Genesis of gene structures and computational analysis of U12-type introns

Matthew Devin Wilkerson

Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd

Part of the Bioinformatics Commons, Genetics and Genomics Commons, and the Molecular Biology Commons

Recommended Citation

Wilkerson, Matthew Devin, "Genesis of gene structures and computational analysis of U12-type introns" (2007). Retrospective Theses and Dissertations. 15615.

https://lib.dr.iastate.edu/rtd/15615

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Genesis of gene structures and computational analysis of U12-type introns

by

Matthew Devin Wilkerson

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Volker Brendel, Co-major Professor
Thomas Peterson, Co-major Professor
  Adam Bogdanove
  David Fernández-Baca
  Carolyn Lawrence

Iowa State University
Ames, Iowa
2007

Copyright © Matthew Devin Wilkerson, 2007. All rights reserved
This is dedicated to my parents.
# TABLE OF CONTENTS

## CHAPTER 1: GENERAL INTRODUCTION
- Introduction 1
- References 3
- Thesis Organization 4

## CHAPTER 2: TRACEMBLER – SOFTWARE FOR IN-SILICO CHROMOSOME WALKING IN UNASSEMBLED GENOMES
- Abstract 6
- Background 8
- Implementation 9
- Results and discussion 11
- Conclusions 15
- Availability and requirements 15
- Authors' contributions 16
- Acknowledgements 16
- Appendix 16
- References 16
- Figures 19

## CHAPTER 3: yrGATE: A WEB-BASED GENE-STRUCTURE ANNOTATION TOOL FOR THE IDENTIFICATION AND DISSEMINATION OF EUKARYOTIC GENES
- Abstract 20
- Rationale 21
- Annotation Tool 23
- Community Annotation Utilities 27
- Implementations and case studies 28
- Community Annotation at PlantGDB 29
- yrGATE with DAS input 31
- Usability and availability 33
- Conclusions 34
- List of abbreviations 34
CHAPTER 1: GENERAL INTRODUCTION

Introduction

Completely sequenced genomes provide a wealth of information that has allowed the exploration of large scale biological questions and continues to provide a critical resource for the advancement of biological research. Previously, the number of completely sequenced genomes was small and was generally limited to the model organisms. Currently, the number of genomes completely or partially sequenced is rapidly increasing, with 338 different eukaryotic genomes available as of October 2007 [1]. With a genome sequence in hand, the typical first step is gene structure annotation, or identifying the location and structural features of genes within the genome sequence, after which functional descriptions of the genes, relationships to homologous genes and other higher level research questions can be investigated. Annotation, then, imparts the biology of the organism onto the genome sequence [2]. The goal of this thesis is to provide useful computational tools for gene annotation in emerging and mature genomes and to analyze a particularly difficult-to-annotate gene feature.

The process of gene structure annotation requires a genomic sequence of sufficient size so that it can contain a full gene, which in eukaryotes can be thousands of nucleotides. The popular method of whole genome shotgun sequencing to furnish genomic sequences produces small sequence fragments of hundreds of nucleotides, which are eventually assembled into chromosome sequences, and can take several years from start to finish. In the interim, these small sequence fragments are deposited into repositories [3] for historical reference and dissemination purposes, but because they are too small to contain a gene, these
fragments are not particularly useful for gene structure annotation purposes. I have
developed a web-based tool, Tracembler, which facilitates dynamic gene annotation of these
fragments through on-the-fly sequence similarity searching and assembly. Hence,
Tracembler allows biologists and interested scientists to immediately create gene annotations
upon the latest sequences from emerging genomes without having to wait for the completion
of the genome sequencing project.

On the other end of the genome maturity spectrum, accurate gene structure
annotation, which includes the biologically-correct specification of exons, introns,
untranslated regions, protein coding regions, and alternatively spliced variants of a gene,
remains a challenge for completely sequenced genomes [4-6]. Pure computational
approaches are excellent for providing an approximate initial summary of an organism's gene
space, but they are not completely accurate or comprehensive [4-6]. Manual annotation by a
human curator, who inspects and reviews the available evidence to make decisions in
constructing a gene structure annotation, is considered the highest quality method [4].
Hindrances to manual annotation are that it is time consuming, has restricted participation,
and is not easy to conduct. To remove these limitations of manual annotation, I developed
the yrGATE ("your Gene structure Annotation Tool for Eukaryotes") software, which
enables individuals to create gene structure annotations using high quality evidence through
an easy-to-use dynamic web browser interface and submit their annotations to a community
database.

A particular category of often mis-annotated genes is those containing U12-type
introns. U12-type introns are a class of introns that have highly conserved sequence features,
have a specific spliceosome that processes their removal from pre-mRNA transcripts, and
comprise less than 1% of the introns in any studied eukaryotic organism [7, 8]. One reason for their mis-annotation is most gene prediction programs are not designed to specifically recognize them [9], which is likely caused by U12-type introns’ unique sequence features and rare occurrence. Apart from their mis-annotation, U12-type introns are intriguing due to their unique proposed evolutionary history [8] and due to their maintenance in organisms at very low frequencies in a seemingly functional redundancy with the major splicing system [10]. In order to further the understanding of this intriguing gene feature, a large-scale annotation and computational investigation of U12-type introns in the context of their host genes and evolution was completed.

References

Thesis Organization

This thesis consists of five chapters. Chapter 1 provides a general introduction. Chapters 2 and 3 present papers published in peer-reviewed journals. Chapter 4 presents a manuscript prepared for submission to a journal. Chapter 5 provides a general conclusion.

Chapter 2, 'Tracembler - software for in silico chromosome walking in unassembled genomes', has been published in *BMC Bioinformatics* in 2007, Volume 8, Number 151. The authors of this publication provided the following contributions. Qunfeng originated the initial idea for the project, contributed to the software, performed analysis, and was responsible for the majority of the writing. Matthew Wilkerson contributed central ideas of the project, wrote the majority of the software, and contributed to the writing and analysis. The contributions of Qunfeng Dong and Matthew Wilkerson were deemed equal, which is the cause for co-first authorship between the authors. Volker Brendel provided supervision, feedback for the project, and contributed to the writing.

Chapter 3, 'yrGATE: a web-based gene-structure annotation tool for the identification and dissemination of eukaryotic genes', has been published in *Genome Biology* in 2006, Volume 7 electronic release 58. The authors of this publication provided the following contributions. Matthew Wilkerson designed and wrote the software, provided the usage cases, and wrote the majority of the manuscript. Shannon Schlueter contributed useful ideas and contributed to an earlier version of the software. Volker Brendel contributed to the writing and provided supplementary ideas for the project.

Chapter 4, 'Computational analysis of U12-type introns within orthologous genes' has been prepared for submission to *PLoS Computational Biology*. The authors of this paper provided the following contributions. Matthew Wilkerson provided the ideas for the project,
designed and wrote the software, performed analysis, and wrote the manuscript. Volker
Brendel provided initial ideas, supporting ideas, and advisory support throughout the project.
CHAPTER 2: TRACEMBLER – SOFTWARE FOR IN-SILICO
CHROMOSOME WALKING IN UNASSEMBLED GENOMES

A paper published in BMC Bioinformatics¹

Qunfeng Dong², Matthew D. Wilkerson², and Volker Brendel²³⁴

Abstract

**Background:** Whole genome shotgun sequencing produces increasingly higher coverage of a
genome with random sequence reads. Progressive whole genome assembly and eventual
finishing sequencing is a process that typically takes several years for large eukaryotic
genomes. In the interim, all sequence reads of public sequencing projects are made available
in repositories such as the NCBI Trace Archive. For a particular locus, sequencing coverage
may be high enough early on to produce a reliable local genome assembly. We have
developed software, Tracembler, that facilitates in silico chromosome walking by recursively
assembling reads of a selected species from the NCBI Trace Archive starting with reads that
significantly match sequence seeds supplied by the user.

¹Reprinted with permission of *BMC Bioinformatics*, 2007, 8:151

²These authors contributed equally to this work

²Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA
50011-3260, USA

³Department of Statistics, Iowa State University, Ames, IA 50011-3260, USA

⁴Author for correspondence
Results. *Tracembler* takes one or multiple DNA or protein sequence(s) as input to the NCBI Trace Archive BLAST engine to identify matching sequence reads from a species of interest. The BLAST searches are carried out recursively such that BLAST matching sequences identified in previous rounds of searches are used as new queries in subsequent rounds of BLAST searches. The recursive BLAST search stops when either no more new matching sequences are found, a given maximal number of queries is exhausted, or a specified maximum number of rounds of recursion is reached. All the BLAST matching sequences are then assembled into contigs based on significant sequence overlaps using the CAP3 program. We demonstrate the validity of the concept and software implementation with an example of successfully recovering a full-length ChrM2 gene as well as its upstream and downstream genomic regions from *Rattus norvegicus* reads. In a second example, a query with two adjacent *Medicago truncatula* genes as seeds resulted in a contig that likely identifies the microsyntenic homologous soybean locus.

Conclusions. *Tracembler* streamlines the process of recursive database searches, sequence assembly, and gene identification in resulting contigs in attempts to identify homologous loci of genes of interest in species with emerging whole genome shotgun reads. A web server hosting *Tracembler* is provided at http://www.plantgdb.org/tool/tracembler/, and the software is also freely available from the authors for local installations.
Background

Comparative genomics is based on the identification and alignment of homologous genes across multiple species and has become a standard, powerful approach in molecular biology for many purposes, including characterization of structurally and functionally important motifs in gene families. Typically, this approach starts with a set of query sequences as input to sequence similarity-based database search programs such as BLAST [1] to identify significantly similar matches in the sequence databases of species of interest. If the species of interest are fully sequenced and evolutionarily close enough to the query species, then this approach will yield the homologous genes in their genomic context. However, the comparative genomics approach is currently limited by the sparse sampling of eukaryotic species from the tree of life that have been sequenced as model organisms. For example, so far only three species (Arabidopsis thaliana, Oryza sativa, and Populus trichocarpa) from the entire plant kingdom have been completely sequenced. At the same time, a large number of on-going sequencing projects (see, e.g., [2, 3]) are generating large numbers of short (yet unassembled) genomic sequences through strategies such as whole-genome shotgun or BAC-by-BAC minimum tiling path sequencing. These sequence reads are continuously made available through the NCBI Trace Archive [4]. In the summer of 2006 the archive topped one billion reads [5].

Because the deposited sequenced reads are short (400-800 bp), a simple query of the repository with a DNA or protein seed (e.g., NCBI's Trace Archive discontiguous Mega BLAST Server [6] or Ensembl's Trace Server [7]) will typically only tag this gene as present in the target genome. Depending on the genome sequence coverage, the query may result in
redundant and overlapping tags. Analysis of the resulting set of reads without the help of an assembly program could become very tedious, particularly if one wishes to obtain the genome context of the tagged gene further upstream or downstream, which would require additional rounds of repository searches.

To facilitate the task of homolog identification in the trace archive repository of an ongoing genome sequencing project, we have developed software that seamlessly integrates recursive database searches and contig assembly and interpretation. Depending on the depth of the current sequencing effort, the final results returned by Tracembler will ideally be full-length genomic sequences that are homologs of the user-supplied query sequences.

Implementation

The Tracembler algorithm is illustrated in Fig. 1. As input, the program takes a single or multiple user-supplied query sequences (either nucleotide or protein), an E-value cutoff, and a user selected Trace Archive database, which contains sequence reads from a particular species deposited at the NCBI Trace Archive. An initial search is then initiated via the remote BLAST service provided by NCBI [6] [8]. Because the searches are always performed directly at NCBI (conducted transparently to the user), users are assured to search against the most up-to-date sequence read repository. If the initial BLAST search returns significant matching sequences (based on the user-specified E-value cutoff), these matching sequences are considered as queries for further database searches, which can potentially extend the initial matching region in both 5' and 3' direction (in-silico chromosome walking).
This process is automatically iterated until either no more new matching sequences are found, a given maximal number of queries is exhausted, or a specified maximum number of rounds of recursion is reached. Only newly identified non-redundant matching sequences from the previous round are used as queries for the next round BLAST search. The ceiling on the number of rounds of recursion is imposed to prevent assembly of more than the local regions surrounding the genes of interest, thus protecting the intellectual properties of the whole-genome sequencing project by preventing large-scale assembly (e.g., [9]).

Additionally, Tracembler has the following rules for polite dynamic NCBI data requests: mandatory pauses between data requests and BLAST job submissions, a maximum number of requests for an individual BLAST result, fixed limits on the BLAST parameter values.

All the obtained BLAST matching sequences are considered as potential genomic constituents of homologous regions of the original user-supplied query sequence and are assembled with the CAP3 program [10]. In addition to the actual sequences, quality scores and mate-pair distance constraints are also critical for high-quality assembly. Therefore, the quality score and mate-pair distance constraint information for each read are dynamically retrieved from the Trace Archive and used in the assembly to evaluate the significance of sequence overlaps. Multiple contigs may result from coverage gaps in one locus or represent duplicated loci. The CAP3-generated contigs are compared with the original user-supplied query sequences using BLAST [1] and GenomeThreader [11] to assess and display the extent of similarity and coverage. After completion of the analysis, an email is sent to the user indicating URLs to view all the results, including the assembled contig sequences, the multiple-sequence alignment underlying the assembly, as well as the pair-wise alignments.
between the original query and the contigs. In addition, the intermediate files (matching sequences, quality scores, and mate-pair distance constraints), the recursive BLAST output, the CAP3 output files, and a log file are included in the result. These additional files permit interested users to download and locally analyze their data further, such as using a different assembly program.

Results and discussion

Validation

To validate Tracembler, we first tested the software by trying to re-assemble a published genome region from trace reads matching an annotated gene. Our test case used the rat (Rattus norvegicus) Chrm2 gene sequence (cholinergic receptor, muscarinic 2; 2,072 bp, [12]) as query against the entire rat whole-genome shotgun sequence reads. The gene was picked randomly. The rat genome has already been fully sequenced and assembled [13], and thus Tracembler was expected to assemble a contig that matches the published genome. As shown in Appendix Figure 2, Tracembler produced a single contig of length 5,068 bp. This contig covers the entire, perfectly matched Chrm2 gene. The entire contig matches very well to chromosome 4 (GenBank accession# NC_005103.2) from base positions 63,909,839 to 63,914,888 (99% identity over the entire match as reported by bl2seq [14]). The Chrm2 gene maps from base positions 63,911,288 to 63,913,359 [15], and thus the contig generated by Tracembler not only recovered the full-length annotated Chrm2 gene but also successfully “walked” 1,449 bp into the 5' upstream and 1,529 bp into the 3' down-stream regions.
Application

Next, we discuss a Tracembler application that revealed microsynteny between Medicago truncatula and Glycine max, which are thought to have diverged through speciation around 50 MYA [16]. *M. truncatula* is an established model organism for the legumes, with a nearly complete sequencing and annotation effort [17]. Whole genome shot-gun sequencing of *G. max* (soybean) has only recently been initiated [18], with currently more than 1.3 million unassembled and unannotated sequence reads deposited in the NCBI Trace Archive. Soybean is the most valuable legume crop [19], and establishing its syntenic relationship with *M. truncatula* is critical for transferring knowledge from this model organism.

In *M. truncatula*, the “SWIM zinc finger” gene (AC146590g10v2) is annotated on a BAC clone (mth2-145p10) from position 50,413 to 49,886 [20]. 3’ downstream of this gene, there is another “hypothetical” gene (AC146590g11v2) annotated from positions 52,448 to 50,777. According to the current *M. truncatula* genome annotation, the “SWIM zinc finger” gene and its neighboring “hypothetical” gene are only 364 bp apart. In order to investigate whether such close distance is likely a result of mis-annotation (“hypothetical” genes are often wrongly predicted by gene-finding software), we took the protein sequences encoded by these two *M. truncatula* genes as input and used Tracembler to search against soybean sequence reads at NCBI Trace Archive. Interestingly, one 4,172 bp soybean genomic contig obtained from the assembly does match well to both the “SWIM zinc finger” and the “hypothetical” protein from positions 1,459 to 1,950 and from positions 2,299 to 2,839 of this contig, respectively (see Appendix Figure 3: Tracembler validation and applications). Thus, there appear to be homologs of the Medicago genes on the soybean genome in similar
proximity (349 bp apart) as on the Medicago genome. Our result provides compelling evidence that the two genes are highly conserved between *M. truncatula* and soybean. In particular, the high conservation of the "hypothetical" gene suggests that it is a true gene.

**Performance**

The performance of *Tracembler* is mainly determined by three factors. The first is the sequencing depth of the target genome, which provides a boundary of the expected extent of read overlaps and therefore assembled contig length. The second factor is the abundance of gene duplications in the genome of interests. For plant genomes, in which many gene duplications have occurred through tandem or whole genome duplication events, multiple homologs of genes of interests may have been sampled by the deposited sequence reads and show up as close BLAST matching sequences in the initial stage of *Tracembler*. If the multiple gene copies are sufficiently diverged, the CAP3 program will split them into different contigs. The pairwise comparison between the original user-supplied queries and the final contigs in the final step of *Tracembler* may identify the likely ortholog of the query based on highest match score. Third, because *Tracembler* relies on the up-to-date NCBI Trace Archive BLAST search engine over the Web, the response time for users will be affected by network traffic as well as the current work-load at the search engine. Various parameter settings deal with the stringency of matching and extent of the search, which will also affect speed and quality of the results.
Other programs

During the preparation of this manuscript, we became aware of a published software package, GENOTRACE, from Berezikov et al. [21] that is similar to our Tracembler. In addition to the choices of embedded external computer programs (e.g., BLAST vs. SSAHA [22] for database searching, CAP3 vs. Phrap [23] for assembly) that likely produce different final outputs, there are several subtle yet important differences between Tracembler and GENOTRACE that matter to the general biology user community. First, GENOTRACE requires maintaining a local copy of NCBI Trace Archive. Although this approach improves the search speed, the required amount of disk space is enormous (currently more than 1.2 TBytes in compressed format at NCBI), which is beyond a typical user’s resources and is superfluous for the task of exploring just a few genes of interest. By contrast, Tracembler takes advantage of the dynamic API provided by NCBI and sends query sequences via the internet to directly search the Trace Archive at NCBI. This not only eases the installation and maintenance for the users, but also ensures that users are always searching the most up-to-date version of Trace Archive. Furthermore, because trace sequences can often accumulate in amounts of hundreds of thousands of sequences per organism per week [24], GENOTRACE’s requirement of a local copy of NCBI’s Trace archive necessitates frequent downloading and processing of local files, which is an obstacle for widespread use. Second, only DNA sequence can be used as query for GENOTRACE, whereas Tracembler can take either DNA or protein sequences as input (the program automatically detects the type of sequences and performs appropriate BLAST-searches, MEGABLAST or TBLASTN, accordingly). Third, GENOTRACE is restricted to one query sequence at a time. As
demonstrated by our application example above, there are instances where it is more convenient to allow multiple seeds spanning one region of interest in one genome to search another genome.

Conclusions

Biologists are often left with an eager sense of anticipation when their species of interest are in the process of being sequenced but the sequencing reads have not yet been assembled. Our Tracembler server, although algorithmically simple, provides an elegant solution for biologists to recover genomic regions of interest from species with on-going sequencing project before the whole genome assemblies are published.

Availability and requirements

The Tracembler program is freely accessible, using a web browser at http://www.plantgdb.org/tool/tracembler. The software, written in Perl and designed for use on Linux machines, is also freely available for local installation by download from http://gremlin1.gdcb.iastate.edu/bgroup/download/tr/download.html. Instructions on obtaining the required external free programs (in particular, CAP3 [10]) are provided with the software.
Authors' contributions

VB designed and supervised the project. QD and MW implemented the computer program. QD analyzed the results and prepared the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported in part by NSF grant DBI-0606909 and by a Specific Cooperative Agreement with the Agricultural Research Service, U.S. Department of Agriculture (no. 58-3625-5-124) awarded to VB. QD is currently supported by the Indiana METACyt Initiative, funded by the Lilly Endowment, Inc. We are grateful for constructive suggestions from two anonymous reviewers.

Appendix


References


7. Ensembl Trace Server [http://trace.ensembl.org/cgi-bin/tracesearch]
25. **NCBI QBlast's URL API. User's Guide**  
Tracembler accepts as input one or more user-supplied query sequences and parameter specifications. The query sequence(s) and associated parameters are submitted using the QBLAST URL API to NCBI [25] (1). Tracembler analyzes these results, and if there are new sequences matching at a significance level below the user supplied E-value parameter, these sequences are used as queries in a new BLAST search (2). One round consists of BLAST searches of all acceptable matching sequences from the previous round. This process is repeated in a recursive manner until either all matching sequences are exhausted, a user-defined maximum round of recursion is reached, or a user defined maximum total number of BLAST queries is reached. For the final set of sequences, quality score and mate-pair distance constraint information is retrieved from NCBI. These sequences are assembled using CAP3 (3). Finally, novel contigs are compared to the query sequences using BLAST for local alignment and GenomeThreader for spliced-alignment (4).
CHAPTER 3: yrGATE: A WEB-BASED GENE-STRUCTURE ANNOTATION TOOL FOR THE IDENTIFICATION AND DISSEMINATION OF EUKARYOTIC GENES

A paper published in Genome Biology¹

Matthew D. Wilkerson², Shannon D. Schlueter², and Volker Brendel²³⁴

Abstract

Your Gene structure Annotation Tool for Eukaryotes (yrGATE) provides an Annotation Tool and Community Utilities for web-based community genome and gene annotation. Annotators can evaluate gene structure evidence derived from multiple sources to create gene structure annotations. Administrators regulate the acceptance of annotations into published gene sets. yrGATE is designed to facilitate rapid and accurate annotation of emerging genomes as well as to confirm, refine, or correct currently published annotations. yrGATE is highly portable and supports different standard input and output formats. The yrGATE software and usage cases are available at http://www.plantgdb.org/prj/yrGATE.

¹Reprinted with permission of Genome Biology 2006, 7, R58

²Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA 50011-3260, USA

³Department of Statistics, Iowa State University, Ames, IA 50011-3260, USA

⁴Author for correspondence
Rationale

Complete and accurate gene structure annotation is a prerequisite for the success of many types of genomic projects. For example, gene expression studies based on gene probes would be misleading unless the gene probes uniquely labelled distinct genes. Identification of potential transcription signals relies on correct determination of transcriptional start and termination sites. Characterization of orthologs or paralogs and other studies of molecular phylogeny are also compromised by incomplete or inaccurate gene structure annotation.

Gene structure determination is particularly difficult for eukaryotic genomes. Here, we focus on protein-coding genes. In higher eukaryotes, most of these genes contain introns, and a large fraction of the genes appear to permit alternative splicing [1-3]. High-throughput computational gene structure annotation has been highly successful in providing a first glimpse of the gene content of a genome, but current methods fall short of the goal of complete and accurate gene structure annotation (e.g., [4-6]). Recent research has focused on improving prediction sensitivity and specificity by combining multiple sources of evidence [7-9]. However, complexities of transcription and pre-mRNA processing such as introns in non-coding regions, non-canonical splice sites, and utilization of alternative splice sites still pose formidable challenges for merely computational methods. Re-annotation efforts for most eukaryotic model genomes have therefore relied in large part on manual inspection of gene structure evidence [5, 10, 11]. However, manual annotation also has shortcomings, such as being typically time-consuming, having exclusive participation, and providing annotations only intermittently [4, 10, 12].
A policy of 'open annotation', using the internet as the forum for annotation, and bringing annotation into the mainstream has been suggested as a means to eliminate the restraints of manual annotation and to develop high quality gene annotation [13-15]. Several systems have successfully adopted this policy for prokaryote gene annotation (ASAP [16], PeerGAD [17], PseudoCAP [18]). Eukaryotic gene annotation projects have not been able to reap the full benefits of community manual annotation because of the absence of an open online community gene annotation system. Here, we describe newly developed software, yrGATE, which seeks to compensate for the inadequacies of traditional manual annotation and to provide a community alternative and/or companion to computational gene annotation, specialized for eukaryotes. yrGATE provides similar functionality as the Apollo annotation tool [19] and NCBI's ModelMaker [20], but includes community utilities, specialized portals to external gene finding and annotation software, and web browser accessibility.

The yrGATE package consists of a web-based Annotation Tool for gene structure annotation creation and Community Utilities for regulating the acceptance of the annotations into a community gene set. The yrGATE Annotation Tool can be used without the Community Utilities for analysis of gene loci independent of a community. The Annotation Tool presents pre-calculated exon evidence in several summaries with different selection mechanisms and provides other methods for specifying custom exons, allowing thorough analysis and quick annotation of loci. Annotators access the tool over the web, where they create an annotation, decide to save the annotation in their personal account, or submit the annotation for review for acceptance into the community gene set. The online nature of yrGATE permits a large and nonexclusive group of annotators, ranging in expertise from
professional curators to students [21]. This also provides a continuous timeframe for gene annotation, allowing annotators to examine new sequence evidence as it becomes available and eliminating the delays of periodic annotation. yrGATE is particularly well suited for emerging genomes that are in the process of being sequenced, such as maize. Additionally, the user-friendly character of the yrGATE system contributes to its accessibility and to its potential for community adoption.

**Annotation Tool**

The Annotation Tool of the yrGATE package is a web-based utility for creating gene structure annotations. The inputs and outputs of the Annotation Tool are depicted in Figure 1. The input consists of a genomic sequence, exon evidence, and evidence references. The output of the Annotation Tool is a gene annotation, which consists of a gene structure (coordinates of exons and introns), the inferred mRNA sequence, a corresponding protein coding region and its associated translation product, evidence attributes, description, and functional information. The input and output can be in several formats (indicated in Figure 1), which will be described in detail in the Implementation Section.

Defining a gene’s exon-intron structure is the central step in creating a eukaryotic gene annotation. The Annotation Tool provides two general categories to specify exons: pre-defined evidence-supported exons and novel user-defined exons. Pre-defined exons are provided by the Annotation Tool from prior computations and are supported by evidence derived from spliced alignments of ESTs and cDNAs, *ab initio* predictions, or a combination
of sources. The evidence is filtered by stringent thresholds to provide exons suggestive of authentic genes. User-defined exons are exons not contained in the pre-defined evidence and are individually specified by the user. Annotators have several channels to designate both categories of exons.

The Annotation Tool contains three representations of the evidence: the Evidence Plot, the Evidence Table, and links to evidence reference files. The Evidence Plot is a clickable graphic that presents evidence in a colour-coded schematic (Fig. 2.A.8). The Evidence Table (Fig. 2.A.11) groups exons into mutually exclusive groups of exon variants. For each exon, the table lists its genomic coordinates, the maximum score from the method that generated the exon, and the evidence sources that support the exon. The evidence identifiers are hyperlinked to reference files for the exon, which could be an alignment or other program output. Annotators can select pre-defined exons by clicking on exon diagrams in the Evidence Plot or clicking on buttons in the Evidence Table. The annotator’s developing gene structure is graphically displayed below the Evidence Plot for visual comparison (Fig. 2.A.10).

User-defined exons are specified through portals to exon-generating programs or through entry of the genomic coordinates of an exon. As these exons are defined, they are listed in the User Defined Exons Table (Fig. 2.A.2). Acting as a type of web service, portals deliver the genome sequence of the annotation region to an online exon-generating program, with appropriate default parameters specified while allowing the user to change these parameters. The program’s output is internally reformatted such that the user can directly add exons from
the program's output window into the current gene structure displayed in the yrGATE Annotation Tool window. Currently, portals are available to the gene prediction programs GENSCAN [22] and GeneMark [23] and to the GeneSeqer spliced alignment web server [24]. Administrators can easily add new portals for other exon-generating programs or sequence analysis programs, such as folding programs for non-coding RNA annotations. A template portal is provided with the package.

As an additional channel provided for designating gene structures, the tool allows pasting a coordinate structure into the mRNA structure field (Fig. 2.A.6). The format for specifying an mRNA structure follows the conventional notation of designating exons by start and end coordinates separated by non-digits, with multiple exons separated by commas (e.g., the Perl regular expression for a two-exon gene structure is \[d+\D+d+,d+\D+d+\]). This channel is appropriate for comparing external gene structures with the evidence. Exons not found in the pre-defined evidence are given an 'unknown' source in the User Defined Exons table.

To document the annotator's procedure and parameters, the Exon Origins attribute of an annotation record automatically stores information about the source of each exon. The following information is stored: the method of exon-generation, a score associated with the method and exon, sequence identifiers used in the method, unique database identifiers to the specific output file or record, and a hyperlink to the program output yielding the exon. Exon Origins allows for complete re-creation of the gene structure annotation and for analysis of manual annotation procedures that could aid in future manual annotation efforts and techniques.
After a gene structure has been defined, a user can specify the protein coding region of the annotation through entry of genomic coordinates (Fig 2.A.4) or by using the ORF Finder [20] portal. The ORF Finder portal (Fig 2.B), operating similar to the User Defined Exons portals, allows a user to select an open reading frame, which upon selection is imported into the Annotation Tool window and is graphically represented in the Preview Structure.

Coordinately with gene structure and protein coding region designation and edits, the mRNA and protein sequence fields are updated (Fig. 2.A.3 and 2.A.5). Hyperlinks, attached to the appropriate sequence, are provided to BLASTN, TBLASTX, BLASTX, TBLASTN and BLASTP at NCBI [20] for an annotator to find similar sequences and/or assign a putative function. Additional pieces of information that can be added to a gene annotation are a description and alternative identifiers.

For cases in which genomic sequence requires editing, such as correction of sequencing errors or annotation of genes undergoing mRNA editing, the Sequence Editor Tool (Fig 2.A.7) enables annotators to insert, delete, or change bases through a web interface. These changes are incorporated into the Annotation Tool and stored with the annotation record.

At the conclusion of a gene annotation session, an annotator decides the outcome of their annotation record (Fig 2.A.1). Annotation records can be saved in the annotator’s personal account, which limits access of the annotation to the owner of the annotation. Annotations can be submitted for review, in which case the annotation is sent to Administrators, who
decide to accept or reject the annotation into a community database for sharing with the community. Alternatively, annotations can be saved locally on the annotator's machine by displaying the annotation in a simple text or GFF3 [25] format. Annotators are also able to delete stored annotations, which have not been accepted.

Community Annotation Utilities

The yrGATE package includes Community Annotation Utilities for sharing annotations among a public or private community. These utilities form a process for annotation management and review (diagrammed in Figure 3) for two different types of users, annotators and administrators. The types of users are distinguished by their actions: annotators create annotations and administrators review these annotations for acceptance into a community gene set. The Community Annotation Process will be described from the perspective of a new annotation submission and review.

A typical annotation submission begins with an annotator logging in to their private account, which contains all of the annotations created by the annotator. Then, the annotator creates a new annotation using the Annotation Tool and decides to submit the annotation to the community.

This newly submitted annotation is listed in the Administration Tool, where an Administrator can 'check out' this annotation for review, so that other administrators do not review this annotation concurrently. The administrator accesses the 'checked-out' annotation in a review
version of the Annotation Tool. Then, the administrator reviews the annotation and is able to edit any attributes of the record. When satisfied with their analysis, the administrator accepts or rejects the annotation. If a decision cannot be reached, the annotation is returned to the to-be-reviewed group. Accepted annotations are added to the public community gene annotation database, where they are presented through the Community Annotation Central and Annotation Record facilities. Rejected annotations can be edited by the annotator to be resubmitted for review.

For specific implementations, the described Community Annotation Process can be adjusted by dropping any of the steps, such as eliminating the user log in or eliminating the review process so that all submitted annotation are published. New steps can also be added to the review process, such as a voting utility for submitted annotations.

**Implementations and case studies**

The yrGATE package can be implemented in different configurations depending on the input and output (Fig.1) and on the annotation review process (Fig 3). The input can be either from a local database or a DAS server. The output can be an entry in a local database or to a simple text or GFF3 file. The optional Community Utilities provide annotation review and community maintenance facilities. Two yrGATE implementations, having different configurations, are described below.
Community Annotation at PlantGDB

PlantGDB includes a family of species-specific databases: AtGDB [26, 27] for Arabidopsis, ZmGDB [28] for maize, and OsGDB [29] for rice. These species-specific databases each have an annotation community and an implementation of yrGATE. Input to the yrGATE annotation tool is supplied by the respective PlantGDB database. Pre-calculated exon evidence consists of spliced alignments of EST and cDNA sequences generated by the GeneSeqer program [30]. Evidence references consist of hyperlinks to GeneSeqer output files, which are a part of the respective databases. Genome sequence segments are also supplied by the database. In these PlantGDB implementations, yrGATE Community Utilities regulate user management and annotation curation according to the described default configuration (Fig. 3). We illustrate yrGATE use at PlantGDB with two gene annotation case studies.

The first case study is a novel maize annotation using the ZmGDB yrGATE implementation. An unannotated genome region, 158659-162032 of BAC 51315585, was chosen by the annotator using the genome browsing function of ZmGDB. A screenshot of the Annotation Tool shows the completed annotation (Fig. 2). Exons were initially selected from the pre-computed evidence. The evidence, though, consists of two separate groups of ESTs (Fig 2.A.9) with no spanning evidence in the region 160260-160664. The annotator decided to use the GENSCAN and the GeneSeqer@PlantGDB portals to explore potential exons in this region (Fig 2.A.2). After adding three User Defined Exons, a gene structure connecting both groups of ESTs was defined (Fig 2.A.6 and 2.A.10). The portal to the ORF Finder was used
to define a protein-coding region, which spanned all eight exons of the putative transcript. Terminal exons, supported by ESTs 71435182 and 32859895, were selected to maximize the untranslated regions. The final step of the annotation session was a BLASTP search at NCBI to compare the novel gene annotation and to assign a putative gene product function. The protein of the annotation had high similarity over most of its length to rice protein NP_915525 and to Arabidopsis protein NP_190282. These proteins provided a putative functional assignment of 'sugar transporter' for the annotation. The annotator was satisfied with the annotation and submitted it for review. Administrators reviewed the annotation and accepted it because it was novel and of good quality. The Annotation, ZM-yrGATE-sugar_transporter, is now accessible from the ZmGDB Community Annotation Central [31].

The second PlantGDB case study concerns alternative splicing and correction of an inaccurate published annotation of an Arabidopsis gene model using the yrGATE implementation at AtGDB. A screenshot of the Transcript View of AtGDB presents two accepted community annotations (green structures in interior window, Figure 4). The annotator decided to investigate this genome region (chromosome 1, segment 30370180-30373939) because, upon visual inspection, the first exon of the published annotation At1g808010.1 conflicts with EST and cDNA evidence (Figure 4.A.3). Initially, the annotator used cDNA 23270370 to define the gene structure and EST 496433 to extend the 3'-untranslated region. Through the Evidence Table and evidence reference links to GeneSeqer output of the Annotation Tool, the annotator recognized exon 11 has an alternate size supported by EST 507078. The annotator examined open reading frames of both transcript structures, and seeing that both protein-coding regions extend over all exons except
for the 5'-most untranslated exon, decided to create two annotations for this locus. An AtGDB administrator reviewed the annotations and accepted both into the community database because they corrected an inaccurate published annotation and captured alternative splicing variants. These alternative splicing variants are displayed in the Transcript View of AtGDB (Fig 4.1), which displays sequence alignments coordinated to a diagram. In the Transcript View, the green vertical rectangle (Fig 4.2) relates the diagram to the multiple sequence alignment, where nucleotides in introns are represented by '>' symbols.

Comparing alignments for sequences 23270370 and 507078, a three base difference in the start of the exon 11 is apparent (Fig 4.4). The upstream intron sequences reveal that both intron variants terminate with the standard AG dinucleotide, which suggests this is a probable alternative splicing event. The Transcript View of AtGDB makes such minute differences distinguishable, which were previously concealed in the diagram.

**yrGATE with DAS input**

DAS servers provide sequence and annotation information that is queryable and in a standard format [32, 33]. The abundance of DAS servers for a variety of organisms provides rich and diverse sources of input for the yrGATE Annotation Tool. An implementation of yrGATE using input data from DAS servers is provided for general use [34]. This implementation, 'yrGATE with DAS input', does not have a community aspect, although a different configuration could add community functionality. The 'yrGATE with DAS input' Selection Page allows an annotator to specify a DAS reference server and DAS evidence sources (Fig. 5.A). The green ‘look up’ buttons beside each text box provide a list for annotators to make
selections. After these selections are stored, the Annotation Tool can be accessed with the
selected input DAS data (Fig. 5.B).

Figure 5 represents a case study of a novel chicken gene structure annotation. The Selection
Page specifies the chicken genome chromosome 3 segment 86850000 – 86990000 as the
genome entry point [35, 36]. The selected evidence sources include primary evidence of
mRNA and EST BLAT alignments and, for comparison, annotations of types RefSeq [37,
38], TWINSCAN [39], Ensembl [40], Geneid [41], and SGP [42]. The published annotation
evidence sources are selected so that the annotator can compare primary evidence against
existing annotations. Inspection of the primary evidence in the Evidence Plot of the
Annotation Tool suggests one gene on the forward strand (approximately 86887000 –
86934000, Fig. 5.B.1) and another gene on the reverse strand (approximately 86853000 –
86975000, Fig. 5.B.2). The gene on the forward strand (Fig 5.B.1; e.g., RefSeq Gene
angiopoietin-2, dark blue, labelled NM_204817.1) is accurately annotated based on mRNA
and EST evidence. Additional alternative variants are also accurately annotated.

The primary evidence also suggests an annotation on the reverse strand that contains the
angiopoietin-2 gene within one of its introns. However, current annotations on the reverse
strand are inaccurate and incomplete based on mRNA and EST evidence (Fig 5.B.3). The
first half of this potential gene is represented in some annotations (Fig 5.B.2; SGP:
chr3_982.1, Geneid: chr3_1361.1, Ensembl: ENSGALT00000026345.2; TWINSCAN:
chr3.87.019.a). Alignments of other species' RefSeq Genes [43] (not pictured) indicate a
larger gene boundary than the displayed annotations, but this boundary is still too short
compared to the primary evidence and does not contain all of the exons supplied by the primary evidence. A novel gene annotation was created on the reverse strand by selecting compatible exons from primary evidence using the Annotation Tool. An open reading frame was designated, and the protein sequence was used to find homologous genes in related species. Based on BLASTP results, this gene was assigned the putative function microcephalin. Interestingly, several species (including human and mouse) have an annotated microcephalin gene with high protein sequence similarity and also maintain the local genome structure of angiopoietin-2 within an intron of the microcephalin gene on the opposite strand.

Links to these case study annotations are provided on the yrGATE website [44].

Usability and availability

The Annotation Tool was designed with emphasis on usability for annotators. Annotators can immediately select from high quality evidence that has a high likelihood of yielding an accurate annotation and can specify new custom evidence for cases where the evidence is inadequate. The two categories provide for a good annotation process where high quality evidence is first examined and then additional evidence is checked, which is completed in a minimal amount of mouse clicks and screen display, achieved by the Tool’s design.

The main components of the tool are contained in one standard 1024x768 resolution screen. The tool is loaded once per genomic region, and the form fields are dynamically updated,
which allows annotators to quickly evaluate the impact of different exon variants and combinations of exons on the gene structure, mRNA sequence, and protein sequence.

yrGATE is compatible with several major operating systems, including Linux, Windows and Macintosh, on several web browsers, of which Mozilla Firefox has the best performance in terms of speed.

yrGATE is available for download [44]. The package consists of Perl, Javascript, HTML, and a MySQL schema. Required Perl libraries for a full implementation are CGI, DBI, LWP, HTTP, PHP::Session, GD, Bio::Graphics, Bio::SeqFeature::Generic, and Bio::Das. Template data is provided for testing and evaluation.

Conclusions

yrGATE opens gene structure annotation to a large, nonexclusive community. The characteristics of yrGATE contribute to its potential for user appeal and community adoption. Among other applications, it is particularly useful for annotating emerging genomes and for correcting inaccurate published annotations. yrGATE is easily adaptable to different input data and can support a community using the Community Utilities.

List of abbreviations

yrGATE: your Gene-structure Annotation Tool for Eukaryotes
Acknowledgements

This work was supported by the National Science Foundation Plant Genome Research Projects grant DBI-0321600 to V.B. Matthew Wilkerson worked in part under a cooperative agreement with University of Missouri, SCA #58 3622-3-152.

References


21. **Annotation for Amateurs** [http://www.plantgdb.org/tutorial/annotatemodule]


27. **An Arabidopsis thaliana Plant Genome Database** [http://www.plantgdb.org/AtGDB]

28. **A Zea mays Plant Genome Database** [http://www.plantgdb.org/ZmGDB]

29. **An Oryza Sativa Genome Database** [http://www.plantgdb.org/OsGDB]


31. **yrGATE @ ZmGDB: Community Annotation Central** [http://www.plantgdb.org/ZmGDB_yrgate-cgi/CommunityCentral.pl]


33. **The distributed annotation system** [http://www.biodas.org]

34. **yrGATE with DAS input** [http://www.plantgdb.org/DAS_yrgate]

36. **The UCSC Genome Database** [http://genome.cse.ucsc.edu/]


38. **UCSC Genome Browser RefSeq Genes Track** [http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=73356660&c=chr3&g=refGene]


43. **UCSC Genome Browser Non-Chicken RefSeq Genes Track** [http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=galGal2&g=xenoRefGene]

44. **Your Gene structure Annotation Tool for Eukaryotes** [http://www.plantgdb.org/prj/yrGATE]
Figures

Figure 1. The applications interface of yrGATE.
Input to yrGATE is derived from either local database tables or distributed DAS sources.
Output is either to local database tables or in the form of simple text or GFF3 files.
Figure 2.
Figure 2. Novel Gene Annotation.
This yrGATE implementation at ZmGDB presents the region 158659-162032 of Zea mays BAC gi 51315585. The main Annotation Tool window (A) contains a completed gene structure annotation. The provided transcript evidence consists of two groups of ESTs (circled as A.9) separated by a region with no spanning evidence, 160260 – 160664 (A.8). User Defined Exons have been designated in this region. The User Defined Exons Table (A.2) lists each exon by coordinates and source. Exon 5, 160575-160721, was defined using portals to GENSCAN (B) and GeneSeqer@PlantGDB (not shown). Yellow buttons in the GENSCAN portal (B) add exons to the gene structure in the Annotation Tool (A.6), which are presented pictorially (A.10) for comparison with the Evidence Plot. A protein-coding region was evaluated using the portal to the ORF Finder (C) and imported into the Annotation Tool (A.4) using the yellow button.
Figure 3. Community annotation review process. Individual Community Utilities are coloured green in this diagram.
Figure 4. Community implementation of yrGATE at the PlantGDB Arabidopsis genome browser, AtGDB, for correction of a public annotation and for alternative splicing. This two-window screenshot depicts yrGATE annotations in the AtGDB browser. The outer window contains a genome context view of AtGDB, which has links to the yrGATE Annotation Tool and to AtGDB’s Transcript View (A.1). The inner window contains the Transcript View, which presents a genome context graphic and sequence alignments represented in the graphic. The graphic has the following color assignments: yrGATE annotations - green, the public annotation - blue, cDNAs - light blue, ESTs – red, annotation protein coding regions – green and red triangles. The multiple sequence alignment in the lower panel of the Transcript View corresponds to the region of graphic contained within the green rectangle (A.2). The first exon (A.3) of the public annotation, At1g80810.1, is not supported by expressed sequence evidence, which instead suggests a downstream exon. There are two yrGATE community annotations, yrGATE-At1g80810-1 and yrGATE-At1g80810-2, both of which contain the first exon supported by the evidence but differ at the 3’-end, because the evidence suggests two alternatives for exon 11 (as seen in the multiple alignment display, A.4).
Figure 5.
Figure 5. yrGATE with DAS input Implementation.
The entrance to yrGATE is a Selection Page where a genome and associated evidence sources are specified (A). Chicken chromosome 3 region 86850000-86990000 is selected. EST and mRNA are primary evidence sources (B.3). Additionally, secondary evidence sources of published annotations are selected for comparison including RefSeq, Ensembl, Twinscan, SGP, and geneid genes. The novel annotation, GG-yrGATE-microcephalin, is based on EST and mRNA evidence and is distinct from all published chicken annotations in this region on this strand (B.2). This novel annotation (B.4) contains a known angiopoietin gene, NM_204817 (B.1), on the opposite strand within its 12th intron.
CHAPTER 4: COMPUTATIONAL ANALYSIS OF U12-TYPE INTRONS WITHIN ORTHOLOGOUS GENES

A paper prepared for submission to *PLoS Computational Biology*

Matthew D. Wilkerson¹, and Volker Brendel¹²³

Abstract

The CIWOG (Common Introns within Orthologous Genes, http://ciwog.gdcb.iastate.edu) database was constructed for the purpose of analyzing U12-type introns in the context of orthologous genes. A large number of U12-type introns were identified in human, mouse, zebrafish, chicken, Drosophila, Arabidopsis, and rice. Novel gene structure annotations were constructed for U12-type introns supported by transcript evidence but lacking an accurate host gene structure annotation relative to published annotation sources. Intron sites were derived from protein sequence multiple alignments and compared by their intron type. The results from this study provide confirmatory support for ancestral genes rich with U12-type introns and for the process of subtype switching. This study provides the first instances of intron type conversion between vertebrate, between plant, and between plant and animal orthologous genes. Between plants and animals, U12-type intron conservation, conversion and loss were detected in an approximate 1:2:4 ratio. Orthologous gene clusters with

¹Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA 50011-3260, USA

²Department of Statistics, Iowa State University, Ames, IA 50011-3260, USA

³Author for correspondence
conserved U12-type introns are shown to experience reduced gene structure, global protein sequence, and local protein sequence evolution. The phase distribution of U12-type introns was found to be strictly conserved in orthologous introns, suggesting that the modern phase distribution was present in the last common ancestor of at least plants and animals. Also, the phase distribution was found to correlate with exonic nucleotide sequence conservation.

Introduction

Among the population of introns in eukaryotic genomes, the minor class of introns, U12-type, has unique characteristics compared to the major class of introns, U2-type. U12-type introns have highly conserved donor site and branch site consensus sequences, 

\[/[GA]TATCCTT \text{ and } TCCTTAAC/\]

respectively, where the exon-intron boundary is indicated by / and [GA] indicates G or A [1, 2]. U12-type introns with a /AT donor site typically have an AC/ acceptor site and those with a /GT donor site typically have an AG/ acceptor site, which are called the AT-AC and GT-AG subtypes, although other terminal dinucleotide combinations have been found [2-6]. In contrast, U2-type introns have less conserved sequences at the donor and branch sites, typically have /GT and AG/ terminal dinucleotides, and have an additional sequence element, the polypyrimidine tract [2, 7]. Comprising less than 1% of the introns of any organism studied [2, 6], U12-type introns have a dramatic difference in abundance compared to U2-type. U12-type introns are spliced from pre-mRNA by a spliceosome that is structurally and mechanistically similar to the U2-type spliceosome (reviewed in [8]). Another distinctive feature of U12-type introns is that their preferred phases are 1 and 2, which interrupt codons, while the preferred phase of U2-type
introns is phase 0, which does not interrupt a codon [2]. Despite the distinctions between the intron classes, host genes of U12-type introns typically have U2-type introns as well [2].

The initial large scale study of U12-type introns [2] established a computational method for identifying U12-type introns and reported intriguing results regarding comparisons of orthologous and paralogous genes. After identifying 60 non-redundant U12-type introns parsed from GenBank gene annotations, the authors compared intron sites within protein sequence alignments of homologous genes. Conservation of U12-type introns was found between organisms as diverse as mammals and jellyfish. Some U12-type introns with /AT-AC/ terminal dinucleotides corresponded with U12-type introns with /GT-AG/ terminal dinucleotides, suggesting a subtype-switch event. Some U12-type introns corresponded with the position of U2-type introns in related species, suggesting a conversion event between intron classes. Some U12-type introns did not have a corresponding intron in related genes, suggesting an intron loss event. Through comparing paralogous genes, the authors found U12-type introns in distinct positions along the protein sequences of the genes, suggesting that an ancestor had multiple U12-type introns. Additionally, 4 out of 56 U12-type intron host genes were found to possess at least 2 U12-type introns each. Combining these observations, the authors proposed a 'fission/fusion' theory for the evolution of the splicing systems.

Under this theory, the splicing systems have a common ancestor and a speciation event allowed introns and spliceosomes to diverge into two distinct classes. After a genome merging event, possibly endosymbiosis, both classes inhabited one organism, after which the U12-type introns have mostly been lost or converted to U2-type, through the proposed pathway model of U12-type AT-AC to U12-type GT-AG to U2-type to loss. This pathway
model is based on the assumption that U12-type introns, having highly conserved donor and branch sites, have a greater likelihood to mutate to the less conserved sites of U2-type introns, than the reverse process. This theory accounts for the elevated concentration of U12-type introns in some genes, since they are descendants of genes with only U12-type introns. However, this theory does not provide a reason that all surveyed eukaryotic organisms, with the exception of nematode and yeast [2, 6, 9], continue to maintain a small number of U12-type introns at the cost of maintaining a parallel splicing system.

There are several possible explanations for the occupancy of the U12-type splicing system in modern eukaryotes alongside the U2-type system. One explanation is that U12-type introns have no distinctive function from U2-type introns, are remnants of a formerly popular splicing system, and are in the process being depleted from genomes through conversion and loss [2]. A second explanation is that U12-type introns do have a distinctive function, which is critical to its host genes. For a few genes, distinctive biological functions have been attributed to possession of a U12-type intron (rate-limiting in gene expression in human E2F2, SmE, and INSIG1 [10], alternatively spliced transcripts developmentally-specific in Drosophila prospero [11] and tissue-specific in rat calcitonin/CGRP [12]). It remains to be seen whether similar functions extend to all U12-type introns or new functions are awaiting discovery. A third explanation is that U12-type introns do not offer a distinctive function, but are contained in genes that experience selection against any sequence change, which relegates U12-type introns to passengers on the gene, and the U12-type splicing system persists to splice these remaining introns.

One complication in studying U12-type introns is that they can be mis-annotated in their host gene structure annotations, because most gene prediction programs are not specifically
designed to recognize them [13]. In addition to the regular challenges of gene prediction, U12-type introns present a unique challenge due to their difference in donor and acceptor site sequences from the U2-type introns which comprise the majority of all introns. Prediction programs may alter a U12-type intron’s boundaries to provide a U2-type intron, may predict the U12-type intron to be exonic, or may miss the U12-type intron altogether. Transcript alignments are an excellent source of evidence for gene structure annotations and introns, since they are a derivative of the gene expression process. In particular for U12-type introns, transcript alignments can be a source of introns that were missed by gene prediction programs and public gene annotation sets.

Since the initial large scale study [2], several genome-wide studies have identified U12-type introns in plant [5, 6, 9] and animal genomes [6, 7, 9, 14], and the list of organisms possessing U12-type introns has just recently grown to include fungi and protists [15].

Recent studies have increased the number of known U12-U2 intron conversions, the number of genes with multiple U12-type introns, [6] and the number of orthologous U12-type introns [9]. The multiple organism studies [2, 6, 7, 9] have used published gene structure annotations as the initial source of introns, but this strategy overlooks potential U12-type intron mis-annotation in genes. Other discoveries include that U12-type introns do not have a positional bias relative to other introns within genes [8], have overall similar length distributions to U2-type introns [5, 14], and tend to be involved in ion transport, protein trafficking, and cell cycle control [6].

Seeking a better understanding of U12-type introns’ persistence in modern genomes, we decided to analyze U12-type introns across 7 organisms in the context of their host genes and of their host gene orthologs. In order to pursue this goal, a new unique resource, the CIWOG
(Common Introns Within Orthologous Genes) database, needed to be developed. We began by identifying U12-type introns in native transcript alignments and gene structure annotations. For those contained only in transcript alignments, we created novel gene structure annotations to accurately capture the U12-type intron. We then created orthologous gene clusters and mapped introns onto protein sequence alignments of each cluster. An algorithm was devised and used to define intron sites within alignments and classify all genes at each intron site.

This study provides confirmatory support for earlier results and made some new discoveries about U12-type introns. We report for the first time U12-type intron sites that support conversion and loss events in orthologous gene clusters among all of the organisms in this study. We detected many orthologous gene clusters with multiple U12-type intron sites, which provide support for ancestral genes rich with U12-type introns. We assembled a large collection of non-canonical (not AT-AC or GT-AG) U12-type introns, including some new varieties such as CT-AC, confirming that the terminal dinucleotides do not define the intron type [4, 16]. We describe how these non-canonicals support the phenomena of subtype switching. We show that orthologous gene clusters with conserved U12-type introns tend to experience slow gene structure, global protein sequence and local protein evolution.

Lastly, we pursued the unique phase preferences of U12-type introns. We find similar phase distributions among conserved U12-type introns between clades, and do not find a single case of a conserved U12-type with multiple phases, suggesting that the present phase distribution of U12-type introns existed in these organisms' last common ancestor. We discovered a correlation between the phase frequencies of U12-type introns and exonic
nucleotide sequence conservation and analyzed this correlation in terms of splicing constraints and protein-coding constraints.

Methods

The central component of this study is a database, CIWOG (Common Introns Within Orthologous Genes), which contains non-redundant introns, gene annotations, orthologous genes, and common introns. The procedure of the construction of CIWOG is illustrated in the flow chart (Figure 1) and the sections below describe the details of each step. In summary, this began with downloading genomic sequence, gene structure annotations, and transcript alignments from publicly available data sources. After loading this data into CIWOG, non-redundant introns, alternative gene groups, and alternative introns were derived. Then, introns were subjected to a classification procedure to determine a predicted intron type. Gene structure annotations were created for U12-type introns that were contained in transcripts but were not accurately contained in a splice junction of any gene annotation. Orthologous gene clusters were computed and multiple alignments were created for each cluster. Using the multiple alignments, intron sites were identified. The following terms are used in this paper:

- OGC: orthologous gene cluster
- uOGC: U12 orthologous gene clusters (OGC containing at least one U12-type intron)
- nOGC: Non-U12 orthologous gene clusters (OGC containing no U12-type introns)
• intron site: distinct intron locations in an OGC
• cintron: common intron between genes, corresponding to one intron site
• absent-intron: classification for a gene not possessing an intron at a particular intron site, but having sufficient protein sequence identity in the region around the intron site.

Unless otherwise specified, the entirety of this study was carried out using ad hoc Perl [17] scripts to perform calculations and interact with CIWOG, which is a MySQL [18] database. The Perl libraries CGI, DBI, and GD and Javascript were used to provide dynamic interfaces to the CIWOG database. R [19] was used for statistical analysis and figure construction.

**Import of external data: genome sequence, transcript alignments, gene structure annotations.** For this study, organisms were required to have a mature genome assembly, mature gene annotation sets, a high number of transcript sequences, and published possession of the U12-type spliceosome. The following organisms were selected: Arabidopsis thaliana, Danio rerio, Drosophila melanogaster, Gallus gallus, Homo sapiens, Mus musculus, and Oryza sativa. Three types of data were downloaded from the following databases: PlantGDB [20], UCSC Genome Browser [21], Ensembl [22], TIGR [23] and TAIR [24]. Genome sequence is chromosome or contig nucleotide sequence. Transcript alignments are spliced alignments of an organism's EST, full-length cDNA, and mRNA sequences having optimal similarity and coverage within the organism's genome, as defined by the respective genome database provider [20, 21]. Gene structure annotations provide the locations of exons, introns, and a protein coding region. In order to attain a complete as possible catalog of an organism's genes, multiple gene structure annotation sets for each
organism were used. A listing of the data sources, types, versions, and dates of download is presented in supplementary table 1.

**Processing of genes, transcripts, and introns.** For this study, an intron is defined as a start and end genomic coordinate pair on a given genomic strand. Introns were parsed from transcript alignments and gene structure annotations and were reduced to a non-redundant set for each organism. Introns overlapping the same genomic sequence and strand were placed into Alternative Intron Groups. For gene structure annotations, transcript and protein sequences were derived from genome sequence using exon structure and protein coding region information. In order to capture alternatively spliced genes and redundancy from the multiple annotation sets per organism, gene structure annotations were placed into alternative gene groups. Pairs of genes were evaluated for the following conditions: they must be located on the same genomic strand, and the genes' genomic position boundaries must overlap at least 20% of the larger gene's boundaries or they must share an intron. Pairs of genes meeting these conditions were grouped through single linkage clustering.

**Intron classification.** A position weight matrix (PWM) classification procedure based on previous studies [2, 5] was used to identify U12-type introns. To construct the PWMs, a set of reference introns were collected for each intron type. The U12-type intron reference set was transcript-supported introns with /ATATCC...AC/ terminal sequences with a length \( \geq 58 \) nt. The U2-type reference set consisted of 10,000 randomly selected introns from gene annotations from each organism that had /GT-AG/ terminal sequences and that lack the U12-type donor site consensus sequence. Using the reference introns, PWMs were created for the donor site (DS) and branch site (BS) sequences of U12-type and U2-type introns. DS sequences, defined as the sequence from positions from -3 to +9 relative to the start of the
intron, were extracted from each reference intron. Branch sites for the U12-type reference introns were identified as the 8-mer in the region -50 to -3 relative to the end of the intron with the highest score using a previously defined U12-type BS PWM [5] that represents U12-type BS consensus TCCTTAAC. A random 8-mer sequence in this region was extracted from each U2-type reference intron to measure the background sequence composition and substitute for the U2-type BS. Then using the reference sequences, the frequency of each nucleotide at each position plus a pseudocount is divided by the total number of reference sequences to provide a probability for the entries in the PWM. The pseudocount is defined as ¼ for each nucleotide and 1 for the total number of reference sequences, which provides a sequence of random composition [2]. Then, each intron was given a DS and BS log score, \( p_{DS} \) and \( p_{BS} \), by taking the sum of the log of the corresponding entries in the PWM. U12 log-odds ratios, defined as \( \log_2(p_{DS}^{U12}/p_{DS}^{U2}) \) and \( \log_2(p_{BS}^{U12}/p_{BS}^{U2}) \), were computed and normalized by subtracting their mean and dividing by their standard deviation. These normalized values are referred to as the Donor Score, \( d_s \), and Branch Score, \( b_s \). Following a previous study [5], minimum \( d_s^{U12} \) and \( b_s^{U12} \) values were defined, 4 and 1.5 respectively, as lower boundaries for U12-type introns. Introns with scores below these minima were classified as U2-type. Also, introns were required meet additional criteria before they were given a score. Introns with length \( \leq 20 \) were not given scores, because the DS and BS regions will overlap. Introns must have one of the following conditions: have /GT...AG/, /AT...AC/, or /GC...AG/ termini, be supported by a gene structure annotation, or have a minimum number of supporting transcripts (3 for plants, 10 for animals. The difference in these numbers is due to different numbers of available transcripts.). Introns not meeting these conditions are less likely to be authentic introns and were classified as type none to be
excluded from further analysis. Following a prior study [5], position 4 in the Donor Site PWMs, which corresponds to the first nucleotide of the intron, was given zero values for all nucleotides, so that introns having the U12-type consensus but not beginning with an A or G are not penalized.

Gene structure annotation of U12-type intron seeds using native transcript sequence. Transcript supported U12-type introns that were not contained in any of the published gene structure annotations (supplementary table 1) were identified and protein coding gene structure annotations were created for as many of these U12-type introns as possible. For plants, genomic regions were evaluated using the xGDB browsers [25] and annotations were constructed manually using the yrGATE Annotation Tool [26]. For animals, the spliced alignment program GenomeThreader [27] was executed using the genomic region surrounding the unannotated U12-type intron and native transcript sequences, which were known to align to this region based on UCSC Genome Browser [21] alignments, and using the options gcmaxgapwidth=50000, gcmincoverage=80, and introncutout. Putative protein coding genes from GenomeThreader output ("predicted protein sequences"), which had a predicted U12-type intron and a protein length of at least 50, were added to CIWOG.

Orthologous gene clusters. For each organism, the gene with the longest protein sequence, without an internal stop codon, from each alternative gene group was designated as the cluster seed. Sequence matches between cluster seeds were calculated by BLASTP [28] in 'all versus all' manner, using a concatenated file of the cluster seeds as input and default parameters except for a minimum e-value of 1e-10. cluster seeds, rather than all protein sequences, were used to remove redundancy within alternative gene groups and to promote
the greatest number of sequence matches to potential orthologs. OGCs were calculated from the 'all versus all' BLASTP output by the program OrthoMCL [29] with default parameters. OrthoMCL creates OGCs on the basis of inter-species reciprocally best hits to identify orthologs and intra-species reciprocally better hits to identify in-paralogs [29].

Representative genes and multiple alignments. OGCs having at least one gene with a U12-type intron were identified (uOGCs). A representative gene was selected from each alternative gene group in each uOGC. The following ordered list of criteria was evaluated to select a representative gene: number of canonical U12-type introns in the protein coding region, number of non-canonical U12-type introns in the protein coding region with supporting alignments, number of canonical U12-type introns in the untranslated region, number of non-canonical U12-type introns in the untranslated region with supporting alignments, number of all U12-type introns, length of protein sequence, number of all introns, size of gene boundaries, and gene source hierarchy. If more than one gene remains after evaluation of this list, a gene was randomly selected. The purpose of the gene source hierarchy criteria was to select genes from sources that typically have more transcript evidence or have greater associated information. The ranked list was (yrGATE and flyBaseGene), (knownGene, refGene, ensembl), and GenomeThreader. Sources in the parentheses have equal rank. For this selection process, canonical introns are those with /GT-AG/, /AT-AC/, or /GC-AG/ terminal dinucleotides, and non-canonical introns are those with other terminal dinucleotides. Supporting alignments refers to alignments containing an intron that has both flanking exons that match genomic sequence 100% and are at least 15 nucleotides in length. Then, protein sequences corresponding to the representative genes of uOGCs were aligned using the program MUSCLE with default parameters [30]. Separately,
the complement of the uOGCs, or non-U12 orthologous gene clusters (nOGCs), were processed using the cluster seeds as representative transcripts.

**Intron site derivation.** An algorithm was created to derive common introns of OGCs using multiple sequence alignments that is inspired by previous studies [31, 32]. First, a gene's protein coding region introns are given a position on the protein sequence, proteinPos, which is the position of amino acid in the protein sequence that is the previous complete codon before the intron. Then, the position of each intron in the alignment, alnPos, is assigned, using proteinPos to translate between the protein sequence and alignment. To enable grouping of introns at close but not identical alnPos, the parameter maxSlide and maxGap define boundary positions, alnPosRight and alnPosLeft, in the alignment that an intron can occupy. maxSlide is a number of amino acids and maxGap is a number of gaps. After defining these values for each intron, introns are grouped together. First, introns with identical alnPos are grouped together, and the group is assigned the minimum alnPosRight and maximum alnPosLeft of those introns. Then, the intron groups are sorted in increasing order of alnPosRight. Each group of introns is evaluated for overlap of alnPosRight and alnPosLeft. If they overlap, the introns are grouped together and the group is assigned the maximum alnPosLeft between the groups. An intron cannot be added to a group if its host gene already has an intron in that group. This process is repeated until all groups have been evaluated for overlap. The resultant groups are referred to as intron sites. The introns within these intron sites are collectively called common introns (cintrons). Then, genes not having an intron in a particular intron sites are evaluated by local protein sequence identity. The parameter flank is a number of amino acids to be compared to the left of alnPosLeft and to the right of alnPosRight. The parameter minSeqSim is the minimum percent of identical
amino acids in cintron flanking regions of size flank. If an intron-less gene has percent identity identities greater than or equal to minSeqSim in both flanking regions with a gene having an intron at intron site, then the cintron-less gene is added to the intron site with a type of 'absent-intron'. Otherwise, if a intron-less gene has aligned amino acids spanning the intron site's alnPosLeft and alnPosRight, the intron-less gene is added to the group with a type of 'internal missing'. Otherwise, the intron-less gene does not have any spanning amino acids and the intron site is outside the intron-less gene's alignment, so the intron-less gene is given a type 'external missing'. Internal missing and external missing are not used in further analysis.

The algorithm considers untranslated region (UTR) introns separately. UTR introns are given a position based on their order from their host gene's translation start or end. For example, a gene that has two 5' UTR introns and two 3' UTR introns, the equivalent proteinPos positions of the intron sites in the order of the introns in the gene is: utr5.2, utr5.1, utr3.1, utr3.2. UTR introns are grouped together solely based on identical proteinPos. This algorithm was executed on OGCs with the following parameter values: maxSlide 1, maxGap 1, flank 10, minSeqSim 0.3.

For the purpose of removing potential false positive cintrons from our dataset, an additional condition was placed on intron sites having a U12-type, and either a U2-type or absent-intron. In these sites, the non-U12-type introns were required to have the exact same alnPos and have minSeqSim of at least .50 with a U12-type intron. Additionally, we manually reviewed these vertebrate intron sites by comparison with the UCSC Genome Browser MULTIZ vertebrate whole genome alignments (http://genome.ucsc.edu/cgi-bin/hgTrackUi?g=multiz28way), which determines orthology using synteny. If there was a
discrepancy in CIWOG introns relative the MULTIZ alignment, the CIWOG introns were
adjusted to have a splice type of 'none' for that particular intron.

Gene structure annotation of orthologs of U12-type intron host genes using
orthologous protein sequence and native transcript sequence. A lack of accurate
transcript alignments and gene structure annotations will prevent identification of U12-type
introns. In an attempt to identify and annotate additional U12-type introns, we used
orthology as a means to identify U12-type intron host genes. The genomic regions
corresponding to all orthologs of U12-type intron host genes were evaluated by native
transcript and orthologous protein sequence GenomeThreader alignment, with options
gcmaxgapwidth=50000, gcmincoverage=10, prseedlength=6, and minmatchlength=6. The
alignments and associated predicted genes, containing a U12-type intron and having a protein
length of at least 50, were added to CIWOG. After this annotation, the representative gene,
alignment and cintron derivation steps were repeated.

Alignment support for non-canonical U12-type introns. In order to remove potential
false positive U12-type intron resulting from annotation or alignment artifacts, we required
U12-type introns, having terminal dinucleotides other than /AT-AC/, /GT-AG/, or /GC-AG/,
to have GenomeThreader transcript or protein alignments that have flanking exons with
100% identity to the genome sequence and length >= 15. Introns from manual annotations
were not limited by this condition.

Information Content of Exonic Nucleotides. Information content provides a measure
of sequence conservation in aligned sequences that is additive across sites, includes samples
size corrections, and provides a consistent measure to compare information among different
sites [33]. This method has been used to measure sequence conservation of splice sites in
several studies [6, 34, 35]. Information content and sequence logos [34] were generated using alpro [36] and makelogo [37] programs.

**Sampling test.** In many of our analyses, we compared mean features of relatively small sets of U12-type host genes with features from genes from a much larger set. To test whether sample size was responsible for the difference of means between the small sample and the much larger sample, we conducted a sampling test. The null hypothesis for this test is that the small sample has a greater mean than the large sample. 10,000 random samples of the same size as the small sample were extracted from the large sample. The p-value was the fraction of random samples with a mean at least as large as the small sample. We declared p-values less than 0.05 to be significant.

**Additional calculations.** Pairwise protein sequence identity, similarity, and coverage were determined from the multiple sequence alignments. Similar amino acids are defined as those with a positive value in the BLOSUM62 similarity matrix [28]. Percent identity and similarity was defined as the number of amino acids over the length of the shorter protein sequence. These same measures were tabulated for the local regions of cintrons: 10 AA to the right of the cintron, the cintron position, and 10 AA to the left of the cintron. Functional annotations were associated with U12-type intron host gene by BLASTP [28] to the UniRef50 [38] subset of the UniProt References Clusters database [39] and extracting the best hit’s UniRef ID and description.
Results

Database Summary and Tools

The CIWOG database contains a non-redundant set of introns, intron supporting evidence, orthologous gene clusters ("uOGCs") (Table 1), intron sites, and other subsidiary data. CIWOG provides an integrated and dynamic multiple alignment display for analysis of intron, gene structure, and protein sequence conservation, that is driven by its database backend (Figure 5). Entry to CIWOG is provided through a summary table that lists each U12-type intron, (http://ciwog.gdcbl.iastate.edu/ciwog-cgi/table1.pl) and through a BLAST utility for searching by sequence. (http://ciwog.gdcbl.iastate.edu/ciwog-cgi/blastCIWOG.pl) CIWOG also includes a utility for users to score their own intron sequence and receive an intron type classification. (http://ciwog.gdcbl.iastate.edu/ciwog-cgi/U12score.pl)

U12-type Intron Identification and Gene Structure Annotation

U12-type introns were identified by sequence composition and orthology criteria. The foundation of this study is a set of predicted U12-type introns. To generate this dataset, we used a two step intron classification procedure that is first a function of intron sequence composition and then a function of orthology. The first classification step began with a query on the CIWOG database to obtain a training set of non-redundant transcript-supported introns having /ATATCC-AC/ flanking sequences to build position weight matrices for the donor and branch sequences. These matrices were used to score the sequence composition of each intron, by assigning a donor and branch score. (Figure 2 - an example of Arabidopsis intron scores) A cluster of AT-AC introns has high scores, indicative of the introns from the
training set, but there is not a clear distinction separating these introns from the rest of the distribution. After manual examination of specific scores and associated intron sequence composition, we decided to establish a conservative minimum donor and branch scores to classify predicted U12-type introns in keeping with a previous study [5]. Introns having a donor score $\geq 4$ and having a branch score $\geq 1.5$ were classified as predicted U12-type, and otherwise as predicted U2-type introns (indicated by the box in Figure 2). These U12-type introns were used as seeds to identify additional U12-type introns in the second classification step.

Manual inspection of U12-type intron sites in uOGCs revealed that some predicted U2-type introns in these intron sites had high donor and branch scores, but did not exceed the minimums for U12-type intron seed classification (see Figure 3). We decided that the cutoff scores used for seed U12-type intron classification were too stringent and, using the evidence of orthology, decided to modify the classification criteria for these introns. For our final predicted U12-type intron set, we empirically set a donor score cutoff of 3 and removed the branch score cutoff, which is justified since branch sequences are known to be more variable and less conserved [2, 7].

However, selecting a cutoff to demarcate the intron types assumes that there are no introns that can be spliced by both spliceosomes. We decided to designate cintrons in the twilight zone between the predicted intron types as members of a dual intron type, and empirically set a minimum donor score of 2 (Figure 3). All remaining introns were classified as U2-type, except non-canonicals which had additional conditions (see Methods). As a result of the twilight zone being occupied by the dual class, comparisons between predicted U12-type and predicted U2-type introns can be made with greater confidence.
The total number of U12-type introns per organism has a large range and more closely related organisms have similar totals. Biological reasons for the difference in U12-type intron abundance across the organisms are gene duplication or loss, and intron gain or loss. Technical reasons for the difference are incomplete genome sequences, incomplete transcript libraries and incomplete or inaccurate gene annotation.

**Novel gene structure annotations were constructed for U12-type introns.** After identifying seed U12-type introns, we identified U12-type introns that were only contained in transcript alignments. We constructed as many gene annotations as possible for these gene-less U12-type introns for plants using the yrGATE manual gene annotation tool [26] and for animals using the GenomeThreader spliced alignment program [27] with native transcripts. After identifying uOGCs, we constructed annotations using native transcript sequences and orthologous protein sequences using GenomeThreader [27]. An example annotation from each method is presented in Figure 4. The significance of our annotation efforts is apparent through the fraction of the number U12-type introns unique to new annotations over the number of total U12-type introns, which ranges from 7% to 24% across the organisms (Table 1). In summary, we used manual annotation, native transcript alignments and orthologous protein alignments to capture new U12-type introns in novel gene structure annotations.

**Terminal dinucleotides of U12-type introns confirm trends.** The majority of the U12-type introns have either GT-AG or AT-AC terminal dinucleotides, also called the subtypes [2, 3], which is consistent with prior studies [2, 5, 6, 14] (Table 2). In addition, we also detected non-canonical varieties, including several new varieties such as one CT-AC intron (Table 2). With the exception of C at position 1, this intron in the mouse gene NM_180662 has the full donor sequence and branch sequence U12-type consensus and is orthologous to a
AT-AC U12-type intron in human (uOGC #4028). Overall, the most non-canonicals were found in mouse and human, which is in part a result of the transcript support condition and mouse and human having a much higher number of transcripts compared to the other organisms (Table 3).

Most of the non-canonicals have variations in the acceptor site dinucleotides (73 of 83), and this agrees with the idea that the branch sequence rather than the acceptor sequence is the main determinant of the 3’ end of U12-type introns [14]. Nearly all (82 of 83) of the non-canonical varieties have 1 nucleotide difference from one of the canonical subtypes, which suggests that multiple mutations in the terminal nucleotides are not favored by the U12-type spliceosome. Likewise, no double mutations in the second intron position and second to last intron position were recorded, which is consistent with in vitro and in vivo studies that found such double mutations to be defective for splicing [4].

Evolutionary Dynamics of U12-type Introns

U12-type introns occur more often than by chance in shared intron sites. A prior study demonstrated that introns occur more often than by chance in the same position in orthologous genes, and the majority of these introns can be deemed orthologous [32]. U12-type introns occupying the same position in genes have also been inferred to be orthologous [2, 5-7, 9]. To test this idea on the CIWOG dataset, we performed a randomization test upon matrices of intron sites of uOGCs. To build these matrices, one gene per organism was selected from each uOGC by the criteria of greatest number of U12-type introns followed by greatest number of U2-type introns. Each OGC was represented as a matrix of coded introns sites, thus preserving the intron structure of each cluster, some of which are intron rich and
some are intron poor. To derive the set of introns that are in confident alignment regions, genes with ‘external missing’, ‘internal missing’, or ‘none’ type at a U12-type intron site were removed and, subsequently, all columns with those types. Dual-type introns were coded as U2-type introns. Then, the intron sites were shuffled within each gene, and the columns having U12-type introns common to a pair of organisms were counted. This shuffling procedure was repeated 10,000 times. The number of shared U12-type intron sites in the original data was compared to the shuffled data, and in no case was the number of common introns greater in the shuffled data than the original data. Therefore, the frequency of shared U12-type intron sites is significant. The cause of these significantly shared U12-type intron sites between organisms is either common ancestry or independent insertions into the same sites. We infer by parsimony that shared U12-type introns are the product of common ancestry and be appropriately called conserved.

Tabulation of U12-type intron sites reveals conservation, conversion, and loss between all clades. The pathway model states that U12-type introns convert to U2-type introns and then are lost. To quantify these dynamics of U12-type introns with the larger intron collection in the CIWOG dataset, we compared intron sites between clades. In this comparison, we assumed the pathway model of U12-type evolution [2], which proceeds U12-type to U2-type to intron loss, and that all U12-type introns are ancestral to the organisms of this study. Following this pathway model for each intron site, we compared the types from each clade and classified the site according to what clade-pair types occur in the site: if at least one pair of U12-type introns occur, the site is counted as “conserved”; otherwise, if at least one pair of a U12-type intron and U2-type intron occur, the site is counted as “converted”; otherwise, if at least one pair of U12-type intron and absent-intron occur, the
site is counted as “lost”. Clades were compared with respect to their last common ancestor according to our species tree (Figure 6).

The conservation of U12-type introns is roughly proportional to the divergence of the organisms in this study (Table 4). Among vertebrates, U12-type introns are very highly conserved, although low levels of conversion and loss were detected. Among plants, U12-type introns are highly conserved, but intron conversion and loss also occur at appreciable levels. Between Drosophila and vertebrates, U12-type introns have typically been lost or converted to U2-type, but nearly all (14 of 15) Drosophila U12-type introns are conserved in vertebrates. Between plants and animals, which have the greatest evolutionary time separation in this study, U12-type introns are conserved, converted and lost in approximately a 1:2:4 ratio, with conversion and loss occurring in both clades.

**Subtype and U12-type intron site classification are related in some clade comparisons.** Possibly, the U12-type intron subtypes are functionally different, such that it affects their rates of conservation, conversion and loss. For example, if the AT-AC subtype had some selective advantage over the GT-AG subtype, one would expect to see a greater proportion of conservation of the AT-AC subtype than of the GT-AG subtype. To test this hypothesis, we divided U12-type intron sites classified as conservation, conversion, and loss, by subtype and by clade comparison (Table 5). \( \chi^2 \) tests were performed to evaluate independence between intron phase and intron site classification. Three clade comparisons, (chicken versus mammals, drosophila versus vertebrates, and Arabidopsis versus rice), had significant (p-value < 0.05) results to allow rejection of independence between intron phase and intron site classification. We conclude that in some clade comparisons there is a
relationship between intron type and intron site classification, but we are unable to draw a
collection over all clade comparisons.

Non-canonical U12-type introns reveal a possible intermediate in the pathway
model of U12-type evolution. The pathway model of U12-type intron evolution specifies
that U12-type introns convert to U2-type. It has been proposed that prior to this conversion
event, the AT-AC subtype mutates to the GT-AG subtype, a so-called subtype switch event
[2]. The shortest mutational path between the subtypes would involve two mutations, first
intron position: A → G and last intron position C → G. U12-type introns without a
common terminal dinucleotide (termdn) between the clades can provide indicators of
mutation events. To evaluate this mutation pathway, we tabulated U12-type introns
between clades in which one clade has at least one non-canonical U12-type intron and there
is no common termdn ("not-conserved" in Table 6). Cintrons with AT-AG and GT-AG
termdns provide evidence for the shortest mutational path for subtype switching, with AT-
AG being a one-step intermediate from AT-AC. Our data shows that these cintrons are
prevalent (Table 7). We do not observe cintrons with AT-AC and GT-AC termdns, with GT-
AC being one-step intermediate. This suggests that the GT-AC intermediate is not a
common intermediate in this mutational pathway. In contrast, the most common variety AT-
AC and AT-AT is not on the shortest path. Perhaps, subtype switch events take several paths
instead of just the shortest path. We conclude that AT-AG is a likely intermediate in the
mutational pathway for subtype-switching, supporting prior studies [2, 6]

A U12-type Intron is extraordinarily conserved in the Fragile Histidine Triad
uO GC. The uOGC 5709 (Figure 7A), annotated as the Fragile Histidine Triad gene,
provides an extraordinary case of U12-type intron conservation. Intron site #7 contains a
GT-AG U12-type intron in human, mouse and chicken, an AT-AC U12-type intron in Arabidopsis and an AT-AA U12-type intron in rice, which suggests one subtype switching event and one acceptor site dinucleotides change have occurred. Both plant U12-type introns are preceded by a particularly short 3-base exon and a U2-type intron, intron site #6. This plant gene structure is accurate because of the following compelling evidence: the U12-type donor consensus follows the 3-base exon, this 3 base exon is supported by several transcript sequence alignments in both species (Arabidopsis: 9; Rice: 8) (Figure 7B), the exons flanking the U12-type and upstream U2-type intron have at least one transcript sequence in both species with 100% genomic identity, and the U12-type intron is conserved in animals.

yrGATE [26] was used to construct these gene structure annotations, because the published annotations do not reflect this U12-type intron and do not conform to this evidence. This U12-type intron is extraordinary because it is maintained in the presence of a very close upstream intron. By parsimony and the model of U12-type evolution, the U12-type intron was in the last common ancestor of the organisms, but the plant-specific upstream U2-type intron could be a result of intron gain in plants or intron loss in animals. In either case, the U2-type intron has existed in the presence of the U12-type intron and has not, over time, claimed the acceptor site of U12-type intron and merged the two introns into one U2-type intron, or vice versa. For this reason, we have highlighted this intron as an example of an extraordinarily conserved U12-type intron.

**Genes Enriched with U12-type Introns**

**U12-type introns are concentrated in 47 uOGCs.** The initial large scale study on U12-type introns identified 4 genes with multiple U12-type introns, which is significant since
U12-type introns are a rare occurrence in genes [2]. Our results confirm the presence of multiple U12-type intron host genes [2, 5, 6], while increasing the species breadth and identifying many more genes (167) relative to prior studies (Table 8; http://ciwog.gdcbeast.edu/ciwog-cgi/table_genpes.pl). Reducing these genes by common ancestry, there are 47 uOGCs containing these genes. (http://ciwog.gdcbeast.edu/ciwog-cgi/table_multiple.pl) In this set of 47, uOGCs #772 and 25391, have 3 U12-type intron sites. uOGC #772 is the previously identified Sodium/hydrogen exchanger gene family [5, 40], where there are two U12-type intron sites shared between plants and animals, one U12-type intron site specific to animals, and one U12-type intron site specific to plants. uOGC #25391, functionally annotated as E3 ubiquitin-protein ligase HOS1, consists of one Arabidopsis gene with 3 U12-type intron intron sites and one rice gene sharing two of these U12-type intron sites.

To test if this observation could be caused by random U12-type occurrence in genes, we performed a randomization test. For this test, each uOGC was represented as a string of intron site classifications: U12-type if there was at least one U12-type intron, otherwise, U2-type. The labels for each intron site were shuffled across all uOGCs and the number of uOGCs having multiple U12-type intron sites were recorded. In 1,000 repetitions, there were no trials having more than the observed number of multiple U12-type intron host genes, indicating that this observation is non-random.

Additionally, 8 uOGCs (#'s 655, 1452, 4569, 6590, 7485, 10939, 25391, and 3949) have adjacent U12-type intron sites. A randomization test, consisting of shuffling the intron sites within these 47 uOGCs and recording adjacent U12-type intron sites, found this adjacency
relationship to be expected by chance (509/1000 trials had at least 8 genes with adjacent U12 introns).

**Five uOGCs have mutually exclusive U12-type intron sites between plants and animals.** There are 5 uOGCs in which one group of genes has one U12-type intron site and another group of genes has a different U12-type intron site. Interestingly, the 5 uOGCs follow the same trend that plants have one U12-type intron site and animals have a different U12-type intron site. For example, uOGC #2929 has two U12-type intron sites, #3 and #5. At intron site #3, human, mouse, and zebrafish have a U12-type intron, and Arabidopsis and rice have a U2-type intron. At intron site #5, Arabidopsis and rice have a U12-type intron, and human and mouse have a U2-type intron. Assuming the pathway model of U12-type evolution, the ancestral sequence of this cluster had U12-type introns at intron sites #3 and #5, which subsequently converted to U2-type introns in plants and animals, respectively. We conclude that these 5 uOGCs had two U12-type intron sites in their ancestral genes.

**Evolution of Genes with U12-type introns**

U12-type introns are present in diverse eukaryotic organisms at very low levels. Assuming that these remaining U12-type introns are remnants of a formerly popular splicing system, one cause for the persistence of the remaining modern U12-type introns over time is that their host genes experience selection against any change. We found gene structure, protein sequence identity, and local protein sequence identity to be greater in U12-type intron host genes than controls.

**Gene structure is more conserved in uOGCs compared to nOGCs.** Several mechanisms for intron loss and gain in genes over evolutionary time have been proposed
[41], but the prevalence and balance of these events is still much debated [42]. One possible explanation for the persistence of U12-type introns is that they reside in genes with generally slowly evolving gene structures, that undergo little intron loss and/or gain. To test this hypothesis, we compared conservation of gene structure within uOGCs to conservation of gene structure within nOGCs. We define a measure of gene structure similarity (GSS) between the sequences of two clades within an OGC as the number of shared intron sites with common introns (at least one sequence in each clade has an intron in this position), divided by the total number of shared intron sites including sites in which one sequence of one clade aligns well to one sequence of the other clade, but introns occur only in sequence of one of the clades (see Figure 10 for example). Our sets for comparison are the uOGCs for which there is at least one U12-type intron between the clades and the nOGCs for which there is at least one U2-type intron between the clades. Figure 11 shows that uOGCs have greater mean GSS than nOGCs in all of the clade comparisons except human versus mouse. This difference is most apparent in the vertebrates versus drosophila in the animals versus plants comparisons (differences greater than 19%). These differences were found to be statistically significant based on a sampling test (see Methods), in which we compared the uOGCs GSS mean to the GSS means from same-sized random samples from the nOGCs. The lack of a significant difference between human and mouse is possibly explained by the fact that there has been less divergence time between the clades to observe many gene structure changes. Overall, we conclude that U12-type introns tend to be in OGCs with slowly evolving gene structures.

Protein sequence is slightly more conserved in uOGCs compared to nOGCs over short evolutionary time spans. Another possible explanation for the conservation of U12-
type introns is that they reside in genes that have slowly evolving protein sequences. We tabulated the maximum protein sequence percent identity between genes from one clade versus genes from another clade for each OGC. Our sets for comparison are the uOGCs in which there is at least one U12-type intron between the clades and the nOGCs in which there is at least one U2-type intron between the clades. For each clade comparison except animals versus plants, the uOGCs have a greater mean protein sequence identity than the nOGCs, although these differences were less than 5% in all cases (Figure 12). These differences, except the vertebrate and drosophila comparison, were found to be statistically significant through a sampling test (see Methods). Between drosophila versus vertebrates and plants versus animals, the divergence times may be too large to detect the relationship between U12-type intron conservation and protein sequence. Overall, we conclude that conserved U12-type introns tend to be in OGCs experiencing slightly reduced protein sequence change over relatively short evolutionary time spans.

**U12-type introns have slightly greater local protein sequence conservation than U2-type introns.** A more specific cause for the conservation of U12-type introns may be that the local protein sequence around the U12-type intron, rather than the entire gene, is slowly evolving or is immutable for the sake of the protein’s function, such as being located in a critical domain. To test this hypothesis, we compared sequence conservation in up to 10 amino acids on both sides of U12-type introns and U2-type introns. For regions where one intron is within 10 amino acids of the beginning or end of the protein sequence, the longest available segment was used. The sets for comparison were U12-type introns and U2-type introns from the uOGCs with U12-type introns, so that all introns are sampled from the same clusters. The mean local protein sequence identity of U12-type introns is slightly
greater than U2-type introns, for all clade comparisons. (Figure 13) These differences, except between animals and plants, were found to be statistically significant through a sampling test (see Methods). This difference is largest between vertebrates and drosophila, with a mean of 65 percent identity for the U12-type cintron set and a mean of 52.2 percent identity for the U2-type cintron set, which equates to a difference of about 2 conserved amino acids in the local region surrounding the cintron. We conclude that cintrons of U12-type introns tend to have a small increase in local protein sequence conservation compared to U2-type cintrons.

Analysis of Phase in U12-type introns

Phase and intron type are related. Over all introns, intron type and phase are not independent ($\chi^2$ test, p-value < 2.2e-16). U2-type introns tend to be phase 0 and lie between codons, and U12-type introns tend to be phase 1 or phase 2 and interrupt codons (Figure 14), confirming earlier reports [2, 14].

Phase and U12-type intron subtype are related. Comparing the subtypes of U12-type introns across all organisms, we found that the subtypes are not independent ($\chi^2$ test, p-value = 9.8e-4). The ranked order of phase frequencies in the AT-AC subtype is 2, 1, 0, while the GT-AG subtype is 1, 2, 0 (Figure 15). Overall, both subtypes of U12-type introns have tendency towards phase 1 or 2.

U12-type cintrons are not enriched for overall U12-type intron phase preferences. Possibly, the U12-type cintrons between clades, since they have been conserved over time, may be enriched for the preferred phases of all U12-type introns. In all clade comparisons, the subtype phase distribution of U12-type cintrons have the same tendency towards phase 1.
and 2 (Figure 16). However, the subtype phase frequencies of nearly all clade comparisons are not significantly different than the overall corresponding subtype phase distribution (by $\chi^2$ test; p-values > 0.05, except in plants versus animals in which the AT-AC subtype has insufficient data). Providing the only significant difference ($\chi^2$ test; p-value = 0.02327), Arabidopsis and rice AT-AC U12-type cintrons have significantly different phase frequencies than overall AT-AC U12-type introns. These cintrons have approximately equal frequencies of phase 0 and phase 2 introns, but this is not a clear enrichment for the overall U12-type intron preferred phases. Overall, we found that U12-type cintrons between nearly all clade comparisons exhibit the same distribution and are not enriched for particular phases relative to all U12-type introns.

**Exonic sequence conservation of U12-type introns correlates to phase frequencies.** Since the intron phases are significantly different between the intron types, the interface of introns with the phase, which is the flanking exonic nucleotides occupying different codons, could explain this difference. To test this hypothesis, we compared exonic sequence conservation within 5 nucleotides on both sides of introns between phases of the following sets of introns: AT-AC U12-type introns, GT-AG U12-type introns, and GT-AG U2-type introns. To prevent redundant sampling of related introns within each set, introns were limited to human genes from the uOGCs and one intron of the appropriate type per intron site was randomly selected. Information content [33, 34] was used to assay sequence conservation within these sets. For U2-type introns, the phase rank by information content is $0 > 1 > 2$ and is the same as the phase rank by frequency of all U2-type introns (Figure 16). This observation confirms an earlier report [35]. In contrast to U2-type introns, GT-AG U12-type introns have the phase rank by information content of $1 > 2 > 0$, which is the same
as the phase rank by frequency of all GT-AG U12-type introns (Figure 16). AT-AC U12-type introns have the same phase rank by information content with $1 > 2 > 0$, but this ranking is different from the phase rank by frequency of all AT-AC U12-type which has 2 before 1. Despite this difference, both U12 subtypes have greater information content in phase 1 and 2 than phase 0, which corresponds to both subtype's overall U12-type phase frequency distributions. Overall, we conclude that exonic sequence conservation is correlated with the phase distribution of U12-type introns.

**Subtypes of U12-type introns and U2-type introns have different patterns of exonic sequence conservation.** GT-AG U12-type introns have greater sequence conservation than AT-AC U12-type introns of corresponding phase, including a preference for T at position -1 in the GT-AG subtype that is absent in the AT-AC subtype overall phases (Figure 16). Also, the position and magnitude of sequence conservation have differences between the subtypes, such as a preference for A at positions +1 in phase 1 introns. Over all phases, the U2-type exonic consensus is AAG|GT which is noticeably different than the U12-type exonic consensus, T|ATAT. We conclude that the phases and subtypes of U12-type introns have different degrees of positional sequence conservation.

**Exonic sequence conservation does not support a U12-type splicing constraint acting on gene sequence.** One explanation for the difference in positional information content among the phases of U12-type introns is that there is a splicing constraint [43] operating on the protein sequence. Such a constraint would be most directly seen through increased conservation at wobble positions in codons. Positions 0 and 3 in phase 1 GT-AG U12-type and positions -1 and 2 in phase 2 GT-AG U12-type, all have relatively high information content and are not in wobble positions. Position -1 in phase 0 GT-AG U12-type and
position 3 in phase 1 GT-AG subtype, are in wobble positions, but their information content is not substantially greater than the previously listed non-wobble positions. We conclude that a splicing constraint is not a strong factor for these different degrees of sequence conservation among the phases of U12-type introns.

**Amino acids flanking U12-type introns are not especially conserved.** Another possible explanation for the phase distribution of U12-type introns is that the immediate amino acids flanking the U12-type intron are significant to the function of the protein. Under this scenario, the protein coding constraint of the gene could favor U12-type introns of different phases, through the relationship of the preferred exonic sequence for U12-type intron splicing and amino acids specified by the codons in these exonic sequences. If such a relationship existed, one would expect greater conservation of the flanking amino acids of U12-type introns in uOGCs relative to other amino acids in the uOGCs. To test this hypothesis, we calculated amino acid conservation between clade pairs for the regions -10 to 10 surrounding U12-type introns. An example clade comparison of mammals and chicken versus zebrafish (Figure 17) indicates that there is no clear increase of conservation in either the preceding codon, interrupted codon in case of phase 1 and 2, or following codon, relative to the other codons in the region. This trend holds through the other clade comparisons (Data not shown). We conclude that amino acid conservation in uOGCs is unrelated to proximity to conserved U12-type introns of any phase.
Discussion

Annotation Accuracy of U12-type Introns

We found substantial numbers of U12-type introns that were not contained in published gene structure annotations, for which we were able to create gene structure annotations. Since we were able to create protein coding gene structure annotations for introns, these introns overall likely represent biologically authentic introns rather than alignment artifacts. Transcript alignments, the source of these introns, are therefore very useful in identifying U12-type introns. It should be noted that other published annotation sources apart from the ones used in this study may have less under-annotation of U12 type introns, but we attempted to pick the best available published annotation sources for each organism. Increases in transcript sequences will likely enable the discovery of further U12-type introns.

Support for Pathway Model of U12-type Intron Evolution

The concentration of U12-type introns in some uOGCs is quite interesting given these introns scarcity in modern genomes. We demonstrated that the uOGCs having genes with more than one U12-type intron site is not the product of chance. Also, the 5 uOGCs with mutually exclusive U12-type intron sites between plants and animals argue for an ancestor with two U12-type intron sites in these uOGCs. These observations support the pathway model of Burge [2] that predicts ancestral genes with only U12-type introns, which would provide a non-random concentration of these introns. One possible explanation for these modern uOGCs with multiple U12-type intron sites involves splicing efficiency of the U12-type spliceosome. After splicing one intron, the spliceosome is in close proximity to the
remaining un-spliced intron in the gene, which is significant because the spliceosome is scarce within the cell [44], and the remaining intron is spliced more easily. This situation would decrease any selective pressure to lose this pair of U12-type intron sites over time, compared to a gene with a single U12-type intron site.

The subtype switching phenomena is supported by our data. The intermediate, AT-AG, that is on the shortest path between AT-AC and GT-AG subtypes, is common among U12-type introns. Other varieties of U12-type introns prevent us from concluding that the shortest path is the dominant path. However, we are able to conclude that subtype switching as well as conversion to U2-type introns does not require selection of a different acceptor site, which would alter the protein sequence, as proposed by an earlier study [5]. This conclusion is based on our detection of intron sites having terminal dinucleotide mutation and intron type conversion, which by definition are located in the same position within genes.

**U12-type introns Evolutionary Dynamics**

We classified U12-type intron sites among uOGCs according to the pathway model and made some new discoveries. Prior studies on the dynamics of U12-type introns have documented conservation, conversion and loss among homologous genes of different species [2], but among orthologous genes, strictly conservation has been observed between human and mouse [6] and among human, mouse, rat and chicken [7]. We detected the first cases of conversion and lost U12-type intron sites between vertebrate, between plant, and between plant and animal orthologous genes. The rates of conservation, conversion and loss vary by clade. Animals and plants have a 1:2:4 ratio of conservation, conversion and loss.
Extrapolating from this clade comparison, we conclude that U12-type introns have the greatest tendency to be lost over time.

**Gene Evolution and U12-type Intron Conservation**

We found that uOGCs with conserved U12-type introns have special characteristics that are associated with conservation of U12-type introns rather than U2-type introns. Gene structure similarity was found to be greater in uOGCs with conserved U12-type introns than nOGCs with conserved U2-type introns. One possible explanation is that U12-type introns are passengers on slowly evolving gene structures, and this would provide an explanation for the continued maintenance of the U12-type intron splicing system over time. An alternative explanation is that U12-type introns reduce the ability for a gene to gain or lose introns. Perhaps, the experimentally verified ability of U12-type introns to experience a reduced rates of splicing than U2-type introns [10] could be extended to an ability to reduce the number of partially or completely spliced mRNA in the cell. This reduction could cause a reduction in the likelihood of reverse transcription of the mRNA, which could then recombine with the genomic gene, causing intron loss, which is one of the mechanisms for intron loss [45, 46].

Global protein sequence identity and local protein sequence identity were found to be slightly greater within uOGCs having conserved U12-type introns than nOGCs having conserved U2-type introns across short evolutionary time spans. A possible interpretation that is overall slow sequence change in these uOGCs is causing persistence of U12-type introns. Over larger time spans, we do not observe this small effect, which would be indistinguishable from other protein sequence change.
Phase and Exonic Sequence Conservation

The difference of phase preferences between intron types provided an intriguing lead that we pursued. We did not find significant differences of specific phases in conserved U12-type introns compared to the overall U12-type intron phase distribution. We also did not find a single U12-type intron with more than one phase. For these reasons, we propose that modern U12-type phase distribution was present in the last common ancestor of the organisms of this study. Exonic sequence conservation and frequency of intron phases were shown to be correlated. This correlation could be explained by increased splicing efficiency for particular exonic sequences. Our data does not support a clear splicing or protein coding constraint. Perhaps such constraints are not observable with modern sequence data, considering these phase preferences were established in the distant past. In any case, this correlation provides grounds for future studies into the early evolution U12-type introns.
References

17. Perl [http://www.cpan.org/]


36. alpro [http://www.lecb.ncifcrf.gov/~toms/delila/alpro.html]

37. makelogo [http://www.lecb.ncifcrf.gov/~toms/delila/makelogo.html]
38. **Uniref50** [ftp://ftp.ebi.ac.uk/pub/databases/uniprot/uniref/uniref50]


50. **AtGDB Data Download** [http://www.plantdb.org/AtGDB/download.php]


52. **UCSC Genome Browser Data Download** [http://hgdownload.cse.ucsc.edu/downloads.html]

53. **Ensembl Genomic Data** [http://www.ensembl.org/info/data/index.html]

54. **OsGDB Data Download** [http://www.plantdb.org/OsGDB/download.php]

Figures

1. **Import of external data:** genome sequence, transcript alignments, gene structure annotations

2. **Processing of genes, transcripts and introns**

3. **Intron classification by PWM**

4. **Gene structure annotation of U12-type introns using native transcript sequence**

5. **Clustering of orthologous genes, selection of representative genes and multiple alignments**

6. **Intron site derivation**

7. **Gene structure annotation of orthologs of U12-type intron host genes using orthologous protein sequence and native transcript sequence**

8. **Repeat representative gene selection, multiple alignment and common intron derivation steps**

9. **Intron classification using homology, and alignment support for non-canonical U12-type introns**

**Figure 1.** The CIWOG database construction flow chart.
Figure 2. Distribution of donor and branch scores of Arabidopsis introns in CTWOG. GT-AG introns are light blue. AT-AC introns are dark blue. Introns with other terminal dinucleotides are green. Donor score is horizontal axis and branch score is vertical axis. The box indicates cutoff scores for U12-type intron seeds.
Figure 3. Scores of introns at an intron site with a seed U12-type intron. U12-type intron seeds are dark blue. U12-type introns by homology are light blue. Dual-type introns are green. U2-type introns are black. Donor score is horizontal axis and branch score is vertical axis. Vertical lines indicate cutoff scores for intron types.
Figure 4. Example of novel annotation of predicted U12-type introns in rice and Drosophila. (A) presents rice genomic region 24,930,600 to 24,934,000 on chromosome 4 from OsGDB, with following color assignments: green – yrGATE manual gene annotation, pink – TIGR v6 gene annotation, light blue – full length cDNA alignments, red – EST alignments. The fourth intron of the yrGATE [26] gene structure is a U12-type intron, which also corresponds to the full-length cDNA and EST alignment evidence. The corresponding intron junction in the TIGR gene annotation misrepresents this intron and does not correspond to the transcript evidence. (B) presents Drosophila genomic region chromosome 3R from 4,835,999 to 4,840,610 nt from the UCSC Genome Browser [21]. The third intron of the GenomeThreader annotation is an AT-AC U12-type intron, which corresponds to the mRNA alignments. The corresponding intron of the FlyBase and RefSeq annotation misrepresents this intron and does not correspond to the transcript alignment evidence.
Figure 5. CIWOG alignment display.

The multiple alignment is presented as a graphic atop the page where each gene is represented as a rectangle with the following color assignments: black – aligned sequence, gray – internal gaps, lines of the same color – cintrons, black asterisks – U12-type introns, brown asterisks – dual-type introns. Numbers flanking genes in the graphic indicate the number of UTR introns. The multiple alignment contains orange rectangles to represent introns and orange perimeter boxes to represent absent-introns. Intron sites are presented in boxes to the right of the alignment. The first line of the intron site box lists the intron site number and alignment position, and the other lines list information corresponding to each intron in the box. When either the intron in the sequence alignment or an intron row within the intron site box is moused-over, both elements are highlighted with a yellow color to indicate their correspondence. By clicking on an intron line within the summary graphic, the display is scrolled to the intron within the multiple alignment. To view a gene in the context of its supporting evidence and other annotation features, the gene names are linked to the genome browser of the parent data source. Other information in this display interface includes related Uniref gene descriptions, each gene’s average protein sequence identity and similarity, the number of genes in each alternative gene group, and a link to download protein sequences. Detailed information about the alternative gene groups and intron sites can be accessed by clicking on the alternative gene group and intron site position hyperlinks.
Figure 6. Species tree. This species tree was obtained from a review [47] and branch order is not proportional to time. Clade nodes are numbered for reference in this paper.
**Figure 7.** Fragile histidine triad orthologous gene cluster has a U12-type cintron flanking a 3 base exon in plants.

(A) Orthologs from human, mouse, chicken, Arabidopsis and rice have a U12-type intron (intron site #7), with animals having the GT-AG subtype, Arabidopsis having the AT-AC subtype and rice having AT-AA termini. Preceding the U12-type intron in Arabidopsis and rice, there is a U12-type intron (intron site #6) and a 3 nucleotide long exon. Only the segment of multiple alignment corresponding to these cintrons is shown. The human gene, NM_002012, is functionally annotated as the fragile histidine triad gene, which has aberrant transcripts found in half of lung, esophageal, stomach, and colon carcinomas [48]. (B) shows a GeneSeqer [49] spliced alignment of an Arabidopsis EST, which includes the U12-type introns and upstream 3 base exon.
Figure 9. Example of a cintron having U12-type introns with non-canonical terminal dinucleotides that are not conserved. This alignment segment corresponding to the cintron with U12-type introns is shown. This cintron is indicated by asterisks in the alignment plot. All three acceptor site intron dinucleotides are different among the three U12-type introns: AT, AC, and AA.
Figure 10. Gene structure similarity in a hypothetical orthologous gene cluster. This diagram represents a protein sequence multiple alignment and introns of a hypothetical orthologous gene cluster with the following conventions: aligned protein sequence - black rectangle, introns – gray rectangle, absent-introns – white rectangle. Intron sites are numbered atop the alignment diagram. Internal missing and external missing types are not shown. This cluster contains three different clades (A, B, and C) and two genes per clade. Gene structure similarity (GSS) is defined as the number of shared intron sites with common introns (at least one sequence in each clade has an intron in this position), divided by the total number of shared intron sites including sites in which one sequence of one clade aligns well to one sequence of the other clade, but introns occur only in sequence of one of the clades. GSS between clade A and B is 2/3. GSS between clade A and C is 3/4. GSS between B and C is 2/2.
Figure 11. Gene structure similarity of orthologs with a U12-type cintron and orthologs with a U2-type cintron. Gene structure similarities are placed into bins of size 5, and each bin is plotted at the midpoint of the bin.
Figure 12. Comparison of protein sequence percent identity between orthologs with a U12-type cintron and orthologs with a U2-type cintron and no U12-type cintron. Orthologous clusters were placed into bins of size 10, and each bin is plotted at its midpoint.
Figure 13. Comparison of local protein sequence percent identity between U12-type cintrons and U2-type cintrons. Local protein sequence identities are placed into bins of size 10, and each bin is plotted at its midpoint.
Figure 14. Comparison of phases of between U12-type introns and U2-type introns overall organisms. The vertical axis is the number of introns with a given phase. Introns are from U12-type intron host genes.
Figure 15. Phases of U12-type cintrons by clade comparison and subtype. The vertical axes are the number of cintrons with a given phase. Only cintrons with common terminal dinucleotides are considered.
Figure 16. Sequence conservation of U12-type introns by intron type and phase. Sequence logos [34] and information content [33] corresponding to the region around introns (-5:+5) were created using the programs alpro [36] and makelogo [37]. These logos show the information content per position in bits, which is represented by the height of the characters. Error bars indicate one standard deviation of the positional information content in both directions. The total information content of a set is $R_s$. The black vertical lines indicate the position of the intron. The rectangles indicate codon positions. The maximum possible bit score for a position is 2 bits.
Figure 17. Amino acid conservation across U12-type cintron junctions.

The percent of cintrons with a conserved amino acid between clades at each position relative to the U12-type cintron are plotted. Conserved amino acids are defined as one gene from each clade having the same amino acid. Position 0 (red) is the amino acid with a complete codon before the intron. Position 1 (green) is the amino acid, whose codon is interrupted by phase 1 and phase 2 introns. Position 2 (blue) is the following codon. The black line indicates the mean for the phase and terminal dinucleotides group. In this example clade comparison of mammals and chicken versus zebrafish, there is not significantly increased conservation of amino acids at positions 0, 1 or 2 relative to the other positions, and this trend holds for all phases and for both subtypes (data not shown).
Tables

<table>
<thead>
<tr>
<th>Predicted U12-type Introns</th>
<th>Arabidopsis</th>
<th>Rice</th>
<th>Drosophila</th>
<th>Zebrafish</th>
<th>Chicken</th>
<th>Mouse</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>U12-type seeds</td>
<td>217</td>
<td>237</td>
<td>16</td>
<td>553</td>
<td>383</td>
<td>549</td>
<td>562</td>
</tr>
<tr>
<td>U12-type using orthology</td>
<td>24</td>
<td>28</td>
<td>2</td>
<td>80</td>
<td>72</td>
<td>81</td>
<td>90</td>
</tr>
<tr>
<td>Total</td>
<td>241</td>
<td>265</td>
<td>18</td>
<td>633</td>
<td>455</td>
<td>630</td>
<td>652</td>
</tr>
<tr>
<td>Predicted Dual-type</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predicted U12-type Introns in Orthologous Gene Cluster ( % of Total)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New U12-type introns in Host Gene Annotations*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>yrGATE %</th>
<th>GenomeThreader %</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Introns</td>
<td>(24%)</td>
<td>(12%)</td>
</tr>
<tr>
<td></td>
<td>(17%)</td>
<td>(18%)</td>
</tr>
<tr>
<td></td>
<td>(15%)</td>
<td>(7%)</td>
</tr>
<tr>
<td></td>
<td>(11%)</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Predicted U12-type introns in seven organisms.
<table>
<thead>
<tr>
<th></th>
<th>Arabidopsis</th>
<th>Rice</th>
<th>Drosophila</th>
<th>Zebrafish</th>
<th>Chicken</th>
<th>Mouse</th>
<th>Human</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT-AC</td>
<td>61</td>
<td>62</td>
<td>9</td>
<td>135</td>
<td>100</td>
<td>154</td>
<td>170</td>
<td>691</td>
</tr>
<tr>
<td>GT-AG</td>
<td>177</td>
<td>200</td>
<td>8</td>
<td>493</td>
<td>346</td>
<td>443</td>
<td>453</td>
<td>2120</td>
</tr>
<tr>
<td>AA-TG</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>AT-AA</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>6</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>AT-AG</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>10</td>
<td>10</td>
<td>21</td>
</tr>
<tr>
<td>AT-AT</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td>CT-AC</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GG-AG</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GT-GG</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>GT-TG</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

**Table 2.** Terminal dinucleotide distribution of predicted U12-type introns. Varieties of terminal dinucleotides new to this study are highlighted in yellow.
Table 3. Summary of Introns, Transcripts, Genes and Orthologous Gene Clusters in CIWOG. The gene count includes genes from external data sources and novel annotations of this study. The intron count only includes those that meet minimum quality criteria described in the methods section. The transcript count only includes those having at least one intron.
<table>
<thead>
<tr>
<th>Clade A</th>
<th>Clade B</th>
<th>conserved U12:U12</th>
<th>converted U12:U2</th>
<th>U2:U12</th>
<th>lost U12:-</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>mouse</td>
<td>425</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>429</td>
</tr>
<tr>
<td>chicken</td>
<td>mammals</td>
<td>358</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>364</td>
</tr>
<tr>
<td>zebrafish</td>
<td>mammals, chicken</td>
<td>373</td>
<td>5</td>
<td>7</td>
<td>3</td>
<td>392</td>
</tr>
<tr>
<td>Drosophila</td>
<td>vertebrates</td>
<td>14</td>
<td>0</td>
<td>49</td>
<td>21.5</td>
<td>228</td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>rice</td>
<td>147</td>
<td>14</td>
<td>13</td>
<td>1</td>
<td>186</td>
</tr>
<tr>
<td>Plants</td>
<td>Animals</td>
<td>34</td>
<td>25</td>
<td>45</td>
<td>53</td>
<td>245</td>
</tr>
</tbody>
</table>

**Table 4.** Evolutionary dynamics of U12-type intron sites between clades. The number of each pair type, conserved (U12:U12), converted (U12:U2), or lost (U12:-) is presented for each clade pair. Each pair type is also expressed as a percentage of the total pair types.
Table 5. The counts of each subtype in intron sites that are classified as conservation (blue), conversion (green), and loss (black) are shown. The row labels correspond to clades comparisons with respect to the species tree in Figure 6. $\chi^2$ statistics from a test of independence of intron type and intron site classification are shown. A count of one was added to the all categories prior to the $\chi^2$ tests.

<table>
<thead>
<tr>
<th>nodeA</th>
<th>nodeB</th>
<th>conserved</th>
<th>converted</th>
<th>lost</th>
<th>conserved</th>
<th>converted</th>
<th>lost</th>
<th>$\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>112</td>
<td>0</td>
<td>0</td>
<td>407</td>
<td>3</td>
<td>1</td>
<td>0.247</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>91</td>
<td>0</td>
<td>2</td>
<td>353</td>
<td>4</td>
<td>0</td>
<td>7.122</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>94</td>
<td>0</td>
<td>0</td>
<td>369</td>
<td>12</td>
<td>7</td>
<td>1.949</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
<td>8</td>
<td>3</td>
<td>44</td>
<td>14</td>
<td>46</td>
<td>123</td>
<td>10.398</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>41</td>
<td>0</td>
<td>3</td>
<td>144</td>
<td>27</td>
<td>9</td>
<td>6.188</td>
</tr>
<tr>
<td>9</td>
<td>12</td>
<td>14</td>
<td>14</td>
<td>32</td>
<td>30</td>
<td>57</td>
<td>107</td>
<td>2.350</td>
</tr>
<tr>
<td>Cluster ID</td>
<td>Cluster ID</td>
<td>Node A</td>
<td>Node B</td>
<td>Terminal Dinucleotides</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>-----------</td>
<td>--------</td>
<td>--------</td>
<td>------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2418</td>
<td>9</td>
<td>1</td>
<td>2</td>
<td>GTAG ATAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4028</td>
<td>34</td>
<td>1</td>
<td>2</td>
<td>ATAC CTAC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4982</td>
<td>16</td>
<td>1</td>
<td>2</td>
<td>ATAC ATAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5711</td>
<td>14</td>
<td>1</td>
<td>2</td>
<td>ATAC ATAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5737</td>
<td>10</td>
<td>1</td>
<td>2</td>
<td>GTAG GTAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5754</td>
<td>14</td>
<td>1</td>
<td>2</td>
<td>ATAC ATAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9654</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>GTAG ATAA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9854</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>GTAG ATAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15124</td>
<td>11</td>
<td>1</td>
<td>2</td>
<td>ATAC GTAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19713</td>
<td>8</td>
<td>1</td>
<td>2</td>
<td>GTAG ATAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1469</td>
<td>40</td>
<td>1</td>
<td>2</td>
<td>ATAT ATAC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4313</td>
<td>20</td>
<td>1</td>
<td>2</td>
<td>GTGG GTAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4350</td>
<td>53</td>
<td>1</td>
<td>2</td>
<td>ATAG GTAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5417</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>ATAT ATAC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3176</td>
<td>20</td>
<td>1</td>
<td>2</td>
<td>ATAT ATAA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>496</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>ATAG GTAT,ATAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1202</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>GTAT GTAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1430</td>
<td>8</td>
<td>1</td>
<td>2</td>
<td>ATAG,ATAC ATAT,ATAC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1756</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>ATAC,ATAA ATAA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1773</td>
<td>12</td>
<td>1</td>
<td>2</td>
<td>ATAG,GTAG ATAG,GTAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2929</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>ATAA ATAA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3172</td>
<td>42</td>
<td>1</td>
<td>2</td>
<td>ATAT ATAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6173</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>GTAT GTAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8031</td>
<td>23</td>
<td>1</td>
<td>2</td>
<td>GCAG GCAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10197</td>
<td>8</td>
<td>1</td>
<td>2</td>
<td>GCAG GCAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13984</td>
<td>12</td>
<td>1</td>
<td>2</td>
<td>GTGG GTGG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15486</td>
<td>13</td>
<td>1</td>
<td>2</td>
<td>ATAG ATAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15808</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>ATAG ATAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1756</td>
<td>5</td>
<td>3</td>
<td>8</td>
<td>GTAG ATAA,ATAC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5928</td>
<td>10</td>
<td>3</td>
<td>8</td>
<td>ATAC ATAA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21569</td>
<td>10</td>
<td>3</td>
<td>8</td>
<td>GTAG GTGG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1778</td>
<td>41</td>
<td>3</td>
<td>8</td>
<td>ATAT ATAC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2533</td>
<td>17</td>
<td>3</td>
<td>8</td>
<td>ATAT ATAC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1202</td>
<td>5</td>
<td>3</td>
<td>8</td>
<td>GTAT GTAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6. U12-type cintrons having non-canonical terminal dinucleotides.

Cintrons consisting of at least one U12-type intron having non-canonical terminal dinucleotides (not AT-AC, or GT-AG) are listed by clade comparison, which is indicated by node A and node B, which correspond to the species tree in Figure 6.
<table>
<thead>
<tr>
<th>Cluster ID</th>
<th>Cintron ID</th>
<th>Node A</th>
<th>Node B</th>
<th>Terminal Dinucleotides</th>
<th>Node A</th>
<th>Node B</th>
<th>Conserved</th>
</tr>
</thead>
<tbody>
<tr>
<td>1469</td>
<td>40</td>
<td>3</td>
<td>8</td>
<td>ATAT</td>
<td>ATAT,ATAC</td>
<td>conserved</td>
<td></td>
</tr>
<tr>
<td>5737</td>
<td>10</td>
<td>3</td>
<td>8</td>
<td>GTAT</td>
<td>GTAT,GTAG</td>
<td>conserved</td>
<td></td>
</tr>
<tr>
<td>496</td>
<td>3</td>
<td>4</td>
<td>10</td>
<td>GTAG</td>
<td>GTAT,ATAG</td>
<td>not conserved</td>
<td></td>
</tr>
<tr>
<td>2929</td>
<td>3</td>
<td>4</td>
<td>10</td>
<td>ATAC</td>
<td>ATAA</td>
<td>not conserved</td>
<td></td>
</tr>
<tr>
<td>3172</td>
<td>42</td>
<td>4</td>
<td>10</td>
<td>ATAC</td>
<td>ATAT</td>
<td>not conserved</td>
<td></td>
</tr>
<tr>
<td>3176</td>
<td>20</td>
<td>4</td>
<td>10</td>
<td>ATAC</td>
<td>ATAA,ATAT</td>
<td>not conserved</td>
<td></td>
</tr>
<tr>
<td>5754</td>
<td>14</td>
<td>4</td>
<td>10</td>
<td>GTAG</td>
<td>ATAT,ATAC</td>
<td>not conserved</td>
<td></td>
</tr>
<tr>
<td>6173</td>
<td>5</td>
<td>4</td>
<td>10</td>
<td>GTAG</td>
<td>GTAT</td>
<td>not conserved</td>
<td></td>
</tr>
<tr>
<td>8031</td>
<td>23</td>
<td>4</td>
<td>10</td>
<td>GTAG</td>
<td>GCAG</td>
<td>not conserved</td>
<td></td>
</tr>
<tr>
<td>10197</td>
<td>8</td>
<td>4</td>
<td>10</td>
<td>GTAG</td>
<td>GCAG</td>
<td>not conserved</td>
<td></td>
</tr>
<tr>
<td>11557</td>
<td>33</td>
<td>4</td>
<td>10</td>
<td>GTAG</td>
<td>ATAG</td>
<td>not conserved</td>
<td></td>
</tr>
<tr>
<td>15486</td>
<td>13</td>
<td>4</td>
<td>10</td>
<td>GTAG</td>
<td>ATAG</td>
<td>not conserved</td>
<td></td>
</tr>
<tr>
<td>8031</td>
<td>23</td>
<td>4</td>
<td>10</td>
<td>GTAG</td>
<td>GCAG</td>
<td>not conserved</td>
<td></td>
</tr>
<tr>
<td>1899</td>
<td>21</td>
<td>4</td>
<td>10</td>
<td>ATAG</td>
<td>ATAC</td>
<td>not conserved</td>
<td></td>
</tr>
<tr>
<td>6591</td>
<td>9</td>
<td>4</td>
<td>10</td>
<td>ATAT</td>
<td>ATAC</td>
<td>not conserved</td>
<td></td>
</tr>
<tr>
<td>6592</td>
<td>1</td>
<td>4</td>
<td>10</td>
<td>ATAA</td>
<td>ATAC</td>
<td>not conserved</td>
<td></td>
</tr>
<tr>
<td>2269</td>
<td>3</td>
<td>6</td>
<td>7</td>
<td>GTAG</td>
<td>GTAT</td>
<td>not conserved</td>
<td></td>
</tr>
<tr>
<td>3249</td>
<td>2</td>
<td>6</td>
<td>7</td>
<td>ATAA</td>
<td>ATAC</td>
<td>not conserved</td>
<td></td>
</tr>
<tr>
<td>5709</td>
<td>7</td>
<td>9</td>
<td>12</td>
<td>ATAA,ATAC</td>
<td>GTAG</td>
<td>not conserved</td>
<td></td>
</tr>
</tbody>
</table>

Table 6. (Continued)
<table>
<thead>
<tr>
<th>Pair</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATAC ATAT</td>
<td>9</td>
</tr>
<tr>
<td>ATAG GTAG</td>
<td>7</td>
</tr>
<tr>
<td>ATAA ATAC</td>
<td>5</td>
</tr>
<tr>
<td>GTAG GTAT</td>
<td>4</td>
</tr>
<tr>
<td>ATAA GTAG</td>
<td>3</td>
</tr>
<tr>
<td>GTAG GTGG</td>
<td>2</td>
</tr>
<tr>
<td>ATAC ATAG</td>
<td>2</td>
</tr>
<tr>
<td>GCAG GTAG</td>
<td>2</td>
</tr>
<tr>
<td>ATAT GTAG</td>
<td>1</td>
</tr>
<tr>
<td>ATAA ATAT</td>
<td>1</td>
</tr>
<tr>
<td>ATAC CTAC</td>
<td>1</td>
</tr>
<tr>
<td>ATAC GTAT</td>
<td>1</td>
</tr>
</tbody>
</table>

**Table 7.** Frequency of U12-type cintrons having non-canonical and not-conserved terminal dinucleotides.

Cintrons having U12-type non-canonical terminal dinucleotides and no common terminal dinucleotide are tailed by pair type and sorted by frequency. Cintrons with multiple terminal dinucleotides are counted pairwise.
# of U12-type introns per gene

<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>Rice</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>Drosophila</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Zebrafish</td>
<td>30</td>
<td>2</td>
</tr>
<tr>
<td>Chicken</td>
<td>24</td>
<td>3</td>
</tr>
<tr>
<td>Mouse</td>
<td>40</td>
<td>3</td>
</tr>
<tr>
<td>Human</td>
<td>39</td>
<td>3</td>
</tr>
<tr>
<td>total</td>
<td>154</td>
<td>13</td>
</tr>
</tbody>
</table>

**Table 8.** Genes with multiple U12-type introns.
All Drosophila U12-type intron host genes have exactly 1 U12-type intron, in agreement with previous studies [6].
<table>
<thead>
<tr>
<th>Organism</th>
<th>Data Source</th>
<th>Version</th>
<th>Data Type</th>
<th>Name</th>
<th>Date Downloaded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis thaliana</td>
<td>PlantGDB[20, 50]</td>
<td>AtGDB155</td>
<td>Genome Sequence</td>
<td></td>
<td>2/21/07</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Transcript Alignments</td>
<td>cdna_good_pg</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>est_good_pg</td>
<td>3/4/07</td>
</tr>
<tr>
<td></td>
<td>TAIR[24, 51]</td>
<td>6</td>
<td>Gene Annotations</td>
<td></td>
<td>2/21/07</td>
</tr>
<tr>
<td>Danio rerio</td>
<td>UCSC G.B.[21, 52]</td>
<td>danRer4</td>
<td>Genome Sequence</td>
<td></td>
<td>3/9/07</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Gene Annotations</td>
<td>knownGene</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ensGene</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Transcript Alignments</td>
<td>all_mrna, all_est</td>
<td></td>
</tr>
<tr>
<td>Drosophila melanogastor</td>
<td>UCSC G.B. [21, 52]</td>
<td>dm2</td>
<td>Genome Sequence</td>
<td></td>
<td>2/18/07</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Gene Annotations</td>
<td>knownGene</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>refGene</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>flyBaseGene</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Transcript Alignments</td>
<td>all_mrna, all_est</td>
<td></td>
</tr>
<tr>
<td>Gallus gallus</td>
<td>UCSC G.B. [21, 52]</td>
<td>galGal3</td>
<td>Genome Sequence</td>
<td></td>
<td>2/18/07</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Gene Annotations</td>
<td>knownGene</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>refGene</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Transcript Alignments</td>
<td>all_mrna, all_est</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ensemble[22, 53]</td>
<td>42</td>
<td>Genome Sequence</td>
<td></td>
<td>2/20/07</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Gene Annotations</td>
<td>knownGene</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>refGene</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Transcript Alignments</td>
<td>all_mrna, all_est</td>
<td></td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>UCSC G.B.[21, 52]</td>
<td>hg18</td>
<td>Genome Sequence</td>
<td></td>
<td>2/18/07</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Gene Annotations</td>
<td>knownGene</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>refGene</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Transcript Alignments</td>
<td>all_mrna, all_est</td>
<td></td>
</tr>
<tr>
<td>Mus musculus</td>
<td>UCSC G.B.[21, 52]</td>
<td>mm8</td>
<td>Genome Sequence</td>
<td></td>
<td>2/18/07</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Gene Annotations</td>
<td>knownGene</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>refGene</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Transcript Alignments</td>
<td>all_mrna, all_est</td>
<td></td>
</tr>
<tr>
<td>Oryza sativa</td>
<td>PlantGDB[20, 54]</td>
<td>OsGDB157</td>
<td>Genome Sequence</td>
<td></td>
<td>2/10/07</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Transcript Alignments</td>
<td>cdna_good_pg</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>est_good_pg</td>
<td>3/03/07</td>
</tr>
<tr>
<td></td>
<td>TIGR[23, 55]</td>
<td>5</td>
<td>Gene Annotations</td>
<td></td>
<td>2/10/07</td>
</tr>
</tbody>
</table>

**Supplementary Table 1.** CIWOG data sources.
This table presents the data sources that were used in the CIWOG construction procedure.
CHAPTER 5: GENERAL CONCLUSIONS

Conclusions

Accurate and prompt gene structure annotation is a necessity for the utilization of genomic sequence data and large scale biological research. As the number of genomes being sequenced rapidly increases, there is a need for immediate gene structure annotation, which is distinct from the whole genome annotation that is conducted at the completion of sequencing. In completely sequenced genomes, inaccuracies in gene structure annotations can make subsequent analysis problematic. For instance, the incorrect assignment of exons to a gene, can change the protein coded for by the gene, which would have effects on the functional and evolutionary descriptions of the gene.

To address both of these needs, the Tracembler and yrGATE tools were developed. Tracembler accepts a nucleotide or protein query sequence(s) submitted by a user, performs a recursive sequence search upon the live Trace Archive databases at NCBI [1], and assembles the read sequences into contigs. Tracembler then aligns the original query sequence(s) to the contigs by local and spliced alignment, which can potentially produce a full length gene structure annotation, possibly including upstream and downstream genomic regions. Discovery of microsynteny in soybean using adjacent Medicago truncatula genes as seeds was achieved using Tracembler. yrGATE provides the ability for users to create their own gene structure annotations using high quality evidence, which can then be submitted to a community for review. yrGATE has been successfully used in re-annotating inaccurate published gene annotation sets [2] and in annotating emerging genomes [3]. yrGATE is implemented at xGDB genome browsers for Arabidopsis, rice, maize, and sorghum, among other plant species [4]. Using both of these tools, on-demand gene annotation is a reality.
A thorough genome-wide identification of U12-type introns across seven eukaryotic organisms revealed that some U12-type introns did not have accurate corresponding gene structure annotations. In these cases, gene structure annotations were created, including using the yrGATE manual annotation tool for plants. Several new discoveries regarding these introns were also made. The first instances of U12-type intron conversion to U2-type and loss were recorded in orthologous genes among vertebrates and among plants. U12-type intron sites suggesting conservation, conversion, and loss between plants and animals approximately were shown to have a 1:2:4 ratio, indicating that U12-type introns have primarily been lost since the divergence of plants and animals. Orthologs with conserved U12-type introns were found to experience reduced evolution in terms of gene structure, global protein sequence, and local protein sequence. New observations regarding the phase frequencies and exonic sequence conservation of U12-type introns were also made. The results of this investigation are presented through the CIWOG ("Common Introns Within Orthologous Genes") database, which will prove to be a useful resource for future studies regarding U12-type intron gene annotation and evolution.

In conclusion, three computational resources were developed to promote rapid, accurate and detailed gene structure annotation suited to emerging and completely sequenced genomes. An emphasis of the yrGATE and Tracembler tools was to make on-demand gene structure annotation available to the everyday researcher. U12-type introns, a particular gene feature that is difficult to annotate, were thoroughly annotated and studied in the context of orthologous genes, which yielded several new discoveries regarding their evolution. The resources developed through this research are freely available and encourage the genesis of accurate gene structure annotations.
References

3. Schlueter JA, Scheffler BE, Schlueter SD, Shoemaker RC: Sequence conservation of homeologous bacterial artificial chromosomes and transcription of homeologous genes in soybean (Glycine max L. Merr.). *Genetics* 2006, **174:**1017-1028.
ACKNOWLEDGEMENTS

First, I would like to thank my major professor Volker Brendel for his guidance, support and encouragement during a great graduate research experience. I also thank my committee members for their input and support: Thomas Peterson, Adam Bogdanove, David Fernández-Baca, and Carolyn Lawrence. I thank the Iowa State University NIH-NSF BBSI Summer Institute in Bioinformatics and Computational Biology for my initial exposure to the university and stimulating my interest in this field. For the scientific and technical collaborations, I thank Shannon Schlueter, Qunfeng Dong and Mike Brekke. Lastly, I thank Loren, Shane, and my parents for their never ending support.