 bp long and is the fourth of eight introns in the *CBP20* gene. The U12 intron from the *GSH2* gene is 108 bp long and is the sixth intron out of 11, and that of *LD* is 235 bp long and is the tenth intron out of 12. All three introns show sequence features characteristic of U12-dependent introns but have different combinations of splice site and branchpoint sequences. The introns from *CBP20* and *GSH2* have AU-AC splice site dinucleotides but vary in their branchpoint sequence. *CBP20* has the branchpoint sequence

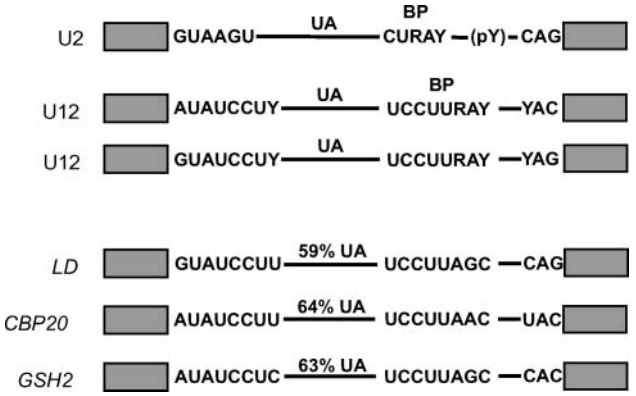

Figure 1. Schematic Representation of Splicing Signals of U2- and U12-Dependent Introns.

Consensus 5′ and 3′ splice sites and branchpoints are shown for plant U2 and U12 introns of both the AU-AC and GU-AG types. The splicing signals of the three plant U12 introns studied here (*LD*, *CBP20*, and *GSH2*) are presented. pY, polypyrimidine tract or U-rich sequence; UA, UA richness.

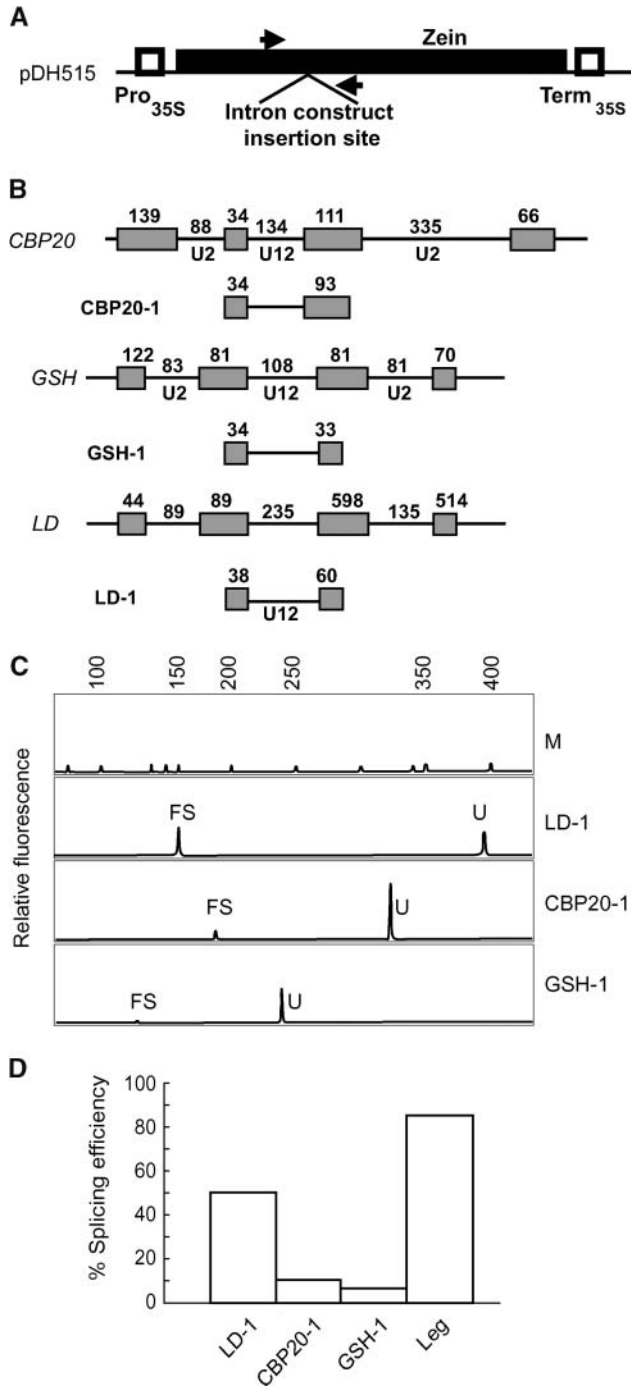

Figure 2. Schematic Diagrams of Intron Constructs to Examine Splicing of U12 Introns.

(A) Transient expression vector pDH5I5, into which intron constructs are placed for splicing analysis. Closed box, zein gene; open boxes, CaMV 35S promoter and terminator regions.

(B) The wild-type exon-intron structures of part of the *CBP20*, *GSH*, and *LD* genes are shown. Boxes, exons; lines, introns. The sizes of authentic exons and exon fragments in the U12 constructs are given.

(C) GeneScan analysis of splicing of the three U12 introns and a U2 control intron from a pea legumin gene in tobacco protoplasts. FS, fully spliced; U, unspliced; M, DNA size markers.

(D) Histogram of splicing efficiency.

UCCUUAAC, whereas *GSH2* has UCCUUAGC. The *LD* U12 intron shows the same branchpoint sequence as *GSH2* but has GU-AG termini (Figure 1). The A/G branchpoint variation (underlined) occurs at the branchpoint nucleotide suggested for animal U12-type introns (Tarn and Steitz, 1996a). However, the adenosine directly upstream has also been shown to act as the branchpoint nucleotide in some vertebrate U12 introns (McConnell et al., 2002). The UA content of the U12-type introns from the *CBP20*, *GSH2*, and *LD* genes is 64, 63, and 59%, respectively, which is equal to or just above the minimum UA requirement for accurate and efficient splicing of plant U2-dependent introns (Goodall and Filipowicz, 1989). The introns were isolated from Arabidopsis genomic DNA by PCR and included >30 bp of authentic 5′ and 3′ exon sequences. Amplified fragments were introduced into the plant expression vector pDH515 (Figure 2A; Simpson et al., 1996), which contains the 35S promoter of *Cauliflower mosaic virus* (CaMV 35S), terminator sequences, and an intronless zein gene, encoding a maize (*Zea mays*) seed storage protein. Each of the three Arabidopsis U12-dependent introns, together with fragments of authentic flanking exons, was cloned into the unique *Bam*HI site within the zein gene, giving constructs pCBP20.1, pGSH.1, and pLD.1 (Figure 2B).

Arabidopsis U12-Dependent Introns Are Spliced Less Efficiently in Tobacco Protoplasts Than U2-Dependent Introns

The U12 intron constructs pCBP20.1, pGSH.1, and pLD.1 were introduced into tobacco (*Nicotiana tabacum*) protoplasts by polyethylene glycol–mediated transfection. As a control, the splicing efficiency of the U2-dependent intron from a pea (*Pisum sativum*) legumin gene was monitored in the same experiments. The efficiency of excision of different introns was analyzed by RT-PCR on total RNA isolated from transfected protoplasts, and splicing efficiency expressed as the mean of three independent experiments. The *CBP20* and *GSH2* introns were spliced with only 10 and 5% efficiency, respectively, whereas the *LD* intron was spliced with 50% efficiency (Figures 2C and 2D). The control U2-type legumin intron was spliced with much higher efficiency (85 to 90%) as found previously. The accuracy of the splicing reactions was confirmed by sequencing of RT-PCR products. Thus, all three U12 introns were spliced inefficiently in the tobacco protoplast system when compared with the efficient splicing of the U2 control intron (Figures 2C and 2D). Moreover, there were notable differences among the splicing efficiencies of the three U12 introns.

AU-AC and GU-AG U12-Dependent Intron Splice Site Dinucleotides Are Equally Efficient Splice Sites

Of the three U12 introns studied in the protoplast system, the *LD* intron was spliced 5- to 10-fold more efficiently than the *CBP20*

and the *GSH2* introns. Both poorly spliced introns have AU-AC at their termini, whereas the *LD* intron has GU-AG intron boundaries normally found in the major U2 introns. To investigate whether the increased splicing efficiency of the intron from *LD* was because of its splice site sequences, the *LD* and *GSH2* U12-dependent introns were mutated from GU-AG to AU-AC and from AU-AC to GU-AG, respectively. Intermediate constructs with only the 5′ or 3′ splice sites mutated were also prepared in both cases (Figure 3A). All mutants were introduced into tobacco protoplasts, and splicing efficiency was measured using RT-PCR. Mutating the GU-AG splice sites of the *LD* intron to AU-AC (pLD.4) had no effect on splicing efficiency (Figures 3B and 3C). On the other hand, mutation of the AU-AC splice sites to GU-AG in the *GSH2* U12 intron (pGSH.4) increased splicing efficiency but remained significantly lower than the splicing efficiency of the *LD* intron (Figure 3). The constructs containing the intermediate intron mutation with AU-AG termini (pLD.2 and pGSH2.2) showed reduced splicing, which in the case of the LD construct was significant (Figures 3B and 3C). The GU-AC splice site combination abolished splicing of the *GSH2* U12 intron (pGSH.3) (Figure 3C). Although the GU-AC combination in the *LD* U12 intron (pLD.3) also abolished splicing to the mutated 3′ splice site, the overall efficiency of splicing remained at 50% because of activation of two cryptic 3′ splice sites (Figure 3D). Thus, the GU-AC combination of terminal dinucleotides is not competent for splicing. Sequencing of the RT-PCR products confirmed that one cryptic AG- lay three nucleotides upstream of the mutated splice site (used in 5% of transcripts), whereas the second 3′ AG- was located nine nucleotides downstream (used in 95% of the spliced transcripts) (Figure 3D). Thus, AU-AC and GU-AG terminal dinucleotides were equally suitable plant U12 intron splice sites, whereas the intermediate combinations, AU-AG and GU-AC, were unable to support splicing.

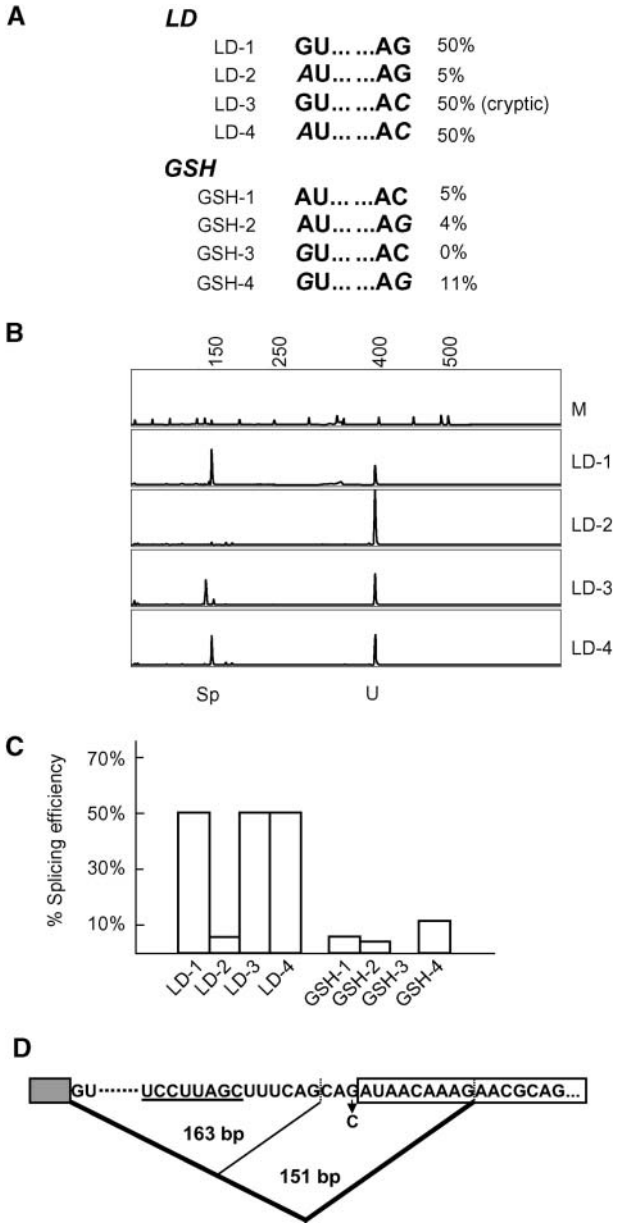

Figure 3. Splicing Analysis of 5′ and 3′ Splice Site Mutations of the *LD* and *GSH* U12 Introns.

(A) Mutations to *LD* and *GSH* to change splice sites from GU-AG to AU-AC and AU-AC to GU-AG, respectively.

(B) GeneScan analysis of splicing of *LD* mutants. Sp, spliced product; U, unspliced; M, DNA size markers.

(C) Histogram of splicing efficiencies of *LD* and *GSH2* mutants.

(D) Cryptic splicing of the *LD* GU-AC mutant (LD-3). The mutation causes selection of either of two cryptic AG dinucleotides, illustrated by lines. The branchpoint sequence is underlined, the G-to-C mutation is indicated, and the sizes of RT-PCR products are given.

Plant U12-Dependent Intron Branchpoint Nucleotides Important for Splicing

Plant U12-type introns contain the conserved UCCUURAY consensus branchpoint sequence found in vertebrate U12 introns. Adenosines at either of the underlined purine positions have been shown to act as branchpoint nucleotides in vertebrate U12 intron splicing (Tarn and Steitz, 1996a; McConnell et al., 2002). Similarly, some plant U12 introns have an adenosine at only one or other position, suggesting some flexibility in branchpoint selection and utilization in U12-type splicing (Zhu and Brendel, 2003; our unpublished results). Of the three plant U12-type introns tested in these studies, the *GSH2* and *LD* introns have the same branchpoint sequence, UCCUUAGC, whereas *CBP20* has the sequence UCCUUAAC (Figures 1 and 4A). The branchpoint sequences are located 11 nucleotides (*LD*), 12 nucleotides (*GSH2*), and 13 nucleotides (*CBP20*) from their associated 3′ splice sites. To investigate whether adenosines at either position could be used in splicing, a series of mutations were made to the branchpoint sequences (Figure 4A). Splicing of the mutated introns was tested in transfected tobacco protoplasts, and the efficiency of splicing was determined (Figure 4A).

In the *LD* intron, the single nucleotide mutation of A to U at the upstream purine position severely reduced splicing to only 5%

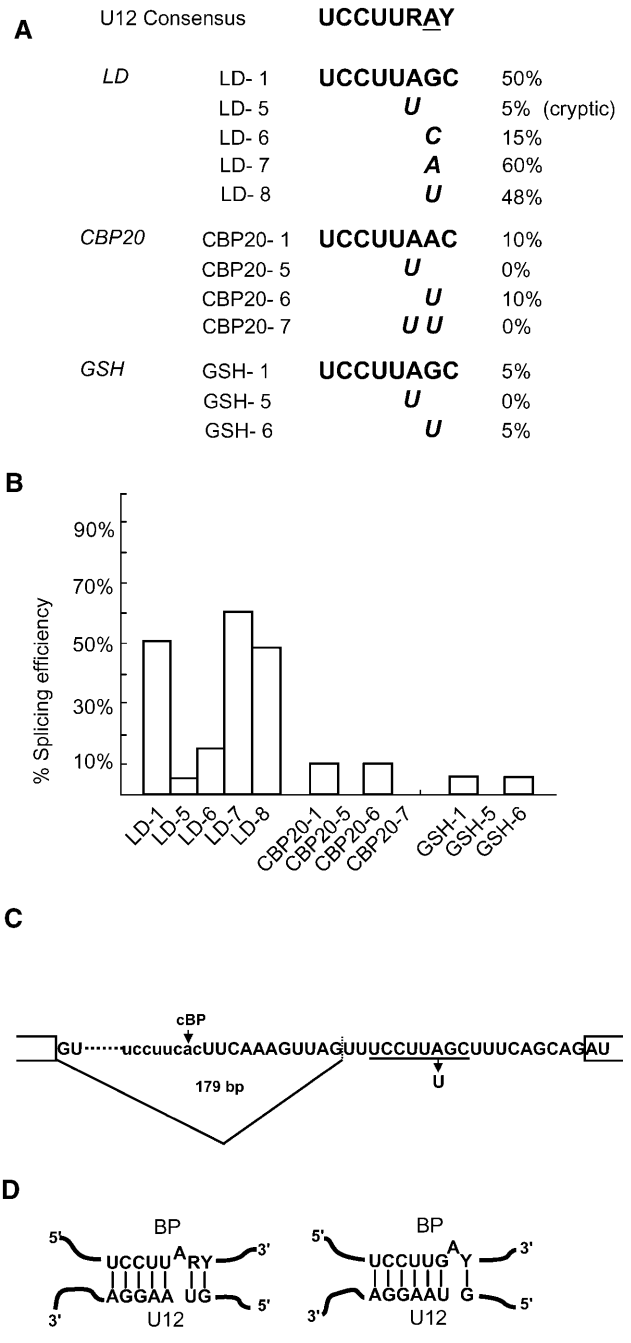

Figure 4. Splicing Analysis of Branchpoint Mutations.

(A) The U12 consensus branchpoint (Zhu and Brendel, 2003) is compared with the branchpoint sequences of *LD*, *CBP20*, and *GSH2*. Single nucleotide mutations are indicated, and efficiencies of splicing are shown next to the mutations.

(B) Histogram of splicing efficiencies of branchpoint mutants.

(C) Cryptic splicing in the branchpoint A-to-U mutation (LD-5) where an upstream cryptic branchpoint and 3′ splice site are activated (lines). The cryptic branchpoint sequence is given in lower case letters, and the putative cryptic branchpoint adenosine (cBP) is indicated by an arrow. The putative authentic branchpoint is underlined, and the A-to-U mutation is indicated.

(D) Models for U12snRNA-branchpoint base-pairing interactions where adenosines at either of two positions act as the branchpoint nucleotide.

(Figure 4B). This residual splicing was because of activation of a cryptic 3′ splice site 19 nucleotides upstream of the mutated 3′ splice site (Figure 4C). Interestingly, a putative U12 branchpoint sequence, UCCUUCAC, was present 13 nucleotides upstream of the cryptic 3′ splice site (Figure 4C), suggesting that the cryptic splicing event was also performed by the U12 spliceosome. Mutation of the G at the downstream purine position in the *LD* intron did not affect selection of the authentic 3′ splice site, but different nucleotides at this position gave different splicing efficiencies. The G-to-U and G-to-A mutations had little effect, but the G-to-C mutation reduced splicing efficiency from 50 to 15%, perhaps by affecting the base-pairing interactions between U12snRNA and the branchpoint sequence (Figure 4D). Thus, only the upstream adenosine in the *LD* U12 intron was essential for splicing.

To examine this further and to confirm the results obtained with the *LD* U12 intron, the nucleotides at both purine positions were also mutated in *CBP20* and *GSH2*. Despite the low splicing efficiency of these introns, mutation of the A to U at the upstream purine position in both introns completely abolished splicing (Figures 4A and 4B, CBP20-5 and GSH-5), whereas mutation of the A at the downstream position in *CBP20* and the G in the *GSH2* intron had no effect (Figures 4A and 4B, CBP20-6 and GSH-6, respectively). Thus, in all three plant U12 introns tested, the adenosine at the upstream purine position was required for splicing. This nucleotide or the purine in the downstream position can be bulged in putative branchpoint/U12snRNA base-pairing interactions (Figure 4D).

Increasing UA Content of U12 Introns Improves Splicing Efficiency

The efficiency of plant U2 intron splicing relies on high intron UA content relative to surrounding exons. Although Arabidopsis U12 introns show a similar UA content range to U2 introns (Zhu and Brendel, 2003; our unpublished results), it is not known whether this UA requirement is important for plant U12 intron splicing. The UA content of the analyzed U12-type introns (59 to 64% UA) lies close to the minimum of 59% UA needed for efficient splicing of U2 introns in tobacco protoplasts (Goodall and Filipowicz, 1989). To examine whether UA-rich elements influence U12 splicing, a series of constructs was made to increase UA content in the poorly spliced *CBP20* U12 intron. The constructs consisted of the *CBP20* U12 intron, with either one, two, or three copies of a U-rich element (UUUUUAU) introduced 67 nucleotides from the 5′ splice site, giving constructs pCBP20.8, pCBP20.9, and pCBP20.10, respectively (Figure 5A). The site of the insertion was selected to maintain the distance between the branchpoint sequence and the 3′ splice site. Constructs were transfected into tobacco protoplasts, and their splicing was assessed by RT-PCR (Figure 5B).

When a single U-rich element was introduced into the *CBP20* U12 intron, its splicing efficiency was unaffected, remaining at 10%. Insertion of two or three copies of the UUUUUAU element

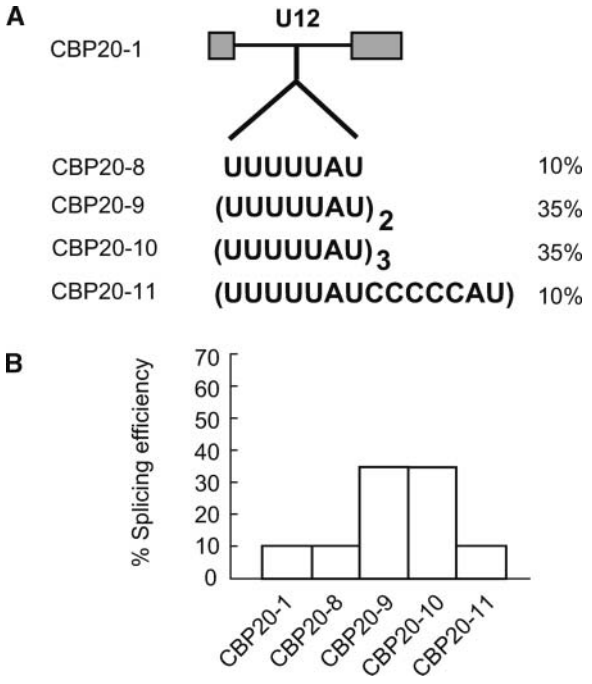

Figure 5. Splicing Analysis of U12 Introns Containing Additional UA-Rich Elements.

(A) One, two, or three copies of the UA-rich element UUUUUAU or a control sequence were inserted into the *CBP20* intron. Efficiencies of splicing are shown next to the insertions.
(B) Histogram of splicing efficiencies of constructs CBP20-8, CBP20-9, CBP20-10, and CBP20-11 in tobacco protoplasts.

led to a 3.5-fold increase in splicing efficiency (35%) of the *CBP20* U12 intron. To confirm that this effect was because of increasing UA content of the intron and not because of an increase in intron length, a control construct containing a 14-nucleotide-long sequence insertion (UUUUUAUCCCCCAU), consisting of one UA-rich element and seven additional nucleotides, was prepared (pCBP20.11). Splicing of this construct showed only poor splicing (10%), demonstrating that the two or three inserted UA-rich elements were responsible for the increased splicing efficiency. Thus, increased UA content improved the poor splicing of the *CBP20* U12-type intron and demonstrated a role for UA sequence elements in plant U12 intron splicing.

Differential Effects of U-Rich Binding Proteins on U12 and U2 Intron Splicing

RBP45 and UBP1 are hnRNP-like RNA binding proteins with affinity for U-rich sequences (Gniadkowski et al., 1996; Lambermon et al., 2000; Lorković et al., 2000). UBP1 is able to improve the splicing efficiency of otherwise poorly spliced U2 plant introns (Lambermon et al., 2000). To determine whether overexpression of UBP1 or RBP45 could increase the splicing efficiency of plant U12 introns in tobacco protoplasts, constructs expressing HA-tagged UBP1 or RBP45 were cotransfected with the three U12 introns (Figure 6). As controls, RT-PCR and protein gel blot analysis with antibody to the HA epitope tag were performed to demonstrate the presence of transcripts and protein produced from the transfected U-rich RNA binding protein constructs (data not shown). Overexpression of these proteins did not affect the splicing efficiency of any of the U12 introns (Figure 6). However, overexpression of UBP1 with the wheat (*Triticum aestivum*) amylase U2 intron, which has a relatively low UA content and splices poorly (5 to 15%) in tobacco protoplasts (Simpson et al., 1996), increased splicing efficiency to 85% (Figure 6). Thus, whereas UBP1 can enhance splicing efficiency of poorly spliced U2 introns, confirming the findings of Lambermon et al. (2000), neither UBP1 nor RBP45 influenced U12 intron splicing directly.

Previously, exon bridging interactions between adjacent U2 introns were shown to enhance splicing of plant pre-mRNAs (McCullough et al., 1996; Simpson et al., 1998, 1999), and flanking U2 introns increased splicing efficiency of the *CBP20* U12 intron (Kmieciak et al., 2002). Overexpression of UBP1 or RBP45 did not enhance U12 intron excision in single intron constructs, but may function in intron recognition and recruitment of splicing factors to splice sites via exon bridging interactions. To examine whether UBP1 or RBP45 could enhance U12 intron splicing in such a transcript, a construct with the poorly spliced *GSH2* U12 intron and its two authentic flanking U2 introns was made (pGSH.9; Figure 7A). In addition, two intermediate constructs with only the 5′ or the 3′ flanking U2 intron were prepared (pGSH.7 and pGSH.8, respectively). These latter two constructs showed increases in splicing efficiency of the U12 intron from 5 to ~10%. When the pGSH.9 construct was expressed in tobacco protoplasts, there were four different RT-PCR products, each with similar abundance of ~20 to 25% (Figures 7B and 7C). These products were characterized by sequencing and represented the unspliced transcript, 564 bp;

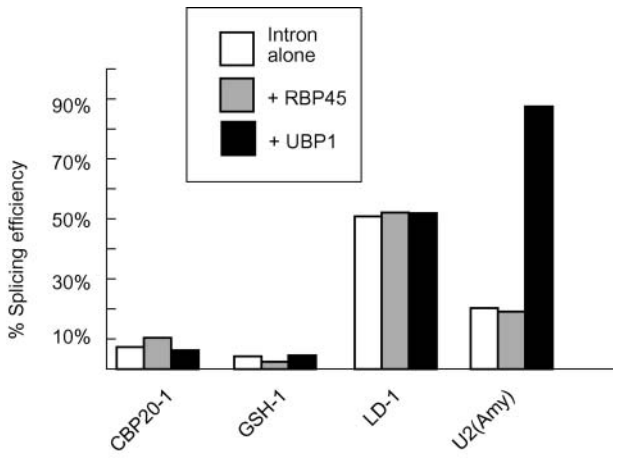

Figure 6. Effect of Overexpression of U-Rich Binding Proteins RBP45 and UBP1 on Splicing of U2 and U12 Introns.

Splicing efficiencies of U12 introns in the absence and presence of overexpressed RBP47 and UBP1 are shown. As a control, the effect of RBP45 and UBP1 on splicing of a poorly spliced U2 intron (Amy U2) was tested.

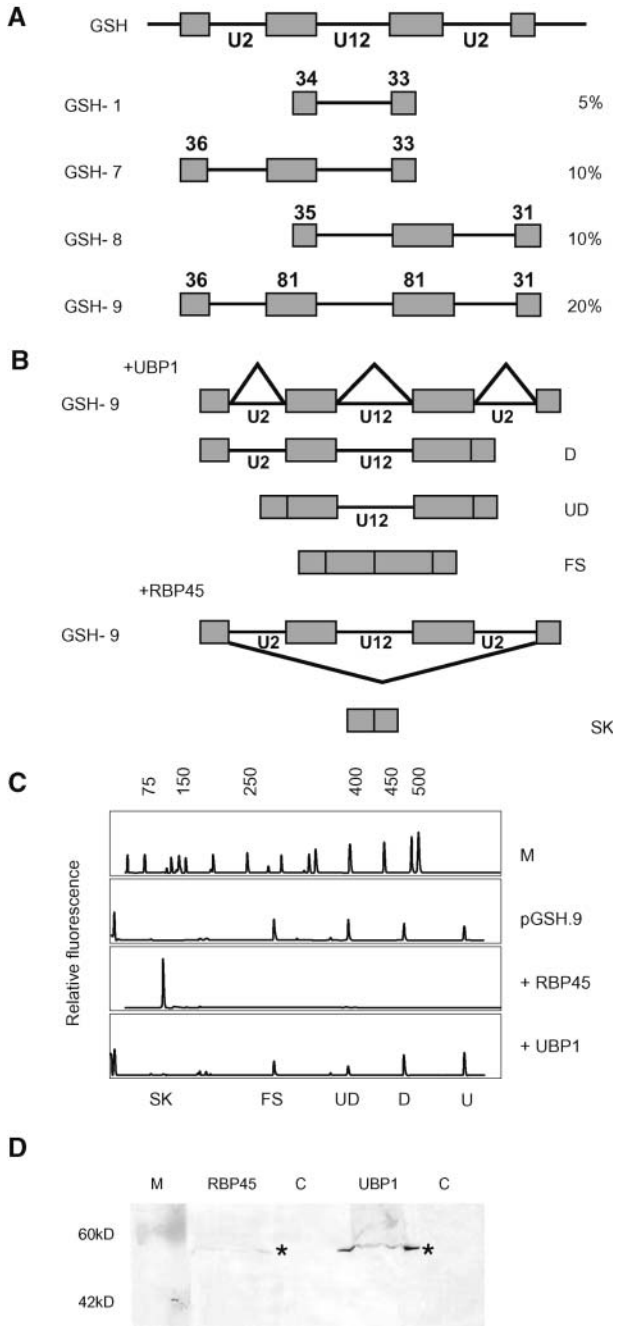

Figure 7. RBP45 Causes Exon Skipping in the Complex GSH Intron Construct.

(A) Constructs of the *GSH2* U12 intron with flanking U2 introns are shown diagrammatically under the wild-type *GSH2* exon-intron structure. GSH.1, U12 intron alone; GSH.7, U12 intron and upstream U2 intron; GSH.8, U12 intron and downstream U2 intron; GSH.9, U12 intron with both upstream and downstream U2 introns.

(B) Splicing of GSH.9 construct gave three different products (in addition to unspliced transcripts) where either the downstream U2 intron was removed (D) or where both U2 introns were removed (UD) and the fully spliced product (FS). Overexpression of UBP1 did not alter the splicing pattern of GSH.9, but overexpression of RBP45 caused exon skipping (SK).

partially spliced transcripts, 479 bp long, where the downstream U2 intron is removed; a partially spliced transcript, 399 bp long, where both U2 introns are removed; and fully spliced transcripts, where both U2 and the U12 intron are removed (290 bp) (Figures 7B and 7C). There was no evidence of partially spliced transcripts, where only the upstream U2 intron had been spliced, suggesting that the pathway of intron removal is splicing of the downstream U2 intron first, followed by the upstream U2 intron and then the U12 intron. The presence of both U2 introns in the triple intron construct increased splicing efficiency of the U12 intron to 20% compared with the 5% efficiency in the single intron construct. This suggested that the splicing efficiency of the U12 intron was enhanced by exon bridging interactions between flanking U2 introns, as observed previously for the *CBP20* U12 intron (Kmieciak et al., 2002).

To examine the effect of overexpression of UBP1 and RBP45 on exon-bridging interactions, the three-intron construct, pGSH.9, was cotransfected with HA-tagged UBP1 or RBP45. Overexpression of UBP1 with pGSH.9 in tobacco protoplasts had little effect on splicing pattern or efficiency. However, overexpression of RBP45 completely abolished the normal splicing pattern, and transcripts showed excision of all three introns and both exons in an exon skipping event, giving an RT-PCR product of 126 bp (Figures 7B and 7C). In addition, no unspliced product was visible (Figure 7C). Therefore, RBP45 increased the overall splicing efficiency of the transcript but blocked recognition of the internal *GSH2* exon sequences. Thus, neither UBP1 nor RBP45 enhances splicing of U12 introns, but RBP45 can influence splice site selection in pre-mRNAs. This, and the ability of UBP1 to enhance removal of poorly spliced U2 but not U12 introns, demonstrates differential roles for UBP1 and RBP45 in pre-mRNA processing, which may reflect specificity of interaction with different sequences in pre-mRNAs.

The High Splicing Efficiency of the *LD* Intron Is Because of Exon Splicing Enhancer Activity

The *LD* U12 intron is spliced much more efficiently in tobacco protoplasts than the *CBP20* or *GSH2* introns. As we have already shown, this enhanced splicing efficiency is not because of splice site dinucleotides, branchpoint sequences, or UA content. To examine whether the higher splicing efficiency of this intron was because of sequences in the flanking exons, the length of the 5′, 3′, or both flanking exons was reduced. In pLD.1, the upstream exon contained 38 nucleotides of the authentic 89 nucleotides, whereas the downstream exon consisted of 60 nucleotides from the original 579 nucleotides (Figure 8A). In pLD.9, both upstream and downstream exons were reduced to 3 and 8 bp, respectively. Intermediate constructs with 3 bp upstream and 60 bp downstream exons and 38 bp upstream and 8 bp

(C) GeneScan analysis of splicing of GSH.9 alone and with overexpressed RBP45 or UBP1 proteins in tobacco protoplasts. M, DNA size markers.

(D) Protein gel blot analysis with anti-HA antibody of HA-tagged RBP45 and UBP1 proteins overexpressed in tobacco protoplasts. Protein bands are indicated with asterisks. C, control untransfected protoplasts.

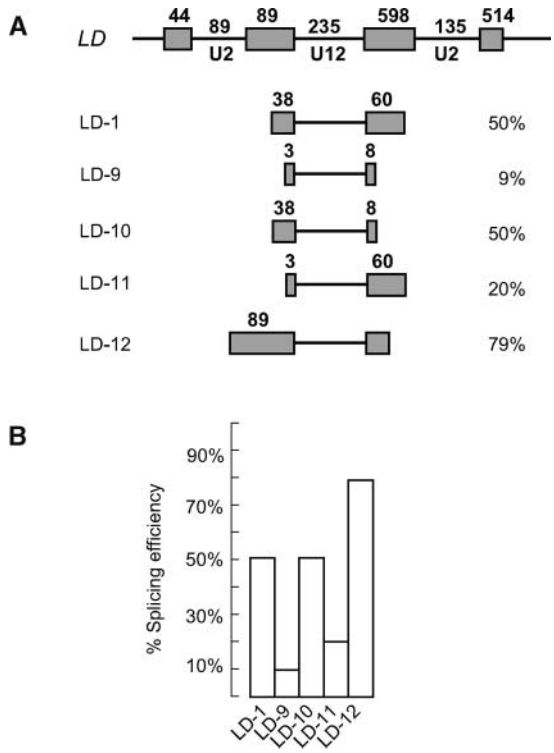

Figure 8. Exon 3 of *LD* Contains Exon Splicing Enhancer Sequences.

(A) Diagrams of *LD* U12 intron constructs with differing lengths of flanking exons (LD-1 and LD-9 to LD-12) are shown below the wild-type exon-intron structure of part of the *LD* gene.

(B) Histogram of splicing efficiencies of the *LD* constructs in tobacco protoplasts.

downstream exons were also generated (pLD.10 and pLD.11). The splicing efficiency of both pLD.9 and pLD.11 dropped significantly to 9 and 20%, respectively, compared with the 50% splicing of the original pLD.1 construct (Figures 8A and 8B). Splicing of pLD.10 remained at 50% efficiency, indicating that sequence(s) between 38 and 3 nucleotides in the upstream exon are important for the higher level of splicing of the *LD* U12 intron in tobacco protoplasts. Furthermore, when the full-length upstream exon was included (Figures 8A and 8B, pLD.12), splicing efficiency of the *LD* U12 intron increased to 79%, suggesting that the region between 89 and 38 nucleotides also contained sequences that increased splicing efficiency. Thus, in the case of the *LD* U12 intron, splicing in tobacco protoplasts required the complete upstream exon (exon 10) sequence for maximal splicing efficiency, and there are potentially two regions that contribute to this enhancement.

DISCUSSION

In the Arabidopsis genome, 165 U12 introns have been predicted by computational analyses, including 113 GT-AG, 50 AT-AC, 1 AT-AA, and 1 GT-AT introns. These introns share similar features with plant U2 introns in terms of length distribution and low GC content (Zhu and Brendel, 2003; our unpublished results). In the absence of a plant in vitro splicing system, we have performed a detailed analysis of the determinants of U12 intron splicing in plants using an in vivo protoplast system. Through extensive mutational analysis, we have examined the influence of splice site and branchpoint signals on splicing efficiency and have identified the likely branchpoint nucleotide used in three plant U12 introns. In particular, we have shown that U12 intron splicing is enhanced by increased UA richness, a characteristic feature of plant U2 and U12 introns, and by the presence of flanking U2 introns, supporting the role of exon bridging interactions between U2 and U12 introns. Finally, exon splicing enhancer sequences may modulate splicing efficiency of U12 introns, raising the possibility of regulation of gene expression of U12 intron–containing genes at the level of pre-mRNA splicing.

Differential Splicing Efficiency of U12 Introns

Two observations were made about splicing efficiency of the plant U12 introns in tobacco protoplasts. First, when introduced as single intron constructs, the efficiency of splicing of all three introns was lower than normally observed for U2 introns. Second, there were significant differences in splicing efficiency among the *LD*, *GSH*, and *CBP20* introns. In the protoplast system, the U12-containing transcripts were overexpressed from the strong CaMV 35S promoter. This could suggest that transfected protoplasts were not competent to splice the U12 introns efficiently because U12 spliceosomes and/or other factors involved in U12 intron splicing were limiting. In the same protoplast system, U2 intron–containing transcripts expressed from the CaMV 35S promoter were efficiently spliced, possibly reflecting the much higher abundance of U2 spliceosomes. However, the *LD* U12 intron construct containing the complete 5′ exon (LD-12) was spliced with 79% efficiency, arguing against the simple interpretation that lower splicing efficiency reflects the lower abundance of U12 splicing machinery components. The efficiency of splicing of pLD.12 demonstrates that the protoplast system has the inherent capability to splice U12 introns efficiently; therefore, the observed differences between *LD*, *CBP20*, and *GSH2* U12 intron–containing pre-mRNAs must reflect variations in splice signal sequences and their ability to interact with splicing factors.

Plant U12 Splice Site Signals

From mutational and computational analyses of vertebrate U12-type introns, it is clear that both AU-AC and GU-AG termini function as U12 intron splice sites (Dietrich et al., 1997). On the basis that the *LD* U12 intron was spliced with much higher efficiency than those of *CBP20* or *GSH2* and that *LD* had GU-AG splice sites, it was possible that the terminal dinucleotides could influence efficiency of plant U12 intron splicing. However, mutational analyses demonstrated that splicing efficiency of plant U12-type introns was unchanged, irrespective of whether the terminal dinucleotides were AU-AC or GU-AG. Moreover, our results indicated that the plant U12-type spliceosome was not able to use the GU-AC–terminal dinucleotide combination, reflecting the lack of plant U12 introns with this sequence (Zhu and Brendel, 2003). The *LD* AU-AG mutant had reduced splicing, similar

to vertebrate mutant U12 introns with AU-AG termini, which splice less efficiently than either AU-AC or AU-AA introns (Wu and Krainer, 1999; Dietrich et al., 2001a).

In the absence of plant in vitro splicing extracts, branchpoint nucleotides required for splicing were examined by in vivo analysis of branchpoint mutations of the three plant U12 introns, as performed previously for plant U2 introns (Simpson et al., 1996, 2000, 2002). The vertebrate and plant consensus branchpoint sequence is UCCUURAY. In vertebrates, the branchpoint nucleotide of the U12 intron in the human *P120* gene was mapped to the adenosine (underlined) (Tarn and Steitz, 1996a). However, some plant and vertebrate U12 introns contain a G or U at this position, with an adenosine immediately upstream, and recently, vertebrate U12 intron branchpoints have been mapped to adenosines in either position (McConnell et al., 2002). For example, the branchpoints of two human U12 introns (from *SME* and *XRP*) were mapped by primer extension to the adenosine in the sequence UCCUUAGC, whereas in the U12 introns from *XTF* and *P120*, branchpoint adenosines mapped to the downstream adenosine in the sequences UCCUUAAA and UCCUUAAC, respectively. Thus, in vertebrates, there is flexibility in branchpoint nucleotide selection, where an A at either position can function as the branchpoint adenosine (McConnell et al., 2002). Plant U12 introns contain an adenosine at one or both purine positions, suggesting that such flexibility may also occur in plant U12 intron splicing (Zhu and Brendel, 2003; our unpublished results). Alternative base-pairing interactions between the U12snRNA and branchpoint sequence (Figure 4D) could cause the purine at either position to be bulged from the duplex and used as the branchpoint nucleotide, as demonstrated for vertebrate U2 introns (Figure 4D; Query et al., 1994). Nevertheless, mutation of the A to U in the upstream position of all three plant U12 introns inhibited splicing. This suggests that either a uridine in this position is not tolerated, perhaps disrupting U12snRNA–branchpoint interactions, or the upstream adenosine is the preferred branchpoint nucleotide. Finally, in the four human introns (above), the use of the upstream or downstream adenosine correlated with their terminal dinucleotides, GU-AG or AU-AC (Tarn and Steitz, 1996a; McConnell et al., 2002), but this correlation did not hold for the plant introns studied here.

A further significant feature of U12 splicing is the short distance between the branchpoint and 3′ splice site and the lack of a polypyrimidine tract (Dietrich et al., 2001a; Levine and Durbin, 2001). In the *LD* U12 intron, the authentic 3′ splice site lies 11 nucleotides downstream of the branchpoint. When mutated (GU-AC), two cryptic 3′ splice site -AGs, nine nucleotides downstream or 3 nucleotides upstream of the mutated 3′ splice site, were activated. These 3′ splice sites were positioned 20 and 8 nucleotides from the branchpoint and were selected in 95 and 5% of transcripts, respectively. The frequencies of selection are consistent with the range of branchpoint to 3′ splice site distances between 10 and 20 nucleotides (with an optimum of 12 to 13 nucleotides) proposed for vertebrate U12 introns and with the observation that when this spacing was altered to <10 nucleotides, normal splicing ceased and cryptic splicing was often activated (Dietrich et al., 2001a). In addition, none of the 165 plant U12 introns had branchpoint to 3′ splice distances <11 nucleotides (mean of 12 nucleotides), and only 12 putative plant U12 introns were reported with a branchpoint to 3′ splice site distance >21 nucleotides (Zhu and Brendel, 2003). Therefore, branchpoint/3′ splice site distance is an important determinant of plant U12 intron 3′ splice site selection. In addition, seven examples of alternative splicing involving plant U12 introns were identified underpinning the potential regulation of splicing and expression of genes containing U12 introns (Zhu and Brendel, 2003).

Splice site selection in both plant and vertebrate U2 introns involves exon and/or intron definition. In the exon definition model, splicing factors recognize splice sites at either end of an exon and assemble a complex on the exon; neighboring exons are then brought together to allow the interaction of 5′ and 3′ splice sites and intron removal. In vertebrates, exon definition has been demonstrated between a U12 intron and a downstream U2 intron (Wu and Krainer, 1996). We have shown that both upstream and downstream U2 introns can increase splicing efficiency of two plant U12 introns: *GSH2* (this article) and *CBP20* (Kmieciak et al., 2002).

UA Richness and U-Rich RNA Binding Proteins

The increased splicing efficiency as a result of inclusion of two or three U-rich elements in the *CBP20* U12 intron parallels the stimulatory effect of such sequences on splicing efficiency of U2 introns (Goodall and Filipowicz, 1989; Gniadkowski et al., 1996). However, splicing efficiency did not increase to the levels observed with some U2 introns. For plant U2 introns, UA-rich binding proteins are thought to function early in spliceosome assembly, to bind to U-rich elements within the intron, and to stabilize and recruit snRNPs and other splicing factors. The only partial improvement in splicing of the U12 intron by insertion of the U-rich sequences therefore may be because of (1) a suboptimal arrangement of U-rich elements within the intron, (2) different functional roles for U-rich elements in U12 and U2 intron splicing, or (3) UA-rich binding factors interacting differentially with U12 and U2 splicing machinery components. Thus, both types of intron may depend on U richness in different ways, and/or different U-rich binding proteins may have specific roles in U2 and U12 intron recognition and splicing. These variations may reflect the need to distinguish between U2 and U12 spliceosomal snRNPs and to recruit, for example, the U11-U12 di-snRNP, which interacts with the U12 5′ splice site and branchpoint (Sharp and Burge, 1997).

In this regard, the two highly related plant U-rich RNA binding proteins, UBP1 and RBP45, showed different effects on U2 and U12 intron splicing. Neither of the proteins affected splicing activity of the three plant U12 introns, but overexpression of UBP1 enhanced splicing of the poorly spliced amylase U2 intron, confirming a role for UBP1 in plant U2 splicing (Lambermon et al., 2000; Lorković et al., 2000). These results suggest that the interactions whereby U2 intron splicing was improved by UBP1 were absent from U12 intron splicing and highlight differences between plant U2 and U12 splicing, most likely in early events in intron recognition and spliceosome assembly. Overexpression of RBP45 had little effect on single intron constructs but affected the splicing efficiency and pattern of a more complex transcript

containing the U12 intron flanked by U2 introns. The exon-skipping phenotype, suggesting disruption of exon bridging interactions, was specific to RBP45 overexpression because the related protein, UBP1, did not have this effect. It remains to be determined whether the effect of RBP45 overexpression is specific to U12-containing transcripts, but these results clearly demonstrate differential functions for these closely related proteins, probably reflecting different binding affinities of these proteins.

U12 Introns and Regulation of Gene Expression

U12 introns are often found in genes that function in DNA replication and RNA metabolism, and in some cases, the presence and positions of U12 introns are conserved across species, suggesting a role for U12 introns in the regulation of expression of these genes (Burge et al., 1998). In vivo splicing of human U12 introns has been shown to proceed more slowly than splicing of U2 introns in the same transcript, which may reflect the lower abundance of the U12 spliceosome or a slower, less efficient splicing reaction (Patel et al., 2002). Furthermore, it has been postulated that U12 intron removal is a rate-limiting step for complete splicing of U12 intron–containing transcripts, with partially spliced transcripts being targeted for degradation. This would maintain appropriate expression levels that could be regulated by the levels of factors involved in U12 intron splicing (Patel et al., 2002). In in vitro extracts, splicing of U12 introns is also slower than that of U2 introns (Tarn and Steitz, 1996b; Wu and Krainer, 1996). Consistent with this mechanism of gene expression regulation, splicing of plant U12 introns, in vivo in tobacco protoplasts, also occurred more slowly or less efficiently than splicing of U2 introns from transcripts containing both U12 and U2 introns (see Kmieciak et al., 2002). In addition, preliminary results showed different efficiencies of *GSH2* U12 intron splicing in different organs of Arabidopsis, being highest in leaves and lowest in roots and flowers (our unpublished results). Thus, it is feasible that plant U12 introns represent one level at which U12 intron–containing genes are regulated posttranscriptionally.

Splicing enhancers play a critical role in the regulation of splicing and in correct splice site recognition of constitutively spliced pre-mRNAs (Hertel et al., 1997; Wang and Manley, 1997; Hertel and Maniatis, 1998; Blencowe, 2000; Smith and Valcárcel, 2000; Graveley, 2001, 2002; Cartegni et al., 2002; Black, 2003). Although they are usually located within the downstream exon, enhancer sequences also have been observed in upstream exons or within the intron itself (Blencowe, 2000). Exon splicing enhancers (ESEs) are often purine rich and are believed to function through the binding of specific protein factors. The best described of these factors are the SR proteins, a large family of polypeptides with one or more RNA binding domains and a variable length domain (the RS domain) containing multiple copies of Arg-Ser dipeptides (Fu, 1995; Valcárcel and Green, 1996; Wang and Manley, 1997; Tacke and Manley, 1999; Black, 2003). Although purine-rich enhancers were identified in many exons flanking U2-dependent introns, it already has been demonstrated, both in vitro and in vivo, that ESEs of this type are also functional in the U12 pre-mRNA splicing pathway (Wu and Krainer, 1998; Graveley, 2000; Dietrich et al., 2001b; Hastings and Krainer, 2001). Many SR proteins have been identified in plants (Lazar et al., 1995; Lopato et al., 1996a, 1996b, 1999a, 1999b, 2002), but to date, only one purine-rich exonic element, which promotes 5′ splice site selection, has been described (McCullough and Schuler, 1997). In this article, we have shown that the exon upstream of the *LD* U12 intron clearly contains two separate regions that increased the efficiency of splicing. Searching this exon sequence for identified vertebrate ESEs and binding sites of SR proteins (http://www.exon.cshl.org/ESE/) revealed two such sequences that might support SR protein binding. Thus, SR protein (or other similar factors) binding to the exon located upstream of the *LD* U12-type intron may stimulate recognition of the 5′ splice site. In addition, the *LD* upstream exon also contained a sequence that has been experimentally demonstrated to act as an ESE in Arabidopsis (S. Mount, personal communication). It is possible that SR (or other) proteins, which recognize ESEs in upstream or downstream exons flanking U12 introns, could contact a component of the U11/U12 di-snRNP, stabilizing its binding to the 5′ splice site (U11) and branchpoint (U12) sequences and promoting U12 spliceosome formation. Interestingly, the *LD* intron has the lowest UA content of those tested but splices with the highest efficiency and therefore may strongly depend on ESEs for regulation of U12 intron splicing. Further investigations are required to determine whether the ESEs are intron specific or can function in other U12 or U2 splicing pathways. Thus, as with plant U2 introns whose splicing efficiency reflects the strength of various signals, splicing of plant U12 introns depends on both intronic and exonic signals (UA content and ESEs) and exon bridging interactions with flanking U2 introns.

METHODS

Isolation of U12 Introns and Construction of Mutants

The *Arabidopsis thaliana* U12 introns from *CBP20* (intron 4) (Kmieciak et al., 2002), *GSH2* (intron 6) (Wang and Oliver, 1996), and *LD* (intron 10) (Lee et al., 1994) and varying amounts of upstream and downstream intron and exon sequences were isolated from Arabidopsis genomic DNA by PCR (Figures 1 and 2B). Fragments were inserted into the unique *Bam*HI site of expression vector pDH515 (Figure 2A; Simpson et al., 1996). The basic set of constructs consisted of the U12 introns from each of the three genes *CBP20*, with 34 and 93 nucleotides of upstream and downstream exons (Figure 2B; pCBP20.1), *GSH2*, with 34 and 33 nucleotides of upstream and downstream exons (Figure 2B; pGSH.1), and *LD*, with 38 and 60 nucleotides of upstream and downstream exons (Figure 2B; pLD.1). PCR amplification of genomic Arabidopsis DNA was also performed to make *GSH2* constructs that contained the upstream-associated U2 intron (pGSH.7; Figure 7A), downstream-associated U2 intron 9 (pGSH.8; Figure 7A), and a construct that contained the U12 intron with both upstream and downstream U2 introns (pGSH.9; Figure 7A). PCR amplification of pLD.1 was performed to reduce the length of both upstream (pLD.9 and pLD.11; Figure 8A) and downstream (pLD.9 and pLD.10; Figure 8A) exon sequences. Genomic Arabidopsis DNA was PCR amplified to extend the length of the upstream exon to its full length of 89 nucleotides (pLD.12; Figure 8A). Site-specific nucleotide substitutions to change the 5′ and 3′ splice site dinucleotides of pLD.1 and pGSH.1 (Figure 3A) and putative branchpoint nucleotides of pLD.1, pCBP20.1, and pGSH.1 (Figure 4A) were performed by site-directed mutagenesis using the Quick-Change site-directed mutagenesis kit

(Stratagene, La Jolla, CA) according to the manufacturer's protocol. One, two, and three copies of the UUUUUAU sequence element (Gniadkowski et al., 1996) were inserted into pCBP20.1, 67 nucleotides downstream from the 5′ splice site and 67 nucleotides upstream of the 3′ splice site, using the Quick-Change site-directed mutagenesis kit (pCBP20.8, pCBP20.9, and pCBP20.10, respectively; Figure 5A). Site-directed mutagenesis was performed on pCBP20.9 to create a pyrimidine-rich control construct (pCBP20.11; Figure 5A). All sequence insertions and mutations were confirmed by sequencing. A list of all the oligonucleotides used for isolation of intron/exon fragments and for mutagenesis is available upon request.

Splicing Analysis by RT-PCR

Intron constructs were transfected into *Nicotiana tabacum* var Xanthi, and total RNA was isolated as described previously (Simpson et al., 1996). RT-PCR analysis was as described (Simpson et al., 2000) using the PCR primers O8, 5′-CCCAATTGTTCAACCCTAC-3′, labeled with the 5′ fluorescent phosphoramidite 6-FAM, and O9, 5′-GGTAAGATGCCT-GTTGCGATTGC-3′. The primer O8 corresponds to the zein sequence 5′ to the site of intron construct insertion, and O9 is complementary to the zein sequence 3′ to the site of intron construct insertion in pDH515 (Simpson et al., 1996). Labeled RT-PCR products were separated on a 4% polyacrylamide denaturing gel on an ABI 377 DNA sequencing machine. Sizes of bands were calculated using GeneScan version 2.1 software by comparison with GeneScan-350 (TAMRA) size standards (Applied Biosystems, Foster City, CA). Quantification of RT-PCR products was by measurement of the fluorescent peak areas of the detected fragments after 24 cycles. Previous quantification established that PCR amplification was linear over a range of 15 to 24 cycles, and Taq polymerase enzyme efficiency was the same for spliced, partially spliced, and unspliced products. Splicing efficiency was calculated from the peak areas for each processed transcript. Each construct was tested at least twice, and standard errors were determined for constructs that were tested three or more times. Standard errors for the poorly spliced *GSH2* and *CBP20* U12 intron constructs were all $<\pm 2\%$, and for the *LD* intron, constructs were $<\pm 6\%$. Novel RT-PCR products (e.g., from cryptic splicing) were isolated from a 6% nondenaturing acrylamide gel and eluted from the gel overnight at 4°C in 10 mM Tris-HCl, pH 8, 1 mM EDTA, 0.5 M ammonium acetate, and 0.1% (w/v) SDS overnight at 4°C. The eluted bands were precipitated and resuspended in water before reamplification with oligonucleotides O8 and O9. These PCR products were either cloned into pGEM-T Easy (Promega, Madison, WI) and sequenced using the standard RS and SS primers, or the PCR product was sequenced directly using O8 and O9.

Cotransfection Analysis

HA-tagged U-rich RNA binding protein expression cassettes, pUBP1-HA and pRBP45-HA (Lambermon et al., 2000; Lorković et al., 2002), were mixed separately with an equal molar amount of pCBP20.1, pGSH.1, pGSH.9, pLD.1, and pA, respectively. Protoplasts were transfected with the plasmid mixtures, and splicing analysis was performed as described above. Expression of the tagged proteins was monitored by protein gel blot and RT-PCR analyses. Protein was extracted from protoplasts by boiling in 50 mM Tris-HCl, pH 6.8, 20% (v/v) glycerol, 1 mM EDTA, 1% (w/v) SDS, and 15% (v/v) β-mercaptoethanol and bromophenol blue. Protein gel blot analysis was performed using rabbit anti-HA antibody (Sigma, St. Louis, MO) as primary antibody and anti-rabbit IgG alkaline phosphatase conjugate (Sigma) as secondary antibody to confirm protein expression in protoplasts. RT-PCR analysis was performed using primers specific to the HA-tagged protein transcripts.

ACKNOWLEDGMENTS

This research was supported by the Scottish Executive Environment and Rural Affairs Department, the Polish Committee for Scientific Research (Grants 0265/P04/2001/21 and 0045/P04/2002/23), and the Royal Society. We thank Witek Filipowicz (Freidrich Miescher Institute, Basel, Switzerland) for the gift of pUBP1-HA and pRBP45-HA expression cassettes.

Received January 6, 2004; accepted February 25, 2004.

REFERENCES

Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408,** 796–815.

Black, D.L. (2003). Mechanisms of alternative pre-messenger RNA splicing. Annu. Rev. Biochem. **72,** 291–336.

Blencowe, B.J. (2000). Exonic splicing enhancers: Mechanism of action, diversity and role in human genetic diseases. Trends Biochem. Sci. **25,** 106–110.

Brown, J.W.S., and Simpson, C.G. (1998). Splice site selection in plant pre-mRNA splicing. Annu. Rev. Plant Physiol. Plant Mol. Biol. **49,** 77–95.

Burge, C.B., Padgett, R.A., and Sharp, P.A. (1998). Evolutionary fates and origins of U12-type introns. Mol. Cell **2,** 773–785.

Cartegni, L., Chew, S.L., and Krainer, A.R. (2002). Listening to silence and understanding nonsense: Exonic mutations that affect splicing. Nat. Rev. Genet. **3,** 285–298.

Dietrich, R.C., Incorvaia, R., and Padgett, R.A. (1997). Terminal intron dinucleotide sequences do not distinguish between U2- and U12-dependent introns. Mol. Cell **1,** 151–160.

Dietrich, R.C., Peris, M.J., Seyboldt, A.S., and Padgett, R.A. (2001a). Role of the 3′ splice site in U12-dependent intron splicing. Mol. Cell. Biol. **21,** 1942–1952.

Dietrich, R.C., Shukla, G.C., Fuller, J.D., and Padgett, R.A. (2001b). Alternative splicing of U12-dependent introns in vivo responds to purine-rich enhancers. RNA **7,** 1378–1388.

Frilander, M.J., and Steitz, J.A. (1999). Initial recognition of U12-dependent introns requires both U11/5′ splice site and U12/branchpoint interactions. Genes Dev. **13,** 851–863.

Fu, X.-D. (1995). The superfamily of arginine/serine-rich splicing factors. RNA **1,** 663–680.

Gniadkowski, M., Hemmings-Mieszcak, M., Klahre, U., Liu, H.-X., and Filipowicz, W. (1996). Characterization of intronic uridine-rich sequence elements acting as possible targets for nuclear proteins during pre-mRNA splicing in Nicotiana plumbaginifolia. Nucleic Acids Res. **24,** 619–627.

Goodall, G.J., and Filipowicz, W. (1989). The AU-rich sequences present in the introns of plant nuclear pre-mRNAs are required for splicing. Cell **58,** 473–483.

Graveley, B.R. (2000). Sorting out the complexity of SR protein functions. RNA **6,** 1197–1211.

Graveley, B.R. (2001). Alternative splicing: Increasing diversity in the proteomic world. Trends Genet. **17,** 100–107.

Graveley, B.R. (2002). Sex, agility, and the regulation of alternative splicing. Cell **109,** 409–412.

Hall, S.L., and Padgett, R.A. (1994). Conserved sequences in a class of rare eukaryotic introns with non-consensus splice sites. J. Mol. Biol. **239,** 357–365.

Hall, S.L., and Padgett, R.A. (1996). Requirement of U12 snRNA for in vivo splicing of a minor class of eukaryotic nuclear pre-mRNA introns. Science **271,** 1716–1718.

Hastings, M.L., and Krainer, A.R. (2001). Pre-mRNA splicing in the new millennium. Curr. Opin. Cell Biol. **13,** 302–309.

Hertel, K.J., Lynch, K.W., and Maniatis, T. (1997). Common themes in the fuction of transcription and splicing enhancers. Curr. Opin. Cell Biol. **9,** 350–357.

Hertel, K.J., and Maniatis, T. (1998). The function of multisite splicing enhancers. Mol. Cell **1,** 449–455.

Incorvaia, R., and Padgett, R.A. (1998). Base pairing with U6atac snRNA is required for 5′ splice site activation of U12-dependent introns in vivo. RNA **4,** 709–714.

Jackson, I.J. (1991). A reappraisal of non-consensus mRNA splice sites. Nucleic Acids Res. **19,** 3795–3798.

Kmieciak, M., Simpson, C.G., Lewandowska, D., Brown, J.W.S., and Jarmolowski, A. (2002). Cloning and characterization of two subunits of Arabidopsis thaliana nuclear cap-binding complex. Gene **283,** 171–183.

Kolossova, I., and Padgett, R.A. (1997). U11 interacts in vivo with the 5′ splice site of U12-dependent (AU-AC) introns. RNA **3,** 227–233.

Lambermon, M.H.L., Simpson, G.G., Kirk, D.A., Hemmings-Mieszczak, M., Klahre, U., and Filipowicz, W. (2000). UBP1, a novel hnRNP-like protein that functions at multiple steps of higher plant nuclear pre-mRNA maturation. EMBO J. **19,** 1638–1649.

Lazar, G., Schaal, T., and Maniatis, T. (1995). Identification of a plant serine-arginine-rich protein similar to the mammalian splicing factor SF2/ASF. Proc. Natl. Acad. Sci. USA **92,** 7672–7676.

Lee, I., Aukerman, M.J., Gore, S.L., Lohman, K.N., Michaels, S.D., Weaver, L.M., John, M.C., Feldmann, K.A., and Amasino, R.M. (1994). Isolation of LUMINIDEPENDENS: A gene involved in the control of flowering time in Arabidopsis. Plant Cell **6,** 75–83.

Levine, A., and Durbin, R. (2001). A computational scan for U12-dependent introns in the human genome sequence. Nucleic Acids Res. **29,** 4006–4013.

Lopato, S., Forstner, C., Kalyna, M., Hilscher, J., Langhammer, U., Indrapichate, K., Lorković, Z.J., and Barta, A. (2002). Network of interactions of a novel plant-specific Arg/Ser-rich protein, atRSZ33, with atSC35-like splicing factors. J. Biol. Chem. **277,** 39989–39998.

Lopato, S., Gattoni, R., Fabini, G., Stevenin, J., and Barta, A. (1999a). A novel family of plant splicing factors with a Zn knuckle motif: Examination of RNA binding and splicing activities. Plant Mol. Biol. **39,** 761–773.

Lopato, S., Kalyna, M., Dorner, S., Kobayahshi, R., Krainer, A.R., and Barta, A. (1999b). atSRp30, one of two SF2/ASF-like proteins from Arabidopsis thaliana, regulates splicing of specific plant genes. Genes Dev. **13,** 987–1001.

Lopato, S., Mayeda, A., Krainer, A.R., and Barta, A. (1996a). Pre-mRNA splicing in plants: Characterization of Ser/Arg splicing factors. Proc. Natl. Acad. Sci. USA **93,** 3074–3079.

Lopato, S., Waigmann, E., and Barta, A. (1996b). Characterization of a novel arginine/serine-rich splicing factor in Arabidopsis. Plant Cell **8,** 2255–2264.

Lorković, Z.J., Kirk, D.A.W., Lambermon, M.H.L., and Filipowicz, W. (2000). Pre-mRNA splicing in higher plants. Trends Plant Sci. **5,** 160–167.

Lorković, Z.J., Wieczorek Kirk, D.A., Lambermon, M.H.L., and Filipowicz, W. (2002). RBP45 and RBP47, two olgouridylate-specific hnRNP-like proteins interacting with poly(A)$^+$ RNA in nuclei of plant cells. RNA **6,** 1610–1624.

Luo, H.R., Moreau, G.A., Levin, N., and Moore, M.J. (1999). The human Prp8 protein is a component of both U2- and U12-dependent spliceosomes. RNA **5,** 893–908.

McConnell, T.S., Cho, S.-J., Frilander, M.J., and Stietz, J.A. (2002). Branchpoint selection in the splicing of U12-dependent introns in vitro. RNA **8,** 579–586.

McCullough, A.J., Baynton, C.E., and Schuler, M.A. (1996). Interactions across exons can influence splice site recognition in plant nuclei. Plant Cell **8,** 2295–2307.

McCullough, A.J., and Schuler, M.A. (1997). Intronic and exonic sequences modulate 5′ splice site selection in plant nuclei. Nucleic Acids Res. **25,** 1071–1077.

Patel, A.A., McCarthy, M., and Steitz, J.A. (2002). The splicing of U12-type introns can be a rate-limiting step in gene expression. EMBO J. **21,** 3804–3815.

Patel, A.A., and Steitz, J.A. (2003). Splicing double: Insights from the second spliceosome. Nat. Rev. Mol. Cell Biol. **4,** 960–970.

Query, C.C., Moore, M.J., and Sharp, P.A. (1994). Branch nucleophile selection in pre-mRNA splicing: Evidence for the bulged duplex model. Genes Dev. **8,** 587–597.

Reddy, A.S.N. (2001). Nuclear pre-mRNA splicing in plants. Crit. Rev. Plant Sci. **20,** 523–571.

Reed, R. (2000). Mechanisms of fidelity in pre-mRNA splicing. Curr. Opin. Cell Biol. **12,** 340–345.

Schneider, C., Will, C.L., Makarova, O.V., Makarov, E.M., and Lührmann, R. (2002). Human U4/U6.U5 and U4atac/U6atac.U5 tri-snRNPs exhibit similar protein compositions. Mol. Cell. Biol. **22,** 3219–3229.

Sharp, P.A., and Burge, C.B. (1997). Classification of introns: U2-type or U12-type. Cell **91,** 875–879.

Shukla, G.C., and Padgett, R.A. (1999). Conservation of functional features of U6atac and U12 snRNAs between vertebrates and higher plants. RNA **5,** 525–538.

Shukla, G.C., and Padgett, R.A. (2001). The intramolecular stem-loop structure of U6 snRNA can functionally replace the U6atac snRNA stem-loop. RNA **7,** 94–105.

Simpson, C.G., Clark, G.P., Davidson, D., Smith, P., and Brown, J.W.S. (1996). Mutation of putative branchpoint consensus sequences in plant introns reduces splicing efficiency. Plant J. **9,** 369–380.

Simpson, C.G., Clark, G.P., Lyon, J.M., Watters, J., McQuade, C., and Brown, J.W.S. (1999). Interactions between introns via exon definition in plant pre-mRNA splicing. Plant J. **18,** 293–302.

Simpson, G.G., and Filipowicz, W. (1996). Splicing of precursors to messenger RNA in higher plants: Mechanism, regulation and subnuclear organisation of the spliceosomal machinery. Plant Mol. Biol. **32,** 1–41.

Simpson, C.G., Hedley, P.E., Watters, J.A., Clark, G.P., McQuade, C., Machray, G.C., and Brown, J.W.S. (2000). Requirements for mini-exon inclusion in potato invertase mRNAs provides evidence for exon-scanning in plants. RNA **6,** 422–433.

Simpson, C.G., McQuade, C., Lyon, J., and Brown, J.W.S. (1998). Characterization of exon skipping mutants of the COP1 gene from Arabidopsis. Plant J. **15,** 125–131.

Simpson, C.G., Thow, G., Clark, G.P., Jennings, S.N., Watters, J.A., and Brown, J.W.S. (2002). Mutational analysis of a plant branchpoint and polypyrimidine tract required for constitutve splicing of a mini-exon. RNA **8,** 47–56.

Smith, C.W.J., and Valcárcel, J. (2000). Alternative pre-mRNA splicing: The logic of combinatorial control. Trends Biochem. Sci. **25,** 381–388.

Tacke, R., and Manley, J.L. (1999). Determinants of SR protein specificity. Curr. Opin. Cell Biol. **11,** 358–362.

Tarn, W.-Y., and Steitz, J.A. (1996a). A novel spliceosome containing U11, U12, and U5snRNPs excises a minor class (AT-AC) intron in vitro. Cell **84,** 801–811.

Tarn, W.-Y., and Steitz, J.A. (1996b). Highly diverged U4 and U6 small nuclear RNAs required for splicing rare AT-AC introns. Science **273,** 1824–1832.

Tarn, W.-Y., and Steitz, J.A. (1997). Pre-mRNA splicing: The discovery of a new spliceosome doubles the challenge. Trends Biochem. Sci. **22,** 132–137.

Valcárel, J., and Green, M.R. (1996). The SR protein family: Pleiotropic functions in pre-mRNA splicing. Trends Biochem. Sci. **21,** 296–301.

Wang, C.L., and Oliver, D.J. (1996). Cloning of the cDNA and genomic clones for glutathione synthetase from Arabidopsis thaliana and complementation of a gsh2 mutant in fission yeast. Plant Mol. Biol. **31,** 1093–1104.

Wang, J., and Manley, J.L. (1997). Regulation of pre-mRNA splicing in metazoa. Curr. Opin. Genet. Dev. **7,** 205–211.

Will, C.L., Schneider, C., MacMillan, A.M., Katopodis, N.F., Neubauer, G., Wilm, M., Lührmann, R., and Query, C.C. (2001). A novel U2 and U11/U12 protein that associates with the pre-mRNA branch site. EMBO J. **20,** 4536–4546.

Will, C.L., Schneider, C., Reed, R., and Lührmann, R. (1999). Identification of both shared and distinct proteins in the major and minor spliceosomes. Science **284,** 2003–2005.

Wu, H.J., Gaubier-Comelia, P., Delseny, M., Grellet, F., Van Montagu, M., and Rouzé, P. (1996). Non-canonical introns are at least 10^9 years old. Nat. Genet. **14,** 383–384.

Wu, Q., and Krainer, A.R. (1996). U1-mediated exon definition interactions between AT-AC and GT-AG introns. Science. **274,** 1005–1008.

Wu, Q., and Krainer, A.R. (1998). Purine-rich enhancers function in the AT-AC pre-mRNA splicing pathway and do so independently of intact U1snRNP. RNA **4,** 1664–1673.

Wu, Q., and Krainer, A.R. (1999). AT-AC pre-mRNA splicing mechanisms and conservation of minor introns in voltage-gated ion channel genes. Mol. Cell. Biol. **19,** 3225–3236.

Yu, Y.-T., and Steitz, J.A. (1997). Site-specific crosslinking of mammalian U11 and U6atac to the 5′ splice site of an AT-AC intron. Proc. Natl. Acad. Sci. USA **94,** 6030–6035.

Zhu, W., and Brendel, V. (2003). Identification, characterization and molecular phylogeny of U12-dependent introns in the Arabidopsis thaliana genome. Nucleic Acids Res. **31,** 1–12.

Appendix B

BIRTH AND DEATH OF GENE OVERLAPS IN VERTEBRATES

BMC Evolutionary Biology

Research article

Birth and death of gene overlaps in vertebrates

Izabela Makałowska*[1], Chiao-Feng Lin[2,3] and Krisitina Hernandez[2]

Address: [1]The Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA 16802, USA, [2]Institute of Molecular Evolutionary Genetics and Department of Biology, The Pennsylvania State University, University Park, PA 16802, USA and [3]Institute of Bioinformatics, University of Muenster, 48149 Muenster, Germany

Email: Izabela Makałowska* - izabelam@psu.edu; Chiao-Feng Lin - cxl46@psu.edu; Krisitina Hernandez - kxh295@psu.edu

* Corresponding author

Published: 16 October 2007

BMC Evolutionary Biology 2007, **7**:193 doi:10.1186/1471-2148-7-193

Received: 26 March 2007
Accepted: 16 October 2007

This article is available from: http://www.biomedcentral.com/1471-2148/7/193

Abstract

Background: Between five and fourteen per cent of genes in the vertebrate genomes do overlap sharing some intronic and/or exonic sequence. It was observed that majority of these overlaps are not conserved among vertebrate lineages. Although several mechanisms have been proposed to explain gene overlap origination the evolutionary basis of these phenomenon are still not well understood. Here, we present results of the comparative analysis of several vertebrate genomes. The purpose of this study was to examine overlapping genes in the context of their evolution and mechanisms leading to their origin.

Results: Based on the presence and arrangement of human overlapping genes orthologs in rodent and fish genomes we developed 15 theoretical scenarios of overlapping genes evolution. Analysis of these theoretical scenarios and close examination of genomic sequences revealed new mechanisms leading to the overlaps evolution and confirmed that many of the vertebrate gene overlaps are not conserved. This study also demonstrates that repetitive elements contribute to the overlapping genes origination and, for the first time, that evolutionary events could lead to the loss of an ancient overlap.

Conclusion: Birth as well as most probably death of gene overlaps occurred over the entire time of vertebrate evolution and there wasn't any rapid origin or 'big bang' in the course of overlapping genes evolution. The major forces in the gene overlaps origination are transposition and exaptation. Our results also imply that origin of overlapping genes is not an issue of saving space and contracting genomes size.

Background

3.2 billion base pairs of the human genome harbor about 23,000 protein coding genes. With the average size of a gene equal to 48 kb, they cover approximately one third of the genome. It seems that there's enough space in the genome for each gene to be separated by a large distance. Yet, between five and fourteen per cent of genes in the vertebrate genomes do overlap [1]. The unexpected abundance of complementary pairs of sense/antisense transcripts poses major challenges to achieve a comprehensive understanding of a gene structure and expression at the genomic level. Studies of individual overlapping gene pairs in eukaryotes have shown that they regulate gene expression by different mechanisms such as genomic imprinting [2], RNA interference and translational regulation [3], transcriptional interference [4], alternative splic-

ing [5], and X-inactivation [6]. Many genes involved in overlaps are known to be involved in disease development, e.g. *CACP* gene is responsible for camptodactyly-arthropathy-coxa vara-pericarditis syndrome [7] or are responsible for some important morphological features, e.g. *SLC24A5* is partially responsible for skin coloration [8].

Several mechanisms have been proposed to explain gene overlap origination. For instance, Keese and Gibbs [9] suggested that overlapping genes arise as a result of overprinting – a process of generating new genes from preexisting nucleotide sequences. This process supposedly took place after divergence of mammals from birds and overlapping genes represent young, phylogeneticaly restricted genes encoding proteins with diverse functions, and are therefore specialized to the present life-style of the organism in which they are found. Shintani et al. [10] suggested that the overlap between genes *ACAT2* (acetyl-Coenzyme A acetyltransferase 2) and *TCP1* (t-complex 1) arose during the transition from therapsid reptiles to mammals in one of two ways. In one scenario, one of genes was translocated and the rearrangement has been accompanied by the loss of a part of the 3' UTR, including the polyadenylation signal, from one gene. By chance the 3' UTR of the new neighbor on the opposite strand contained all the signals necessary for transcription termination so that the translocated gene could continue to function. Alternatively, the two genes become neighbors through the rearrangement but at first did not overlap. Later, one of the genes lost its original polyadenylation signal, but was able to use a signal that happened to be present on the noncoding strand of the other gene. The *ACAT2-TCP1* overlap evolution was placed, similarily as in Keese and Gibbs hypothesis, after the divergence of mammals from birds. Dahary et al. [11] place the origin of most vertebrate overlaps much earlier. They found that human antisense genes have largely conserved linkage in torafugu which may imply that big fraction of human overlapping genes represents vertebrates' ancestral overlaps. However, our previous study of human and mouse overlapping genes showed that even between closely related species overlaps are not that well conserved [12]. Out of 255 cases in which both members of the human overlapping gene pair had mouse orthologs, only 95 were overlapping in both species. In addition, significant fraction of these 95 gene pairs show different overlap patterns in the two genomes. Lack of the overlap conservation was also observed in other studies [13-15].

Here we present results of the comparative analysis of seven vertebrate genomes: human, chimpanzee, mouse, rat, chicken, fugu, and zebrafish. This comparative study shows that on one hand, many of the vertebrate gene overlaps are not conserved and are lineage specific. On the other hand, this work reveals new, not published before, cases of genes overlap conservation in vertebrates. We also show new mechanisms of overlapping genes evolution and demonstrate, for the first time, that evolutionary events could not only lead to the new gene overlaps origin but also to the loss of an ancient overlap. Therefore lack of strong overlaps conservation between even closely related species may result from the origin of a new, lineage specific overlap as well as from the loss of overlaps in many lineages. Findings about evolutionary changes in the gene structure and organization are very important in our quest toward understanding genomes and genes expression. Changes in the gene structures may lead to modifications in the gene expression and expression correlation between involved genes, which may further explain some differences between species such as discrepancies in the orthologous genes expression patterns [16].

Results

Conservation of overlaps in vertebrate genomes

Fraction of the overlapping genes in various vertebrates differs significantly (Table 1). In tetrapoda over 10% of all genes are involved in some type of overlap, while in fish only 5–7%. The exception is the rat genome where only 4.87% of all genes are overlapping with another gene. However, this is most likely due to the annotation incompleteness and not a specific feature of the rat genome. Many rat genes do not have UTRs annotated and as we learned from this and other studies [12,17,18] the majority of gene overlaps are in the UTR regions. Incomplete annotations are likely to be responsible for some discrepancies between human and chimpanzee and could also be true for fish genomes. Another possibility is that many overlaps evolved after *Actinopterigii* diverged from *Sacropterigii* and most overlaps observed in these lineages arose independently. Differences in overlapping genes frequencies between human and other species were tested using chi-square test (Table 1). In all cases but mouse the chi-square is higher than critical value at $\alpha = 0.0005$ and therefore differences are statistically significant. For mouse the difference is significant at $\alpha = 0.05$.

Table 2 shows results of the analysis of overlap conservation, which indicate that sets of overlapping genes differ between species. This demonstrates that many overlaps are species or lineage specific and only a small fraction of them are shared among vertebrates. Although some conserved overlaps are not observed due to missing data, we can assume that this would affect both, species specific and conserved overlaps in similar way and therefore the proportions and the general picture are not affected. Similar disproportion in the sense-antisense transcripts abundance was also shown by Zhang et al. [19].

Table 1: Overlapping genes in vertebrate genomes

	Number of genes analyzed	Number and fraction (in parentheses) of genes involved in overlaps	Chi-square test value (when compared to human)	Number of overlaps	Nested genes	Exon/exon overlaps (NATs) *		Intron/exon overlaps*
						Total	CDS involved	
Human	22,291	2,978 (13.4)	NA	1,766	972	634	417	160
Chimpanzee	21,506	2,219 (10.3)	73.0888	1,276	665	479	317	132
Mouse	25,383	3,456 (13.6)	5.0750	2,053	1,071	819	565	163
Rat	22,159	1,080 (4.9)	895.1585	607	458	102	100	47
Chicken	17,709	1,960 (11.1)	32.2585	1,135	474	511	471	150
Fugu	20,796	993 (4.8)	880.5199	556	174	290	290	92
Zebrafish	23,524	1,625 (6.9)	472.4534	1,026	767	98	85	161

* excluding nested genes
Number of genes involved in overlaps is smaller than the number of overlaps multiplied by two since some genes are involved in more than one overlap. Multiple genes may be nested in one host gene as well as a gene may be overlapping other genes on both ends as reported previously [12]. Differences in overlapping genes frequencies between human and other species were tested using chi-square test. In all cases but mouse the chi-square is higher than critical value at $\alpha = 0.0005$ and therefore differences are statistically significant. For mouse the difference is significant at $\alpha = 0.05$. However, no definite conclusions can be drawn by this test results since some differences may result from the annotation problems but not real differences in the overlapping genes fraction.

Patterns of human overlapping genes evolution

Using Ensembl gene homology data [20] we identified homologs of human overlapping genes in other species. Out of 2,978 human genes involved in overlaps, 264 were human specific and had no homologs in any other analyzed genome, including chimpanzee. Interestingly, we couldn't find a rodent homolog for about 25% of human overlapping genes, whereas genome wide comparison shows that 89–90% of rat genes possess a single ortholog in the human genome [21]. Similarly, it was observed that 25% of human genes do not have torafugu orthologs [22], while our study shows that in the case of the overlapping genes 46.17% of human genes lack a torafugu ortholog. These results imply that a lot of genes involved in overlaps are young, lineage specific genes and do not have orthologs in other lineages. This supports the 'overprinting' hypothesis but is in sharp contrast to observation made by Dahary et al. who based on comparison of the human and torafugu genomes concluded that most human overlaps are ancient [11]. However, their conclusion may be an artifact of the applied method, because they analyzed only those human genes that have identifiable orthologs in the torafugu genome.

Based on the presence or absence of an ortholog in species representing two other lineages, i.e. rodents and fish, we divided human overlapping genes pairs into those that have: both orthologs in both lineages (476 pairs); both orthologs in rodents and only one in fish (279 pairs); both orthologs in rodents and none in fish (111 gene pairs); an ortholog of one gene only in both lineages (466 pairs); ortholog of one gene in rodents and none in fish (200 pairs); and no orthologs in neither lineage (92 pairs). Next, we analyzed genomic arrangement in all cases where both orthologs of human overlapping genes were found. According to our results we divided gene pairs into: overlapping (if they also overlap in particular species), neighboring (if they were not overlapping but

Table 2: Overlapping genes conserved between species

	Human	Chimpanzee	Mouse	Rat	Chicken	Fugu	Zebrafish
Human	-	1100	274	98	64	23	17
Chimpanzee	477	-	NA	NA	NA	NA	NA
Mouse	146	NA	-	141	76	26	16
Rat	11	NA	48	-	45	19	6
Chicken	9	NA	10	2	-	22	13
Fugu	1	NA	0	0	0	-	13
Zebrafish	5	NA	5	0	0	1	-

Above diagonal shows total number of conserved overlaps, and below diagonal shows numbers of conserved exon/exon overlaps. For chimpanzee the data is provided only in relation to human, gene orthology relation between chimpanzee and other than human species is not established and annotated in the Ensemble database yet.

placed one next to each other without any gene between them), and separated (if they were on different chromosomes, contigs or were separated by other genes).

Considering previously published major events leading to the gene overlaps; genes rearrangements or transposition, extension of genes by adoption of signals or new exons, and new genes origination [9,10] we developed 15 theoretical scenarios of overlapping genes evolution (Figure 1). Results from the above analysis provide support for every theoretical pattern of overlap evolution. Presented examples, for each scenario, were carefully examined and confirmed by the presence of cDNA and/or EST sequences. Also, sequence analysis was performed to ensure the missing orthologs are truly not present in a given species or lack of overlaps is not resulting from annotations problem. Our requirements here were very conservative and each gene pair where there was disagreement between species from the same lineage was removed from studies.

Mechanisms leading to gene overlaps

Gene overlaps evolve by a variety of mechanisms and not by a single universal mechanism. Essentially, any mechanism that gives rise to a new gene, such as gene duplication or retroposition, may result in a gene overlap. Alternative splicing represents another major source of proteome diversity in mammals and origination of a new splice form may lead to a gene overlap as well.

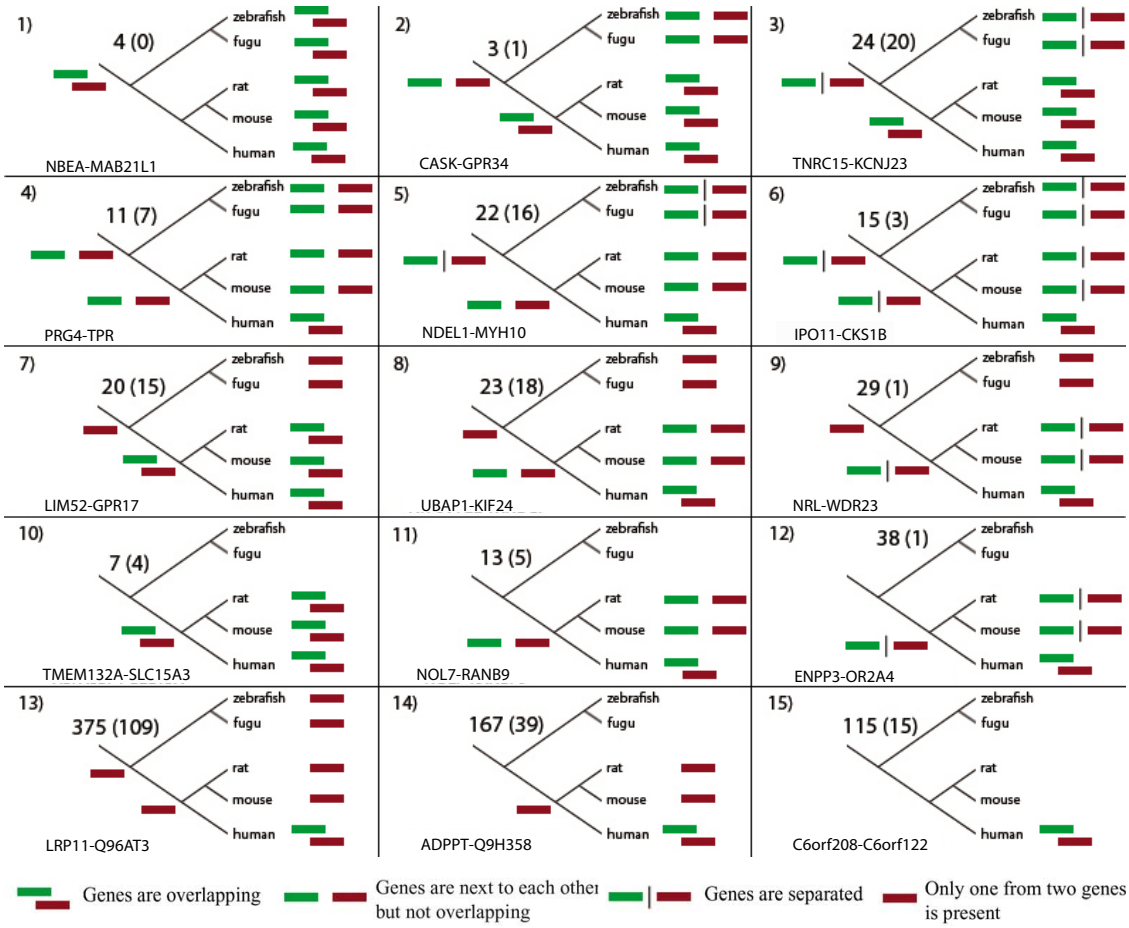

Figure 1
Putative patterns of human overlapping genes evolution together with examples from our data set. Numbers in parenthesis represent cases of exon/exon overlaps in each category. In each case an example from our studies is given. Analysis was done based on the October 2004 Ensembl release. The bar between two genes indicates that the genes are located on different chromosomes.

In the analyzed data, we found cases supporting all the proposed hypotheses of gene overlap origination, i.e. overprinting, a gene translocation, or adoption of a new transcription termination signal (changing a gene structure in more general terms). Among identified human overlapping genes in 115 cases both genes involved in the overlap did not have the ortholog in neither rodent nor fish lineage, in 64 cases one ortholog was present in rodents and none in fish, and only in 68 cases both orthologs were present in rodent and fish genomes. Figure 1 shows fifteen scenarios of overlapping gene evolution. Patterns 7, 13, and 14 (Figure 1) fit exactly the overprinting hypothesis because they show human overlaps where one of gene is an old gene present in all vertebrates and the second gene is a young one not present in fish nor rodents. Dan et al. [13] showed that a recently evolved overlap between MINK and CHRNE genes resulted from mutations in the polyadenlylation signal and acquisition of a new downstream signal within a neighboring locus. Evolutionary scenarios represented by models 2–6, 8, 9, 11, and 12 (Figure 1) indicate involvement of translocation and possible signal adoption in the overlapping genes origin.

Although our models support published hypotheses we should consider much broader range of events which could lead to genes overlaps. Summarizing published hypotheses in a more general way we can say that major events playing a role in overlapping genes evolution are: translocation (or transposition), change in the gene structure (extension of UTR would fall into this category), and development of a new gene or a new splice variant.

Development of a new splice variant

Gene overlaps might not be conserved among species due to different gene structures [12]. In addition to adopting a new termination and an extension of the last exon, conversion of the previously unused genetic material in the form of a new splicing variant may lead to the gene overlap. There are two possible scenarios, an additional splice variant arises or the ancestral variant may be replaced by a new one.

Developing a new, additional, splice variant may be considered as a special case of overprinting since the new splice variant represents a new transcript. In fact, the case described by Keese and Gibbs [9] falls into this category because reported the new gene is just a new splice variant of TRalpha (TRHA) gene. Comparative analysis of the genomic region containing TRHA and NR1D1 (nuclear receptor subfamily 1, group D, member 1) genes revealed that the overlap is conserved among placental mammals, who have two splice variants of TRHA. Only one of these, the one which does not overlap with NR1D1, was identified in marsupials and all non-mammalian lineages. Close examination of the genomic region alignments showed that an insertion of new genetic material occurred some time after divergence of placental mammals and this inserted sequence was used for a new splice variant. This finding disagrees with the overprinting hypothesis as a new variant wasn't built from old existing material but rather new genetic information, not present in other genomes. However, we can not exclude possibility that this genomic fragment wasn't lost in other genomes.

The same mechanism can be attributed to the origination of ITFG3 (integrin alpha FG-GAP repeat containing 3) and RGS11 (regulator of G-protein signalling 11) overlap in primates. ITFG3 has two splice variants in primates and one of them is overlapping at 3' end with RGS11 gene. Figure 2 shows genomic organization of both genes in human; similar organization is observed in the chimpanzee and macaque genomes. In all non-primate species only the non-overlapping (shorter) variant is present. To exclude the possibility that the overlapping splice variant was missed in annotations in non primate genomes we investigated alignments of human and other genomes in two genomic browsers, Ensemble and UCSC browser. In both cases there was not a good alignment in the region occupied by the primate specific exon. Similarly, search against GenBank databases did not reveal any similarity between proteins encoded by the overlapping exon and any other non primate protein, genomic or EST sequence. This clearly shows that overlapping splice variant of gene encoding ITFG3 is lineage specific and arose recently after divergence of primates.

Another example of primate specific overlap that resulted from a new splice variant is a pair of genes THAP3, THAP domain containing, apoptosis associated protein 3, and DNAJC11, DnaJ (Hsp40) homolog, subfamily C, member 11. Both THAP3 and DNAJC11 homologs were found in

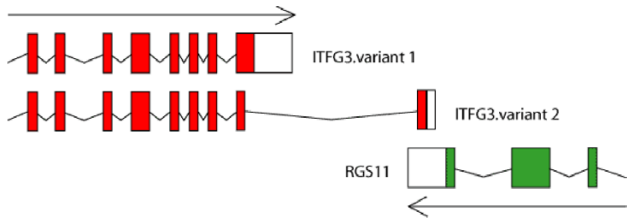

Figure 2
Genomic organization of ITFG3 and RGS11 in primates. In other mammals, although, these genes are neighbors on the same chromosome, variant 2 of ITGG3 is not observed and genes do not overlap. Coding sequences are colored and empty boxes denote 3'UTRs of both genes. 5' end of both genes were trimmed for the presentation purpose.

a majority of analyzed vertebrate species: human, macaque, chimpanzee, mouse, rat, dog, cow, opossum, and zebrafish. Interestingly, the THAP3 was missing in chicken, frog, and tetraodon. We couldn't identify the gene in these genomes by any standard comparative methods including BLASTn and tBLASTn. However, zebrafish is apparently not the only fish species THAP3 gene. EST sequences DT157701, DT154094, DT175180, DT175179, and DT157700 from *Pimephales promelas* show high similarity to zebrafish THAP3 protein (64–77% identity) and likely represent THAP3 transcript. In primates, THAP3 has two splice variants one of which overlaps with DNAJC11. In all other species only one, shorter variant is present and it is not overlapping with DNAJC11. Comparative analysis of genomic sequences in the region of the overlap shows that there is no conservation in this region and most likely the longer variant is primate specific. We also did not identify any EST sequence which could show the presence of longer (overlapping) variant in non primate vertebrates. This analysis led to conclusion that longer splice variant of THAP3 and THAP3-DNAJC11 overlap are primate specific.

Development of a new gene

The new splice variant origination seems to be one of the most common events leading to the lineage specific overlaps. However, our data strongly suggest that this is not the only case where we observe 'overprinting'. The powerful evidence that origination of new lineage specific genes plays a big role in the evolution of overlapping genes comes from data that many human genes do not have orthologs in other lineages. It is known that the many of human genes are not found in rodent [21], chicken [23] or fish [22] genomes, so our result could just reflect these findings. Interestingly, the fraction of human genes with missing orthologs is higher for overlapping genes than non-overlapping genes. Approximately 10% of human genes are missing in mouse genome while 27% of human overlapping genes are missing in mouse. Similarly, about 20% and 24% of human genes are not found in chicken and torafugu but this fraction is much higher in our studies: 44.9% and 46.17% in chicken and torafugu, respectively.

A confirmation for new gene origination as a source of gene overlap comes from the case of TMEM16C, transmembrane protein 16C, and MUC15, mucin 15; located on 11p14.2-3. Human TMEM16C has 27 exons spanning 331,865 base pairs. Two splice variants of MUC15 are embedded in the TMEM16C gene occupying introns 13 and 14 and overlapping, in the 3'UTR area, with exon 14 (Figure 3). Gene MUC15 is present only in mammalian genomes and there is experimental evidence that it is overlapping with TMEM16C at least in primates, rodents and cow. This gene is not present in any other lineages including chicken, xenopus and zebrafish and alignment of genomic sequences shows no conservation in areas covered by MUC15. However alignments at the protein level show some traces of similarity in chicken in MUC15 exon four of splice variant 2 and in xenopus in the same exon four in part of exon three. We could not detect any similarity in zebrafish as well as in invertebrates. The MUC15 sequence appears to be specific for vertebrates only and there are two possible evolutionary scenarios. In one, the sequence was present in early chordates and in the process of neutral mutation gained the coding potential which was used to build a new gene in mammalian ancestor. Another possibility is that this gene, was lost in majority of lineages with traces left in few genomes, and was maintained only in mammals. At any rate TMEM16C and MUC15 represent an overlap between an old gene TMEM16C and a newer mammalian specific gene (MUC15).

Changes in the gene structure

In the cases described above the gene overlap evolved through the origin of a new, longer splice variant. In many instances we observed a slightly different situation, a new variant arose and replaced the ancient one, so the number of variants was the same in analyzed lineages; however, they differ in their genomic organization. Examples of ACAT2-TCP1 [10] and MINK-CHRNE [13] overlaps are simple cases of changes in the gene structure where the most 3' exon was extended as a result of adopting the closest polyA signal after the original one was lost.

Example of BLZF1 gene (basic leucine zipper nuclear factor 1), overlapping at 3' end with the gene C1orf114 (open reading frame 114 on human chromosome 1, position 167603818–167663296) shows a more drastic shift in the gene structure. This overlap exists in the human and chimpanzee genomes but the two genes are neighbors, and do not overlap in other mammals including opossum where they are located on chromosome 2 about 6 kb apart. A similar arrangement is observed in chicken (chromosome 1, 2 kb separation) and in Xenopus (the same

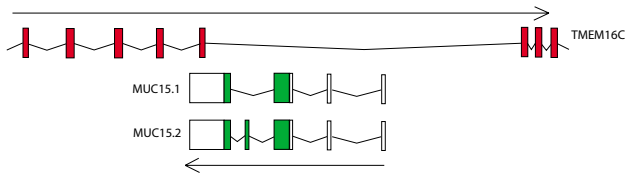

Figure 3
Genomic organization of TMEM16C and MUC15 in mammals. In chicken, Xenopus and zebrafish MUC15 is not observed.

scaffold, 9 kb separation). Interestingly, in zebrafish these two genes are located on different chromosomes – BLZF1 on chromosome 1 and a homolog of C1orf114 most likely is located on chromosome 8. Although there's no gene annotated in the cognate region, both human protein similarity and zebrafish EST alignments strongly suggest existence of the C1orf114 gene in this region. Figure 4 shows genomic organization of these overlapping genes in human and mouse. Analysis of multiple alignments of the region containing the last exon of BLZF1 in human and corresponding region in other vertebrates revealed that this fragment is not conserved among vertebrate lineages. Clearly there was an insertion in an ancient primate genome; a fragment belonging to the last, primate specific, exon does not align with non-primate genomes. Analysis of this fragment showed that the very 3' end of human BLZF1 is occupied by AluS, an old primate specific retroelement. Similarly, middle part of the last BLZF1 exon in mouse, contains rodent specific element B4A. Apparently independent insertions of repetitive elements occurred in both, primate and rodent, genomes and both are associated with exonification and new splice variant development.

Gene duplication and retrotransposition

Gene duplication is a common mechanism for the origin of new genes [24,25]. Retrotransposition is an interesting mechanism that allows a gene to move to a distant location on the same or different chromosome. Retro(pseudo)genes are products of reverse transcription of a spliced (mature) mRNA and they are characterized by lack of introns, presence of polyA track, and flanking direct repeats. Because they are copies of mature mRNAs, they usually lack promoters and cannot be transcribed. However, in some rare instances, after insertion near an existing promoter or exaptation of anonymous sequence as a promoter, they can gain transcriptional activity and create a new functional gene [24,26].

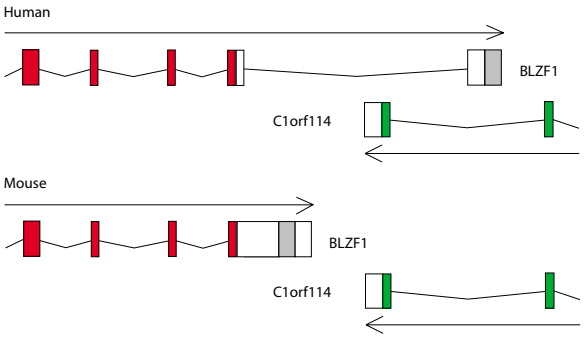

Figure 4
Genomic organization of human and mouse BLZF1 and C1orf114 genes. Shaded areas of 3'UTRs represent repetitive elements location.

An example of a new gene overlap due to formation of a new gene origination comes from the ribosomal protein RPS27 retrogene and TSPAN9 (tetraspanin 9, known also as NET-5) gene. RPS27 has two intron-containing paralogs: RPS27 and RPS27L, and both of them gave rise to multiple retrocopies in the human genome. We identified 24 retro(pseudo)genes of RPS27; ten of them are nested in another gene. Although multiple RPS27 retrogenes can be identified in other mammalian genomes, none of the host-nested gene pairs are the same in the human and rodent genomes. The aforementioned retrocopy of RPS27, nested in the human tetraspanin 9 gene, has an intact open reading frame and potentially encodes for 84 amino acid protein 100% identical to the spliced version of the gene on chromosome 1. We also identified two EST sequences, AV763564 and CD386048 that are 99% identical to this gene and show weaker similarity to other RPS27 genes, which may imply that this gene is expressed. However, because of relatively low quality of EST sequences, these results are not conclusive and further analysis would be required to confirm expression of this gene. This retrosequence is present in the human and chimp genomes but missing from orthologous location in macaque and all other vertebrates that we analyzed. This confirms recent origin of the RPS27 retrosequence and makes its expression assessment based on an intact ORF impossible. However, this example demonstrates a potential route to new overlaps in the vertebrate genomes.

Loss of gene overlaps

While young gene overlaps can arise from new splice variants, ancient overlaps can disappear due to the loss of the overlapping splice variant. In fact, we did observe such cases in our dataset that showed that analysis of narrow range of vertebrate or mammalian lineages may lead to false conclusions due to incompleteness of the data. The human NDEL1 gene coding for a thiol-activated peptidase and MYH10 gene that codes for myosin heavy chain 10 reside on chromosome 17 and they overlap in primates but not in rodents, chicken or fish. In primates there are three splice variants of NDEL1 and only one is overlapping with MYH10. In mouse and rat only two non overlapping transcripts are present, and in chicken we observe only one variant. The same variant as in chicken is present in zebrafish, however NDEL1 and MYH10 genes are separated. In addition, in zebrafish NDEL1 has two copies representing the same splice variant, one located at chromosome 12, position 35 35,318,946–35,338,461 and another located at chromosome 3 at position 56,442,046–56,469,437. MYH10 in zebrafish is on chromosome 6, position 5,082,544–5,106,092. Based on these observations the most parsimonious explanation

would be that one of these genes was translocated after fish divergence and later on the overlapping splice variant arose in primates. However, when we looked at the genomic arrangement of these genes in other species we observed that the dog genome has all three primate's splice variants. Further analysis showed that *Xenopus* has even more, four splice variants, two of them confirmed by cDNA and EST sequences and two predicted. One of the predicted variants is overlapping with MYH10 and protein coded by this variant shows high similarity to protein encoded by overlapping transcript observed in primates and in dog. Analysis of all known vertebrate transcripts and proteins encoded by them suggests that in a tetrapoda ancestor genome there were at least three splice variants of NDEL1. During the evolution one or two of them were lost in most lineages. Only primates retained three variants and all of them are confirmed by EST or cDNA sequences. In dog and *Xenopus* overlapping variant was predicted but it is shorter than the one in primates and up to date there is no experimental confirmation for its expression. However, the analysis of proteins coded by these variants (Figure 5) clearly shows that the overlapping splice variant in *Xenopus* is the same as the one in dogs and primates. In zebrafish both copies of NDEL1 represent variant 1 that seems to be the most common in vertebrate genomes. Variants of NDEL1 most probably evolved at early chordates because only one is present in *Ciona intestinalis*. Sequence divergence did not allow us to establish which, if any, of three variants is shared with *Ciona intestinalis*. Figure 6 shows the phylogenetic tree with all analyzed species. Bars to the right represent splice variants of NDEL1. The most likely mechanism of the variant loss in some lineages is a weakening of the splicing signal that leads to skipping of an exon. Similar effect can be observed in case when the other exon acquires signal so strong that dominates completely splicing and results in a constitutive exon. Another possibility is that this transcript is using signals from L1 elements which are in its 3' UTR. [27] These elements are in human sequence but are partially or completely lost in other genomes. However, more extensive analyses that would include some wet lab experiments are required in order to determine what the case is here.

Time of overlapping genes evolution

Equally important, to the mechanism leading to overlaps, is the time when particular event occurred. Two widely accepted hypotheses of vertebrate overlapping genes evolution [9,10] assume that explosion of gene overlaps occurred in early mammals. On the contrary, Dahary et al. [11] and Zhang et al. [19] suggested that most naturally occurring anti-transcripts observed in human represent ancient vertebrate gene overlaps. To answer this question, we looked for evidence of putative gene overlap before and after mammalian radiation. We studied cases where the pattern of gene arrangement differs in fish and rodent, i.e. two genes overlap in rodents but not in fish. As shown in Figure 7, genes arrangement in the chicken genome in some cases is similar to the one in rodents in other to the one in fish. For example genes CIO32 and ASB6 overlap in mammals only but genes RNF123 and GNPPB overlap also in chicken. In another example gene pair UBAP1 and KIF24 is overlapping in human but not overlapping, although located next to each other, in rodents and chicken. Only one of these genes, UBAP1 is present in fish. On the other hand, genes RFESD and SPATA9 over-

```
            ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
               335        345        355        365        375        385        395        405        415        425        435
human.1     FSRSGHTSFF DKG-AVNGFD PAPPPPGLGS SRPSSAPGML PLSV
chimp.1     FSRSGHTSFF DKG-AVNGFD PAPPPPGLGS SRPSSAPGML PLSV
mouse.1     CPRSGRATFF HKG-AVNGFD PAPPPPGLGS SRPSSAPGML PLSV
rat.1       CPRSGRATFF HKG-AVNGFD PAPPPPGLGS SRPSSAPGML PLSV
rabbit.1    FSRSGHTSFF DKG-AVNGFD PAPPPPGLGS SRPLSAPGML PLSV
dog.1       FSRSGHTSFF DKG-AVNGFD PAPPPPGLGS SRPSSAPGML PLSV
cow.1       FSRSGHTSFF DKG-AVNGFD PAPPPPGLGS SRPSSAPGML PLSV
chicken.1   YPHPGHTSFF DKG-AVNGFD QG--TPGLGA SRPSSAPGML PLSV
xenopus.1   YSHAGHTSFF DKG-AVNGYD ----PPGVLG SRPPSPPGML PLSV
zebrafish.  FSHSLHTTYF DKT-TMNGLD PGALSAAIAS PRAVSPPGLL PLSV
zebrafish.  YSHSLHTSYF DKARTVNGLD PGDGTHITAP PRSNSPSGLV -LSV
fugu.1      FPHALHTAYF DKT-TVNGLD PGALTAS--- -RAMSPPGML PLSV
human.2     FSRSGHTSFF DKG------- ---------- ---------- ----QEKVIF PTLFMGQ
chimp.2     FSRSGHTSFF DKG------- ---------- ---------- ----QEKVIF PTLFMGQ
mouse.2     CPRSGRATFF HKG------- ---------- ---------- ----QEKVIF PTLFMGQ
dog.2       FSRSGHTSFF DKG------- ---------- ---------- ----QEKVIF PTLFMGQ
cow.2       FSRSGHTSFF DKG------- ---------- ---------- ----QEKVIF PTLFMGQ
Xenopus.2   YSHAGHTSFF DKG------- ---------- ---------- ----QEKVIF PTLFMG
human.3     FSRSGHTSFF DKG------- ---------- ---------- ---------- -------SSS SCAIRASRRR FSSSSADLAG RGG-GDGSIP AQHSTGASGK APPPTLLLET
chimp.3     FSRSGHTSFF DKG------- ---------- ---------- ---------- -------SSS SCAIRASRRR FSSSSADLAG RGG-GDGSIP AQHSTGASGK APPPTLLLET
dog.3       FSRSGHTSFF DKG------- ---------- ---------- ---------- -------SSS SCAIRASRRR FSSSSADLAG AGREGDGSTP ---------- ----------
Xenopus.3   YSHAGHTSFF DKG------- ---------- ---------- ---------- -------SSS NWAILASRRL FSSRRADL-- ---------- ---------- ----------
Xenopus.4   YSHAGHTSFF DKG
ciona       TPHEGSTTDS SKN------- ---------- ---------- ---------- -------SS- -VKIRV
```

Figure 5
Alignments of vertebrate NDEL1 proteins coded by all splice variants.

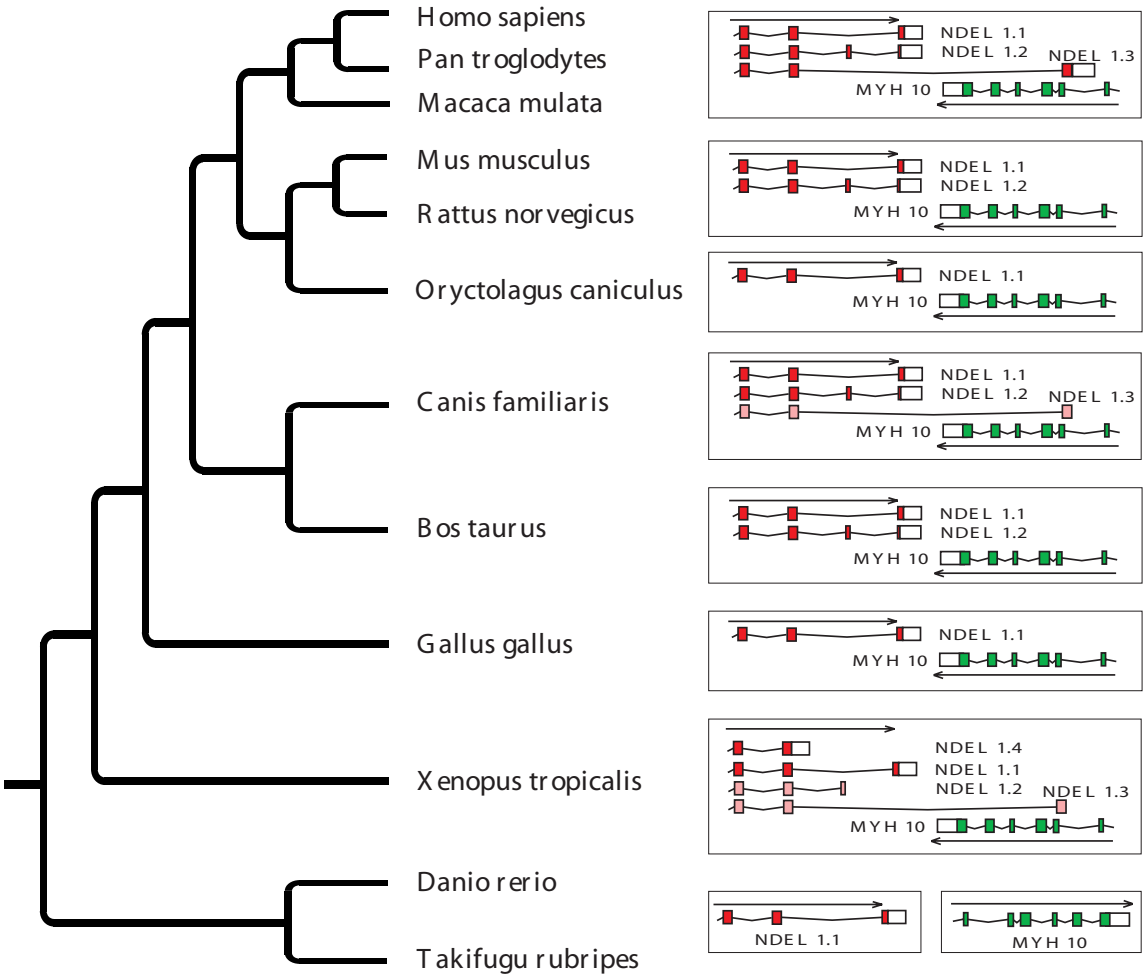

Figure 6
Phylogenetic tree with MYH10 and NDEL1 variants and genomic organization in vertebrates.

lap in human but not in rodents where they are located next to each other. In chicken, as well as in fish gene SPATA9 is not found. This demonstrates that the events leading to gene overlaps happened before and after mammalian radiation. The development of gene overlaps is a long, continuous process, not a 'big bang' that happened during a specific short period of vertebrate evolution. Also examples discussed in previous sections showed that gene overlaps were evolving at different stages of vertebrate evolution and could arise in early chordates, early mammals, or just recently in primates and other mammalian lineages.

Discussion

Although large scale studies of overlapping genes have been available since 2002 we still do not understand how these overlaps evolved and what, if any, is the functional meaning of sharing the genomic locus between genes in eukaryotic genomes. Results published so far show evidence for both, relatively new, lineage specific [9,10,13] as well as conserved overlaps among vertebrate [11,19] and even all eukaryotes [28] gene overlaps. However, none of the papers, even those with gene overlaps origin hypotheses, fully explains this evolutionary phenomenon. This study brings us a little closer to that goal and the major conclusion is that there's no single mechanism responsible for the overlap origination. In principle, any mechanism of a new exon or a new gene origination may lead to a gene overlap. In the light of presented results, we can conclude that the major forces in the overlapping genes evolution are transposition and exaptation – a process that gives rise to new genes or new variants from preexisting nucleotide sequences. [29]. UTR extension in the course of new polyA signal adoption [10,13] or new splice

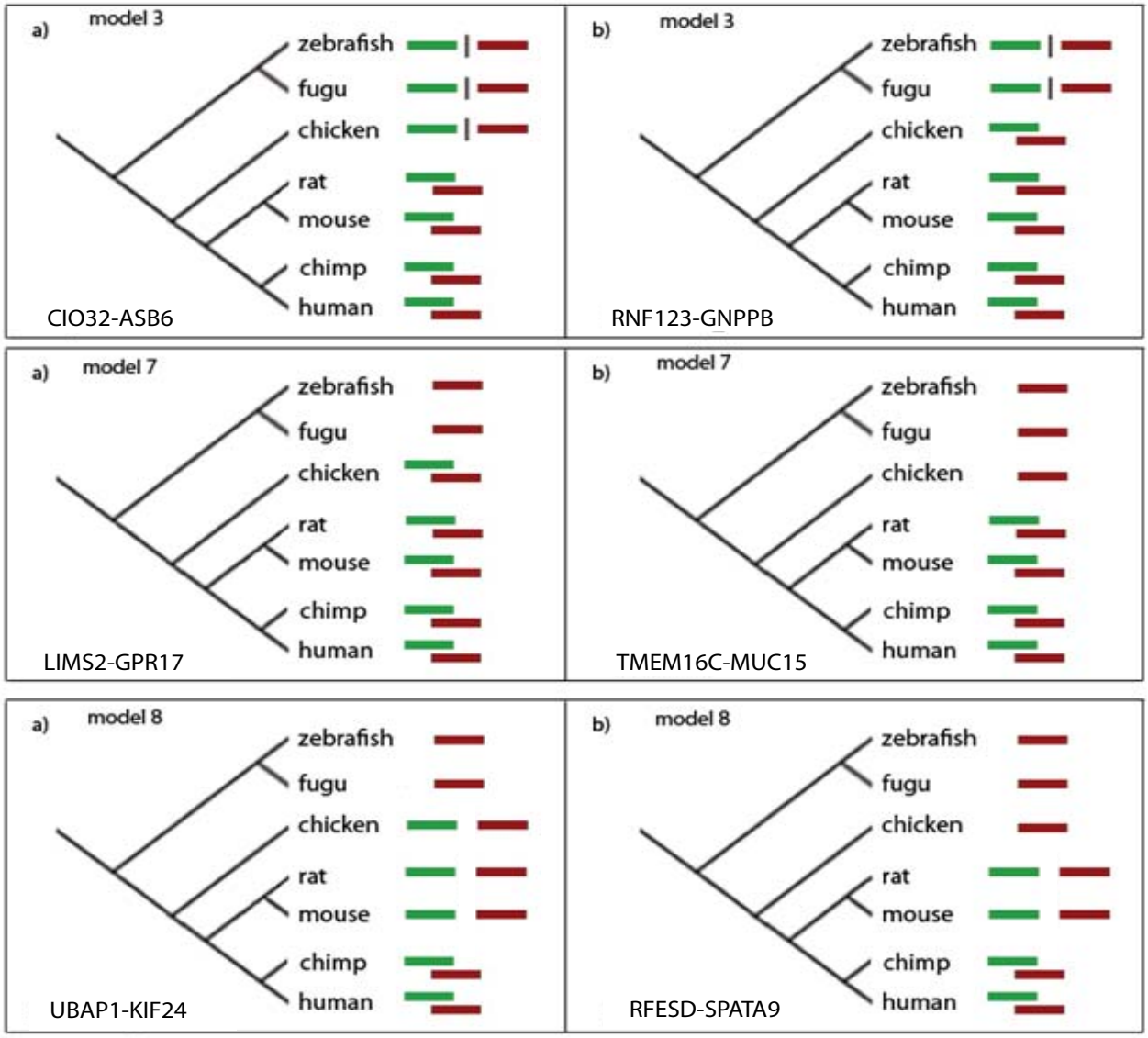

Figure 7
Examples of genomic arrangements of human overlapping genes orthologs in chicken. In left column genes arrangement in chicken resembles this in primates and rodents, in the right column genes in chicken are arranged the same way as in fish.

variant development [9] are perfect examples of exaptation. Another type of exaptation is building a new gene structure by adopting an inserted transposable element. Transposable elements are known to contribute to host gene regulation [30,31] or structure [32,33]. Our study on BLZF1 and C1orf114 showed that transposable elements also contribute to origin of a new class of genetic novelties, namely overlapping genes. The analyzed data also provided evidence that new gene origination is truly observed in the process of overlaps evolution supporting even further hypothesis by Keese and Gibbs [9]. 'Overprinting' hypothesis was constructed based on the new splice variant origination. Overlap between TMEM16C and mammalian specific gene MUC15 showed that overlaps may involve pairs of an old ancient gene and a new lineage specific gene. However, we cannot agree that this is true for all vertebrate overlaps as hypothesized. In many

of analyzed gene pairs both genes were old and conserved through eukaryotes.

Nevertheless, our study shows a number of gene overlaps that are lineage specific and are not conserved among vertebrates which supports our earlier studies of overlaps in human and mouse [12]. This is in a contradiction with the study of Dahary et al. [11] on human and fugu genomes that concludes that most human overlaps are ancient. However, they analyzed only those human genes that have identifiable orthologs in the fugu genome and therefore, young overlaps involving lineage specific genes were excluded from the study by definition. Also, their judgment was based on cases of overlapping human genes that were on average closer to each other in the fugu genome then other genes, and not based on true conserved overlaps. So, some of the gene pairs in fugu although close one to each other, are not necessary overlapping as is clear from our analysis.

In summary, we should emphasize that overlapping genes do not present any special case in regard to mechanisms of evolution. Events like gene translocation or exaptation, driving forces in genome evolution, are also common and major mechanisms in gene overlaps origin. There wasn't also any rapid origin or a 'big bang' of the overlapping genes after the split of bird and mammal lineages as suggested by Keese and Gibbs [9], nor are most of the human overlaps ancient as described by Dahary et al [11]. Birth as well as most likely death of gene overlaps is a continue process that occurred over the entire time of vertebrate evolution, similarly like any other genes arose or die over a long process of the eukaryotic genomes evolution [34].

Our results also imply that origin of overlapping genes is not an issue of saving space and contracting genomes size. Although there are some implications on functional importance of overlapping genes, the present analysis shows that most gene overlaps evolve stochastically, the same way as other genomic features, and without any positive pressure on the overlap presence. If overlaps have some functional meaning it is not a common case and most likely this function evolved by chance as a consequence of new genes arrangement.

This study also demonstrates that in order to fully understand the evolution of overlapping genes one has to study many genomes in minute details. Studies on a limited number of species may lead to false conclusions as shown in the case of NDEL1 and in many other cases we investigated during this study. Many gene pairs were moved from one category to another as a result of detailed examination of annotations and additional analysis. This shows that although human and other genomes are considered to be complete, their annotation is still far from final and in many cases cannot be trusted. Therefore, careful examination of any gene pair by a human expert followed by, in an ideal world, some wet-lab experiments is a key to sound results. We are very well aware that the present study did not solve all the questions regarding overlapping genes evolution and their origins. However, it did shed a light on how some of these overlaps evolved, provided a strong confirmation for lineage specific overlaps, and delivered firsthand evidence of gene overlap loss in the vertebrate lineage.

Methods

Sequence data

Assembled sequences and annotations of seven analyzed genomes were downloaded from Ensembl [20] and stored in a local mySQL database. We used following versions of the genomes: human-24.34e (NCBI 34), chimp-24.1 (CHIMP 1), mouse-24.33 (NCBI m33), rat-24.3c (RGSC 3.1), chicken-24.1a (WASHUC 1), fugu-24.2c (FUGU 2.0), and zebrafish-24.4 (Zv 4).

Identification of the overlapping genes

For practical reasons, we applied an operational definition of a gene, as a part of the genomic region from the beginning to the end of an annotated transcript. Any two genes, defined as above, whose coordinates overlap and are transcribed from the different DNA strand, are considered as overlapping.

Identification of orthologous genes and mapping information

Orthology inference was done based on any two genomes homology information provided in Ensembl. The set of overlapping genes for a given species was always a starting point for each orthology analysis. As a result seven by seven orthology matrix was created. It is important to stress that orthology relationship provided by Ensembl is not a simple one-to-one relationship. Whenever lineage specific gene duplication is detected one-to-many orthologs are provided. In these cases, each of several orthologs was checked for the overlaps. The detailed description of the method is available at Ensembl webpage [35]. Additionally, we used conserved synteny information of the neighboring genes to enhance reliability of the orthology inference. However, not all the genes have had their orthologs listed. In these cases, we assumed that a cognate gene is missing from a given genome.

The mapping information of each orthologous gene was downloaded from the Ensemble. For each pair of overlapping genes in one genome, e.g. human, spatial relationship of their orthologs in the other six genomes was checked based on existing annotation.

Extending neighboring but not overlapping genes

We mapped TIGR gene indices [36] to all neighboring but not overlapping orthologs of human overlapping genes to check for possible extensions. In each case, we extracted genomic fragments containing a pair of neighboring genes and BLAST [37] against corresponding TGI sequences. Next we mapped transcripts to genomic fragments together with TGI sequences obtained from BLAST search. Only sequences showing similarity over 98% and fully aligning with the genomic fragment were used in order to avoid false positive hits to repetitive elements and ESTs from related genes. Results were stored in ASN.1 format and examined in Sequin [38].

Multiple alignments

MultiZ alignments of genomic sequences were obtained from UCSC genome browser [39]. Protein multiple alignments were constructed using Clustalw [40].

Acknowledgements

We gratefully thank Stephen Schaeffer for reading the manuscript and for his valuable comments, and Narayanan Veeraraghavan for his assistance.

References

1. Makalowska I, Lin CF, Makalowski W: **Overlapping genes in vertebrate genomes.** *Comput Biol Chem* 2005, **29(1):**1-12.
2. Rougeulle C, Heard E: **Antisense RNA in imprinting: spreading silence through Air.** *Trends Genet* 2002, **18(9):**434-437.
3. Brantl S: **Antisense-RNA regulation and RNA interference.** *Biochim Biophys Acta* 2002, **1575(1-3):**15-25.
4. Prescott EM, Proudfoot NJ: **Transcriptional collision between convergent genes in budding yeast.** *Proc Natl Acad Sci U S A* 2002, **99(13):**8796-8801.
5. Hastings ML, Ingle HA, Lazar MA, Munroe SH: **Post-transcriptional regulation of thyroid hormone receptor expression by cis-acting sequences and a naturally occurring antisense RNA.** *J Biol Chem* 2000, **275(15):**11507-11513.
6. Ogawa Y, Lee JT: **Antisense regulation in X inactivation and autosomal imprinting.** *Cytogenet Genome Res* 2002, **99(1-4):**59-65.
7. Marcelino J, Carpten JD, Suwairi WM, Gutierrez OM, Schwartz S, Robbins C, Sood R, Makalowska I, Baxevanis A, Johnstone B, Laxer RM, Zemel L, Kim CA, Herd JK, Ihle J, Williams C, Johnson M, Raman V, Alonso LG, Brunoni D, Gerstein A, Papadopoulos N, Bahabri SA, Trent JM, Warman ML: **CACP, encoding a secreted proteoglycan, is mutated in camptodactyly-arthropathy-coxa vara-pericarditis syndrome.** *Nature genetics* 1999, **23(3):**319-322.
8. Lamason RL, Mohideen MA, Mest JR, Wong AC, Norton HL, Aros MC, Jurynec MJ, Mao X, Humphreville VR, Humbert JE, Sinha S, Moore JL, Jagadeeswaran P, Zhao W, Ning G, Makalowska I, McKeigue PM, O'Donnell D, Kittles R, Parra EJ, Mangini NJ, Grunwald DJ, Shriver MD, Canfield VA, Cheng KC: **SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans.** *Science* 2005, **310(5755):**1782-1786.
9. Keese PK, Gibbs A: **Origins of genes: "big bang" or continuous creation?** *Proc Natl Acad Sci U S A* 1992, **89(20):**9489-9493.
10. Shintani S, O'HUigin C, Toyosawa S, Michalova V, Klein J: **Origin of gene overlap: the case of TCP1 and ACAT2.** *Genetics* 1999, **152(2):**743-754.
11. Dahary D, Elroy-Stein O, Sorek R: **Naturally occurring antisense: transcriptional leakage or real overlap?** *Genome Res* 2005, **15(3):**364-368.
12. Veeramachaneni V, Makalowski W, Galdzicki M, Sood R, Makalowska I: **Mammalian overlapping genes: the comparative perspective.** *Genome Res* 2004, **14(2):**280-286.
13. Dan I, Watanabe NM, Kajikawa E, Ishida T, Pandey A, Kusumi A: **Overlapping of MINK and CHRNE gene loci in the course of mammalian evolution.** *Nucleic Acids Res* 2002, **30(13):**2906-2910.
14. Kasper G, Taudien S, Staub E, Mennerich D, Rieder M, Hinzmann B, Dahl E, Schwidetzky U, Rosenthal A, Rump A: **Different structural organization of the encephalopsin gene in man and mouse.** *Gene* 2002, **295(1):**27-32.
15. Steigele S, Nieselt K: **Open reading frames provide a rich pool of potential natural antisense transcripts in fungal genomes.** *Nucleic Acids Res* 2005, **33(16):**5034-5044.
16. Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, Patapoutian A, Hampton GM, Schultz PG, Hogenesch JB: **Large-scale analysis of the human and mouse transcriptomes.** *Proc Natl Acad Sci U S A* 2002, **99(7):**4465-4470.
17. Yelin R, Dahary D, Sorek R, Levanon EY, Goldstein O, Shoshan A, Diber A, Biton S, Tamir Y, Khosravi R, Nemzer S, Pinner E, Walach S, Bernstein J, Savitsky K, Rotman G: **Widespread occurrence of antisense transcription in the human genome.** *Nat Biotechnol* 2003, **21(4):**379-386.
18. Lehner B, Williams G, Campbell RD, Sanderson CM: **Antisense transcripts in the human genome.** *Trends Genet* 2002, **18(2):**63-65.
19. Zhang Y, Liu XS, Liu QR, Wei L: **Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species.** *Nucleic Acids Res* 2006, **34(12):**3465-3475.
20. Birney E, Andrews TD, Bevan P, Caccamo M, Chen Y, Clarke L, Coates G, Cuff J, Curwen V, Cutts T, Down T, Eyras E, Fernandez-Suarez XM, Gane P, Gibbins B, Gilbert J, Hammond M, Hotz HR, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Keenan S, Lehvaslaiho H, McVicker G, Melsopp C, Meidl P, Mongin E, Pettett R, Potter S, Proctor G, Rae M, Searle S, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Ureta-Vidal A, Woodwark KC, Cameron G, Durbin R, Cox A, Hubbard T, Clamp M: **An overview of Ensembl.** *Genome Res* 2004, **14(5):**925-928.
21. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE, Okwuonu G, Hines S, Lewis L, DeRamo C, Delgado O, Dugan-Rocha S, Miner G, Morgan M, Hawes A, Gill R, Celera, Holt RA, Adams MD, Amanatides PG, Baden-Tillson H, Barnstead M, Chin S, Evans CA, Ferriera S, Fosler C, Glodek A, Gu Z, Jennings D, Kraft CL, Nguyen T, Pfannkoch CM, Sitter C, Sutton GG, Venter JC, Woodage T, Smith D, Lee HM, Gustafson E, Cahill P, Kana A, Doucette-Stamm L, Weinstock K, Fechtel K, Weiss RB, Dunn DM, Green ED, Blakesley RW, Bouffard GG, De Jong PJ, Osoegawa K, Zhu B, Marra M, Schein J, Bosdet I, Fjell C, Jones S, Krzywinski M, Mathewson C, Siddiqui A, Wye N, McPherson J, Zhao S, Fraser CM, Shetty J, Shatsman S, Geer K, Chen Y, Abramzon S, Nierman WC, Havlak PH, Chen R, Durbin KJ, Egan A, Ren Y, Song XZ, Li B, Liu Y, Qin X, Cawley S, Worley KC, Cooney AJ, D'Souza LM, Martin K, Wu JQ, Gonzalez-Garay ML, Jackson AR, Kalafus KJ, McLeod MP, Milosavljevic A, Virk D, Volkov A, Wheeler DA, Zhang Z, Bailey JA, Eichler EE, Tuzun E, Birney E, Mongin E, Ureta-Vidal A, Woodwark C, Zdobnov E, Bork P, Suyama M, Torrents D, Alexandersson M, Trask BJ, Young JM, Huang H, Wang H, Xing H, Daniels S, Gietzen D, Schmidt J, Stevens K, Vitt U, Wingrove J, Camara F, Mar Alba M, Abril JF, Guigo R, Smit A, Dubchak I, Rubin EM, Couronne O, Poliakov A, Hubner N, Ganten D, Goesele C, Hummel O, Kreitler T, Lee YA, Monti J, Schulz H, Zimdahl H, Himmelbauer H, Lehrach H, Jacob HJ, Bromberg S, Gullings-Handley J, Jensen-Seaman MI, Kwitek AE, Lazar J, Pasko D, Tonellato PJ, Twigger S, Ponting CP, Duarte JM, Rice S, Goodstadt L, Beatson SA, Emes RD, Winter EE, Webber C, Brandt P, Nyakatura G, Adetobi M, Chiaromonte F, Elnitski L, Eswara P, Hardison RC, Hou M, Kolbe D, Makova K, Miller W, Nekrutenko A, Riemer C, Schwartz S, Taylor J, Yang S, Zhang Y, Lindpaintner K, Andrews TD, Caccamo M, Clamp M, Clarke L, Curwen V, Durbin R, Eyras E, Searle SM, Cooper GM, Batzoglou S, Brudno M, Sidow A, Stone EA, Venter JC, Payseur BA, Bourque G, Lopez-Otin C, Puente XS, Chakrabarti K, Chatterji S, Dewey C, Pachter L, Bray N, Yap VB, Caspi A, Tesler G, Pevzner PA, Haussler D, Roskin KM, Baertsch R, Clawson H, Furey TS, Hinrichs AS, Karolchik D, Kent WJ, Rosenbloom KR, Trumbower H, Weirauch M, Cooper DN, Stenson PD, Ma B, Brent M, Arumugam M, Shteynberg D, Copley RR, Taylor MS, Riethman H, Mudunuri U, Peterson J, Guyer M, Felsenfeld A, Old S, Mockrin S, Collins F: **Genome sequence of the Brown Norway rat yields insights into mammalian evolution.** *Nature* 2004, **428(6982):**493-521.
22. Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, Gelpke MD, Roach J, Oh T, Ho IY, Wong M, Detter C, Verhoef F, Predki P, Tay A, Lucas S, Richardson

P, Smith SF, Clark MS, Edwards YJ, Doggett N, Zharkikh A, Tavtigian SV, Pruss D, Barnstead M, Evans C, Baden H, Powell J, Glusman G, Rowen L, Hood L, Tan YH, Elgar G, Hawkins T, Venkatesh B, Rokhsar D, Brenner S: **Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes.** *Science* 2002, **297(5585):**1301-1310.

23. **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.** *Nature* 2004, **432(7018):**695-716.
24. Long M, Betran E, Thornton K, Wang W: **The origin of new genes: glimpses from the young and old.** *Nat Rev Genet* 2003, **4(11):**865-875.
25. Ohno S: **Evolution by gene duplication.** Berlin; New York , Springer-Verlag; 1970.
26. Brosius J: **RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements.** *Gene* 1999, **238(1):**115-134.
27. Makalowski W: **Genomics. Not junk after all.** *Science* 2003, **300(5623):**1246-1247.
28. van Duin M, van Den Tol J, Hoeijmakers JH, Bootsma D, Rupp IP, Reynolds P, Prakash L, Prakash S: **Conserved pattern of antisense overlapping transcription in the homologous human ERCC-1 and yeast RAD10 DNA repair gene regions.** *Mol Cell Biol* 1989, **9(4):**1794-1798.
29. Brosius J Gould, SJ: **On "genomencature": a comprehensive (and respectful) taxonomy for pseudogenes amd other "junk DNA".** *Proc Natl Acad Sci* 1992, **89(22):**10706-10710.
30. Thornburg BG, Gotea V, Makalowski W: **Transposable elements as a significant source of transcription regulating signals.** *Gene* 2006, **365:**104-110.
31. Romanish MT, Lock WM, de Lagemaat LN, Dunn CA, Mager DL: **Repeated Recruitment of LTR Retrotransposons as Promoters by the Anti-Apoptotic Locus NAIP during Mammalian Evolution.** *PLoS Genet* 2007, **3(1):**e10.
32. Lorenc A, Makalowski W: **Transposable elements and vertebrate protein diversity.** *Genetica* 2003, **118(2-3):**183-191.
33. Gotea V, Makalowski W: **Do transposable elements really contribute to proteomes?** *Trends Genet* 2006, **22(5):**260-267.
34. Nei M, Rooney AP: **Concerted and birth-and-death evolution of multigene families.** *Annual review of genetics* 2005, **39:**121-152.
35. Ensembl: **Gene Orthology/Paralogy prediction method.** [http://www.ensembl.org/info/data/compara/homology_method.html].
36. Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, Tsai J, Quackenbush J: **TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets.** *Bioinformatics* 2003, **19(5):**651-652.
37. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215(3):**403-410.
38. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank.** *Nucleic Acids Res* 2005, **33(Database issue):**D34-8.
39. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, Hillman-Jackson J, Kuhn RM, Pedersen JS, Pohl A, Raney BJ, Rosenbloom KR, Siepel A, Smith KE, Sugnet CW, Sultan-Qurraie A, Thomas DJ, Trumbower H, Weber RJ, Weirauch M, Zweig AS, Haussler D, Kent WJ: **The UCSC Genome Browser Database: update 2006.** *Nucleic Acids Res* 2006, **34(Database issue):**D590-8.
40. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22(22):**4673-4680.

Appendix C

EVOLUTION OF GENES AND GENOMES ON THE *DROSOPHILA* PHYLOGENY

Evolution of genes and genomes on the *Drosophila* phylogeny

Drosophila 12 Genomes Consortium*

Comparative analysis of multiple genomes in a phylogenetic framework dramatically improves the precision and sensitivity of evolutionary inference, producing more robust results than single-genome analyses can provide. The genomes of 12 *Drosophila* species, ten of which are presented here for the first time (*sechellia*, *simulans*, *yakuba*, *erecta*, *ananassae*, *persimilis*, *willistoni*, *mojavensis*, *virilis* and *grimshawi*), illustrate how rates and patterns of sequence divergence across taxa can illuminate evolutionary processes on a genomic scale. These genome sequences augment the formidable genetic tools that have made *Drosophila melanogaster* a pre-eminent model for animal genetics, and will further catalyse fundamental research on mechanisms of development, cell biology, genetics, disease, neurobiology, behaviour, physiology and evolution. Despite remarkable similarities among these *Drosophila* species, we identified many putatively non-neutral changes in protein-coding genes, non-coding RNA genes, and *cis*-regulatory regions. These may prove to underlie differences in the ecology and behaviour of these diverse species.

As one might expect from a genus with species living in deserts, in the tropics, on chains of volcanic islands and, often, commensally with humans, *Drosophila* species vary considerably in their morphology, ecology and behaviour[1]. Species in this genus span a wide range of global distributions: the 12 sequenced species originate from Africa, Asia, the Americas and the Pacific Islands, and also include cosmopolitan species that have colonized the planet (*D. melanogaster* and *D. simulans*) as well as closely related species that live on single islands (*D. sechellia*)[2]. A variety of behavioural strategies is also encompassed by the sequenced species, ranging in feeding habit from generalist, such as *D. ananassae*, to specialist, such as *D. sechellia*, which feeds on the fruit of a single plant species.

Despite this wealth of phenotypic diversity, *Drosophila* species share a distinctive body plan and life cycle. Although only *D. melanogaster* has been extensively characterized, it seems that the most important aspects of the cellular, molecular and developmental biology of these species are well conserved. Thus, in addition to providing an extensive resource for the study of the relationship between sequence and phenotypic diversity, the genomes of these species provide an excellent model for studying how conserved functions are maintained in the face of sequence divergence. These genome sequences provide an unprecedented dataset to contrast genome structure, genome content, and evolutionary dynamics across the well-defined phylogeny of the sequenced species (Fig. 1).

Genome assembly, annotation and alignment

Genome sequencing and assembly. We used the previously published sequence and updated assemblies for two *Drosophila* species, *D. melanogaster*[3,4] (release 4) and *D. pseudoobscura*[5] (release 2), and generated DNA sequence data for 10 additional *Drosophila* genomes by whole-genome shotgun sequencing[6,7]. These species were chosen to span a wide variety of evolutionary distances, from closely related pairs such as *D. sechellia*/*D. simulans* and *D. persimilis*/*D. pseudoobscura* to the distantly related species of the *Drosophila* and *Sophophora* subgenera. Whereas the time to the most recent common ancestor of the sequenced species may seem small on an evolutionary timescale, the evolutionary divergence spanned by the genus *Drosophila* exceeds that of the entire mammalian radiation when generation time is taken into account, as discussed further in ref. 8. We sequenced seven of the new species (*D. yakuba*, *D. erecta*, *D. ananassae*, *D. willistoni*, *D. virilis*, *D. mojavensis* and *D. grimshawi*) to deep coverage (8.4× to 11.0×) to produce high quality draft sequences. We sequenced two species, *D. sechellia* and *D. persimilis*, to intermediate coverage (4.9× and 4.1×, respectively) under the assumption that the availability of a sister species sequenced to high coverage would obviate the need for deep sequencing without sacrificing draft genome quality. Finally, seven inbred strains of *D. simulans* were sequenced to low coverage (2.9× coverage from w^{501} and ~1× coverage of six other strains) to provide population variation data[9]. Further details of the sequencing strategy can be found in Table 1, Supplementary Table 1 and section 1 in Supplementary Information.

We generated an initial draft assembly for each species using one of three different whole-genome shotgun assembly programs (Table 1). For *D. ananassae*, *D. erecta*, *D. grimshawi*, *D. mojavensis*, *D. virilis* and *D. willistoni*, we also generated secondary assemblies; reconciliation of these with the primary assemblies resulted in a 7–30% decrease in the estimated number of misassembled regions and a 12–23% increase in the N50 contig size[10] (Supplementary Table 2). For *D. yakuba*, we generated 52,000 targeted reads across low-quality regions and gaps to improve the assembly. This doubled the mean contig and scaffold sizes and increased the total fraction of high quality bases (quality score (Q) > 40) from 96.5% to 98.5%. We improved the initial 2.9× *D. simulans* w^{501} whole-genome shotgun assembly by filling assembly gaps with contigs and unplaced reads from the ~1× assemblies of the six other *D. simulans* strains, generating a 'mosaic' assembly (Supplementary Table 3). This integration markedly improved the *D. simulans* assembly: the N50 contig size of the mosaic assembly, for instance, is more than twice that of the initial w^{501} assembly (17 kb versus 7 kb).

Finally, one advantage of sequencing genomes of multiple closely related species is that these evolutionary relationships can be exploited to dramatically improve assemblies. *D. yakuba* and *D. simulans* contigs and scaffolds were ordered and oriented using pairwise alignment to the well-validated *D. melanogaster* genome

*A list of participants and affiliations appears at the end of the paper.

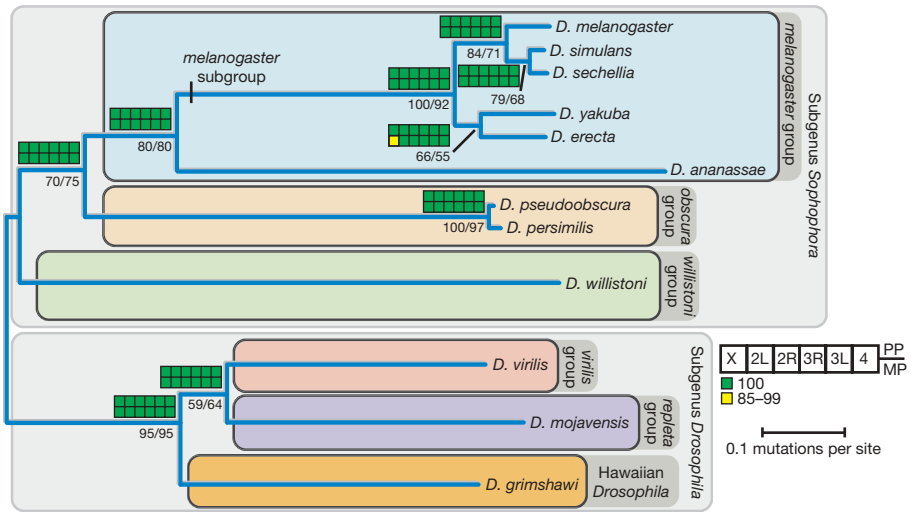

Figure 1 | Phylogram of the 12 sequenced species of *Drosophila*. Phylogram derived using pairwise genomic mutation distances and the neighbour-joining method[152,153]. Numbers below nodes indicate the per cent of genes supporting a given relationship, based on evolutionary distances estimated from fourfold-degenerate sites (left of solidus) and second codon positions (right of solidus). Coloured blocks indicate support from bayesian (posterior probability (PP), upper blocks) and maximum parsimony (MP; bootstrap values, lower blocks) analyses of data partitioned by chromosome arm. Branch lengths indicate the number of mutations per site (at fourfold-degenerate sites) using the ordinary least squares method. See ref. 154 for a discussion of the uncertainties in the *D. yakuba*/*D. erecta* clade.

sequence (Supplementary Information section 2). Likewise, the 4–5× *D. persimilis* and *D. sechellia* assemblies were improved by assisted assembly using the sister species (*D. pseudoobscura* and *D. simulans*, respectively) to validate both alignments between reads and linkage information. For the remaining species, comparative syntenic information, and in some cases linkage information, were also used to pinpoint locations of probable genome misassembly, to assign assembly scaffolds to chromosome arms and to infer their order and orientation along euchromatic chromosome arms, supplementing experimental analysis based on known markers (A. Bhutkar, S. Russo, S. Schaeffer, T. F. Smith and W. M. Gelbart, personal communication) (Supplementary Information section 2).

The mitochondrial (mt)DNA of *D. melanogaster*, *D. sechellia*, *D. simulans* (siI), *D. mauritiana* (maII) and *D. yakuba* have been previously sequenced[11,12]. For the remaining species (except *D. pseudoobscura*, the DNA from which was prepared from embryonic nuclei), we were able to assemble full mitochondrial genomes, excluding the A+T-rich control region (Supplementary Information section 2)[13]. In addition, the genome sequences of three *Wolbachia* endosymbionts (*Wolbachia wSim, Wolbachia wAna* and *Wolbachia wWil)* were assembled from trace archives, in *D. simulans*, *D. ananassae* and *D. willistoni*, respectively[14]. All of the genome sequences described here are available in FlyBase (www.flybase.org) and GenBank (www.ncbi.nlm.nih.gov) (Supplementary Tables 4 and 5).

Repeat and transposable element annotation. Repetitive DNA sequences such as transposable elements pose challenges for whole-genome shotgun assembly and annotation. Because the best approach to transposable element discovery and identification is still an active and unresolved research question, we used several repeat libraries and computational strategies to estimate the transposable element/repeat content of the 12 *Drosophila* genome assemblies (Supplementary Information section 3). Previously curated transposable element libraries in *D. melanogaster* provided the starting point for our analysis; to limit the effects of ascertainment bias, we also developed *de novo* repeat libraries using PILER-DF[15,16] and ReAS[17]. We used four transposable element/repeat detection methods (RepeatMasker, BLASTER-TX, RepeatRunner and CompTE) in conjunction with these transposable element libraries to identify repetitive elements in non-*melanogaster* species. We assessed the accuracy of each method by calibration with the estimated 5.5% transposable element content in the *D. melanogaster* genome, which is based on a high-resolution transposable element annotation[18] (Supplementary Fig. 1). On the basis of our results, we suggest a hybrid strategy for new genome sequences, employing translated BLAST with general transposable element libraries and RepeatMasker with species-specific ReAS libraries to estimate the upper and lower bound on transposable element content.

Protein-coding gene annotation. We annotated protein-coding sequences in the 11 non-*melanogaster* genomes, using four different *de novo* gene predictors (GeneID[19], SNAP[20], N-SCAN[21] and CONTRAST[22]); three homology-based predictors that transfer annotations from *D. melanogaster* (GeneWise[23], Exonerate[24], GeneMapper[25]); and one predictor that combined *de novo* and homology-based evidence (Gnomon[26]). These gene prediction sets

Table 1 | A summary of sequencing and assembly properties of each new genome

Final assembly	Genome centre	Q20 coverage (×)	Assembly size (Mb)	No. of contigs ≥2 kb	N50 contig ≥2 kb (kb)	Per cent of base pairs with quality >Q40
D. simulans	WUGSC*	2.9	137.8	10,843	17	90.3
D. sechellia	Broad†	4.9	166.6	9,713	43	90.6
D. yakuba	WUGSC*	9.1	165.7	6,344	125	98.5
D. erecta	Agencourt†	10.6	152.7	3,283	458	99.2
D. ananassae	Agencourt†	8.9	231.0	8,155	113	98.5
D. persimilis	Broad†	4.1	188.4	14,547	20	93.3
D. willistoni	JCVI‡	8.4	235.5	6,652	197	97.4
D. virilis	Agencourt†	8.0	206.0	5,327	136	98.7
D. mojavensis	Agencourt†	8.2	193.8	5,734	132	98.6
D. grimshawi	Agencourt†	7.9	200.5	9,632	114	97.1

Contigs, contiguous sequences not interrupted by gaps; N50, the largest length *L* such that 50% of all nucleotides are contained in contigs of size ≥*L*. The Q20 coverage of contigs is based on the number of assembled reads, average Q20 readlength and the assembled size excluding gaps. Assemblers used: *PCAP6, †ARACHNE4.5 and ‡Celera Assembler 7.

Table 2 | A summary of annotated features across all 12 genomes

	Protein-coding gene annotations			Non-coding RNA annotations				Repeat coverage (%)*	Genome size (Mb; assembly†/flow cytometry‡)
	Total no. of protein- coding genes (per cent with *D. melanogaster* homologue)	Coding sequence/ intron (Mb)	tRNA (pseudo)	snoRNA	miRNA	rRNA (5.8S + 5S)	snRNA		
D. melanogaster	13,733 (100%)	38.9/21.8	297 (4)	250	78	101	28	5.35	118/200
D. simulans	15,983 (80.0%)	45.8/19.6	268 (2)	246	70	72	32	2.73	111/162
D. sechellia	16,884 (81.2%)	47.9/21.9	312 (13)	242	78	133	30	3.67	115/171
D. yakuba	16,423 (82.5%)	50.8/22.9	380 (52)	255	80	55	37	12.04	127/190
D. erecta	15,324 (86.4%)	49.1/22.0	286 (2)	252	81	101	38	6.97	134/135
D. ananassae	15,276 (83.0%)	57.3/22.3	472 (165)	194	76	134	29	24.93	176/217
D. pseudoobscura	16,363 (78.2%)	49.7/24.0	295 (1)	203	73	55	31	2.76	127/193
D. persimilis	17,325 (72.6%)	54.0/21.9	306 (1)	199	75	80	31	8.47	138/193
D. willistoni	15,816 (78.8%)	65.4/23.5	484 (164)	216	77	76	37	15.57	187/222
D. virilis	14,680 (82.7%)	57.9/21.7	279 (2)	165	74	294	31	13.96	172/364
D. mojavensis	14,849 (80.8%)	57.8/21.9	267 (3)	139	71	74	30	8.92	161/130
D. grimshawi	15,270 (81.3%)	54.9/22.5	261 (1)	154	82	70	32	2.84	138/231

* Repeat coverage calculated as the fraction of scaffolds >200 kb covered by repeats, estimated as the midpoint between BLASTER-tx + PILER and RepeatMasker + ReAS (Supplementary Information section 3). †Total genome size estimated as the sum of base pairs in genomic scaffold >200,000 bp. ‡Genome size estimates based on flow cytometry[38].

were combined using GLEAN, a gene model combiner that chooses the most probable combination of start, stop, donor and acceptor sites from the input predictions[27,28]. All analyses reported here, unless otherwise noted, relied on a reconciled consensus set of predicted gene models—the GLEAN-R set (Table 2, and Supplementary Information section 4.1).

Quality of gene models. As the first step in assessing the quality of the GLEAN-R gene models, we used expression data from microarray experiments on adult flies, with arrays custom-designed for *D. simulans*, *D. yakuba*, *D. ananassae*, *D. pseudoobscura*, *D. virilis* and *D. mojavensis*[29] (GEO series GSE6640; Supplementary Information section 4.2). We detected expression significantly above negative controls (false-discovery-rate-corrected Mann–Whitney U (MWU) $P < 0.001$) for 77–93% of assayed GLEAN-R models, representing 50–68% of the total GLEAN-R predictions in each species (Supplementary Table 6). Evolutionarily conserved gene models are much more likely to be expressed than lineage-specific ones (Fig. 2). Although these data cannot confirm the detailed structure of gene models, they do suggest that the majority of GLEAN-R models contain sequence that is part of a poly-adenylated transcript. Approximately 20% of transcription in *D. melanogaster* seems to be unassociated with protein-coding genes[30], and our microarray experiments fail to detect conditionally expressed genes. Thus, transcript abundance cannot conclusively establish the presence or absence of a protein-coding gene. Nonetheless, we believe these expression data increase our confidence in the reliability of the GLEAN-R models, particularly those supported by homology evidence (Fig. 2).

Because the GLEAN-R gene models were built using assemblies that were not repeat masked, it is likely that some proportion of gene models are false positives corresponding to coding sequences of transposable elements. We used RepeatMasker with *de novo* ReAS libraries and PFAM structural annotations of the GLEAN-R gene set to flag potentially transposable element-contaminated gene models (Supplementary Information section 4.2). These procedures suggest that 5.6–32.3% of gene models in non-*melanogaster* species correspond to protein-coding content derived from transposable elements (Supplementary Table 7); these transposable element-contaminated gene models are almost exclusively confined to gene predictions without strong homology support (Fig. 2). Transposable element-contaminated gene models are excluded from the final gene prediction set used for subsequent analysis, unless otherwise noted.

Homology assignment. Two independent approaches were used to assign orthology and paralogy relationships among euchromatic *D. melanogaster* gene models and GLEAN-R predictions. The first approach was a fuzzy reciprocal BLAST (FRB) algorithm, which is an

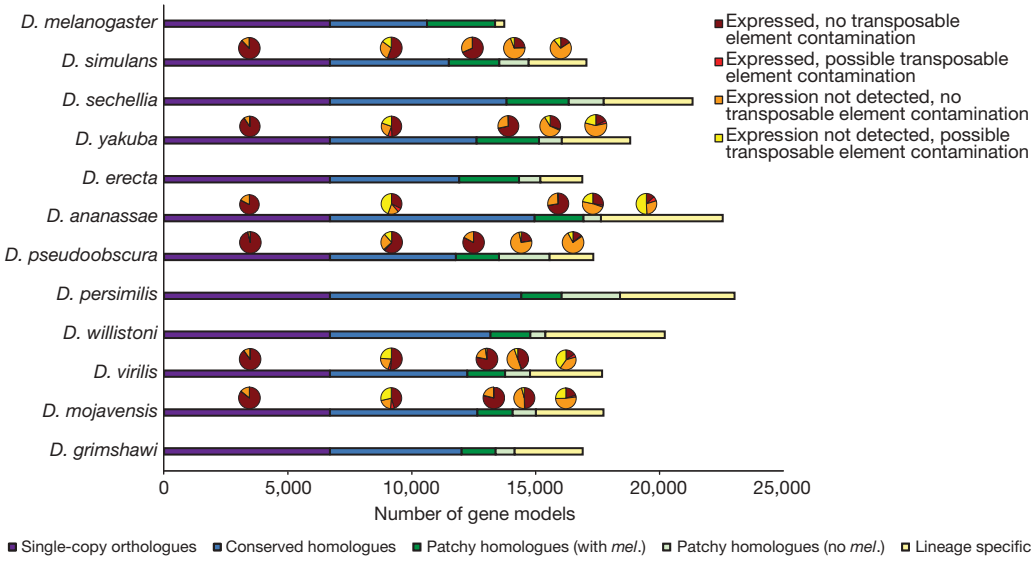

Figure 2 | Gene models in 12 *Drosophila* genomes. Number of gene models that fall into one of five homology classes: single-copy orthologues in all species (single-copy orthologues), conserved in all species as orthologues or paralogues (conserved homologues), a *D. melanogaster* homologue, but not found in all species (patchy homologues with *mel.*), conserved in at least two species but without a *D. melanogaster* homologue (patchy homologues, no *mel.*), and found only in a single lineage (lineage specific). For those species with expression data[29], pie charts indicate the fraction of genes in each homology class that fall into one of four evidence classes (see text for details).

extension of the reciprocal BLAST method[31] applicable to multiple species simultaneously (Supplementary Information section 5.1). Because the FRB algorithm does not integrate syntenic information, we also used a second approach based on Synpipe (Supplementary Information section 5.2), a tool for synteny-aided orthology assignment[32]. To generate a reconciled set of homology calls, pairwise Synpipe calls (between each species and *D. melanogaster*) were mapped to GLEAN-R models, filtered to retain only 1:1 relationships, and added to the FRB calls when they did not conflict and were non-redundant. This reconciled FRB + Synpipe set of homology calls forms the basis of our subsequent analyses. There were 8,563 genes with single-copy orthologues in the *melanogaster* group and 6,698 genes with single-copy orthologues in all 12 species; similar numbers of genes were also obtained with an independent approach[33]. Most single-copy orthologues are expressed and are free from potential transposable element contamination, suggesting that the reconciled orthologue set contains robust and high-quality gene models (Fig. 2).

Validation of homology calls. Because both the FRB algorithm and Synpipe rely on BLAST-based methods to infer similarities, rapidly evolving genes may be overlooked. Moreover, assembly gaps and poor-quality sequence may lead to erroneous inferences of gene loss. To validate putative gene absences, we used a synteny-based GeneWise pipeline to find potentially missed homologues of *D. melanogaster* proteins (Supplementary Information section 5.4). Of the 21,928 cases in which a *D. melanogaster* gene was absent from another species in the initial homology call set, we identified plausible homologues for 13,265 (60.5%), confirmed 4,546 (20.7%) as genuine absences, and were unable to resolve 4,117 (18.8%). Because this approach is conservative and only confirms strongly supported absences, we are probably underestimating the number of genuine absences.

Coding gene alignment and filtering. Investigating the molecular evolution of orthologous and paralogous genes requires accurate multi-species alignments. Initial amino acid alignments were generated using TCOFFEE[34] and converted to nucleotide alignments (Supplementary Table 8). To reduce biases in downstream analyses, a simple computational screen was developed to identify and mask problematic regions of each alignment (Supplementary Information section 6). Overall, 2.8% of bases were masked in the *melanogaster* group alignments, and 3.0% of bases were masked in the full 12 species alignments, representing 8.5% and 13.8% of alignment columns, respectively. The vast majority of masked bases are masked in no more than one species (Supplementary Fig. 3), suggesting that the masking procedure is not simply eliminating rapidly evolving regions of the genome. We find an appreciably higher frequency of masked bases in lower-quality *D. simulans* and *D. sechellia* assemblies, compared to the more divergent (from *D. melanogaster*) but higher-quality *D. erecta* and *D. yakuba* assemblies, suggesting a higher error rate in accurately predicting and aligning gene models in lower-quality assemblies (Supplementary Information section 6 and Supplementary Fig. 3). We used masked versions of the alignments, including only the longest *D. melanogaster* transcripts for all subsequent analysis unless otherwise noted.

Annotation of non-coding (nc)RNA genes. Using *de novo* and homology-based approaches we annotated over 9,000 ncRNA genes from recognized ncRNA classes (Table 2, and Supplementary Information section 7). In contrast to the large number of predictions observed for many ncRNA families in vertebrates (due in part to large numbers of ncRNA pseudogenes[35,36]), the number of ncRNA genes per family predicted by RFAM and tRNAscan in *Drosophila* is relatively low (Table 2). This suggests that ncRNA pseudogenes are largely absent from *Drosophila* genomes, which is consistent with the low number of protein-coding pseudogenes in *Drosophila*[37]. The relatively low numbers of some classes of ncRNA genes (for example, small nucleolar (sno)RNAs) in the *Drosophila* subgenus are likely to be an artefact of rapid rates of evolution in these types of genes and the limitation of the homology-based methods used to annotate distantly related species.

Evolution of genome structure

Coarse-level similarities among Drosophilids. At a coarse level, genome structure is well conserved across the 12 sequenced species. Total genome size estimated by flow cytometry varies less than three-fold across the phylogeny, ranging from 130 Mb (*D. mojavensis)* to 364 Mb (*D. virilis*)[38] (Table 2), in contrast to the order of magnitude difference between *Drosophila* and mammals. Total protein-coding sequence ranges from 38.9 Mb in *D. melanogaster* to 65.4 Mb in *D. willistoni*. Intronic DNA content is also largely conserved, ranging from 19.6 Mb in *D. simulans* to 24.0 Mb in *D. pseudoobscura* (Table 2). This contrasts dramatically with transposable element-derived genomic DNA content, which varies considerably across genomes (Table 2) and correlates significantly with euchromatic genome size (estimated as the summed length of contigs > 200 kb) (Kendall's $\tau = 0.70$, $P = 0.0016$).

To investigate overall conservation of genome architecture at an intermediate scale, we analysed synteny relationships across species using Synpipe[32] (Supplementary Information section 9.1). Synteny block size and average number of genes per block varies across the phylogeny as expected, with the number of blocks increasing and the average size of blocks decreasing with increasing evolutionary distance from *D. melanogaster* (A. Bhutkar, S. Russo, T. F. Smith and W. M. Gelbart, personal communication) (Supplementary Fig. 4). We inferred 112 syntenic blocks between *D. melanogaster* and *D. sechellia* (with an average of 122 genes per block), compared to 1,406 syntenic blocks between *D. melanogaster* and *D. grimshawi* (with an average of 8 genes per block). On average, 66% of each genome assembly was covered by syntenic blocks, ranging from 68% in *D. sechellia* to 58% in *D. grimshawi*.

Similarity across genomes is largely recapitulated at the level of individual genes, with roughly comparable numbers of predicted protein-coding genes across the 12 species (Table 2). The majority of predicted genes in each species have homologues in *D. melanogaster* (Table 2, Supplementary Table 9). Moreover, most of the 13,733 protein-coding genes in *D. melanogaster* are conserved across the entire phylogeny: 77% have identifiable homologues in all 12 genomes, 62% can be identified as single-copy orthologues in the six genomes of the *melanogaster* group and 49% can be identified as single-copy orthologues in all 12 genomes. The number of functional non-coding RNA genes predicted in each *Drosophila* genome is also largely conserved, ranging from 584 in *D. mojavensis* to 908 in *D. ananassae* (Table 2).

There are several possible explanations for the observed interspecific variation in gene content. First, approximately 700 *D. melanogaster* gene models have been newly annotated since the FlyBase Release 4.3 annotations used in the current study, reducing the discrepancy between *D. melanogaster* and the other sequenced genomes in this study. Second, because low-coverage genomes tend to have more predicted gene models, we suspect that artefactual duplication of genomic segments due to assembly errors inflates the number of predicted genes in some species. Finally, the non-*melanogaster* species have many more predicted lineage-specific genes than *D. melanogaster*, and it is possible that some of these are artefactual. In the absence of experimental evidence, it is difficult to distinguish genuine lineage-specific genes from putative artefacts. Future experimental work will be required to fully disentangle the causes of interspecific variation in gene number.

Abundant genome rearrangements during *Drosophila* evolution. To study the structural relationships among genomes on a finer scale, we analysed gene-level synteny between species pairs. These synteny maps allowed us to infer the history and locations of fixed genomic rearrangements between species. Although *Drosophila* species vary in their number of chromosomes, there are six fundamental chromosome arms common to all species. For ease of denoting

chromosomal homology, these six arms are referred to as 'Muller elements' after Hermann J. Muller, and are denoted A–F. Although most pairs of orthologous genes are found on the same Muller element, there is extensive gene shuffling within Muller elements between even moderately diverged genomes (Fig. 3, and Supplementary Information section 9.1).

Previous analysis has revealed heterogeneity in rearrangement rates among close relatives: careful inspection of 29 inversions that differentiate the chromosomes of *D. melanogaster* and *D. yakuba* revealed that 28 were fixed in the lineage leading to *D. yakuba*, and only one was fixed on the lineage leading to *D. melanogaster*[39]. Rearrangement rates are also heterogeneous across the genome among the 12 species: simulations reject a random-breakage model, which assumes that all sites are free to break in inversion events, but fail to reject a model of coldspots and hotspots for breakpoints (S. Schaeffer, personal communication). Furthermore, inversions seem to have played important roles in the process of speciation in at least some of these taxa[40].

One particularly striking example of the dynamic nature of genome micro-structure in *Drosophila* is the homeotic *homeobox (Hox)* gene cluster(s)[41]. *Hox* genes typically occur in genomic clusters, and this clustering is conserved across many vertebrate and invertebrate taxa, suggesting a functional role for the precise and collinear arrangement of these genes. However, several cluster splits have been previously identified in *Drosophila*[42,43], and the 12 *Drosophila* genome sequences provide additional evidence against the functional importance of *Hox* gene clustering in *Drosophila*. There are seven different gene arrangements found across 13 *Drosophila* species (the 12 sequenced genomes and *D. buzzatii*), with no species retaining the inferred ancestral gene order[44]. It thus seems that, in *Drosophila, Hox* genes do not require clustering to maintain proper function, and are a powerful illustration of the dynamism of genome structure across the sequenced genomes.

Transposable element evolution. Mobile, repetitive transposable element sequences are a particularly dynamic component of eukaryotic genomes. Transposable element/repeat content (in scaffolds >200 kb) varies by over an order of magnitude across the genus, ranging from ~2.7% in *D. simulans* and *D. grimshawi* to ~25% in *D. ananassae* (Table 2, and Supplementary Fig. 1). These data support the lower euchromatic transposable element content in *D. simulans* relative to *D. melanogaster*[45], and reveal that euchromatic transposable element/repeat content is generally similar within the *melanogaster* subgroup. Within the *Drosophila* subgenus, *D. grimshawi* has the lowest transposable element/repeat content, possibly relating to its ecological status as an island endemic, which may minimize the chance for horizontal transfer of transposable element families. Finally, the highest levels of transposable element/repeat content are found in *D. ananassae* and *D. willistoni*. These species also have the highest numbers of pseudo-transfer (t)RNA genes (Table 2), indicating a potential relationship between pseudo-tRNA genesis and repetitive DNA, as has been established in the mouse genome[36].

Different classes of transposable elements can vary in abundance owing to a variety of host factors, motivating an analysis of the intragenomic ecology of transposable elements in the 12 genomes. In *D. melanogaster*, long terminal repeat (LTR) retrotransposons have the highest abundance, followed by LINE (long interspersed nuclear element)-like retrotransposons and terminal inverted repeat (TIR) DNA-based transposons[18]. An unbiased, conservative approach (Supplementary Information section 3) for estimating the rank order abundance of major transposable element classes suggests that these abundance trends are conserved across the entire genus (Supplementary Fig. 5). Two exceptions are an increased abundance of TIR elements in *D. erecta* and a decreased abundance of LTR elements in *D. pseudoobscura*; the latter observation may represent an assembly artefact because the sister species *D. persimilis* shows typical LTR abundance. Given that individual instances of transposable element repeats and transposable element families themselves are not conserved across the genus, the stability of abundance trends for different classes of transposable elements is striking and suggests common mechanisms for host–transposable element co-evolution in *Drosophila*.

Although comprehensive analysis of the structural and evolutionary relationships among families of transposable elements in the 12 genomes remains a major challenge for *Drosophila* genomics, some initial insights can be gleaned from analysis of particularly well-characterized transposable element families. Previous analysis has shown variable dynamics for the most abundant transposable element family *(DINE-1)*[46] in the *D. melanogaster* genome[18,47]: although inactive in *D. melanogaster*[48], *DINE-1* has experienced a recent transpositional burst in *D. yakuba*[49]. Our analysis confirms that this element is highly abundant in all of the other sequenced genomes of *Drosophila*, but is not found outside of Diptera[50,51]. Moreover, the inferred phylogenetic relationship of *DINE-1* paralogues from several *Drosophila* species suggests vertical transmission as the major mechanism for *DINE-1* propagation. Likewise, analysis of the *Galileo*

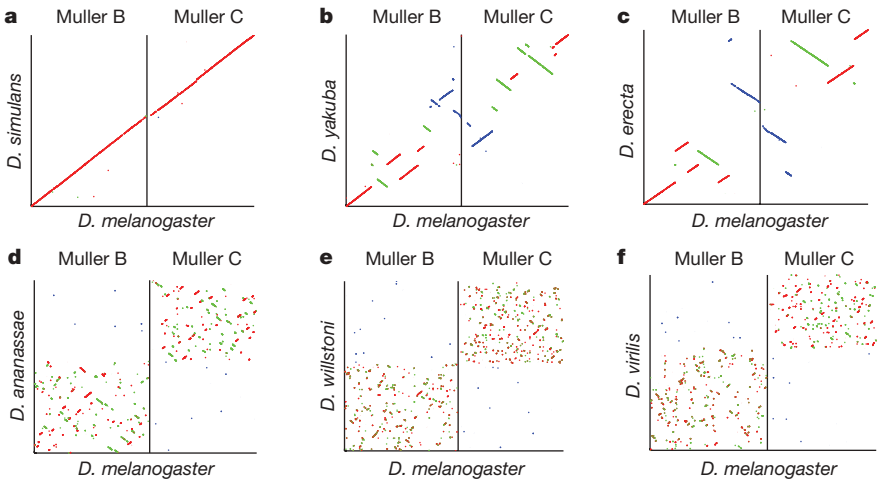

Figure 3 | Synteny plots for Muller elements B and C with respect to *D. melanogaster* gene order. The horizontal axis shows *D. melanogaster* gene order for Muller elements B and C, and the vertical axis maps homologous locations[32,155] in individual species (**a**–**f** in increasing evolutionary distance from *D. melanogaster)*. Left to right on the *x* axis is from telomere to centromere for Muller element B, followed by Muller element C from centromere to telomere. Red and green lines represent syntenic segments in the same or reverse orientation along the chromosome relative to *D. melanogaster*, respectively. Blue segments show gene transposition of genes from one element to the other.

and *1360* transposons reveals a widespread but discontinuous phylogenetic distribution for both families, notably with both families absent in the geographically isolated Hawaiian species, *D. grimshawi*[52]. These results are consistent with an ancient origin of the *Galileo* and *1360* families in the genus and subsequent horizontal transfer and/or loss in some lineages.

The use of these 12 genomes also facilitated the discovery of transposable element lineages not yet documented in *Drosophila*, specifically the P instability factor (*PIF*) superfamily of DNA transposons. Our analysis indicates that there are four distinct lineages of this transposon in *Drosophila*, and that this element has indeed colonized many of the sequenced genomes[53]. This superfamily is particularly intriguing given that *PIF*-transposase-like genes have been implicated in the origin of at least seven different genes during the *Drosophila* radiation[53], suggesting that not only do transposable elements affect the evolution of genome structure, but that their domestication can play a part in the emergence of novel genes.

D. melanogaster maintains its telomeres by occasional targeted transposition of three telomere-specific non-LTR retrotransposons (*HeT-A*, *TART* and *TAHRE*) to chromosome ends[54,55] and not by the more common mechanism of telomerase-generated G-rich repeats[56]. Multiple telomeric retrotransposons have originated within the genus, where they now maintain telomeres, and recurrent loss of most of the ORF2 from telomeric retrotransposons (for example, *TAHRE*) has given rise to half-telomeric-retrotransposons (for example, *HeT-A*) during *Drosophila* evolution[57]. The phylogenetic relationship among these telomeric elements is congruent with the species phylogeny, suggesting that they have been vertically transmitted from a common ancestor[57].

ncRNA gene family evolution. Using ncRNA gene annotations across the 12-species phylogeny, we inferred patterns of gene copy number evolution in several ncRNA families. Transfer RNA genes are the most abundant family of ncRNA genes in all 12 genomes, with 297 tRNAs in *D. melanogaster* and 261–484 tRNA genes in the other species (Table 2). Each genome encodes a single selenocysteine tRNA, with the exception of *D. willistoni*, which seems to lack this gene (R. Guigo, personal communication). Elevated tRNA gene counts in *D. ananassae* and *D. willistoni* are explained almost entirely by pseudo-tRNA gene predictions. We infer from the lack of pseudo-tRNAs in most *Drosophila* species, and from similar numbers of tRNAs obtained from an analysis of the chicken genome ($n = 280$)[58], that the minimal metazoan tRNA set is encoded by ~300 genes, in contrast to previous estimates of 497 in human and 659 in *Caenorhabditis elegans*[59,60]. Similar numbers of snoRNAs are predicted in the *D. melanogaster* subgroup ($n = 242$–255), in which sequence similarity is high enough for annotation by homology, with fewer snoRNAs ($n = 194$–216) annotated in more distant members of the *Sophophora* subgenus, and even fewer snoRNAs ($n = 139$–165) predicted in the *Drosophila* subgenus, in which annotation by homology becomes much more difficult.

Of 78 previously reported micro (mi)RNA genes, 71 (91%) are highly conserved across the entire genus, with the remaining seven genes (*mir-2b-1*, *-289*, *-303*, *-310*, *-311*, *-312* and *-313*) restricted to the subgenus *Sophophora* (Supplementary Information section 7.2). All the species contain similar numbers of spliceosomal snRNA genes (Table 2), including at least one copy each of the four U12-dependent (minor) spliceosomal RNAs, despite evidence for birth and death of these genes and the absence of stable subtypes[61]. The unusual, lineage-specific expansion in size of U11 snRNA, previously described in *Drosophila*[61,62], is even more extreme in *D. willistoni*. We annotated 99 copies of the 5S ribosomal (r)RNA gene in a cluster in *D. melanogaster*, and between 13 and 73 partial 5S rRNA genes in clusters in the other genomes. Finally, we identified members of several other classes of ncRNA genes, including the RNA components of the RNase P (1 per genome) and the signal recognition particle (SRP) RNA complexes (1–3 per genome), suggesting that these functional RNAs are involved in similar biological processes throughout the genus. We were only able to locate the *roX* (RNA on X)[63,64] genes involved in dosage compensation using nucleotide homology in the *melanogaster* subgroup, although analyses incorporating structural information have identified *roX* genes in other members of the genus[65].

We investigated the evolution of rRNA genes in the 12 sequenced genomes, using trace archives to locate sequence variants within the transcribed portions of these genes. This analysis revealed moderate levels of variation that are not distributed evenly across the rRNA genes, with fewest variants in conserved core coding regions, more variants in coding expansion regions, and higher still variant abundances in non-coding regions. The level and distribution of sequence variation in rRNA genes are suggestive of concerted evolution, in which recombination events uniformly distribute variants throughout the rDNA loci, and selection dictates the frequency to which variants can expand[66].

Protein-coding gene family evolution. For a general perspective on how the protein-coding composition of these 12 genomes has changed, we examined gene family expansions and contractions in the 11,434 gene families (including those of size one in each species) predicted to be present in the most recent common ancestor of the two subgenera. We applied a maximum likelihood model of gene gain and loss[67] to estimate rates of gene turnover. This analysis suggests that gene families expand or contract at a rate of 0.0012 gains and losses per gene per million years, or roughly one fixed gene gain/loss across the genome every 60,000 yr[68]. Many gene families (4,692 or 41.0%) changed in size in at least one species, and 342 families showed significantly elevated ($P < 0.0001$) rates of gene gain and loss compared to the genomic average, indicating that non-neutral processes may play a part in gene family evolution. Twenty-two families exhibit rapid copy number evolution along the branch leading to *D. melanogaster* (eighteen contractions and four expansions; Supplementary Table 10). The most common Gene Ontology (GO) terms among families with elevated rates of gain/loss include 'defence response', 'protein binding', 'zinc ion binding', 'proteolysis', and 'trypsin activity'. Interestingly, genes involved in 'defence response' and 'proteolysis' also show high rates of protein evolution (see below). We also found heterogeneity in overall rates of gene gain and loss across lineages, although much of this variation could result from interspecific differences in assembly quality[68].

Lineage-specific genes. The vast majority of *D. melanogaster* proteins that can be unambiguously assigned a homology pattern (Supplementary Information section 5) are inferred to be ancestrally present at the genus root (11,348/11,644, or 97.5%). Of the 296 non-ancestrally present genes, 252 are either *Sophophora*-specific, or have a complicated pattern of homology requiring more than one gain and/or loss on the phylogeny, and are not discussed further. The remaining 44 proteins include 14 present in the *melanogaster* group, 23 present only in the *melanogaster* subgroup, 3 unique to the *melanogaster* species complex, and 4 found in *D. melanogaster* only. Because we restricted this analysis to unambiguous homologues of high-confidence protein-coding genes in *D. melanogaster*[8], we are probably undercounting the number of genes that have arisen *de novo* in any particular lineage. However, ancestrally heterochromatic genes that are currently euchromatic in *D. melanogaster* may spuriously seem to be lineage-specific.

The 44 lineage-specific genes (Supplementary Table 11) differ from ancestrally present genes in several ways. They have a shorter median predicted protein length (lineage-specific median 177 amino acids, other median 421 amino acids, MWU, $P = 3.6 \times 10^{-13}$), are more likely to be intronless (Fisher's exact test (FET), $P = 6.2 \times 10^{-6}$), and are more likely to be located in the intron of another gene on the opposite strand (FET, $P = 3.5 \times 10^{-4}$). In addition, 18 of these 44 genes are testis- or accessory-gland-specific in *D. melanogaster*, a significantly greater fraction than is found in the ancestral set (FET, $P = 1.25 \times 10^{-4}$). This is consistent with previous observations that novel genes are often testis-specific in *Drosophila*[69–73] and

expression studies on seven of the species show that species-restricted genes are more likely to exhibit male-biased expression[29]. Further, these genes are significantly more tissue-specific in expression (as measured by τ; ref. 74) (MWU, $P = 9.6 \times 10^{-6}$), and this pattern is not solely driven by genes with testis-specific expression patterns.

Protein-coding gene evolution

Positive selection and selective constraints in *Drosophila* genomes. To study the molecular evolution of protein-coding genes, we estimated rates of synonymous and non-synonymous substitution in 8,510 single-copy orthologues within the six *melanogaster* group species using PAML[75] (Supplementary Information section 11.1); synonymous site saturation prevents analysis of more divergent comparisons. We investigate only single-copy orthologues because when paralogues are included, alignments become increasingly problematic. Rates of amino acid divergence for single-copy orthologues in all 12 species were also calculated; these results are largely consistent with the analysis of non-synonymous divergence in the *melanogaster* group, and are not discussed further.

To understand global patterns of divergence and constraint across functional classes of genes, we examined the distributions of ω ($=d_N/d_S$, the ratio of non-synonymous to synonymous divergence) across Gene Ontology categories (GO)[76], excluding GO annotations based solely on electronic support (Supplementary Information section 11.2). Most functional categories of genes are strongly constrained, with median estimates of ω much less than one. In general, functionally similar genes are similarly constrained: 31.8% of GO categories have significantly lower variance in ω than expected (q-value true-positive test[77]). Only 11% of GO categories had statistically significantly elevated ω (relative to the median of all genes with GO annotations) at a 5% false-discovery rate (FDR), suggesting either positive selection or a reduction in selective constraint. The GO categories with elevated ω include the biological process terms 'defence response', 'proteolysis', 'DNA metabolic process' and 'response to biotic stimulus'; the molecular function terms 'transcription factor activity', 'peptidase activity', 'receptor binding', 'odorant binding', 'DNA binding', 'receptor activity' and 'G-protein-coupled receptor activity'; and the cellular location term 'extracellular' (Fig. 4, and Supplementary Table 12). Similar results are obtained when d_N is compared across GO categories, suggesting that in most cases differences in ω among GO categories is driven by amino acid rather than synonymous site substitutions. The two exceptions are the molecular function terms 'transcription factor activity' and 'DNA binding activity', for which we observe significantly decelerated d_S (FDR $= 7.2 \times 10^{-4}$ for both; Supplementary Information section 11.2) and no significant differences in d_N.

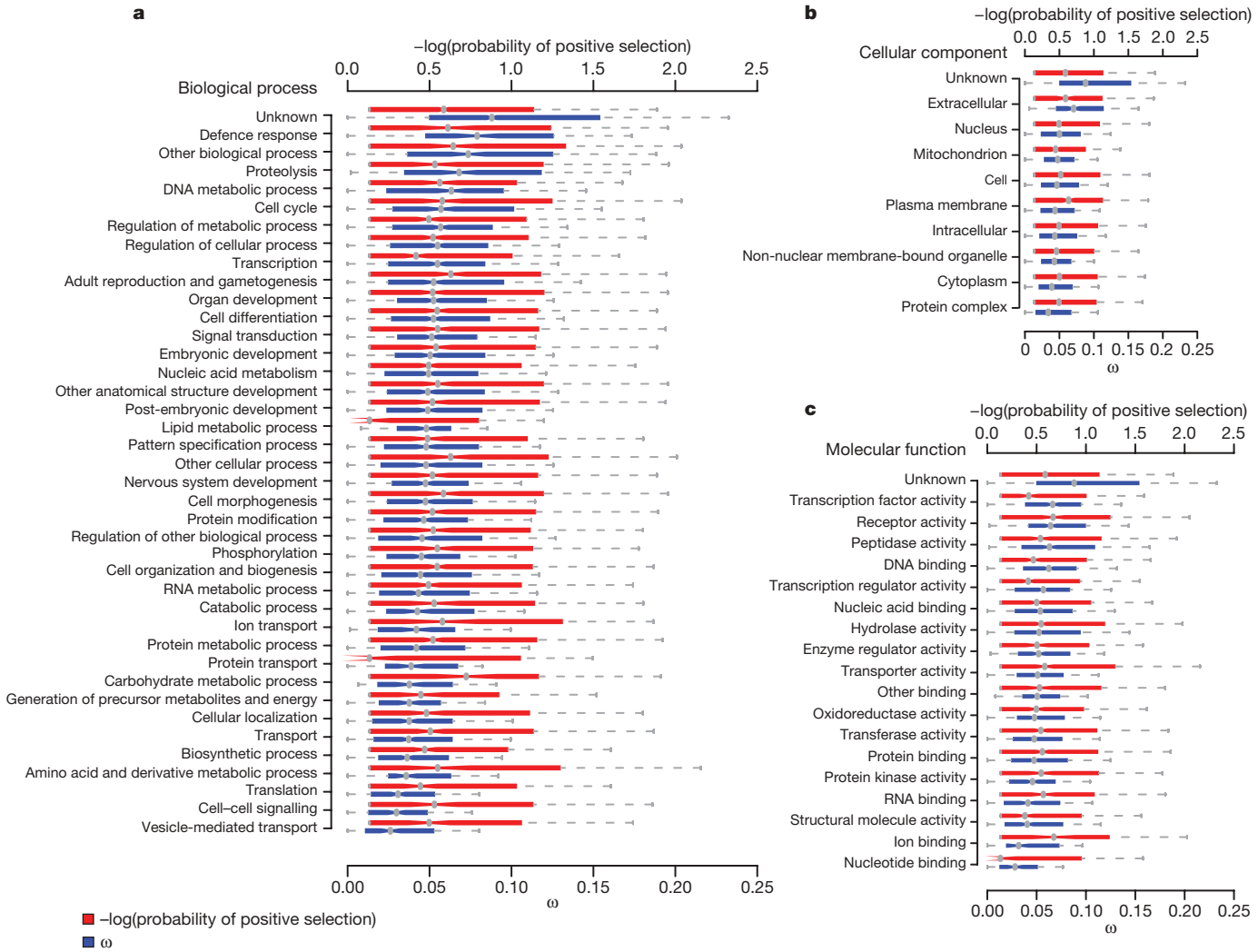

Figure 4 | Patterns of constraint and positive selection among GO terms. Distribution of average ω per gene and the negative $\log_{10}$ of the probability of positive selection (Supplementary Information section 11.2) for genes annotated with: **a**, biological process GO terms; **b**, cellular component GO terms; and **c**, molecular function GO terms. Only GO terms with 200 or more genes annotated are plotted. See Supplementary Table 12 for median values and significance. Note that most genes evolve under evolutionary constraint at most of their sites, leading to low values of ω; even genes that experience positive selection do not typically have an average ω across all codons that exceeds one.

To distinguish possible positive selection from relaxed constraint, we tested explicitly for genes that have a subset of codons with signatures of positive selection, using codon-based likelihood models of molecular evolution, implemented in PAML[78,79] (Supplementary Information section 11.1). Although this test is typically regarded as a conservative test for positive selection, it may be confounded by selection at synonymous sites. However, selection at synonymous sites (that is, codon bias, see below) is quite weak. Moreover, variability in ω presented here tends to reflect variability in d_N. We therefore believe that it is appropriate to treat synonymous sites as nearly neutral and sites with $\omega > 1$ as consistent with positive selection. Despite a number of functional categories with evidence for elevated ω, 'helicase activity' is the only functional category significantly more likely to be positively selected (permutation test, $P = 2 \times 10^{-4}$, FDR = 0.007; Supplementary Table 12); the biological significance of this finding merits further investigation. Furthermore, within each GO class, there is greater dispersion among genes in their probability of positive selection than in their estimate of ω (MWU one-tailed, $P = 0.011$; Supplementary Information section 11.1), suggesting that although functionally similar genes share patterns of constraint, they do not necessarily show similar patterns of positive selection (Fig. 4).

Interestingly, protein-coding genes with no annotated ('unknown') function in the GO database seem to be less constrained (permutation test, $P < 1 \times 10^{-4}$, FDR = 0.006)[80] and to have on average lower *P*-values for the test of positive selection than genes with annotated functions (permutation test, $P = 0.001$, FDR = 0.058). It is unlikely that this observation results entirely from an over-representation of mis-annotated or non-protein-coding genes in the 'unknown' functional class, because this finding is robust to the removal of all *D. melanogaster* genes predicted to be non-protein-coding in ref. 8. The bias in the way biological function is ascribed to genes (to laboratory-induced, easily scorable functions) leaves open the possibility that unannotated biological functions may have an important role in evolution. Indeed, genes with characterized mutant alleles in FlyBase evolve significantly more slowly than other genes (median $\omega_{\text{with alleles}} = 0.0525$ and $\omega_{\text{without alleles}} = 0.0701$; MWU, $P < 1 \times 10^{-16}$).

Previous work has suggested that a substantial fraction of non-synonymous substitutions in *Drosophila* were fixed through positive selection[81–85]. We estimate that 33.1% of single-copy orthologues in the *melanogaster* group have experienced positive selection on at least a subset of codons (*q*-value true-positive tests[77]) (Supplementary Information section 11.1). This may be an underestimate, because we have only examined single-copy orthologues, owing to difficulties in producing accurate alignments of paralogues by automated methods. On the basis of the 878 genes inferred to have experienced positive selection with high confidence (FDR < 10%), we estimated that an average of 2% of codons in positively selected genes have $\omega > 1$. Thus, several lines of evidence, based on different methodologies, suggest that patterns of amino acid fixation in *Drosophila* genomes have been shaped extensively by positive selection.

The presence of functional domains within a protein may lead to heterogeneity in patterns of constraint and adaptation along its length. Among genes inferred to be evolving by positive selection at a 10% FDR, 63.7% (*q*-value true-positive tests[77]) show evidence for spatial clustering of positively selected codons (Supplementary Information section 11.2). Spatial heterogeneity in constraint is further supported by contrasting ω for codons inside versus outside defined InterPro domains (genes lacking InterPro domains are treated as 'outside' a defined InterPro domain). Codons within InterPro domains were significantly more conserved than codons outside InterPro domains (median ω: 0.062 InterPro domains, 0.084 outside InterPro domains; MWU, $P < 2.2 \times 10^{-16}$; Supplementary Information section 11.2). Similarly, there were significantly more positively selected codons outside of InterPro domains than inside domains (FET $P < 2.2 \times 10^{-16}$), suggesting that in addition to being more constrained, codons in protein domains are less likely to be targets of positive selection (Supplementary Fig. 6).

Factors affecting the rate of protein evolution in *Drosophila*. The sequenced genomes of the *melanogaster* group provide unprecedented statistical power to identify factors affecting rates of protein evolution. Previous analyses have suggested that although the level of gene expression consistently seems to be a major determinant of variation in rates of evolution among proteins[86,87], other factors probably play a significant, if perhaps minor, part[88–91]. In *Drosophila*, although highly expressed genes do evolve more slowly, breadth of expression across tissues, gene essentiality and intron number all also independently correlate with rates of protein evolution, suggesting that the additional complexities of multicellular organisms are important factors in modulating rates of protein evolution[78]. The presence of repetitive amino acid sequences has a role as well: non-repeat regions in proteins containing repeats evolve faster and show more evidence for positive selection than genes lacking repeats[92].

These data also provide a unique opportunity to examine the impact of chromosomal location on evolutionary rates. Population genetic theory predicts that for new recessive mutations, both purifying and positive selection will be more efficient on the X chromosome given its hemizygosity in males[93]. In contrast, the lack of recombination on the small, mainly heterochromatic dot chromosome[94,95] is expected to reduce the efficacy of selection[96]. Because codon bias, or the unequal usage of synonymous codons in protein-coding sequences, reflects weak but pervasive selection, it is a sensitive metric for evaluating the efficacy of purifying selection. Consistent with expectation, in all 12 species, we find significantly elevated levels of codon bias on the X chromosome and significantly reduced levels of codon bias on the dot chromosome[97]. Furthermore, X-chromosome-linked genes are marginally over-represented within the set of positively selected genes in the *melanogaster* group (FET, $P = 0.055$), which is consistent with increased rates of adaptive substitution on this chromosome. This analysis suggests that chromosomal context also serves to modulate rates of molecular evolution in protein-coding genes.

To examine further the impact of genomic location on protein evolution, we examined the subset of genes that have moved within or between chromosome arms[32,98]. Genes inferred to have moved between Muller elements have a significantly higher rate of protein evolution than genes inferred to have moved within a Muller element (MWU, $P = 1.32 \times 10^{-14}$) and genes that have maintained their genomic position (MWU, $P = 0.008$) (Supplementary Fig. 7). Interestingly, genes that move within Muller elements have a significantly lower rate of protein evolution than those for which genomic locations have been maintained (MWU, $P = 3.85 \times 10^{-14}$). It remains unclear whether these differences reflect underlying biases in the types of genes that move inter- versus intra-chromosomally, or whether they are due to *in situ* patterns of evolution in novel genomic contexts.

Codon bias. Codon bias is thought to enhance the efficiency and/or accuracy of translation[99–101] and seems to be maintained by mutation–selection–drift balance[101–104]. Across the 12 *Drosophila* genomes, there is more codon bias in the *Sophophora* subgenus than in the *Drosophila* subgenus, and a previously noted[105–109] striking reduction in codon bias in *D. willistoni*[110,111] (Fig. 5). However, with only minor exceptions, codon preferences for each amino acid seem to be conserved across 11 of the 12 species. The striking exception is *D. willistoni*, in which codon usage for 6 of 18 redundant amino acids has diverged (Fig. 5). Mutation alone is not sufficient to explain codon-usage bias in *D. willistoni*, which is suggestive of a lineage-specific shift in codon preferences[111,112]. We found evidence for a lineage-specific genomic reduction in codon bias in *D. melanogaster* (Fig. 5), as has been suggested previously[113–119]. In addition, maximum-likelihood estimation of the strength of selection on synonymous sites in 8,510 *melanogaster* group single-copy orthologues revealed a marked reduction in the number of genes under selection

for increased codon bias in *D. melanogaster* relative to its sister species *D. sechellia*[120].

Evolution of genes associated with ecology and reproduction. Given the ecological and environmental diversity encompassed by the 12 *Drosophila* species, we examined the evolution of genes and gene families associated with ecology and reproduction. Specifically, we selected genes with roles in chemoreception, detoxification/metabolism, immunity/defence, and sex/reproduction for more detailed study.

Chemoreception. *Drosophila* species have complex olfactory and gustatory systems used to identify food sources, hazards and mates, which depend on odorant-binding proteins, and olfactory/odorant and gustatory receptors (*Or*s and *Gr*s). The *D. melanogaster* genome has approximately 60 *Or*s, 60 *Gr*s and 50 odorant-binding protein genes. Despite overall conservation of gene number across the 12 species and widespread evidence for purifying selection within the *melanogaster* group, there is evidence that a subset of *Or* and *Gr* genes experiences positive selection[121–123]. Furthermore, clear lineage-specific differences are detectable between generalist and specialist species within the *melanogaster* subgroup. First, the two independently evolved specialists (*D. sechellia* and *D. erecta*) are losing *Gr* genes approximately five times more rapidly than the generalist species[121,124]. We believe this result is robust to sequence quality, because all pseudogenes and deletions were verified by direct re-sequencing and synteny-based orthologue searches, respectively. Generalists are expected to encounter the most diverse set of tastants and seem to have maintained the greatest diversity of gustatory receptors. Second, *Or* and *Gr* genes that remain intact in *D. sechellia* and *D. erecta* evolve significantly more rapidly along these two lineages ($\omega = 0.1556$ for *Or*s and 0.1874 for *Gr*s) than along the generalist lineages ($\omega = 0.1049$ for *Or*s and 0.1658 for *Gr*s; paired Wilcoxon, $P = 0.0003$ and 0.003, respectively[124]). There is some evidence that odorant-binding protein genes also evolve significantly faster in specialists compared to generalists[122]. This elevated ω reflects a trend observed throughout the genomes of the two specialists and is likely to result, at least in part, from demographic phenomena. However, the difference between specialist and generalist ω for *Or*/*Gr* genes (0.0292) is significantly greater than the difference for genes across the genome (0.0091; MWU, $P = 0.0052$)[121], suggesting a change in selective regime. Moreover, the observation that elevated ω as well as accelerated gene loss disproportionately affect groups of *Or* and *Gr* genes that respond to specific chemical ligands and/or are expressed during specific life stages suggests that rapid evolution at *Or*/*Gr* loci in specialists is related to the ecological shifts these species have sustained[121].

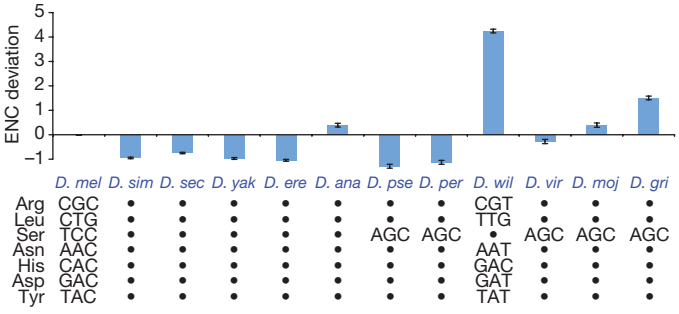

Figure 5 | Deviations in codon bias from *D. melanogaster* in 11 *Drosophila* species. The upper panel depicts differences in ENC (effective number of codons) between *D. melanogaster* and the 11 non-*melanogaster* species, calculated on a gene-by-gene basis. Note that increasing levels of ENC indicates a decrease in codon bias. The *Sophophora* subgenus in general has higher levels of codon bias than the *Drosophila* subgenus with the exception of *D. willistoni,* which shows a dramatic reduction in codon bias. The lower panel shows the 7 codons for which preference changes across the 12 *Drosophila* species. A dot indicates identical codon preference to *D. melanogaster*; otherwise the preferred codon is indicated.

Detoxification/metabolism. The larval food sources for many *Drosophila* species contain a cocktail of toxic compounds, and consequently *Drosophila* genomes encode a wide variety of detoxification proteins. These include members of the cytochrome P450 (P450), carboxyl/choline-esterase (CCE) and glutathione *S*-transferase (GST) multigene families, all of which also have critical roles in resistance to insecticides[125–127]. Among the P450s, the five enzymes associated with insecticide resistance are highly dynamic across the phylogeny, with 24 duplication events and 4 loss events since the last common ancestor of the genus, which is in striking contrast to genes with known developmental roles, eight of which are present as a single copy in all 12 species (C. Robin, personal communication). As with chemoreceptors, specialists seem to lose detoxification genes at a faster rate than generalists. For instance, *D. sechellia* has lost the most P450 genes; these 14 losses comprise almost one-third of all P450 loss events (Supplementary Table 13) (C. Robin, personal communication). Positive selection has been implicated in detoxification-gene evolution as well, because a search for positive selection among GSTs identified the parallel evolution of a radical glycine to lysine amino acid change in GSTD1, an enzyme known to degrade DDT[128]. Finally, although metabolic enzymes in general are highly constrained (median $\omega = 0.045$ for enzymes, 0.066 for non-enzymes; MWU, $P = 5.7 \times 10^{-24}$), enzymes involved in xenobiotic metabolism evolve significantly faster than other enzymes (median $\omega = 0.05$ for the xenobiotic group versus $\omega = 0.045$ overall, two-tailed permutation test, $P = 0.0110$; A. J. Greenberg, personal communication).

Metazoans deal with excess selenium in the diet by sequestration in selenoproteins, which incorporate the rare amino acid selenocysteine (Sec) at sites specified by the TGA codon. The recoding of the normally terminating signal TGA as a Sec codon is mediated by the selenocystein insertion sequence (SECIS), a secondary structure in the 3′ UTR of selenoprotein messenger RNAs. All animals examined so far have selenoproteins; three have been identified in *D. melanogaster* (SELG, SELM and SPS2[129,130]). Interestingly, although the three known *melanogaster* selenoproteins are all present in the genomes of the other *Drosophila* species, in *D. willistoni* the TGA Sec codons have been substituted by cysteine codons (TGT/TGC). Consistent with this finding, analysis of the seven genes implicated to date in selenoprotein synthesis including the Sec-specific tRNA suggests that most of these genes are absent in *D. willistoni* (R. Guigo, personal communication). *D. willistoni* thus seems to be the first animal known to lack selenoproteins. If correct, this observation is all the more remarkable given the ubiquity of selenoproteins and the selenoprotein biosynthesis machinery in metazoans, the toxicity of excess selenium, and the protection from oxidative stress mediated by selenoproteins. However, it remains possible that this species encodes selenoproteins in a different way, and this represents an exciting avenue of future research.

Immunity/defence. *Drosophila*, like all insects, possesses an innate immune system with many components analogous to the innate immune pathways of mammals, although it lacks an antibody-mediated adaptive immune system[131]. Immune system genes often evolve rapidly and adaptively, driven by selection pressures from pathogens and parasites[132–134]. The genus *Drosophila* is no exception: immune system genes evolve more rapidly than non-immune genes, showing both high total divergence rates and specific signs of positive selection[135]. In particular, 29% of receptor genes involved in phagocytosis seem to evolve under positive selection, suggesting that molecular co-evolution between *Drosophila* pattern recognition receptors and pathogen antigens is driving adaptation in the immune system[135]. Somewhat surprisingly, genes encoding effector proteins such as antimicrobial peptides are far less likely to exhibit adaptive sequence evolution. Only 5% of effector genes (and no antimicrobial peptides) show evidence of adaptive evolution, compared to 10% of genes genome-wide. Instead, effector genes seem to evolve by rapid duplication and deletion. Whereas 49% of genes genome-wide, 63%

of genes involved in pathogen recognition and 81% of genes implicated in immune-related signal transduction can be found as single-copy orthologues in all 12 species, only 40% of effector genes exist as single-copy orthologues across the genus ($\chi 2 = 41.13$, $P = 2.53 \times 10^{-8}$), suggesting rapid radiation of effector protein classes along particular lineages[135]. Thus, much of the *Drosophila* immune system seems to evolve rapidly, although the mode of evolution varies across immune-gene functional classes.

Sex/reproduction. Genes encoding sex- and reproduction-related proteins are subject to a wide array of selective forces, including sexual conflict, sperm competition and cryptic female choice, and to the extent that these selective forces are of evolutionary consequence, this should lead to rapid evolution in these genes[136] (for an overview see refs 137, 138). The analysis of 2,505 sex- and reproduction-related genes within the *melanogaster* group indicated that male sex- and reproduction-related genes evolve more rapidly at the protein level than genes not involved in sex or reproduction or than female sex- and reproduction-related genes (Supplementary Fig. 8). Positive selection seems to be at least partially responsible for these patterns, because genes involved in spermatogenesis have significantly stronger evidence for positive selection than do non-spermatogenesis genes (permutation test, $P = 0.0053$). Similarly, genes that encode components of seminal fluid have significantly stronger evidence for positive selection than 'non-sex' genes[139]. Moreover, protein-coding genes involved in male reproduction, especially seminal fluid and testis genes, are particularly likely to be lost or gained across *Drosophila* species[29,139].

Evolutionary forces in the mitochondrial genome. Functional elements in mtDNA are strongly conserved, as expected: tRNAs are relatively more conserved than the mtDNA overall (average pairwise nucleotide distance = 0.055 substitutions per site for tRNAs versus 0.125 substitutions per site overall). We observe a deficit of substitutions occurring in the stem regions of the stem-loop structure in tRNAs, consistent with strong selective pressure to maintain RNA secondary structure, and there is a strong signature of purifying selection in protein-coding genes[13]. However, despite their shared role in aerobic respiration, there is marked heterogeneity in the rates of amino acid divergence between the oxidative phosphorylation enzyme complexes across the 12 species (NADH dehydrogenase, 0.059 > ATPase, 0.042 > CytB, 0.037 > cytochrome oxidase, 0.020; mean pairwise d_N), which contrasts with the relative homogeneity in synonymous substitution rates. A model with distinct substitution rates for each enzyme complex rather than a single rate provides a significantly better fit to the data ($P < 0.0001$), suggesting complex-specific selective effects of mitochondrial mutations[13].

Non-coding sequence evolution

ncRNA sequence evolution. The availability of complete sequence from 12 *Drosophila* genomes, combined with the tractability of RNA structure predictions, offers the exciting opportunity to connect patterns of sequence evolution directly with structural and functional constraints at the molecular level. We tested models of RNA evolution focusing on specific ncRNA gene classes in addition to inferring patterns of sequence evolution using more general datasets that are based on predicted intronic RNA structures.

The exquisite simplicity of miRNAs and their shared stem-loop structure makes these ncRNAs particularly amenable to evolutionary analysis. Most miRNAs are highly conserved within the *Drosophila* genus: for the 71 previously described miRNA genes inferred to be present in the common ancestor of these 12 species, mature miRNA sequences are nearly invariant. However, we do find a small number of substitutions and a single deletion in mature miRNA sequences (Supplementary Table 14), which may have functional consequences for miRNA–target interactions and may ultimately help identify targets through sequence covariation. Pre-miRNA sequences are also highly conserved, evolving at about 10% of the rate of synonymous sites.

To link patterns of evolution with structural constraints, we inferred ancestral pre-miRNA sequences and deduced secondary structures at each ancestral node on the phylogeny (Supplementary Information section 12.1). Although conserved miRNA genes show little structural change (little change in free energy), the five *melanogaster* group-specific miRNA genes (*miR-303* and the *mir-310/311/312/313* cluster) have undergone numerous changes across the entire pre-miRNA sequence, including the ordinarily invariant mature miRNA. Patterns of polymorphism and divergence in these lineage-specific miRNA genes, including a high frequency of derived mutations, are suggestive of positive selection[140]. Although lineage-specific miRNAs may evolve under less constraint because they have fewer target transcripts in the genome, it is also possible that recent integration into regulatory networks causes accelerated rates of miRNA evolution.

We further investigated patterns of sequence evolution for the subset of 38 conserved pre-miRNAs with mature miRNA sequences at their 3′ end by calculating evolutionary rates in distinct site classes (Fig. 6, and Supplementary Information section 12.2). Outside the mature miRNA and its complementary sequence, loops had the highest rate of evolution, followed by unpaired sites, with paired sites having the lowest rate of evolution. Inside the mature miRNA, unpaired sites evolve more slowly than paired sites, whereas the opposite is true for the sequence complementary to the mature miRNA. Surprisingly, a large fraction of unpaired bulges or internal loops in the mature miRNA seem to be conserved—a pattern which may have implications for models of miRNA biogenesis and the degree of mismatch allowed in miRNA–target prediction methods. Overall these results support the qualitative model proposed in ref. 141 for the canonical progression of miRNA evolution, and show that functional constraints on the miRNA itself supersede structural constraints imposed by maintenance of the hairpin-loop.

To assess constraint on stem regions of RNA structures more generally, we compared substitution rates in stems (*S*) to those in nominally unconstrained loop regions (*L*) in a wide variety of ncRNAs (Supplementary Information section 12.3). We estimated substitution rates using a maximum likelihood framework, and compared the observed *L*/*S* ratio with the average *L*/*S* ratio estimated from published secondary structures in RFAM, which we normalized to 1.0. *L*/*S* ratios for *Drosophila* ncRNA families range from a highly constrained 2.57 for the nuclear RNase P family to 0.56 for the 5S ribosomal RNA (Supplementary Table 15).

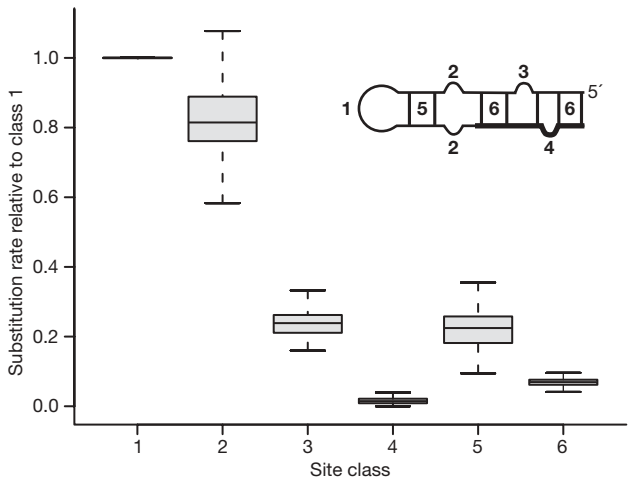

Figure 6 | Substitution rate of site classes within miRNAs. Bootstrap distributions of miRNA substitution rates. Structural alignments of miRNA precursor hairpins were partitioned into six site-classes (inset): (1) hairpin loops; unpaired sites (2) outside, (3) in the complementary region of, and (4) inside the miRNA; and base pairs (5) adjacent to and (6) involving the miRNA. Whiskers show approximate 95% confidence intervals for median differences, boxes show interquartile range.

Finally, we predicted a set of conserved intronic RNA structures and analysed patterns of compensatory nucleotide substitution in *D. melanogaster*, *D. yakuba*, *D. ananassae*, *D. pseudoobscura*, *D. virilis* and *D. mojavensis* (Supplementary Information section 13). Signatures of compensatory evolution in RNA helices are detected as covarying nucleotide sites or 'covariations' (that is, two Watson–Crick bases that interact in species A replaced by a different Watson–Crick pair in species B). The number of covariations (per base pair of a helix) depends on the physical distance between the interacting nucleotides (Supplementary Fig. 9), as has been observed for the RNA helices in the *Drosophila bicoid* 3′ UTR region[142]. Short-range pairings exhibit a higher average number of covariations with a larger variance among helices than longer-range pairings. The decrease in rate of covariation with increasing distance may be explained by physical properties of a helix, which may impose selective constraints on the evolution of covarying nucleotides within a helix. Alternatively, if individual mutations at each locus are deleterious but compensated by mutations at a second locus, given sufficiently strong selection against the first deleterious mutation these epistatic fitness interactions could generate the observed distance effect[143].

Evolution of *cis*-regulatory DNAs. Comparative analyses of *cis*-regulatory sequences may provide insights into the evolutionary forces acting on regulatory components of genes, shed light on the constraints of the *cis*-regulatory code and aid in annotation of new regulatory sequences. Here we rely on two recently compiled databases, and present results comparing *cis*-regulatory modules[144] and transcription factor binding sites (derived from DNase I footprints)[145] between *D. melanogaster* and *D. simulans* (Supplementary Information section 8). We estimated mean selective constraint (*C*, the fraction of mutations removed by natural selection) relative to the 'fastest evolving intron' sites at the 5′ end of short introns, which represent putatively unconstrained neutral standards (Supplementary Information section 8.2)[146]. Note that this approach ignores the contribution of positively selected sites, potentially underestimating the fraction of functionally relevant sites[147].

Consistent with previous findings, *Drosophila cis*-regulatory sequences are highly constrained[148,149]. Mean constraint within *cis*-regulatory modules is 0.643 (95% bootstrap confidence interval = 0.621–0.662) and within footprints is 0.692 (0.655–0.723), both of which are significantly higher than mean constraint in non-coding DNA overall (0.555 (0.546–0.563)) and significantly lower than constraint at non-degenerate coding sites (0.862 (0.856–0.868)) and ncRNA genes (0.864 (0.846–0.880)) (Supplementary Fig. 10). The high level of constraint in *cis*-regulatory sequences also extends into flanking sequences, only declining to constraint levels typical of non-coding DNA 40 bp away. This is consistent with previous findings that transcription factor binding sites tend to be found in larger blocks of constraint that cluster to form *cis*-regulatory modules[150]. To understand selective constraints on nucleotides within *cis*-regulatory sequences that have direct contact with transcription factors, we estimated the selective constraint for the best match to position weight matrices within each footprint[151]; core motifs in transcription-factor-binding sites have a mean constraint of 0.773 (0.729–0.814), significantly greater than the mean for the footprints as a whole, and approaching the level of constraint found at non-degenerate coding sites and in ncRNA genes (Supplementary Fig. 10).

We next examined the variation in selective constraint across *cis*-regulatory sequences. Surprisingly, we find no evidence that selective constraint is correlated with predicted transcription-factor-binding strength (estimated as the position weight matrix score *P*-value) (Spearman's $r = 0.0681$, $P = 0.0609$). We observe significant variation in constraint both among target genes (Kruskal–Wallis tests, footprints, $P < 0.0001$; and position weight matrix matches within footprints, $P = 0.0023$) and among chromosomes (*cis*-regulatory modules, $P = 0.0186$; footprints, $P = 0.0388$; and position weight matrix matches within footprints, $P = 0.0108$; Supplementary Table 16).

Discussion and conclusion

Each new genome sequence affords novel opportunities for comparative genomic inference. What makes the analysis of these 12 *Drosophila* genomes special is the ability to place every one of these genomic comparisons on a phylogeny with a taxon separation that is ideal for asking a wealth of questions about evolutionary patterns and processes. It is without question that this phylogenomic approach places additional burdens on bioinformatics efforts, multiplying the amount of data many-fold, requiring extra care in generating multi-species alignments, and accommodating the reality that not all genome sequences have the same degree of sequencing or assembly accuracy. These difficulties notwithstanding, phylogenomics has extraordinary advantages not only for the analyses that are possible, but also for the ability to produce high-quality assemblies and accurate annotations of functional features in a genome by using closely related genomes as guides. The use of multi-species orthology provides especially convincing evidence in support of particular gene models, not only for protein-coding genes, but also for miRNA and other ncRNA genes.

Many attributes of the genomes of *Drosophila* are remarkably conserved across species. Overall genome size, number of genes, distribution of transposable element classes, and patterns of codon usage are all very similar across these 12 genomes, although *D. willistoni* is an exceptional outlier by several criteria, including its unusually skewed codon usage, increased transposable element content and potential lack of selenoproteins. At a finer scale, the number of structural changes and rearrangements is much larger; for example, there are several different rearrangements of genes in the *Hox* cluster found in these *Drosophila* species.

The vast majority of multigene families are found in all 12 genomes, although gene family size seems to be highly dynamic: almost half of all gene families change in size on at least one lineage, and a noticeable fraction shows rapid and lineage-specific expansions and contractions. Particularly notable are cases consistent with adaptive hypotheses, such as the loss of *Gr* genes in ecological specialists and the lineage-specific expansions of antimicrobial peptides and other immune effectors. All species were found to have novel genes not seen in other species. Although lineage-specific genes are challenging to verify computationally, we can confirm at least 44 protein-coding genes unique to the *melanogaster* group, and these proteins have very different properties from ancestral proteins. Similarly, although the relative abundance of transposable element subclasses across these genomes does not differ dramatically, total genomic transposable element content varies substantially among species, and several instances of lineage-specific transposable elements were discovered.

There is considerable variation among protein-coding genes in rates of evolution and patterns of positive selection. Functionally similar proteins tend to evolve at similar rates, although variation in genomic features such as gene expression level, as well as chromosomal location, are also associated with variation in evolutionary rate among proteins. Whereas broad functional classes do not seem to share patterns of positive selection, and although very few GO categories show excesses of positive selection, a number of genes involved in interactions with the environment and in sex and reproduction do show signatures of adaptive evolution. It thus seems likely that adaptation to changing environments, as well as sexual selection, shape the evolution of protein-coding genes.

Annotation of ncRNA genes across all 12 species allows comprehensive analysis of the evolutionary divergence of these genes. MicroRNA genes in particular are more conserved than protein-coding genes with respect to their primary DNA sequence, and the substitutions that do occur often have compensatory changes such that the average estimated free energy of the folding structures remains remarkably constant across the phylogeny. Surprisingly,

mismatches in miRNAs seem to be highly conserved, which may impact models of miRNA biogenesis and target recognition. Lineage-restricted miRNAs, however, have considerably elevated rates of change, suggesting either reduced constraint due to novel miRNAs having fewer targets, or adaptive evolution of evolutionarily young miRNAs.

Virtually any question about the function of genome features in *Drosophila* is now empowered by being embedded in the context of this 12 species phylogeny, allowing an analysis of the ways by which evolution has tuned myriad biological processes across the hundreds of millions of years spanned in total by this phylogeny. The analyses presented herein have generated more questions than they have answered, and these results represent a small fraction of that which is possible. Because much of this rich and extraordinary comparative genomic dataset remains to be explored, we believe that these 12 *Drosophila* genome sequences will serve as a powerful tool for gleaning further insight into genetic, developmental, regulatory and evolutionary processes.

METHODS

The full methods for this paper are described in Supplementary Information. Here, we describe the datasets generated by this project and their availability.

Genomic sequence. Scaffolds and assemblies for all genomic sequence generated by this project are available from GenBank (Supplementary Tables 4 and 5), and FlyBase (ftp://ftp.flybase.net/12_species_analysis/). Genome browsers are available from UCSC (http://genome.ucsc.edu/cgi-bin/hgGateway?hgsid = 98180333&clade = insect&org = 0&db = 0) and Flybase (http://flybase.org/cgi-bin/gbrowse/dmel/). BLAST search of these genomes is available at FlyBase (http://flybase.org/blast).

Predicted gene models. Consensus gene predictions for the 11 non-*melanogaster* species, produced by combining several different GLEAN runs that weight homology evidence more or less strongly, are available from FlyBase as GFF files for each species (ftp://ftp.flybase.net/12_species_analysis/). These gene models can also be accessed from the Genome Browser in FlyBase (Gbrowse; http://flybase.org/cgi-bin/gbrowse/dmel/). Predictions of non-protein-coding genes are also available in GFF format for each species, from FlyBase (ftp://ftp.flybase.net/12_species_analysis/).

Homology. Multiway homology assignments are available from FlyBase (ftp://ftp.flybase.net/12_species_analysis/), and also in the Genome Browser (Gbrowse).

Alignments. All alignment sets produced are available in FASTA format from FlyBase (ftp://ftp.flybase.net/12_species_analysis/).

PAML parameters. Output from PAML models for the alignments of single copy orthologues in the *melanogaster* group, including the *q*-value for the test for positive selection, are available from FlyBase (ftp://ftp.flybase.net/12_species_analysis/).

Received 19 July; accepted 5 October 2007.

1. Markow, T. A. & O'Grady, P. M. Drosophila biology in the genomic age. *Genetics* doi:10.1534/genetics.107.074112 (in the press).
2. Powell, J. R. *Progress and Prospects in Evolutionary Biology: The* Drosophila *Model* (Oxford Univ. Press, Oxford, 1997).
3. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287,** 2185–2195 (2000).
4. Celniker, S. E. *et al.* Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome Biol.* **3,** research0079.1–0079.14 (2002).
5. Richards, S. *et al.* Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and *cis*-element evolution. *Genome Res.* **15,** 1–18 (2005).
6. Myers, E. W. *et al.* A whole-genome assembly of *Drosophila*. *Science* **287,** 2196–2204 (2000).
7. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269,** 496–512 (1995).
8. Stark et al.. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* doi:10.1038/nature06340 (this issue).
9. Begun, D. J. *et al.* Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol.* **5,** e310, doi:10.1371/journal.pbio.0050310 (2007).
10. Zimin, A. V., Smith, D. R., Sutton, G. & Yorke, J. A. Assembly reconciliation. *Bioinformatics* (in the press).
11. Clary, D. O. & Wolstenholme, D. R. The mitochondrial DNA molecule of *Drosophila yakuba*: nucleotide sequence, gene organization, and genetic code. *J. Mol. Evol.* **22,** 252–271 (1985).
12. Ballard, J. W. When one is not enough: introgression of mitochondrial DNA in *Drosophila*. *Mol. Biol. Evol.* **17,** 1126–1130 (2000).
13. Montooth, K. L., Abt, D. N., Hoffman, J. & Rand, D. M. Evolution of the mitochondrial DNA across twelve species of *Drosophila*. *Mol. Biol. Evol.* (submitted).
14. Salzberg, S. *et al.* Serendipitous discovery of *Wolbachia* genomes in multiple *Drosophila* species. *Genome Biol.* **6,** R23 (2005).
15. Edgar, R. C. & Myers, E. W. PILER: identification and classification of genomic repeats. *Bioinformatics* **21,** i152–i158 (2005).
16. Smith, C. D. *et al.* Improved repeat identification and masking in Dipterans. *Gene* **389,** 1–9 (2007).
17. Li, Q. *et al.* ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole shotgun. *PloS Comput. Biol.* **1,** e43 (2005).
18. Bergman, C. M., Quesneville, H., Anxolabehere, D. & Ashburner, M. Recurrent insertion and duplication generate networks of transposable element sequences in the *Drosophila melanogaster* genome. *Genome Biol.* **7,** R112 (2006).
19. Guigo, R., Knudsen, S., Drake, N. & Smith, T. Prediction of gene structure. *J. Mol. Biol.* **226,** 141–157 (1992).
20. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5,** 59 (2004).
21. Gross, S. S. & Brent, M. R. Using multiple alignments to improve gene prediction. *J. Comput. Biol.* **13,** 379–393 (2006).
22. Gross, S. S., Do, C. B. & Batzoglou, S. in *BCATS 2005 Symposium Proc.* 82 (2005).
23. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14,** 988–995 (2004).
24. Slater, G. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6,** 31 (2005).
25. Chatterji, S. & Pachter, L. Reference based annotation with GeneMapper. *Genome Biol.* **7,** R29 (2006).
26. Souvorov, A. *et al.* in *NCBI News Fall/Winter, NIH Publication No. 04-3272* (eds Benson, D & Wheeler, D) (2006).
27. Honeybee Genome Sequencing Consortium. Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* **443,** 931–949 (2006).
28. Elsik, C. G. *et al.* Creating a honey bee consensus gene set. *Genome Biol.* **8,** R13 (2007).
29. Zhang, Y., Sturgill, D., Parisi, M., Kumar, S. & Oliver, B. Constraint and turnover in sex-biased gene expression in the genus *Drosophila*. *Nature* doi:10.1038/nature06323 (this issue).
30. Manak, J. R. *et al.* Biological function of unannotated transcription during the early development of *Drosophila melanogaster*. *Nature Genet.* **38,** 1151–1158 (2006).
31. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278,** 631–637 (1997).
32. Bhutkar, A., Russo, S., Smith, T. F. & Gelbart, W. M. Techniques for multi-genome synteny analysis to overcome assembly limitations. *Genome Informatics* **17,** 152–161 (2006).
33. Heger, A. & Ponting, C. Evolutionary rate analyses of orthologues and paralogues from twelve *Drosophila* genomes. doi:10.1101/gr6249707 *Genome Res.* (in the press).
34. Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302,** 205–217 (2000).
35. Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428,** 493–521 (2004).
36. Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
37. Harrison, P. M., Milburn, D., Zhang, Z., Bertone, P. & Gerstein, M. Identification of pseudogenes in the *Drosophila melanogaster* genome. *Nucleic Acids Res.* **31,** 1033–1037 (2003).
38. Bosco, G., Campbell, P., Leiva-Neto, J. & Markow, T. Analysis of *Drosophila* species genome size and satellite DNA content reveals significant differences among strains as well as between species. *Genetics* doi:10.1534/Genetics107.075069 (in the press).
39. Ranz, J. *et al.* Principles of genome evolution in the *Drosophila melanogaster* species group. *PLoS Biol.* **5,** e152, doi:10.1371/journal.pbio.0050152 (2007).
40. Noor, M. A. F., Garfield, D. A., Schaeffer, S. W. & Machado, C. A. Divergence between the *Drosophila pseudoobscura* and *D. persimilis* genome sequences in relation to chromosomal inversions. *Genetics* doi:10.1534/genetics.107.070672 (in the press).
41. Lewis, E. B. A gene complex controlling segmentation in *Drosophila*. *Nature* **276,** 565–570 (1978).
42. Negre, B., Ranz, J. M., Casals, F., Caceres, M. & Ruiz, A. A new split of the *Hox* gene complex in *Drosophila*: relocation and evolution of the gene labial. *Mol. Biol. Evol.* **20,** 2042–2054 (2003).
43. Von Allmen, G. *et al.* Splits in fruitfly *Hox* gene complexes. *Nature* **380,** 116 (1996).
44. Negre, B. & Ruiz, A. HOM-C evolution in *Drosophila*: is there a need for *Hox* gene clustering? *Trends Genet.* **23,** 55–59 (2007).
45. Dowsett, A. P. & Young, M. W. Differing levels of dispersed repetitive DNA among closely related species of *Drosophila*. *Proc. Natl Acad. Sci.* **79,** 4570–4574 (1982).
46. Kapitonov, V. V. & Jurka, J. DNAREP1_DM. (Repbase Update Release 3.4, 1999).
47. Kapitonov, V. V. & Jurka, J. Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. *Proc. Natl Acad. Sci. USA* **100,** 6569–6574 (2003).

48. Singh, N. D., Arndt, P. F. & Petrov, D. A. Genomic heterogeneity of background substitutional patterns in *Drosophila melanogaster. Genetics* **169,** 709–722 (2004).
49. Yang, H.-P., Hung, T.-L., You, T.-L. & Yang, T.-H. Genomewide comparative analysis of the highly abundant transposable element *DINE-1* suggests a recent transpositional burst in *Drosophila yakuba*. *Genetics* **173,** 189–196 (2006).
50. Yang, H.-P. & Barbash, D. Abundant and species-specific miniature inverted-repeat transposable elements in 12 *Drosophila* genomes. *Genome Biol.* (submitted).
51. Wilder, J. & Hollocher, H. Mobile elements and the genesis of microsatellites in dipterans. *Mol. Biol. Evol.* **18,** 384–392 (2001).
52. Marzo, M., Puig, M. & Ruiz, A. The foldback-like element *Galileo* belongs to the P superfamily of DNA transposons and is widespread within the genus *Drosophila*. *Proc. Natl Acad. Sci. USA* (submitted).
53. Casola, C., Lawing, A., Betran, E. & Feschotte, C. PIF-like transposons are common in *Drosophila* and have been repeatedly domesticated to generate new host genes. *Mol. Biol. Evol.* **24,** 1872–1888 (2007).
54. Abad, J. P. *et al.* Genomic analysis of *Drosophila melanogaster* telomeres: full-length copies of *HeT-A* and *TART* elements at telomeres. *Mol. Biol. Evol.* **21,** 1613–1619 (2004).
55. Abad, J. P. *et al. TAHRE*, a novel telomeric retrotransposon from *Drosophila melanogaster*, reveals the origin of *Drosophila* telomeres. *Mol. Biol. Evol.* **21,** 1620–1624 (2004).
56. Blackburn, E. H. Telomerases. *Annu. Rev. Biochem.* **61,** 113–129 (1992).
57. Villasante, A. *et al. Drosophila* telomeric retrotransposons derived from an ancestral element that as recruited to replace telomerase. *Genome Res.* (in the press).
58. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432,** 695–716 (2004).
59. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
60. *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282,** 2012–2018 (1998).
61. Mount, S. M., Gotea, V., Lin, C. F., Hernandez, K. & Makalowski, W. Spliceosomal small nuclear RNA genes in 11 insect genomes. *RNA* **13,** 5–14 (2007).
62. Schneider, C., Will, C. L., Brosius, J., Frilander, M. J. & Luhrmann, R. Identification of an evolutionarily divergent U11 small nuclear ribonucleoprotein particle in *Drosophila. Proc. Natl Acad. Sci. USA* **101,** 9584–9589 (2004).
63. Deng, X. & Meller, V. H. Non-coding RNA in fly dosage compensation. *Trends Biochem. Sci.* **31,** 526–532 (2006).
64. Amrien, H. & Axel, R. Genes expressed in neurons of adult male *Drosophila. Cell* **88,** 459–469 (1997).
65. Park, S.-W. *et al.* An evolutionarily conserved domain of roX2 RNA is sufficient for induction of H4-Lys16 acetylation on the *Drosophila* X chromosome. *Genetics* (in the press).
66. Stage, D. E. & Eickbush, T. H. Sequence variation within the rRNA gene loci of 12 *Drosophila* species. *Genome Res.* (in the press).
67. Hahn, M. W., De Bie, T., Stajich, J. E., Nguyen, C. & Cristianini, N. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res.* **15,** 1153–1160 (2005).
68. Hahn, M. W., Han, M. V. & Han, S.-G. Gene family evolution across 12 *Drosophila* genomes. *PLoS Biol.* **3,** e197 (2007).
69. Levine, M. T., Jones, C. D., Kern, A. D., Lindfors, H. A. & Begun, D. J. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proc. Natl Acad. Sci. USA* **103,** 9935–9939 (2006).
70. Ponce, R. & Hartl, D. L. The evolution of the novel *Sdic* gene cluster in *Drosophila melanogaster. Gene* **376,** 174–183 (2006).
71. Arguello, J. R., Chen, Y., Tang, S., Wang, W. & Long, M. Originiation of an X-linked testes chimeric gene by illegitimate recombination in *Drosophila. PLoS Genet.* **2,** e77 (2006).
72. Begun, D. J., Lindfore, H. A., Thompson, M. E. & Holloway, A. K. Recently evolved genes identified from *Drosophila yakuba* and *D. erecta* accessory gland expressed sequence tags. *Genetics* **172,** 1675–1681 (2006).
73. Betran, E., Thornton, K. & Long, M. Retroposed new genes out of the X in *Drosophila. Genome Res.* **12,** 1854–1859 (2002).
74. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21,** 650–659 (2005).
75. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13,** 555–556 (1997).
76. The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genet.* **25,** 25–29 (2000).
77. Storey, J. D. A direct approach to false discovery rates. *J. R. Stat. Soc. B* **64,** 479–498 (2002).
78. Larracuente, A. M. *et al.* Evolution of protein-coding genes in *Drosophila. Trends Genet.* (submitted).
79. Yang, Z., Nielsen, R., Goldman, N. & Pedersen, A. M. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155,** 431–449 (2000).
80. Bergman, C. M. *et al.* Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome. *Genome Biol.* **3,** research0086.1–0086.20 (2002).
81. Bierne, N. & Eyre Walker, A. C. The genomic rate of adaptive amino acid substitution in *Drosophila. Mol. Biol. Evol.* **21,** 1350–1360 (2004).
82. Sawyer, S. A., Kulathinal, R. J., Bustamante, C. D. & Hartl, D. L. Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. *J. Mol. Evol.* **57** (suppl. 1), S154–S164 (2003).
83. Sawyer, S. A., Parsch, J., Zhang, Z. & Hartl, D. L. Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila. Proc. Natl Acad. Sci. USA* **104,** 6504–6510 (2007).
84. Smith, N. G. & Eyre-Walker, A. Adaptive protein evolution in *Drosophila. Nature* **415,** 1022–1024 (2002).
85. Welch, J. J. Estimating the genomewide rate of adaptive protein evolution in *Drosophila. Genetics* **173,** 821–837 (2006).
86. Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O. & Arnold, F. H. Why highly expressed proteins evolve slowly. *Proc. Natl Acad. Sci. USA* **102,** 14338–14343 (2005).
87. Drummond, D. A., Raval, A. & Wilke, C. O. A single determinant dominates the rate of yeast protein evolution. *Mol. Biol. Evol.* **23,** 327–337 (2006).
88. Pal, C., Papp, B. & Hurst, L. D. Highly expressed genes in yeast evolve slowly. *Genetics* **158,** 927–931 (2001).
89. Pal, C., Papp, B. & Lercher, M. J. An integrated view of protein evolution. *Nature Rev. Genet.* **7,** 337–348 (2006).
90. Wall, D. P. *et al.* Functional genomic analysis of the rates of protein evolution. *Proc. Natl Acad. Sci. USA* **102,** 5483–5488 (2005).
91. Rocha, E. P. The quest for the universals of protein evolution. *Trends Genet.* **22,** 412–416 (2006).
92. Huntley, M. A. & Clark, A. G. Evolutionary analysis of amino acid repeats across the genomes of 12 *Drosophila* species. *Mol. Biol. Evol.* (in the press).
93. Charlesworth, B., Coyne, J. A. & Barton, N. H. The relative rates of evolution of sex chromosomes and autosomes. *Am. Nat.* **130,** 113–146 (1987).
94. Larsson, J. & Meller, V. H. Dosage compensation, the origin and the afterlife of sex chromosomes. *Chromosome Res.* **14,** 417–431 (2006).
95. Riddle, N. C. & Elgin, S. C. The dot chromosome of *Drosophila*: insights into chromatin states and their change over evolutionary time. *Chromosome Res.* **14,** 405–416 (2006).
96. Gordo, I. & Charlesworth, B. Genetic linkage and molecular evolution. *Curr. Biol.* **11,** R684–R686 (2001).
97. Singh, N. D., Larracuente, A. M. & Clark, A. G. Contrasting the efficacy of selection on the X and autosomes in *Drosophila. Mol. Biol. Evol.* (submitted).
98. Bhutkar, A., Russo, S. M., Smith, T. F. & Gelbart, W. M. Genome scale analysis of positionally relocated genes. *Genome Res.* (in the press).
99. Akashi, H. & Eyre-Walker, A. Translational selection and molecular evolution. *Curr. Opin. Genet. Dev.* **8,** 688–693 (1998).
100. Akashi, H., Kliman, R. M. & Eyre-Walker, A. Mutation pressure, natural selection, and the evolution of base composition in *Drosophila. Genetica (Dordrecht)* **102–103,** 49–60 (1998).
101. Bulmer, M. The selection–mutation–drift theory of synonymous codon usage. *Genetics* **129,** 897–908 (1991).
102. McVean, G. A. T. & Charlesworth, B. A population genetic model for the evolution of synonymous codon usage: Patterns and predictions. *Genet. Res.* **74,** 145–158 (1999).
103. Sharp, P. M. & Li, W. H. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.* **24,** 28–38 (1986).
104. Akashi, H. & Schaeffer, S. W. Natural selection and the frequency distributions of "silent" DNA polymorphism in *Drosophila. Genetics* **146,** 295–307 (1997).
105. Powell, J. R., Sezzi, E., Moriyama, E. N., Gleason, J. M. & Caccone, A. Analysis of a shift in codon usage in *Drosophila. J. Mol. Evol.* **57,** S214–S225 (2003).
106. Anderson, C. L., Carew, E. A. & Powell, J. R. Evolution of the *Adh* locus in the *Drosophila willistoni* group: The loss of an intron, and shift in codon usage. *Mol. Biol. Evol.* **10,** 605–618 (1993).
107. Rodriguez-Trelles, F., Tarrio, R. & Ayala, F. J. Switch in codon bias and increased rates of amino acid substitution in the *Drosophila saltans* species group. *Genetics* **153,** 339–350 (1999).
108. Rodriguez-Trelles, F., Tarrio, R. & Ayala, F. J. Evidence for a high ancestral GC content in *Drosophila. Mol. Biol. Evol.* **17,** 1710–1717 (2000).
109. Rodriguez-Trelles, F., Tarrio, R. & Ayala, F. J. Fluctuating mutation bias and the evolution of base composition in *Drosophila. J. Mol. Evol.* **50,** 1–10 (2000).
110. Heger, A. & Ponting, C. Variable strength of translational selection among twelve *Drosophila* species. *Genetics* (in the press).
111. Vicario, S., Moriyama, E. N. & Powell, J. R. Codon Usage in Twelve Species of *Drosophila. BMC Evol. Biol.* (submitted).
112. Singh, N. D., Arndt, P. F. & Petrov, D. A. Minor shift in background substitutional patterns in the *Drosophila saltans* and *willistoni* lineages is insufficient to explain GC content of coding sequences. *BMC Biol.* **4,** 10.1186/1741–7007–4-37 (2006).
113. Akashi, H. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. *Genetics* **139,** 1067–1076 (1995).
114. Akashi, H. Molecular evolution between *Drosophila melanogaster* and *D. simulans*: reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster. Genetics* **144,** 1297–1307 (1996).

115. Akashi, H. *et al.* Molecular evolution in the *Drosophila melanogaster* species subgroup: Frequent parameter fluctuations on the timescale of molecular divergence. *Genetics* **172,** 1711–1726 (2006).
116. Bauer DuMont, V., Fay, J. C., Calabrese, P. P. & Aquadro, C. F. DNA variability and divergence at the *Notch* locus in *Drosophila melanogaster* and *D. simulans*: a case of accelerated synonymous site divergence. *Genetics* **167,** 171–185 (2004).
117. McVean, G. A. & Vieira, J. The evolution of codon preferences in *Drosophila*: a maximum-likelihood approach to parameter estimation and hypothesis testing. *J. Mol. Evol.* **49,** 63–75 (1999).
118. Nielsen, R., Bauer DuMont, V., Hubisz, M. J. & Aquadro, C. F. Maximum likelihood estimation of ancestral codon usage bias parameters in *Drosophila*. *Mol. Biol. Evol.* **24,** 228–235 (2007).
119. Begun, D. J. The frequency distribution of nucleotide variation in *Drosophila* simulans. *Mol. Biol. Evol.* **18,** 1343–1352 (2001).
120. Singh, N. S., Bauer DuMont, V. L., Hubisz, M. J., Nielsen, R. & Aquadro, C. F. Patterns of mutation and selection at synonymous sites in *Drosophila*. *Mol. Biol. Evol.* doi:10.1093/mobev.msm196 (in the press).
121. McBride, C. S. & Arguello, J. R. Five *Drosophila* genomes reveal non-neutral evolution and the signature of host specialization in the chemoreceptor superfamily. *Genetics* (in the press).
122. Vieira, F. G., Sanchez-Gracia, A. & Rozas, J. Comparative genomic analysis of the odorant-binding protein family in 12 *Drosophila* genomes: Purifying selection and birth-and-death evolution. *Genome Biol.* **8,** 235 (2007).
123. Gardiner, A., Barker, D., Butilne, R. K., Jordan, W. C. & Ritchie, M. G. *Drosophila* chemoreceptor evolution: Selection, specialisation and genome size. *Genome Biol.* (submitted).
124. McBride, C. S. Rapid evolution of smell and taste receptor genes during host specialization in *Drosophila sechellia*. *Proc. Natl Acad. Sci. USA* **104,** 4996–5001 (2007).
125. Ranson, H. *et al.* Evolution of supergene families associated with insecticide resistance. *Science* **298,** 179–181 (2002).
126. Tijet, N., Helvig, C. & Feyereisen, R. The cytochrome P450 gene superfamily in *Drosophila melanogaster*. *Gene* **262,** 189–198 (2001).
127. Claudianos, C. *et al.* A deficit of detoxification enzymes: pesticide sensitivity and environmental response in the honeybee. *Insect Mol. Biol.* **15,** 615–636 (2006).
128. Low, W. L. *et al.* Molecular evolution of glutathione *S*-transferases in the genus *Drosophila*. *Genetics* (in the press).
129. Castellano, S. *et al. In silico* identification of novel selenoproteins in the *Drosophila melanogaster* genome. *EMBO Rep.* **2,** 697–702 (2001).
130. Martin-Romero, F. J. *et al.* Selenium metabolism in *Drosophila*: selenoproteins, selenoprotein mRNA expression, fertility, and mortality. *J. Biol. Chem.* **276,** 29798–29804 (2001).
131. Lemaitre, B. & Hoffmann, J. The host defense of *Drosophila melanogaster*. *Annu. Rev. Immunol.* **25,** 697–743 (2007).
132. Hughes, A. L. & Nei, M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335,** 167–170 (1988).
133. Murphy, P. M. Molecular mimicry and the generation of host defense protein diversity. *Cell* **72,** 823–826 (1993).
134. Schlenke, T. A. & Begun, D. J. Natural selection drives *Drosophila* immune system evolution. *Genetics* **164,** 1471–1480 (2003).
135. Sackton, T. B. *et al.* The evolution of the innate immune system across *Drosophila*. *Nature Genet.* (submitted).
136. Civetta, A. & Singh, R. S. High divergence of reproductive tract proteins and their association with postzygotic reproductive isolation in *Drosophila melanogaster* and *Drosophila virilis* group species. *J. Mol. Evol.* **41,** 1085–1095 (1995).
137. Civetta, A. Shall we dance or shall we fight? Using DNA sequence data to untangel controversies surrounding sexual selection. *Genome* **46,** 925–929 (2003).
138. Clark, N. L., Aagard, J. E. & Swanson, W. J. Evolution of reproductive proteins from animals and plants. *Reproduction* **131,** 11–22 (2006).
139. Haerty, W. *et al.* Evolution in the fast lane: rapidly evolving sex- and reproduction-related genes in *Drosophila* species. *Genetics* (in the press).
140. Lu, J. *et al.* Adaptive evolution of newly-emerged microRNA genes in *Drosophila*. *Mol. Biol. Evol.* (submitted).
141. Lai, E. C., Tomancak, P., Williams, R. W. & Rubin, G. M. Computational identification of *Drosophila* microRNA genes. *Genome Biol.* **4,** R42 (2003).
142. Parsch, J., Braverman, J. M. & Stephan, W. Comparative sequence analysis and patterns of covariation in RNA secondary structures. *Genetics* **154,** 909–921 (2000).
143. Stephan, W. The rate of compensatory evolution. *Genetics* **144,** 419–426 (1996).
144. Gallo, S. M., Li, L., Hu, Z. & Halfon, M. S. REDfly: a Regulatory Element Database for *Drosophila*. *Bioinformatics* **22,** 381–383 (2006).
145. Bergman, C. M., Carlson, J. W. & Celniker, S. E. *Drosophila* DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, *Drosophila melanogaster*. *Bioinformatics* **21,** 1747–1749 (2005).
146. Halligan, D. L. & Keightley, P. D. Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. *Genome Res.* **16,** 875–884 (2006).
147. Andolfatto, P. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* **437,** 1149–1152 (2005).
148. Bird, C. P., Stranger, B. E. & Dermitzakis, E. T. Functional variation and evolution of non-coding DNA. *Curr. Opin. Genet. Dev.* **16,** 559–564 (2006).
149. Wittkopp, P. J. Evolution of *cis*-regulatory sequence and function in Diptera. *Heredity* **97,** 139–147 (2006).
150. Ludwig, M. Z., Patel, N. H. & Kreitman, M. Functional analysis of *eve stripe 2* enhancer evolution in *Drosophila*. *Development* **125,** 949–958 (1998).
151. Down, A. T. A., Bergman, C. M., Su, J. & Hubbard, T. J. P. Large scale discovery of promoter motifs in *Drosophila melanogaster*. *PloS Comput. Biol.* **3,** e7 (2007).
152. Tamura, K., Subramanian, S. & Kumar, S. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Mol. Biol. Evol.* **21,** 36–44 (2004).
153. Kumar, S., Tamura, K. & Nei, M. MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief. Bioinform.* **5,** 150–163 (2004).
154. Pollard, D. A., Iyer, V. N., Moses, A. M. & Eisen, M. B. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. *PLoS Genet.* **2,** e173 (2006).
155. Bhutkar, A., Gelbart, W. M. & Smith, T. F. Inferring genome-scale rearrangement phylogeny and ancestral gene order: A *Drosophila* case study. *Genome Biol.* (in the press).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements Agencourt Bioscience Corporation, The Broad Institute of MIT and Harvard and the Washington University Genome Sequencing Center were supported by grants and contracts from the National Human Genome Research Insititute (NHGRI). T.C. Kaufman acknowledges support from the Indian Genomics Initiative.

Author Contributions The laboratory groups of A. G. Clark (including A. M. Larracuente, T. B. Sackton, and N. D. Singh) and Michael B. Eisen (including V. N. Iyer and D. A. Pollard) played the part of coordinating the primary writing and editing of the manuscript with the considerable help of D. R. Smith, C. M. Bergman, W. M. Gelbart, B. Oliver, T. A. Markow, T. C. Kaufman and M. Kellis. D. R. Smith served as primary coordinator for the assemblies. The remaining authors contributed either through their efforts in sequence production, assembly and annotation, or in the analysis of specific topics that served as the focus of more than 40 companion papers.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.G.C. (ac347@cornell.edu), M.B.E. (mbeisen@lbl.gov), D.R.S. (douglas.smith@agencourt.com), C.M.B. (casey.bergman@manchester.ac.uk), W.G. (gelbart@morgan.harvard.edu), B.O. (oliver@helix.nih.gov), T.A.M. (tmarkow@public.arl.arizona.edu), T.C.K. (kaufman@indiana.edu), M.K. (manoli@mit.edu), V.N.I. (venky@berkeley.edu), T.B.S. (tbs7@cornell.edu), A.M.L. (aml69@cornell.edu), D.A.P. (danielapollard@alum.bowdoin.edu), N.D.S. (nds25@cornell.edu), or collectively to 12flies@morgan.harvard.edu.

***Drosophila* 12 Genomes Consortium**

Project Leaders Andrew G. Clark[1], Michael B. Eisen[2,3], Douglas R. Smith[4], Casey M. Bergman[5], Brian Oliver[6], Therese A. Markow[7], Thomas C. Kaufman[8], Manolis Kellis[9,10] & William Gelbart[11,12]

Annotation Coordination Venky N. Iyer[13] & Daniel A. Pollard[14]

Analysis/Writing Coordination Timothy B. Sackton[1,15], Amanda M. Larracuente[1] & Nadia D. Singh[1]

Sequencing, Assembly, Annotation and Analysis Contributors Jose P. Abad[16], Dawn N. Abt[17], Boris Adryan[18], Montserrat Aguade[19], Hiroshi Akashi[20], Wyatt W. Anderson[21], Charles F. Aquadro[1], David H. Ardell[22], Roman Arguello[23], Carlo G. Artieri[24], Daniel A. Barbash[1], Daniel Barker[25], Paolo Barsanti[26], Phil Batterham[27], Serafim Batzoglou[28], Dave Begun[29], Arjun Bhutkar[11,30], Enrico Blanco[31], Stephanie A. Bosak[4], Robert K. Bradley[32], Adrianne D. Brand[4], Michael R. Brent[33], Angela N. Brooks[13], Randall H. Brown[33], Roger K. Butlin[34], Corrado Caggese[26], Brian R. Calvi[35], A. Bernardo de Carvalho[36], Anat Caspi[32], Sergio Castrezana[37], Susan E. Celniker[2], Jean L. Chang[10], Charles Chapple[31], Sourav Chatterji[38,39], Asif Chinwalla[40], Alberto Civetta[41], Sandra W. Clifton[40], Josep M. Comeron[42], James C. Costello[43], Jerry A. Coyne[23], Jennifer Daub[44], Robert G. David[4], Arthur L. Delcher[45], Kim Delehaunty[40], Chuong B. Do[28], Heather Ebling[4], Kevin Edwards[46], Thomas Eickbush[47], Jay D. Evans[48], Alan Filipski[49], Sven Findeiß[49,50], Eva Freyhult[22], Lucinda Fulton[40], Robert Fulton[40], Ana C. L. Garcia[51], Anastasia Gardiner[25], David A. Garfield[52], Barry E. Garvin[4], Greg Gibson[53], Don Gilbert[8], Sante Gnerre[10], Jennifer Godfrey[40], Robert Good[27], Valer Gotea[20], Brenton Gravely[54], Anthony J. Greenberg[1], Sam Griffiths-Jones[5,44], Samuel Gross[28], Roderic Guigo[31,55], Erik A. Gustafson[4], Wilfried Haerty[24], Matthew W. Hahn[8,43], Daniel L. Halligan[56], Aaron L. Halpern[57], Gillian M. Halter[20], Mira V. Han[43], Andreas Heger[58,59], LaDeana Hillier[40], Angie S. Hinrichs[60], Ian Holmes[32], Roger A. Hoskins[2], Melissa J. Hubisz[61], Dan Hultmark[62], Melanie A. Huntley[1], David B. Jaffe[10], Santosh Jagadeeshan[24], William R. Jeck[63], Justin Johnson[57], Corbin D. Jones[63], William C. Jordan[64], Gary H. Karpen[13,65], Eiko Kataoka[66], Peter D. Keightley[56], Pouya Kheradpour[9], Ewen F. Kirkness[57], Leonardo B. Koerich[36], Karsten Kristiansen[67], Dave

Kudrna[68], Rob J. Kulathinal[69], Sudhir Kumar[49,70], Roberta Kwok[8], Eric Lander[10], Charles H. Langley[29], Richard Lapoint[71], Brian P. Lazzaro[72], So-Jeong Lee[68], Lisa Levesque[41], Ruiqiang Li[67,73], Chiao-Feng Lin[20], Michael F. Lin[9,10], Kerstin Lindblad-Toh[10], Ana Llopart[42], Manyuan Long[23], Lloyd Low[27], Elena Lozovsky[69], Jian Lu[23], Meizhong Luo[68], Carlos A. Machado[7], Wojciech Makalowski[20], Mar Marzo[74], Muneo Matsuda[66], Luciano Matzkin[7], Bryant McAllister[42], Carolyn S. McBride[29], Brendan McKernan[4], Kevin McKernan[4], Maria Mendez-Lago[16], Patrick Minx[40], Michael U. Mollenhauer[20], Kristi Montooth[17], Stephen M. Mount[45,75], Xu Mu[20], Eugene Myers[76], Barbara Negre[77], Stuart Newfeld[70], Rasmus Nielsen[78], Mohamed A. F. Noor[52], Patrick O'Grady[71], Lior Pachter[38], Montserrat Papaceit[19], Matthew J. Parisi[4], Michael Parisi[6], Leopold Parts[9], Jakob S. Pedersen[60,79], Graziano Pesole[80], Adam M. Phillippy[45], Chris P. Ponting[58,59], Mihai Pop[45], Damiano Porcelli[26], Jeffrey R. Powell[81], Sonja Prohaska[49,82], Kim Pruitt[83], Marta Puig[74], Hadi Quesneville[84], Kristipati Ravi Ram[1], David Rand[17], Matthew D. Rasmussen[9], Laura K. Reed[53], Robert Reenan[85], Amy Reily[40], Karin A. Remington[57], Tania T. Rieger[86], Michael G. Ritchie[25], Charles Robin[27], Yu-Hui Rogers[57], Claudia Rohde[87], Julio Rozas[19], Marc J. Rubenfield[4], Alfredo Ruiz[74], Susan Russo[11,12], Steven L. Salzberg[45], Alejandro Sanchez-Gracia[19,88], David J. Saranga[4], Hajime Sato[66], Stephen W. Schaeffer[20], Michael C. Schatz[45], Todd Schlenke[89], Russell Schwartz[90], Carmen Segarra[19], Rama S. Singh[24], Laura Sirot[1], Marina Sirota[91], Nicholas B. Sisneros[68], Chris D. Smith[65,92], Temple F. Smith[30], John Spieth[40], Deborah E. Stage[47], Alexander Stark[9,10], Wolfgang Stephan[93], Robert L. Strausberg[57], Sebastian Strempel[93], David Sturgill[6], Granger Sutton[57], Granger G. Sutton[57], Wei Tao[4], Sarah Teichmann[18], Yoshiko N. Tobari[94], Yoshihiko Tomimura[95], Jason M. Tsolas[4], Vera L. S. Valente[51], Eli Venter[57], J. Craig Venter[57], Saverio Vicario[81], Filipe G. Vieira[19], Albert J. Vilella[19,96], Alfredo Villasante[16], Brian Walenz[57], Jun Wang[67,73], Marvin Wasserman[97], Thomas Watts[7], Derek Wilson[18], Richard K. Wilson[40], Rod A. Wing[68], Mariana F. Wolfner[1], Alex Wong[1], Gane Ka-Shu Wong[73,98], Chung-I Wu[23], Gabriel Wu[32], Daisuke Yamamoto[99], Hsiao-Pei Yang[1], Shiaw-Pyng Yang[40], James A. Yorke[100], Kiyohito Yoshida[101], Evgeny Zdobnov[102], Peili Zhang[11,12], Yu Zhang[6], Aleksey V. Zimin[100], Broad Institute Genome Sequencing Platform* & Broad Institute Whole Genome Assembly Team*

[1]Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853, USA. [2]Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. [3]Center for Integrative Genomics, Department of Molecular and Cell Biology, University of California at Berkeley, Berkeley, California 94720, USA. [4]Agencourt Bioscience Corporation, Beverly, Massachusetts 01915, USA. [5]Faculty of Life Sciences, University of Manchester, Manchester M13 9PT, UK. [6]Laboratory of Cellular and Developmental Biology, National Institutes of Health, Bethesda, Maryland 20892, USA. [7]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA. [8]Department of Biology, Indiana University, Bloomington, Indiana 47405, USA. [9]Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts 02139, USA. [10]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [11]Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [12]FlyBase, The Biological Laboratories, Harvard University, Cambridge, Massachusetts 02138, USA. [13]Department of Molecular and Cell Biology, University of California at Berkeley, Berkeley, California 94720, USA. [14]Biophysics Graduate Group, University of California at Berkeley, Berkeley, California 94720, USA. [15]Field of Ecology and Evolutionary Biology, Cornell University, Ithaca, New York 14853, USA. [16]Centro de Biología Molecular Severo Ochoa, Universidad Autónoma de Madrid, Madrid 28049, Spain. [17]Department of Ecology and Evolutionary Biology, Brown University, Providence, Rhode Island 02912, USA. [18]Structural Studies Division, MRC Laboratory of Molecular Biology, Cambridge CB2 2QH, UK. [19]Departament de Genètica, Universitat de Barcelona, Barcelona 08071, Spain. [20]Department of Biology, Pennsylvania State University, University Park, Pennsylvania 16802, USA. [21]Department of Genetics, University of Georgia, Athens, Georgia 30602, USA. [22]Linnaeus Centre for Bioinformatics, Uppsala Universitet, Uppsala, SE-75124, Sweden. [23]Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637, USA. [24]Department of Biology, McMaster University, Hamilton, Ontario, L8S 4K1, Canada. [25]School of Biology, University of St. Andrews, Fife KY16 9TH, UK. [26]Dipartimento di Genetica e Microbiologia dell'Università di Bari, Bari, 70126, Italy. [27]Department of Genetics, University of Melbourne, Melbourne 3010, Australia. [28]Computer Science Department, Stanford University, Stanford, California 94305, USA. [29]Section of Evolution and Ecology and Center for Population Biology, University of California at Davis, Davis, California 95616, USA. [30]BioMolecular Engineering Research Center, Boston University, Boston, Massachusetts 02215, USA. [31]Research Group in Biomedical Informatics, Institut Municipal d'Investigacio Medica, Universitat Pompeu Fabra, Barcelona 08003, Catalonia, Spain. [32]Department of Bioengineering, University of California at Berkeley, Berkeley, California 94720, USA. [33]Laboratory for Computational Genomics, Washington University, St Louis, Missouri 63108, USA. [34]Animal and Plant Sciences, The University of Sheffield, Sheffield S10 2TN, UK. [35]Department of Biology, Syracuse University, Syracuse, New York 13244, USA. [36]Departamento de Genética, Universidade Federal do Rio de Janeiro, Rio de Janeiro 21944-970, Brazil. [37]Tucson Stock Center, Tucson, Arizona 85721, USA. [38]Department of Mathematics, University of California at Berkeley, Berkeley, California 94720, USA. [39]Genome Center, University of California at Davis, Davis, California 95616, USA. [40]Genome Sequencing Center, Washington University School of Medicine, St Louis, Missouri 63108, USA. [41]Department of Biology, University of Winnipeg, Winnipeg, Manitoba R3B 2E9, Canada. [42]Department of Biological Sciences, University of Iowa, Iowa City, Iowa 52242, USA. [43]School of Informatics, Indiana University, Bloomington, Indiana 47405, USA. [44]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [45]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland 20742, USA. [46]Department of Biological Sciences, Illinois State University, Normal, Illinois 61790, USA. [47]Department of Biology, University of Rochester, Rochester, New York 14627, USA. [48]Bee Research Lab, USDA-ARS, Beltsville, Maryland 20705, USA. [49]Center for Evolutionary Functional Genomics, Biodesign Institute, Arizona State University, Tempe, Arizona 85287, USA. [50]Department of Computer Science, University of Leipzig, Leipzig 04107, Germany. [51]Departamento de Genética, Universidade Federal do Rio Grande do Sul, Porto Alegre/RS 68011, Brazil. [52]Department of Biology, Duke University, Durham, New Carolina 27708, USA. [53]Department of Genetics, North Carolina State University, Raleigh, North Caroline 27695, USA. [54]Health Center, University of Connecticut, Farmington, Connecticut 06030, USA. [55]Center of Genomic Regulation, Barcelona 8003, Catalonia, Spain. [56]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, UK. [57]J. Craig Venter Institute, Rockville, Maryland 20850, USA. [58]MRC Functional Genetics Unit, University of Oxford, Oxford OX1 3QX, UK. [59]Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford OX1 3QX, UK. [60]Center for Biomolecular Science and Engineering, University of California at Santa Cruz, Santa Cruz, California 95064, USA. [61]Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA. [62]Umeå Center for Molecular Pathogenesis, Umeå University, Umeå SE-90187, Sweden. [63]Department of Biology and Carolina Center for Genome Sciences, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [64]Institute of Zoology, Regent's Park, London NW1 4RY, UK. [65]Drosophila Heterochromatin Genome Project, Department of Genome and Computational Biology, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. [66]Kyorin University, School of Medicine, Mitaka, Tokyo 181-8611, Japan. [67]Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense M DK-5230, Denmark [68]Arizona Genomics Institute, Department of Plant Sciences and BIO5, University of Arizona, Tucson, Arizona 85721, USA. [69]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [70]School of Life Sciences, Arizona State University, Tempe, Arizona 85287, USA. [71]Department of Environmental Science, Policy and Management, University of California at Berkeley, Berkeley, California 94720, USA. [72]Department of Entomology, Cornell University, Ithaca, New York 14853, USA. [73]Beijing Genomics Institute at ShenZhen, ShenZhen 518083, China. [74]Departament Genètica i Microbiologia, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain. [75]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, Maryland 20742, USA. [76]Janelia Farm Research Campus, Howard Hughes Medical Institue, Ashburn, Viginia 20147-2408, USA. [77]Department of Zoology, University of Cambridge, Cambridge, CB2 3EJ, UK. [78]Institue of Biology, University of Copenhagen, DK-2100 Copenhagen Ø, Denmark. [79]Bioinformatics Centre, Department of Molecular Biology, University of Copenhagen, DK-2200 Copenhagen N, Denmark. [80]Dipartimento di Biochimica e Biologia Molecolare, Università di Bari and Istituto Tecnologie Biomediche del Consiglio Nazionale delle Ricerche, Bari 70126, Italy. [81]Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut 06520, USA. [82]Department of Biomedical Informatics, Arizona State University, Tempe, Arizona 85287, USA. [83]National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland 20894, USA. [84]Bioinformatics and Genomics Laboratory, Institut Jacques Monod, Paris, 75251, France. [85]Department of Molecular Biology, Cell Biology and Biochemistry, Brown University, Providence, Rhode Island 02912, USA. [86]Departamento de Genética, Centro de Ciências Biológicas, Universidade Federal de Pernambuco, Recife/PE 68011, Brazil. [87]Centro Acadêmico de Vitória, Universidade Federal de Pernambuco, Vitória de Santo Antão/PE, Brazil. [88]Cajal Institute, CSIC, Madrid 28002, Spain. [89]Department of Biology, Emory University, Atlanta, Georgia 30322, USA. [90]Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA. [91]Biomedical Informatics, Stanford University, Stanford, California 94305, USA. [92]Department of Biology, San Francisco State University, San Francisco, California 94132, USA. [93]Department of Biology, University of Munich, 82152 Planegg-Martinsried, Germany. [94]Institute of Evolutionary Biology, Setagaya-ku, Tokyo 158-0098, Japan. [95]Shiba Gakuen, Minato-ku, Tokyo 105-0011, Japan. [96]European Bioinformatics Institute, Hinxton, CB10 1SD, UK. [97]Department of Biology, City University of New York at Queens, Flushing, New York 11367, USA. [98]Department of Biological Sciences and Department of Medicine, University of Alberta, Edmonton, Alberta T6G 2E9, Canada. [99]Department of Developmental Biology and Neurosciences, Tohoku University, Sendai 980-8578, Japan. [100]Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742, USA. [101]Hokkaido University, EESBIO, Sapporo, Hokkaido 060-0810, Japan. [102]Faculty of Medecine, Universite de Geneve, Geneva CH-1211, Switzerland.

*__Broad Institute Genome Sequencing Platform__ Jennifer Baldwin[10], Amr Abdouelleil[10], Jamal Abdulkadir[10], Adal Abebe[10], Brikti Abera[10], Justin Abreu[10], St Christophe Acer[10], Lynne Aftuck[10], Allen Alexander[10], Peter An[10], Erica Anderson[10], Scott Anderson[10], Harindra Arachi[10], Marc Azer[10], Pasang Bachantsang[10], Andrew Barry[10], Tashi Bayul[10], Aaron Berlin[10], Daniel Bessette[10], Toby Bloom[10], Jason Blye[10], Leonid Boguslavskiy[10], Claude Bonnet[10], Boris Boukhgalter[10], Imane Bourzgui[10], Adam Brown[10], Patrick Cahill[10], Sheridon Channer[10], Yama Cheshatsang[10], Lisa Chuda[10], Mieke Citroen[10], Alville Collymore[10], Patrick Cooke[10], Maura Costello[10], Katie D'Aco[10], Riza Daza[10], Georgius De Haan[10], Stuart DeGray[10], Christina DeMaso[10], Norbu Dhargay[10], Kimberly Dooley[10], Erin Dooley[10], Missole Doricent[10], Passang Dorje[10], Kunsang Dorjee[10], Alan Dupes[10], Richard Elong[10], Jill Falk[10], Abderrahim Farina[10], Susan Faro[10], Diallo Ferguson[10], Sheila Fisher[10], Chelsea D. Foley[10], Alicia Franke[10], Dennis Friedrich[10], Loryn Gadbois[10], Gary Gearin[10], Christina R. Gearin[10], Georgia Giannoukos[10], Tina Goode[10], Joseph Graham[10], Edward Grandbois[10], Sharleen Grewal[10], Kunsang Gyaltsen[10], Nabil Hafez[10], Birhane Hagos[10], Jennifer Hall[10], Charlotte Henson[10], Andrew Hollinger[10], Tracey Honan[10], Monika D. Huard[10], Leanne Hughes[10], Brian Hurhula[10], M Erii Husby[10], Asha Kamat[10], Ben Kanga[10],

Seva Kashin[10], Dmitry Khazanovich[10], Peter Kisner[10], Krista Lance[10], Marcia Lara[10], William Lee[10], Niall Lennon[10], Frances Letendre[10], Rosie LeVine[10], Alex Lipovsky[10], Xiaohong Liu[10], Jinlei Liu[10], Shangtao Liu[10], Tashi Lokyitsang[10], Yeshi Lokyitsang[10], Rakela Lubonja[10], Annie Lui[10], Pen MacDonald[10], Vasilia Magnisalis[10], Kebede Maru[10], Charles Matthews[10], William McCusker[10], Susan McDonough[10], Teena Mehta[10], James Meldrim[10], Louis Meneus[10], Oana Mihai[10], Atanas Mihalev[10], Tanya Mihova[10], Rachel Mittelman[10], Valentine Mlenga[10], Anna Montmayeur[10], Leonidas Mulrain[10], Adam Navidi[10], Jerome Naylor[10], Tamrat Negash[10], Thu Nguyen[10], Nga Nguyen[10], Robert Nicol[10], Choe Norbu[10], Nyima Norbu[10], Nathaniel Novod[10], Barry O'Neill[10], Sahal Osman[10], Eva Markiewicz[10], Otero L. Oyono[10], Christopher Patti[10], Pema Phunkhang[10], Fritz Pierre[10], Margaret Priest[10], Sujaa Raghuraman[10], Filip Rege[10], Rebecca Reyes[10], Cecil Rise[10], Peter Rogov[10], Keenan Ross[10], Elizabeth Ryan[10], Sampath Settipalli[10], Terry Shea[10], Ngawang Sherpa[10], Lu Shi[10], Diana Shih[10], Todd Sparrow[10], Jessica Spaulding[10], John Stalker[10], Nicole Stange-Thomann[10], Sharon Stavropoulos[10], Catherine Stone[10], Christopher Strader[10], Senait Tesfaye[10], Talene Thomson[10], Yama Thoulutsang[10], Dawa Thoulutsang[10], Kerri Topham[10], Ira Topping[10], Tsamla Tsamla[10], Helen Vassiliev[10], Andy Vo[10], Tsering Wangchuk[10], Tsering Wangdi[10], Michael Weiand[10], Jane Wilkinson[10], Adam Wilson[10], Shailendra Yadav[10], Geneva Young[10], Qing Yu[10], Lisa Zembek[10], Danni Zhong[10], Andrew Zimmer[10] & Zac Zwirko[10] **Broad Institute Whole Genome Assembly Team** David B. Jaffe[10], Pablo Alvarez[10], Will Brockman[10], Jonathan Butler[10], CheeWhye Chin[10], Sante Gnerre[10], Manfred Grabherr[10], Michael Kleber[10], Evan Mauceli[10] & Iain MacCallum[10]

VITA

Chiao-Feng Lin 208 Mueller Lab
CXL46@psu.edu University Park, PA 16802

EDUCATION

Ph.D.	Biology (Molecular Evolutionary Biology option), Pennsylvania State University, University Park, PA	expected Summer 2008
MLIS	Information Science, University of Texas, Austin, TX	2000
BS	Electronic Engr., Chung-Yuan Christian University, Chung-Li, Taiwan	1992

ACADEMIC AND PROFESSIONAL EXPERIENCE

2001-2002	Project Asst. with Dr. Hiroshi Akashi, Dept. of Biology, Penn State Univ.
2000	Research Asst., Tarlton Law Library, Univ. of Texas School of Law, Austin, TX
1995-1999	Technician, Computer Center, China Medical College, Taichung Taiwan

AWARDS AND HONORS

Graduate School of Library and Information Science Academic Competitive Scholarship, University of Texas at Austin. 2000

PEER-REVIEWED PUBLICATIONS

Drosophila 12 Genomes Consortium
Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 2007 Nov 8; 450, 203–218.

Makalowska I, **Lin CF**, Hernandez K.
Birth and death of gene overlaps in vertebrates. BMC *Evol Biol.* 2007 Oct 16;7(1):193

Mount SM, Gotea V, **Lin CF**, Hernandez K, Makalowski W.
Spliceosomal small nuclear RNA genes in 11 insect genomes. *RNA*. 2007 Jan;13(1):5-14.

Akashi H, Ko WY, Piao S, John A, Goel P, **Lin CF**, Vitins AP.
Molecular evolution in the *Drosophila melanogaster* species subgroup: frequent parameter fluctuations on the timescale of molecular divergence. *Genetics*. 2006 Mar;172(3):1711-26.

Makalowska I, **Lin CF,** Makalowski W.
Overlapping genes in vertebrate genomes. *Computational Biology and Chemistry*. Volume 29, Issue 1, February 2005, Pages 1-12

Lewandowska D, Simpson CG, Clark GP, Jennings NS, Barciszewska-Pacak M, **Lin CF**, Makalowski W, Brown JW, Jarmolowski A.
Determinants of plant U12-dependent intron splicing efficiency. *Plant Cell*. 2004 May;16(5):1340-52.

REVIEWER OF MANUSCRIPTS

Gene

TEACHING EXPERIENCE

Biology 12 - Introductory Biology, Teaching Assistant (Spring 2004, 2005, 2006)
Biology 110 - Biology: Basic Concepts and Biodiversity, Teaching Assistant (Fall 2004, 2006)
Biology 439 - Practical Bioinformatics, Teaching Assistant (Spring 2007)
Bioinformatics Workshop at PSU, Teaching Assistant (Summer 2003, 2004)