Genome-wide analysis of mutually exclusive splicing

Dissertation

zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades

"Doctor rerum naturalium"

der Georg-August-Universität Göttingen

vorgelegt von

Klas Hatje

aus Göttingen

Göttingen, 2012

Betreuungsausschuss

PD Dr. Martin Kollmar (Referent)
Forschungsgruppe Systembiologie der Motor-Proteine
Max-Planck-Institut für biophysikalische Chemie, Göttingen

Prof. Dr. Burkhard Morgenstern (Co-Referent)
Institut für Mikrobiologie und Genetik, Abteilung Bioinformatik
Georg-August-Universität Göttingen

Prof. Dr. Bert L. de Groot
Forschungsgruppe Computergestützte biomolekulare Dynamik
Max-Planck-Institut für biophysikalische Chemie, Göttingen

Tag der mündlichen Prüfung: 29. Januar 2013

I hereby declare that this thesis was written independently and with no other sources and aids than quoted.

Göttingen, December 20th, 2012

Klas Hatje

Publications

Hatje K & Kollmar M (2012). A phylogenetic analysis of the Brassicales clade based on an alignment-free sequence comparison method. *Front Plant Sci* 192(3), pp. 1-12.

Hatje K & Kollmar M (2011). Predicting Tandemly Arrayed Gene Duplicates with WebScipio, Gene Duplication. Felix Friedberg (Ed.), ISBN: 978-953-307-387-3, pp. 59-76. *InTech*.

Hatje[*] K, Keller[*] O, Hammesfahr B, Pillmann H, Waack S, Kollmar M (2011). Cross-species protein sequence and gene structure prediction with fine-tuned Webscipio 2.0 and Scipio. *BMC Res Notes* 265(4), pp. 1-20.

Pillmann[*] H, Hatje[*] K, Odronitz F, Hammesfahr B, Kollmar M (2011). Predicting mutually exclusive spliced exons based on exon length, splice site and reading frame conservation, and exon sequence homology. *BMC Bioinformatics* 270(12), pp. 1-16.

Talks

2012, October
MPI Campus Seminar, Max Planck Institutes at Fassberg Campus, Göttingen, Germany
Genome-wide analysis of mutually exclusive splicing

2012, July
Bioinformatics Seminar, Göttingen, Germany
The mutually exclusive spliced exonome of *Drosophila melanogaster*

2012, June
GGNB Biomolecules Retreat, St. Andreasberg, Germany
The mutually exclusive spliced exonome of *Drosophila melanogaster*

2010, August
GGNB Biomolecules Retreat, Reinhausen, Germany
WebScipio: A web tool for gene structure prediction including alternatively spliced exons

[*] Contributed equally

Poster presentations

2012, September
German Conference on Bioinformatics, Jena, Germany
The mutually exclusive spliced exome of *Drosophila melanogaster*

2012, September
11th European Conference on Computational Biology, Basel, Switzerland
Predicting mutually exclusive spliced exons and tandem gene duplicates with WebScipio

2011, November
GGNB Science Day, Göttingen, Germany
Predicting mutually exclusive spliced exons and tandem gene duplicates

2011, October
3rd Bio-IT World Europe Conference & Expo, Hannover, Germany
Predicting mutually exclusive spliced exons and tandem gene duplicates

2011, July
19th Annual International Conference on Intelligent Systems for Molecular Biology
and 10th European Conference on Computational Biology, Vienna, Austria
WebScipio: A web tool for gene structure prediction

2010, October
2nd Bio-IT World Europe Conference & Expo, Hannover, Germany
WebScipio: A web tool for gene structure prediction

Teaching

2012, October-December
Supervision of a practical: Development of a tool to analyse sequence similarities in multiple sequence alignments

2012, October
Methods course: Protein family analysis as basis for experiments and experimental data interpretation

2012, March-September
Supervision of a practical: Development of a tool to analyse coiled-coil motif predictions

2012, March
Methods course: Protein family analysis as basis for experiments and experimental data interpretation

2011, November
Methods course: Protein family analysis as basis for experiments and experimental data interpretation

Abstract

In recent years, advances in sequencing techniques resulted in an explosive increase in sequencing data. Here, computational methods and bioinformatical analyses are presented that provide approaches to keep pace with the growing amount of data.

In the post-genomic era, an important step to derive knowledge from sequence information is to find protein-coding genes in the genomes. Scipio, a tool to reconstruct exon-intron gene structures, was improved for accurate cross-species gene reconstruction. It performed best in comparison to other tools in reconstructing the dynein heavy chain genes in the whole *Loxodonta africana* (elephant) genome based on human protein sequences. Only eleven of 1,202 exons were missed and six exons were predicted wrongly. Scipio is specialised to cope with sequencing errors and incomplete assembled genomes. The web interface WebScipio provides direct access to almost all public available eukaryotic genome sequences (December 2012: ~3,200 genome files of ~1,000 species).

Alternative splicing is a wide-spread mechanism to increase the protein inventory. About 95% of the multi-exon genes are spliced alternatively in human. A new computational method was developed to predict a special type of alternatively spliced exons, mutually exclusive exons (MXEs). In the case of mutually exclusive splicing exactly one exon of a cluster of neighbouring exons is retained in the mRNA. Those exons code for the same region in the three-dimensional structure of the protein, and therefore are predicted based on similarity and length constraints as well as compatible splice sites. The new algorithm reconstructed the MXEs in diverse genes, for example in a dynein heavy chain gene of the human parasite *Schistosoma mansoni*, in the myosin heavy chain gene of the waterflea *Daphnia magna* and in the Dscam genes of several *Drosophila* species. In addition, all but two of 28 MXEs annotated in the *Drosophila melanogaster* X chromosome were identified correctly. The algorithm was integrated int the WebScipio interface.

The continuous process of whole genome sequencing paves the way for genome-wide analyses of gene expression mechanisms like mutually exclusive splicing. The database application Kassiopeia was implemented to provide genome-wide analyses of MXEs in several organisms. It contains the mutually exclusive exomes of human, the fruit fly *Drosophila melanogaster*, eleven additional *Drosophila* species, the flatworm *Caenorhabditis elegans*, and the thale cress *Arabidopsis thaliana*. Further datasets of several species are in preparation. For each cluster of mutually exclusive exons, Kassiopeia provides EST validation data, cross-species support data, protein secondary structure predictions, and RNA secondary structure predictions. All gene annotations are searchable by BLAST and linked to organism-specific databases, like Flybase. Kassiopeia includes diverse parameters to filter the predicted exon candidates.

The detailed analysis of mutually exclusive splicing in the model organism *Drosophila melanogaster* is presented. The high-quality gene annotation of Flybase (release r5.36) was used to evaluate the quality of the prediction method. 218 of 261 annotated MXEs could be reconstructed, resulting in a sensitivity of 83.5%. The study reports 44 newly predicted exon candidates, of which five are annotated in the current release of Flybase (r5.48), eight are supported by RNA-Seq or EST data, and 29 seem to be conserved in related Arthropods.

Another algorithm was implemented that reconstructs tandem gene duplicates. Gene duplications play an important role in the origin of new genes. The algorithm is able to identify putative tandem gene duplicates which can be encoded on the forward or reverse strand or which are spread over hundreds of thousands of nucleotides. The algorithms has also been integrated into the WebScipio interface.

Meaningful evolutionary information can be derived from genomic sequences alone. An alignment-free method based on Chaos Game Representations (CGRs) was used to derive phylogentic trees of the Brassicales clade. Two algorithms, Fitch-Margoliash and Neighbour joining, and the bootstrapping method were applied to three different kinds of data: whole genome sequences, expressed sequence tag data and mitochondrial genome sequences. The methods gave reasonable results in comparison to reference trees derived from established alignment methods. The study provides a reference to evaluate further alignment-free approaches.

Table of contents

1 Introduction

In the post-genomic era, creating knowledge from genome sequence information is one of the major challenges in biology. An important step to gain biological insight is to identify parts of the genome, which encode for proteins, the molecular machines in the cell. In this work, methods to reconstruct protein-coding regions in genomes and to predict alternative exons, which lead to variations in proteins due to alternative splicing, are introduced. Another key question in biology is, how the different species living on earth have evolved. An approach is presented, which utilises genomic sequences to determine the evolutionary relations of species.

The main part of this work is composed of four publications that were published in the years 2011 and 2012, and two manuscripts that will be submitted soon. The studies are based on the experience that protein-coding regions in the genome and meaningful evolutionary information can be derived from genomic sequences. This work shows that the continuous process of whole genome sequencing paves the way for genome-wide analyses of alternative splicing mechanisms like mutually exclusive splicing, which would be unfeasible otherwise.

1.1 Background

The blueprint of living cells on earth is encoded by deoxyribonucleic acid (DNA) and it is preserved in generations of organisms by replication. The genomic DNA contains genes, the coding regions of the genome, which are transcribed into ribonucleic acid (RNA). Some genes code for functional RNA others for proteins. This work deals with protein-coding genes, which are transcribed into messenger RNA (mRNA) and then translated into proteins.

The sequencing of whole genomes made it possible to get detailed and exhaustive insight into the genetic inventory of diverse species. For years, the most common method to sequence whole genomes was the Sanger method [1]. In 1996 the first completely sequenced eukaryotic genome was published, the genome of the yeast *Saccharomyces cerevisiae* [2]. The human genome sequence was completed and published in 2001 [3]. Since then the number of human genes had been an unfeasible question. Estimations ranged from 50,000 to 100,000 in 1996, when the human genome project was started [4]. Nowadays, the number is narrowed down to 20,687 protein-coding genes and 9,640 long noncoding RNA loci [5]. In recent years new sequencing methods were developed, which allow higher throughput and have led to an exponential growth in the amount of sequencing data (Figure 1.1-1).

In eukaryotes genes are interrupted by intronic regions that do not code for proteins and are spliced out after transcription. Introns make the gene annotation a challenging task. In contrast, the demand for those annotations increases with the number of sequenced eukaryotes. The sequencing data allows genome-wide analyses based on the annotations. In addition, the

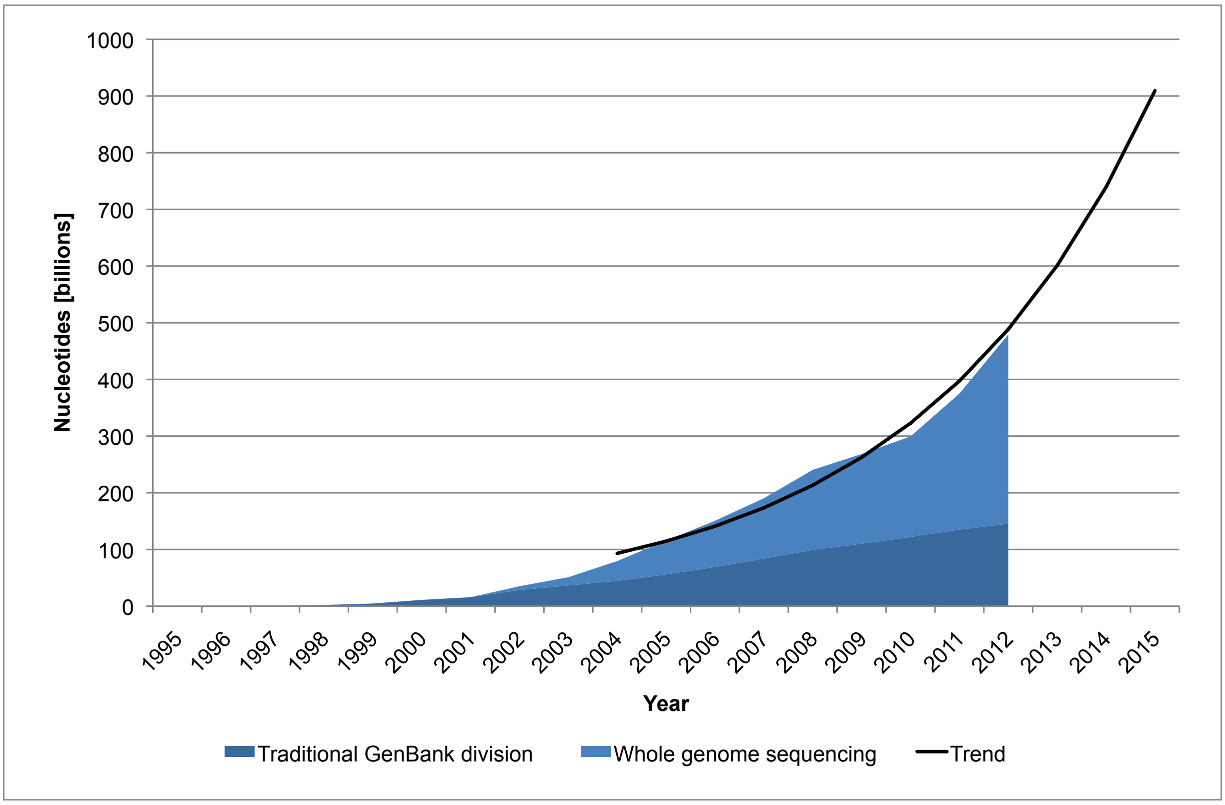


Figure 1.1-1 | Growth of sequencing data. The diagram illustrates the amount of sequencing data stored in the GenBank database of the National Center for Biotechnology Information (NCBI) from 1995 to 2012. The database divides into a traditional division and a whole genome sequencing division. An exponential trend of the data growth from 2013 to 2015 was calculated based on the numbers of the last nine years. The numbers were obtained from NCBI GenBank release notes 192 (ftp://ftp.ncbi.nih.gov/genbank/release.notes/gb192.release.notes).

sequences open up the possibility of phylogenetic analyses based on the whole genomic information.

One characteristic of life is reproduction. The living organisms conserve their blueprints by copying their DNA. During the reproduction process mutations are introduced into the DNA by chance. If a subset of a species population is isolated or partially separated, those slight variations can lead to speciation of the population, which means that a new species arises. The general aim of phylogenetics is the reconstruction of the time points of speciation events, from the first common ancestor to species that live or lived on earth.

1.2 Sequencing methods

The most commonly used method to sequence DNA was the dideoxy or chain termination sequencing method published by F. Sanger *et. al.* in 1977 [1]. This method was used to sequence several genomes, for example the human reference genome [3]. In recent years new sequencing methods were developed that are faster and less expensive. Those are called next generation sequencing (NGS) methods (reviewed in [6]). The most common NGS methods are the Illumina/Solexa [7] and the Roche/454 [8] sequencing systems. Nowadays, most eukaryotic genomes are sequenced with those NGS methods (Figure 1.2-1).

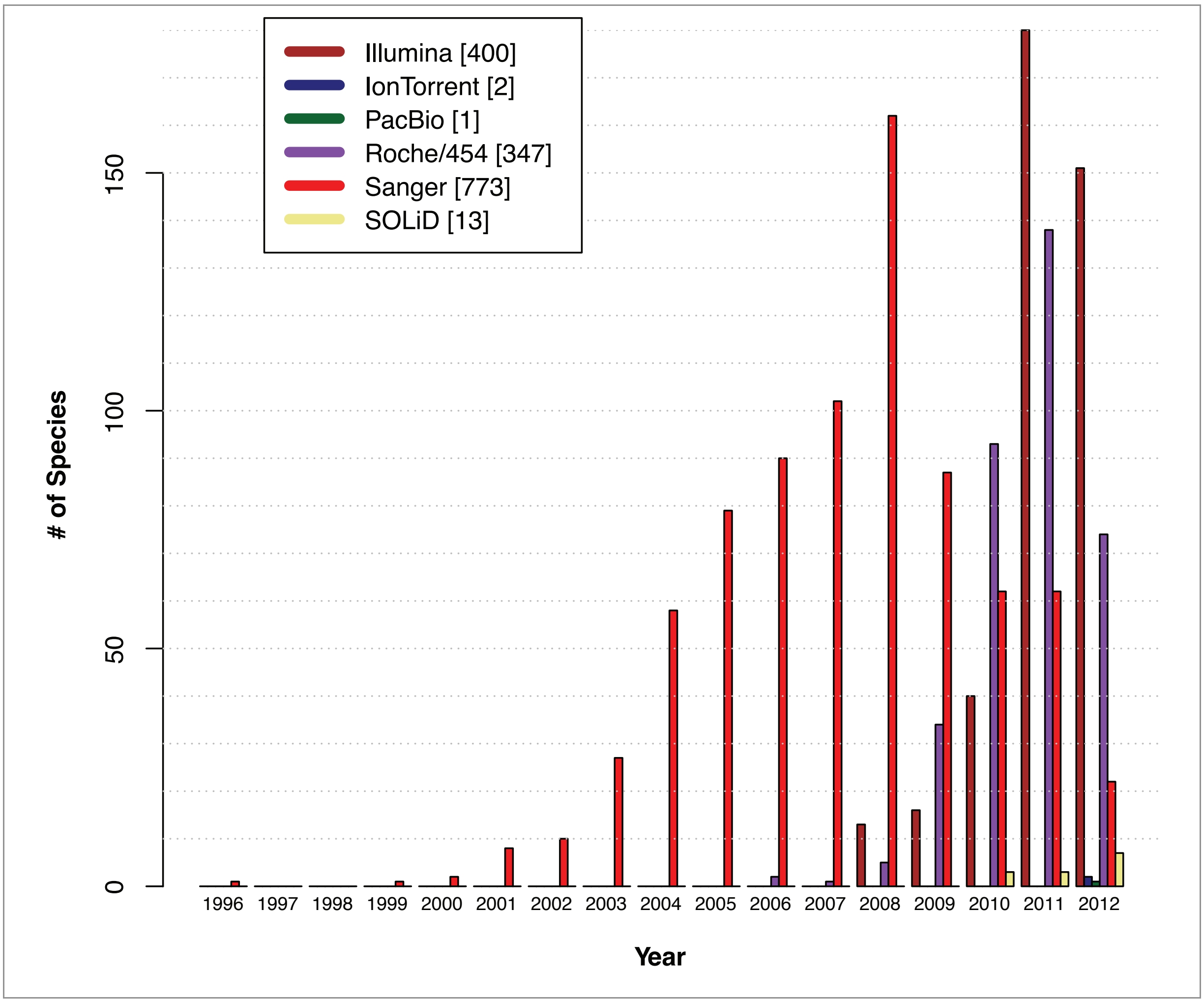


Figure 1.2-1 | Usage of sequencing methods. The diagram illustrates how many eukaryotic species were sequenced in recent years using different sequencing methods. The diagram was obtained from diArk (http://www.diark.org/diark/statistics).

NGS methods produce shorter reads (Illumina/Solexa: 50 to 250 bp; Roche/454: up to 700 bp) than the Sanger method (about 1000 bp). Short reads are difficult to assemble into long contiguous sequences. Beside whole genomes, the NGS methods also improved the possibility to sequence transcriptomes. In the RNA-Seq technique all RNA is extracted from the cell, reverse transcribed into complementary DNA (cDNA), sequenced, and mapped onto the genome [9]. RNA-Seq opens many knew insights into the transcriptome [10]. Previously, sequencing of expressed sequence tags (EST) was the method of choice to investigate mature mRNA. Here, Sanger sequencing is used to sequence one or both ends of 200 to 500 nucleotide long pieces of the mRNA molecules.[1] EST libraries contain many cDNAs that were sequenced just from the 5'-end and therefore those libraries are biased against the 5'-end of genes.

[1] http://www.ncbi.nlm.nih.gov/About/primer/est.html

The best studied transcriptomes are those of human, mouse, the fruit fly *Drosophila melanogaster* and the nematode *Caenorhabditis elegans*, which were sequenced using the RNA-Seq technique in the ENCyclopedia Of DNA Elements (ENCODE) projects. The human ENCODE project[2] was started in September 2003 to annotate all elements in the human genome, which are transcribed or which belong to transcriptional regulatory regions [11–13]. The same goal was pursued for the mouse by the mouse ENCODE project[3] [14] as well as for the model organisms *Drosophila melanogaster* and *Caenorhabditis elegans* by the modENCODE project[4] [15, 16]. The transcriptome sequencing resulted in high-quality annotations of genes in those organisms. The ENCODE project showed that three-quarter of the human genome is capable of being transcribed [17].

EST and RNA-Seq data is used to identify alternative splice forms (see section 1.4, p. 6). Here, a problem arises: It was proposed that if the sequencing coverage is increased more and more, one will find every possible splice site to be used alternatively [18, 19]. This might not mean that each of those splice variants have a function in the cell.

1.3 Gene annotation

The determination of the coding regions in a genome is done in the process of gene annotation. In the following, the term gene annotation is used to describe the task of finding genes in genomic sequences including the reconstruction of the exon-intron structures. It does not mean the functional annotation of those genes to biological processes, molecular functions, diseases or phenotypes. The term genome annotation refers to the determination of key features of the genome. This includes the annotation of genes, their products and associated biological processes [20].

Genes can be discovered by extraction of spliced RNA or by computer-based prediction methods. Gene prediction approaches divide into two major types: *Ab initio* and homology-based. *Ab initio* gene prediction methods are based on the genomic sequence alone and use statistical models for the nucleotide composition of exons in genes in contrast to introns or non-genetic regions. Those algorithms are trained by known annotations and additional biological knowledge like full-length cDNA or EST data. A comparison of *ab initio* gene annotation tools was undertaken in the EGASP project [21]. Homology-based prediction methods reconstruct genes based on the annotation of closely related species. Here, the gene annotation of a related species is mapped onto the genome sequence of interest. Gene prediction tools, which identify eukaryotic protein-coding genes are reviewed in [22].

[2] http://www.genome.gov/ENCODE

[3] http://www.mouseencode.org

[4] http://www.modencode.org

In the course of this work a homology-based approach to reconstruct the exon-intron structure of genes was developed and evaluated (see section 2.1, p. 23 or [23]). For this task the tool Scipio that reconstructs exon-intron gene structures based on a protein query sequence and a genomic target sequence was improved. Especially, the ability to reconstruct gene structures from the protein sequences of related species was developed further. This task is called cross-species search. On the one hand, the tool was extended by a new algorithm, which calculates an alignment of the protein query sequence to the genome target sequence by considering intron positions. On the other hand, the web interface was improved by a new workflow and higher configurability for cross-species searches.

The focus of this work was the improvement of cross-species searches in Scipio and its web interface WebScipio to make them usable for homology-based gene prediction. The improved version of Scipio was compared to the first version as well as to additional current gene prediction tools. We report a very good performance of Scipio in the application to cross-species gene reconstruction. Due to the exponential growth of sequencing data and the lack of quality in gene annotations, the importance to provide tools for exact gene annotation is high. Scipio and WebScipio were used in diverse studies to analyse the exon-intron structure of genes (see for example [24–28]).

Scipio uses BLAT [29] for initial gene finding and then refines the results of BLAT to reconstruct exact intron borders and to fill gaps with a Needleman-Wunsch-like algorithm [30] that calculates a spliced alignment. The web interface WebScipio gives easy access to almost all eukaryotic genomes sequenced so far, because it accesses the diArk database [31, 32]. At the moment (December 2012), about 3,200 genome files of more than 1,000 species are available. In 2012 WebScipio had more than 200 users per month. Beside Scipio, there are other tools for homology-based gene prediction using protein sequences, for example Prosplign [33], Exonerate [34] and Prot_map [35]. In comparison to other tools, Scipio performed best in reconstructing genes in whole genomes. The different tools were evaluated in reconstructing the dynein heavy chain (DHC) genes in the genome of the elephant *Loxodonta africana* using human protein sequences (Table 2.1-2, p. 48). The genes are composed of 1,202 exons in total. The new version of Scipio missed only eleven exons and predicted six wrong exons. The sensitivity of Scipio to find exons was improved from 86.1% to 93.4% compared to version 1.0 [36], and the specificity from 83.2% to 93.3%.

Protein family analyses rely on homology-based gene prediction tools. Here, homologous protein sequences in a branch of life are collected. The analysis starts with a specific protein in one species and goes on with the step-by-step reconstruction of homologous proteins in related species. All protein sequences are stored in a multiple sequence alignment. The exon-intron gene structure, which can be computed with Scipio, provides important information about the reliability of those sequences. Intron positions are well conserved, which makes it possible to validate the multiple sequence alignment using intron positions. A not-conserved intron position provides an indication for a wrong annotation.

For the initial annotation of sequenced genomes, tools are used, which calculate spliced alignments of cDNA to the genome instead of protein sequences [37]. Examples are Sim4cc [38] and Pairagon [39] which were especially designed for the cross-species case.

1.4 Alternative splicing

Alternative splicing makes it possible to derive different transcripts from a gene. This increases the variability of proteins in the cell and therefore could boost speciation events during evolution due to increases in the phenotypic complexity. In higher eukaryotes most of the protein-coding genes are alternatively spliced [40]. For example, in human up to 95% of the multi-exon genes are estimated to undergo alternative splicing [41–43].

1.4.1 The splicing process

The main actor in splicing is the spliceosome that catalyses the splicing reactions. In addition, self-splicing introns exist that are spliced without a spliceosome. The spliceosome is a complex that contains small nuclear ribonucleoproteins (snRNPs), which are composed of protein and RNA molecules. Catalytically active RNA is a special feature of the spliceosome. The main parts of the spliceosome, the snRNPs, are called subunits U1, U2, U4/U6, and U5. During splicing the snRNPs assemble to and disassemble from the pre-mRNA in a stepwise process (Figure 1.4-1), which is very flexible [44, 45]: First, U1 recognises the 5' splice site due to an RNA-RNA interaction. Then U2 binds to a branch point upstream of the 3’ splice site, and recruits U4/U6 and U5. After dissociation of U1 and U4 the spliceosome is activated and the catalytical steps to cut out the intron and join the flanking exons are processed. The splicing machinery is stabilised and regulated by several additional proteins that associate and dissociate during the splicing process. In case of long introns the spliceosome complex binds across an exon, and a rearrangement with a spliceosome on the next exon is needed to splice out the intron in between. This process is not well understood [45].

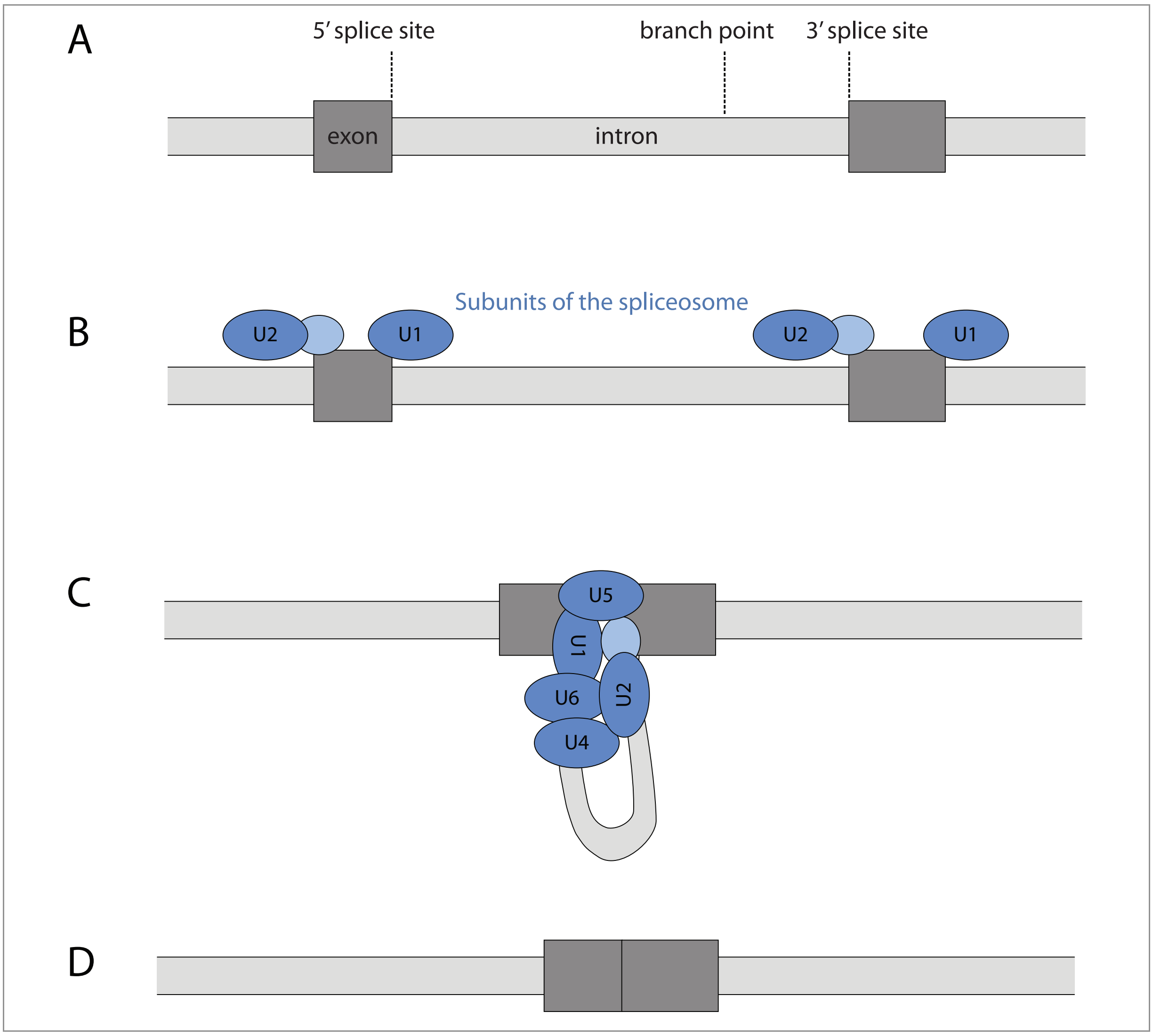


Figure 1.4-1 | The splicing process. A) Eukarytic genes are composed of exons and introns. B) The subunits of the spliceosome bind to the 5' splice site, the branch point and the 3' splice site of the intron. C) The spliceosome catalyses the splicing reaction. D) The intron is spliced out and the flanking exons are joined. This figure is based on Figure 1 of [44].

Most introns start with the two nucleotides GT at the 5’ splice site and end with the nucleotides AG at the 3’ splice site. This pattern is recognised by the major spliceosome. In addition, a second variant of the spliceosome exists. This minor spliceosome recognises only introns starting with AT at the 5’ splice site and ending with AC at the 3’ splice site, because it is composed of other snRNPs. The counterparts of the subunits U1 and U2 of the major spliceosome are called U11 and U12 in the minor spliceosome [46–48].

1.4.2 Types of alternative splicing

The process of splicing is flexible, but also well regulated. The flexibility allows exons to be spliced alternatively. Alternative splicing events can be divided into several types. Commonly used types are illustrated in Figure 1.4-2 and Figure 1.4-3.

Exon skipping / differentially included exons

In the case of exon skipping, an exon is either retained in the mRNA or it is spliced out together with the flanking introns (Figure 1.4-2A). Those exons are called differentially included exons. This is the most prevalent type of alternative splicing in higher eukaryotes. It counts for about 30% to 40% of all alternative splicing events in vertebrates and invertebrates, but for less than 5% in plants [44, 49].

Intron retention

In the case of intron retention, an intron acts either as a normal intron or is completely retained in the mRNA and so joins the flanking exons to one exon (Figure 1.4-2B). This type seems to be most prevalent in plants (~30%), fungi and protozoa, but counts only for less than 5% in vertebrates and invertebrates [44, 49].

Alternative 5’ and 3’ splice sites

Introns could have more than one start site (5’ splice site) or end site (3’ splice site). In the case of an alternative 5’ splice site the exon in front is shortened or enlarged in the mRNA (Figure 1.4-2C) and in the case of an alternative 3’ splice site the exon behind is shortened or enlarged (Figure 1.4-2D). The alternative 5’ splice sites type accounts for 18.4% of all alternative splicing events in higher eukaryotes and the alternative 3’ splice sites type for 7.9% [44].

Mutually exclusive splicing

In the case of mutually exclusive splicing exactly one exon of two or more consecutive exons is retained in each transcript (Figure 1.4-3A). The consecutive exons in a cluster of mutually exclusive exons (MXEs) exclude each other and it is not possible that the cluster is spliced out as a whole. This type seems to be rare in all organisms studied so far [40, 50].

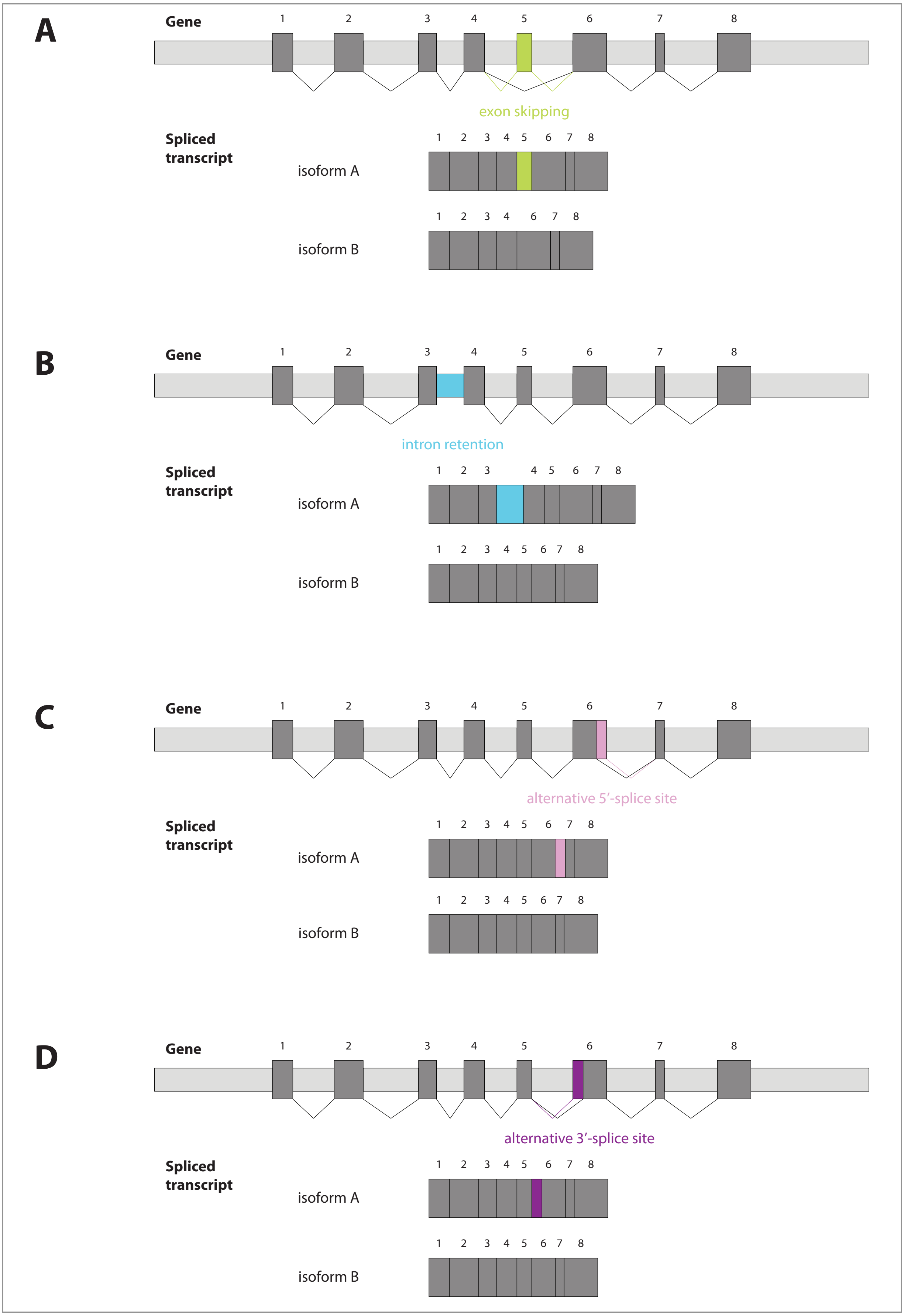


Figure 1.4-2 | Types of alternative splicing I

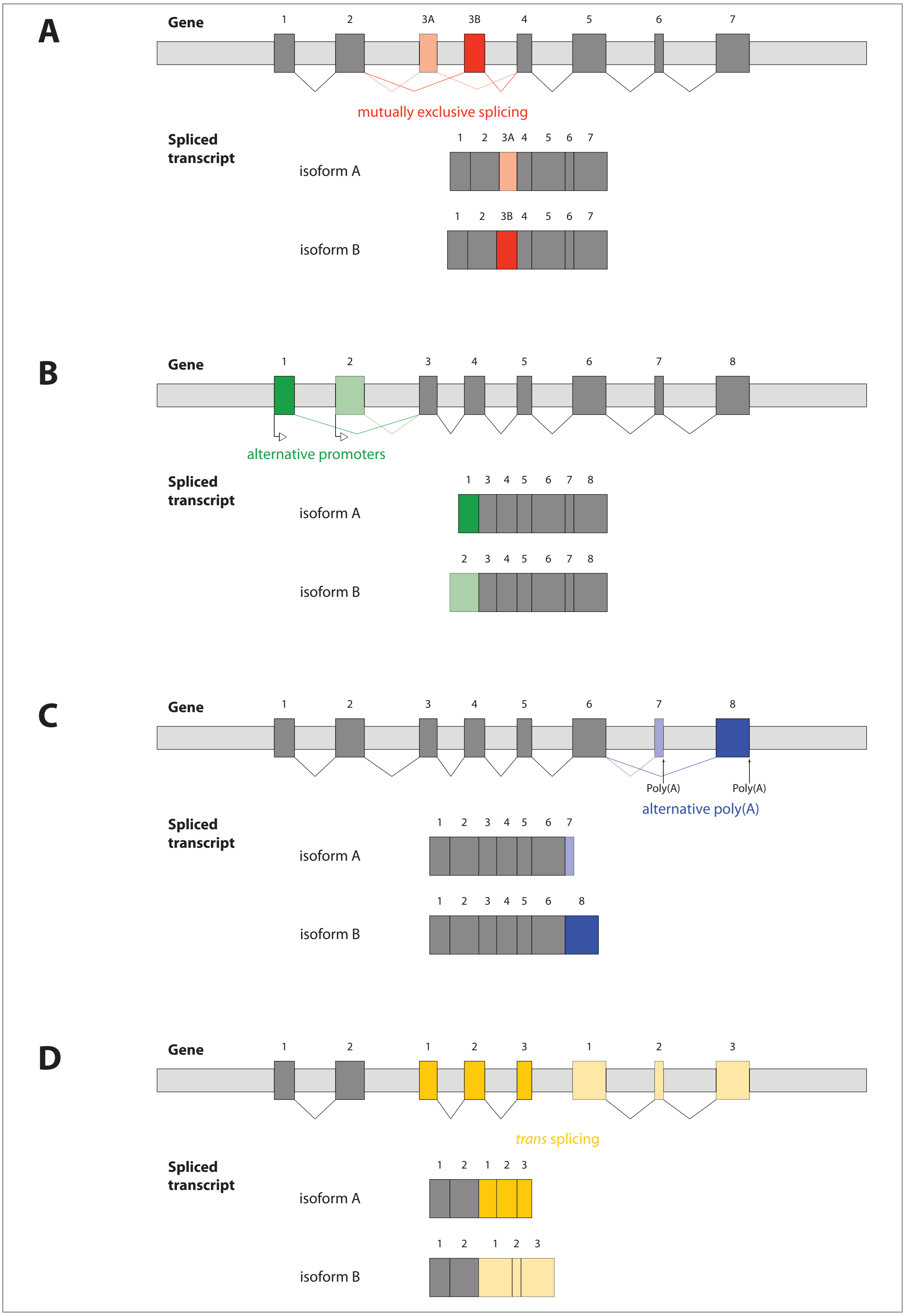


Figure 1.4-3 | Types of alternative splicing II

Multiple promoters and multiple poly(A) sites

Similar cases to mutually exclusive splicing are multiple promoters and multiple poly(A) sites. In the case of multiple promoters, the first exon of the transcript is defined by alternative promoter sites (that are start sites for transcription), leading to alternative first exons that are mutually exclusive (Figure 1.4-3B), but not spliced in the same manner as MXEs. In the case of multiple poly(A) sites the transcripts end at alternative poly(A) sites (that define the end of the transcript) due to differentially included last exons (Figure 1.4-3C).

Trans-splicing

The previous types of alternative splicing are characterised as *cis*-splicing, because the alternative exons belong to the same transcript. In the case of *trans*-splicing transcripts of different genes are expressed and than spliced to one transcript. This mechanism can also lead to alternative transcripts (Figure 1.4-3D). *Trans*-splicing seems to be an extensively used mechanism in nematodes. In the flatworm *Caenorhabditis elegans* about 70% of all genes are *trans*-spliced [51]. In contrast to this it is rare in arthropods, where 58 events are reported for the silkworm *Bombyx mori* [52] and 80 events for the fruit fly *Drosophila melanogaster* [53].

Prediction of alternative splicing

Annotating a whole eukaryotic genome is a complex task especially if alternative transcripts are considered [54]. A lot of effort is done in manually annotating alternative splicing in human [55]. It is also possible to predict alternative transcripts *ab initio*. Those tools annotate alternative splicing in selecting not only the best scoring hit, but several transcripts, which have a high score or are consistent with EST or cDNA data [56, 57]. In this work we present another method that predicts MXEs based on a different approach. It uses an initial annotation and a homology-based search algorithm to produce biological meaningful, alternative transcripts.

1.5 Mutually exclusive exons

Mutually exclusive splicing constitutes the major part of this work. This type of splicing seems to be underestimated in the literature so far, and therefore is ignored in some studies that analyse the major types of alternative splicing (see for example Figure 1 of [58]). The molecular mechanism making sure that exactly one exon of a cluster of MXEs is retained in the transcript during splicing is in dispute. Different mechanisms are proposed and were already verified for single cases [59], but there is no general mechanism that was shown to hold for the majority of mutually exclusively spliced genes. Mutually exclusive splicing plays a role in human diseases. A single mutation in a MXE of the CaV1.2 calcium channel gene leads to the Timothy syndrome [60] and a mutation in the phosphate carrier SLC25A3 gene

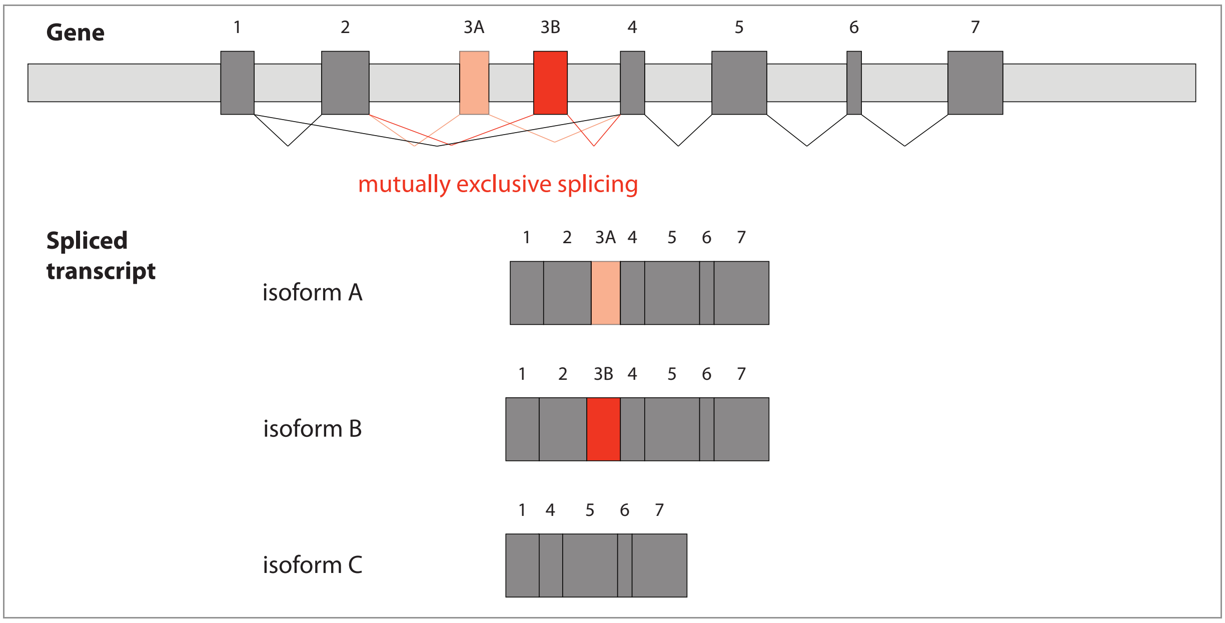


Figure 1.5-1 | Exception to the general definition of mutually exclusive exons.

leads to a myophathy [61]. The misregulation of MXEs in the glycolytic enzyme pyruvate kinase gene had been proposed to play a role in cancer formation [62–64], but a recent study did not find any evidence for an exchange in the expression of the two different isoforms during cancer formation [65].

The definition of MXEs in literature is not clear. In this work they are defined by the following characteristics: MXEs of one cluster are located consecutive in one gene. In each spliced transcript exactly one exon out of a cluster is retained in the transcript that means that no mature transcript exists that contains more than one exon of the cluster, or no exon of the cluster. These characteristics are illustrated in Figure 1.4-3 (p. 10). There is one exception: If in a spliced transcript the whole cluster of MXEs is spliced out together with a neighbouring non-MXE, the exons of the cluster are still called mutually exclusive even though mature transcripts exist that do not contain any exon of the cluster. This exception is illustrated in Figure 1.5-1. Exons in a cluster that do not contain the first exon or the last exon of the gene are called internal MXEs. Multiple promoter exons and multiple poly(A) site exons also meet the mentioned characteristics, but the splicing mechanism is very different, so they are not categorised as MXEs in this work.

1.5.1 Prediction of mutually exclusive exons

A new algorithm that predicts MXEs by finding biological meaningful transcripts was developed, evaluated and applied in genome-wide analyses. The algorithm is integrated into WebScipio to make it accessible and easy to use. The algorithm, its application to sample genes and its evaluation on the whole X chromosome of *Drosophila melanogaster* are described in [66], which is part of this work (section 2.2, p. 55). The study revealed a high sensitivity of the prediction, because all MXEs in the example genes could be reconstructed, like in the myosin heavy chain gene of *Daphnia pulex* with 9 clusters [67] or in the Dscam gene including up to 98 MXEs in *Drosophila virilis* [68–70]. In the whole X chromosome of *Drosophila melanogaster* all but two of the 28 MXEs were recognized.

The main result is that mutually exclusive splicing can be predicted with traceable criteria and it was shown that the prediction algorithm is applicable to the whole genome scale. The next logical step was to apply the algorithm to the whole *Drosophila melanogaster* genome, and genomes of species in other branches of the tree of life.

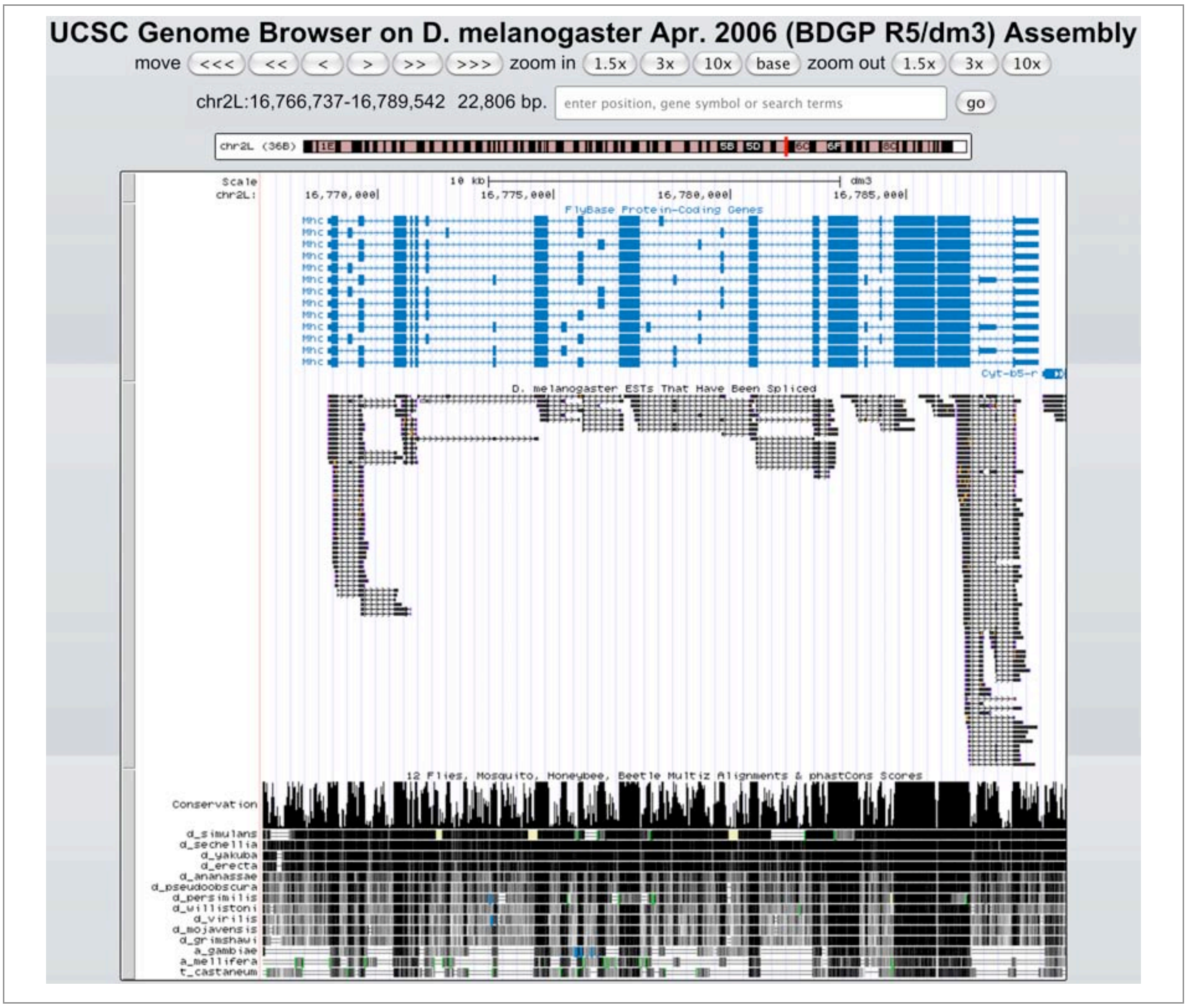
1.5.2 Genome-wide analysis of mutually exclusive exons

The prediction algorithm is parameterised. Different parameters result in different predictions. If less restricted parameters are used, more already known annotations can be reconstructed resulting in a higher sensitivity, but also more false positive predictions are introduced resulting in a lower specificity. We could determine reasonable parameters from the application of the prediction algorithm to some example genes and the X chromosome of *Drosophila melanogaster*. Those parameters needed further evaluation. Very low parameters were used during the search in the whole fruit fly genome to examine the limits of these parameters with respect to the sensitivity. All predictions were stored in a database. A corresponding web application was developed to analyse the results based on different parameters that can be chosen after the prediction process. This application, called Kassiopeia, is able to store the genome-wide analyses of mutually exclusive splicing in different organisms and to make those accessible. The development of Kassiopeia was part of this work (section 3.1, p. 119).

Kassiopeia

One could imagine to use an already established tool like the UCSC genome browser [71, 72] instead of developing a new application. This genome browser is a popular tool used in many web applications to visualise annotations of genomes as shown in Figure 1.5-2 for the myosin heavy chain gene of *Drosophila melanogaster*. It allows adding multiple annotation tracks that contain position specific information related to the genome. Examples are the positions of exons as well as expression patterns or sequence conservation in different species.

In the case of the Kassiopeia database we could not follow this approach and decided to develop a new application based on the WebScipio source code. The main reason was the demand to allow filtering of the MXE candidates after the prediction process (Figure 3.1-2, p. 124). To our knowledge, this is not possible in any tool published so far. Each gene entry in



UCSC Genome Browser on D. melanogaster Apr. 2006 (BDGP R5/dm3) Assembly

chr2L:16,766,737-16,789,542 22,806 bp.

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
enter position, gene symbol or search terms
go
chr2L (36B)
Scale
chr2L:
10 kb
dm3
16,770,000
16,775,000
16,780,000
16,785,000
FlyBase Protein-Coding Genes
Mhc
Cyt-b5-r
D. melanogaster ESTs That Have Been Spliced
12 Flies, Mosquito, Honeybee, Beetle Multiz Alignments & phastCons Scores
Conservation
d_simulans
d_sechellia
d_yakuba
d_erecta
d_ananassae
d_pseudoobscura
d_persimilis
d_willistoni
d_virilis
d_mojavensis
d_grimshawi
a_gambiae
a_mellifera
t_castaneum

Figure 1.5-2 | UCSC genome browser. The figure shows the genomic region of the myosin heavy chain (Mhc) gene of *Drosophila melanogaster* in the UCSC genome browser (http://genome.ucsc.edu). The annotation tracks FlyBase Genes, Spliced ESTs and Conservation are selected.

Kassiopeia is linked to other tools and databases to make their data easily accessible. The *Drosophila* genes are linked to the corresponding Flybase[5] [73] entry, the modENCODE data in the UCSC genome browser[6] and to WebScipio[7].

Drosophila melanogaster

We chose the *Drosophila melanogaster* genome for the first genome-wide analysis and prediction of MXEs. Since the first classical genetic experiments with fruit flies by Thomas Hunt Morgan in 1908, *Drosophila melanogaster* developed to one of the best-analysed model organisms for genetic studies. The annotation of its genes is in an advanced state, due to cDNA

[5] http://flybase.org
[6] http://flybase.org/cgi-bin/gbrowse/dmelrnaseq
[7] http://www.webscipio.org

sequencing, whole genome sequencing of closely related *Drosophila* species, transcriptome sequencing using RNA-Seq and additional computational methods. In addition, *Drosophila melanogaster* was the main object in the modENCODE project[8] [15, 74]. This makes it possible to validate our prediction approach with reliable annotations.

Compared to human and mouse, which also have high-quality annotations, the fruit fly genome contains shorter introns making the analysis less complex and the prediction more robust. To evaluate the sensitivity of our prediction method, the most important advantage of *Drosophila melanogaster* in contrast to the model organisms *Arabidopsis thaliana* and *Caenorhabditis elegans* is that a lot of mutually exclusive splicing events were already reported:

- *Drosophila melanogaster*: 102-251 events [40, 50, 74]
- *Arabidopsis thaliana*: 3-4 events [40, 50]
- *Caenorhabditis elegans*: 30-55 events [40, 50, 75]
- *Homo sapiens*: 124-212 events [40, 50]

Twelve *Drosophila* species

Our prediction pipeline was applied to eleven additional *Drosophila* species besides *Drosophila melanogaster* (section 3.1, p. 119). This enables the analysis of the evolution of MXE clusters. The main result was that these clusters evolved very fast in the past 50 million years (section 3.2, p. 137). The mechanism seems to be frequently inserted in a wide range of genes. The analyses of the other *Drosophilas* also showed how accurate the predictions are for species that do not have a good gene annotation.

Arabidopsis thaliana

Intron retention is the most prevalent type of alternative splicing in plants, in contrast to exon skipping in Metazoa [44, 49]. Mutually exclusive splicing events seem to be very rare in plants [50] and overlooked by some studies up to know as in [76]. In the model organism *Arabidopsis thaliana* three to four events of mutually exclusive splicing were reported [40, 50] and 14 events are annotated in release 10 of The Arabidopsis Information Resource (TAIR) database [77]. Based on this release, our prediction pipeline found 99 internal MXE candidates (section 3.1, p. 119). Therefore, we expect the number of mutually exclusive splicing events in plants to be underestimated.

[8] http://www.modencode.org

Caenorhabditis elegans

Another model organism that has an accurate gene annotation is the namatode *Caenorhabditis elegans*. So far 30 to 55 events of mutually exclusive splicing were reported [40, 50, 75] and 35 are annotated in the WormBase release 230. Based on this release, our predictions suggest 283 internal MXE candidates (section 3.1, p. 119).

Homo sapiens

The organism of highest interest in science is the human. This results in an accurate annotation of human genes, the basis of our prediction pipeline. In the human genome less MXEs are annotated than in the *Drosophila melanogaster* genome[9], even though the total number of alternative splicing events is much higher [40]. In human 124 to 212 events of mutually exclusive splicing are reported [40, 42, 50].

1.5.3 Mechanisms of mutually exclusive splicing

Different molecular mechanisms were proposed, which make sure that exactly one exon of a MXE cluster is retained in the mature transcript. The first two mechanisms hold only for two MXEs in a cluster. In the first mechanism the intron between the two exons is so short that the two subunits U1 and U2 of the spliceosome cannot bind to the intron at the same time, they inhibit each other in binding due to steric interference (Figure 1.5-3A). If the intron in between cannot be spliced out, the two neighbouring exons are spliced in a mutually exclusive manner. This mechanism was shown for example in the alpha-tropomyosin of human [78].

Eukaryotic cells include two different types of spliceosomes [46–48], the major one contains subunits U1 and U2, and the minor one contains subunits U11 and U12. The subunits of the different spliceosomes are not compatible to each other and bind to different sequence motifs at the intron 5’ and 3’ splice sites. If the intron between the two MXEs has a 5’ splice site that can be bound by the major spliceosome (subunit U1) and a 3‘ splice site that can be bound by the minor spliceosome (subunit U12), it is not possible to splice this intron out [79]. The same holds for the contrary case as shown in Figure 1.5-3B.

[9] http://www.motorprotein.de/kassiopeia

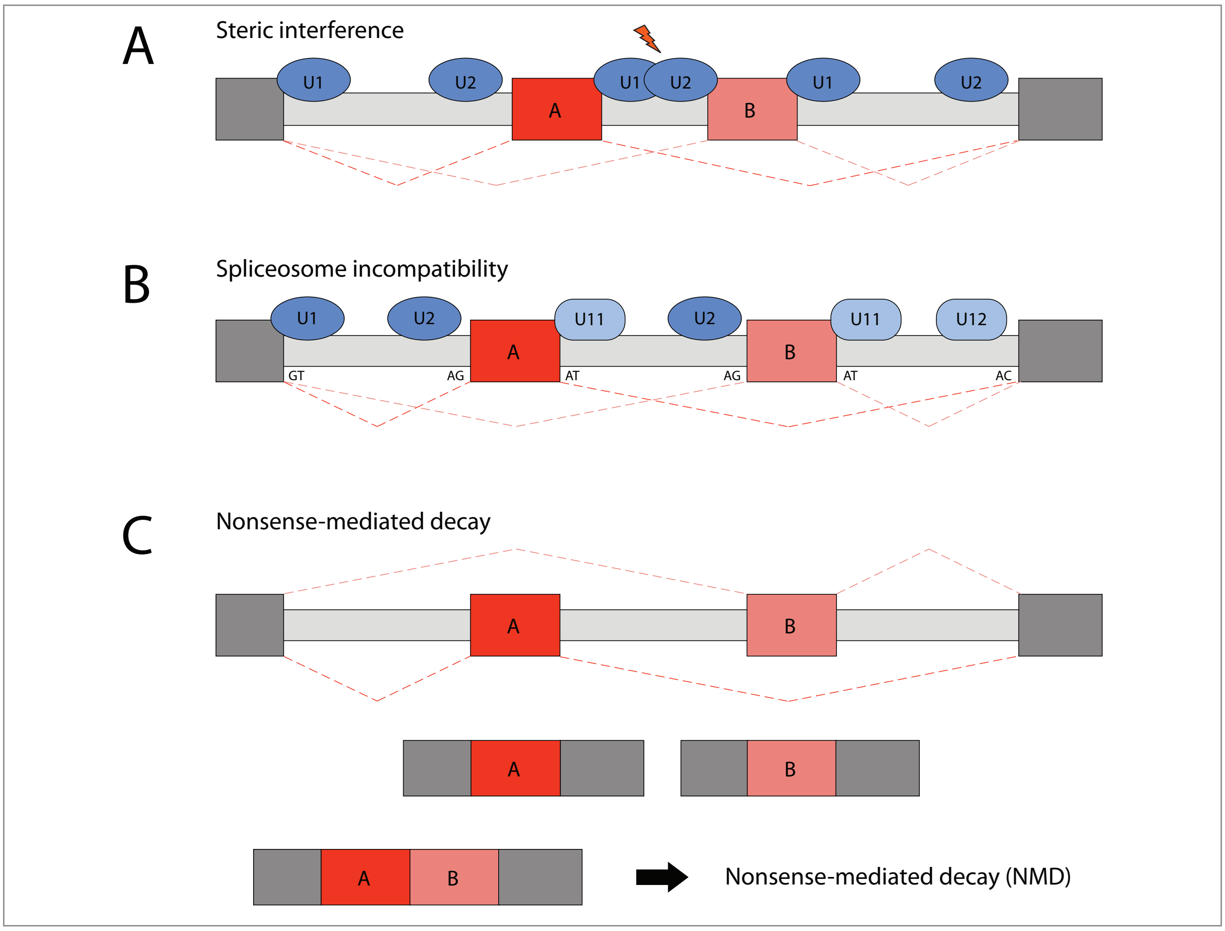


Figure 1.5-3 | Mechanisms of mutually exclusive splicing. The figure illustrates three molecular mechanisms (A-C), which lead to mutually exclusive splicing. This figure is based on Figure 1 of [59].

Another mechanism prevents two or more MXEs from being spliced into one meaningful transcript. Here, if two exons of the cluster are retained in the transcript, the second exon includes a shift in the reading frame, which leads to a premature stop codon in the mRNA sequence, and the mRNA is degraded in a process called nonsense-mediated decay (Figure 1.5-3C). This mechanism was for example found in the human fibroblast growth factor receptor 2 (FGFR2) gene [80].

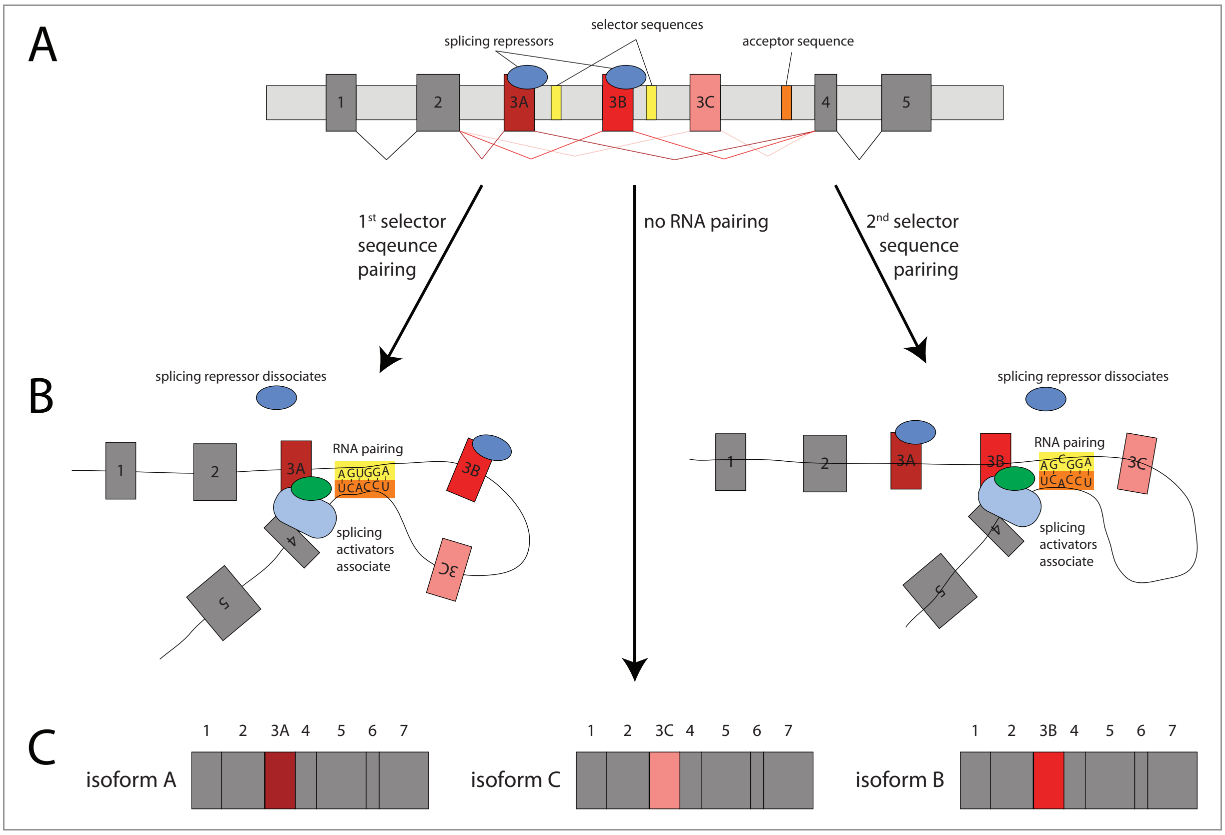


Figure 1.5-4 | RNA secondary structure in mutually exclusive splicing. A) The figure shows a gene including three MXEs: 3A, 3B and 3C. The exons 3A and 3B are bound by splicing repressors, which cause the exons to be spliced out. B) Sequence motifs following exon 3A and exon 3B (selector sequences) could pair with a complementary motif following exon 3C (acceptor sequence). The RNA pairings lead to loops in the transcript, which allow binding of splicing activators, and lead to dissociation of the splicing repressors. C) All exons, but the activated exon, are spliced out. This figure is based on Figure 8a of [81].

A fourth mechanism was shown for the down syndrome cell adhesion molecule (Dscam) gene of *Drosophila melanogaster*, and is proposed as well for other genes and organisms including mammals [68, 69, 81, 82]. In this mechanism the RNA secondary structure of the transcript plays an important role (Figure 1.5-4). Conserved sequence motifs were found in the introns between MXEs, which could bind to a complementary motif in the preceding or following intron of the MXE cluster. The competing binding sites lead to different loops in the RNA secondary structure. These loops activate the neighbouring MXEs by releasing a splicing repressor that is bound to each exon of the cluster. The exons in the loop are spliced out, because the whole loop is spliced out, and the other MXEs are spliced out due to the splicing repressors. The conserved sequence motifs were found in the Dscam exon 6 cluster [68, 69]. The heterogeneous nuclear ribonucleoprotein hrp36 could be detected to be the corresponding repressor [83]. It was shown that these RNA binding sites really play a role in the splicing of the exon 6 cluster *in vivo* [84]. Later those complementary motifs could also be detected in different Arthropods for the clusters 4 and 9 of Dscam [81, 85], for the 14-3-3ζ gene [81, 86] and for the myosin heavy chain gene [81]. One goal of the genome-wide analyses is to find those complementary elements that form a RNA secondary structure in the reconstructed clusters of MXEs.

1.6 Tandem gene duplications

The prediction algorithm, which searches for MXEs, can be applied to the up- and downstream region of a gene. This allows finding MXE candidates if there is an additional noncoding exon in front of or behind the gene, or if only a fragment of the gene is annotated. The analysis of those predictions showed that many of those candidates in the up- and downstream regions belong to gene duplications and are not members of MXE clusters. This led to a new algorithm, which uses similar criteria to find tandemly arrayed gene duplicates [87] and is part of this work (section 2.3, p. 79).

This homology-based algorithm is able to reconstruct several consecutive gene duplicates, to cope with intron losses and gains, and to report the completeness of the gene reconstruction. The algorithm is integrated into WebScipio to make it accessible and easy to use. The artificial fusion of two genes is a common problem in the automatic annotation of genes (see Additional data file 1 of [88]). Scipio is susceptible to fuse tandem gene duplicates as shown for the human muscle myosin in Figure 2.3-6 (p. 92). Those duplicates can easily be reconstructed using the new algorithm.

The major result is that the approach to reconstruct tandem gene duplicates by searching for homologous exons was successful. The genome-wide application of this algorithm was already computed for several species. The next step will be the integration of this data into the Kassiopeia interface, and detailed, genome-wide analyses of tandem gene duplicates.

1.7 Phylogeny

There is a great potential to get new insights from the large amount of sequence data, which are accessible nowadays. The analysis of this data is lacking behind the pace of sequencing. A big challenge is the finding of genes as well as functional annotation of the genes in the next step. Another potential not fully tapped is the reconstruction of the tree of life based on this large amount of genome sequences that contain not only the blueprint of the species, but also a lot of evolutionary information.

Phylogenetic trees are reconstructed from differences between species of interest. The general assumption is that more distant species separated earlier in the evolution. There are a lot of criteria for measuring the differences between species. The major ones are morphologic and genetic differences. A common approach for the calculation of phylogenetic tree is to collect representative protein sequences that appear in all species of interest, align them and calculate phylogenetic trees based on distance, maximum likelihood, maximum parsimony or Bayesian methods (reviewed in [89]). This approach incorporates two challenges: The protein sequences must be available for each species in full length and they must be aligned properly.

To overcome the alignment problem we used an alignment-free method to derive phylogenetic trees from sequence. In addition, the method compares whole genome sequences and is not dependent on single proteins. The application and evaluation of this approach to a branch of the plants, the Brassicales clade, is described in [90], which is part of this work (section 2.4, p. 99). In our study differences in Chaos Game Representation (CGR) pictures generated from genomic sequences were used, to derive phylogenetic trees. An advantage of this method is that people are able to retrace the magnitude of the differences between those pictures. It is not possible to compare whole genomes by just looking at their sequences. CGRs were already used to reconstruct the phylogeny of 20 birds [91] and of 26 eukaryotes using their mitochondrial genomes [92]. Furthermore, the approach was used for HIV-I sub-typing [93]. Alignment-free sequence comparison methods are reviewed in [94].

The general question is how to use the ever-increasing amount of sequencing data directly to derive a detailed picture of the tree of life. At the moment it is computational expensive to align whole genomes and it does not make sense for distant relatives. An alignment-free method that is comprehensible as well as easily interpretable, and incorporates the whole sequence information would be very convincing. Using CGRs or word counts is a first step into this direction. The advantage of the visual representation of the CGR method is that it clearly shows the differences and therefore it is human interpretable. Our study showed that the CGR method is an appropriate method to reconstruct phylogenetic trees from very divergent data sources, but needs further improvements.

1.8 Computational approaches

In this work, new approaches to decipher the blueprint of life and to reconstruct the tree of life are presented. These approaches are based on the ever increasing amount of sequencing data, and try to tap the full potential of this data to answer basic biological questions. The presented methods are computer-based approaches. The explosion in the amount of sequencing data and the ever increasing computational power make it possible to clarify biological questions systematically on a genome-wide scale, and across species. To evaluate the biological insights it is important that the data is accessible and refined for non-computer scientists, and that newly developed methods are user-friendly and convenient. In our studies, the processed data, the scientific results and the developed algorithms are provided through web interfaces that are accessible worldwide, highly configurable, and whose usage is straightforward.

The presented computational approaches generate diverse requirements for software, libraries and frameworks that are necessary during development and execution. Ruby on Rails[10] is used as a framework for the web applications. A PostgreSQL database[11] constitutes the data back-end. The operating system of the development server and the production servers is Ubuntu Linux[12]. Most of the algorithms are implemented in the Ruby programming language[13] or if run time is important in C/C++[14]. The Scipio script is written in Perl[15]. Genome-wide prediction pipelines are parallelised to be executed on computer clusters with multiple processors. The implemented tools make extensive use of the BioRuby [95] and the SeqAn [96] libraries. The YAML file format[16] is mainly used to store structured result data and the SVG file format[17] is used to visualise the results.

Despite the increase of computational possibilities to solve biological questions, the performance of lab-based methods to understand the blueprint of life also increases as shown in the ENCODE[18] and modENCODE[19] projects. Nowadays, those methods need to be complemented by computer-based methods that cope with the huge amount of data produced in the experiments. The big challenge is to interpret these data.

[10] http://rubyonrails.org
[11] http://www.postgresql.org
[12] http://www.ubuntu.com
[13] http://www.ruby-lang.org
[14] http://www.stroustrup.com/C++.html
[15] http://www.perl.org
[16] http://www.yaml.org
[17] http://www.w3.org/Graphics/SVG
[18] http://www.genome.gov/ENCODE
[19] http://www.modencode.org

2 Publications

2.1 Cross-species protein sequence and gene structure prediction with fine-tuned Webscipio 2.0 and Scipio

Klas Hatje[1,*], Oliver Keller[1,*], Björn Hammesfahr[1], Holger Pillmann[1], Stephan Waack[2] and Martin Kollmar[1]

1 Abteilung NMR-basierte Strukturbiologie, Max-Planck-Institut für biophysikalische Chemie, Am Fassberg 11, D-37077 Göttingen, Germany
2 Institute of Computer Science, University of Göttingen, Goldschmidtstr. 7, 37077 Göttingen, Germany
* Contributed equally

BMC Research Notes 2011, 4(265)

Published: 28 July 2011
doi: 10.1186/1756-0500-4-265
This article is available from http://www.biomedcentral.com/1756-0500/4/265

2.1.1 Abstract

Background

Obtaining transcripts of homologs of closely related organisms and retrieving the reconstructed exon-intron patterns of the genes is a very important process during the analysis of the evolution of a protein family and the comparative analysis of the exon-intron structure of a certain gene from different species. Due to the ever-increasing speed of genome sequencing, the gap to genome annotation is growing. Thus, tools for the correct prediction and reconstruction of genes in related organisms become more and more important. The tool Scipio, which can also be used via the graphical interface WebScipio, performs significant hit processing of the output of the Blat program to account for sequencing errors, missing sequence, and fragmented genome assemblies. However, Scipio has so far been limited to high sequence similarity and unable to reconstruct short exons.

Results

Scipio and WebScipio have fundamentally been extended to better reconstruct very short exons and intron splice sites and to be better suited for cross-species gene structure predictions. The Needleman-Wunsch algorithm has been implemented for the search for short parts of the query sequence that were not recognized by Blat. Those regions might either be short exons, divergent sequence at intron splice sites, or very divergent exons. We have shown the benefit and use of new parameters with several protein examples from completely different protein families in searches against species from several kingdoms of the eukaryotes. The performance of the new Scipio version has been tested in comparison with several similar tools.

Conclusions

With the new version of Scipio very short exons, terminal and internal, of even just one amino acid can correctly be reconstructed. Scipio is also able to correctly predict almost all genes in cross-species searches even if the ancestors of the species separated more than 100 Myr ago and if the protein sequence identity is below 80%. For our test cases Scipio outperforms all other software tested. WebScipio has been restructured and provides easy access to the genome assemblies of about 640 eukaryotic species. Scipio and WebScipio are freely accessible at http://www.webscipio.org.

2.1.2 Background

Whole genome sequences of eukaryotes are generated with increasing speed [97]. While the focus at the beginning of high-throughput DNA sequencing was on model organisms and the human genome, for which tremendous amounts of secondary data was available, the aims have shifted to organisms of medical or economic relevance (e.g. *Plasmodium falciparum* [98] or *Phytophthora ramorum* [99]), to the comparative analysis of entire taxa (e.g. the Drosophila clade [100] or Candida species [101]), and, very recently, to organisms of evolutionary interest (e.g. *Trichoplax adhaerens* [102] or *Volvox carteri* [103]). However, gene catalogues are only available for a small part of the sequenced organisms and a precise and complete set of genes is still unavailable for even a single species. In the first instance the gene annotation is done with automatic gene prediction programs that either predict only isolated exons, or reconstruct the complete exon-intron structures of the protein-coding genes, or even try to predict 5’ and 3’ untranslated regions. *Ab-initio* gene prediction programs only use the assembled DNA sequences as input, having precomputed models for nucleotide distributions, while evidence-based programs consider alignments of ESTs, cDNAs, or annotated sequences from closely related organisms, with the target sequence (reviewed in [104]). The highest accuracy is reached by programs that combine model-based and alignment-based approaches [105, 106].

For many biological applications like the phylogenetic analysis of a protein family (e.g. [88]) or the comparative analysis of the exon-intron structure of a certain gene from different species (e.g. [107]), it is necessary to obtain translated transcripts of homologs of closely related organisms or the reconstructed exon-intron patterns of the genes, respectively. The protein sequences of homologs of a certain protein can be obtained in several ways. Annotations based on *ab-initio* gene predictions, sometimes supplemented by EST data, are available for about half of the sequenced eukaryotic genomes, although it is often tedious to find the corresponding data via the FTP-pages of the sequencing centers. In addition, automatic predictions are not complete and in many cases not correct. For very few eukaryotes, full-length cDNA data can be accessed. However, these data never cover the complete transcriptome of the species. Another possibility is to manually annotate the protein homologs in the genomes of choice by comparative genomics. This is certainly the most accurate way. By this approach a multiple sequence alignment of as many as possible homologs is created, and based on this sequence alignment mispredicted sequence regions (insertions and missing regions) are easily detected. Further homologs are added by manual inspection of the corresponding genomic DNA regions and manual reconstruction of intron splice sites. Splice sites are in most cases conserved throughout the eukaryotes [108] and therefore their position and frame can be used for gene reconstructions by comparing gene structures from known and to be annotated genes.

To assist in the task of the manual annotation of eukaryotic genomes, and to provide options for genomes for which gene prediction data is not available, we have recently developed Scipio [36, 109]. Scipio is a post-processing script for the Blat output [29] and maps a protein sequence to a genomic DNA sequence. Blat has been developed for the fast alignment of very similar DNA or protein sequences. However, Blat is not able to identify very short exons (two or three amino acids, or exons of just the N-terminal methionine), it is not able to assemble genes spread on more than one contiguous DNA sequence, it misses exons that are too divergent, it does not apply biological sequence models to determine exact splice site locations on nucleotide level, or to distinguish introns from insertions caused by frameshifts or in-frame stop codons [36, 109, 110]. Scipio is able to address most of these issues resulting in considerably improved gene structure reconstructions [36, 109]. Its initial intention was to cope with sequencing errors, to assemble genes from highly fragmented genome assemblies, and to reconstruct intron splice sites. Scipio was not able to correctly reconstruct very short exons or to correctly reconstruct genes in cross-species searches if these were not highly identical.

Here, we present the fundamentally improved version of the Scipio software that has been extended for the use in cross-species searches. In addition, very short exons and divergent regions at intron borders are now correctly reconstructed. Scipio can be used via the web-interface WebScipio that provides access to 2111 genome assembly files for 592 species (end of February 2011).

2.1.3 Methods

The presented software consists of two programs that form a pipeline for the output of the external program Blat, which is executed first. The Blat results are post-processed by the Scipio script written in Perl[20]. WebScipio provides a graphical user interface for Scipio that we have developed using the web framework Ruby on Rails[21,22]. The workflow was optimized to direct the user to the necessary input parameters. This was implemented with the technique of Asynchronous Javascript and XML (AJAX). Visual effects were realized with the help of Prototype[23] and script.aculo.us[24] that are JavaScript libraries, which are integral parts of Ruby on Rails.

Scipio

The Scipio Perl script itself, which can also be run standalone, has undergone numerous extensions that are based on our extensive experience in manual gene annotation [88, 111][25]. The general setup of the script that aimed to handle all the various sequencing and assembly errors has already been described [36]; here, we present an implementation of the Needleman-Wunsch algorithm which is the main extension to the previous version.

The Needleman-Wunsch Algorithm used in Scipio

In the updated version of Scipio, we use a modified Needleman-Wunsch style dynamic programming (DP) algorithm to perform an exhaustive search for the best-scoring spliced alignment between the query and target sequence fragments that were left unmatched by Blat. Like the original Needleman-Wunsch algorithm, it calculates an optimal global alignment between the sequences, but it is adjusted to find an optimal *spliced* alignment between a protein query sequence s and a genomic target sequence t. Given the computational cost of $|s|$ and $|t|$, it is executed only on very short sequence fragments s and t. We introduce different categories of penalties depending on the type of matching. Any alignment can be represented by a *parse* Φ: a collection of pairs of strings (s_1, t_1), …, (s_k, t_k), such that the aligned sequences are the concatenations: $s = s_1 \ldots s_r$, $t = t_1 \ldots t_r$. A *penalty score* $p(s_k, t_k)$ is assigned to each pair as follows:

- if s_k is a single residue and t_k a string of length 3 (codon), then $p(s_k, t_k) = p_{\mathrm{MAP}}(s_k, t_k)$ is a *match/mismatch* penalty:

[20] http://www.perl.org
[21] http://www.ruby-lang.org
[22] http://rubyonrails.org
[23] http://www.prototypejs.org
[24] http://script.aculo.us
[25] http://www.cymobase.org

$$p_{\mathrm{MAP}}(s_k, t_k) = \begin{cases} 0, & \text{if } t_k \text{ translates to } s_k \\ p_{\mathrm{MISM}}, & \text{if not} \end{cases}$$

- an *insertion* penalty $p(s_k, t_k) = p_{INS}$ is assigned to them if s_k is a single residue and t_k is empty
- a *gap* penalty $p(s_k, t_k) = p_{GAP}$ is assigned to them if t_k is a codon and s_k is empty
- a *frameshift* penalty $p(s_k, t_k) = p_{FS}$ is assigned to them if t_k consists of 1 or 2 nucleotides, and s_k is empty or a single residue

To cover the case of introns, in addition we define intron penalties based on the donor and acceptor splice sites:

$$p_{\mathrm{INTRON}}\left(n_1 \ldots n_\ell\right) = p_{\mathrm{DSS}}\left(n_1 n_2\right) + p_{\mathrm{INT}} + p_{\mathrm{ASS}}\left(n_{\ell-1} n_\ell\right)$$

with a constant value p_{INT} for any sequence of nucleotides $n_1 \ldots n_\ell$ and zero splice site penalties if $n_1 n_2$ = "GT", and $n_{\ell-1} n_\ell$ = "AG". We distinguish two cases: in-frame introns, and introns splitting codons:

- if s_k is empty and t_k exceeds the minimum intron length, then $p(s_k, t_k) = p_{INTRON}(t_k)$ is the *intron* penalty
- if s_k is a single residue $n_1 n_2 n_3$, and $t_k = n_1 \omega n_2 n_3$, or $t_k = n_1\ n_2 \omega n_3$ with single residues and ω a string exceeding the minimum intron length, then the penalty is a combined match/intron penalty: $p(s_k, t_k) = p_{INTRON}(\omega) + p_{MAP}(s_k, n_1 n_2 n_3)$. Here, two different penalties are defined (depending on the frame of the intron), and thus the minimum of them is taken.

If (s_k, t_k) does not satisfy any of these conditions, no penalty is defined resulting in an invalid parse. By combining insertions, deletions, and frameshifts, there is always some valid parse for any given pair of sequences. The cost of a parse Φ is the sum of the penalties: $p(\Phi) = p(s_1, t_1) + \ldots + p(s_r, t_r)$, and we calculate

$$p\left(s,t\right) = \min\left\{p\left(\Phi\right) \mid \text{is a valid parse aligning } s \text{ and } t\right\}$$

by computing the DP matrix (M_{ij}) containing the minimal score for an alignment of the subsequences $s_{[0..j\text{-}1]}$ and $t_{[0..i\text{-}1]}$, using the following recursions:

$$M_{ij} = \min \begin{cases} M_{(i-3)(j-1)} + p_{\mathrm{MAP}}\left(s_{[j-1]}, t_{[i-3,i-2,i-1]}\right), \\ M_{i(j-1)} + p_{\mathrm{INS}}, \\ M_{(i-3)j} + p_{\mathrm{GAP}}, \\ \min\left\{M_{(i-1)j}, M_{(i-2)j}, M_{(i-1)(j-1)}, M_{(i-2)(j-1)}\right\} + p_{\mathrm{FS}}, \\ \min\limits_{i' \le i-\ell_{\min}} \left\{M_{i'j} + p_{\mathrm{INTRON}}\left(t_{[i'..i-1]}\right)\right\}, \\ \min\limits_{i' \le i-\ell_{\min}-3} \left\{M_{i'(j-1)} + p_{\mathrm{MAP}}\left(s_{[j-1]}, t_{[i']} t_{[i-2,i-1]}\right)\right\} + p_{\mathrm{INTRON}}\left(t_{[i'+1..i-3]}\right), \\ \min\limits_{i' \le i-\ell_{\min}-3} \left\{M_{i'(j-1)} + p_{\mathrm{MAP}}\left(s_{[j-1]}, t_{[i',i'+1]} t_{[i-1]}\right)\right\} + p_{\mathrm{INTRON}}\left(t_{[i'+2..i-2]}\right) \end{cases}$$

where each of these expressions corresponds to one of the possible penalty types for the last segment of the parse.

The last three lines cover introns, one for each reading frame, with $l_{\min}$ denoting the minimum intron length. To avoid having to iterate over all values for *i'* in these cases, we precompute nine variants of the score matrix with partial intron penalties added (indexed by a nucleotide *n* if it splits a codon) as follows:

$$\begin{aligned} M^{(0)}_{i,j} &= \min_{i' \le i-\ell_{\min}} \left\{M_{i'j} + p_{\mathrm{DSS}}\left(t_{[i',i'+1]}\right)\right\} + p_{\mathrm{INT}} \\ M^{(1)}_{i,j,n} &= \min_{\substack{i' \le i-\ell_{\min}-3 \\ t_{[i']}=n}} \left\{M_{i'(j-1)} + p_{\mathrm{DSS}}\left(t_{[i'+1,i'+2]}\right)\right\} + p_{\mathrm{INT}} \\ M^{(2)}_{i,j,n} &= \min_{i' \le i-\ell_{\min}-3} \left\{M_{i'(j-1)} + p_{\mathrm{MAP}}\left(s_{[j-1]}, t_{[i,i'+1]} n\right) + p_{\mathrm{DSS}}\left(t_{[i'+2,i'+3]}\right)\right\} + p_{\mathrm{INT}} \end{aligned}$$

Note that *n* denotes the nucleotide before the intron in $M^{(1)}$, and the nucleotide after it in $M^{(2)}$. The latter contains already the mismatch penalty, while the former does not. With *i'* the latest segment start allowed ($i' = i - l_{\min}$ for an intron scored by $M^{(0)}$, and $i' = i - l_{\min} - 3$ for a codon split by an intron), the intron variables are given recursively by

$$\begin{aligned} M^{(0)}_{i,j} &= \min\left\{M_{i'j} + p_{\mathrm{DSS}}\left(t_{[i',i'+1]}\right) + p_{\mathrm{INT}}, M^{(0)}_{i-1,j}\right\} \\ M^{(1)}_{i,j,n} &= \min\left\{M_{i'(j-1)} + p_{\mathrm{DSS}}\left(t_{[i'+1,i'+2]}\right) + p_{\mathrm{INT}}, M^{(1)}_{i-1,j,n}\right\} & (n = t_{[i']}) \\ M^{(1)}_{i,j,n} &= M^{(1)}_{i-1,j,n} & (n \ne t_{[i']}) \\ M^{(2)}_{i,j,n} &= \min\left\{M_{i'(j-1)} + p_{\mathrm{MAP}}\left(s_{[j-1]}, t_{[i,i'+1]} n\right) + p_{\mathrm{DSS}}\left(t_{[i'+2,i'+3]}\right) + p_{\mathrm{INT}}, M^{(2)}_{i-1,j,n}\right\} \end{aligned}$$

and then replace the last three lines in the recursion for M_{ij}:

$$M_{ij} = \min \begin{cases} \ldots, \\ M_{i,j}^{(0)} + p_{\mathrm{ASS}}\left(t_{[i-2,i-1]}\right), \\ \min\limits_{n=a,c,g,t} \left\{ M_{i,j,n}^{(1)} + p_{\mathrm{MAP}}\left(s_{[j-1]}, n\, t_{[i-2,i-1]}\right)\right\} + p_{\mathrm{ASS}}\left(t_{[i-4,i-3]}\right), \\ M_{i,j,n}^{(2)} + p_{\mathrm{ASS}}\left(t_{[i-3,i-2]}\right) \end{cases}$$

The penalties for the Needleman-Wunsch algorithm can be adjusted manually in the Scipio command-line version but not via the WebScipio web-interface. The penalties need to be well balanced so that the Needleman-Wunsch search does not result for example in a number of artificial short exons where a long exon is missing due to a gap in the genome assembly. Based on extensive tests with in-house test data we set the following values as default: mismatch-penalty: 1.0; insertion-penalty: 1.5; gap-penalty: 1.1; frameshift-penalty: 2.5; intron-penalty: 2.0 + the respective penalties for donor and acceptor splice sites.

WebScipio

At present, the web interface offers 2272 genome files of 643 eukaryotic organisms. Metadata corresponding to the species, like assembly versions, sequencing centers, and assembly coverage, is available from the diArk database [32]. WebScipio reads the metadata out of a periodically updated text file generated from diArk, or queries the diArk database directly with SQL.

The gene structure schemes resulting from the Scipio run are generated and displayed in the Scalable Vector Graphics (SVG) format[26]. This allows scaling the graphics while retaining their resolution and to show tooltips generated with JavaScript and HTML for each element of the gene structure schemes. For browsers not supporting SVG, a fallback solution is implemented, which uses the Portable Network Graphics (PNG) format. The PNG files are generated by Inkscape[27].

Internally, the sequence data is processed with the help of BioRuby [95]. Results are saved in the YAML format[28], but are also available for download in the GFF format. The web application runs the Blat and Scipio jobs in the background, which was implemented using the Rails plug-in Workling[29] in combination with Spawn[30]. The server-side stored session data is increasing with every extension of WebScipio. To make the session storage fast, flexible, and

[26] http://www.w3.org/Graphics/SVG

[27] http://inkscape.org

[28] http://www.yaml.org

[29] http://github.com/purzelrakete/workling

[30] http://github.com/tra/spawn

scalable we use a database backend called Tokyo Cabinet[31]. It offers a simple key-value store, also called hash store, for accessing different data objects with the help of a unique key for each object. Tokyo Tyrant is the network interface to Tokyo Cabinet and allows storing data across the network on several servers. It is used in WebScipio for scalability reasons.

External Tools

We use Hoptoad[32] for error reporting. It is a web application that collects errors generated by WebScipio, aggregates them to the detailed error reports for developer review, and sends email notifications. We use a behaviour-driven testing strategy to validate the functionality and behaviour of WebScipio. For the automation of these tests we use RSpec[33], which is a behaviour-driven development framework for the Ruby programming language. Our intention for this test implementation was the need of reliability and accuracy within the continuously extended software. Application tests are run with Selenium[34], a test system for web applications. This offers the opportunity to test the web-interface as a whole. Selenium integrates into the Mozilla Firefox browser as a plug-in that records the user interaction in the form of a Ruby script. To run the test scripts without user-interaction, Selenium starts and controls the browser automatically. We integrated the user-interface tests into our automated test environment as additional RSpec test cases.

2.1.4 Results and Discussion

Scipio and WebScipio workflow, and general parameters for fine-tuning gene predictions

The general workflow of Scipio and WebScipio is similar to that described previously [36, 109]. Scipio provides some general search parameters that filter the Blat output for further post-processing, and offers several expert options that influence the post-processing steps. In the new Scipio version, especially the part of the gap-closing (mapping the parts of the query sequence to the target sequence that Blat failed to recognize) and hit extension (modelling the regions at exon borders, including terminal exons, where homology was too low to be identified by Blat) has been improved (Figure 2.1-1, see also Supplementary information 2.1-1). This has been done by implementing the Needleman-Wunsch algorithm for the search of unmapped query sequence in respective target regions and by introducing parameters that

[31] http://fallabs.com/tokyocabinet
[32] http://hoptoadapp.com
[33] http://rspec.info
[34] http://seleniumhq.org

allow a higher divergence from the exon border regions predicted by Blat. All new parameters are adjustable by the user although the default values should be good enough for most cases. However, especially when searching for very divergent homologs or when searching for homologs of very divergent species, these parameters might need manual adaptation. Figure 2.1-1 shows a detailed scheme of the Scipio workflow including all parameters that can manually be adjusted. Also, some of the most important decisions are outlined that Scipio makes to provide the best possible result. The detailed scheme should allow the experienced user to fine-tune the search in especially difficult cases. The rationale for implementing each of the parameters and its consequences are explained below.

The new web-interface

Because we wanted to offer most of the new parameters to the experienced user via the web-interface WebScipio, and we planned to introduce searches for alternatively spliced exons, we had to redesign the WebScipio workflow. The goal was to keep it well structured, intuitive and clear. We have also improved the usability for new and less experienced users by providing more examples, help pages, and documentation. The general design of selecting one target sequence for the search for multiple query sequences has been retained. Next, the experienced user can adjust many of the Scipio variables, and, also at this stage, many of the parameters for searches for alternative exons (those parameters are described elsewhere). We provide some default values for cross-species searches that are based on our experience in working with and knowledge about eukaryotic genomes [88]. For example, some genomes are known to contain only small numbers of introns while others are known to contain only short introns. Special settings for cross-species searches are provided for several specific taxa but the default cross-species parameters should be applicable for most genomes. Having selected a specific set of parameters every single parameter can still be adjusted individually.

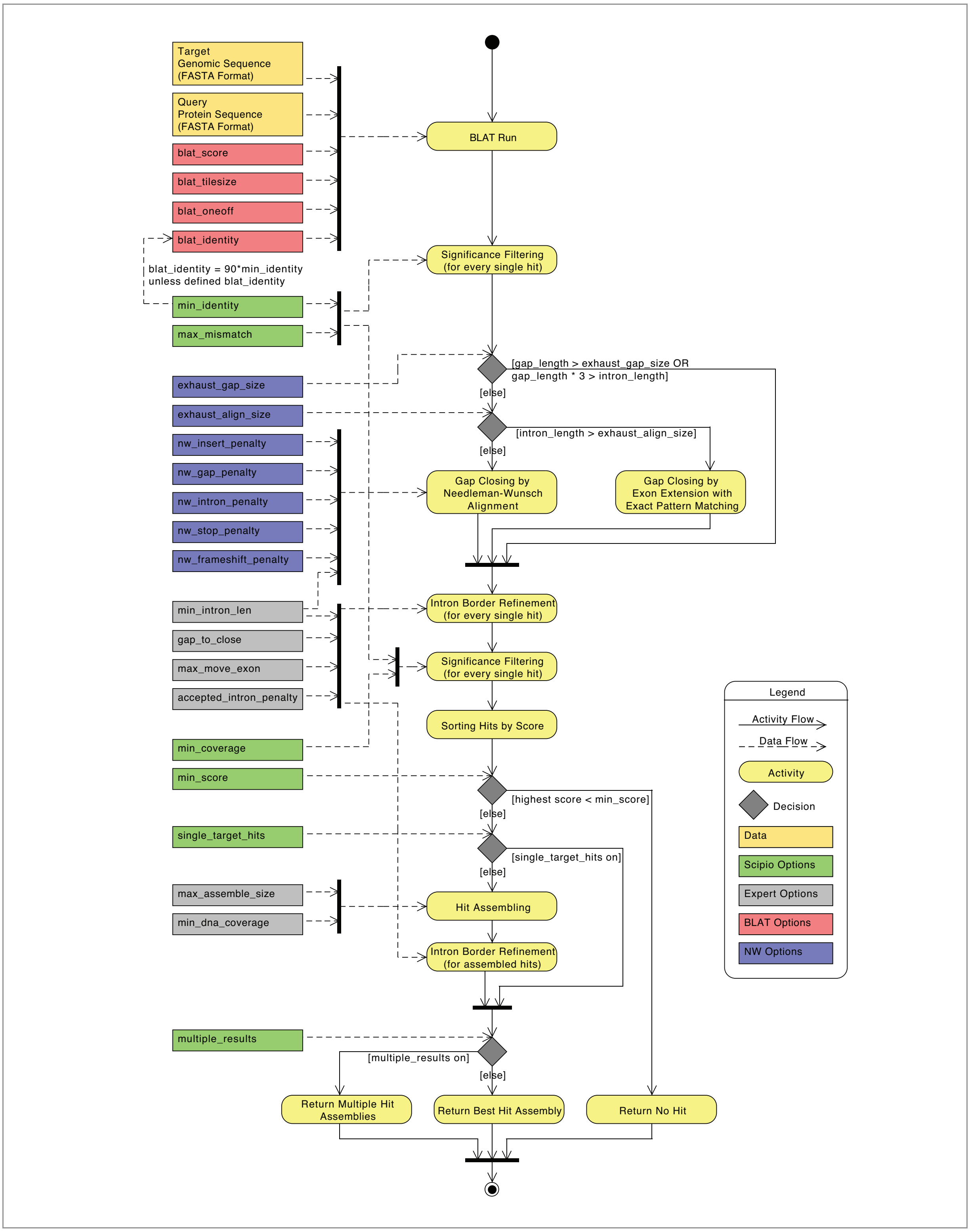
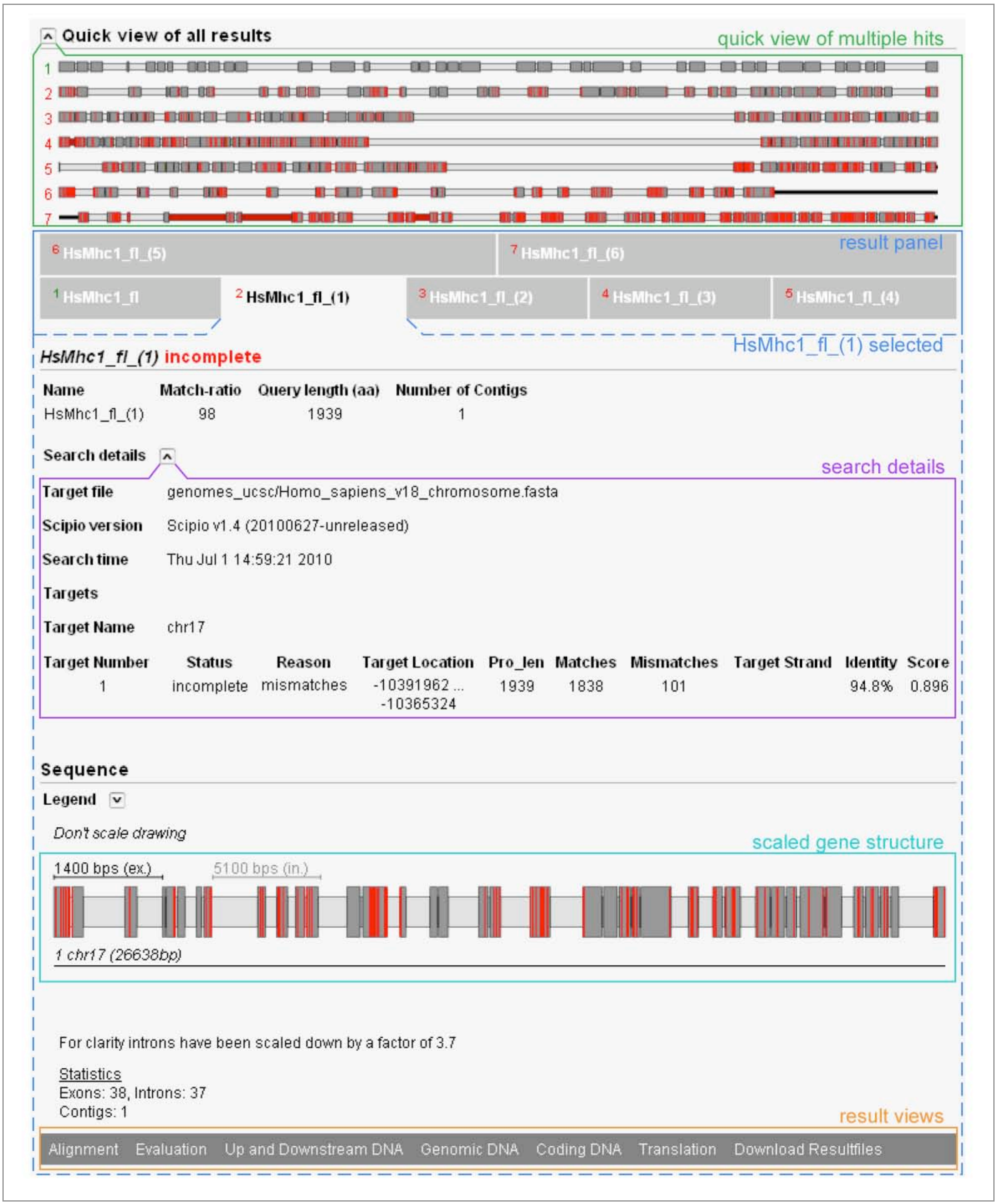


Figure 2.1-1 | The extended Scipio workflow. This diagram depicts the activity and data flow of a Scipio run. Scipio needs a protein and a target genome sequence, both in FASTA format, as input to start a Blat run. Every single Blat hit is subsequently processed and filtered, and assembled in the case of hits on multiple targets. The gap_length describes the number of amino acids of an unmatched query subsequence. The intron_length is the corresponding length of the unmatched target subsequence in nucleotides.

As before, the most important result view is the scheme of the exon-intron structure of the search result. In this scheme, all information regarding the quality of the result (complete versus incomplete, containing gaps, i.e. unmatched parts of the query sequence, questionable introns, mismatches, frame-shifts, in-frame stop-codons, etc.) is included. Opening the "Search details" box provides further information concerning the search parameters, and additional data regarding the aligned query sequence is available from the different result views.

Due to gene and whole genome duplications during eukaryotic evolution there are often two or more closely related homologs of a certain protein per genome. This might cause some problems for cross-species searches if the paralogs in the target genome are about equally homologous to the query sequence. Therefore, we implemented a --multiple_results parameter. Switching --multiple_results off is the best way to get the exact gene structure for an intra-species search. Switching --multiple_results on (default setting in cross-species searches) allows retrieving all possible results depending on the general search parameters (like --min_score or --min_identity, Figure 2.1-2). If multiple hits are found they will be listed separately and can be analysed using the various result views. In addition, we implemented a quick view showing the gene structure schemes as a fast overview. As example for the benefit and limitation of this parameter, we searched for class-II myosin heavy chain homologs in humans (Figure 2.1-2). It is known that vertebrate genomes contain several muscle myosin heavy chain genes (belonging to the class-II myosin heavy chains) that are specialised for certain tissues like heart muscles or skeletal muscles [88]. Six of these genes are encoded in a cluster [112]. The example search shows the gene structure corresponding to the query sequence (*Hs*Mhc1_fl) and the gene structures of six homologs of varying degree of divergence. While the closest homolog (*Hs*Mhc1_fl_(1)) only contains mismatches compared to the query sequence, the three next closest homologs have severely deviating gene structures. They contain very long introns in the middle of the genes indicating that they are mixed genes assembled from the N-terminal half from one gene of the muscle myosin heavy chain cluster and the C-terminal half taken from the following gene of the cluster. The next two homologs are already very divergent so that parts of the genes cannot be reconstructed leaving many and long gaps.



HsMhc1_fl_(1) incomplete

Name	Match-ratio	Query length (aa)	Number of Contigs
HsMhc1_fl_(1)	98	1939	1

Search details

Target file	genomes_ucsc/Homo_sapiens_v18_chromosome.fasta
Scipio version	Scipio v1.4 (20100627-unreleased)
Search time	Thu Jul 1 14:59:21 2010
Targets	
Target Name	chr17

Target Number	Status	Reason	Target Location	Pro_len	Matches	Mismatches	Target Strand	Identity	Score
1	incomplete	mismatches	-10391962 ... -10365324	1939	1838	101		94.8%	0.896

Sequence

Legend

Don't scale drawing

For clarity introns have been scaled down by a factor of 3.7

Statistics
Exons: 38, Introns: 37
Contigs: 1

Figure 2.1-2 | Screenshot of the multiple results view of WebScipio. The screenshot shows the result of the search for multiple homologs of one of the muscle class-II myosins from human in the human genome. The search parameters were --min_identity = 60%, --max_mismatch = infinite, and --multiple_results = yes to get as many homologs as possible. On top, the opened quick view of all reconstructed gene structures is shown. Next, a panel with the different results is shown. Green numbers mark complete results (100% of the query sequence reconstructed) while red numbers mark incomplete results (might contain gaps, mismatches, frameshifts, etc.). Result hit number 2 was selected and shows the result for the closest homolog to the query sequence with no gaps (unmapped query sequence) but 101 mismatches.

Use of WebScipio to produce publication-quality figures of gene structures

WebScipio can be used to easily produce publication-quality figures of gene structures. Either, these figures can be produced in the described way, or the user can upload an own genomic DNA sequence for use as target sequence. This is interesting in the case that the whole genome sequence is not known but only the genomic sequence of a certain region. SVGs can be downloaded and further processed in many graphics programs.

New general and expert search parameters

The parameters --min_score (previously: --best_size), --min_identity, and --max_mismatch have already been described [109] and define the threshold for the Blat hits to be processed by Scipio. To reduce or even abolish the artificial assembly of contigs that by chance contain some identical residues we have introduced the parameter --min_coverage that applies to every single Blat hit. The coverage is the number of mapped residues (as match or mismatch) divided by the query length of the (possibly partial) hit. By default, Scipio rejects hits with coverage of less then 60%.

In addition to these general parameters we have introduced several expert options most of which will be described in detail below. One of the parameters is --transtable that allows the user to specify a non-standard translation table, for the use with species like Candida species, *Tetrahymena thermophila* and others that would otherwise lead to mismatches. Another parameter called --accepted_intron_penalty is used to define valid splice sites. By default, GT---AG and GC---AG are accepted, whereas, for example, introns with the pattern AT---AC would be classified as doubtful (“intron?”). By adjusting the --accepted_intron_penalty parameter those introns will also be accepted instead of defining those introns as “intron?”.

Parameters to account for additional/missing bases in predicted exons

Gene homologs even from very closely related species are often too divergent to be completely identified by Blat. While the core building block of the proteins and the functional sites are often strongly conserved, low homology is especially found at the surface of the proteins. Thus, loop regions are often sites of amino acid substitutions, insertions of long stretches of residues, and deletions. In addition, since the terminal regions of most proteins are at the surface, they are also often very divergent. Short stretches of nucleotides whose lengths are multiples of three and whose translations do not result in any in-frame stop codons are most likely to be insertions rather than true introns.

A parameter --min_intron_len has been implemented to distinguish introns from insertions, with a default minimum intron length of 22 nucleotides. A minimum intron length of 22 nucleotides is a rather conservative estimate given the minimum intron length of 35-40 nucleotides based on a test set of about 17,000 introns of genes of 10 model organisms [113]. Thus, by default additional coding sequence for up to seven amino acids (= 21 nucleotides) will be treated as exon sequence and joined with the surrounding exons into a single exon.

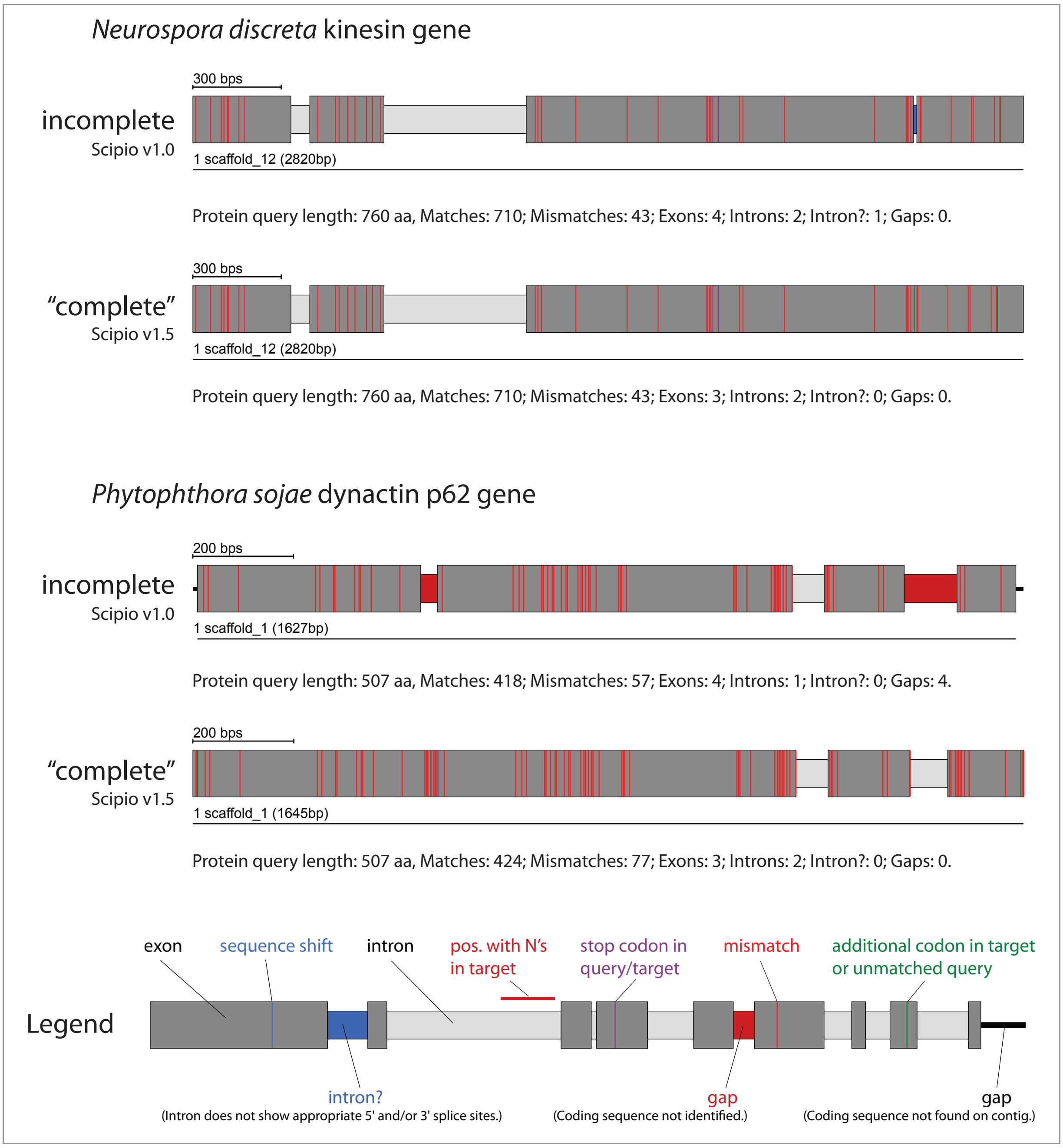

Neurospora discreta kinesin gene
incomplete
Scipio v1.0
300 bps
1 scaffold_12 (2820bp)
Protein query length: 760 aa, Matches: 710; Mismatches: 43; Exons: 4; Introns: 2; Intron?: 1; Gaps: 0.
"complete"
Scipio v1.5
300 bps
1 scaffold_12 (2820bp)
Protein query length: 760 aa, Matches: 710; Mismatches: 43; Exons: 3; Introns: 2; Intron?: 0; Gaps: 0.
Phytophthora sojae dynactin p62 gene
incomplete
Scipio v1.0
200 bps
1 scaffold_1 (1627bp)
Protein query length: 507 aa, Matches: 418; Mismatches: 57; Exons: 4; Introns: 1; Intron?: 0; Gaps: 4.
"complete"
Scipio v1.5
200 bps
1 scaffold_1 (1645bp)
Protein query length: 507 aa, Matches: 424; Mismatches: 77; Exons: 3; Introns: 2; Intron?: 0; Gaps: 0.
Legend
exon
sequence shift
intron
pos. with N's in target
stop codon in query/target
mismatch
additional codon in target or unmatched query
intron?
(Intron does not show appropriate 5' and/or 3' splice sites.)
gap
(Coding sequence not identified.)
gap
(Coding sequence not found on contig.)

Figure 2.1-3 | Modelling of additional/missing bases in gene. Case A shows the result of the search of a kinesin from *Neurospora crassa* (query sequence) in *Neusrospora discreta* (target sequence) using the old and the new Scipio version. The --min_intron_len parameter has been set to 22. Case B shows the result of a search of the dynactin p62 homolog from Phytophthora ramorum (query sequence) in *Phytophthora sojae* (target sequence). To get the correct gene prediction the following Scipio parameters have been used: --min_identity = 60%, --min_score = 0.3, --max_mismatch = ∞, --gap_to_close = 15, --min_intron_length = 22. The colour coding is explained in the legend and applies to all gene structure figures. For further information see Supplementary information 2.1-2.

The opposite case of extra amino acids in the query is dealt with by the parameter --gap_to_close. By default, a mapping of up to six additional amino acids from the query sequence to the exon borders will be enforced at the cost of further mismatches, in order to eliminate a gap (of unmatched query sequence). This parameter also effects the modelling of the intron borders (see below). Figure 2.1-3 shows two examples of cross-species searches in which the target sequence contains additional or less amino acids in conserved exons. Case A shows the results of a search for a kinesin homolog from *Neurospora crassa* (query sequence) in the closely related organism *Neurospora discreta* (target sequence, see also Supplementary information 2.1-2). Because of the relatively high homology of the two sequences, Blat has already retained the additional residues of the query sequence so that they are included in the result of the old Scipio version. However, a questionable intron (called intron? in Scipio) was introduced in the region that contained additional nucleotides in the target sequence leading to missing residues in the target translation. With the new parameter --min_intron_len these additional nucleotides are correctly treated as exonic sequence. Case B shows an example of two divergent homologs of the dynactin p62 gene of *Phytophthora ramorum* (query sequence) and *Phytophthora sojae* (target sequence, see also Supplementary information 2.1-2). These two homologs contain a long divergent region with many consecutive mismatches in the first exon that is not identified by Blat and introduces a long gap of unmatched residues. In addition, the N- and C-termini have divergent sequences and different lengths. With the new parameters, Scipio can correctly model the target gene.

Parameters to identify divergent exons and very short exons ignored by Blat

To identify exons that contain too many mismatches to be identified by Blat, and to correctly annotate very short exons, the Needleman-Wunsch algorithm described above forces an alignment of unmatched query sequence to spare target sequence. Very short exons of one to four amino acids are only reconstructed if they are identical to the query sequence and contain valid splice sites while short exons of five to seven amino acids are also often correctly reconstructed if they contain mismatches between query and target sequence (e.g. in cross-species searches). The maximal lengths of query and target sequence fragments to be aligned with Needleman-Wunsch are controlled by the parameters --exhaust_align_size and --exhaust_gap_size, respectively. By default, the exhaustive search is restricted to query gaps of 21 amino acides (three times the default Blat tilesize), since we expect Blat to successfully discover at least parts of any longer exons, and to a target subsequence of 15,000 bps. The restriction of the latter value is caused by the exponentially increased run time with increased target subsequence so that for example the potentially very long introns in mammalian genomes are only searched after manual increase of this value. Other parameters affecting the Needleman-Wunsch algorithm, such as the penalties mentioned above, can be adjusted by the command line version only, and not via WebScipio. However, the default values have extensively been tested with in-house data and should not require changes in most if not all cases.

The effect of the new parameters on the search results is demonstrated with the examples shown Figure 2.1-4 (see also Supplementary information 2.1-2). In case A, the human dynactin p50 gene contains two very short exons of 3 and 2 amino acids. These two short exons are conserved in all vertebrates (B. Hammesfahr and M. Kollmar, unpublished data). Case B shows the coronin gene from the basidiomycote fungi *Puccinia graminis* encoding a short 3 amino acid exon (Figure 2.1-4, see also Supplementary information 2.1-2). In addition, the codons at the exon/intron junctions of this short exon are split. In most of the other basidiomycotes sequenced so far, this short exon is part of one of the neighbouring exons, or part of a longer exon that includes both neighbouring exons. However, it also exists in the basidiomycote *Melampsora laricis-populina*. Thus, this short exon is not an artificial creation but a true exon. Case C presents the dynactin p150 gene that contains three short exons of 7, 6, and 7 amino acids at the beginning of the gene (Figure 2.1-4, see also Supplementary information 2.1-2). Even with the Blat-tilesize set to 5 those exons are not recognized in the search against the chromosome assembly. This example best demonstrates the effect of the --exhaust_align_size (default setting 15,000 bps) and the --exhaust_gap_size (default setting 21 aa) parameters to completely reconstruct the respective part of the gene. At the 3'-end of the p150 gene, there is another very short exon that shows some homology to the beginning of the preceding intron and is therefore added to the 3’-end of the preceding exon although this results in some mismatches. This behaviour has also been corrected in the new Scipio version by some other parameters (see below).

Genes might not only contain very short exons between other exons but also at gene termini. Scipio uses an exact pattern search for N-terminal and C-terminal exons. Terminal exons will only be accepted if they match the query sequence and if the resulting intron borders agree with the two most common splice site patterns (GT---AG and GC---AG). The length of the terminal exons searched for is limited by the --gap_to_close parameter that is by default six residues.

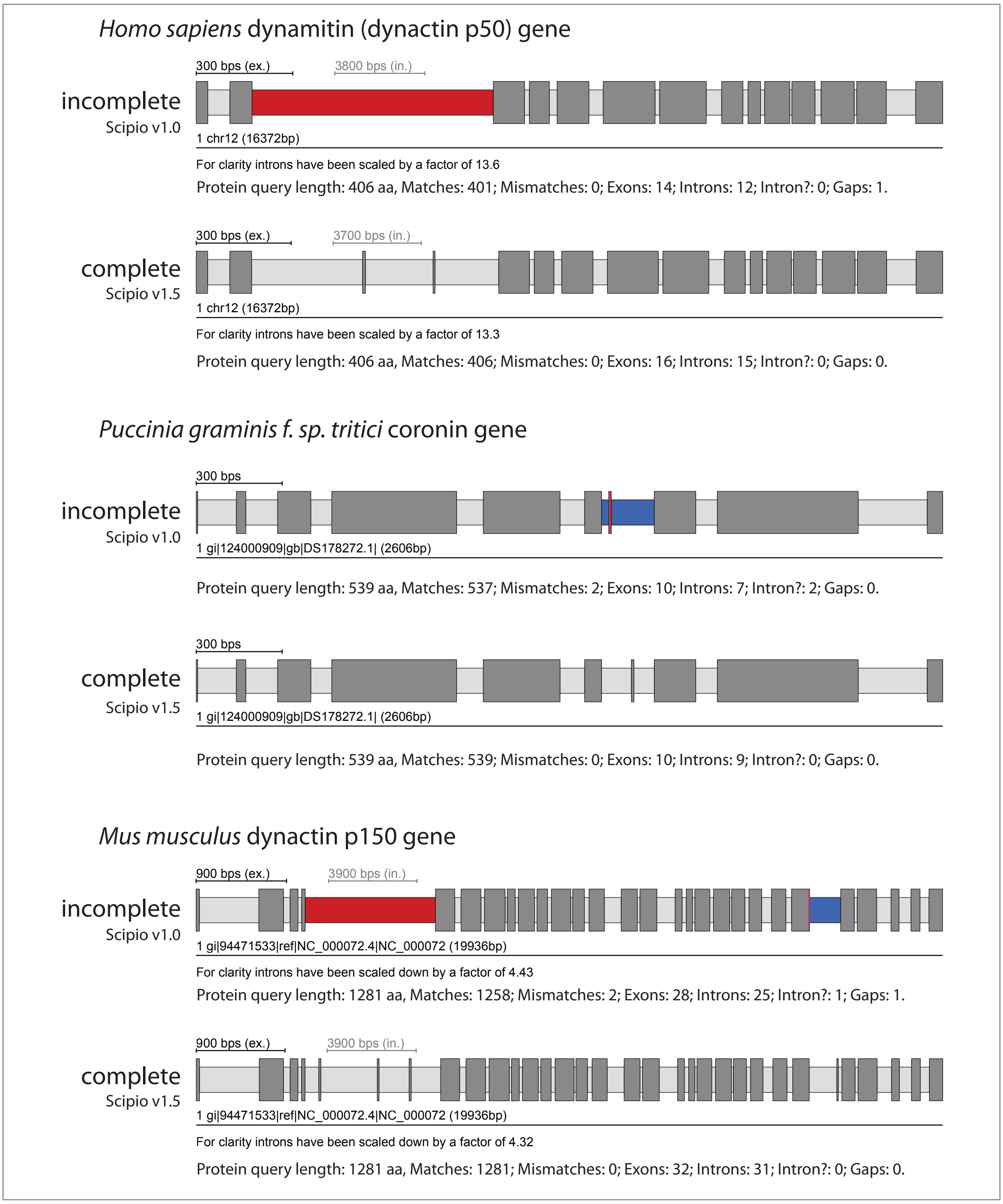

Homo sapiens dynamitin (dynactin p50) gene
300 bps (ex.)
3800 bps (in.)
incomplete
Scipio v1.0
1 chr12 (16372bp)
For clarity introns have been scaled by a factor of 13.6
Protein query length: 406 aa, Matches: 401; Mismatches: 0; Exons: 14; Introns: 12; Intron?: 0; Gaps: 1.
300 bps (ex.)
3700 bps (in.)
complete
Scipio v1.5
1 chr12 (16372bp)
For clarity introns have been scaled by a factor of 13.3
Protein query length: 406 aa, Matches: 406; Mismatches: 0; Exons: 16; Introns: 15; Intron?: 0; Gaps: 0.
Puccinia graminis f. sp. tritici coronin gene
300 bps
incomplete
Scipio v1.0
1 gi|124000909|gb|DS178272.1| (2606bp)
Protein query length: 539 aa, Matches: 537; Mismatches: 2; Exons: 10; Introns: 7; Intron?: 2; Gaps: 0.
300 bps
complete
Scipio v1.5
1 gi|124000909|gb|DS178272.1| (2606bp)
Protein query length: 539 aa, Matches: 539; Mismatches: 0; Exons: 10; Introns: 9; Intron?: 0; Gaps: 0.
Mus musculus dynactin p150 gene
900 bps (ex.)
3900 bps (in.)
incomplete
Scipio v1.0
1 gi|94471533|ref|NC_000072.4|NC_000072 (19936bp)
For clarity introns have been scaled down by a factor of 4.43
Protein query length: 1281 aa, Matches: 1258; Mismatches: 2; Exons: 28; Introns: 25; Intron?: 1; Gaps: 1.
900 bps (ex.)
3900 bps (in.)
complete
Scipio v1.5
1 gi|94471533|ref|NC_000072.4|NC_000072 (19936bp)
For clarity introns have been scaled down by a factor of 4.32
Protein query length: 1281 aa, Matches: 1281; Mismatches: 0; Exons: 32; Introns: 31; Intron?: 0; Gaps: 0.

Figure 2.1-4 | Reconstruction of very short exons. Case A shows the result for the reconstruction of the human dynamitin (dynactin p50) gene, that contains a 3 amino acid exon and a following 2 amino acid exon that are differentially included in the final transcript. These exons could not be reconstructed with Blat and the old Scipio version, but using the new Scipio version that enables Needlman-Wunsch searches. The --exhaust_align_size parameter has been set to 15,000 bp because of the length of the intron. Case B shows the result of the reconstruction of the coronin gene from *Puccinia graminis f. sp. tritici*. The small but evolutionarily conserved exon 7 can now correctly be reconstructed. Case C shows the result of the reconstruction of the mouse dynactin p150 gene that contains three short exons of 7, 6, and 7 amino acids close to the 5'-end of the gene. For the correct reconstruction, the --exhaust_align_size parameter has been increased to 10,000 bp, because of the length of the intron, and the --exhaust_gap_size has been set to 21 because of the length of the query that could not be mapped. The colour coding of the scheme is the same as in Figure 2.1-3. For further information see Supplementary information 2.1-2.

Parameters to account for low homology at intron borders

The correct prediction of exact intron borders is one of the most difficult tasks in protein-based gene-prediction, especially those intron borders next to small exons, because their residues might be falsely assigned to neighbouring exons, or when homology is low, as in cross-species applications. Here, divergent residues at intron borders are often not recognized by Blat, or conversely, intronic sequence is falsely assigned to the exon. To deal with the latter case, Scipio cuts off the marginal parts of Blat matches and realigns them. The parameter --max_move_exon allows increasing the default value of six residues that are cut off from the marginal parts. Figure 2.1-5 shows the effect of this parameter in some representative examples (see also Supplementary information 2.1-2). In the case of the human class-19 myosin gene, Blat and the old Scipio version were not able to reconstruct the 5'-end of the gene correctly, because the intron in front of the second exon of the gene ends with the translated sequence LFQ that is very homologous to the real sequence LQQ. Blat added these residues to exon 2 albeit introducing a mismatch. With the new parameter --max_move_exon (default setting is 6), Scipio is now able to resolve this misalignment and to subsequently identify the correct exon 1. Case B shows the reconstruction of the actin capping protein α from *Theileria heterothallica* (Figure 2.1-5, see also Supplementary information 2.1-2). Here, by chance the intergenic region before exon 3 shows some homology to exon 2 (3 matches and 3 mismatches) and thus the exon 2 sequence was erroneously joined to exon 3. This happened irrespectively of lowering the Blat tilesize or adjusting any of the other Scipio parameters. By setting the --max_move_exon to 6 (default setting), the new version of Scipio is now able to correctly reconstruct the CAPα gene.

Parameters to adjust searches on chromosomes or highly fragmented data

Scipio is able to reconstruct genes that are spread on several contigs or supercontigs of highly fragmented genomes. As we have shown, this feature is one of the most important strengths of Scipio [36] that other programs do not offer. However, this feature is not needed in chromosomal assemblies, and might lead, especially in the case of cross-species searches, to composed hits that stretch across multiple chromosomes, one of them being false positive (Figure 2.1-6). Hence, it can be switched off with the parameter --single_target_hits (or --chromosome), which is the default setting when selecting a chromosome assembly as genome target file in WebScipio.

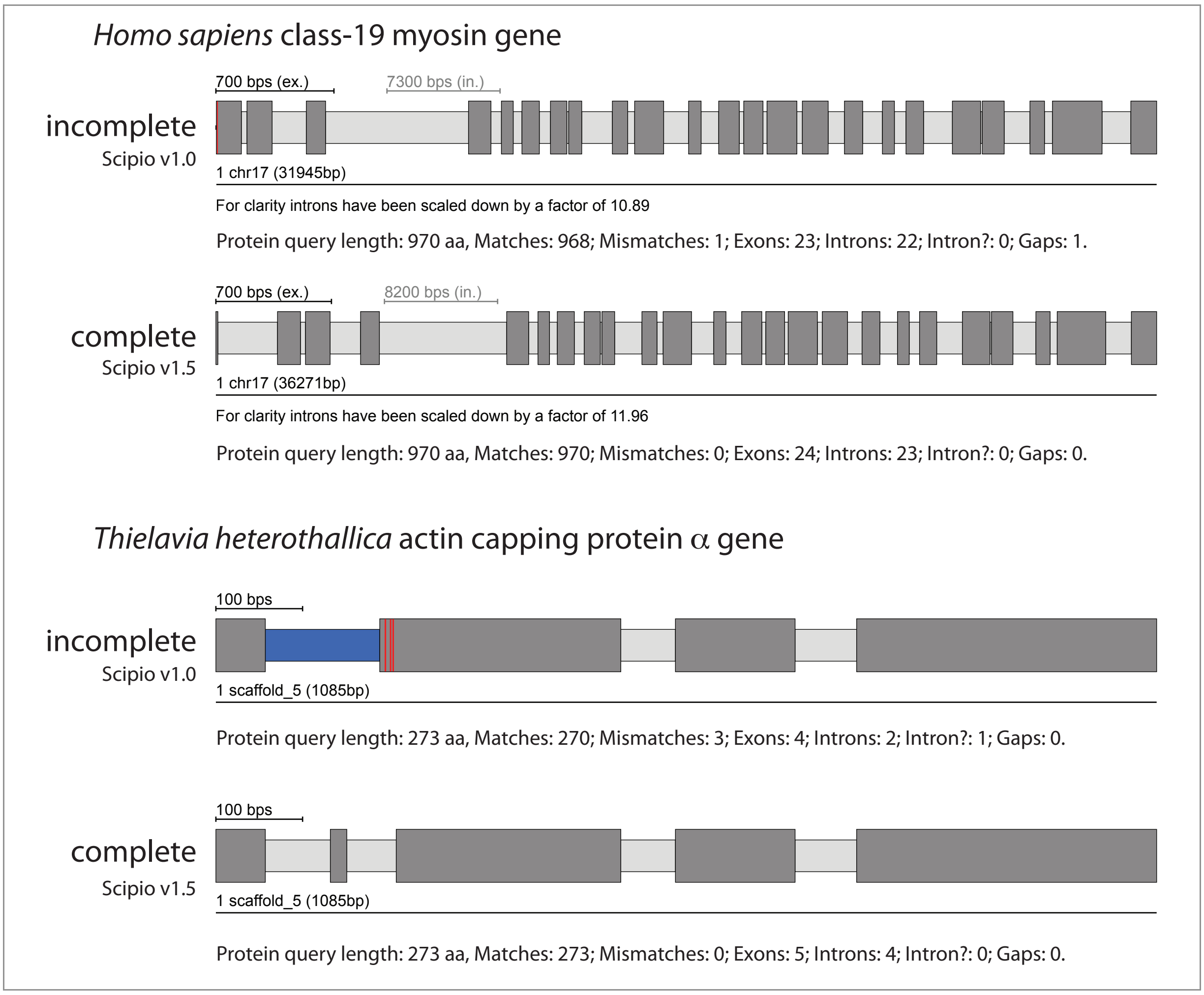

Homo sapiens class-19 myosin gene
700 bps (ex.)
7300 bps (in.)
incomplete
Scipio v1.0
1 chr17 (31945bp)
For clarity introns have been scaled down by a factor of 10.89
Protein query length: 970 aa, Matches: 968; Mismatches: 1; Exons: 23; Introns: 22; Intron?: 0; Gaps: 1.
700 bps (ex.)
8200 bps (in.)
complete
Scipio v1.5
1 chr17 (36271bp)
For clarity introns have been scaled down by a factor of 11.96
Protein query length: 970 aa, Matches: 970; Mismatches: 0; Exons: 24; Introns: 23; Intron?: 0; Gaps: 0.
Thielavia heterothallica actin capping protein α gene
100 bps
incomplete
Scipio v1.0
1 scaffold_5 (1085bp)
Protein query length: 273 aa, Matches: 270; Mismatches: 3; Exons: 4; Introns: 2; Intron?: 1; Gaps: 0.
100 bps
complete
Scipio v1.5
1 scaffold_5 (1085bp)
Protein query length: 273 aa, Matches: 273; Mismatches: 0; Exons: 5; Introns: 4; Intron?: 0; Gaps: 0.

Figure 2.1-5 | Reconstructing short exons at low homology intron borders. The scheme shows two examples for the reconstruction of short exons in regions where the intron borders of the neighbouring exons show some homology to the unmatched query sequence. The value for the --max_move_exon parameter has been set to 3 (case A) and 6 (case B), respectively. The colour coding of the scheme is the same as in Figure 2.1-3. For further information see Supplementary information 2.1-2.

For highly fragmented genomes it is still useful to allow gene reconstructions across several contigs. But also in this case one would want to exclude the assembly of hits that would introduce extremely long introns between exons on different contigs. To accomplish for those cases we have introduced the --max_assemble_size parameter that adjusts the maximum size of intron parts at target boundaries. If an intron would have to be created between two partial hits across two contigs that exceeds the given size (default: 75000 nucleotides), the two hits will not appear together as parts of one composed hit; rather, the lower-scoring contig will be discarded unless --multiple_results is enabled. Alternatively, the parameter --min_dna_coverage can be used to limit the length of introns stretching across contig boundaries, by specifying a minimum query/target length ratio for composed hits, in percent.

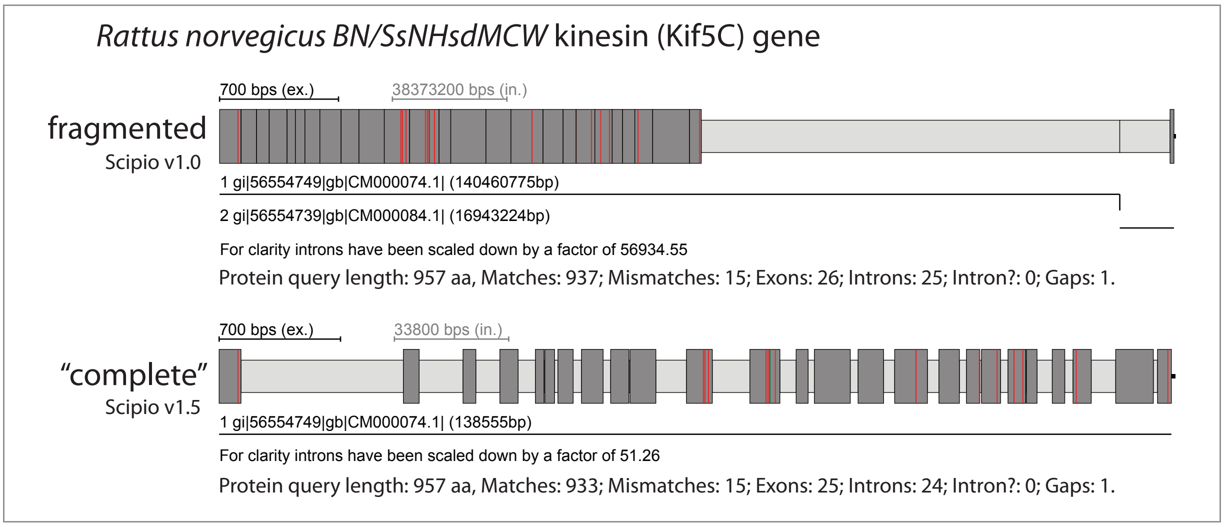

Rattus norvegicus BN/SsNHsdMCW kinesin (Kif5C) gene
fragmented
Scipio v1.0
700 bps (ex.)
38373200 bps (in.)
1 gi|56554749|gb|CM000074.1| (140460775bp)
2 gi|56554739|gb|CM000084.1| (16943224bp)
For clarity introns have been scaled down by a factor of 56934.55
Protein query length: 957 aa, Matches: 937; Mismatches: 15; Exons: 26; Introns: 25; Intron?: 0; Gaps: 1.
"complete"
Scipio v1.5
700 bps (ex.)
33800 bps (in.)
1 gi|56554749|gb|CM000074.1| (138555bp)
For clarity introns have been scaled down by a factor of 51.26
Protein query length: 957 aa, Matches: 933; Mismatches: 15; Exons: 25; Introns: 24; Intron?: 0; Gaps: 1.

Figure 2.1-6 | Reconstructing genes on chromosome assemblies. The scheme shows an example of the search for the rat homolog (target sequence) of the human Kif5C kinesin motor protein. The C-terminal about 25 amino acids of the rat Kif5C homolog are missing in the respective chromosome assembly. Using Scipio v1.0 a very short identical stretch of four amino acids, found on a different chromosome, has artificially been added to the 3'-end of the gene generating an "intron" of millions of base pairs (Note the scale of the introns!). The new parameter --single_target_hits now prevents this mis-assembly. The colour coding of the scheme is the same as in Figure 2.1-3.

Improved gene structure reconstruction in cross-species searches

To test the sensitivity and specificity of the new Scipio version we performed a cross-species search of the dynein heavy chain (DHC) genes of *Homo sapiens* in *Loxodonta africana*. The dynein heavy chain genes have been chosen because they belong to the longest genes in eukaryotic genomes and thus contain many exons spread on several hundred thousands of base pairs (Table 2.1-1). In addition, the dynein heavy chain family members show different degrees of identity in mammals and are therefore very suitable to test the limits of Scipio. Afrotheria (to which the elephants belong) and the Euarchontoglires separated about 100 million years ago [114]. The DHC query sequence test set and the longer time the species have split up should be a better test for the cross-species search capabilities of Scipio compared to the cross-species search of human myosin heavy chain genes in the mouse genome that we performed earlier [109].

Table 2.1-1 | Details of the dynein heavy chain genes used for the cross-species search

Protein	*Homo*	*Loxodonta*	Status*	*Loxodonta*	*Loxodonta*
Name	Length [aa]	Length [aa]		Length [bp]	Exons
DHC1	4646	4561	P	62248	77
DHC2	4307	4234	P	413085	89
DHC3A	4707	4690	✓	358204	92
DHC3B	4624	4582	P	298827	78
DHC4A	4507	4508	✓	361115	82
DHC4B	4462	4428	P	115251	79
DHC4C	4486	4339	P	486267	69
DHC5	4589	4584	✓	140290	79
DHC6	4509	4457	P	134640	86
DHC7A	4024	4019	✓	253605	62
DHC7B	4070	3966	P	187156	60
DHC7C	3960	3960	✓	201577	73
DHC8	4265	4064	F	80895	73
DHC9A	4158	4062	P	302568	75
DHC9B	4612	4597	P	369531	85
DHC11	4779	4779	✓	81205	43

* Status: P = partial sequence (short part of the sequence missing); F = sequence fragment (large region of the gene missing in the genome); ✓ = sequence complete

Figure 2.1-7 shows some example results of the cross-species search with genes of decreasing identity. The class-1 dynein heavy chain genes (DHC1) are very conserved between mammals, and the *Loxodonta* DHC1 could perfectly be reconstructed (except for the N-terminus that is not covered in the genome assembly). The DHC4A protein of *Loxodonta* has about 88 percent identity to the human homolog, and could also completely be reconstructed. In contrast, the DHC9B protein has only about 78 percent identity to the human homolog and the reconstructed gene still contains several gaps. The figure shows the result of the search using the old Scipio version compared to the result of the search with the new Scipio version. As reference, the result of the manual annotation of the gene is shown. It is very obvious that the new Scipio version provides a dramatically improved reconstruction of the *Loxodonta* DHC9B gene. More than 1,000 additional residues could be mapped corresponding to an increase in completeness by about 25 percent. The number of reconstructed exons increased from 62 to 80, which is close to the optimally reconstructed number of 85.

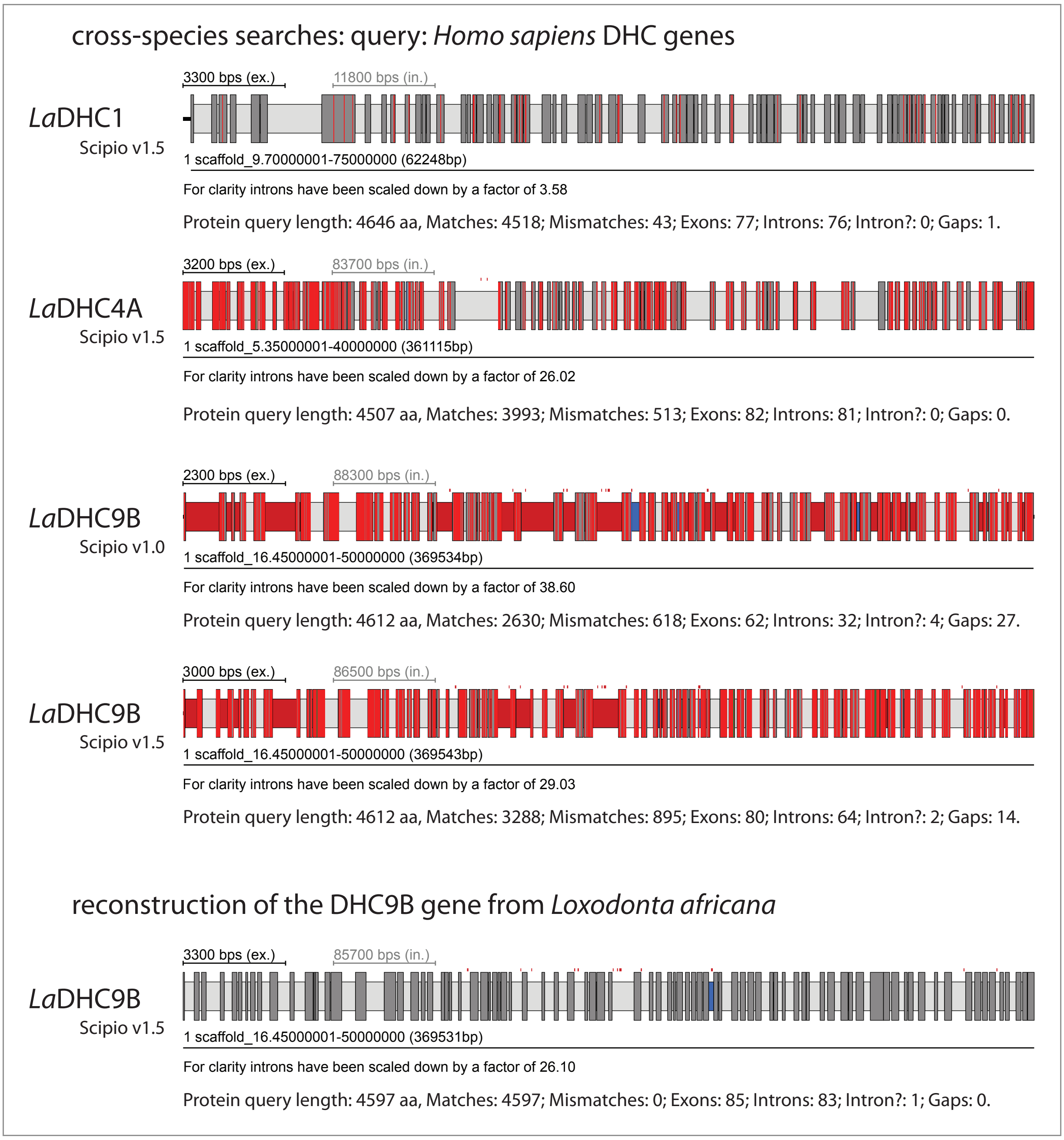

cross-species searches: query: *Homo sapiens* DHC genes
3300 bps (ex.)
11800 bps (in.)
*La*DHC1
Scipio v1.5
1 scaffold_9.70000001-75000000 (62248bp)
For clarity introns have been scaled down by a factor of 3.58
Protein query length: 4646 aa, Matches: 4518; Mismatches: 43; Exons: 77; Introns: 76; Intron?: 0; Gaps: 1.
3200 bps (ex.)
83700 bps (in.)
*La*DHC4A
Scipio v1.5
1 scaffold_5.35000001-40000000 (361115bp)
For clarity introns have been scaled down by a factor of 26.02
Protein query length: 4507 aa, Matches: 3993; Mismatches: 513; Exons: 82; Introns: 81; Intron?: 0; Gaps: 0.
2300 bps (ex.)
88300 bps (in.)
*La*DHC9B
Scipio v1.0
1 scaffold_16.45000001-50000000 (369534bp)
For clarity introns have been scaled down by a factor of 38.60
Protein query length: 4612 aa, Matches: 2630; Mismatches: 618; Exons: 62; Introns: 32; Intron?: 4; Gaps: 27.
3000 bps (ex.)
86500 bps (in.)
*La*DHC9B
Scipio v1.5
1 scaffold_16.45000001-50000000 (369543bp)
For clarity introns have been scaled down by a factor of 29.03
Protein query length: 4612 aa, Matches: 3288; Mismatches: 895; Exons: 80; Introns: 64; Intron?: 2; Gaps: 14.
reconstruction of the DHC9B gene from *Loxodonta africana*
3300 bps (ex.)
85700 bps (in.)
*La*DHC9B
Scipio v1.5
1 scaffold_16.45000001-50000000 (369531bp)
For clarity introns have been scaled down by a factor of 26.10
Protein query length: 4597 aa, Matches: 4597; Mismatches: 0; Exons: 85; Introns: 83; Intron?: 1; Gaps: 0.

Figure 2.1-7 | Example cross-species searches. The results of four searches with dynein heavy chain sequences from *Homo sapiens* in the elephant (*Loxodonta africana*) genome are shown. All genes are spread on several hundred thousands of base pairs. Statistics to the sequence results are given below the gene structure cartoons. An "intron?" is an intron for which the borders do not correspond to the standard splice sites GT---AG or GC---AG. The colour coding of the scheme is the same as in Figure 2.1-3.

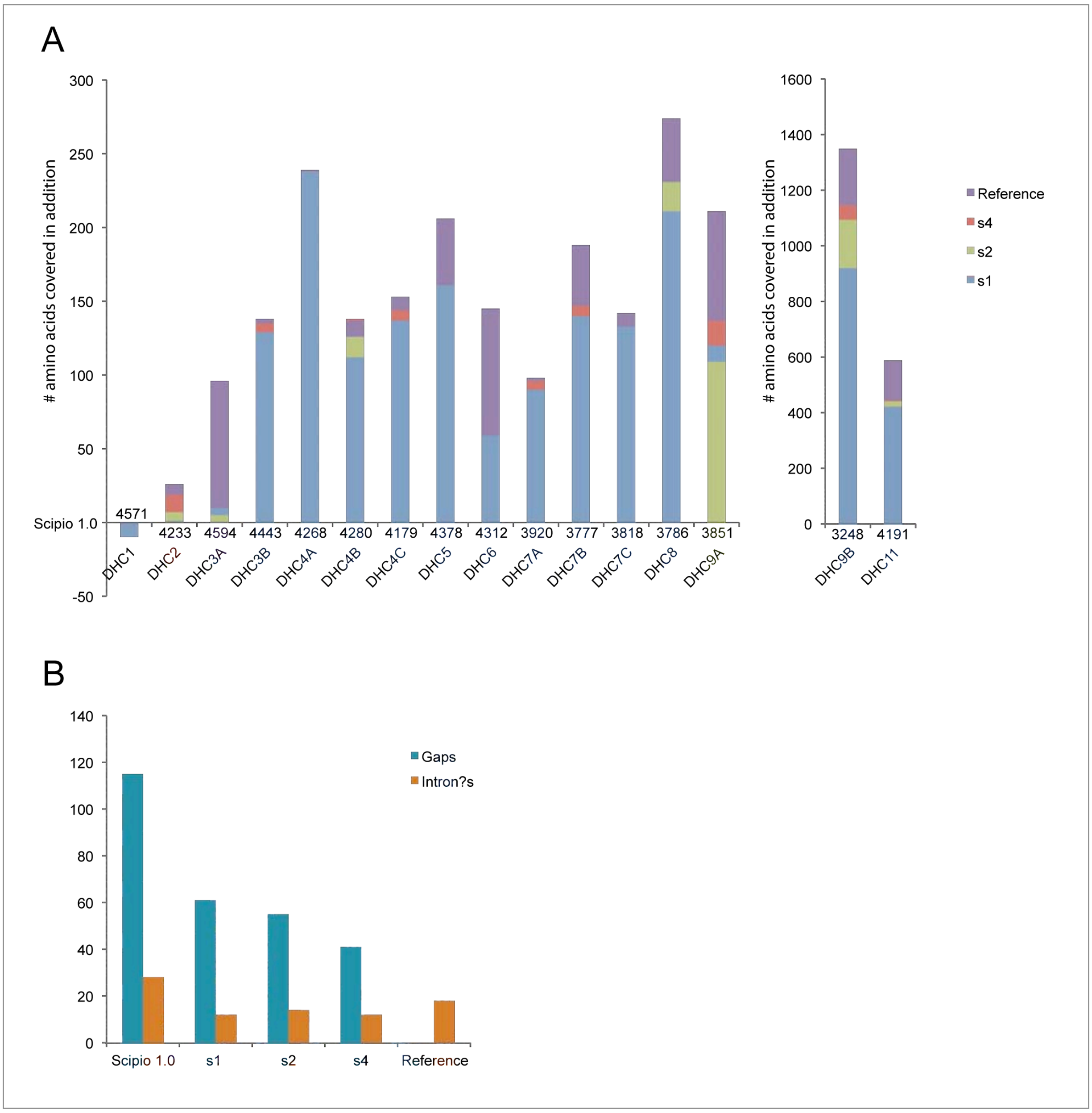


Figure 2.1-8 | Diagrams of the improvements introduced with the new Scipio version. The diagrams describe the improvement of the gene reconstructions of the DHC genes in the cross-species search of the human homologs (query sequences) in elephant (target sequence) using different Scipio versions and parameters. (A) The base-line is the result of the search using the old Scipio v1.0. The maximal possible annotation is represented by the gene reconstructions based on the manually annotated elephant DHC genes (reference dataset, purple). The blue bars show the reconstruction with Scipio v1.5 using --blat_tilesize = 7, --exhaust_align_size = 500 and --exhaust_gap_size = 21 (dataset s1). Green bars are results from the second search (dataset s2) with same parameters as for the first search, except for --blat_tilesize = 6 and --exhaust_gap_size = 18 (three times the tilesize). This dataset represents improvements independent of Scipio. The red bars represent searches with same parameters as for dataset s1, except for the increased parameters --exhaust_align_size = 5,000 and --exhaust_gap_size = 25 (dataset s4). This data takes far longer to compute compared to the first search, because of the Needleman-Wunsch search in longer regions. For the DHC1 gene Scipio v1.0 maps too many amino acids of the human query sequence to the elephant genome. So the negative bar representing the other datasets shows that these datasets cover the right number of 4561 amino acids. (B) This diagram depicts the number of gaps (human query sequence not matched in the elephant genome) and questionable introns (intron?; introns with uncommon splice sites) for the searches with the old Scipio version and the new version applying different parameters as in (A). The detailed values of the diagrams are shown in tables in Supplementary information 2.1-3.

The diagrams in Figure 2.1-8 show the improvements in gene reconstruction of the new Scipio version compared to the old version for the complete DHC dataset (see also Supplementary information 2.1-3). The reference for the perfectly reconstructed gene is the manual annotation based on the comparative annotation of more than 2,000 dynein heavy chain genes. The basis in diagram A is the reconstruction with Scipio v1.0, and shown are the improvements in the completeness of the annotation with Scipio v1.5 using different search parameters. In general, with the new Scipio version the reconstructions in these cross-species searches could considerably be improved. Lowering the tilesize, a Blat parameter to search with smaller fragments, further improved the results in only two cases. This corresponds to improvements independent of Scipio. However, extending the search frame for the exon search with the Needleman-Wunsch algorithm (parameter --exhaust_align_size) further completed the reconstruction in almost all cases demonstrating the effect of the newly introduced Needleman-Wunsch search for short or divergent exons.

Comparison of gene reconstruction and prediction tools

We compared Scipio to other tools that reconstruct and predict genes based on a protein sequence, and to general gene prediction tools. The tools can be ordered in three categories. The tools of the first category reconstruct the exon-intron structure of the protein-coding genes based on a genomic sequence and a provided protein sequence. Scipio [36], Prosplign [115], Exonerate [34], and Prot_map [35] belong to this category. The second category includes the tools Fgenesh+ [116], GeneWise/Wise2 [117], and GenomeScan [118], that combine homology based gene reconstructions taking advantage of given protein sequences, and *ab initio* gene prediction approaches. The third group of software packages consists of *ab initio* gene prediction tools like Augustus [119], Fgenesh [116], and Genscan [120]. The latter tools are not really comparable with the other ones in the task of reconstructing single genes, but the comparison illustrates the differences of *ab initio* and homology based gene predictions. In addition to Blat, we tested Blast [121], which can also be used as an initial search for the Prosplign tool. However, for our test cases this approach did not improve the results of Prosplign (see Supplementary information 2.1-4).

To evaluate the performance of Scipio in comparison to the other tools, four test scenarios have been designed. The DHC proteins have been chosen as a large general test set, while the other examples used for the explanation of the new Scipio parameters have been used as a test set for genes difficult to reconstruct. Both test data sets have been explored in reconstructions/predictions out of whole genome assemblies and respective gene regions. This differentiation has been done because only a few of the above-mentioned tools could be used in searches against whole genomes due to the limited upload possibilities of the respective web-interfaces while command-line versions of the tools were not available for every software. Thus we tested the performance of all tools against the gene regions of the test data that correspond to the nucleotide sequence of the reference annotation plus 2,000 additional base pairs up- and downstream. To make the execution times comparable, the genome wide runs were performed on a dedicated server, which contains four 2.2Ghz AMD Opteron 6174 processors, with 12 cores each, and 128 GByte of memory.

Scenario 1

In the first scenario, the tools had to reconstruct the dynein heavy chain genes in the whole *Loxodonta africana* genome assembly based on the human protein sequences (Table 2.1-2). Besides Scipio, only Exonerate and Augustus were able to produce reasonable results. Prot_map, Fgenesh, and Fgenesh+ could not be tested in this scenario because the command-line versions are proprietary and it is not possible to upload whole genome sequences via their web-interfaces. WebScipio is the only tool available, which already provides genome sequences. The dynein heavy chain genes contain 1,202 annotated exons including 209,486 nucleotides. The *Loxodonta africana* genome contains 3,271,792,967 nucleotides including N's. For the DHC1 gene the N-terminus cannot be found in the genome sequence because of a gap in the genome assembly. We adjusted the start of the first known exon in the reference annotation to the predicted exon for each tool, because the start depends on whether a tool found an exon in front of the first known exon. The results of the first test scenario are presented in Table 2.1-2 (for more data see Supplementary information 2.1-4). Both Scipio and Exonerate in the standard mode are comparable in exon sensitivity (93.4% and 94.8%, respectively) and missed a similar amount of exons (11 exons and 6 exons, respectively). However, Exonerate predicted many wrong exons (5669 exons) resulting in a low specificity (16.5%, compared to 93.3% exon specificity by Scipio). Exonerate can be configured to report only the best hit by setting the --bestn option to 1. While this option increased the specificity (from 16.5% to 90.2%), the sensitivity decreased (from 94.8% to 73.4%). Also, the number of missing exons increased to 287.

Comparing the results of Scipio and Blat illustrates that Blat found almost all exons, but that Scipio is needed to refine the exon borders as well as to exclude hits not related to the query sequence. Using the new Needleman-Wunsch algorithm Scipio v1.5 closes many gaps by adding and extending exons to the hits found by Blat. The number of missing exons is lower in Blat (9 exons missing) than in Scipio (11 exons missing), because Blat maps parts of the protein sequence to the genomic sequence, although these hits are not in the same order as in the protein sequence. Scipio excludes these hits. The results also show the great improvement of Scipio v1.5 compared to Scipio v1.0 in sensitivity (93.4% and 86.1%, respectively) and specificity (93.3% and 83.2%, respectively). Altogether, these results show that Scipio v1.5 is the only free tool that is able to reconstruct the genes nearly complete in this scenario.

Table 2.1-2 | Test scenario 1: Reconstruction of the *Loxodonta africana* dynein heavy chain genes in the whole genome sequence based on human protein sequences

Tool	Predicted genes	Missing exons[1]	Wrong exons[2]	Exon sens. %	Exon sens. (ov.)[3] %	Exon spec. %	Nucl. sens. %	Nucl. spec.%	Execution time per prot. seq.
Scipio 1.5[4]	16	11	6	93.4	99.1	93.3	98.7	99.8	70m 46s
Exonerate[5]	2145	6	5669	94.8	99.5	16.5	99.7	18.5	123m 23s
Exonerate[6]	16	287	62	73.4	76.1	90.2	76.1	94.3	121m 27s
Augustus[7]	1374928	0	3909434	47.9	100.0	0.0	100.0	0.3	> 10 days
BLAT[8]	-	9	264228	19.6	99.3	0.1	97.4	2.6	7m 24s
Scipio 1.0[4]	16	14	46	86.1	98.8	83.2	97.9	99.4	8m 24s

[1] Number of annotated exons, which are not overlapped by any predicted exon
[2] Number of predicted exons, which are not overlapped by any annotated exon
[3] Number annotated exons, which are overlapped by at least one predicted exon divided by the number of annotated exons
[4] Mammalia cross species default options (for detailed parameters see Supplementary information 2.1-5)
[5] Parameters: --model protein2genome
[6] Parameters: --model protein2genome --bestn 1
[7] Parameters: --species=human --genemodel=exactlyone (for more parameters see Supplementary information 2.1-5)
[8] Parameters as in 4: -tileSize=7 -minIdentity=54 -minScore=15 -oneOff=1

Scenario 2

The results of the second scenario are shown in Table 2.1-3. All above-mentioned tools were compared except for GenomeScan. Although GenomeScan produced results with the data provided on the respective webpage it did not work with our protein examples. The data show that Scipio performed in the same range as the other tools with respect to sensitivity and specificity. Scipio, Prosplign, and Exonerate revealed the highest sensitivity (94.7%, 95.7%, and 94.8%, respectively). Although Prosplign missed only one exon it also mis-predicted 41 exons. The homology based *ab initio* tools Fgenesh+ and Wise2 also provided almost complete reconstructions. Especially Fgenesh+ achieved high and balanced values for sensitivity (94.9%) and specificity (94.8%). The number of predicted genes illustrates that Exonerate without the --bestn option and Wise2 tend to divide long genes (32 and 39 genes predicted, respectively, instead of 16). The *ab initio* tools did not show comparable performance to the other tools in this scenario resulting in sensitivities of 76 – 82% and specificities of 58 – 83%. Augustus outperforms Fgenesh and Genscan with (Table 2.1-3) or without (see Supplementary information 2.1-4) the option to predict exactly one gene. Augusuts with the restriction to predict exactly one gene resulted in more accurate reconstructions. As in the whole genome scenario, the new Scipio v1.5 (93.1% sensitivity and 93.1% specificity) provides far better gene predictions than Blat and Scipio v1.0 (sensitivity of 19.9% and 86.2%, and specificity of 19.4% and 85.9%, respectively).

Table 2.1-3 | Test scenario 2: Reconstruction of the *Loxodonta africana* dynein heavy chain gene structures in the respective gene regions based on human protein sequences

Tool	Predicted genes	Missing exons[1]	Wrong exons[2]	Exon sens. %	Exon sens. (ov.)[3] %	Exon spec. %	Nucl. sens. %	Nucl. spec.%
Scipio 1.5[4]	16	13	6	93.1	98.9	93.1	98.6	99.8
Scipio 1.5[5]	16	4	7	94.7	99.7	93.7	99.2	99.8
Prosplign[6]	16	1	41	95.7	99.9	92.6	99.9	98.7
Exonerate[7]	32	7	6	94.8	99.4	94.6	99.6	99.5
Exonerate[8]	16	255	4	75.7	78.8	95.6	79.2	99.7
Prot_map[9]	16	4	27	91.7	99.7	86.2	99.3	99.7
Fgenesh+[10]	16	10	10	94.9	99.2	94.8	99.0	99.7
Wise2[11]	39	3	16	93.3	99.8	91.2	99.7	98.9
Augustus[12]	16	132	111	81.9	89.0	83.2	89.9	88.7
Fgenesh[10]	161	111	342	80.2	90.8	67.3	91.8	62.3
Genscan[13]	194	138	520	76.3	88.5	57.9	90.4	55.3
BLAT[14]	-	16	19	19.9	98.7	19.4	97.0	98.9
Scipio 1.0[4]	16	16	10	86.2	98.7	85.9	97.8	99.8

[1] Number of annotated exons, which are not overlapped by any predicted exon
[2] Number of predicted exons, which are not overlapped by any annotated exon
[3] Number annotated exons, which are overlapped by at least one predicted exon divided by the number of annotated exons
[4] Mammalia cross species default options (for detailed parameters see Supplementary information 2.1-5)
[5] Mammalia cross species default options; -tileSize=6 (for detailed parameters see Supplementary information 2.1-5)
[6] Parameters: -full -two_stages
[7] Parameters: --model protein2genome
[8] Parameters: --model protein2genome --bestn 1
[9] Similarity: Weak; Search for one best alignment only (for more parameters see Supplementary information 2.1-5)
[10] Organism: Human
[11] Parameters: -both
[12] Parameters:--species=human --genemodel=exactlyone (for more parameters see Supplementary information 2.1-5)
[13] Organism: Vertebrate; Suboptimal exon cutoff: 1.00
[14] Parameters as in [4]: -tileSize=7 -minIdentity=54 -minScore=15 -oneOff=1

Scenario 3 and 4

In the third and forth scenario we compared the tools in their performance to reconstruct the difficult cases, which we introduced above by describing the new parameters of Scipio v1.5. In scenario 3 a search in the whole genome and in scenario 4 a search in the respective gene regions (as in scenario 2) was performed. Table 2.1-4 summarizes the results of the third and forth scenario. Only when using the latest version of Scipio the genes of the test data set could correctly be reconstructed and predicted in the whole genome assemblies as well as in the gene region. None of the other tools was able to reconstruct all genes correctly, even if the gene region was given as in the forth scenario.

Table 2.1-4 | Test scenario 3 and 4: Difficult cases for reconstruction of gene structures

Tool	Ned kinesin	Phs dynactin p62	Hs dynactin p50	Pug coronin	Mm dynactin p150	Hs myosin	Th CAPα
Scipio 1.5[1]	✓	✓	✓	✓	✓	✓	✓
Prosplign[2]	o	o	✓	–	✓	–	✓
Exonerate[3]	o	o	–	–	✓	–	–
Prot_map[4]	o	o	–	–	–	o	–
Fgenesh+[5]	o	o	–	–	–	o	o
Wise2[6]	o	o	–	–	–	–	–
Augustus[7]	o	o	–	–	–	–	–
Fgenesh[5]	o	o	–	–	–	–	–
Genscan[8]	o	o	–	–	–	–	–
BLAT[9]	o	o	–	–	–	–	–
Scipio 1.0[1]	o	o	–	–	–	–	–

Ned: *Neurospora discreta*, Phs: *Phytophthora sojae*, Hs: *Homo sapiens*, Pug: *Puccinia graminis*, Mm: *Mus musculus*, Th: *Thielavia heterothallica*

✓All exons are reconstructed correctly

o All annotated exons are matched by or overlap with predicted exons

– Exons are missing

[1] Ned, Phs: cross species default options; Hs, Mm: default options, --exhaust_align_size=15000; Pug, Th: default options (for detailed parameters see Figure 2.1-3, Figure 2.1-4, and Figure 2.1-5 and Supplementary information 2.1-5)

[2] Parameters: -full -two_stages

[3] Parameters: --model protein2genome

[4] Similarity: Weak; Search for one best alignment only (for more parameters see Supplementary information 2.1-5)

[5] Organisms: Ned, Th: Neurospora crassa; Phs: Phytophtora; Hs: Human; Pug: Puccina; Mm: Mouse

[6] Parameters: -both

[7] Parameters: --genemodel=exactlyone; Organisms: Ned: --species=neurospora; Phs, Pug, Th: --species=generic; Hs, Mm: --species=human (for more parameters see Supplementary information 2.1-5)

[8] Organism: Vertebrate; Suboptimal exon cutoff: 1.00

[9] Parameters: -minScore=15; Ned, Phs: -tileSize=5 -minIdentity=54 -oneOff=1; Hs, Pug, Mm, Th: -tileSize=7 -minIdentity=81

2.1.5 Conclusions

Scipio and its graphical web-interface WebScipio are tools for the reconstruction of gene structures in eukaryotes. Scipio is based on the widely used program Blat that has been developed for aligning sequences of very high similarity. However, for the correct reconstruction of intron splice sites, very short exons, genes spread on several contigs, and the handling of sequencing errors a lot of post-processing is required. This is done by Scipio. Here, we present the fundamentally updated versions of Scipio and WebScipio, with an improved reconstruction of very short exons and intron splice sites, especially for the case of cross-species searches. To this end, we introduced a version of the Needleman-Wunsch algorithm that was shown to find a higher number of short exons previously missed, and to correct intron boundaries, especially in cases of lower sequence similarity. Furthermore, gaps in the mapping are now more frequently explained by divergent sequences, allowing for longer regions of insertions or deletions predicted on the same exon. Several parameters were introduced that can be used to fine-tune this behaviour if necessary. The sequence similarity between query and target sequence decreases with increasing evolutionary distance. While Blat is in principle able to locate hits for more distant species, the results become more and more incomplete, raising the importance of the post-processing. We could show that Scipio is now able to almost completely reconstruct genes from species whose ancestors separated more than 100 Myr ago. WebScipio allows easy access to Scipio and genome assemblies of about 640 eukaryotic species. This is unique to all gene reconstruction/prediction tools available and allows easy identification and reconstruction of protein homologs in related organisms. We compared the performance of Scipio to many other tools using our test data. While there are only minor differences in the reconstruction of the mammalian dynein heavy chain genes between Scipio, Exonerate, Prosplign, and Fgenesh+, the other software tools were not able to correctly reconstruct the more difficult cases encoding very short exons and showing strong sequence divergence at intron borders or inside of exons. Also unique to Scipio, this is the only tool available that is able to correctly reconstruct and predict genes that are spread on several contigs.

2.1.6 Availability and requirements

Project name: WebScipio, Scipio

Project home page: http://www.webscipio.org

Operating system: Platform independent

Programming languages: Ruby, Perl

Software requirements: Installation of Blat and BioPerl for using Scipio as command-line tool. WebScipio has been tested with InternetExplorer, Firefox, Chrome, Safari, and Opera.

License: WebScipio and Scipio may be obtained upon request and used under a GNU General Public License.

Any restrictions to use by non-academics: Using WebScipio and Scipio by non-academics requires permission.

2.1.7 List of abbreviations

Blat: BLAST like alignment tool; FTP: File transfer protocol; HTML: Hypertext markup language; SVG: Scalable vector graphics; PNG: Portable network graphics; YAML: YAML ain't markup language

2.1.8 Acknowledgements and Funding

MK has been funded by grants KO 2251/3-1, KO 2251/3-2, and KO 2251/6-1, and SW by grant WA 766/6-1 of the Deutsche Forschungsgemeinschaft. We thank Florian Odronitz for helpful discussions and support, and all the known and unknown users of WebScipio for their testing and feedback.

2.1.9 Authors' contributions

KH and MK set the requirements for the system and wrote the manuscript. KH and BH wrote the WebScipio software. HP implemented the test environment. OK wrote the Scipio source code and assisted in writing the manuscript. SW supervised the implementation of Scipio. KH, HP, OK, and MK performed extensive testing. KH performed the comparative software analysis. All authors read and approved the final version of the manuscript.

2.1.10 Supplementary information

Supplementary information 2.1-1 | Activity flow of the hit processing step. The scheme shows a detailed activity flow of the hit processing step. Here, the experienced user can see, where and how the various expert parameters modulate Scipio's hit processing, and can thus adjust these parameters to get the best result possible. This supplementary file is available from the corresponding publication.[35]

Supplementary information 2.1-2 | Protein – DNA alignments corresponding to the example searches. Here, additional data corresponding to the example searches is provided. This supplementary file is available from the corresponding publication.[36]

Supplementary information 2.1-3 | Table with detailed data of the results of the cross-species search of the human DHC genes in the elephant genome. The table provides detailed data to the cross-species searches including numbers of matches and mismatches, gaps and intron?'s, for the searches with different parameters. This supplementary file is available from the corresponding publication.[37]

Supplementary information 2.1-4 | Detailed evaluation values used for Table 2.1-2, Table 2.1-3, and Table 2.1-4. This file provides a description of each evaluation parameter and the values obtained with each software tool for all sequence predictions. The values highlighted in yellow were used for Table 2.1-2, Table 2.1-3, and Table 2.1-4. ,This supplementary file is available from the corresponding publication.[38]

Supplementary information 2.1-5 | Software versions and run parameters of the gene reconstruction and prediction tools. The tables shows the exact versions and run parameters, which were used for the comparison, for each scenario. This supplementary file is available from the corresponding publication.[39]

[35] http://www.biomedcentral.com/content/supplementary/1756-0500-4-265-s1.pdf
[36] http://www.biomedcentral.com/content/supplementary/1756-0500-4-265-s2.pdf
[37] http://www.biomedcentral.com/content/supplementary/1756-0500-4-265-s3.pdf
[38] http://www.biomedcentral.com/content/supplementary/1756-0500-4-265-s4.pdf
[39] http://www.biomedcentral.com/content/supplementary/1756-0500-4-265-s5.pdf

2.2 Predicting mutually exclusive spliced exons based on exon length, splice site and reading frame conservation, and exon sequence homology

Holger Pillmann[*], Klas Hatje[*], Florian Odronitz, Björn Hammesfahr, and Martin Kollmar

Abteilung NMR-basierte Strukturbiologie, Max-Planck-Institut für biophysikalische Chemie, Am Fassberg 11, D-37077 Göttingen, Germany
* Contributed equally

BMC Bioinformatics 2011, 12(270)

Published: 30 June 2011
doi: 10.1186/1471-2105-12-270
This article is available from http://www.biomedcentral.com/1471-2105/12/270

2.2.1 Abstract

Background

Alternative splicing of pre-mature RNA is an important process eukaryotes utilize to increase their repertoire of different protein products. Several types of different alternative splice forms exist including exon skipping, differential splicing of exons at their 3'- or 5'-end, intron retention, and mutually exclusive splicing. The latter term is used for clusters of internal exons that are spliced in a mutually exclusive manner.

Results

We have implemented an extension to the WebScipio software to search for mutually exclusive exons. Here, the search is based on the precondition that mutually exclusive exons encode regions of the same structural part of the protein product. This precondition provides restrictions to the search for candidate exons concerning their length, splice site conservation and reading frame preservation, and overall homology. Mutually exclusive exons that are not homologous and not of about the same length will not be found. Using the new algorithm, mutually exclusive exons in several example genes, a dynein heavy chain, a muscle myosin heavy chain, and Dscam were correctly identified. In addition, the algorithm was applied to the whole *Drosophila melanogaster* X chromosome and the results were compared to the

Flybase annotation and an *ab initio* prediction. Clusters of mutually exclusive exons might be subsequent to each other and might encode dozens of exons.

Conclusions

This is the first implementation of an automatic search for mutually exclusive exons in eukaryotes. Exons are predicted and reconstructed in the same run providing the complete gene structure for the protein query of interest. WebScipio offers high quality gene structure figures with the clusters of mutually exclusive exons colour-coded, and several analysis tools for further manual inspection. The genome scale analysis of all genes of the *Drosophila melanogaster* X chromosome showed that WebScipio is able to find all but two of the 28 annotated mutually exclusive spliced exons and predicts 39 new candidate exons. Thus, WebScipio should be able to identify mutually exclusive spliced exons in any query sequence from any species with a very high probability. WebScipio is freely available to academics at http://www.webscipio.org.

2.2.2 Background

Eukaryotes can enhance their repertoire of different protein products by alternative splicing of the corresponding genes [122]. Since the first description of alternative splicing of precursor mRNA almost 30 years ago [123, 124] the suggested and verified percentage of human genes that are spliced into alternative transcripts has steadily risen (for reviews see for example [125, 126]). Very recently, two studies using high-throughput sequencing indicate that every single human gene containing more than one exon is transcribed and processed to yield multiple mRNAs [41, 42].

Mainly, five different types of alternative splicing affect the resulting translated protein product [44, 127, 128]: The first type is exon skipping, in which an exon, also called cassette exon, is spliced out of the transcript together with its flanking introns. The second and third types are the alternative splicing of the 3' splice site and 5' splice site, respectively. Here, two or more splice sites are recognized at one end of the exon. The fourth type is intron retention in which part of an exon is either spliced (like a regular intron) or retained in the mature mRNA transcript. While exon skipping and alternative 3' splice site selection account for most alternative splicing events in higher eukaryotes [129, 130], the most prevalent type of alternative splicing in plants, fungi, and protozoa is intron retention [49]. The fifths type is called mutually exclusive splicing and is used for clusters of internal exons that are spliced in a mutually exclusive manner. It is important to note that the term mutually exclusive splicing is only used for these specific clusters of exons. Mutually exclusive splicing demands a specific mechanism for the regulated splicing of exactly one of the exons of such a cluster. Recent analyses have shown that this mechanism might be based on intra-intronic RNA pairings that are conserved at the secondary structure level [68, 69, 81]. These alternatively spliced exons must not be mixed up with exons that seem to be spliced in a mutually exclusive manner based on their annotation. This especially accounts for terminal exons that are alternatively spliced in conjunction with the use of alternative promoters or 3'-end processing sites (for a review see for example [131]). The regulation of the splicing of these types need not be at the level of splicing.

To our knowledge, the only study to identify and predict regions *in silico* that might contain mutually exclusive spliced exons used a method of local similarity of genomic regions at the nucleotide level [132]. Assuming that clusters of mutually exclusive exons evolved by one or several rounds of single-exon duplications, given gene locations were self-aligned using a pairwise local alignment algorithm to derive similar regions. Those regions were regarded as candidate regions, and mutually exclusive exons were only predicted by verification through EST and cDNA data. The method itself cannot determine exons including intron splice sites, and is not able to identify mutually exclusive exons whose DNA sequences have diverged considerably. False positive candidates are detected in regions that contain clusters of duplicated genes, and in regions containing pseudo-exons (e.g. exons that are in the process of being lost containing frame-shifts and in-frame stop codons, and missing correct splice sites).

Here, we propose a different approach that is based on the knowledge of creating meaningful transcripts. We presume that most mutually exclusive exons encode the same region of the resulting protein structure. These regions are embedded in the surrounding three-dimensional structure and thus alternative exons must preserve all structurally important contacts between the corresponding local structure elements. A demonstrative example is the alternatively spliced motor domain of the muscle myosin heavy chain in arthropods [67]. In *Drosophila*, four clusters of mutually exclusive spliced exons encode regions of the motor domain, and the variability of creating different transcripts and further fine-tune the motor domain function is even enhanced in the waterflea *Daphnia magna* by four additional clusters. One of the clusters contains exons encoding the so-called relay helix and subsequent relay loop, a structural element that starts at switch-2 embedded in the middle of the motor domain and ends at the connection to the converter domain. This whole relay element converts small conformational changes at the ATP-binding site to large movements of the lever arm [133]. Retaining structural integrity is therefore indispensible for mutually exclusive exons. Of course, parts of the exons might also encode loop regions, but also those parts must at least partly be conserved to retain their general function.

Based on these preconditions we apply the following constrains to our search for mutually exclusive exons: A) Mutually exclusive exons must have about the same length (allowing some length difference for e.g. parts encoding loop regions). B) They must have conserved splice site patterns (e.g. a GT 5' intron splice site cannot be combined with a AC 3' splice site) and the reading frame of the exon must be conserved. C) They must show sequence similarity. These features have been implemented in an extension to the WebScipio software. The application of the algorithm to various genes from several eukaryotes, and to all genes of the X chromosome of *Drosophila melanogaster* is demonstrated.

2.2.3 Methods

The search algorithm has been implemented as an extension to the WebScipio web application [109]. It is based on the exon-intron gene structure reconstructed by Scipio [36]. The extension is written in the Ruby programming language[40] and fully integrated into WebScipio to facilitate user interaction, and visualization and analysis of the results. WebScipio uses the web framework Ruby on Rails[41]. To make the session storage fast, flexible, and scalable a database backend consisting of Tokyo Cabinet and Tokyo Tyrant[42] is used. To run jobs in background the Rails plug-in Workling[43] in combination with Spawn[44] is applied.

Search algorithm

The new algorithm divides into several steps, which are executed for each original exon (Figure 2.2-1, a detailed activity diagram is available as Supplementary information 2.2-1). It assumes that mutually exclusive spliced exons share the following features: Firstly, mutually exclusive spliced exons have a similar length; secondly, their splice sites and reading frames are conserved; thirdly, they are homologous.

[40] http://www.ruby-lang.org

[41] http://rubyonrails.org

[42] http://fallabs.com/tokyocabinet

[43] http://github.com/purzelrakete/workling

[44] http://github.com/tra/spawn

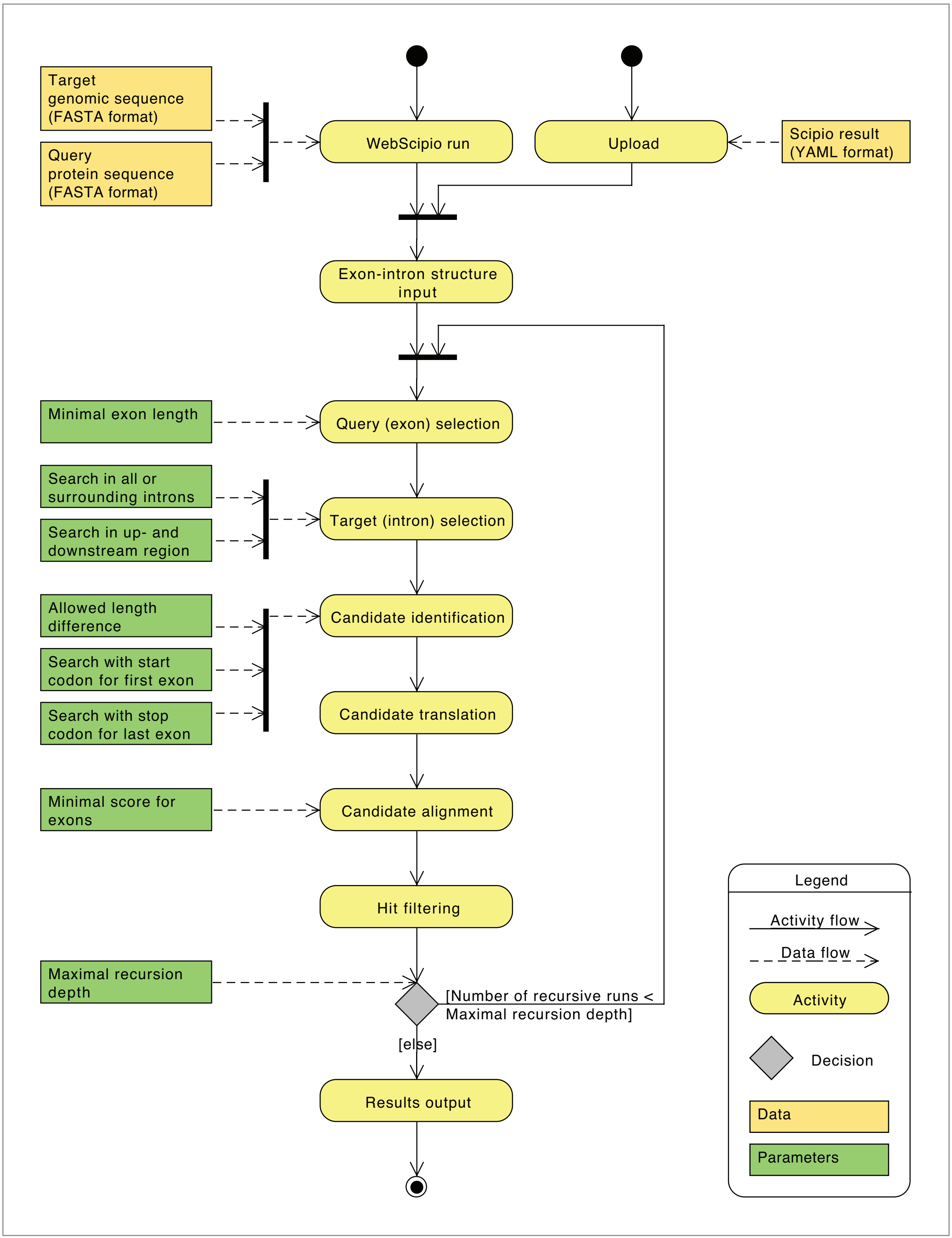

Target genomic sequence (FASTA format)
Query protein sequence (FASTA format)
WebScipio run
Upload
Scipio result (YAML format)
Exon-intron structure input
Minimal exon length
Query (exon) selection
Search in all or surrounding introns
Search in up- and downstream region
Target (intron) selection
Allowed length difference
Search with start codon for first exon
Search with stop codon for last exon
Candidate identification
Candidate translation
Minimal score for exons
Candidate alignment
Hit filtering
Maximal recursion depth
[Number of recursive runs < Maximal recursion depth]
[else]
Results output
Legend
Activity flow
Data flow
Activity
Decision
Data
Parameters

Figure 2.2-1 | Activity diagram of the search algorithm. The activity diagram shows the processing steps of the search algorithm and the influence of the parameters on each step. The run starts with an exon-intron gene structure determined by Scipio. Based on the chosen parameters the exons and corresponding introns are selected and searched for mutually exclusive spliced exon candidates. The candidates are processed and filtered. These steps are repeated in the case of a recursive run. In the end, the algorithm outputs the exon-intron structure including mutually exclusive spliced exons.

For each internal exon ("original exon") the two surrounding introns (or optionally all introns of the gene) are scanned for exon candidates that have a similar length. These exon candidates must introduce introns with the following splice site pattern: GT---AG, GC---AG, GG---AG, and AT---AC. Firstly, the algorithm looks for the nucleotide pairs AG or AC in the intron sequence, which define start sites of exon candidates and 3' splice sites of the proposed intron. If the intron in front of the original exon starts with GT, GC or GG the algorithm searches for AG, if it starts with an AT the algorithm searches for AC. Secondly, the algorithm looks for the nucleotide pairs GT, GC, GG and AT in the intron sequence, which define ends of exon candidates and 5' splice sites of the proposed intron. If the intron following the original exon ends with AG the algorithm searches for GT, GC and GG, if it ends with AC the algorithm searches for AT. The nucleotide sequences between two possible 3' and 5' splice sites of the scanned intron that have a length similar to the length of the original exon are considered as exon candidates. The maximum length difference between an exon and its candidate can be adjusted by the *allowed length difference* parameter in number of amino acids. The default value of this parameter is 20 aa.

For terminal exons, the algorithm is able to scan the up- and downstream regions of the gene for exon candidates. The first exon of a protein-coding gene has to start with the start codon ATG. Thus, for the first exon, alternative candidates must start with ATG instead of sharing a theoretical splice site pattern with the first exon. The last exon is followed by a stop codon (TAG, TAA, or TGA) and all exon candidates must be followed by a stop codon instead of sharing a splice site pattern with the last exon. The use of the start codon and stop codon instead of the splice sites can be adjusted by the *search with start codon for first exon* and *search with stop codon for last exon* parameters. For example it would be useful to release this restriction in the case where the algorithm searches for alternative exons in a protein fragment. The default of these parameters is to search with a start codon if the first amino acid of the user-provided protein query sequence starts with methionine, and to search with stop codons if the last exon is followed by a stop codon. To reduce the number of candidates it is possible to set the *minimal exon length* parameter. Original exons, which are shorter than this length, are not considered in the candidate search. The default value for this parameter is 15 aa.

The nucleotide sequences of the exon candidates are translated into amino acid sequences using the BioRuby library [95]. The candidates are translated in the same reading frame as the original exon, because their nucleotide sequences appear mutually exclusive in the resulting mRNA and thus share the same reading frame. If the translation results in an in-frame stop codon, the candidate is rejected.

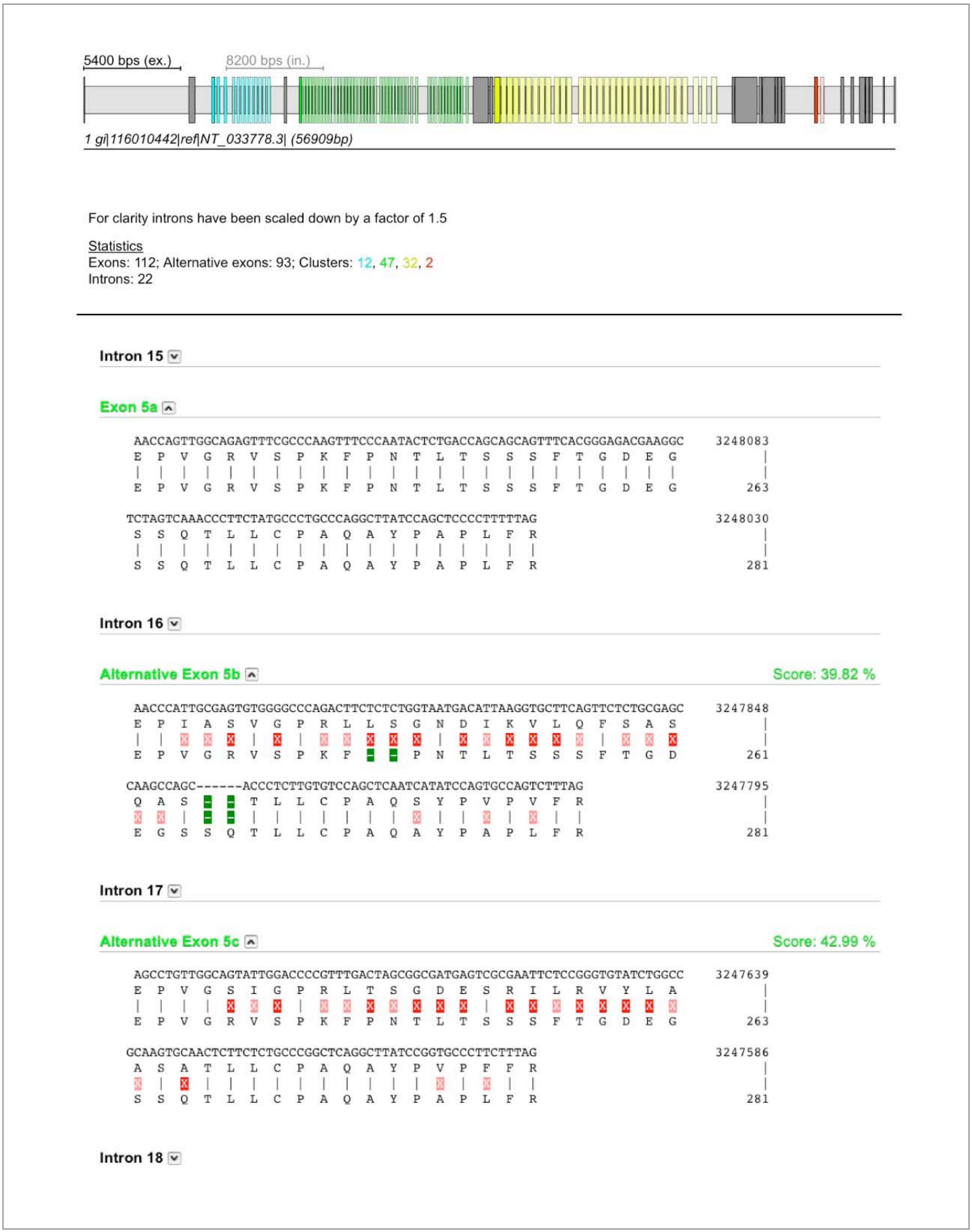
Each candidate sequence is aligned to the original exon sequence. If the alignment score is high, the probability that the two exons are homologous is high as well. The optimal global alignment of the two amino acid sequences is calculated with the Gotoh algorithm, which extends the Needleman-Wunsch algorithm by affine gap costs [30, 134]. For this task, the pair_align program of the SeqAn package [96] is used. The gap penalties are set to -10 for initial gaps and -2 for extending gaps. The Blosum62 matrix is used as substitution matrix [135, 136]. Because of differences in length and amino acid composition of the clusters of mutually exclusive exons the resulting global alignment scores are not directly comparable. To normalise the alignment scores each score is divided by the score of the alignment of the original exon sequence to itself. This relative score shows the similarity of the two sequences on a scale from zero to one. Candidates, which have a low alignment score, are rejected. The threshold for rejection can be adjusted in percent by the *minimal score for exons* parameter (default: 15%). If candidate regions overlap the highest scoring candidates are retained or, if scores are identical, the longest candidates.

An optional recursive search was implemented to find less similar alternative exons. If this option is selected, the search is repeated with the found alternatively spliced exons as query exons. The number of recursive runs can be adjusted with the *maximal recursion depth* parameter up to three rounds of recursion (default: recursive search disabled).

WebScipio integration

The WebScipio tool allows reconstructing an exon-intron gene structure based on a protein sequence query. This reconstruction step is the basis for the mutually exclusive spliced exon search. The user can enable the search and adjust several parameters in the Advanced Options section of WebScipio. The search will run subsequently to the gene structure reconstruction step. In addition, the user can enable the search after uploading a previously calculated and downloaded Scipio result.

The result of the search is displayed in the Result section of the WebScipio interface (Figure 2.2-2, top). The standard gene structure picture is extended by the predicted mutually exclusive spliced exons. The alternative exons corresponding to the same original exon constitute a cluster. Exons of a cluster get the same colour. The original exon is dark coloured and the corresponding predicted ones are lighter coloured depending on their similarity with respect to the original exon. In the Statistics section the number of exons in each cluster is shown in colour.



For clarity introns have been scaled down by a factor of 1.5

Statistics
Exons: 112; Alternative exons: 93; Clusters: 12, 47, 32, 2
Introns: 22

Intron 15

Exon 5a

```
 AACCAGTTGGCAGAGTTTCGCCCAAGTTTCCCAATACTCTGACCAGCAGCAGTTTCACGGGAGACGAAGGC    3248083
 E  P  V  G  R  V  S  P  K  F  P  N  T  L  T  S  S  S  F  T  G  D  E  G            |
 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |            |
 E  P  V  G  R  V  S  P  K  F  P  N  T  L  T  S  S  S  F  T  G  D  E  G          263

TCTAGTCAAACCCTTCTATGCCCTGCCCAGGCTTATCCAGCTCCCCTTTTTAG                        3248030
 S  S  Q  T  L  L  C  P  A  Q  A  Y  P  A  P  L  F  R                              |
 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |                              |
 S  S  Q  T  L  L  C  P  A  Q  A  Y  P  A  P  L  F  R                            281
```

Intron 16

Alternative Exon 5b Score: 39.82 %

```
 AACCCATTGCGAGTGTGGGGCCCAGACTTCTCTCTGGTAATGACATTAAGGTGCTTCAGTTCTCTGCGAGC    3247848
 E  P  I  A  S  V  G  P  R  L  L  S  G  N  D  I  K  V  L  Q  F  S  A  S            |
 |  |  X  X  X  |  X  |  X  X  X  X  X  |  X  X  X  X  X  X  |  X  X  X            |
 E  P  V  G  R  V  S  P  K  F  -  -  P  N  T  L  T  S  S  S  F  T  G  D          261

CAAGCCAGC------ACCCTCTTGTGTCCAGCTCAATCATATCCAGTGCCAGTCTTTAG                  3247795
 Q  A  S  -  -  T  L  L  C  P  A  Q  S  Y  P  V  P  V  F  R                        |
 X  X  |  -  -  |  |  |  |  |  |  |  X  |  |  X  |  X  |  |                        |
 E  G  S  S  Q  T  L  L  C  P  A  Q  A  Y  P  A  P  L  F  R                      281
```

Intron 17

Alternative Exon 5c Score: 42.99 %

```
 AGCCTGTTGGCAGTATTGGACCCCGTTTGACTAGCGGCGATGAGTCGCGAATTCTCCGGGTGTATCTGGCC    3247639
 E  P  V  G  S  I  G  P  R  L  T  S  G  D  E  S  R  I  L  R  V  Y  L  A            |
 |  |  |  |  X  X  X  |  X  X  X  X  X  X  X  |  X  X  X  X  X  X  X  X            |
 E  P  V  G  R  V  S  P  K  F  P  N  T  L  T  S  S  S  F  T  G  D  E  G          263

GCAAGTGCAACTCTTCTCTGCCCGGCTCAGGCTTATCCGGTGCCCTTCTTTAG                        3247586
 A  S  A  T  L  L  C  P  A  Q  A  Y  P  V  P  F  F  R                              |
 X  |  X  |  |  |  |  |  |  |  |  |  |  X  |  X  |  |                              |
 S  S  Q  T  L  L  C  P  A  Q  A  Y  P  A  P  L  F  R                            281
```

Intron 18

Figure 2.2-2 | Gene structure representation and detailed alignment view. The figure shows the WebScipio gene structure representation of the *Drosophila melanogaster* Dscam gene with mutually exclusive spliced exons and a section of the alignment view including exon 5 and the first two identified alternative exon candidates. The colours in the gene structure figure are the same as the colours of the exon identifiers in the text alignment. The opacity of the colours of each alternative exon corresponds to the alignment score of the alternative exon to the original one. This score is shown in the detailed alignment view next to the exon identifier. For each exon the genomic sequence, its translation, and the translation of the original exon is shown. Identical residues are illustrated as dashes and mismatches as red highlighted crosses. The crosses are highlighted in light red for amino acids, which are chemically similar. Gaps are marked as green hyphens.

The Alignment view (Figure 2.2-2, bottom) offers a detailed analysis at the sequence level. For each alternative exon the genomic sequence, its translation, and the alignment to the original translated exon are shown. The alignment score is given in percent. The alternative exons are also marked in the Genomic DNA result view. In the Coding DNA and Translation result view the user can choose the alternative exons that should build the alternative coding DNAs or protein sequences. The results can be downloaded in several data formats. The YAML[45] file contains all corresponding information and can later be uploaded and used for future analysis. Additionally, the results can be downloaded as General Feature Format (GFF) file[46]. The figures can be downloaded in the Scalable Vector Graphics (SVG) format[47] for further high quality processing. Example searches as well as further descriptions of the search parameters are provided on the help pages of WebScipio.

2.2.4 Results and Discussion

Identification of mutually exclusive spliced exons

The search for mutually exclusive spliced exons is based on three criteria: (1) The lengths of the mutually exclusive exons must be very similar, because these exons are supposed to code for the same part in the resulting protein structure, including identical secondary structural elements. (2) To be spliced in a mutually exclusive way, the exons must have similar splice sites and reading frames to be compatible with the previous and following exons. (3) The exons must encode homologous protein sequences, because their inclusion into the protein structure must be compatible with the corresponding local structural environment. The search implemented in WebScipio is based on the availability of the gene structure. Firstly, mutually exclusive exon candidates are searched for using corresponding splice sites to the query exons and restricting the candidate length to similar reading frames (e.g. split codons in the query exon must result in split codons in the candidate exons). Total length difference is less restricted allowing length differences between query and candidate exons at the DNA level in multiples of three for each additional or missing codon. These candidate exons are then filtered and scored based on the Blosum62 matrix. The best scoring, non-overlapping candidates are proposed to be alternative exons to the respective query exon, resulting in a cluster of mutually exclusive exons. With this approach, the absolute necessary constraints at the DNA-level that can be obtained by bioinformatics means are combined with biological information. Based on these criteria several cases can be distinguished: (A) alternative exons found in the surrounding introns of single internal exons should form true clusters of mutually exclusive

[45] http://www.yaml.org

[46] http://www.sanger.ac.uk/resources/software/gff/spec.html

[47] http://www.w3.org/Graphics/SVG

exons, (B) alternative exons found for terminal exons most probably constitute multiple promoters or multiple poly(A) sites, (C) clusters of several exons in combination, which can be found by searching for candidates for all exons in all introns and up- and downstream regions, most probably represent cases of tandemly arrayed gene duplications or *trans*-spliced genes.

Example genes with clusters of mutually exclusive exons

To test the quality of the new algorithm, several well-known genes with clusters of mutually exclusive exons with different characteristics were analysed (Figure 2.2-3). The first test case is the cytoplasmic dynein heavy chain from *Schistosoma mansoni* (*Sm*DHC1). Dynein heavy chains belong to the longest genes in eukaryotes encoding 4000 – 5000 residues and are spread over several dozens of exons. The mutually exclusive exon is clearly identified in the middle of the gene, encoding split codons at the 3'- and 5'-end of the exon. The query exon and the candidate exon have identical lengths and show strong homology. Based on the multiple sequence alignment of more than 2000 DHCs these exons are mutually exclusive and not constitutive or differentially included. The second case represents the muscle myosin heavy chain gene from the waterflea *Daphnia magna* [67]. The arthropod muscle myosin heavy chain genes contain several clusters of mutually exclusive exons to fine tune the mechanochemical characteristics of the motor domain that are needed to accomplish the different tasks in the various muscle types [137]. The *Dap*Mhc1 is an example with nine clusters of mutually exclusive exons of which several are adjacent and not interrupted by constitutive exons. The new algorithm found all mutually exclusive exons that have manually been identified previously [67]. The two example alignments show that the new algorithm is able to correctly identify even short exons with limited complexity, and subsequent clusters of mutually exclusive exons encoded in different reading frames. The third example shows the prediction of the mutually exclusive exons in Dscam (Down syndrome cell adhesion molecule) from *Drosophila melanogaster*, which is known to encode the largest set of mutually exclusive exons of any gene analysed so far [138, 139]. The potentially 95 mutually exclusive exons of the Dscam gene are organized into four clusters that are separated by constitutive exons. The exon 4, 6, 9, and 17 clusters are supposed to contain 12, 48, 33, and 2 exons, respectively [139]. In the publicly available *Drosophila melanogaster* reference genome sequence (chromosome assembly version 4.1 as provided by Flybase [140, 141]) mutually exclusive exons were searched using a gene translation containing the first exons of each of the clusters as query sequence. If clusters contain that many exons as are found in the Dscam genes it might be possible that the exon, that has been included in the query sequence, is the most divergent of the exons of the cluster. Therefore, a parameter to the search algorithm that enforces recursive searches in all introns with the newly identified exon candidates was introduced. Exons that might not be identified in the first round might then be found in the second, third, or later round. Of course, the recursive depth should not be too large to avoid the inclusion of false positive exons because of the decreasing stringency of the query exons. Including every first exon of the Dscam mutually exclusive exon clusters in the query sequence, all twelve exons of the exon 4 cluster were identified, both exons of the exon 17 cluster, and 46 and 32 exons for the exon 6 and exon 9 cluster, respectively (Figure 2.2-3, Table 2.2-1). Increasing the recursive depth to one also revealed exon 6.11, which is the most divergent exon of the cluster, and which has not been detected in transcriptome studies yet [142–144]. Exon 6.47 was not identified because the intron before exon 6.47 does not have an "AG" at the 3'-end and is therefore not compatible with the "GT" at the 5'-end of the intron succeeding exon 5. The supposed 5'-end sequence of exon 6.47 is different to the published sequence [139] but is supported by many genomic DNA reads available from GenBank (a genomic DNA read identical to the published sequence was not found). Exon 9.13 was also not identified because it contains a frame shift in the *Drosophila* reference genome assembly, supported by many genomic DNA reads. Therefore, the translations of the predicted transcripts containing exon 9.13 all stop shortly behind this frame shift (e.g. NM_001043054.1, NM_001043034.1, and NM_001043065.1). However, both exon 6.47 and exon 9.13 were identified in many transcripts [142–144]. Thus, either the genome assembly based on the many genomic DNA reads is wrong, which is unlikely, or the many EST/cDNA-reads are wrong, which is also unlikely, or the genomic DNA has been obtained from a different strain than the one that has been used in the transcriptome studies. WebScipio is, however, not able to identify mutually exclusive exons if those do not correspond to the exon length (e.g. frame shifts will result in other reading frames and exon lengths) and corresponding splice site restrictions. The strength of the new algorithm is illustrated at the exon 17 cluster that encodes two highly divergent but mutually exclusive spliced exons (Figure 2.2-3). When applying the search for mutually exclusive exons in the Dscam gene against the published genomic sequence (NCBI accession number AF260530 [139]) all proposed 95 mutually exclusive exons were identified (Table 2.2-1). Less mutually exclusive exons in the search against the *Drosophila melanogaster* reference genome sequence compared to the search against the published sequence are therefore not due to problems with the search algorithm.

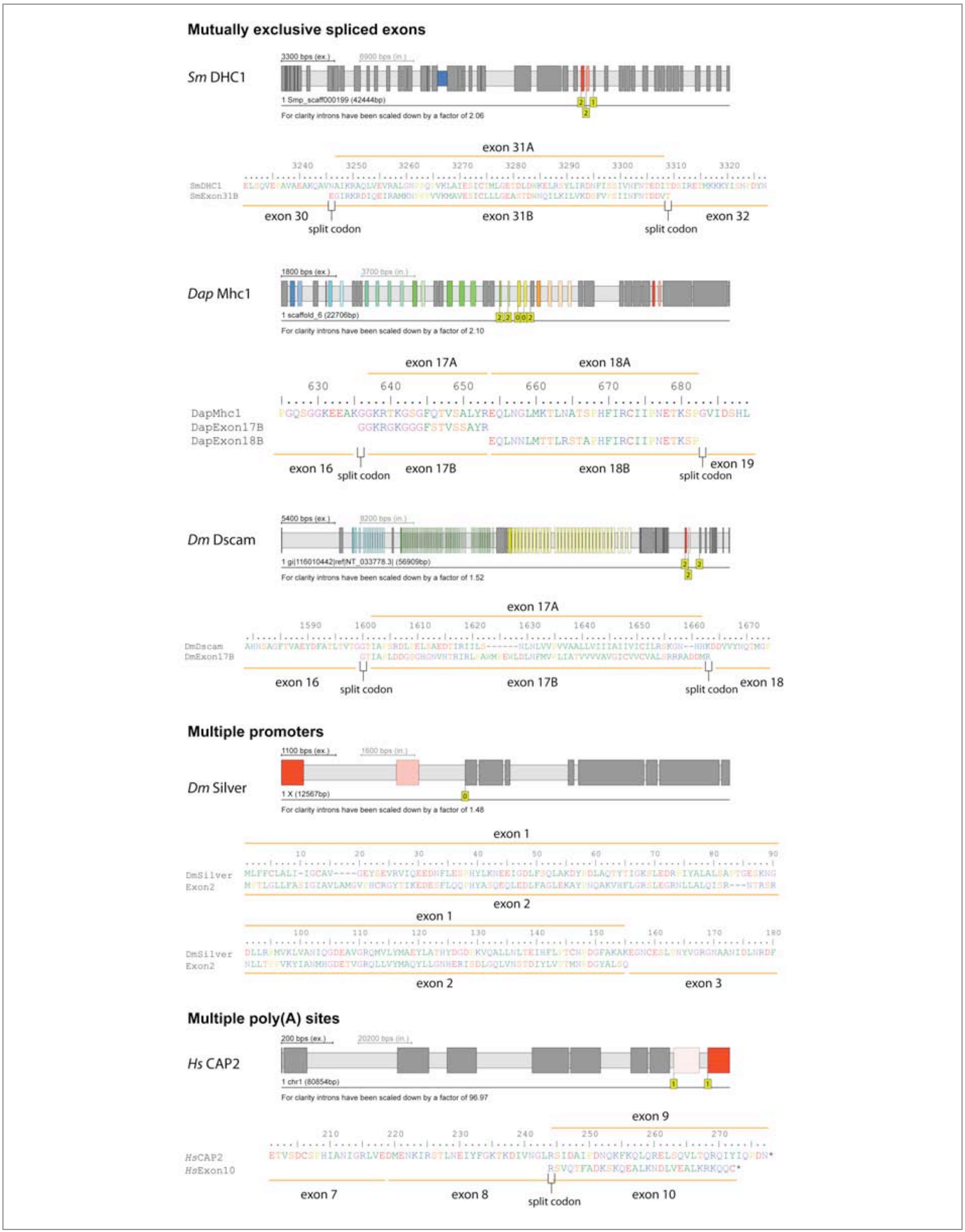


Figure 2.2-3 | Example cases of mutually exclusive spliced exons, multiple promoters and multiple poly(A) sites. The figure illustrates three examples of genes containing mutually exclusive spliced exons, one example containing multiple promoters, and one containing multiple poly(A) sites. Dark grey bars and light grey bars mark exons and introns, respectively. The small blue bar represents an "intron?" that does not have canonical splice sites because an exon is missing in the assembly. Coloured big bars represent mutually exclusive exons found by the new algorithm. The darkest coloured bar is the exon that was included in the query sequence, while the lighter coloured bars represent identified mutually exclusive exons. The higher the similarity between the candidate and the query exon the darker will be the colour of the candidate (100% identity would result in the same colour). Yellow boxes with numbers indicate the reading frame of the corresponding exon.

Table 2.2-1 | Mutually exclusive exons in the *Drosophila* species Dscam genes

exon	Dm	AF260530	Dse[a]	Dy	Der	Da	Dp	Drp[a]	Dw	Dmo	Dv	Dg
4	12	12	12	12	12	12	12	10	12	12	12	12
6[b]	46/47	47/48	46[c]	39/40	44	47	49	49	48/49	50	52	53
9	32	33	29	32	33	33	32	29	29	32	32	32
17	2	2	2	2	2	2	2	2	2	2	2	2
total	92/93	94/95	90	85/86	91	94	95	90	91/92	96	98	99
[68, 70]	95	95	95	87	94	93	94	95	95	95	98	94
[69]	95	95		88		93	94			95	98	

Dm = *Drosophila melanogaster*; Dse = *Drosophila sechellia* Rob3c; Dy = *Drosophila yakuba* Tai18E2; Der = *Drosophila erecta* TSC#14021-0224.01; Da = *Drosophila ananassae* TSC#14024-0371.13; Dp = *Drosophila pseudoobscura* MV2-25; Drp = *Drosophila persimilis* MSH-3; Dw = *Drosophila willistoni* TSC#14030-0811.24; Dmo = *Drosophila mojavensis* TSC#15081-1352.22; Dv = *Drosophila virilis* TSC#15010-1051.87; Dg = *Drosophila grimshawi* TSC#15287-2541.00.

[a] Genomes are fragmented and contain gaps in the Dscam genes.

[b] The first number corresponds to a search with standard parameters, the second to searches with one round of recursion.

[c] One of the exons is a pseudo-exon because it misses the last forth of the exon because of an in-frame stop codon.

In addition, mutually exclusive exons in the Dscam genes of the other sequenced *Drosophila* species were searched ([100]; Table 2.2-1). Here, all mutually exclusive exons were found immediately, and only three further exons were identified by a second recursive round of exon search. As found for the *Drosophila melanogaster* gene, WebScipio identified sometimes more sometimes less exons compared to the published analyses [68–70]. However, the WebScipio searches were performed against the official reference genome assemblies, while the published analyses were based on manually performed genomic clone assemblies of the Dscam gene regions. Therefore, the differences in exon numbers do not result from shortcomings of the search algorithm, but from differences in the assembly of the reference genome data and the manually assembled genomic regions.

Example genes encoding 5'- and 3'-terminal exons with features of mutually exclusive spliced exons

Terminal exons are often not selected at the level of splicing. Instead, initial (5’-terminal) exons are most probably selected at the level of transcription that starts at different promoters. Terminal exons (or better alternative exons encoding for the terminal stop codon) might either be spliced as differentially included exons, like in the case of the *Drosophila* muscle myosin heavy chain gene [67], or as multiple poly(A) sites. Nevertheless, these terminal exons might contain an important structural part of the encoded protein and thus often have similar length and show sequence similarity. Figure 2.2-3 shows two examples of genes that contain 5’- and 3’-terminal exons sharing the described features of mutually exclusive exons, but are spliced as multiple promoters or multiple poly(A) sites. The silver protein of *Drosophila melanogaster* illustrates a case where two initial exons, which are transcribed/spliced as multiple promoters, share the features of mutually exclusive exons. The capping protein beta (Capβ) from *Homo sapiens* represents a case where homologous 3'-terminal exons containing multiple poly(A) sites are found. The detection of these cases can be suppressed by disabling the search for mutually exclusive exons for 5'- and 3'-terminal exons. By default, WebScipio enables the search for homologous exons for all exons, because it is not known whether the user is searching with a complete, partial or fragmented query sequence. In the case of partial and fragmented sequences the search would provide significant results. Also, genes sometimes contain untranslated 5'- and/or 3'-terminal exons whereby the first translated exon could well be part of a cluster of mutually exclusive spliced exons. In addition, alternative terminal exons by themselves might provide interesting perspectives to the corresponding genes independently of whether they are mutually exclusively spliced or not. WebScipio cannot distinguish between the described cases and thus the user has to be careful when alternative terminal exons are proposed.

Detection of *trans*-spliced genes and arrays of tandem gene duplications

The *trans*-splicing of separate pre-mRNAs involving coding exons to reveal a joined transcript is a relatively uncommon event [53]. In general, *trans*-spliced genes in *Drosophila melanogaster* can be distinguished into those with multiple first exons or multiple 3'-terminal exons, or those with very large introns. Many of the *trans*-spliced genes contain variable single terminal exons (e.g. *mod(mdg4)* [145, 146] or *lola* [147]) or alternative terminal exon groups (e.g. CG42235 [53]). When searching for mutually exclusive spliced exons based on one of the annotated isoforms of a *trans*-spliced gene potentially alternative exons of internal exons might be identified. An example of the *trans*-spliced *Drosophila melanogaster* gene CG1637 is shown in Figure 2.2-4A. Three isoforms of the CG1637 gene exist (Isoform A, B, and C) that result in transcripts of a common 5' exon spliced to isoform-specific sets of three 3' exons. The sequences of the isoform-specific sets are homologous although the intron positions are different between the isoform A/B exons and the isoform C exons. When searching with the isoform A exons for mutually exclusive exons in surrounding introns the homologous exon of isoform B is found for the first of the three isoform A-specific exons (Figure 2.2-4A-I). When only searching in surrounding introns (search in up- and downstream regions disabled) further exons are not found for isoform B (homologous exons would only exist in the downstream region, Figure 2.2-4A-II) and for isoform C (the introns are at different positions so that the similar-length condition does not apply anymore, Figure 2.2-4A-III). Thus, if only isoform A were known a mutually exclusive exon would have been proposed. To avoid the mis-annotation of exons of *trans*-spliced clusters a parameter was introduced that allows searching for candidate exons not only in the neighbouring introns but also in all introns. In Figure 2.2-4A-IV the exons of isoform B were identified by searching with the exons of isoform A in all introns revealing the *trans*-spliced nature of the cluster.

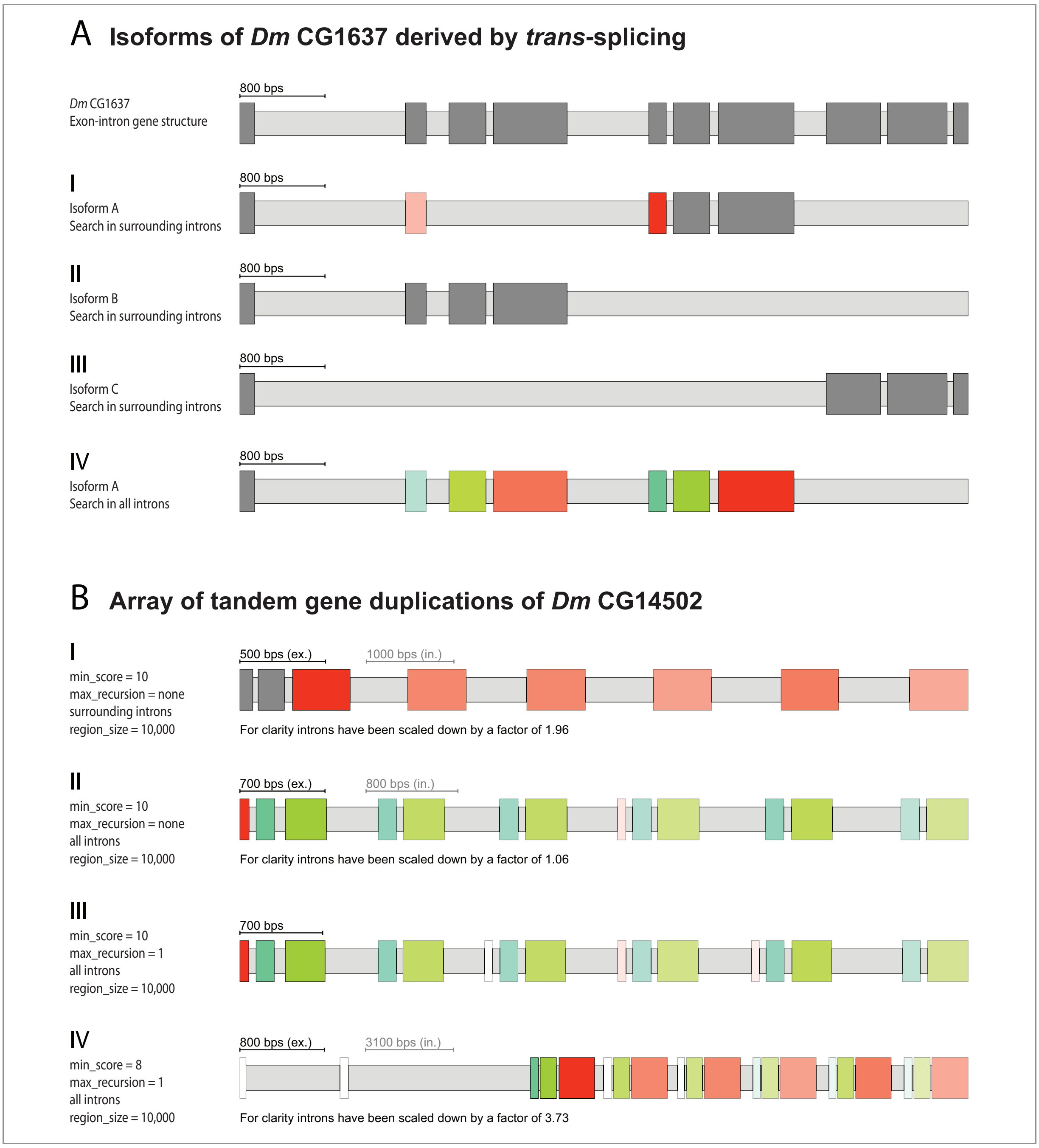


Figure 2.2-4 | Examples of a *trans*-spliced gene and an array of tandem gene duplications. A) Schematic representation of the *trans*-spliced *Drosophila melanogaster* CG1637 gene. The three annotated isoforms A-C are shown consisting of the common 3'-terminal start exon and different groups of alternative exons. If only isoform A were known a potentially mutually exclusive exon would have been found by a search for candidates in surrounding introns (case I). However, a search for candidates of all exons in all introns reveals the two groups of homologous exons that are *trans*-spliced in isoform A and B (case IV). Isoform C also encodes a cluster of *trans*-spliced exons whose sequence is homologous to that of isoform A/B. However, the exonic sequence is interrupted at different intron positions (case III). Note, that the gene structure annotated by Flybase (shown here) is different to the published one ([70], supplementary Figure 3). B) Gene duplications of the *Drosophila melanogaster* CG14502 gene. The figure shows the tandem arrangement of the duplicated genes of the *Drosophila melanogaster* CG14502 gene as found by WebScipio. The parameters *minimal score for exons*, *maximal recursion depth*, *search in all introns* and *region size* were adjusted for each search. With less restrict parameters less similar exons are found.

If searching in up- and downstream regions for alternatively spliced exons, it is possible that candidate exons belong to gene duplicates (Figure 2.2-4B). In this case, the WebScipio option to search for candidates in all introns including up- and downstream regions and not only in surrounding introns helps identifying exons of gene duplications. In many cases, gene duplications result in genes arranged in tandem. Those gene duplicates often share the complete gene structure meaning that for every exon there is a corresponding exon in the duplicated gene. Figure 2.2-4B illustrates this behaviour and provides means by which users can judge between a true cluster of mutually exclusive exons belonging to one gene and a set of duplicated genes. If the search for candidate exons is only allowed in surrounding introns, a set of six homologous exons is found for the *Drosophila melanogaster* gene CG14502 (Figure 2.2-4B-I). Performing the search in all introns results in five homologous exons also for the second exon of the CG14502 gene, and shows one homologous exon for exon 1 (Figure 2.2-4B-II). The first exons of the genes seem to be very divergent. Allowing one additional recursive round of candidate search reveals the first exons for two additional gene homologs (Figure 2.2-4B-III). In addition lowering the score reveals the exon 1 candidates of the remaining two gene homologs, although two further regions with very low homology to exon 1 appear in the upstream region of the CG14502 gene (Figure 2.2-4B-IV). This example illustrates the use of the search parameters so that gene duplications can be identified. Gene duplicates that are not arranged in tandem but are distributed in the genome do not provide problems in evaluating exon candidates, because the search is restricted to a certain size of the up- and downstream regions. If needed, these gene duplicates can be identified with WebScipio using the general *multiple results* option.

Application of the search algorithm for mutually exclusive exons to genome scale data

The described search algorithm identifies three types of exons as described above: (A) mutually exclusive exons, (B) terminal exons that are spliced as multiple promoters or multiple poly(A) sites but share similar length, reading frame, and sequence homology, and (C) exons with the characteristics of mutually exclusive exons that are actually part of tandemly arrayed gene duplicates or groups of alternative exons in *trans*-spliced genes. Type B and type C exons are false positives, when looking for mutually exclusive exons. In addition, false positive exons are those exons that show all characteristics of type A exons but are constitutively or differentially included spliced. False negatives exons, which are not identified by WebScipio, are those mutually exclusive exons that do not have similar length and sequence homology. To quantify the amount of each of these exon types we searched the complete X chromosome of the fruit fly *Drosophila melanogaster* for mutually exclusive spliced exons with WebScipio and compared the results to the Flybase annotation.

Protein sequences for the search were obtained from the Flybase annotation (version 5.27) and mapped to the genomic sequence of the X chromosome using Scipio. 2,967 transcripts containing more than one exon were derived from 1,705 genes. For each exon mutually exclusive alternative splice variants have been searched for in the surrounding introns. The search parameters were set to 20 amino acids for the *allowed length difference*, to 15% for the *minimal score for exons*, and to 15 amino acids for the *minimal exon length*. We did not search for alternative exons in up- and downstream regions of genes, and we did not apply the recursive search, which means the repeated search for further alternative exons with the newly identified exons that we demonstrated for Dscam (see above). Three genes (lethal (1) G0193, CG1637, and CG42249), in which mutually exclusive exons were found, were excluded from the analysis, because the respective exons are spliced in a mutually exclusive manner in groups of two, three, and four exons, instead of single exons within a cluster. Those genes are probably *trans*-spliced (for an example see Figure 2.2-4A).

Search for non-mutually exclusive exons sharing similar length, same reading frame, and sequence homology

It could well be possible that internal exons with similar length, same reading frame, and showing sequence homology are not mutually exclusive spliced exons, but constitutive exons or exons spliced by one of the other types of alternative splicing. To get a statistically relevant number of these types of exons we collected all genes of the *Drosophila melanogaster* X chromosome containing at least two exons based on the Flybase annotation version 5.27. The transcripts of each gene were analysed independently because alternative splicing produces different exon neighbours. Thus exons are counted for each transcript (not each gene) even if the transcripts have the same start and end points in the genomic sequence. In total, the 2,967 transcripts of the *Drosophila melanogaster* X chromosome include 16,180 exons. All neighbouring exons were compared with respect to having similar length (*allowed length difference* 20 aa), sharing high similarity (*minimal score for exons* 15%), coding for at least fifteen amino acids (*minimal exon length* 15 aa), and encoding the same reading frame. The results are summarized in Table 2.2-2 (for detailed information see Supplementary information 2.2-2). Only 0.56% of the non-mutually exclusive exons (90 out of 16180) share the features of mutually exclusive exons. These exons are located in only six genes out of 1705 (0.35%). In one of the six genes (*Ciboulot*) the two homologous exons are terminal exons and would represent a case of multiple poly(A) sites if alternatively spliced. This analysis shows that the chance that the exons predicted by WebScipio as mutually exclusive exons will later (e.g. after obtaining cDNAs) be reannotated as constitutive or differentially included exons, is very low.

Table 2.2-2 | Search for exons annotated as constitutively spliced or differentially included sharing similar length, same reading frame and sequence homology in the *Drosophila melanogaster* X chromosome

	Total	Hits[a]	Percentage
Exons	16180	90	0.56%
Transcripts	2967	20	0.67%
Genes	1705	6	0.35%

[a] Exons (or transcripts/genes containing exons) which share similar length, same reading frame and sequence homology.

Search for mutually exclusive spliced exons in the *Drosophila melanogaster* X chromosome

Some categories have to be defined to separate true (annotated) mutually exclusive spliced exons from predicted ones and false positives and false negatives. As real mutually exclusive exons we regard those with the following criteria: An exon is part of a cluster of mutually exclusive spliced exons if each transcript of the gene contains exactly one exon of the cluster (not none or more than one), the cluster contains at least two exons, the exons of the cluster are neighbouring exons, and the cluster is surrounded by further exons. The latter criterion distinguishes the mutually exclusive spliced exons from clusters of initial exons (5'-terminal exons) and 3'-terminal exons that are spliced in a mutually exclusive manner and share sequence similarity, similar length, and splice site conservation. In contrast to real mutually exclusive spliced exons the exons of these clusters appear mutually exclusive in the transcripts but their transcription and splicing is regulated in a different way. These clusters are therefore regarded as types of multiple promoters and types of multiple poly(A) sites, and are false positives. Other types of false positives are those exons that are predicted by WebScipio but overlap with already annotated exons and do not match exactly the positions of these exons. False negatives are those exons that do not meet the preconditions of similar length and sequence homology. However, if those exons are mutually exclusive spliced they must have conserved splice sites and reading frames.

In total, 94 exons of similar length, same splice sites and reading frames, and sequence homology have been identified by WebScipio, of which 65 are potentially in clusters of mutually exclusive exons, 21 in clusters of multiple promoters, and 8 in clusters of multiple poly(A) sites (Figure 2.2-5). Of the 65 exons predicted to belong to internal clusters of mutually exclusive spliced exons, 26 exons are already annotated in Flybase. 39 exons are predictions by WebScipio that have not been described before. These 39 exons are distributed into 18 clusters that belong to 17 genes. Thus, there are several clusters with more than two alternative exons, and one gene with two clusters. If the Flybase based annotation is assumed to represent true mutually exclusive exons, the chart represents the specificity of our method. The 26 already known mutually exons divided by 65 predicted exons result in 40% specificity. However, the value for the specificity is misleading because it depends on the "known" mutually exclusive exons. We expect that many of the additional exons predicted by Web-

Scipio will be experimentally confirmed in the future and thus will become "known" mutually exclusive exons. The true specificity will therefore be much higher than the value of 40% suggests. To analyse whether the additional exons predicted by WebScipio contain general features of exons (for example a higher GC content than the surrounding region), the found exons were compared to those of an *ab initio* prediction performed by AUGUSTUS [119] (Figure 2.2-5). In many cases the WebScipio predictions are supported by the *ab initio* prediction, which is based on the genomic sequence alone. The AUGUSTUS prediction matches 27 of the 94 exons with exact exon borders (Figure 2.2-5, orange numbers) and overlaps with 46 of them (Figure 2.2-5, yellow numbers).

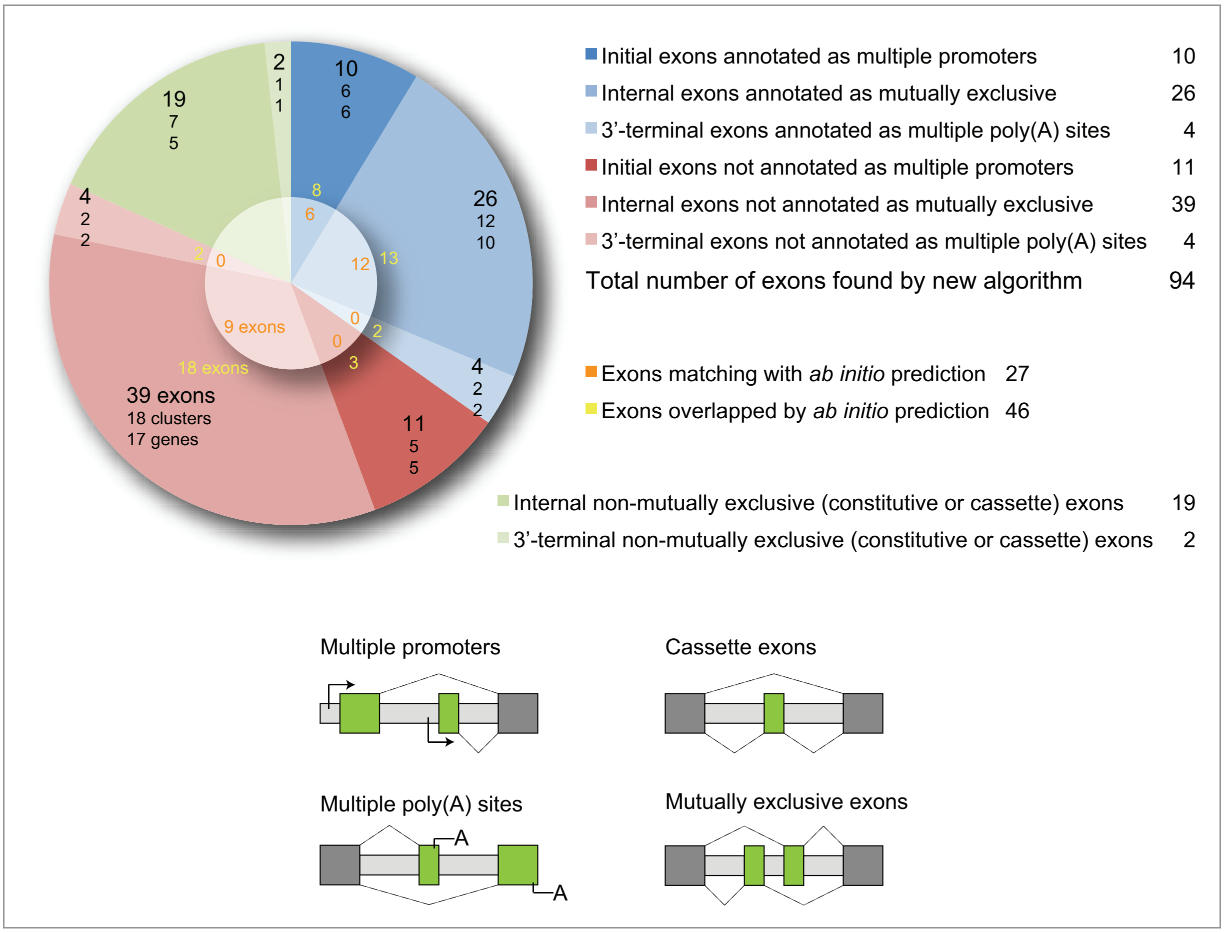


Figure 2.2-5 | Exons located on the *Drosophila melanogaster* X chromosome sharing similar length, same splice sites and reading frames, and sequence similarity. The pie chart shows the total number of exons of the *Drosohpila melanogaster* X chromosome, which share the features used by the new search algorithm. The blue and red slices represent the number of exons found by the new algorithm compared to existing annotations and to the *ab initio* prediction by AUGUSTUS, shown in the middle. The blue part illustrates the exons already annotated by Flybase, in contrast to the exons in clusters additionally predicted by WebScipio in red. The pie is divided in slices for initial, internal, and 3'-terminal exons. In addition to the number of exons, the chart indicates the number of clusters and genes, in which these exons were found. The orange numbers in the middle part of the pie indicate how many of the respective exons are found and reconstructed with correct exon borders by the *ab initio* prediction with AUGUSTUS, while the yellow numbers reveal the number of exons to which exons predicted by AUGUSTUS at least overlap. The green slices indicate constitutive exons, which share the features of mutually exclusive exons. These are the same exons, clusters, and genes as listed in Table 2.2-2 and Supplementary information 2.2-2. At the bottom, the figure illustrates the different types of alternatively spliced exons (multiple promoters, multiple poly(A) sites, mutually exclusive exons) in comparison with the cassette exon type.

The results show that about 70% of all predicted exons (65 out of 94) comprise clusters of internal mutually exclusive exons. The false positive prediction of 5'- and 3'-terminal exons as mutually exclusive exons, which comprise the remaining 30% of predicted exons, could even be suppressed by a WebScipio option. We can also conclude that WebScipio correctly identifies all but one (see following section) of the annotated mutually exclusive exons. This suggests that most of the WebScipio predictions of new mutually exclusive exon candidates will also be real mutually exclusive exons. This is supported by the *ab initio* exon prediction by AUGUSTUS that showed exon probability for about 50% of the newly predicted exons, which is comparable to the *ab initio* prediction of the already annotated exons. However, we cannot completely exclude the possibility that some of the newly predicted exons might in truth be constitutive or differentially included exons (see previous section).

False negatives would be those mutually exclusive spliced exons that do not share a similar length and sequence similarity. To figure out how often clusters of mutually exclusive exons with such characteristics exist in comparison to mutually exclusive exons with similar length and sequence similarity, all internal clusters of exons on the X chromosome that were annotated as mutually exclusive based on Flybase transcripts were manually analysed (Figure 2.2-6). Of the annotated genes only the Phosphorylase kinase γ gene contains two mutually exclusive spliced exons that do not have similar length and sequence (Figure 2.2-6, bottom). If the Flybase annotation is assumed as true, the chart in Figure 2.2-6 represents the sensitivity of the algorithm. 26 predicted mutually exclusive spliced exons divided by 28 annotated exons results in 93% sensitivity for internal exons. These data likely indicate that not many mutually exclusive spliced exons will be missed given the constraints of similar length and sequence similarity as implemented in WebScipio.

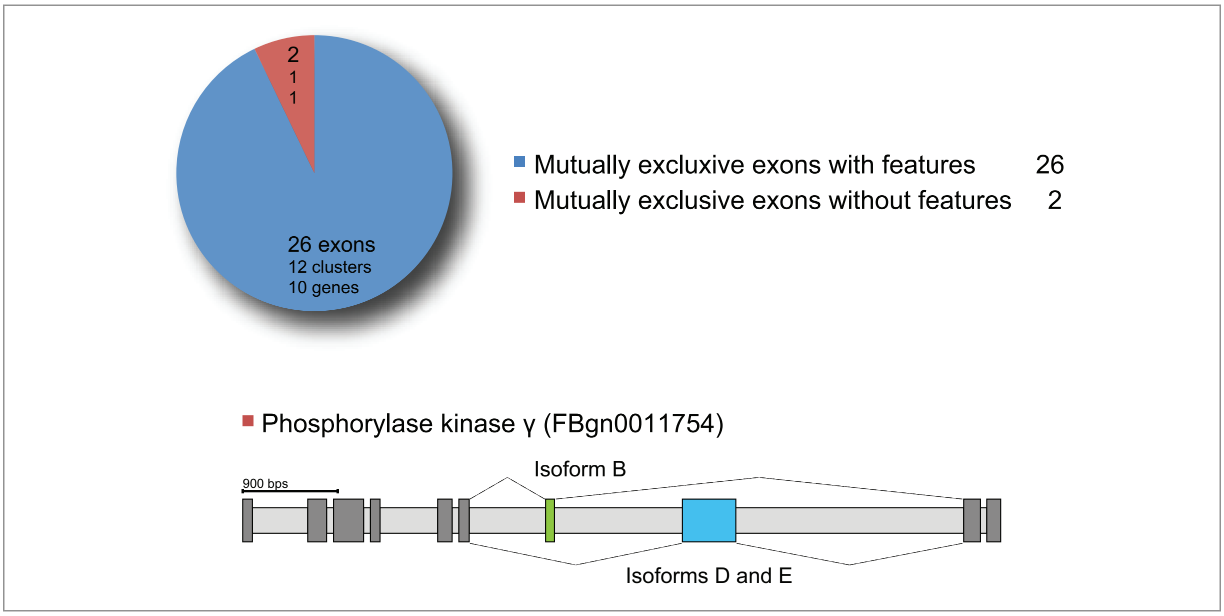


Figure 2.2-6 | Mutually exclusive exons in genes of the *Drosophila melanogaster* X chromosome. The figure illustrates how many of the mutually exclusive exons, which were annotated based on Flybase transcripts, share the following features: high sequence similarity, similar length, same reading frame, and a minimal exon length (15 residues). The blue slice indicates exons characterised by these features and found by the new algorithm. The red slice indicates exons not sharing these features. At the bottom, the figure shows the exon-intron structure of the Phosphorylase kinase gamma gene, which includes the only cluster of mutually exclusive exons that was not found by the new algorithm.

Mutually exclusive exons predicted for 5'- and 3'-terminal exons were regarded as false positives because these rather present cases of multiple promoters, multiple poly(A) sites, and differentially included exons. However, additional untranslated terminal exons might exist that were not analysed here, and in those cases the exons, based on the translation predicted as terminal, become internal and thus true mutually exclusive exons. For comparison all terminal exons annotated as transcribed or spliced in a mutually exclusive manner have been analysed (Figure 2.2-7). Of the 101 terminal exons only 14 terminal exons share the features of mutually exclusive spliced exons. A reason for the sequence and length variability of terminal exons is that the N- and C-termini of proteins are not as restricted in their structure as internal parts. Thus, the number of false positives predicted by WebScipio is rather low.

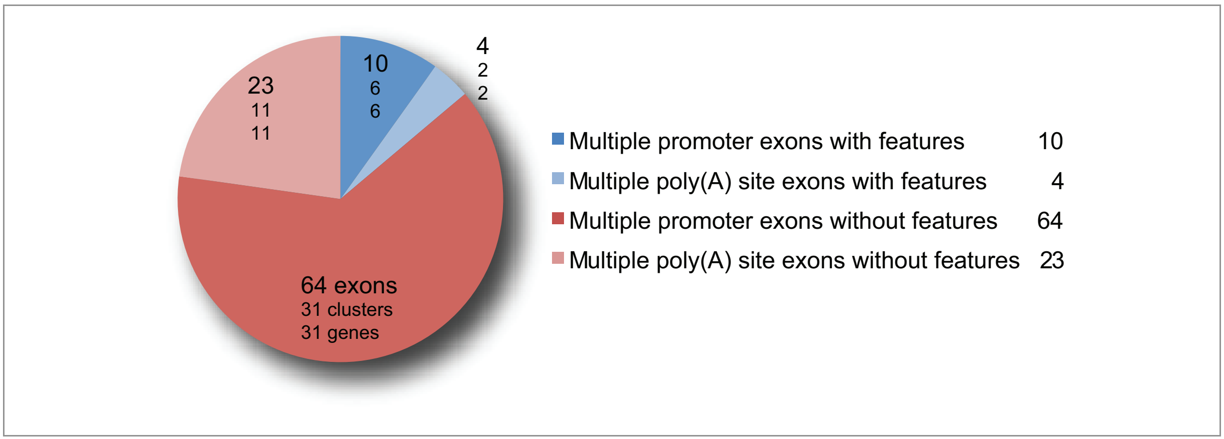


Figure 2.2-7 | Exons belonging to clusters of multiple promoters and multiple poly(A) sites in the *Drosophila melanogaster* X chromosome. The figure shows the number of multiple promoter exons and multiple poly(A) sites exons based on the Flybase annotation and illustrates how many of these exons share the following features: high sequence similarity, similar length, same reading frame, and a minimal exon length (15 residues). Blue slices indicate exons characterised by these features, and red slices indicate exons not sharing these features.

Future developments and applications

Due to the precondition that mutually exclusive exons encode the same part of the protein product, we also want to include the comparison of the prediction of secondary structural elements for the query and the candidate exons as an additional scoring, analysis, and validation parameter. Also, other substitution matrices might be offered for the scoring of the aligned query and candidate exons. Scipio and WebScipio have been shown to be suitable for the prediction of genes in cross-species searches [36, 109]. Of course, both approaches can be combined and users can search, for example, with a human protein query sequence in other mammals to identify homologous genes and simultaneously predict mutually exclusive exons in the target sequence. Because the search for mutually exclusive exons relies on the translation of the exons as found in the genomic DNA, it does not depend on the initial query sequence but on the quality of the exons identified in the cross-species search. Another application would be to search for mutually exclusive spliced genes in the complete genomes of sequenced eukaryotes.

2.2.5 Conclusions

The extension of WebScipio to search for mutually exclusive exons is based on the precondition that these exons encode regions of the same structural part of the protein product. This precondition provides restrictions to the search for candidate exons concerning their length, splice site conservation and reading frame preservation, and overall homology. The implemented algorithm has been shown to identify all known mutually exclusive spliced exons in many example genes from various species, like the muscle myosin heavy chain gene of *Daphnia pulex* or the Dscam gene of *Drosophila melanogaster*. The search for homologs of terminal exons might, however, result in the prediction of multiple promoters, multiple poly(A) sites, groups of *trans*-spliced exons, or tandemly arrayed gene duplicates, and can therefore optionally be disabled. To quantify the quality of WebScipio to correctly predict already annotated mutually exclusive exons and to predict so far unrecognized exon candidates, an analysis of the whole X chromosome of *Drosophila melanogaster* has been performed. All but two of the 28 annotated mutually exclusive exons were found by WebScipio. In addition, WebScipio predicts 39 new mutually exclusive exon candidates of which about 50% are supported by an *ab initio* exon prediction by AUGUSTUS. In conclusion, WebScipio should be able to identify mutually exclusive spliced exons in any query sequence from any species with a very high probability.

2.2.6 Abbreviations

DHC: Dynein heavy chain; Dscam: Down Syndrome Cell Adhesion Molecule; GFF: General Feature Format; Mhc: Myosin heavy chain; SVG: Scalable Vector Graphics; YAML: YAML ain’t markup language

2.2.7 Acknowledgements

MK has been funded by grants KO 2251/3-1, KO 2251/3-2, and KO 2251/6-1 of the Deutsche Forschungsgemeinschaft.

2.2.8 Authors' contributions

HP, FO, and MK set the requirements for the system. HP and KH wrote the software. FO assisted in and supervised the implementation of the software. BH implemented improvements to the software, and KH committed the final version. KH performed the whole chromosome search and evaluation. HP, KH, BH, and MK performed extensive testing. KH and MK wrote the manuscript. All authors read and approved the final version of the manuscript.

2.2.9 Supplementary information

Supplementary information 2.2-1 | Detailed activity diagram. The detailed activity diagram shows each step of the search algorithm including points of decision and loops. This supplementary file is available from the corresponding publication.[48]

Supplementary information 2.2-2 | Search for non-mutually exclusive exons sharing similar length, same reading frame and sequence homology. The file provides detailed information of the found genes and their gene structures. This supplementary file is available from the corresponding publication.[49]

[48] http://www.biomedcentral.com/content/supplementary/1471-2105-12-270-s1.pdf

[49] http://www.biomedcentral.com/content/supplementary/1471-2105-12-270-s2.pdf

2.3 Predicting tandemly arrayed gene duplicates with WebScipio

Klas Hatje and Martin Kollmar

Abteilung NMR-basierte Strukturbiologie, Max-Planck-Institut für biophysikalische Chemie, Am Fassberg 11, D-37077 Göttingen, Germany

Gene Duplication, Prof. Felix Friedberg (Ed.), *InTech* 2011

Published: 17 October 2011
doi: 10.5772/24240
This article is available from
http://www.intechopen.com/books/gene-duplication/predicting-tandemly-arrayed-gene-duplicates-with-webscipio

2.3.1 Introduction

Since the first high-quality eukaryotic genome assemblies became available the large scale analysis of the origin of new genes came into the focus of many studies [148, 149]. New genes can originate through multiple mechanisms including gene duplication, gene fusion/fission, exon shuffling, retroposition, horizontal gene transfer, and de novo from noncoding sequences [150]. Although initial models proposed that new copies of genes soon become nonfuntional [151, 152] it has since been shown for numerous genes that they retain function through creating redundancy, subfunctionalization, and neofunctionalization [153–155]. While de novo origination from noncoding sequence has been shown to play an unexpectedly important role [149] most of the new genes are derived through duplications. Gene duplicates are normally classified into dispersed and tandem duplicates. Tandem duplications of clusters of genes, single genes, groups of exons, or single exons are thought to be formed by unequal crossing-over events, or misaligned homologous recombinational repair [156, 157]. A comparative analysis of the human, mouse, and rat genome has shown that about 15% of all genes represent tandemly arrayed genes [148]. A similar number of about 20% has been found for the fruit fly *Drosophila melanogaster* [158]. All these analyses rely on the particular dataset of annotated genes used and the specific methods for defining genes as tandem genes. However, first annotations of genomes are in most cases done by automatic gene prediction programs, nowadays often supported by incorporating additional EST data, and therefore miss many genes, include artificially fused neighbouring genes, and contain mis-predicted exons and introns. Although these errors seem small, in the case of distinguishing tandem gene duplicates from genomic region duplication and *trans*-spliced genes they are essential. In addition, defining tandem genes by a certain number of nucleotides appearing in-between cannot separate tandem gene duplicates from duplications of small genomic regions. Tandemly arrayed gene duplicates are often conserved between species. Examples are the olfactory receptor genes that constitute a very large gene family of several hundred genes per species in vertebrates [159] and the HOX genes [160, 161]. While algorithms have been developed to reconstruct the history and evolution of tandemly arrayed genes [162, 163] specific programs are not available for the prediction and local reconstruction of these gene arrays.

WebScipio is a web application to reconstruct genes based on a given protein query sequence and a genomic DNA target sequence [109]. The reconstruction is done with Scipio [36], a post-processing script for the output of a BLAT run [29]. BLAT is a very fast tool for the alignment of protein or DNA sequences if these sequences are almost identical. However, BLAT is not able to reconstruct intron and exon borders, it does not identify very short exons and very divergent exons, and it is not able to reconstruct genes spread on several pieces of contiguous DNA (contigs), which is very common in low-coverage genome assemblies. Furthermore, BLAT is not able to identify sequencing and assembly errors like additional or missing bases in exon regions or base substitutions leading to in-frame stop codons. Scipio is able to correct all these errors and extend the BLAT output for the missing sequences of short or divergent exons and of exon borders. In addition, Scipio assembles genes spread on several contigs. WebScipio has been developed as a web interface to Scipio so that the user does not have to install scripts and libraries. Moreover, WebScipio offers access to about 2300 genome assembly files of more than 650 sequenced eukaryotes (July 2011), and provides graphical and human-readable analyses of the results.

Here, we present an extension to the WebScipio web application to search for and predict tandemly arrayed gene duplicates for a given query sequence. This extension is not available via the Scipio command-line script. The user can search for gene duplicates in hundreds of species for which reliable annotations are not available yet, because WebScipio provides access to thousands of genome files.

2.3.2 Implementation

The new algorithm to predict tandemly arrayed gene duplicates is fully integrated into the web application WebScipio to make it usable for the inexperienced user and to visualize the results for immediate analysis. It was implemented in the Ruby programming language[50] using the BioRuby library [95] to handle sequences. WebScipio is based on the web frame-

[50] http://www.ruby-lang.org

work Ruby on Rails[51], which includes the Javascript libraries Prototype[52] and Scriptaculous[53]. To keep the web application responsive, the search algorithm runs in the background with the help of the Ruby on Rails plug-ins Workling[54] and Spawn[55]. To store the user session data, the database backend Tokyo Tyrant is used in combination with Tokyo Cabinet[56]. The results of the search are presented as SVG[57] pictures and several human-readable representations, most notably a detailed alignment of protein query, target DNA sequence, and target translation. The raw results can be downloaded as General Feature Format (GFF) files or as YAML[58] files for future upload and analysis. Specific results are available in various formats for further inspection, like the human-readable log-files, or publication quality figures, like the SVGs.

Search algorithm

The overall workflow of the search algorithm is shown in Figure 2.3-1. The search for tandem gene duplications is based on the exon-intron structure of a gene generated by Scipio. Thus the first step of the algorithm includes a WebScipio run generating a new gene structure or the upload of an existing Scipio result.

Query and target selection

The next steps are the selection of the query and the target for the search. All exons, which are longer than a minimal length, are selected as query. The minimal length can be adjusted by the *minimal exon length* parameter, which is given in number of amino acids coded by the exon. In addition, the algorithm is able to generate exon tuples by the fusion of neighbouring exons to one exon. This means that all pairs (2-tuples) of consecutive exons, triplets (3-tuples), 4-tuples, 5-tuples, up to all exons are concatenated and used as query exons. This option can be enabled by the *search for concatenated exons* parameter. The nucleotide sequences of the up- and downstream regions of the gene are used as target sequences. The lengths of these sequences are determined by the Scipio parameter *region size* in number of nucleotides. The up- and downstream sequences are scanned in forward and reverse direction. For the reverse strand the reverse complements of the given target sequences are created.

[51] http://rubyonrails.org
[52] http://www.prototypejs.org
[53] http://script.aculo.us
[54] http://github.com/purzelrakete/workling
[55] http://github.com/tra/spawn
[56] http://fallabs.com/tokyocabinet
[57] http://www.w3.org/Graphics/SVG
[58] http://www.yaml.org

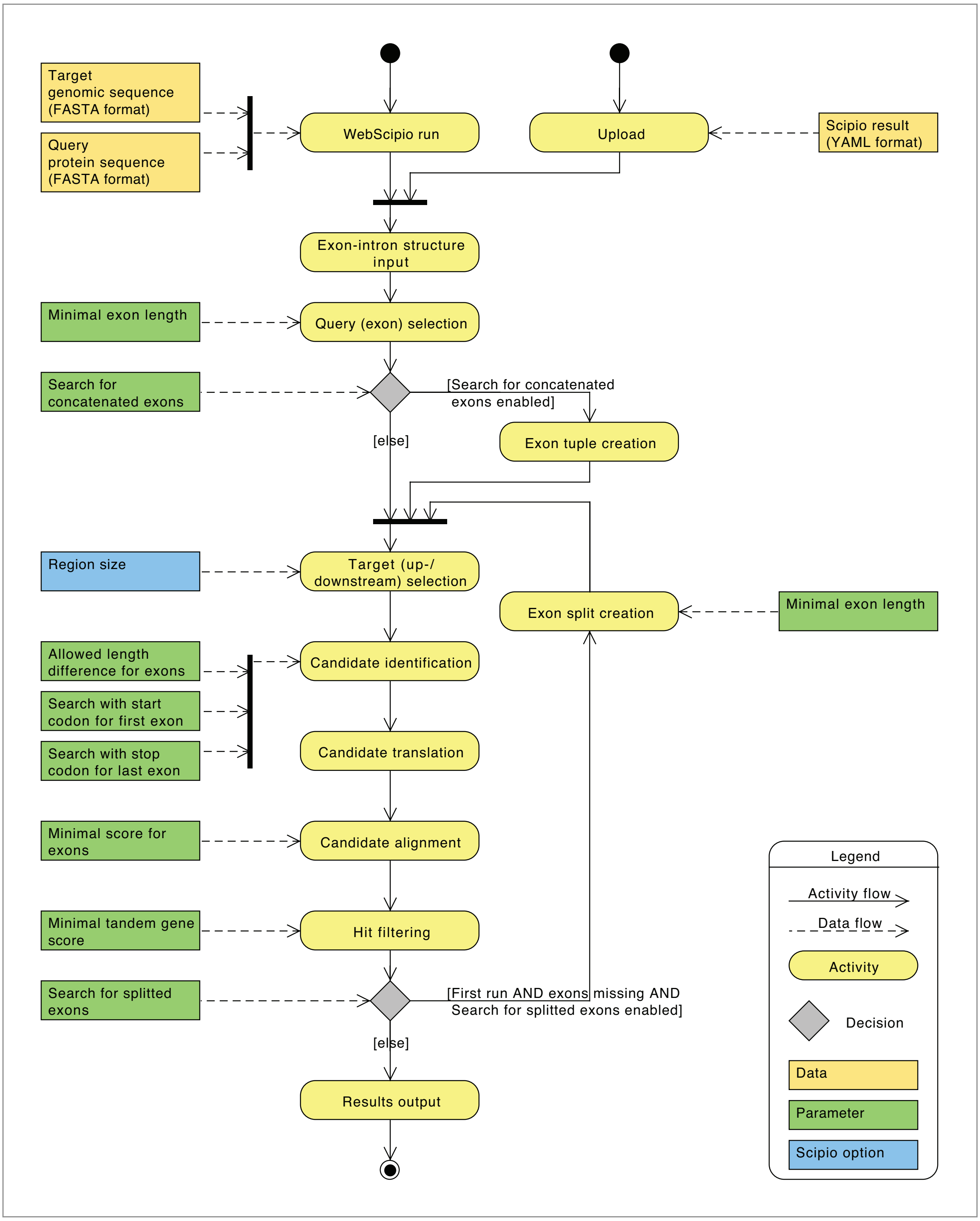

Target genomic sequence (FASTA format)
Query protein sequence (FASTA format)
WebScipio run
Upload
Scipio result (YAML format)
Exon-intron structure input
Minimal exon length
Query (exon) selection
Search for concatenated exons
[Search for concatenated exons enabled]
[else]
Exon tuple creation
Region size
Target (up-/ downstream) selection
Exon split creation
Minimal exon length
Allowed length difference for exons
Candidate identification
Search with start codon for first exon
Candidate translation
Search with stop codon for last exon
Minimal score for exons
Candidate alignment
Minimal tandem gene score
Hit filtering
Search for splitted exons
[First run AND exons missing AND Search for splitted exons enabled]
[else]
Results output
Legend
Activity flow
Data flow
Activity
Decision
Data
Parameter
Scipio option

Figure 2.3-1 | Activity flow diagram of the search for tandem gene duplications. The activity diagram shows the processing steps of the search algorithm and the influence of the parameters on each step. The run starts with an exon-intron gene structure determined by Scipio. Based on the chosen parameters the exons and up- and downstream regions are selected and searched for candidate exons of gene duplicates. The candidates are processed and filtered. These steps are repeated for exons that have not been found. Those exons are splitted and the search is repeated with fragments. In the end, the algorithm outputs the exon-intron structure of the original gene and all gene duplicates.

Candidate identification

The query and target selection steps are followed by the search for exon candidates in the target sequences. The search algorithm assumes that exons of gene duplications have a similar length, share sequence similarity, are translated in the same reading frame and have conserved splice sites. Candidate exons are determined in the target sequences for each exon of the original gene and each exon tuple. The target nucleotide sequences are scanned for sequence sections, which do not differ more than a maximal number of nucleotides from the original exon length. This maximal difference is given by the *allowed length difference* for exons parameter in number of amino acids. In addition, the sequence section, which determines an exon candidate, must be flanked by a splice site pattern that corresponds to the introns surrounding the original exon or exon tuple. Allowed splice site patterns for the first two and last two nucleotides of these introns are GT---AG, GC---AG, GG---AG, and AT---AC. The first exon of a gene must start with the start codon ATG and the last exon must be followed by one of the stop codons TAG, TAA, or TGA. To allow searches for partial genes, the algorithm is able to find candidates corresponding to the first and last exon of the gene fragment that share splice site patterns instead of having a start codon or stop codon. This behaviour can be adjusted by the *search with start codon for first exon* and *search with stop codon for last exon* parameters.

Candidate translation and alignment

Candidate sequences are translated to amino acids in the same reading frame as the original exon. If a candidate sequence includes a stop codon, the candidate is rejected immediately. The translations of the candidate exons are aligned to the original exon translations by a global alignment algorithm. The pair_align tool of the SeqAn package [96] is used for this task. The resulting alignment score is divided by the score resulting from the alignment of the original exon translation to itself. This normalised score makes exons of different lengths and amino acid compositions comparable. Finally, exon candidates having a score lower than the score given by the *minimal score for exons* parameter are rejected.

Hit filtering

The resulting candidate hits are filtered. If candidate sequences are overlapping, the lower scoring candidates are rejected. Neighbouring candidate exons are combined to genes if they are in the same order as the original exons. For each identified tandem gene a score is calculated that reveals how many residues of the original gene were found in the tandem gene duplication. The score is calculated as the number of residues of the original gene that are aligned to residues in the tandem gene duplicate (and not to gaps) divided by the number of all residues of the original gene. The tandem gene duplications that have a low score are rejected. This behaviour can be adjusted by the *minimal tandem gene score* parameter.

Exon split run

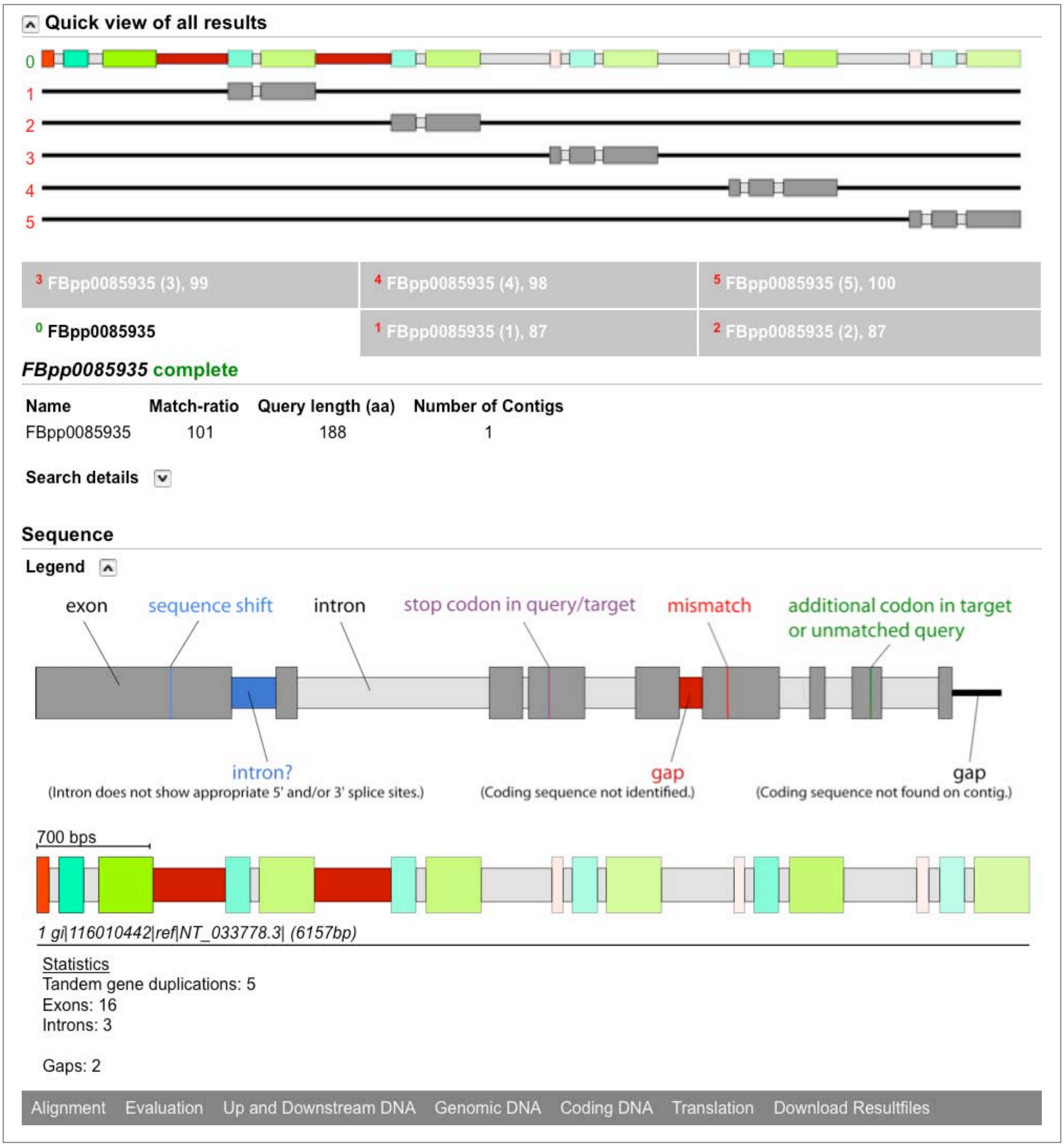
If exons of a duplicated gene are missing, either in between two neighbouring exons, at the start of the gene or at the end, the search is repeated for these exons by splitting the missing original exons into pieces. The original exon sequences are split in two parts at each nucleotide as long as the smaller part is longer than the minimum exon length. The algorithm scans the intron regions of the duplicated genes that miss exons for candidates corresponding to these exon splits, each composed of two parts. Thus, exons, which are split by an intron in the duplicated gene, are found too. This option can be enabled by the *search for splitted exons* parameter.

Results output

The output of the search algorithm is the exon-intron structure of all identified tandem gene duplications combined in one result, and the exon-intron structure of each duplicated gene alone. For every result a gene structure drawing is shown, as well as several options to further examine gene details like the alignment of the query sequence to the translation of the hit and the hit itself (Figure 2.3-2).

WebScipio integration

The search algorithm is fully integrated into the web interface of WebScipio. The search for tandem gene duplications can be enabled in the Advanced Options section. WebScipio provides an interface to easily set the parameters, suggests default parameters, which will be suitable for most cases, and offers documentation at several help pages and examples. The raw results for the gene cluster can be downloaded all together in one YAML file or the result for each gene of the cluster in a separate file. In addition to the raw data, the SVG figures of the gene structures and FASTA files of the sequences (cDNA, genomic DNA, exons, introns, target translation) are available for download. WebScipio provides an upload option for downloaded YAML files to let the user analyse his results at a later date.



Quick view of all results

3 FBpp0085935 (3), 99	4 FBpp0085935 (4), 98	5 FBpp0085935 (5), 100
0 FBpp0085935	1 FBpp0085935 (1), 87	2 FBpp0085935 (2), 87

FBpp0085935 complete

Name	Match-ratio	Query length (aa)	Number of Contigs
FBpp0085935	101	188	1

Search details

Sequence

Legend

exon
sequence shift
intron
stop codon in query/target
mismatch
additional codon in target or unmatched query
intron?
(Intron does not show appropriate 5' and/or 3' splice sites.)
gap
(Coding sequence not identified.)
gap
(Coding sequence not found on contig.)
700 bps
1 gi|116010442|ref|NT_033778.3| (6157bp)

Statistics
Tandem gene duplications: 5
Exons: 16
Introns: 3

Gaps: 2

Alignment
Evaluation
Up and Downstream DNA
Genomic DNA
Coding DNA
Translation
Download Resultfiles

Figure 2.3-2 | WebScipio result view of the search for tandem gene duplications of the *Drosophila melanogaster* CG14502 gene (Flybase sequence accession FBpp0085935). Exons are illustrated as coloured rectangles, introns as grey narrow rectangles, and gaps as red narrow rectangles. Gaps indicate missing exons of the tandem gene duplicates. For the search the default parameters were used except for the *minimal score for exons* parameter that was set to 5% to find some exon duplicates of the first exon.

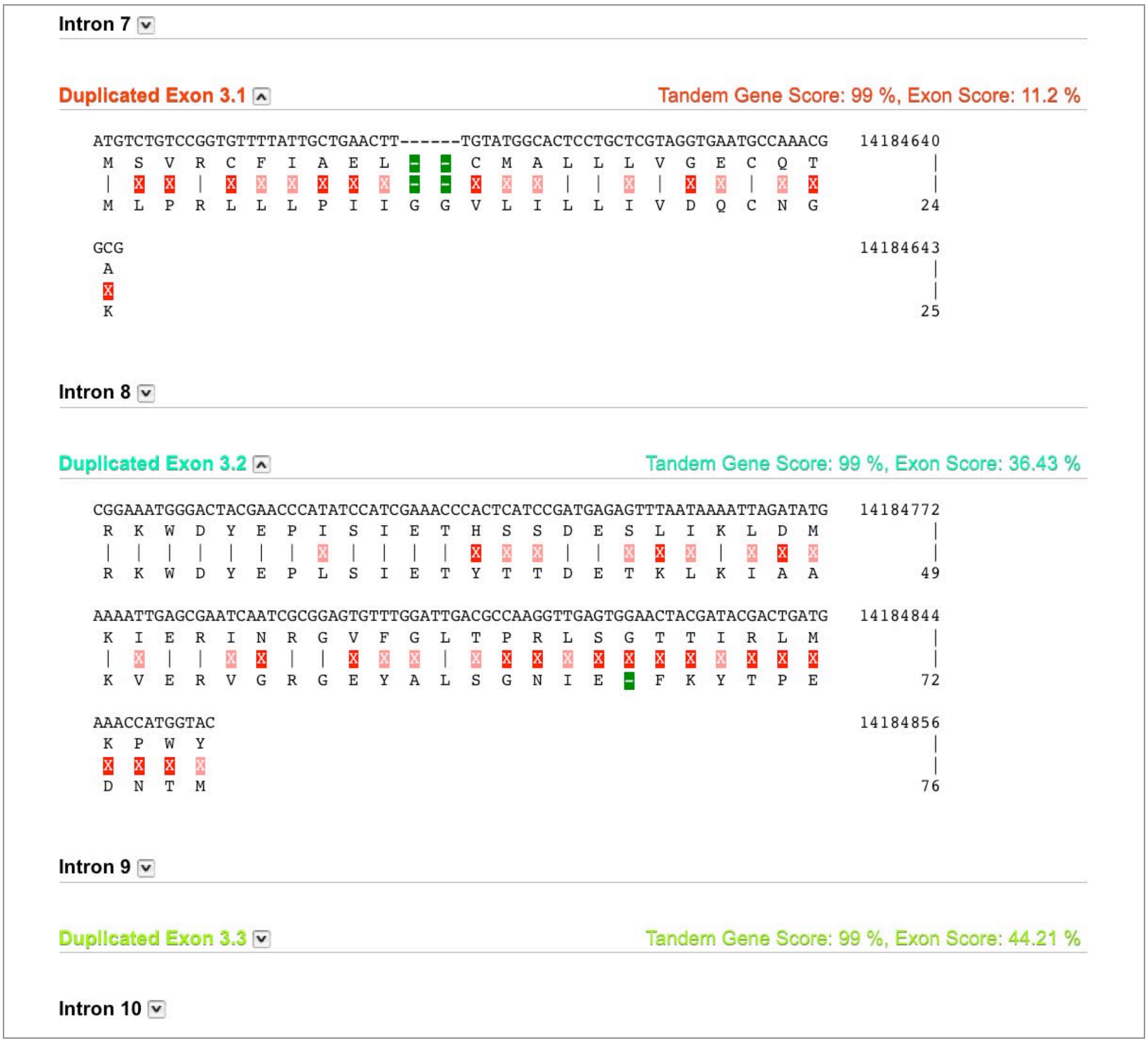
2.3.3 Results and discussion

WebScipio uses the command-line tool Scipio to reconstruct the gene structures of given protein sequences based on the available eukaryotic genome assemblies. Scipio has been developed for the case that protein sequences and target genome sequence are from the same organism. Nevertheless, Scipio allows several mismatches that might result from sequencing and assembly errors like missing or additional bases, which lead to frame-shifts, or in-frame stop codons that would lead to premature gene stops. Mismatches might also be the result from differences in the source of the protein sequence, which might have been obtained from cDNA libraries of a certain strain, and the specific sequenced strain of the species. To accomplish this task, Scipio relies on BLAT, which is one of the fastest tools available for the alignment of almost similar protein or DNA sequences. As Scipio tolerates a certain amount of mismatches between query and target sequence it can also successfully be used for cross-species gene reconstructions and predictions [109]. Because Scipio relies on BLAT the success of the cross-species search depends on the difference between the query and target gene. If genes are highly conserved in evolution Scipio is able to correctly reconstruct genes in species that diverged hundreds of million years ago. If genes evolve fast Scipio can predict genes only in very related organisms. This behaviour can also be used to predict gene duplicates in the same organism, and is implemented as *multiple results* parameter in the Scipio options. Again, because Scipio relies on BLAT, only those duplicates will be identified that are very similar. An advantage of this option is that Scipio is able to find dispersed as well as tandem duplicates.

In an analysis of the origin of new genes in the *Drosophila* species complex [149] it has been shown that the majority of the constrained functional new genes are dispersed duplicates. In contrast, tandem duplications were found to be young events and to lead to lower survival rates. Thus, tandem duplicates are often pseudogenes most probably because the introduction of frame shifts and in-frame stop codons does not demand too many mutations to destroy the transcription and expression of the new gene. If duplicates are kept in the genome they acquire new functions through neofunctionalization and subfunctionalization by accumulation of many substitutions [152]. Those genes are too divergent to be identified by the *multiple results* option of Scipio. However, although accumulating many substitutions tandem duplicates very often retain the gene structure of the original gene including intron splice sites and reading frames of exons. Occasionally further introns might be introduced or prior existing introns lost because these changes would not destroy transcription and translation. To use this knowledge in tandem gene duplicate identification we developed an algorithm that searches for duplicates of a query sequence based on the restrictions imposed by its gene structure. Every piece of DNA in the up- and downstream region of the original exon that has the same splice sites and shares sequence homology to the original exon, when translated in the same reading frame, is thought to be a candidate for an exon of a duplicated gene. In the case that introns have been lost or gained in the duplicated genes the splice site restrictions apply to the outer borders of the fused or split exons. WebScipio is able to correctly reconstruct the gene structure for a given protein sequence and is thus very suited as starting point for searches for candidate exons of duplicated genes.

To search for tandem gene duplicates an extension to WebScipio was implemented providing several parameters to adjust the search according to users or genome-specific needs. In most cases, however, the standard parameters will provide reasonable and interpretable results. As soon as the search is done, WebScipio shows an overview of the results as small gene structure pictures (Quick View), which reveal the exon regions of the found tandem genes (Figure 2.3-2). For convenient analysis the genomic region comprising the gene structure of the query sequence and the exons of the predicted tandem genes is shown in a combined graph and provided as one YAML file. The exons of the original gene are dark coloured and the corresponding predicted exons have the same but lighter colour. The darkness of the colour relates to the similarity of the predicted exon to the original one. The same colour scheme is used to highlight the various exons in the Alignment view of the genomic regions (Figure 2.3-3). The Alignment view shows the nucleotide sequence of the gene ordered in exons and introns. For every exon the genomic DNA and the corresponding translation are shown, as well as the alignment of the query sequence to the translation.

To demonstrate the application, quality, and limitations of the new algorithm we provide some example searches in the following sections. Tandemly arrayed gene duplicates have several characteristics that need to be considered. Gene duplications can be found on both the forward and the reverse strand. The duplicated genes might contain fused exons or might contain additional introns. In the case of retroposed genes, which are derived from the reverse transcription and insertion of processed genes, gene duplications do not have any introns. Although gene duplications are more often found for small genes consisting of one or only a few exons, gene duplicates can also be identified for genes consisting of dozens of exons spanning large genomic regions. Because tandem gene duplicates are defined by being located next to each other in the genome, intergenic regions are expected to be short. This is also the reason why the parameter for bordering the search in up- and downstream regions of the original gene limits this region to 300,000 nucleotides. However, WebScipio cannot exclude that there may be additional genes in-between gene duplicates. An example for such a scenario would be the duplication or multiple duplications of small genomic regions that encode several genes. In most cases we considered examples from the fruit fly *Drosophila melanogaster* and sequences from Flybase [140], because the corresponding genome is of high quality and the annotation of the genome is already at a very advanced stage. Fragmented genomes, like draft genomes for which only short contigs are available, or chromosome assemblies containing many gaps, are useful to screen for interesting candidates but do not provide the reliability needed for tests of the algorithms quality and limitations. An advanced annotation provides the advantage that genomic locations of most genes have already been identified. Thus the gene order is already established although there might still be errors in the annotation of single exons.



Intron 7

Duplicated Exon 3.1 Tandem Gene Score: 99 %, Exon Score: 11.2 %

```
ATGTCTGTCCGGTGTTTTATTGCTGAACTT------TGTATGGCACTCCTGCTCGTAGGTGAATGCCAAACG   14184640
 M  S  V  R  C  F  I  A  E  L  -  -  C  M  A  L  L  L  V  G  E  C  Q  T           |
 |  X  X  |  X  X  X  X  X  X  -  -  X  X  X  |  |  X  |  X  X  |  X  X           |
 M  L  P  R  L  L  L  P  I  I  G  G  V  L  I  L  L  I  V  D  Q  C  N  G          24

GCG                                                                        14184643
 A                                                                                |
 X                                                                                |
 K                                                                               25
```

Intron 8

Duplicated Exon 3.2 Tandem Gene Score: 99 %, Exon Score: 36.43 %

```
CGGAAATGGGACTACGAACCCATATCCATCGAAACCCACTCATCCGATGAGAGTTTAATAAAATTAGATATG   14184772
 R  K  W  D  Y  E  P  I  S  I  E  T  H  S  S  D  E  S  L  I  K  L  D  M           |
 |  |  |  |  |  |  |  X  |  |  |  |  X  X  X  |  |  X  X  X  |  X  X  X           |
 R  K  W  D  Y  E  P  L  S  I  E  T  Y  T  T  D  E  T  K  L  K  I  A  A          49

AAAATTGAGCGAATCAATCGCGGAGTGTTTGGATTGACGCCAAGGTTGAGTGGAACTACGATACGACTGATG   14184844
 K  I  E  R  I  N  R  G  V  F  G  L  T  P  R  L  S  G  T  T  I  R  L  M           |
 |  X  |  |  X  X  |  |  X  X  X  |  X  X  X  X  X  X  X  X  X  X  X  X           |
 K  V  E  R  V  G  R  G  E  Y  A  L  S  G  N  I  E  -  F  K  Y  T  P  E          72

AAACCATGGTAC                                                               14184856
 K  P  W  Y                                                                       |
 X  X  X  X                                                                       |
 D  N  T  M                                                                      76
```

Intron 9

Duplicated Exon 3.3 Tandem Gene Score: 99 %, Exon Score: 44.21 %

Intron 10

Figure 2.3-3 | Alignment view of the first two exons of the third gene duplicate of the *Drosophila melanogaster* CG14502 gene (see also Figure 2.3-2). Each exon is named by the tandem gene number and the exon number. In addition, the tandem gene score and the exon score is given for each exon in percentage. The alignment indicates the positions of the sequences in the genome and the protein. The first line of the alignment represents the nucleotide sequence of the gene and the second line the translation of this sequence. The third line shows how the amino acids of this translation match the amino acids coded by the original exon, which are shown in the forth line. Mismatches are represented by an X in dark red or, if amino acids are chemically similar, in light red. Gaps in the alignment are shown as hyphens in green. The Duplicated Exon 3.3 alignment has been closed for representation purposes.

Examples of tandemly arrayed gene duplicates

Gene duplicates on both the forward and the reverse strand

The WebScipio tandem gene duplication extension has been developed to find tandem gene duplications on the forward as well as on the reverse strand in relation to the query gene. The example in Figure 2.3-4 shows five gene duplicates of the *Drosophila melanogaster* heat shock protein 23 gene (Hsp23), which consists of one exon. The first duplicate (Hsp67Bc) and the forth duplicate (Hsp26) in the genomic region are on the reverse strand, the other duplications Hsp22, CG4461, and Hsp27 are in the same reading direction as Hsp23. This search was performed with default parameters except increasing the *allowed length difference for exons* parameter to 30 amino acids. The most divergent gene duplication Hsp67Ba (Table 2.3-1), which is encoded in the genomic region between Hsp26 and Hsp23, was not found. This example shows that although the sequence identity is very low between the duplicates and the Hsp23 search sequence (Table 2.3-1), five duplicates could be identified. The length difference between Hsp23 and Hsp67Ba was too large so that candidates of the length of Hsp67Ba were not included in the search with the given search parameters.

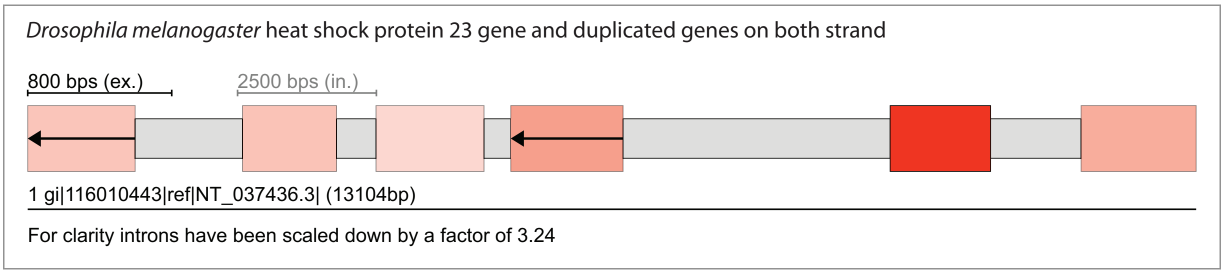

Drosophila melanogaster heat shock protein 23 gene and duplicated genes on both strand
800 bps (ex.)
2500 bps (in.)
1 gi|116010443|ref|NT_037436.3| (13104bp)
For clarity introns have been scaled down by a factor of 3.24

Figure 2.3-4 | *Drosophila melanogaster* heat shock protein gene duplicates. The figure shows the duplications found by the algorithm with Hsp23 as query. The genomic region contains, from the left to the right side in the drawing, the identified genes Hsp67Bc, Hsp22, CG4461, Hsp26, the query gene Hsp23, and another gene duplicate Hsp27. Gene duplications on the reverse strand are marked by an arrow in reverse direction.

Table 2.3-1 | Comparison of the length, similarity, and reading direction of the genes of the *Drosophila melanogaster* heat shock protein cluster

	Hsp67Bc	**Hsp22**	**CG4461**	**Hsp26**	**Hsp67Ba**	**Hsp23**	**Hsp27**
Length [aa]	199	174	200	208	445	186	213
Identity to Hsp23	0.29	0.31	0.26	0.49	0.15	1.00	0.41
Strand	rev	for	for	rev	rev	for	for

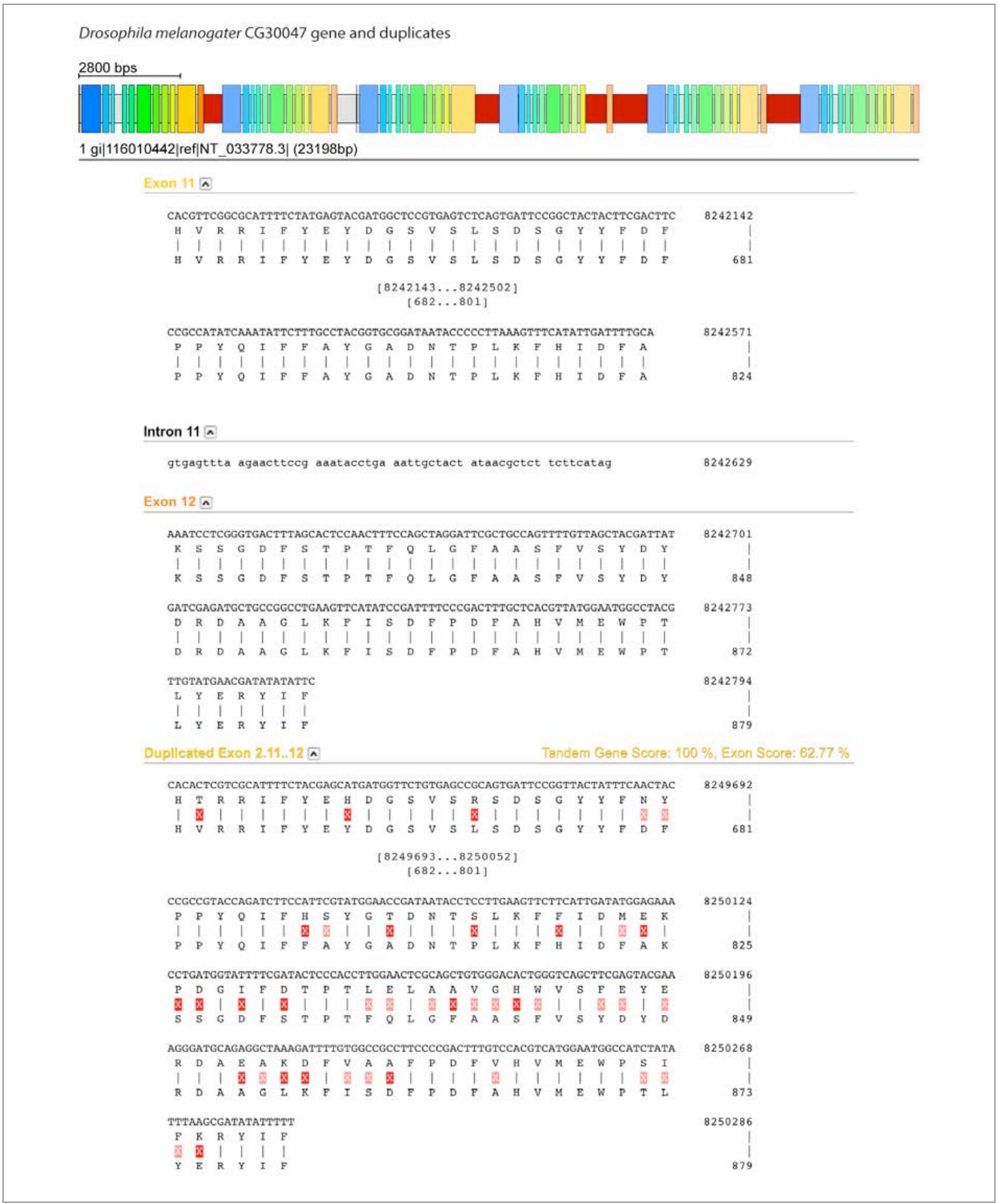
Duplicated exons in six tandemly arrayed genes including a lost intron and a pseudogene

The new algorithm is able to reconstruct tandemly arrayed gene duplications containing many exons and gene duplicates. The *Drosophila melanogaster* CG30047 gene includes 12 exons.

Five duplicates of this gene could be identified with the algorithm (Figure 2.3-5, top). In the second duplicated gene an intron loss could be identified. The exons 11 and 12 of CG30047 are translated as one exon in this duplicate (Figure 2.3-5, bottom). To find such lost introns the option to *search for concatenated exons* has been enabled. The third duplicate most probably represents a pseudogene, because exon 11 contains a frame shift and could thus not be found. Other reasons for the frame shift could be sequencing and assembly errors. However, the *Drosophila melanogaster* genome [141] is one of the best available and a lot of effort has been spent in the finishing process. Thus, it is more probable that the third duplicate is a pseudogene. Exon 1, which codes for seven amino acids, has low complexity and could therefore only be identified in the second gene duplication by setting the *minimal exon length* parameter to 7 aa.

Myosin heavy chain gene duplicates

Mammals encode two clusters of muscle myosin heavy chain genes, one cluster containing the α- and β-cardiac muscle myosin heavy chain genes [164, 165], and one cluster containing six skeletal muscle myosin heavy chain genes in the order embryonic, 2a, 2x, 2b, perinatal, and extraocular [165, 166]. These myosin genes consist of 38 exons each. Based on their gene size and number of exons the genes of the muscle myosin gene cluster should be on the upper limit of the complexity of a search for tandem gene duplicates. With the new WebScipio extension all genes of the muscle myosin cluster in *Homo sapiens* could be identified (Figure 2.3-6). For the search the *region size* parameter was set to 300,000 nucleotides and the *minimal score for exons* to 50%. This example also shows the advantage of the new WebScipio extension compared to the *multiple results* option in Scipio. When searching with the *multiple results* option of Scipio and the 2a gene as starting sequence, mixed genes are found for every additional gene candidate (Figure 2.3-6). Scipio does not know about gene borders and analyses all BLAT hits according to their score. Therefore, Scipio combines the highest scoring hits to gene candidate one (2a), the next highest scoring hits to gene candidate two (2x), and so on. The third gene candidate, for example, mainly consists of the exons of the perinatal muscle myosin heavy chain gene, but the N-terminus of the 2b gene has a higher homology to the 2a gene than the N-terminus of the perinatal gene and therefore the 2b N-terminus is combined with the C- terminus of perinatal.



Drosophila melanogater CG30047 gene and duplicates

Exon 11 ⌃

```
CACGTTCGGCGCATTTTCTATGAGTACGATGGCTCCGTGAGTCTCAGTGATTCCGGCTACTACTTCGACTTC   8242142
 H  V  R  R  I  F  Y  E  Y  D  G  S  V  S  L  S  D  S  G  Y  Y  F  D  F          |
 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |          |
 H  V  R  R  I  F  Y  E  Y  D  G  S  V  S  L  S  D  S  G  Y  Y  F  D  F        681

                         [8242143...8242502]
                            [682...801]

CCGCCATATCAAATATTCTTTGCCTACGGTGCGGATAATACCCCCTTAAAGTTTCATATTGATTTTGCA      8242571
 P  P  Y  Q  I  F  F  A  Y  G  A  D  N  T  P  L  K  F  H  I  D  F  A             |
 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |             |
 P  P  Y  Q  I  F  F  A  Y  G  A  D  N  T  P  L  K  F  H  I  D  F  A           824
```

Intron 11 ⌃

```
gtgagttta agaacttccg aaatacctga aattgctact ataacgctct tcttcatag            8242629
```

Exon 12 ⌃

```
AAATCCTCGGGTGACTTTAGCACTCCAACTTTCCAGCTAGGATTCGCTGCCAGTTTTGTTAGCTACGATTAT   8242701
 K  S  S  G  D  F  S  T  P  T  F  Q  L  G  F  A  A  S  F  V  S  Y  D  Y          |
 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |          |
 K  S  S  G  D  F  S  T  P  T  F  Q  L  G  F  A  A  S  F  V  S  Y  D  Y        848

GATCGAGATGCTGCCGGCCTGAAGTTCATATCCGATTTTCCCGACTTTGCTCACGTTATGGAATGGCCTACG   8242773
 D  R  D  A  A  G  L  K  F  I  S  D  F  P  D  F  A  H  V  M  E  W  P  T          |
 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |          |
 D  R  D  A  A  G  L  K  F  I  S  D  F  P  D  F  A  H  V  M  E  W  P  T        872

TTGTATGAACGATATATATTC                                                      8242794
 L  Y  E  R  Y  I  F                                                              |
 |  |  |  |  |  |  |                                                              |
 L  Y  E  R  Y  I  F                                                            879
```

Duplicated Exon 2.11..12 ⌃ Tandem Gene Score: 100 %, Exon Score: 62.77 %

```
CACACTCGTCGCATTTTCTACGAGCATGATGGTTCTGTGAGCCGCAGTGATTCCGGTTACTATTTCAACTAC   8249692
 H  T  R  R  I  F  Y  E  H  D  G  S  V  S  R  S  D  S  G  Y  Y  F  N  Y          |
 |  ☒  |  |  |  |  |  |  ☒  |  |  |  |  |  ☒  |  |  |  |  |  |  |  ☒  ☒          |
 H  V  R  R  I  F  Y  E  Y  D  G  S  V  S  L  S  D  S  G  Y  Y  F  D  F        681

                         [8249693...8250052]
                            [682...801]

CCGCCGTACCAGATCTTCCATTCGTATGGAACCGATAATACCTCCTTGAAGTTCTTCATTGATATGGAGAAA   8250124
 P  P  Y  Q  I  F  H  S  Y  G  T  D  N  T  S  L  K  F  F  I  D  M  E  K          |
 |  |  |  |  |  |  ☒  ☒  |  |  ☒  |  |  |  ☒  |  |  |  ☒  |  |  ☒  ☒  |          |
 P  P  Y  Q  I  F  F  A  Y  G  A  D  N  T  P  L  K  F  H  I  D  F  A  K        825

CCTGATGGTATTTTCGATACTCCCACCTTGGAACTCGCAGCTGTGGGACACTGGGTCAGCTTCGAGTACGAA   8250196
 P  D  G  I  F  D  T  P  T  L  E  L  A  A  V  G  H  W  V  S  F  E  Y  E          |
 ☒  ☒  |  ☒  |  ☒  |  |  |  ☒  ☒  |  ☒  ☒  ☒  ☒  ☒  ☒  |  |  ☒  ☒  |  ☒          |
 S  S  G  D  F  S  T  P  T  F  Q  L  G  F  A  A  S  F  V  S  Y  D  Y  D        849

AGGGATGCAGAGGCTAAAGATTTTGTGGCCGCCTTCCCCGACTTTGTCCACGTCATGGAATGGCCATCTATA   8250268
 R  D  A  E  A  K  D  F  V  A  A  F  P  D  F  V  H  V  M  E  W  P  S  I          |
 |  |  |  ☒  ☒  ☒  ☒  |  ☒  ☒  ☒  |  |  |  |  ☒  |  |  |  |  |  |  ☒  ☒          |
 R  D  A  A  G  L  K  F  I  S  D  F  P  D  F  A  H  V  M  E  W  P  T  L        873

TTTAAGCGATATATTTTT                                                         8250286
 F  K  R  Y  I  F                                                                 |
 ☒  ☒  |  |  |  |                                                                 |
 Y  E  R  Y  I  F                                                               879
```

Figure 2.3-5 | *Drosophila melanogaster* CG30047. Five gene duplications were found for the CG30047 gene. In the second duplication the intron between exon 11 and 12 was lost as shown in the alignment. The alignment of exon 11 (CG30047) and the alignment of the corresponding region in the duplicated gene were shortend by amino acids 682 to 801 for representation purposes.

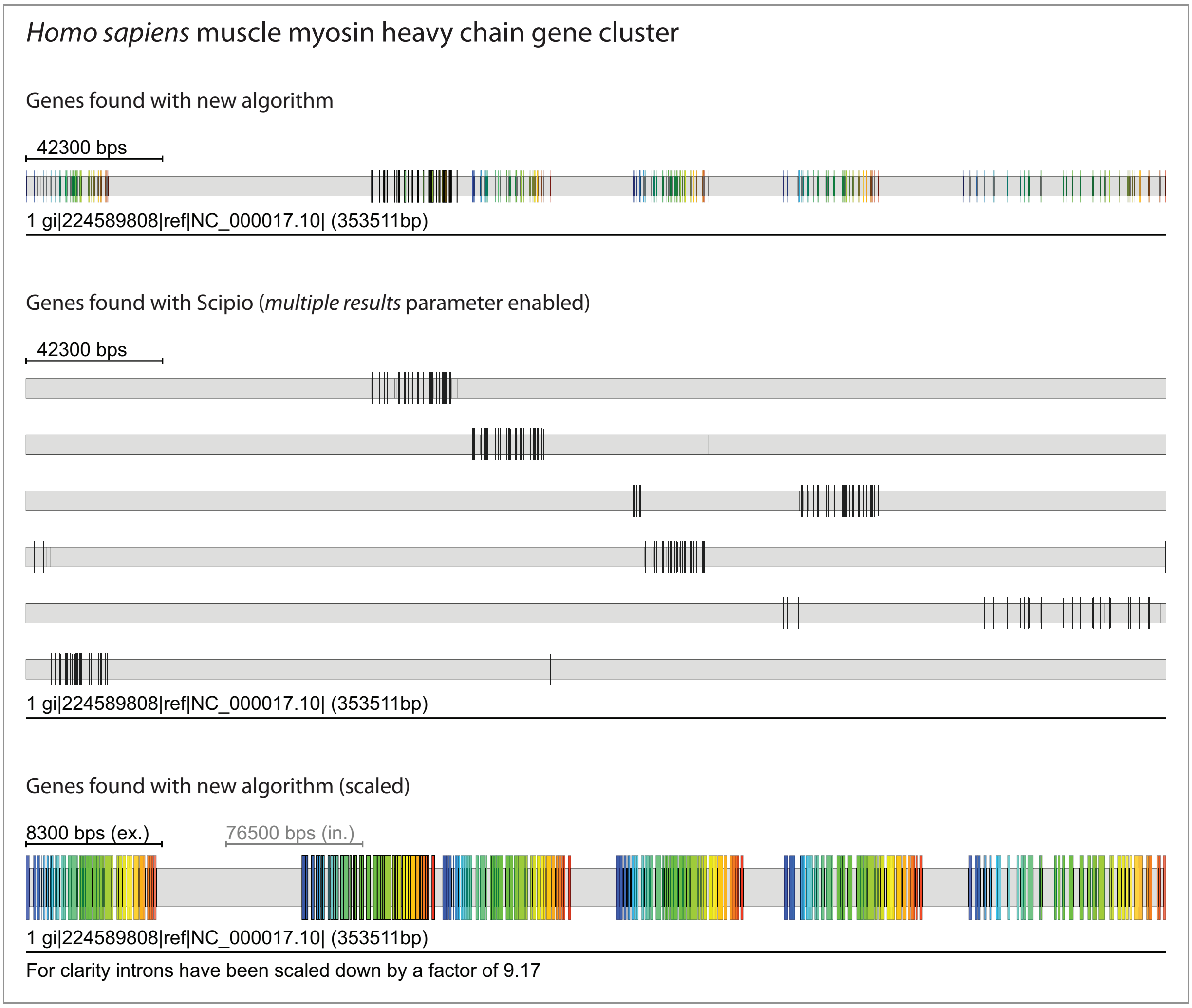

Homo sapiens muscle myosin heavy chain gene cluster
Genes found with new algorithm
42300 bps
1 gi|224589808|ref|NC_000017.10| (353511bp)
Genes found with Scipio (*multiple results* parameter enabled)
42300 bps
1 gi|224589808|ref|NC_000017.10| (353511bp)
Genes found with new algorithm (scaled)
8300 bps (ex.)
76500 bps (in.)
1 gi|224589808|ref|NC_000017.10| (353511bp)
For clarity introns have been scaled down by a factor of 9.17

Figure 2.3-6 | *Homo sapiens* muscle myosin heavy chain gene cluster. The skeletal muscle myosin heavy chain cluster consists of the genes embryonic, 2a, 2x, 2b, perinatal, and extraocular, from left (5' end) to the right (3' end). The WebScipio search for tandem gene duplicates based on the 2a gene identifies all other genes of the cluster. The Scipio search with the parameter *multiple results* also identifies six gene candidates but only the search sequence (the 2a gene) is found correctly while the other gene candidates consist of fusions of different parts of the other muscle myosin heavy chain genes.

The nile tilapia *Oreochromis niloticus* contains another type of a muscle myosin heavy chain gene cluster (Figure 2.3-7). Here, two genes (Mhc6 and Mhc13) are encoded on the forward strand, and Mhc7 is encoded on the reverse strand. Nevertheless, WebScipio correctly reconstructed the complete cluster when searching with the Mhc13 gene. When searching with Mhc6 or Mhc7, the small C-terminal exons of the respective other genes could not be identified. These examples demonstrate that WebScipio with the new extension is able to correctly identify arrays of very large and complex genes. For this search the minimal score for exons parameter was set to 30% and the region size parameter to 50,000 nucleotides.

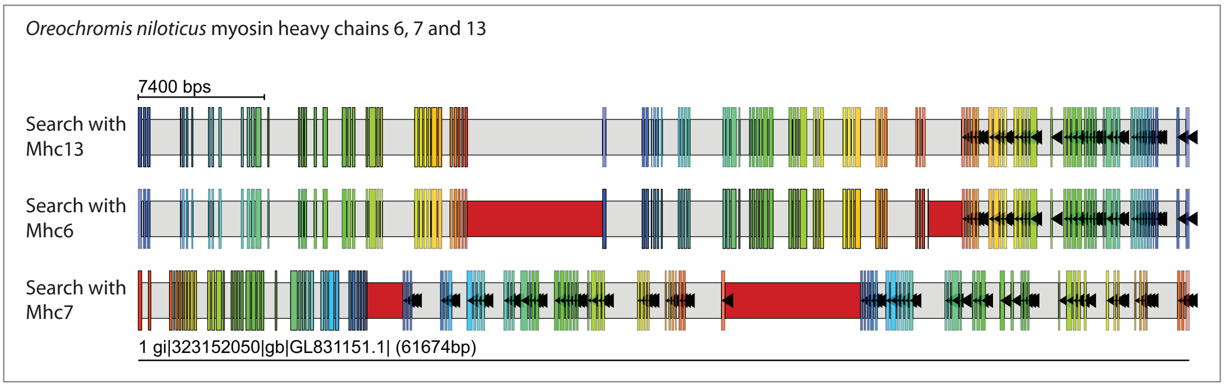

Oreochromis niloticus myosin heavy chains 6, 7 and 13
7400 bps
Search with Mhc13
Search with Mhc6
Search with Mhc7
1 gi|323152050|gb|GL831151.1| (61674bp)

Figure 2.3-7 | *Oreochromis niloticus* muscle myosin heavy chain gene cluster. The nile tilapia contains a cluster of three muscle myosin heavy chain genes (Mhc13, Mhc6, and Mhc7) of which Mhc7 is encoded in the opposite direction. The last exon is too divergent to be identified in most cases. Only when searching with the Mhc13 gene, the tandem genes Mhc6 and Mhc7 are reconstructed completely.

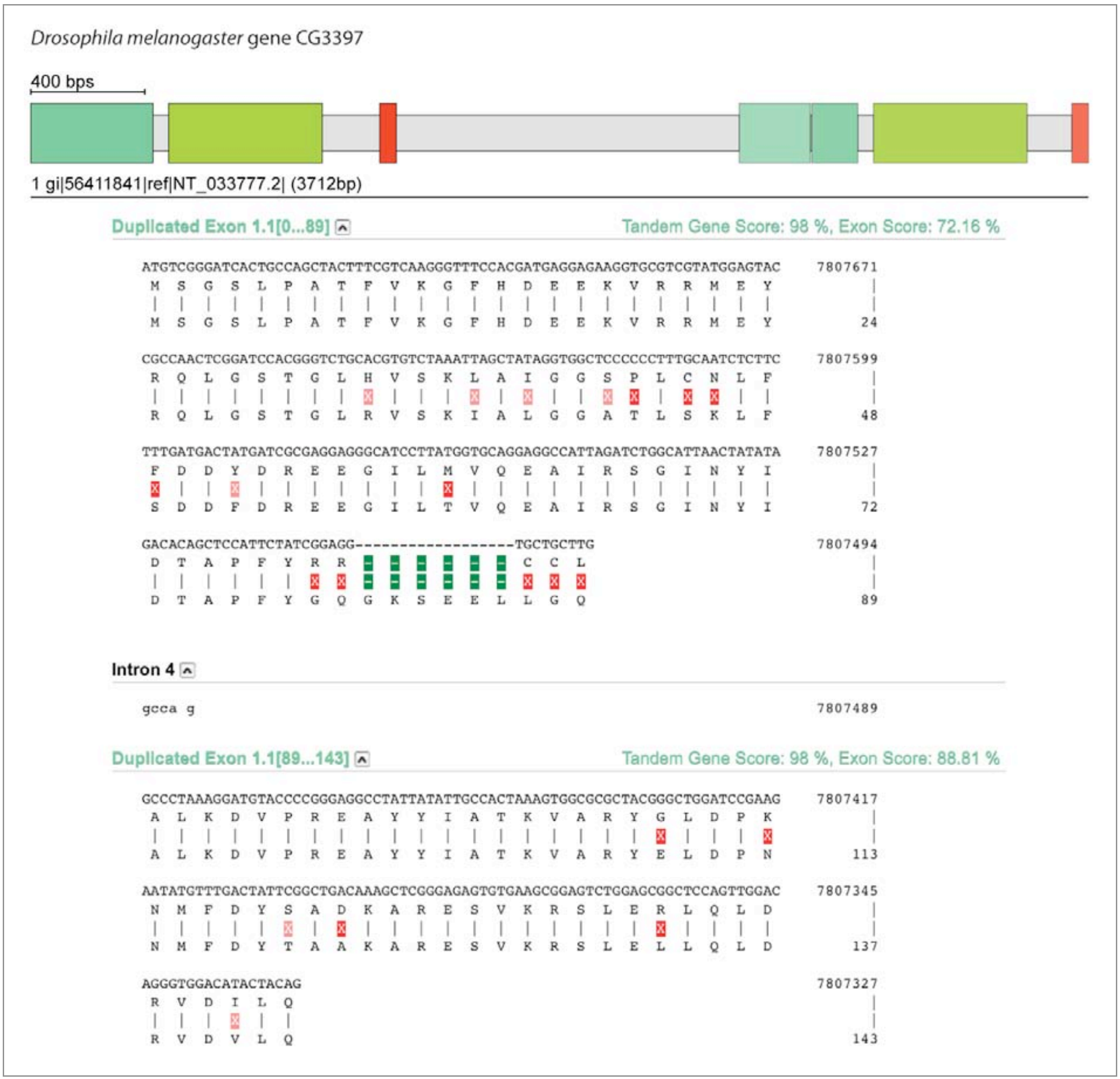
Revealing a pseudogene

For the *Drosophila melanogaster* CG3397 gene the first exon is splitted into two exons in the prediction of the gene duplication. For this search the default parameters were used and the option to *search for splitted exons* was enabled. The predicted gene is most probably a pseudogene, because either the predicted intron between the two splitted exons is too short to be spliced, or the exon translation results in a frame shift if both parts are considered as one potential exon. The details are shown in the alignment (Figure 2.3-8).

Examples of non-tandemly arrayed gene duplicates

Duplicated gene regions

Tandemly arrayed genes evolve through unequal recombination. In this process not only single genes might be duplicated but small genomic regions containing several genes. The result would be a tandemly arrayed group of genes. Because WebScipio is searching for each gene separately it cannot separate a group of duplicated genes from a tandem array of single genes. An example for duplicated genomic regions is the region in *Drosophila melanogaster* containing genes coding for histones (Figure 2.3-9). The new algorithm identified many duplicates for each of the His1, His2A, His2B, His3, and His4 genes in the *Drosophila* genome. As query the genes CG33825 (His1), CG33826 (His2A), CG33894 (His2B), CG33827 (His3), and CG33893 (His4) were used. The His2B and His4 genes are on the reverse strand in comparison to the other genes. The genes are very similar (some code for the same protein sequence) resulting in alignment scores between 99% and 100%. Only two more divergent gene duplicates were found for the His2A gene. The first two gene duplicates of His2A have alignment scores of 79%.



Drosophila melanogaster gene CG3397

1 gi|56411841|ref|NT_033777.2| (3712bp)

Duplicated Exon 1.1[0...89] Tandem Gene Score: 98 %, Exon Score: 72.16 %

```
ATGTCGGGATCACTGCCAGCTACTTTCGTCAAGGGTTTCCACGATGAGGAGAAGGTGCGTCGTATGGAGTAC    7807671
 M   S   G   S   L   P   A   T   F   V   K   G   F   H   D   E   E   K   V   R   R   M   E   Y          |
 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |          |
 M   S   G   S   L   P   A   T   F   V   K   G   F   H   D   E   E   K   V   R   R   M   E   Y         24

CGCCAACTCGGATCCACGGGTCTGCACGTGTCTAAATTAGCTATAGGTGGCTCCCCCTTTGCAATCTCTTC    7807599
 R   Q   L   G   S   T   G   L   H   V   S   K   L   A   I   G   G   S   P   L   C   N   L   F          |
 |   |   |   |   |   |   |   |   X   |   |   |   X   |   X   |   |   X   X   |   X   X   |   |          |
 R   Q   L   G   S   T   G   L   R   V   S   K   I   A   L   G   G   A   T   L   S   K   L   F         48

TTTGATGACTATGATCGCGAGGAGGGCATCCTTATGGTGCAGGAGGCCATTAGATCTGGCATTAACTATATA    7807527
 F   D   D   Y   D   R   E   E   G   I   L   M   V   Q   E   A   I   R   S   G   I   N   Y   I          |
 X   |   |   X   |   |   |   |   |   |   |   X   |   |   |   |   |   |   |   |   |   |   |   |          |
 S   D   D   F   D   R   E   E   G   I   L   T   V   Q   E   A   I   R   S   G   I   N   Y   I         72

GACACAGCTCCATTCTATCGGAGG------------------TGCTGCTTG                          7807494
 D   T   A   P   F   Y   R   R   -   -   -   -   -   -   C   C   L                                      |
 |   |   |   |   |   |   X   X   -   -   -   -   -   -   X   X   X                                      |
 D   T   A   P   F   Y   G   Q   G   K   S   E   E   L   L   G   Q                                     89
```

Intron 4

```
gcca g                                                                         7807489
```

Duplicated Exon 1.1[89...143] Tandem Gene Score: 98 %, Exon Score: 88.81 %

```
GCCCTAAAGGATGTACCCCGGGAGGCCTATTATATTGCCACTAAAGTGGCGCGCTACGGGCTGGATCCGAAG    7807417
 A   L   K   D   V   P   R   E   A   Y   Y   I   A   T   K   V   A   R   Y   G   L   D   P   K          |
 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   X   |   |   |   X          |
 A   L   K   D   V   P   R   E   A   Y   Y   I   A   T   K   V   A   R   Y   E   L   D   P   N        113

AATATGTTTGACTATTCGGCTGACAAAGCTCGGGAGAGTGTGAAGCGGAGTCTGGAGCGGCTCCAGTTGGAC    7807345
 N   M   F   D   Y   S   A   D   K   A   R   E   S   V   K   R   S   L   E   R   L   Q   L   D          |
 |   |   |   |   |   X   |   X   |   |   |   |   |   |   |   |   |   |   |   X   |   |   |   |          |
 N   M   F   D   Y   T   A   A   K   A   R   E   S   V   K   R   S   L   E   L   L   Q   L   D        137

AGGGTGGACATACTACAG                                                           7807327
 R   V   D   I   L   Q                                                                                  |
 |   |   |   X   |   |                                                                                  |
 R   V   D   V   L   Q                                                                                143
```

Figure 2.3-8 | *Drosophila melanogaster* CG3397 gene. A gene duplication could be identified downstream of the CG3397 gene, which, however, most probably is a pseudogene.

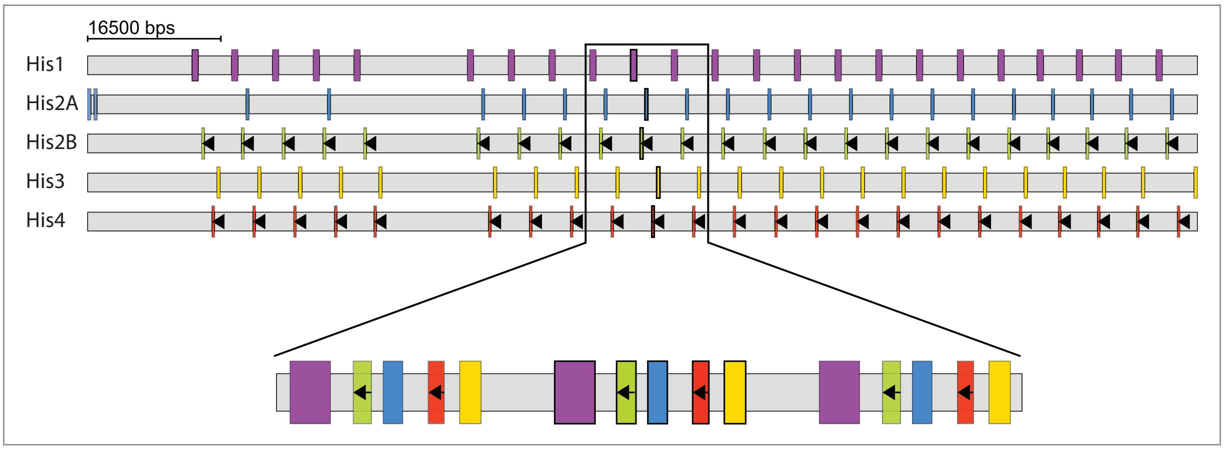


Figure 2.3-9 | *Drosophila melanogaster* histones. The results for the separate searches for gene duplicates of the histones His1, His2A, His2B, His3, and His4 are shown. Based on the results of the search for each single gene it is not possible to distinguish between a gene and a genomic region duplication. The results of all searches at the same scale shows that not single genes but a genomic region containing all five histone genes has been duplicated several times.

***Trans*-spliced genes**

Tandem gene duplicates and *trans*-spliced genes could evolve through the same gene duplication process during evolution, except that only part of the gene is duplicated instead of the complete gene. The exon-intron structure of tandem gene duplicates and *trans*-spliced genes look very similar, which complicates their differentiation during the process of gene identification. If, for example, the constitutive part of the *trans*-spliced gene consists of only one exon while the *trans*-spliced part consists of groups of similar alternative exons the correct reconstruction of the *trans*-spliced gene would not look different compared to a partial reconstruction of a cluster of duplicated genes for which the first (or last) exons were not found because of low similarity. The gene CG1637 of Drosphila is a *trans*-spliced gene [53]. The WebScipio algorithm predicts tandemly arrayed genes for isoform A and B of CG1637, although the first exons of the potential tandem gene candidates were not found (Figure 2.3-10). The close inspection of the three isoforms shows that the predicted exons do not belong to duplicated genes, but to *trans*-spliced variants of the same gene. Another type of problem is demonstrated by the dynein intermediate chain gene of *Drosophila melanogaster*. Here, the dynein intermediate chain gene is annotated as four separate genes (Sdic1, Sdic2, Sdic3 and Sdic4) in Flybase (version of June 24th, 2011). The problem is, however, that the real first two exons of the gene are not annotated in Flybase.

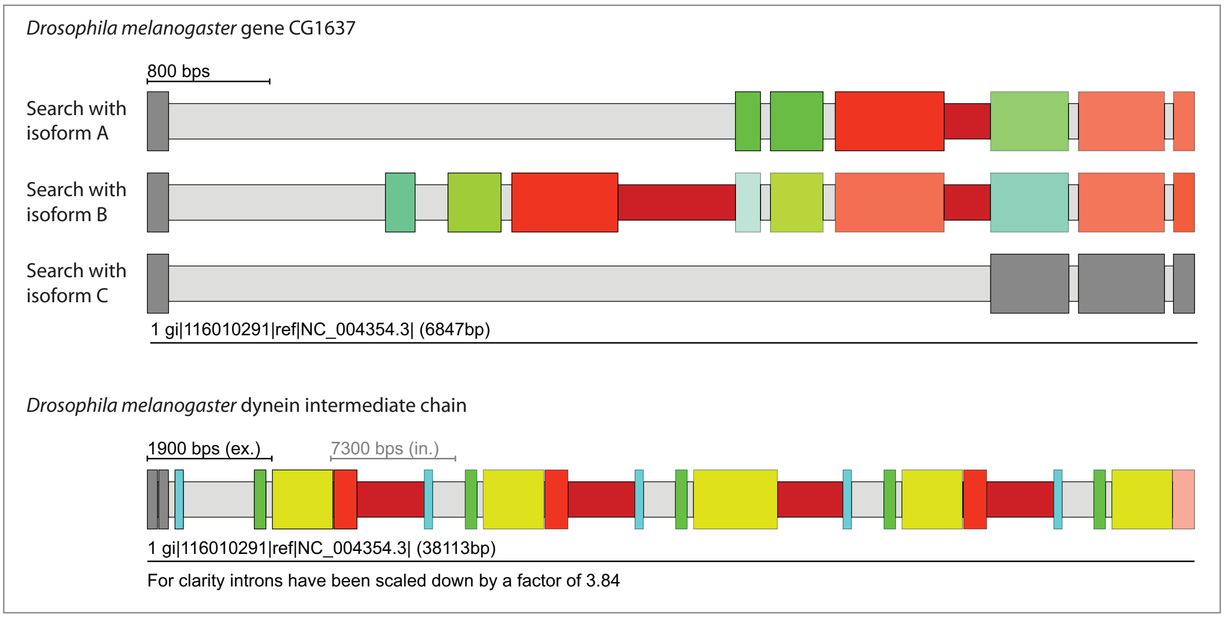

Drosophila melanogaster gene CG1637
800 bps
Search with isoform A
Search with isoform B
Search with isoform C
1 gi|116010291|ref|NC_004354.3| (6847bp)
Drosophila melanogaster dynein intermediate chain
1900 bps (ex.)
7300 bps (in.)
1 gi|116010291|ref|NC_004354.3| (38113bp)
For clarity introns have been scaled down by a factor of 3.84

Figure 2.3-10 | *Drosophila melanogaster* CG1737 gene and *Drosophila melanogaster* dynein intermediate chain. The algorithm identified duplicated exons in the *trans*-spliced CG1737 and dynein intermediate chain genes. The search was done with default parameters and the *search for concatenated exons* and *search for splitted exons* options were enabled. To reveal the last and most divergent exon the *region size* parameter was set to 35,000 nucleotides and the *allowed length difference* parameter to 30 amino acids for the dynein intermediate chain gene.

The sequence encoded by the true first exons is conserved throughout all major branches of the eukaryotic tree of life that express a cytoplasmic dynein, in chromalveolates, Excavata, and Opisthokonta. In addition, this N-terminal part of the dynein intermediate chain is of high functional importance because it connects dynein to dynactin by interacting with the dynactin p150 gene. Based on these facts and the found exon order of the genomic region, we expect the gene to be *trans*-spliced (Figure 2.3-10, bottom).

2.3.4 Conclusion

Our algorithm provides a method to consistently predict and reconstruct tandemly arrayed gene duplicates. It has been integrated into the web interface of WebScipio allowing the search for gene duplicates of a given query protein sequence in the respective genome assemblies. WebScipio provides access to more than 2300 genome assembly files from more than 650 eukaryotes (July 2011) and is updated as soon as further genome assemblies become available whether from newer versions of already sequenced species or from newly sequenced genomes. The search results are presented in drawings coloured according to the sequence similarity of the gene duplicate to the search sequence, and in several human-readable formats like detailed alignments of the found exons to the genomic DNA. Sequences and figures can be downloaded, as well as the complete raw data for later upload or further computational analysis. The new algorithm is based on the precondition that gene duplicates rather retain the gene structure of the original gene than the sequence. We could show that the new extension to WebScipio is able to correctly predict and reconstruct gene duplicates on both the forward and the reverse strand. Also, the new algorithm is able to correctly reconstruct complicated gene structures spread over hundreds of thousands of nucleotides like the skeletal muscle myosin heavy chain gene cluster in mammals. Gene duplications often accumulate gene function destroying mutations that lead to frame shifts and in-frame stop codons. Those potential pseudogenes are identified by WebScipio but the user has to carefully inspect the results to distinguish between sequencing errors and real pseudogenes. WebScipio cannot distinguish between gene duplicates and duplications of small genomic regions that might encode several genes. Here, WebScipio can identify and reconstruct the duplicates of one gene but does not provide any hints about other genes in the intergenic regions. *Trans*-spliced genes often contain clusters of alternative exons. Those clusters will be identified by WebScipio, but again the user needs to evaluate the results to distinguish between cases of *trans*-spliced genes, where the constitutive part is encoded by just a few exons, or real gene duplications, for which some terminal exons could not be identified because of very low sequence similarity or even assembly gaps. Altogether, WebScipio provides an easy to use way to analyse the genomic region of every gene of interest for the very common event of tandem gene duplication.

2.3.5 Acknowledgements

MK has been funded by grant KO 2251/6-1 of the Deutsche Forschungsgemeinschaft. We thank Björn Hammesfahr for fruitful discussions.

2.3.6 Authors' contributions

Both authors set the requirements for the system, performed extensive testing, wrote the manuscript, and approved the final version of the manuscript. KH implemented the algorithm and performed the computer calculations of the examples.

2.4 A phylogenetic analysis of the Brassicales clade based on an alignment-free sequence comparison method

Klas Hatje and Martin Kollmar

Abteilung NMR-basierte Strukturbiologie, Max-Planck-Institut für biophysikalische Chemie, Am Fassberg 11, D-37077 Göttingen, Germany

Frontiers in Plant Science 2012, 3(192)

Published: 29 August 2012
doi: 10.3389/fpls.2012.00192
This article is available from
http://www.frontiersin.org/Plant_Genetics_and_Genomics/10.3389/fpls.2012.00192/full

2.4.1 Abstract

Phylogenetic analyses reveal the evolutionary derivation of species. A phylogenetic tree can be inferred from multiple sequence alignments of proteins or genes. The alignment of whole genome sequences of higher eukaryotes is a computational intensive and ambitious task as is the computation of phylogenetic trees based on these alignments. To overcome these limitations, we here used an alignment-free method to compare genomes of the Brassicales clade.

For each nucleotide sequence a Chaos Game Representation (CGR) can be computed, which represents each nucleotide of the sequence as a point in a square defined by the four nucleotides as vertices. Each CGR is therefore a unique fingerprint of the underlying sequence. If the CGRs are divided by grid lines each grid square denotes the occurrence of oligonucleotides of a specific length in the sequence (Frequency Chaos Game Representation, FCGR).

Here, we used distance measures between FCGRs to infer phylogenetic trees of Brassicales species. Three types of data were analysed because of their different characteristics: A) Whole genome assemblies as far as available for species belonging to the Malvidae taxon. B) EST data of species of the Brassicales clade. C) Mitochondrial genomes of the Rosids branch, a supergroup of the Malvidae. The trees reconstructed based on the Euclidean distance method are in general agreement with single gene trees. The Fitch-Margoliash and Neighbour joining algorithms resulted in similar to identical trees. Here, for the first time we have applied the bootstrap re-sampling concept to trees based on FCGRs to determine the support of the

branchings. FCGRs have the advantage that they are fast to calculate, and can be used as additional information to alignment based data and morphological characteristics to improve the phylogenetic classification of species in ambiguous cases.

2.4.2 Introduction

Phylogenetic analyses reveal the evolutionary derivation of species. A phylogenetic tree can be inferred from multiple sequence alignments of proteins or genes, which assume the conservation and contiguity over the total sample length between homologous sequences [167]. The alignment of whole genome sequences of eukaryotes is a computational intensive and ambitious task as is the computation of phylogenetic trees based on these alignments [168]. In particular, genetic recombination and shuffling during species evolution complicate whole genome alignments limiting species genome versus single gene, multiple gene or transcriptome comparisons. However, it would be beneficial for the significance of the species trees, if also whole genome assembly data were taken into account. In the past two decades several methods have been suggested for alignment-free sequence analyses that mainly group into word (oligomer) frequency methods and methods that do not resolve the fixed word-length distance measures and are thus absolutely independent from the assumption of conservation and contiguity (reviewed in [94]). The latter category includes the Chaos Theory [169] and the theoretical concept of Kolomogorov complexity [170]. More recent methods include the alignment-free estimation of the number of substitutions per site [171] and feature frequency profiles [172].

The Chaos Game Representation (CGR) denotes an algorithm, which produces fractal pictures and can be adapted to reveal patterns in DNA [170] and even protein sequences [173, 174]. These CGR pictures exhibit the fractal property that the overall pattern of the CGR picture is repeated in smaller parts of the picture. It has been shown that this self-similarity even holds for whole genome sequences and its sub-sequences, like single chromosomes, contigs or genes [175–177]. Commonly, the pictures of DNA sequences are generated as squares such that the lower (A + T) and the upper (C + G) halves indicate the base composition and the diagonals the purine/pyrimidine composition. CGRs are unique descriptions of each DNA sequence and, in the case of whole genome sequences, can therefore be regarded as genomic fingerprints. However, the CGRs are not directly comparable. If the CGR pictures are divided into smaller squares by grid lines, each grid square represents the frequencies of the respective oligonucleotides as found in the whole sequence [175, 176]. These frequencies can be represented in Frequency Chaos Game Representation (FCGR) pictures with a grey scale to express the number of points within each grid square and with pictures for each length k oligonucleotide (with k = 1,2,3…). FCGRs are numerical matrices and can be used to infer phylogenetic trees based on distance methods [92]. So far this approach has only been applied to reconstruct the phylogeny of 20 birds using nuclear genome data [91], to analyse the mitochondrial genomes of 26 sample eukaryotes [92], and to sub-typing of HIV-I [93]. One of the advantages of using FCGRs for phylogenetic reconstructions is that sequence, which cannot be aligned, can be used.

Here, we performed phylogenetic analyses based on three different types of data. Firstly we used the whole genomic sequence assemblies of all so far sequenced species in the taxon Malvidae, including that of *Brassica rapa*. Because a reference tree including all these species was not available we assembled and annotated all actin-capping (CP) protein sequences [178] and the sequences of the actin-related proteins Arp2 and Arp3 [179]. These proteins are present in all eukaryotes and as single copies in the *Arabidopsis thaliana* genome. Thus they were not expected to exist in duplicates in the other analyzed species avoiding the ortholog-paralog problem. To infer the phylogeny of the different *Brassica* species, for which whole genome assemblies have not yet been produced, we used EST and mitochondrial genome DNA. The quality of the phylogenetic analyses depends on the resolution of the FCGRs (length of *k*) and thus on the length of the nucleotide sequences. Thus we only included those species for which a considerable number of EST clones were available. To estimate the support for the branchings, here, we apply the concept of bootstrap re-sampling to the comparison of FCGRs for the first time.

2.4.3 Material and methods

Data acquisition

The genome files were retrieved from diArk[59] [31], and the mitochondrial genomes and EST reads from the NCBI database, each in FASTA format (Table 2.4-1). For the generation of the CGRs the contigs and reads of each dataset were concatenated. The whole genome assemblies as available from the sequencing centers contain both the nuclear and mitochondrial genomes, and potentially still some contaminations from other species' DNA. However, given the sizes of the whole genome datasets, the contributions of the mitochondrial genomes and contaminating DNA to the FCGRs are negligible. The FCGRs of the whole genome data can thus be regarded as identical to the FCGRs of the nuclear genomes.

[59] http://www.diark.org

Table 2.4-1 | List of the species used in the analysis

	Whole genome			EST		Mitochondrial genome		
Species	**Contigs**	**Nucleotides**	**Accession Numbers**	**Reads**	**Nucleotides**	**Contigs**	**Nucleotides**	**Accession Numbers**
Arabidopsis lyrata	695	206667935	GL348713-GL349407					
Arabidopsis thaliana	5	119145879	NC_003070-NC_003071, NC_003074-NC_003076	1529700	400512451	1	366924	NC_001284
Brassica rapa	51658	273071614	AENI01000001-AENI01051658	213605	122970377	1	219747	NC_016125
Capsella rubella	853	134834574						
Carica papaya	3207	331271729	DS981520-DS984726	77393	54789864			
Citrus clementina	1128	295550349						
Citrus sinensis	12574	319231331						
Eucalyptus camaldulensis	274001	654922307	DF097775-DF126446					
Eucalyptus grandis	4952	691297852						
Eutrema halophilum	639	243117811		38022	20080214			
Eutrema parvulum	7	114396853	CM001187-CM001193					
Gossypium raimondii	1448	763818933						
Theobroma cacao	1782	351351221	FR720657-FR725448					
Vitis vinifera	33	486265422	FN597015-FN597047	446643	284204927	1	773279	NC_012119
Brassica napus				643437	381399492	1	221853	NC_008285
Brassica oleracea				179150	125257248	1	360271	NC_016118
Limnanthes alba				15331	8582959			
Raphanus raphanistrum				164119	104536170			
Raphanus sativus				150680	97973638			
Tropaeolum majus				10507	6436290			
Brassica carinata						1	232241	NC_016120
Brassica juncea						1	219766	NC_016123
Lotus japonicus						1	380861	NC_016743
Millettia pinnata						1	425718	NC_016742
Ricinus communis						1	502773	NC_015141

The number of contigs/reads and the number of nucleotides for the whole genome, EST and mitochondrial genome data files are given. In addition, for whole genome and mitochondrial genome data the NCBI accession numbers are given if available.

Implementation of the algorithm

The algorithm to calculate CGRs and FCGRs was implemented in C/C++. CGR positions were generated as lists in plain text and plotted for graphical presentations in the Scalable Vector Graphics (SVG) format[60]. Based on the CGR position values, FCGRs were calculated for each *k* in 1…8. Distance calculations were implemented in Ruby[61].

[60] http://www.w3.org/Graphics/SVG

[61] http://ruby-lang.org

Generating Chaos Game Representations

CGRs of the nucleotide sequences were generated by the following algorithm. A 1 × 1 square is drawn and each vertex labelled by a nucleotide. In agreement with other analyses we placed C in the upper left, G in the upper right, A in the lower left, and T in the lower right vertex. The starting point is defined as the geometric center of the square at position (0.5,0.5). The respective nucleotide sequences are then plotted sequentially. For the first nucleotide a point is plotted on half the distance between the starting point (0.5,0.5) and the vertex corresponding to this nucleotide. Subsequently for each following nucleotide a point is placed as midpoint between the previously plotted point and the vertex corresponding to the nucleotide (Figure 2.4-1A).

The algorithm can be expressed by the following equations:

$$(1) \qquad CGR_0 = (0.5{,}0.5)$$

$$(2) \qquad CGR_i = \begin{cases} CGR_{i-1} + 0.5 \cdot \left(CGR_{i-1} + (0.0{,}0.0)\right) \text{ if } seq_i = \text{'C'} \\ CGR_{i-1} + 0.5 \cdot \left(CGR_{i-1} + (1.0{,}0.0)\right) \text{ if } seq_i = \text{'G'} \\ CGR_{i-1} + 0.5 \cdot \left(CGR_{i-1} + (0.0{,}1.0)\right) \text{ if } seq_i = \text{'A'} \\ CGR_{i-1} + 0.5 \cdot \left(CGR_{i-1} + (1.0{,}1.0)\right) \text{ if } seq_i = \text{'T'} \end{cases}$$

The resulting plot is unique for each sequence. The overall pattern of points is repeated in each sub-square of the plot (Figure 2.4-1B). In addition, each plot based on a sub-sequence of the whole sequence has a similar appearance. Thus similar sequences result in similar CGR plots. Figure 2.4-1B shows the CGR of the first 1,000,000 nucleotides of the *Brassica rapa* genome sequence.

The calculation of the frequencies of points within each sub-square results in an FCGR. Thus each FCGR represents the occurrence of oligonucleotides in the whole sequence. For dinucleotides (*k* = 2) the binary square is divided into a 4 × 4 grid, for trinucleotides (*k* = 3) into an 8 × 8 grid, and in general into a $2^k \times 2^k$ grid. Figure 2.4-1C shows an FCGR (*k* = 3) of the whole *Brassica rapa* genome sequence.

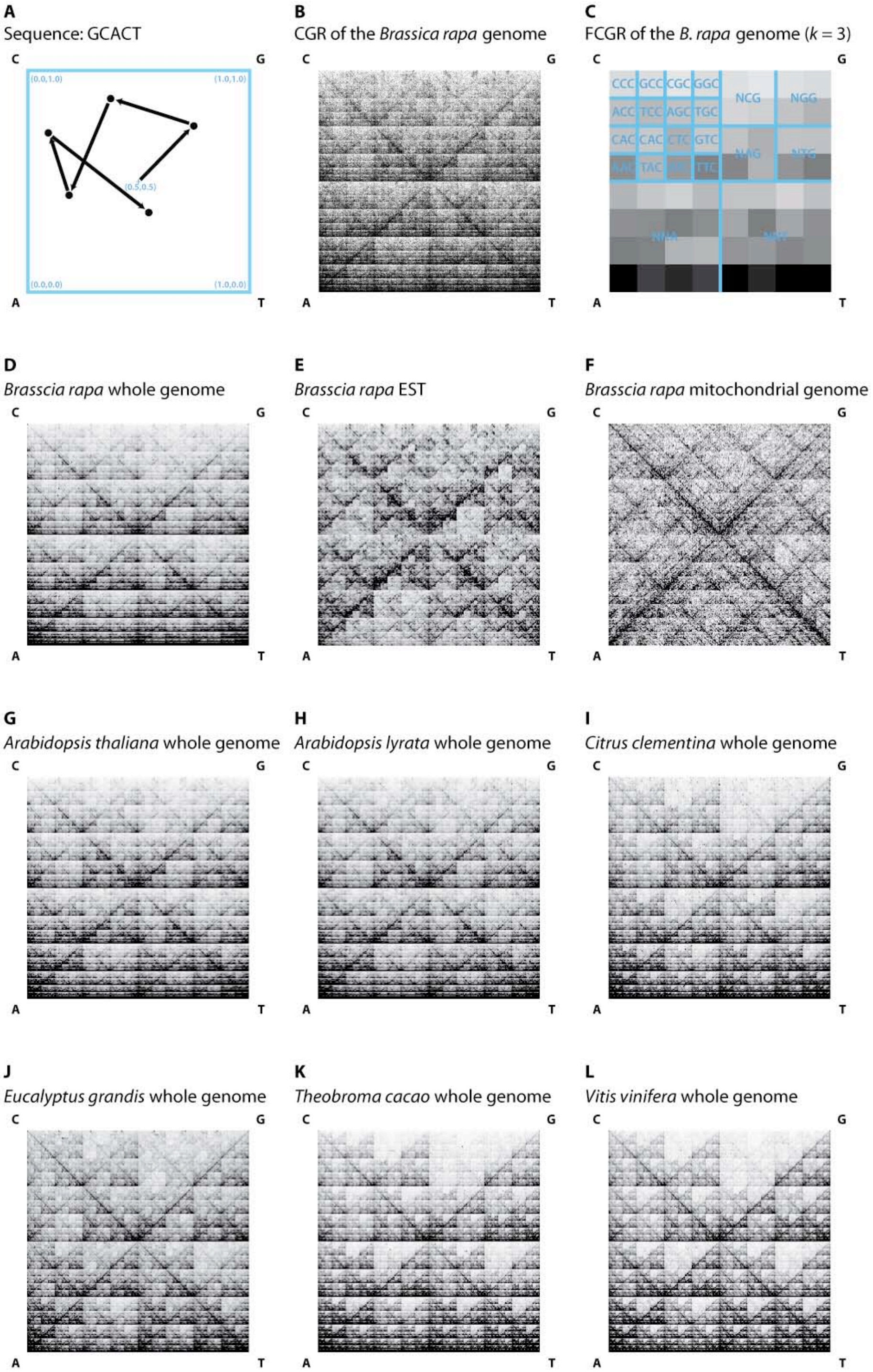

A
Sequence: GCACT
C
G
(0.0,1.0)
(1.0,1.0)
(0.5,0.5)
(0.0,0.0)
(1.0,0.0)
A
T
B
CGR of the Brassica rapa genome
C
G
A
T
C
FCGR of the B. rapa genome (k = 3)
C
G
CCC
GCC
CGC
GGC
NCG
NGG
ACC
TCC
AGC
TGC
CAC
CAC
CTC
GTC
NAG
NTG
TAC
TTC
A
T
D
Brasscia rapa whole genome
C
G
A
T
E
Brasscia rapa EST
C
G
A
T
F
Brasscia rapa mitochondrial genome
C
G
A
T
G
Arabidopsis thaliana whole genome
C
G
A
T
H
Arabidopsis lyrata whole genome
C
G
A
T
I
Citrus clementina whole genome
C
G
A
T
J
Eucalyptus grandis whole genome
C
G
A
T
K
Theobroma cacao whole genome
C
G
A
T
L
Vitis vinifera whole genome
C
G
A
T

Figure 2.4-1 | Generating Chaos Game Representations. A) A chaos game representation (CGR) image is generated by drawing a unit square and, starting at the center (0.5,0.5), plotting for each nucleotide of the sequence a point on half the distance to the corresponding vertex. In this example the CGR for the sequence ‘GCACT’ was drawn. B) The image shows the CGR of the first 1,000,000 nucleotides of the *Brassica rapa* genome. C) The figure shows an FCGR (k = 3) of the whole *Brassica rapa* genome illustrating the frequencies of points in the CGR in an 8 × 8 grid. The squares of the grid represent the occurrence of specific trinucleotides, which are labelled in the figure. In D) to L) the FCGRs (k = 8) of the whole genome (D), EST (E) and mitochondrial genome sequences (F) of *Brassica rapa* and the FCGRs (k = 8) of the whole genome sequences of some representatives of the different clades (G – L) are shown for visual comparison.

If the nucleotide sequences differ in length, the resulting FCGRs will also differ in there overall frequencies. To overcome this sequence-length bias each FCGR was standardized [92]. If the FCGR is represented as for example a $2^k \times 2^k$ matrix, the matrix $A = (a)_{2^k \times 2^k}$ is transformed to a standardized FCGR as follows:

$$(3) \qquad \overline{A} = \frac{4^k}{\sum_{i=1}^{k} \sum_{j=1}^{k} a_{i,j}} A$$

The nucleotide sequences of each data file (whole genome, EST or mitochondrial genome data) were concatenated and the reverse complement of the concatenated sequence was appended. Characters other than ‘C’, ‘G’, ‘A’ or ‘T’ were ignored. Some example FCGRs generated with $k = 8$ are shown in Figure 2.4-1D-L. Already by visual inspection it is obvious, that whole genome, EST, and mitochondrial genome FCGRs have distinct patterns (Figure 2.4-1D-F), while the FCGRs generated from the same data type of closely related species are very similar (Figure 2.4-1G-L). EST data disproportionately contain poly-A sequences, resulting in unusually high frequency values in the FCGRs. These subsequently dominate the distance matrix calculation for higher order FCGRs ($k > 5$) and misdirect the calculation of the phylogenetic trees (data not shown). Therefore, in the case of EST data, the two entries in each FCGR that contain poly-A and poly-T stretches were set to zero.

Distances

In order to reveal the phylogenetic relation between the analyzed species we calculated pairwise distances between the FCGRs. In general all distances that are applicable to matrices could be used. The following distances have already been described for comparing FCGRs: The Hamming distance [92, 180], the Euclidean distance [91–94], the Image distance defined in Wang et al., 2005 and the Pearson distance [92, 94, 176]. Here, we chose the Pearson distance as a statistical distance and the Euclidean distance as a geometrical distance, which performed best in a comparison of difference distance methods [92]. The Euclidean distance between two points in 2-dimensional space is defined as the length of the line segment between these two points and can be calculated using the Pythagorean equation. This concept can be adapted to calculate the distance between two FCGRs. The Euclidean distance between two standardized FCGRs $A = (a)_{2^k \times 2^k}$ and $B = (b)_{2^k \times 2^k}$ is defined as follows:

$$(4) \qquad d_{Euclidean}(\overline{A}, \overline{B}) = \frac{\sqrt{2^k}}{4^k} \sqrt{\sum_{i=1}^{2^k} \sum_{j=1}^{2^k} (a_{i,j} - b_{i,j})^2}$$

The Pearson distance is based on a weighted Pearson correlation coefficient [92, 176]. To calculate the Pearson distance, the FCGRs are represented as lists of the frequencies with

$n = 4^k$ values. The Pearson distance between the non-standardized FCGRs $A = (x_1,\ldots,x_n)$ and $B = (y_1,\ldots,y_n)$ is defined as follows:

$$nw = \sum_{i=1}^{n} x_i \cdot y_i$$

$$(5) \qquad \bar{x}w = \frac{\sum_{i=1}^{n} x_i^2 \cdot y_i}{nw}, \quad \bar{y}w = \frac{\sum_{i=1}^{n} y_i^2 \cdot x_i}{nw}, \quad sx = \frac{\sum_{i=1}^{n} (x_i - \bar{x}w)^2 \cdot x_i \cdot y_i}{nw}, \quad sy = \frac{\sum_{i=1}^{n} (y_i - \bar{y}w)^2 \cdot x_i \cdot y_i}{nw}$$

$$d_{Pearson} = 1 - \frac{\sum_{i=1}^{n} \frac{x_i - \bar{x}w}{\sqrt{sx}} \cdot \frac{y_i - \bar{y}w}{\sqrt{sy}} \cdot x_i \cdot y_i}{nw}$$

Generating phylogenetic trees

To generate the phylogenetic trees, pair-wise distance matrices were calculated for each *k* in 1...8 with the Euclidean distance method as defined in (4) and the Pearson distance as defined in (5). The distance matrices were subjected to the Neighbour joining (NJ) and Fitch-Margoliash algorithms as implemented in the Phylip package[62]. Statistical support for branchings was obtained by applying the bootstrap re-sampling method. For each FCGR, 500 datasets were generated by random sampling with replacement. Based on these re-sampled FCGRs 500 phylogenetic trees were reconstructed for each *k* in 1...8. The trees of each dataset were summarized to consensus trees using the *consense* program of the Phylip package. The topologies of the consensus trees were fixed and the branch lengths calculated with the Fitch-Margoliash algorithm. In the case of the NJ trees, a bootstraped tree was chosen that had the same topology as the consensus tree and the bootstrap values were plotted onto this tree. The bootstrap values represent the percentage each interior branch has the same partition as the consensus tree.

Generation of the reference tree for the whole genome analysis

For the reference tree of those species for which whole genome assemblies are available we identified, assembled and annotated the sequences of the heterodimeric actin capping protein (CAP), α-CAP and β-CAP, and the sequences of the actin-related proteins Arp2 and Arp3. The *Brassica rapa* and *Gossypium raimondii* genomes contain duplicates of these genes due to species-specific duplications. Therefore, only one of the duplicates had been used for the phylogenetic tree reconstructions. The CAP and Arp sequences were aligned, concatenated, and phylogenetic trees reconstructed using the NJ and the Maximum likelihood (ML) method. The NJ tree was unrooted and generated using ClustalW [181] with standard settings and the

[62] http://evolution.genetics.washington.edu/phylip.html

Bootstrap (1,000 replicates) method. The ML tree was calculated using the JTT [182] substitution model as suggested by ProtTest [183] with estimated proportion of invariable sites and bootstrapping (1,000 replicates) using RAxML [184].

2.4.4 Results

Phylogenetic trees based on whole genome, mitochondrial genome, and EST data were generated using the Euclidean or Pearson distance methods in combination with the NJ or the Fitch-Margoliash tree reconstruction algorithms. In order to reveal the influence of the lengths of the oligonucleotides we report trees of FCGRs generated with $k = 3$ (trinucleotides, 64 data points) and $k = 8$ (octanucleotides, 65,536 data points).

Influence of sequence lengths on the phylogenetic trees

First we tested whether different sequence lengths have an influence on the results (Figure 2.4-2). For the whole genome assemblies and the EST datasets, sub-sections of the sequences were generated with lengths of 10^6, 10^7, and 10^8 nucleotides. For that purpose the contigs or EST entries of each organism were shuffled, concatenated, and subsequently the subsequences generated by cutting the sequences at the respective positions. In the case of the whole genome data (Figure 2.4-2A), the FCGRs of the whole genome assemblies and the subsequences of each organism grouped together forming clusters. The only exceptions were the shortest 10*6*-nucleotide sequences of *Citrus sinensis*, *Citrus clementina*, *Arabidopsis lyrata*, and *Arabidopsis thaliana*, which group to different species. The FCGRs of the EST data group together for each species independently of the lengths of the sequences (Figure 2.4-2B). For the mitochondrial genomes datasets with shorter sequences of 10^4 and 10^5 nucleotides were generated. Here the FCGRs of the 10^4-nucleotide sequences do not cluster together with those of the longer sequences of the corresponding species. The FCGRs of the mitochondrial sequences have been calculated based on hexanucleotides ($k = 6$; 4,096 data points). Here, $k = 6$ was chosen, because in the case of higher k values ($k = 7$ or $k = 8$), the sequence length of the shortest sequences (10^4 nucleotides) would be less than the number of data points in the FCGRs. In the shortest sequences (10^4 nucleotides) many of the hexanucleotides are not covered at all resulting in many zero values for frequency positions, which lead to the unusual grouping of these FCGRs.

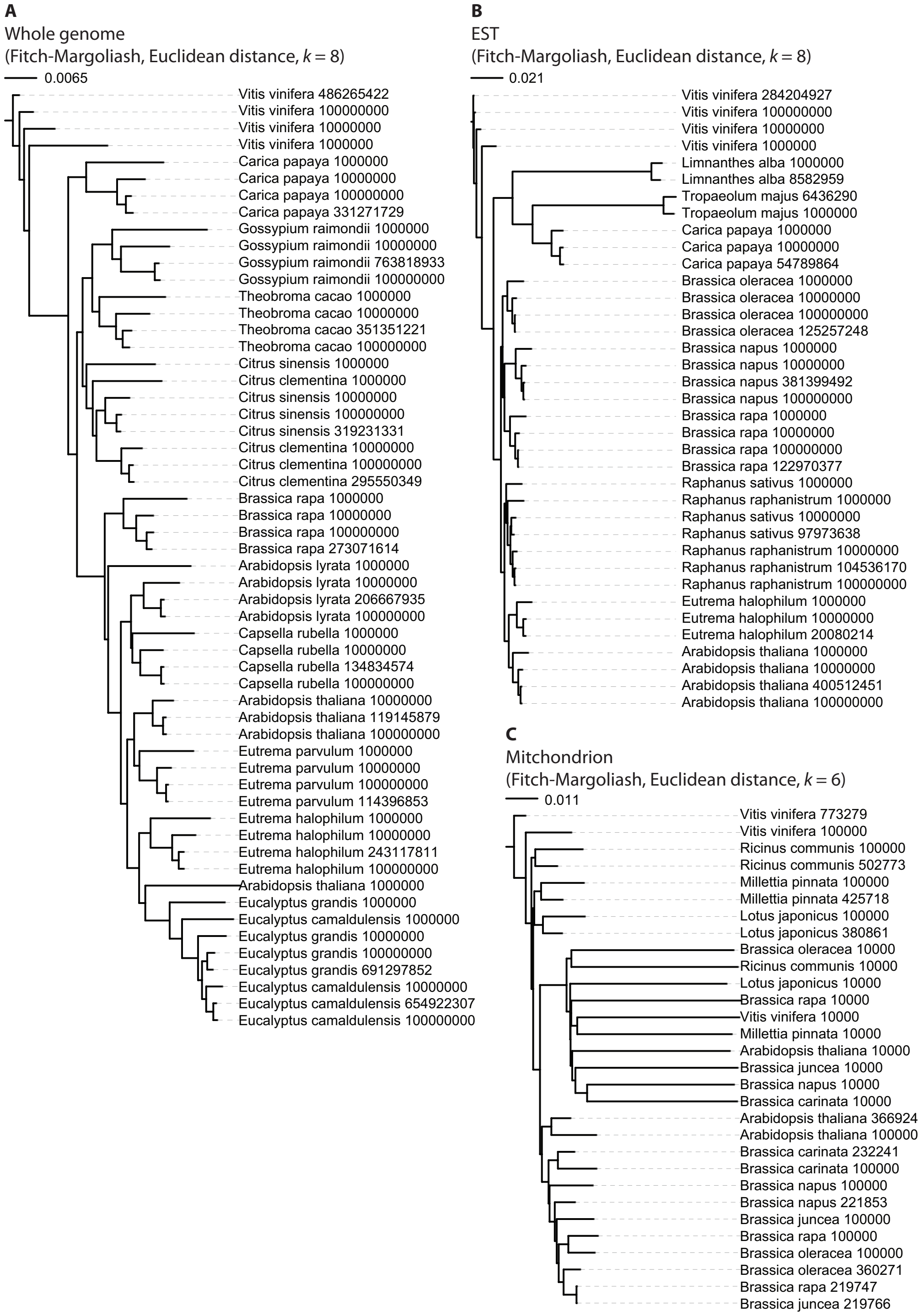

A
Whole genome
(Fitch-Margoliash, Euclidean distance, k = 8)
0.0065
Vitis vinifera 486265422
Vitis vinifera 100000000
Vitis vinifera 10000000
Vitis vinifera 1000000
Carica papaya 1000000
Carica papaya 10000000
Carica papaya 100000000
Carica papaya 331271729
Gossypium raimondii 1000000
Gossypium raimondii 10000000
Gossypium raimondii 763818933
Gossypium raimondii 100000000
Theobroma cacao 1000000
Theobroma cacao 10000000
Theobroma cacao 351351221
Theobroma cacao 100000000
Citrus sinensis 1000000
Citrus clementina 1000000
Citrus sinensis 10000000
Citrus sinensis 100000000
Citrus sinensis 319231331
Citrus clementina 10000000
Citrus clementina 100000000
Citrus clementina 295550349
Brassica rapa 1000000
Brassica rapa 10000000
Brassica rapa 100000000
Brassica rapa 273071614
Arabidopsis lyrata 1000000
Arabidopsis lyrata 10000000
Arabidopsis lyrata 206667935
Arabidopsis lyrata 100000000
Capsella rubella 1000000
Capsella rubella 10000000
Capsella rubella 134834574
Capsella rubella 100000000
Arabidopsis thaliana 10000000
Arabidopsis thaliana 119145879
Arabidopsis thaliana 100000000
Eutrema parvulum 1000000
Eutrema parvulum 10000000
Eutrema parvulum 100000000
Eutrema parvulum 114396853
Eutrema halophilum 1000000
Eutrema halophilum 10000000
Eutrema halophilum 243117811
Eutrema halophilum 100000000
Arabidopsis thaliana 1000000
Eucalyptus grandis 1000000
Eucalyptus camaldulensis 1000000
Eucalyptus grandis 10000000
Eucalyptus grandis 100000000
Eucalyptus grandis 691297852
Eucalyptus camaldulensis 10000000
Eucalyptus camaldulensis 654922307
Eucalyptus camaldulensis 100000000
B
EST
(Fitch-Margoliash, Euclidean distance, k = 8)
0.021
Vitis vinifera 284204927
Vitis vinifera 100000000
Vitis vinifera 10000000
Vitis vinifera 1000000
Limnanthes alba 1000000
Limnanthes alba 8582959
Tropaeolum majus 6436290
Tropaeolum majus 1000000
Carica papaya 1000000
Carica papaya 10000000
Carica papaya 54789864
Brassica oleracea 1000000
Brassica oleracea 10000000
Brassica oleracea 100000000
Brassica oleracea 125257248
Brassica napus 1000000
Brassica napus 10000000
Brassica napus 381399492
Brassica napus 100000000
Brassica rapa 1000000
Brassica rapa 10000000
Brassica rapa 100000000
Brassica rapa 122970377
Raphanus sativus 1000000
Raphanus raphanistrum 1000000
Raphanus sativus 10000000
Raphanus sativus 97973638
Raphanus raphanistrum 10000000
Raphanus raphanistrum 104536170
Raphanus raphanistrum 100000000
Eutrema halophilum 1000000
Eutrema halophilum 10000000
Eutrema halophilum 20080214
Arabidopsis thaliana 1000000
Arabidopsis thaliana 10000000
Arabidopsis thaliana 400512451
Arabidopsis thaliana 100000000
C
Mitchondrion
(Fitch-Margoliash, Euclidean distance, k = 6)
0.011
Vitis vinifera 773279
Vitis vinifera 100000
Ricinus communis 100000
Ricinus communis 502773
Millettia pinnata 100000
Millettia pinnata 425718
Lotus japonicus 100000
Lotus japonicus 380861
Brassica oleracea 10000
Ricinus communis 10000
Lotus japonicus 10000
Brassica rapa 10000
Vitis vinifera 10000
Millettia pinnata 10000
Arabidopsis thaliana 10000
Brassica juncea 10000
Brassica napus 10000
Brassica carinata 10000
Arabidopsis thaliana 366924
Arabidopsis thaliana 100000
Brassica carinata 232241
Brassica carinata 100000
Brassica napus 100000
Brassica napus 221853
Brassica juncea 100000
Brassica rapa 100000
Brassica oleracea 100000
Brassica oleracea 360271
Brassica rapa 219747
Brassica juncea 219766

Figure 2.4-2 | Phylogenetic trees to reveal the potential influence of sequence length. For each dataset subsequences with defined lengths were generated and FCGRs calculated. The lengths of the sequences were supposed to be sufficient for reliable tree reconstructions if the datasets generated from the same species grouped together. For whole genome and EST data 10,000,000 nucleotides should be sufficient while the full-length mitochondrial genomes are needed for tree reconstructions.

Whole genome analysis

In order to analyse the phylogenetic grouping of *Brassica rapa* in a whole-genome context we searched for closely related plant species, for which whole-genome assemblies are available. According to diArk [31], that comprises the most reliable and complete compilation of eukaryotic genome projects for which genome assemblies are available, the genomes of thirteen different species (excluding different *Arabidopsis thaliana* strains) of the taxon Malvidae have been sequenced and assembled: *Arabidopsis lyrata* [185], *Arabidopsis thaliana* (thale cress; [186]), *Brassica rapa* subsp. *pekinensis* (Chinese cabbage; [187]), *Capsella rubella*, *Carica papaya* [188], *Citrus clementina*, *Citrus sinensis* (sweet orange), *Eucalyptus camaldulensis* (Murray red gum), *Eucalyptus grandis* (Flooded gum), *Eutrema halophilum* (salt cress), *Eutrema parvulum* [189], *Gossypium raimondii*, and *Theobroma cacao* (cacao plant; [190]). In addition, the genome of *Vitis vinifera* (grape vine; [191, 192]) was chosen as outgroup to root the trees. A species tree including all these organisms is not available. For comparison we therefore reconstructed trees of these species based on the alignment of the concatenated protein sequences of the actin capping protein (CAP), Arp2, and Arp3 proteins (Figure 2.4-3A and B). The trees based on the NJ and ML methods are almost identical and differ only in the grouping of the two *Citrus* species (Sapindales clade) as independent clade (NJ, Figure 2.4-3A) or as sister clade of the Malvales (ML, Figure 2.4-3B). While the bootstrap support for all branchings is high, the support for the grouping of the *Citrus* clade is low in both trees (68.6% in the NJ and 66% in the ML tree, respectively). Both trees are in general agreement with phylogenetic analyses of the mitochondrial matR proteins [193] and 61 chloroplast protein-coding genes [194], and the combined analysis of 10 plastid and 2 nuclear (18S and 26S rDNA) genes [195] that also show different groupings of the Sapindales clade. All trees agree with the grouping of the Malvales, Sapindales, and Brassicales into one clade and the grouping of the Myrtales as a sister clade, *Carica papaya* being the most divergent of the analysed Brassicales species and *Capsella rubella* being the closest relative of the *Arabidopsis* species. Except for the grouping of the two *Citrus* species the topology of the tree based on the ubiquitous cytoskeletal proteins CAP and Arp2/3 can thus be regarded as reference.

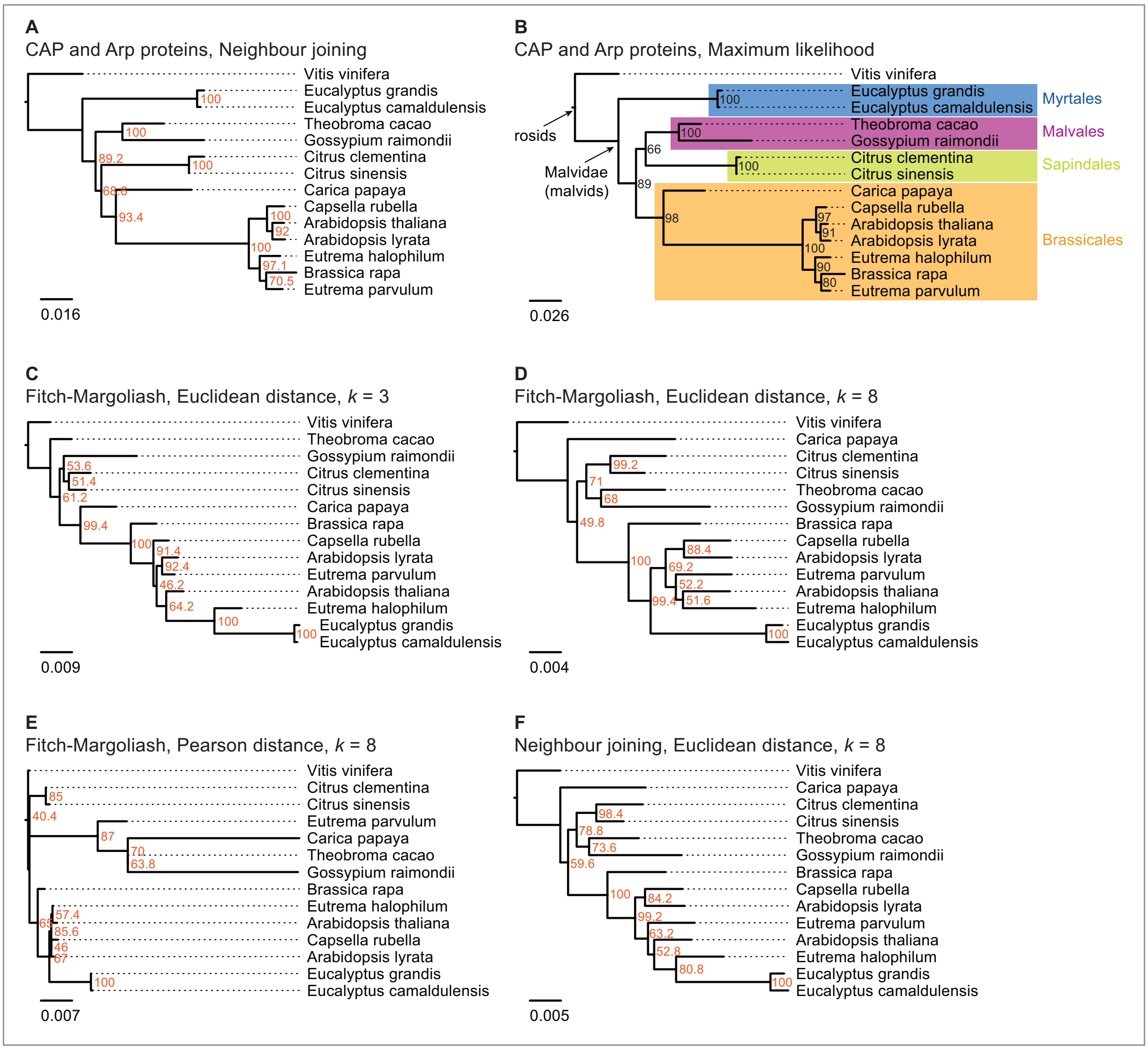


Figure 2.4-3 | Whole genome analysis. The trees in A) and B) are based on a multiple sequence alignment of manually assembled CAP and Arp2/3 protein sequences. The trees were calculated using the Neighbour joining and the Maximum likelihood method, respectively, with 1,000 bootstraps for each tree. In C) to F) phylogenetic trees were generated applying different methods on FCGRs of whole genome sequence data of species of the taxon Malvidae. In C), D) and E) the Fitch-Margoliash algorithm was used to calculate trees for 500 re-sampled datasets. Subsequently, a consensus tree was built and branch lengths were calculated based on the fixed consensus tree. The method used for the distance calculation and the resolution of the FCGRs are given on top of the trees. In F) the Neighbour joining algorithm was used to calculate the tree.

The resulting phylogenetic trees of the FCGRs differ as a function of data and methods used (Figure 2.4-3C-F). We reconstructed two trees based on the Euclidean distance and the Fitch-Margoliash algorithm but based on FCGRs with different resolution (k = 3 and k = 8 in Figure 2.4-3C and D, respectively), a tree using a different method for the distance calculation, the Pearson distance (Figure 2.4-3E), and a tree by applying a different method for the tree reconstruction, the NJ method (Figure 2.4-3F). In general, the trees agree with the reference tree except for the *Eucalyptus* species, which are either placed as sister group to *Eutrema halophilum* (Figure 2.4-3C and F) or at the base of the Brassicales (Figure 2.4-3D and E) and thus far from their position according to the reference tree. In addition, *Theobroma cacao* in Figure

2.4-2C, *Carica papaya* in Figure 2.4-3D to F, and *Eutrema parvulum* in Figure 2.4-2E are in wrong positions. None of the combinations of methods and data resulted in a correct resolution of the very closely related *Arabidopsis*, *Eutrema*, and *Capsella* species.

The tree based on the Pearson distance method (Figure 2.4-3E) contains the most deviations from the reference tree and this method therefore seems to be the least appropriate for reconstructing phylogenetic trees of whole genome sequences. This observation is in accordance with Wang et al., 2005. In addition, the bootstrap values do not provide reasonable support for most of the branchings except for the monophyly of the Citrus and the *Eucalyptus* clades. The trees based on high-resolution FCGRs (k = 8) using the Euclidean distance method (Figure 2.4-3D and 3) have identical topologies except for the *Eucalyptus* outliers. In both trees *Carica papaya* is placed as closest species to *Vitis vinifera* and not at the base of the Brassicales, *Arabidopsis thaliana* grouped to the *Eutrema* species instead to its closest relative *Arabidopsis lyrata*, and *Brassica rapa* is found at the base of the Brassicales instead of grouping to the *Eutrema* species. However, the misplacement of *Carica* and *Arabidopsis thaliana* is not well supported (bootstrap values of 50-60%). Thus, the considerably faster NJ algorithm is a good alternative to the Fitch-Margoliash algorithm if run time is important. In contrast, the phylogenetic tree based on the low-resolution FCGRs ($k = 3$) contains more differences compared to the reference tree (Figure 2.4-3C).

EST data analysis

For this analysis related species of *Brassica rapa* were chosen, for which more than 1,000 EST entries are available in the EST database of NCBI. There are ten species that belong to the Brassicales taxon and match this criteria: *Arabidopsis thaliana*, *Brassica napus*, *Brassica oleracea*, *Brassica rapa*, *Carica papaya*, *Eutrema halophilum*, *Limnanthes alba*, *Raphanus raphanistrum*, *Raphanus sativus*, and *Tropaeolum majus* (Table 2.4-1). Again, *Vitis vinifera* was included as outgroup. The trees reconstructed from the FCGRs of the EST datasets are shown in Figure 2.4-4. The tree based on the Pearson distance and calculated with the Fitch-Margoliash algorithm (Figure 2.4-4C) shows many deviations from the known relationships of the species but also low support for the branchings. Like for the whole genome analysis, the Pearson distance concept is not appropriate for the reconstruction of reliable phylogenetic trees based on FCGRs. The trees based on the Euclidean distance (Figure 2.4-4A, B and D) have almost identical (low-resolution $k = 3$ compared to high-resolution data $k = 8$) to identical topologies (Fitch-Margoliash compared to NJ algorithm). Especially the species of the Brassicaceae clade are well resolved and their topology is highly supported in all trees. The Limnanthaceae, Tropaeolaceae, and Caricaceae are sistergroups of the Brassicaceae. To our knowledge there is no highly resolved tree of these groups available that we could use as reference. Based on our experience with the whole genome data we suppose that the trees based on high-resolution data represent the more reliable topologies.

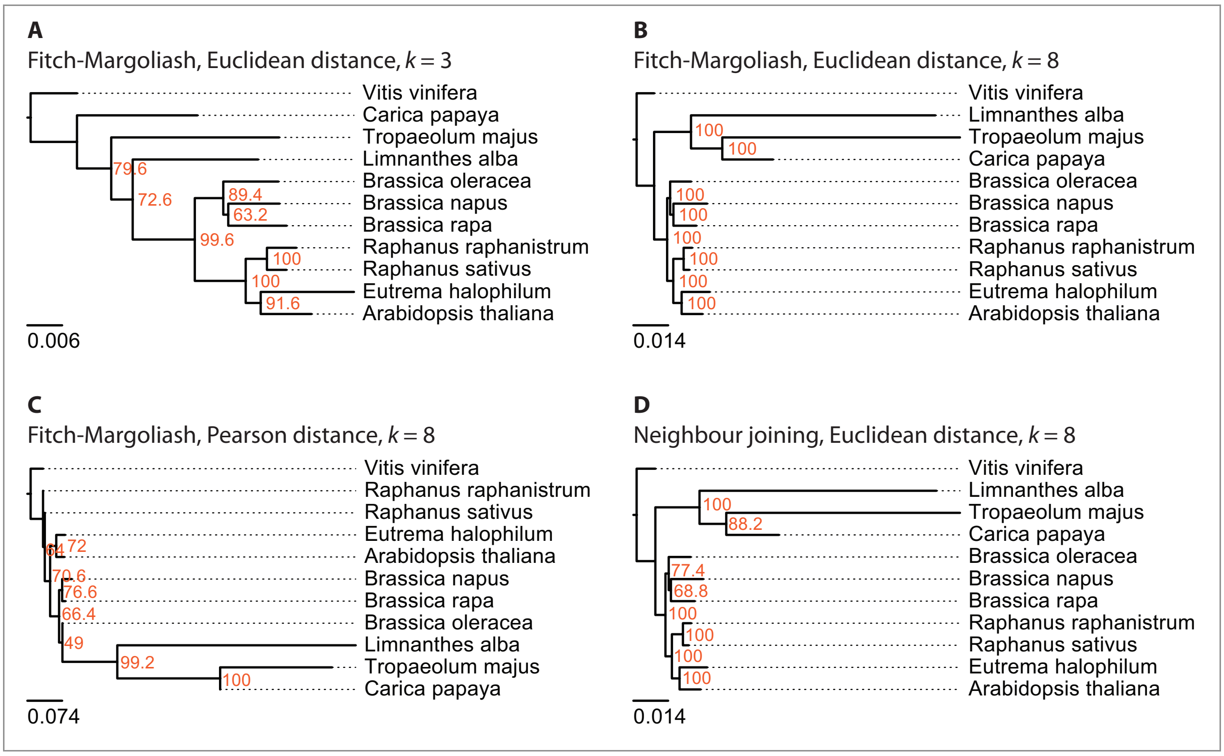


Figure 2.4-4 | EST data analysis. The phylogenetic trees were generated applying different methods on FCGRs of public available EST data of the Brassicales taxon. In A), B) and C) the Fitch-Margoliash algorithm was used to calculate trees for 500 re-sampled datasets. Subsequently, a consensus tree was built and branch lengths were calculated based on the fixed consensus tree. The methods used for the distance calculation and the resolution of the FCGRs are given on top of the trees. In D) the Neighbour joining algorithm was used to calculate the tree.

Mitochondrial genome analysis

For this analysis close relatives of *Brassica rapa* were chosen, for which sequenced mitochondria are available from NCBI. There were nine species in the rosids taxon, whose mitochondrial genome sequences were available: *Arabidopsis thaliana*, *Brassica carinata*, *Brassica juncea*, *Brassica napus*, *Brassica oleracea*, *Brassica rapa*, *Lotus japonicus*, *Millettia pinnata*, and *Ricinus communis* (Table 2.4-1). The mitochondrial genome of *Vitis vinifera* was used as outgroup. In contrast to the analyses of the other datasets, the trees based on the FCGRs of the mitochondrial genomes were very similar for the four different methods (Figure 2.4-5). Especially the sub-branches containing the five closely related *Brassica* species show exactly the same topology supported by high bootstrap values. While the topology of the Brassicales subfamily tree is well resolved the grouping of the Fabales *Lotus japonicus* and *Millettia pinnata* and the Malpighiales *Ricinus communis*, which all belong to the fabids, is different in the four trees. Here, the trees based on the Euclidean distance with high resolution FCGRs ($k = 8$) have the same well-supported topology grouping the Fabales together (Figure 2.4-5B and D) independently which method has been used for the tree reconstruction. This is in agreement with the results from the whole genome and EST analysis that the use of FCGRs with high resolution results in more reasonable trees, and that the Euclidean method for the calculations of the distances is more appropriate than the Pearson method.

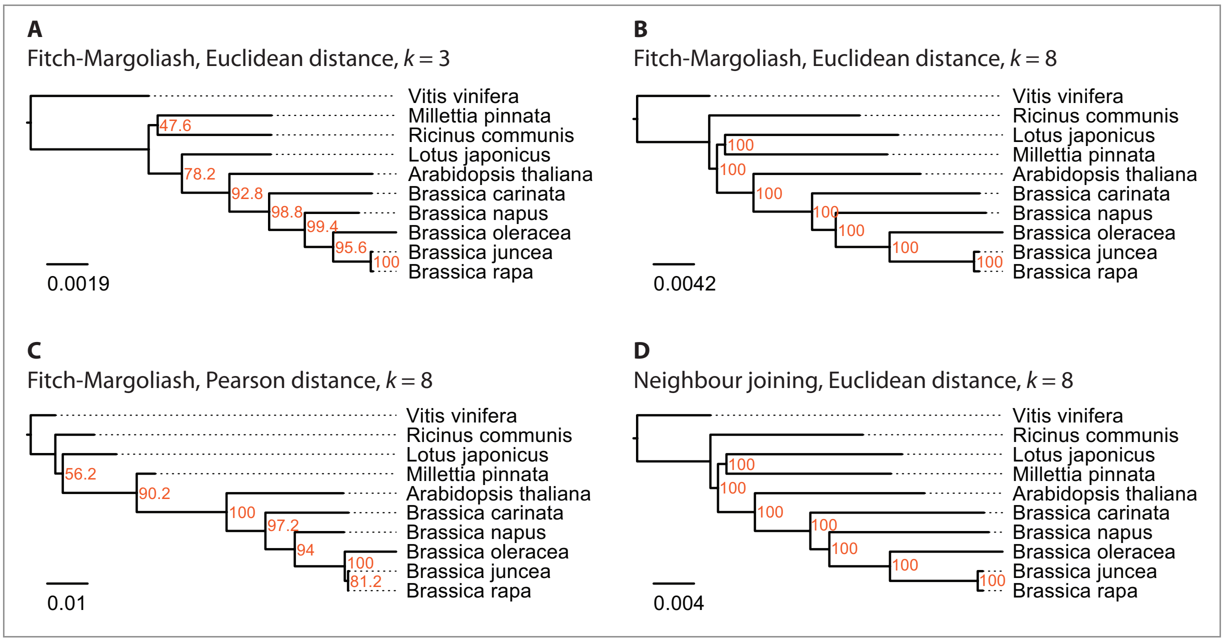


Figure 2.4-5 | Mitochondrial genome analysis. The phylogenetic trees were generated applying different methods on FCGRs of available mitochondrial genome sequence data of the Rosids taxon. In A), B) and C) the Fitch-Margoliash algorithm was used to calculate trees for 500 re-sampled datasets. Subsequently, a consensus tree was built and branch lengths were calculated based on the fixed consensus tree. The method used for the distance calculation and the resolution of the FCGRs are given on top of the trees. In D) the Neighbour joining algorithm was used to calculate the tree.

Computational resource comparison

The algorithm to calculate the CGRs and FCGRs has linear time complexity $O(L)$ and space constant complexity $O(1)$, where L is the length of the nucleotide sequence. In the case of whole genomes, the calculation of the CGRs and FCGRs took about 7,600 s for each genome, for EST data 2,800 s for each species, and 140 s for each mitochondrial genome. The time the algorithm needs to calculate the phylogenetic trees mainly depends on the distance matrix calculated for each species against each other species. This calculation has time complexity $O(4^k s^2)$ and space complexity $O(s^2)$, where s is the number of species and k is the length of the oligonucleotide. The reconstructions of the phylogenetic trees took 98 s for $k = 8$ and the whole genome datasets ($k = 7$: 41 s, $k = 6$: 10 s, $k = 3$: 4 s), 86 s with $k = 8$ for the EST datasets ($k = 7$: 22 s, $k = 3$: 2 s) and 58 s for $k = 8$ and the mitochondrial genome datasets ($k = 7$: 13 s, $k = 3$: 1 s). These values refer to one round of bootstrapping. For comparison, one of the fastest whole genome alignment tools, called Mugsy, needs 45,000 s (ca. 12 h) to align the human and the mouse genomes [196]. However, whole genomes can only be aligned if they are from closely related species and, to our knowledge, phylogenies of multiple sequence alignments of the whole genomes from different eukaryotes have not been reconstructed yet.

2.4.5 Discussion

In general, phylogenetic trees of species are reconstructed from amino acid or nucleotide sequence data, by comparing morphological characteristics, or by combining these data. While most of the sequence-based analyses are built on single genes, concatenated sequences are increasingly used, which could consist of even whole transcriptomes (phylogenomics). Here, we wanted to reconstruct the phylogeny of selected Brassicales species based on alignment-free sequence data. As approach we chose CGRs, which are scale-independent representations for genomic sequences [169]. Because CGRs are unique fingerprints of the corresponding sequences they cannot be compared directly. To reconstruct phylogenetic trees we therefore generated FCGRs at different resolutions. For the calculation of the distances between FCGRs we used the Euclidean (a geometric distance) and the Pearson (a statistical distance) method, and trees were reconstructed with the Fitch-Margoliash and the NJ algorithm.

Because of their different characteristics we compared three types of nucleotide sequences, nuclear genome sequences, mitochondrial genome sequences, and EST reads. Nuclear and mitochondrial genomes have been shown to have different GC contents and codon usage patterns [197]. EST data just comprise the exons and thus only part of the nuclear genome sequences. In addition EST data are potentially biased towards highly abundant genes and 5'- and 3'-terminal sequences. In order to reduce this bias we decided to include only those species for which at least 1,000 EST clones were available. Unfortunately, appropriate species from the Brassicales clade are not available for which all three types of nucleotide data have been sequenced. Therefore, we compared different sets of species for the three data types. Also, it is not known whether the mitochondrial genome data have been extracted from the whole genome datasets. As most of these are denoted as “draft assembly” we assume that the whole genome datasets still contain mitochondrial data. However, because of the very small size of the mitochondrial genomes compared to the nuclear genomes the results should be identical to those obtained from pure nuclear genome data. We would have liked to compare the results of each type of nucleotide sequence with the results of combined datasets but appropriate sequence data is not available. However, the EST and mitochondrial data do not comprise 1% of the whole genome data (Table 2.4-1) and a combined analysis should therefore be dominated by and be identical to the whole genome data.

The mitochondrial and whole genomes of the analysed Brassicales species are of considerably different size, and different amounts of EST data are available. FCGRs naturally depend on the presence and frequency of the respective oligonucleotides and thus on the length of the analysed sequence. For a reasonable result it is therefore essential to find the best balance between sequence length and FCGR resolution (oligonucleotide length), which represents the number of data available for the tree calculations and is also the main determinant for computing time. To exclude that the lengths of the concatenated sequences have an influence on the phylogenetic tree reconstructions of the Brassicales species at high FCGR resolution we calculated trees including the full-lengths sequences and specific defined subsets (Figure 2.4-2). At the resolution of octanucleotides, all partial sequences of whole genome assemblies containing more than ten million nucleotides of each species group together while sets with one million nucleotides result in the ambiguous grouping of some species. In contrast, one million nucleotides of EST data, which correspond to the exon sequences, already result in consistent monophyly of all datasets of each species. Remarkably, this holds even true for the closely related *Brassica* species. The mitochondrial genomes of the analysed species have sizes of 220 to 780 kbp. Thus, at the resolution of hexanucleotides it is not surprising that many oligonucleotides do not exist in subsections of 10 kbp leading to the artificial attraction of all these datasets in the reconstructed tree. Also, datasets of 100 kbp of the different *Brassica* species do not consistently group to the full-length mitochondrial genomes. Therefore, for mitochondrial data the resolution has to be reduced or full-length data to be used. As outgroup we choose *Vitis vinifera* in all analyses.

According to the diArk database, whole genome assemblies are available for 34 species belonging to the Malvidae/malvids [31]. 22 of them are *Arabidopsis thaliana* strains of which we only included the reference strain into the analysis. A species tree including all these sequenced Malvidae is not available. Therefore, we assembled and annotated the capping proteins α- and β-CAP, and the actin related proteins Arp2 and Arp3 to generate a reference tree. The CAP and Arp proteins have been chosen for the reference tree because they are ubiquitous and well conserved in all eukaryotes [178, 179], and duplicates were most probably removed after the many whole genome duplication events that happened in plant evolution [198]. For example, the *Arabidopsis thaliana* genome has experienced two duplications since its divergence from *Carica* [199], but has retained single copies of the CAP and Arp genes [28]. Nevertheless, duplicated CAP and Arp2/3 genes have been identified in the *Brassica rapa* and *Gossypium raimondii* genomes that are, however, the result of species-specific duplications. Only one of each duplicate has been used in this analysis. The phylogenetic tree of the concatenated CAP and Arp proteins is in agreement with other recent analyses containing part of the species [193, 194, 200] and can thus be regarded as reference tree. Compared to this reference tree, the FCGR tree based on the Pearson distance displays the most discrepancies followed by the tree based on low-resolution data ($k = 3$, trinucleotides). In addition, most of the branchings have low bootstrap values. The trees based on high-resolution data ($k = 8$; 65,536 data points) and the Euclidean distance method show overall agreement with the reference trees independent of the method used for the tree reconstruction. Notably, *Carica papaya* and *Brassica rapa* group wrongly, although both are only shifted by one branching event. Most surprisingly, the *Eucalyptus* species are completely wrongly grouped in all FCGR trees. Their exclusion from the tree calculation did not change the grouping of the other species (data not shown). However, the grouping of the Myrtales branch, which contains the *Eucalyptus* species, is different in all published trees [193, 194, 200] and their wrong placement in the FCGR trees might be due to some unknown characteristics of the genomes.

Probably, they would group better, if species from other branches like the Crossosomatales, Geraniales, and Fabidae branches were included in the analysis. The phylogenetic trees of the FCGRs of the mitochondrial genomes are very similar independently of the resolution, distance measure, and tree reconstruction method. Therefore either the species selection was fortunate or mitochondrial genome data is less sensitive with respect to these parameters.

When working with the EST data we observed disproportionate high frequencies for poly-A and poly-T oligonucleotides in the FCGRs. Probably, the poly-A tails were not consistently removed during the cDNA library construction. For low-resolution data (up to $k = 5$) the differences of the frequencies of these oligonucleotides to the next-highest values were not large enough to considerably bias the phylogenetic tree reconstructions. However, the topologies of trees based on high-resolution data ($k > 5$) are strongly disturbed. Therefore, we set the values for the frequencies of the poly-A and poly-T oligonucleotides to zero before we started the tree calculations. The artificial oligonucleotides generated at the boundaries of the concatenated EST reads apparently do not influence the resulting trees. The phylogeny of the *Brassica* species is slightly different compared to that obtained from the mitochondrial genome data. The genus *Brassica* includes 41 species [201] the six with the highest economic importance being *Brassica rapa* (A), *Brassica nigra* (B), *Brassia oleracea* (C), *Brassica napus* (AC), *Brassica juncea* (AB), and *Brassica carinata* (BC). The first three comprise the three elementary species while the other three are amphidiploids that originated from natural hybridizations between two of the elementary species [201]. Thus the amphidiploid EST data contain mixtures of the hybridized species and dependent on which part is overrepresented in the data they will look closer related to one of their parent species. Although the distance in the phylogenetic tree is very small, *Brassica napus* seems to be closer to *Brassica rapa* based on the mitochondrial data. Based on the EST data, the hybrids *Brassica juncea* and *Brassica carinata* are more divergent than the parent species *Brassica rapa* and *Brassia oleracea*. Probably the part of the more divergent parent species *Brassica nigra* is dominating in this case.

In general we could show that FCGRs are well suited to phylogenetically group plant genomes and exonomes from even closely related species. We assume that FCGRs could also be used to group all eukaryotes provided that a balanced set of species from all lineages is taken. This has in part already been demonstrated on the phylogeny of 26 mitochondrial genomes of which only three were placed completely wrong when using the Euclidean distance method [92]. However, this analysis was solely based on data from mitochondrions and biased against fish and mammalian species. Our analysis of the Brassicales clade has shown that high-resolution data (octanucleotides and longer sequences) result in better tree topologies and higher support for branchings. Trees based on the Pearson distance, which is a statistical distance measure, are less reliable than those based on Euclidean distances. The Fitch-Margoliash and NJ algorithms result in similar to identical trees. We have shown for the first time that the bootstrap concept to determine the support of the branchings in the tree, which is well established for trees based on sequence alignments since decades ("taxon-by-character" data matrix; [202]), can also be applied to trees based on FCGRs. In another study it has been shown that although longer word lengths could reveal the correct clustering of the HIV-I subtypes in contrast to shorter word lengths [93] the grouping within the subtypes was always different. Also in this case a bootstrap analysis could have helped in the interpretation of the various branchings and we would recommend applying the bootstrap concept to all phylogenies based on FCGRs. FCGRs are fast to calculate and could be used in combination with alignment based data and morphological characteristics to improve the phylogenetic classification in ambiguous cases.

2.4.6 Acknowledgements

We would like to thank Björn Hammesfahr for helpful suggestions and discussion, and Prof. Christian Griesinger for continuous support.

2.4.7 Authors' contributions

Both authors designed the study, performed data analysis, wrote the manuscript, and approved the final version of the manuscript. KH implemented the software system and performed the phylogenetic computer calculations.

3 Manuscripts

3.1 Kassiopeia: A database for mutually exclusive exomes of eukaryotes

Klas Hatje and Martin Kollmar

Abteilung NMR-basierte Strukturbiologie, Max-Planck-Institut für biophysikalische Chemie, Am Fassberg 11,
D-37077 Göttingen, Germany

Manuscript in preparation

3.1.1 Abstract

Alternative splicing is an important process in higher eukaryotes that allows generating several transcripts out of one gene. One type of alternative splicing is mutually exclusive splicing, which refers to the splicing of exactly one exon out of a cluster of neighbouring exons into the mature transcript. Mutations in one of the exons can lead to human diseases. Recently, a new algorithm for the prediction of these exons has been developed based on the preconditions that the exons of the cluster have similar lengths, sequence homology, and conserved splice sites, and that they are translated in the same reading frame.

In this contribution we introduce Kassiopeia, a web application for the generation, storage, and presentation of genome-wide analyses of mutually exclusive exomes. Currently, Kassiopeia provides the mutually exclusive exomes of twelve sequenced *Drosophila* species, of the thale cress *Arabidopsis thaliana*, of the flatworm *Caenorhabditis elegans*, and of human. All genes were reconstructed with Scipio. Based on the standard prediction parameters, with which 85.7% of the annotated mutually exclusive exons (MXEs) of *Drosophila melanogaster* were found, the exomes contain surprisingly more MXEs than previously supposed and identified. The user can search Kassiopeia using BLAST or browse the genes of each species optionally adjusting the parameters used for the prediction to reveal more divergent or only very similar exon candidates.

3.1.2 Background

Alternative splicing is an important mechanism to increase and regulate the protein content of eukaryotic cells. There is evidence that about 95% of human multi-exon genes undergo alternative splicing [41]. One type of alternative splicing is mutually exclusive splicing, in which exactly one exon of a cluster of several exons is kept in the messenger RNA. The splicing of these mutually exclusive exons (MXEs) is highly regulated in a tissue-specific manner. In humans, missense mutations in MXEs can lead to disease [61, 203]. At the molecular level the splicing is regulated by the secondary structure of the RNA [85, 204].

Different approaches have been developed to identify alternatively spliced isoforms of genes. There are many whole genome studies based on transcriptome sequencing (RNA-Seq), cDNA sequencing, and tiling microarrays (see for example: [50, 205, 206]). The search of tandem mass spectra against genomic databases is also increasingly been used to identify alternatively spliced genes [207]. In contrast to these high-throughput experimental data methods, computer based *de novo* predictions of alternative splicing events are rather complicated. In one approach support vector machine classifiers have been built from gene features that have been experimentally shown to effect alternative splicing [208]. Others used bayesian networks to predict NAGNAG tandem acceptor splice sites [209], genetic programming to classify cassette exons versus retained introns [210], and *ab initio* gene prediction methods [211]. Further, virtual genetic coding schemes combined with time series analyses have been used to predict alternatively spliced genes in *Caenorhabditis elegans* [212].

Recently, we introduced a new method to predict MXEs based on several preconditions to create biological meaningful transcripts [66]. In general exons of a cluster of MXEs encode the same region and thus the same secondary structural elements of the resulting protein structure. Two prominent examples are the arthropod muscle myosin heavy chain [67] and DSCAM genes [213]. The preconditions for MXEs are therefore similar length (sequence length should be fixed in regions forming α-helices and β-strands but slightly flexible in loop regions), conserved splice site patterns (only certain combinations of 5'- and 3'-splice sites are possible), the preservation of the reading frame, and sequence homology. These conditions have been implemented into an algorithm with which many new exon candidates were proposed in an analysis of the genes of the *Drosophila melanogaster* X-chromosome [66].

In order to facilitate the production of datasets of mutually exclusive exomes and to provide a helpful interface for their analysis and presentation we have developed a web application, which we called Kassiopeia. We generated and incorporated data for twelve *Drosophila* species, which are well known to contain many mutually exclusively spliced genes including the highly complex DSCAM gene [214], and for the plant *Arabidopsis thaliana* and the nematode *Caenorhabditis elegans*, for which reports about mutually exclusively spliced genes are rare. Additionally, a preliminary analysis of the human genome was performed. The Kassiopeia web application can be accessed at www.motorprotein.de/kassiopeia.

3.1.3 Construction and content

The database

The database management system is PostgreSQL[63]. The table *proteins* is in the center of the data model with one entry for each protein. Each *proteins* entry contains a key for the dataset, the name of the protein, and additional identifiers like the Genbank Id, the genome target and the locus that codes for the protein. The entries also contain some annotations like the completeness of the Scipio gene structure reconstruction, and the presence of predicted MXEs and of constitutive exons sharing the criteria of MXEs. Each protein belongs to a gene, which is saved in the *genes* table. For each gene the dataset, the target, the locus, the name, additional identifiers and whether the gene contains MXEs based on the protein isoforms of the original annotation are saved. The table *dataset_properties* contains the species scientific name, the taxonomic grouping, the species abbreviation and the release version of the original protein annotations. Each protein is connected to its gene structure and, if available, further data like EST data, cross-species search results and RNA secondary structure predictions for the introns within the cluster of MXEs. The gene structures of all annotated proteins were reconstructed with Scipio [23].

The predicted MXEs are saved in the *exons* table. Each exon entry belongs to a protein, and includes the 3'- and 5'-end position of the exon with respect to the contig/chromosome, the exon number of the original exon, and the score and the length difference as parameters for the significance of the predicted exon. An exon entry might contain further annotations like the overlapping with an exon of another annotated isoform of the gene or with an exon of a neighbouring gene, verification data (e.g. cDNA), and evidence for *trans*-splicing. In order to retain annotations on the same genome target sequences in the case that the predictions are repeated with different parameters or based on new releases of protein annotations, target specific exon annotations like location specific comments, manually verified exon positions and manually entered *trans*-spliced exons are stored in independent database tables.

The web interface

As web application framework we chose Ruby on Rails[64] since it has the advantage of rapid and agile development while keeping the code well organized. The site makes extensive use of Ajax (Asynchronous JavaScript and XML) in order to present the user a feature rich interface while minimizing the amount of transferred data. All technologies used are freely available and open source. The system is running on a Linux machine.

[63] http://www.postgresql.org

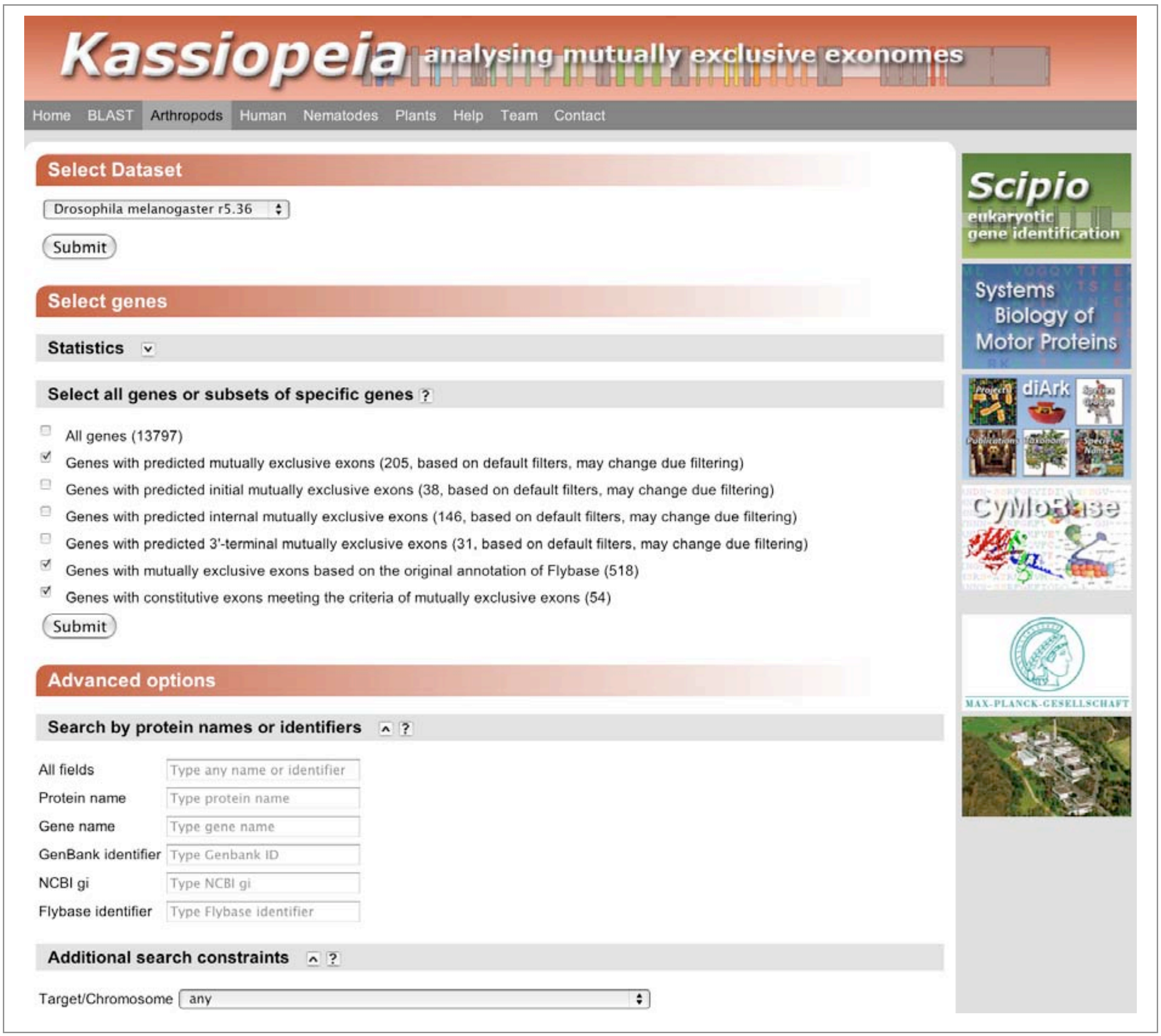
[64] http://rubyonrails.org

Search options

The web interface has been designed to provide easy access to the data while providing specific search and filter options for the expert (Figure 3.1-1). A BLAST [121] service provides a homology-based search against all datasets. The results are linked to gene-specific pages. The entry to whole genome analyses is via taxon-specific pages. Here, a dataset corresponding to one of the available species is chosen (Figure 3.1-1, top). Subsequently all genes can be selected, or single or combined subsets of genes for which MXEs were predicted, genes which are mutually exclusive spliced based on the original annotation from Flybase/Phytozome/Wormbase/NCBI, and/or genes containing annotated constitutive and cassette exons matching the standard prediction criteria (Figure 3.1-1, middle). The latter exons indicate either potentially false positive predictions or false annotations. False annotations might be the case if at least two exons of a cluster are included in one of the annotated transcript isoforms although external evidence like EST and cDNA data is not available. The set of selected genes can be searched by protein name, gene name, and specific identifiers as used in other databases (Figure 3.1-1, bottom). Autocomplete widgets suggest matching names. In addition specific targets can be selected for the analysis of for example a specific chromosome.

Exon filtering

The standard parameters for the prediction of MXEs are reliable to reproduce existing annotations. By applying these parameters in the whole genome searches many new candidates were already predicted. By relaxing any of the parameters both more divergent candidates might be identified as well as wrong exons be predicted. In order not to force users to repeat searches with relaxed parameters, we therefore used more permissive parameters in the Kassiopeia prediction pipeline. Within the advanced options in the Kassiopeia web interface the standard parameters can freely be changed to more restrict or relaxed values.



Select Dataset

Drosophila melanogaster r5.36

Select genes

Statistics

Select all genes or subsets of specific genes ?

- ☐ All genes (13797)
- ☑ Genes with predicted mutually exclusive exons (205, based on default filters, may change due filtering)
- ☐ Genes with predicted initial mutually exclusive exons (38, based on default filters, may change due filtering)
- ☐ Genes with predicted internal mutually exclusive exons (146, based on default filters, may change due filtering)
- ☐ Genes with predicted 3'-terminal mutually exclusive exons (31, based on default filters, may change due filtering)
- ☑ Genes with mutually exclusive exons based on the original annotation of Flybase (518)
- ☑ Genes with constitutive exons meeting the criteria of mutually exclusive exons (54)

Advanced options

Search by protein names or identifiers ?

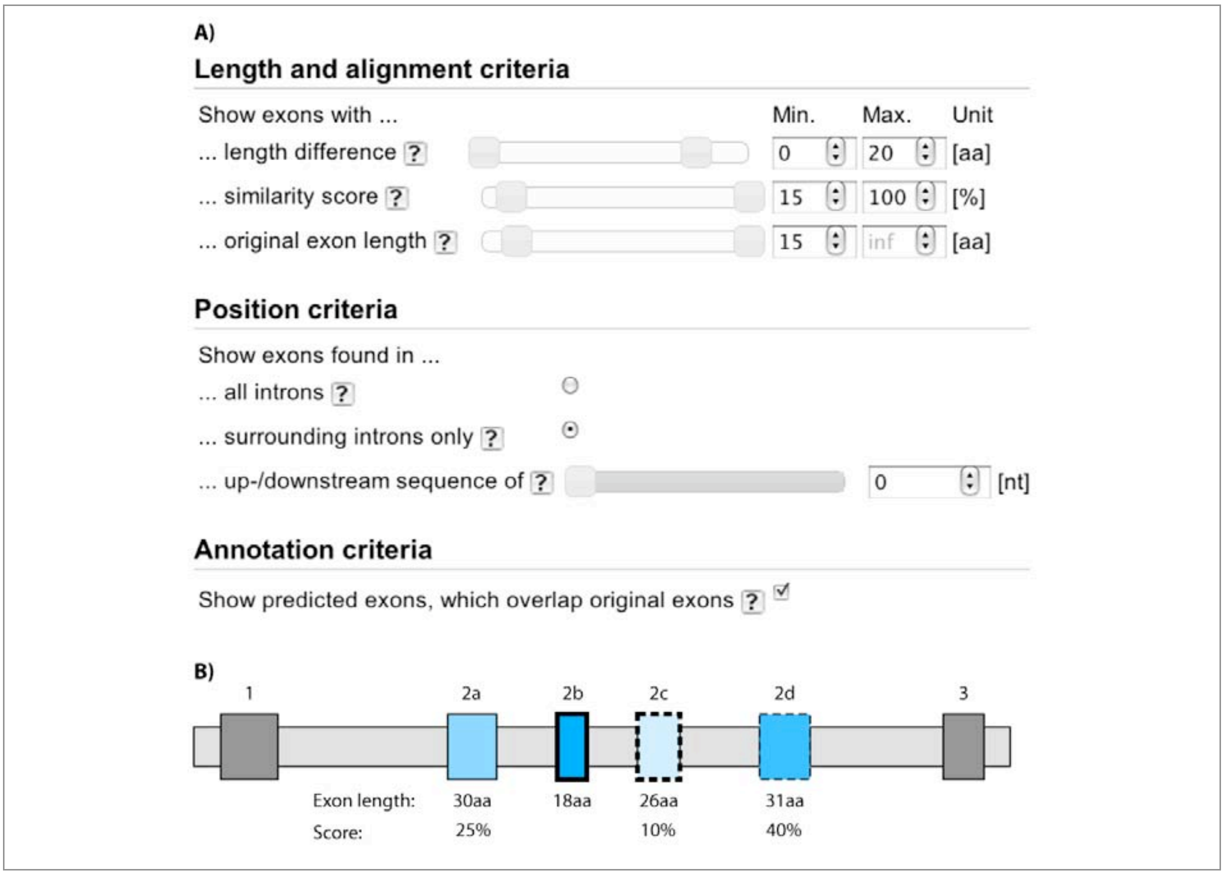
All fields	Type any name or identifier
Protein name	Type protein name
Gene name	Type gene name
GenBank identifier	Type Genbank ID
NCBI gi	Type NCBI gi
Flybase identifier	Type Flybase identifier

Additional search constraints ?

Target/Chromosome: any

Figure 3.1-1 | Dataset selection and search options. The Kassiopeia web application provides an interface to select a dataset from various species and taxa, to choose a specific set of genes, and to search for specific gene names and identifiers. In the example shown the *D. melanogaster* dataset with more than 13,000 genes, of which more than 200 contain predicted MXEs, was selected.

The selected set of predicted MXEs can be filtered by the following criteria (Figure 3.1-2A), which will be explained on the example of a hypothetical cluster of four MXEs as shown in Figure 3.1-2B. In this example the original annotation contains the exons 1, 2b, and 3. Based on the second exon 2b, one alternative exon 5' (exon 2a) and two alternative exons 3' (exons 2c and 2d) were predicted in the introns between exons 1 and 2b, and exons 2b and 3, respectively. If the maximal allowed length difference between the original annotated exon (exon 2b) and the predicted exons (exons 2a, 2c, and 2d) is changed to less than 12 amino acids, exons 2a and 2d would be filtered out. The similarity score is given in percent and defined by the alignment score of the amino acid sequence coded by the original exon to the one of the predicted exon divided by the alignment score of the amino acid sequence coded by the original exon to itself. Given the standard minimal score with 15%, exon 2c in the example would be filtered out (Figure 3.1-2B). The minimal original exon length filter allows preventing predictions based on very short exons. If the minimal exon length were set to a value higher dictions based on very short exons. If the minimal exon length were set to a value higher than 18 amino acids, all alternative exons of the original exon 2b would all be filtered out.



A)

Length and alignment criteria

Show exons with ...	Min.	Max.	Unit
... length difference ?	0	20	[aa]
... similarity score ?	15	100	[%]
... original exon length ?	15	inf	[aa]

Position criteria

Show exons found in ...

... all introns ? ○

... surrounding introns only ? ●

... up-/downstream sequence of ? 0 [nt]

Annotation criteria

Show predicted exons, which overlap original exons ? ☑

Figure 3.1-2 | Exon filtering. A) The Kassiopeia web application provides an interface to filter the predicted MXEs by the parameters of the MXEs search algorithm and a filter to exclude predicted exons, which overlap with exons of other isoforms of the original annotation. B) The effects of the different filter parameters are demonstrated on the example of a hypothetical gene containing a cluster of four MXEs. The gene includes three exons in its original annotation, exons 1, 2b, and 3 (constitutive exons are displayed as dark grey boxes; light-gray boxes denote introns). The algorithm found alternative exon 2a 5' of 2b, and the two alternative exons 2c and 2d in the intron between exons 2b and 3. The exon candidates of the cluster of MXEs are drawn in blue. Scores and lengths of the predicted exons are given to demonstrate the potential effect of the filters. Dashed borderlines around MXEs indicate predicted exons that are not present in any annotated isoform, in contrast to continuous lines that indicate exons already annotated in at least one isoform. Exons with a thick borderline are manually verified by EST data, cross-species gene data, or have already been described in the literature.

MXEs are expected to be located next to each other as part of a cluster. Because annotations might contain mis-predicted exons within a cluster of MXEs the Kassiopeia prediction pipeline was set up to search for exon candidates in all introns. By default, only those MXEs are selected that were predicted in the introns surrounding the original exon (Figure 3.1-2A). To allow the identification of MXEs of partial genes, in which the 5'- and/or 3'-ends of the genes are missing, the exon prediction has been extended into the up- and downstream regions of the genes. The length of these regions, for which predicted exon candidates are displayed, can be varied. However, this option must be treated with caution, because the number of false positive predictions might increase. Cases for false positives are clusters of terminal exons, whose inclusion in the transcripts is regulated by multiple promoters and multiple poly(A) sites and not at the level of splicing, and exons from gene duplicates and *trans*-spliced genes [66]. Exons from gene duplicates and *trans*-spliced genes can be distinguished from MXEs if copies of several exons are found in the up- or downstream regions and if these copies are in the same order as in the original gene. Although not directly related to MXEs, these potentially *trans*-spliced genes and tandem gene duplicates can be displayed by selecting predicted exons found in all introns.

If the original annotations contain several isoforms of a gene, predicted exons in one isoform might overlap with exons of another isoform. If these predicted exons overlap but do not exactly match an exon of the original annotation in another isoform they are most probably false positive predictions and can be deselected (Figure 3.1-2A).

View options and statistics

In the view options section the width of the exons in the graphical output can be scaled and some statistics based on the search results are provided.

Graphical output

The search results are shown as lists of genes represented by the exon-intron structures (Figure 3.1-3A). The gene structure schemes are generated and displayed in the Scalable Vector Graphics (SVG) format[65] for resolution-independent scaling and for convenient interaction with specific graphical elements using JavaScript. For gene colouring we adopted the system used in WebScipio [23]. Exons in a cluster of MXEs get the same colour and the brightness denotes the similarity to the original search exon. Dashed lines around exons indicate newly predicted MXEs and continuous lines mark exons that have already been annotated as MXEs in Flybase/Phytozome/Wormbase/NCBI. Thick lines indicate exons that were verified as MXEs by manually inspecting matching EST data, cross-species search results or literature mining. Constitutive exons with a thick green border represent exons that meet our criteria of MXEs based on the default parameters. If several isoforms for one gene were contained in the annotation datasets, an additional exon-intron structure picture is shown for each isoform. Above the gene structure schemes, a label indicates the completeness or incompleteness of the exon-intron structure. Complete denotes genes for which all amino acids of the protein sequence from the annotation dataset could be mapped onto the genomic sequence. Incomplete gene structures contain gaps (protein sequence not found in the target genome), mismatches or sequence shifts. Details of the gene structures can be analysed by clicking on the WebScipio link on top of the gene structure picture. Below the gene structure schemes, sequence

[65] http://www.w3.org/Graphics/SVG

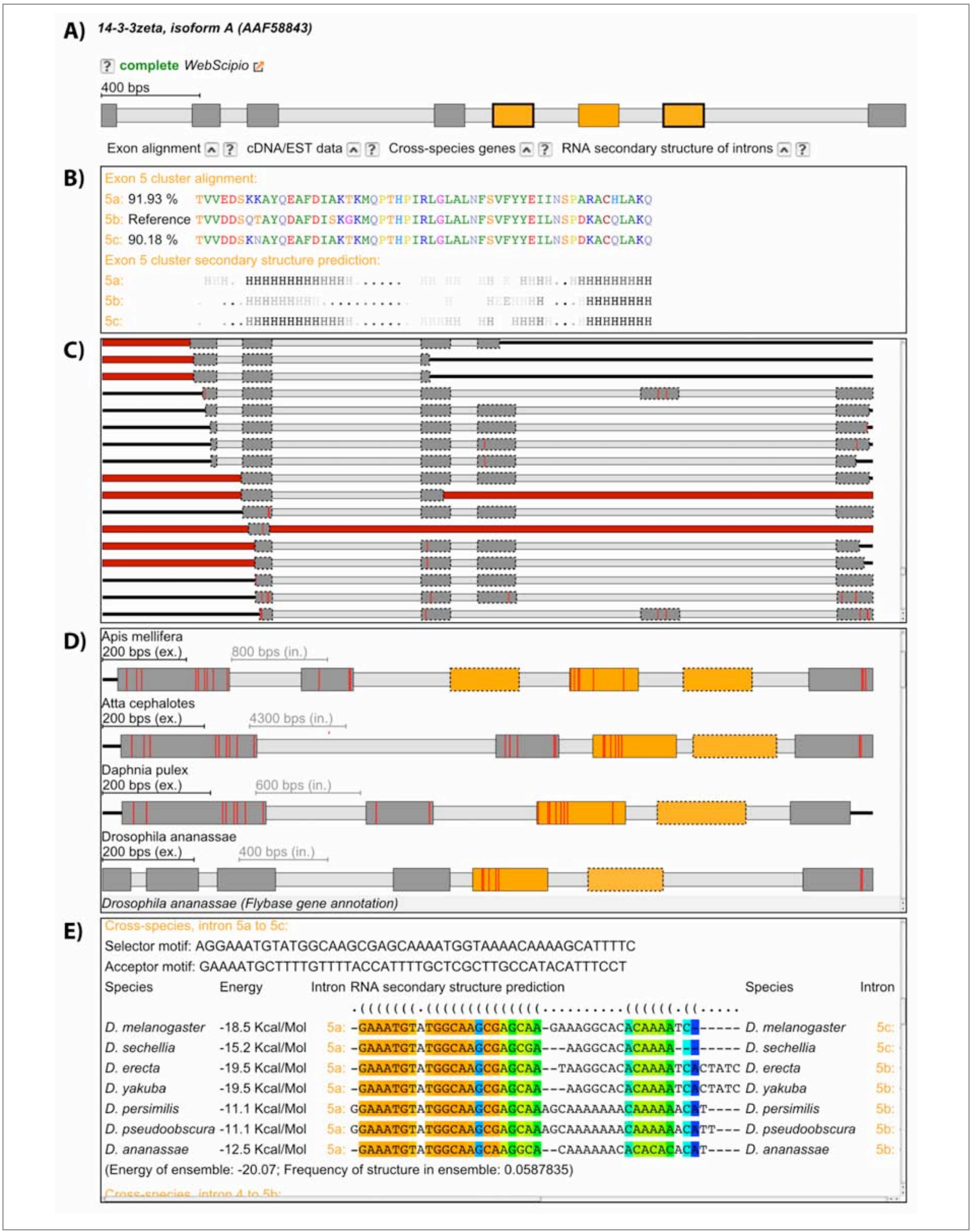
alignments and secondary structure comparisons of the exon candidates are shown (Figure 3.1-3B) and, if available, additional evidence for the MXEs. The alignments of the amino acid sequences coded by the exons in the cluster were calculated with MUSCLE [215, 216] and the secondary structure predictions were done with PSIPRED [217]. The gene structure schemes of the genes can be downloaded directly and those of the isoforms via WebScipio.

Additional evidence for mutually exclusive exons

Experimental validation for the MXEs can be obtained from Expressed Sequence Tags (EST), cDNA and RNA-Seq data. Therefore, we mapped EST data onto the respective gene regions and list hits below the gene structure schemes (Figure 3.1-3C). EST data for these comparisons were retrieved from the EST database of NCBI. The mapping was done by an internally developed method that uses BLAT [29].

Further confidence to the predicted MXEs can be obtained from similar searches in the homologous genes of related organisms. Thus we used Scipio's cross-species search option [23] to identify and reconstruct orthologous genes in related species (Figure 3.1-3D). These genes were then used as basis for the prediction of MXEs. Here, the default parameters were used for the prediction, except that MXEs were searched not only in the surrounding introns of the exons but also in all introns. These predictions are therefore independent of the ones in the original species.

Recently, it has been found that mutually exclusive splicing can be directed by competing intron RNA secondary structures, which was first observed in *Drosophila* [69, 84, 214, 218], but might also exist in mammalian species [82]. Although such competing RNA secondary structures are not found in all clusters of MXEs [84, 218], their identification would provide strong further confidence to any prediction. Therefore we started the prediction of sites in the introns, which could build RNA secondary structures to regulate splicing (Figure 3.1-3E). The binding windows were calculated using a genetic programming algorithm [69]. The first step in this process is the identification of binding windows within the intron preceding the cluster and the internal introns of the cluster, and within the intron following the cluster and the internal introns. Binding windows were predicted for all candidate clusters of MXEs using the SeqAn [96] and the ViennaRNA [219] packages, and, subsequently, also for the available exon-intron gene structures from the related species as obtained in the cross-species searches. For the latter, the identified binding windows of all homologous genes from the different species were aligned using MUSCLE [215, 216] and the RNA secondary structures predicted by RNAalifold [220] from the ViennaRNA package.



```
Exon 5 cluster alignment:
5a: 91.93 %    TVVEDSKKAYQEAFDIAKTKMQPTHPIRLGLALNFSVFYYEIINSPARACHLAKQ
5b: Reference  TVVDDSQTAYQDAFDISKGKMQPTHPIRLGLALNFSVFYYEILNSPDKACQLAKQ
5c: 90.18 %    TVVDDSKNAYQEAFDIAKTKMQPTHPIRLGLALNFSVFYYEILNSPDKACQLAKQ
Exon 5 cluster secondary structure prediction:
5a:              HHH. HHHHHHHHHHHHH......  HH HH  HH E HHHH..HHHHHHHHH
5b:            .   ...HHHHHHHHH..........    H     HHEHHHH ...HHHHHHHH
5c:            .   ...HHHHHHHHHHHHH...... HHHHH  HH  HHHH...HHHHHHHH
```

```
Cross-species, intron 5a to 5c:
Selector motif: AGGAAATGTATGGCAAGCGAGCAAAATGGTAAAACAAAAGCATTTTC
Acceptor motif: GAAAATGCTTTTGTTTTACCATTTTGCTCGCTTGCCATACATTTCCT
Species          Energy          Intron RNA secondary structure prediction              Species          Intron
                                     .(((((((.((((((((((((((...........((((((.((.....
D. melanogaster  -18.5 Kcal/Mol  5a: -GAAATGTATGGCAAGCGAGCAA-GAAAGGCACACAAAATCA----- D. melanogaster  5c:
D. sechellia     -15.2 Kcal/Mol  5a: -GAAATGTATGGCAAGCGAGCGA---AAGGCACACAAAA--A----- D. sechellia     5c:
D. erecta        -19.5 Kcal/Mol  5a: -GAAATGTATGGCAAGCGAGCAA--TAAGGCACACAAAATCACTATC D. erecta        5b:
D. yakuba        -19.5 Kcal/Mol  5a: -GAAATGTATGGCAAGCGAGCAA---AAGGCACACAAAATCACTATC D. yakuba        5b:
D. persimilis    -11.1 Kcal/Mol  5a: GGAAATGTATGGCAAGCGAGCAAAGCAAAAAAACAAAAAACAT---- D. persimilis    5b:
D. pseudoobscura -11.1 Kcal/Mol  5a: GGAAATGTATGGCAAGCGAGCAAAGCAAAAAAACAAAAAACATT--- D. pseudoobscura 5b:
D. ananassae     -12.5 Kcal/Mol  5a: -GAAATGTATGGCAAGCAAGGCA--CAAAAAACACACACACAT---- D. ananassae     5b:
(Energy of ensemble: -20.07; Frequency of structure in ensemble: 0.0587835)
Cross-species, intron 4 to 5b:
```

Figure 3.1-3 | Results of the *Drosophila melanogaster* 14-3-3zeta protein as available in Kassiopeia. The scheme of the exon-intron structure contains exons as dark gray boxes and introns as light gray boxes (A). The exons of a cluster of MXEs have the same colour. The brightness of the predicted exons indicates the similarity to the original exon. The sequence alignments and secondary structure predictions (B), and additional evidence by EST data mapping (C), cross-species search results (D), and RNA secondary structure predictions (E) can be opened below the gene structure scheme.

Analysis of mutually exclusive exomes

For the search for MXEs, annotations for 12 *Drosophila* species, for *Arabidopsis thaliana*, for *Caenorhabditis elegans* and for Homo sapiens were obtained from Flybase , Phytozome, Wormbase and NCBI, respectively:

- ftp://ftp.flybase.net/genome:
  dmel_r5.36_FB2011_04, dana_r1.3_FB2011_07, dere_r1.3_FB2011_08,
  dgri_r1.3_FB2010_02, dmoj_r1.3_FB2011_05, dper_r1.3_FB2010_02,
  dpse_r2.25_FB2011_10, dsec_r1.3_FB2011_08, dsim_r1.3_FB2011_08,
  dvir_r1.2_FB2011_07, dwil_r1.3_FB2010_02, dyak_r1.3_FB2011_08
- ftp://ftp.arabidopsis.org/home/tair/Genes: TAIR10_genome_release
- ftp://ftp.wormbase.org/pub/wormbase/releases: WS230
- ftp://ftp.ncbi.nih.gov/genomes/H_sapiens: Build 37.3

To standardize the procedure for the predictions a pipeline was developed and run for each organism. The pipeline was designed as general as possible to incorporate any annotated genome sequence in the future. As input the pipeline requires the genome sequence and the annotated protein sequences, both in FASTA format. During the prediction process several scripts are started, which were written in the Ruby programming language[66] and C/C++. Within Ruby we use BioRuby [95] to handle the sequences. The outputs of the prediction pipeline are YAML files[67].

Reconstruction of exon-intron gene structures

The first step in the prediction process is the generation of the exome of each organism by mapping the protein sequences onto the genomes using Scipio [36]. Scipio is able to recognize and report shifts in the reading frames of translated genomic sequences, mismatches between the protein query sequence and the translation of the genome sequence, questionable introns that do not match the prevalent intron splice site patterns GT---AG or GC---AG, and missing stop codons (Supplementary information 3.1-1). In some cases small parts of the protein sequences could not be identified in the gene regions due to missing or strongly differing nucleotides in the genome sequence resulting in gaps in the reconstructed genes. These data are missed in the predictions but are, however, insignificant. For example, 64 out of 13,797 reconstructed genes in *D. melanogaster* contain a gap (Supplementary information 3.1-1). Gene reconstructions that include these features are marked as incomplete in the results section of Kassiopeia.

[66] http://www.ruby-lang.org
[67] http://yaml.org

Prediction of mutually exclusive exons

MXEs were predicted in each reconstructed gene using the algorithm described in [66]. If a gene codes for several isoforms, the predictions were done independently for each isoform. The parameters of the prediction pipeline were chosen to be slightly more permissive than the default parameters of WebScipio, which were used in the analyses. This means that more distantly related exons, being true MXEs or potentially false positive predictions, were predicted during the process and are stored in Kassiopeia. The intention is to allow the user a to apply appropriate filters to balance the amount of false positive and false negative predictions during the analysis without having to repeat the overall prediction. In the prediction pipeline the following parameters were used: a maximal length difference of 20 amino acids; a minimal score of 10%; a minimal original exon length of 10 amino acids; exons were predicted in all introns for each exon, and in 20,000 nucleotides up- and downstream of the gene. Only in the preliminary human dataset, the up- and downstream regions were excluded and the maximal length difference was set to 10 amino acids. The analyses shown here (Table 3.1-1, Figure 3.1-4 and Figure 3.1-5) are based on the default parameters of the MXEs search of WebScipio, which are the following: a maximal length difference of 20 amino acids; a minimal score of 15%; a minimal original exon lengh of 15 amino acids; exons are predicted in surrounding introns only and not in the up- and downstream regions. The default parameters are rather strict and more distantly related exons might be missed.

Compiling evidence for mutually exclusive exons

For all genes, that contain candidates of MXEs, EST data were mapped onto the gene region, cross-species searches were executed, and sites to build RNA secondary structures were predicted as described above. These analyses to add confidence to the predicted exon candidates were performed for all twelve *Drosophila* datasets and the *A. thaliana* dataset.

3.1.4 Utility and Discussion

Here, we present the web application Kassiopeia that allows exploring the content of MXEs in whole genomes. Currently, analyses of twelve *Drosophila* genomes, the *Arabidopsis thaliana* genome, the *Caenorhabditis elegans* genome and a preliminary analysis of the human genome are available. A pipeline for the standardized prediction of MXEs has been implemented. The main part of the pipeline is the algorithm for the prediction of MXEs, which is implemented in WebScipio [66]. The predictions were compared with annotations as available from the respective species databases. Further evidence for predicted exons was obtained *in silico* through validation with EST data, comparison with predictions in orthologous genes of related species, and RNA secondary structure predictions. Kassiopeia allows homology-based searching, and selecting and filtering specific parts of the data. Thus, the user can browse the data for specific genes as well as for lists of candidates depending on the prediction parameters. Kassiopeia has been designed to easily adopt the data of any further analysed species, and the data from upcoming versions of genome annotations without loosing the results from the validations and annotations.

Mutually exclusive splicing in twelve *Drosophila* species

The exomes of the twelve completely sequenced *Drosophila* species [221], *D. melanogaster* (dmel), *D. ananassae* (dana), *D. erecta* (dere), *D. grimshawi* (dgri), *D. mojavensis* (dmoj), *D. persimilis* (dper), *D. pseudoobscura* (dpse), *D. sechellia* (dsec), *D. simulans* (dsim), *D. virilis* (dvir), *D. willistoni* (dwil), and *D. yakuba* (dyak), were reconstructed to subsequently predict exons that are spliced in a mutually exclusive manner. The annotations from Flybase contain between 13,797 and 16,639 genes for each species (Table 3.1-1 and Supplementary information 3.1-1).

Table 3.1-1 | Mutually exclusive splicing in twelve *Drosophila* species

Species	dmel	dana	dere	dgri	dmoj	dper	dpse	dsec	dsim	dvir	dwil	dyak
Genes	13797	14917	14842	14635	14431	16639	15805	15936	15261	14353	15359	15845
Proteins	23456	15067	15046	14982	14590	16858	16594	16460	15353	14488	15507	16074
Genes with …												
… multiple exons	11043	11760	11541	11464	11214	12693	11952	12251	11798	11267	11549	12262
… predicted MXEs	205	153	134	168	181	178	171	127	137	166	191	167
… MXEs based on the original annotation	518	0	0	0	0	0	0	0	0	0	0	0
… constitutive exons sharing the criteria of MXEs	54	95	75	87	87	77	93	79	69	51	86	87
Exons in original annotation	60064	55971	55563	55602	54355	58060	57671	57240	52756	54441	55934	57989
Predicted MXEs	763	514	450	551	612	524	453	387	335	524	574	511
MXEs based on the original annotation	1296	0	0	0	0	0	0	0	0	0	0	0
Constitutive exons sharing the criteria of MXEs	129	141	130	162	163	130	248	151	133	91	137	153

The table shows Kassiopeia's content of mutually exclusive exons (MXEs) in twelve *Drosophila* species: *Drosophila melanogaster* (dmel), *D. ananassae* (dana), *D. erecta* (dere), *D. grimshawi* (dgri), *D. mojavensis* (dmoj), *D. persimilis* (dper), *D. pseudoobscura* (dpse), *D. sechellia* (dsec), *D. simulans* (dsim), *D. virilis* (dvir), *D. willistoni* (dwil), and *D. yakuba* (dyak).

Alternative splice forms are well annotated in *D. melanogaster* (23,456 protein isoforms), but almost absent in the datasets of the other *Drosophila* species. The *Drosophila* species contain 52,756 to 60,064 annotated exons. 335 to 763 exons were predicted to be candidates for MXEs (Table 3.1-1). In the *D. melanogaster* genome 1,296 exons of the 60,064 exons were already annotated as MXEs (Table 3.1-1). Here, MXEs were defined as being annotated if the exons are in a cluster of neighbouring exons and each of the annotated isoforms of the corresponding gene includes exactly and only one of the exons of the cluster independently of the position of the cluster within the gene. However, most of these exons are terminal exons, which are alternatively spliced in conjunction with the use of alternative transcriptional initiation or 3'-end processing sites, whose regulation need not be at the level of splicing [131]. Of the 1,296 annotated MXEs in *D. melanogaster* only 259 are internal exons, whose splicing is supposed to be regulated by the formation of specific RNA secondary structures [84, 218]. With WebScipio we predicted 222 exons out of these 259 resulting in a sensitivity of 85.7%. Figure 3.1-4 displays the number of predicted MXEs of all twelve *Drosophila* species divided into three types: initial 5'-end exons, internal exons, and 3'-terminal exons. Additionally, the number of exons that have been annotated as constitutive or cassette exons, but match the criteria of MXEs, are shown.

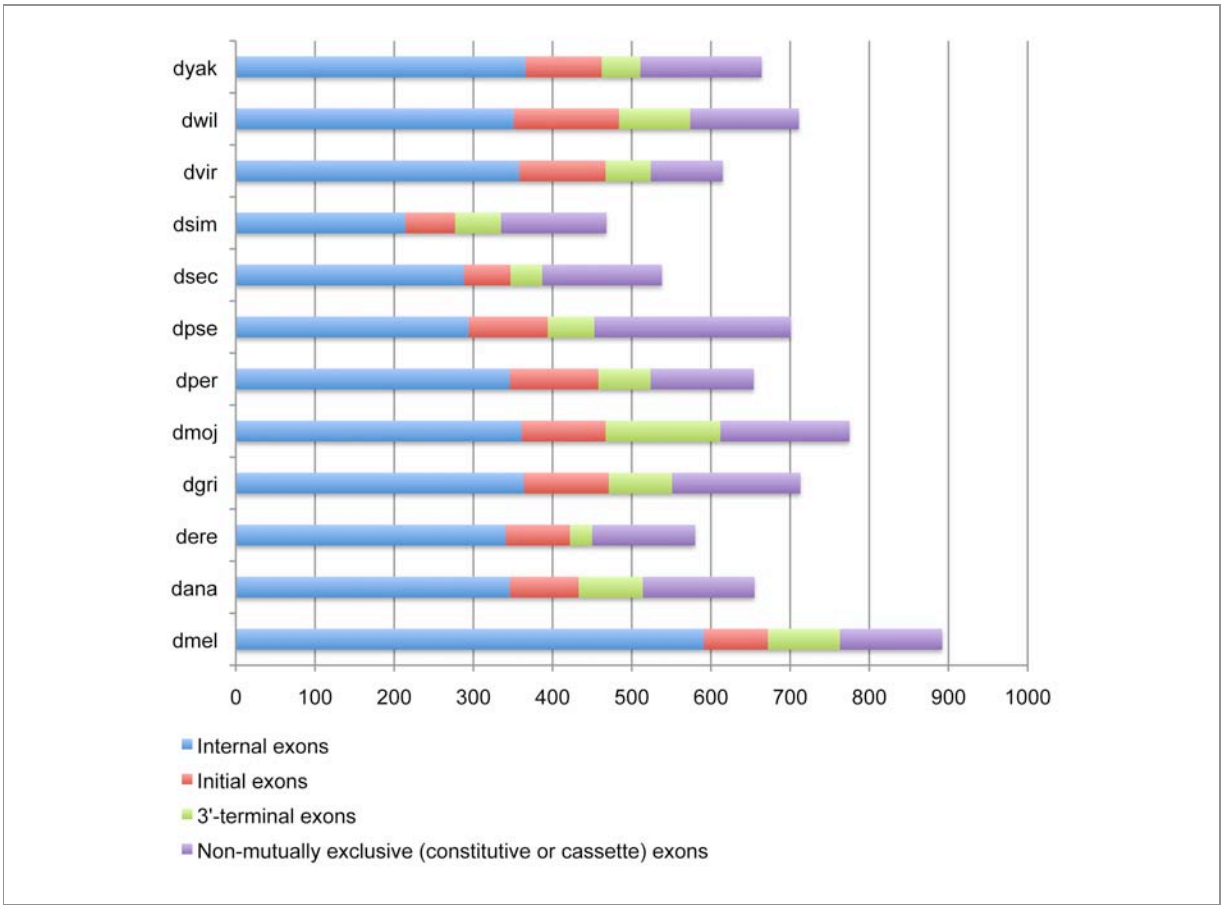


Figure 3.1-4 | Exons in the *Drosophila* genomes that appear in clusters of exons with same reading frames, splice sites, similar lengths and sequence similarity. The numbers for internal exons, initial exons and 3'-terminal exons were derived from the predictions. The exon denoted as non-mutually exclusive meet the criteria of MXEs, but have been annotated as constitutive or cassette exons.

In contrast to the sensitivity, we cannot determine a reliable estimate for the specificity, which considers the false positive predictions. Evaluating the specificity would require a perfectly annotated genome including the knowledge that specific introns, for which we predict MXEs, definitively do not contain any further exons. Future experiments providing further cDNA and EST data could help in determining the specificity by either confirming the predictions or by assigning the exons as constitutive or cassette types.

The annotations available for the other *Drosophila* species do not contain any annotated MXE (Table 3.1-1). Therefore, many of the potentially MXEs were annotated as constitutive exons. For example, all identified exons of the clusters of MXEs in the well-known muscle myosin heavy chain [67] and DSCAM [66, 214] were annotated as constitutive. Accordingly, for most of the other *Drosophila* species the number of constitutive exons that meet the criteria of MXEs is considerably higher than for *D. melanogaster*. We have already shown that many of the predicted MXEs of the *D. melanogaster* X chromosome were also identified as exons in an *ab initio* gene prediction with AUGUSTUS [66]. Therefore we suppose that most of the

129 exons in *D. melanogaster*, which are annotated as constitutive but are not supported by cDNA/EST data yet, might also constitute MXEs.

Mutually exclusive splicing in *Arabidopsis thaliana* and *Caenorhabditis elegans*

Arabidopsis thaliana and *Caenorhabditis elegans* were chosen as representatives for plants and nematodes, respectively, because they are designated model species and many single gene studies as well as whole transcriptome analyses have been performed. Thus, their annotations are supposed to belong to the best available. In the *A. thaliana* genome 166 exons were predicted to be mutually exclusively spliced belonging to 66 genes. 26 of them are initial exons, which are supposed to be spliced by the multiple promoters mechanism, and 41 are 3'-terminal exons containing multiple poly(A) sites (Figure 3.1-5). Thus, 99 exons are candidates for MXEs. In TAIR (The Arabidopsis Information Resource) 139 exons are annotated as MXEs, including 14 internal exons, whose splicing is supposed to be regulated by the formation of specific RNA secondary structures. Those exons are, however, of very different length escaping WebScipio's search algorithm. In the *A. thaliana* gene dataset only four initial exons, but no internal or 3'-terminal exons of the exons predicted by Kassiopeia were already annotated as mutually exclusive (Figure 3.1-5). Our analysis provides the first evidence, that mutually exclusive splicing is also a widely used mechanism to increase the potential number of transcripts in plants. Within PubMed and ArabiTag, a database to a recent very comprehensive analysis of alternative splicing events in *A. thaliana* [222], mutually exclusive spliced genes in *A. thaliana* are not described at all.

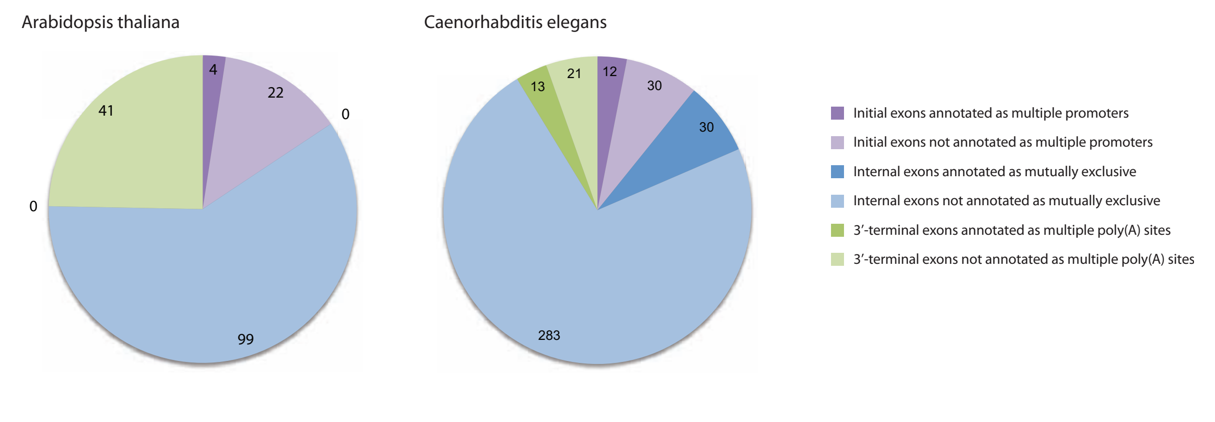


Figure 3.1-5 | Exons in the *Arabidopsis thaliana* and *Caenorhabditis elegans* genomes annotated and predicted as mutually exclusive exons. The graphs represent the number of predicted initial, internal and 3'-terminal exons. Some of these predicted exons are already included in the annotations from Phytozome or Wormbase, especially in the *C. elegans* annotation. The initial exons are supposed to be spliced by the multiple promoters mechanism and the 3'-terminal exons by the multiple poly(A) site mechanism.

In the *C. elegans* genome 389 exons were predicted to be mutually exclusive spliced belonging to 138 genes. 42 of them are initial exons, 313 are internal exons and 34 are 3'-terminal exons (Figure 3.1-5). In the case of *C. elegans* many of the predicted exons are already annotated in the Wormbase: 12 initial exons, 30 internal exons, and 13 3'-terminal exons. However, apart from the terminal exons we identified 283 new candidates for MXEs in internal clusters, about five times more than the largest number reported (55 exons; [75]). These examples show that with Kassiopeia it is possible to identify many new candidates for mutually exclusively spliced genes that were not covered by exhaustive EST data sequencing yet.

3.1.5 Conclusions

Mutually exclusive splicing is a highly regulated mechanism leading to the inclusion of one exon of a cluster of neighbouring exons into the final transcript. We have set up a pipeline to predict MXEs in the whole genomes of several model organisms based on conserved splice sites, same reading frame, sequence similarity and similar length. To make these data easily accessible and informative, we constructed Kassiopeia, a web interface in which researchers can BLAST and search for specific proteins, or browse through whole genomes or chromosomes. For each gene Kassiopeia provides a comprehensive gene structure scheme, sequence and predicted secondary structure alignments, and, if available, further confidence to putative MXEs from cDNA/EST data, comparative predictions in closely related species, and RNA secondary structure information. As standard parameters for the search Kassiopeia offers those with which we could reproduce well-described genes like the DSCAM and the muscle myosin heavy chain gene. However, the user can loosen these parameters to search for more divergent candidate exons.

3.1.6 Availability and requirements

Kassiopeia is maintained under the GPL license and can be accessed at http://www.motorprotein.de/kassiopeia.

3.1.7 Acknowledgements

We would like to thank Björn Hammesfahr, Christian Voss and Prof. Stephan Waack for helpful suggestions and discussions, and Prof. Christian Griesinger for continuous generous support. MK has been funded by grant KO 2251/6-1 of the Deutsche Forschungsgemeinschaft.

3.1.8 Authors' contributions

KH and MK set the requirements for the system. KH wrote the software. KH and MK extensively tested the software, performed all analyses and wrote the manuscript. All authors read and approved the final version of the manuscript.

3.1.9 Supplementary information

Supplementary information 3.1-1 | Detailed statistics of Kassiopeia's content of mutually exclusive splicing in twelve *Drosophila* species

Species	dmel	dana	dere	dgri	dmoj	dper	dpse	dsec	dsim	dvir	dwil	dyak
Genes	13797	14917	14842	14635	14431	16639	15805	15936	15261	14353	15359	15845
Genes with …												
… multiple exons	11043	11760	11541	11464	11214	12693	11952	12251	11798	11267	11549	12262
… predicted mutually exclusive exons	205	153	134	168	181	178	171	127	137	166	191	167
… mutually exclusive exons based on the original annotation	518	0	0	0	0	0	0	0	0	0	0	0
… constitutive exons sharing the criteria of mutually exclusive exons	54	95	75	87	87	77	93	79	69	51	86	87
… gap	64	117	102	109	107	282	122	204	202	111	99	129
... sequence shift	19	34	28	31	37	170	31	105	96	27	30	41
… mismatches	46	37	35	48	42	65	50	55	45	41	45	47
… questionable intron	39	136	126	162	158	311	125	229	278	130	176	157
… missing stopcodon	63	10	2	4	2	10	22	9	7	4	11	5
Proteins	23456	15067	15046	14982	14590	16858	16594	16460	15353	14488	15507	16074
Proteins with …												
… multiple exons	19920	11795	11602	11596	11275	12775	12645	12415	11816	11294	11610	12411
… predicted mutually exclusive exons	797	153	134	169	181	180	203	128	137	166	192	168
… mutually exclusive exons based on the original annotation	2274	0	0	0	0	0	0	0	0	0	0	0
… constitutive exons sharing the criteria of mutually exclusive exons	147	96	75	88	87	77	107	79	69	51	86	87
… gap	139	117	102	109	107	282	129	204	202	111	99	131
... sequence shift	33	34	28	36	37	173	31	105	96	27	31	41
… mismatches	67	37	35	48	42	66	53	56	45	41	46	48
… questionable intron	83	147	127	163	158	311	128	235	278	130	176	160
… missing stopcodon	77	10	2	4	2	10	23	9	7	4	12	5
Exons in original annotation	60064	55971	55563	55602	54355	58060	57671	57240	52756	54441	55934	57989
Predicted mutually exclusive exons	763	514	450	551	612	524	453	387	335	524	574	511
Mutually exclusive exons based on the original annotation	1296	0	0	0	0	0	0	0	0	0	0	0
Constitutive exons sharing the criteria of mutually exclusive exons	129	141	130	162	163	130	248	151	133	91	137	153

3.2 Continuous rapid expansion of the mutually exclusive spliced exome in *Drosophila* species

Klas Hatje and Martin Kollmar

Abteilung NMR-basierte Strukturbiologie, Max-Planck-Institut für biophysikalische Chemie, Am Fassberg 11,
D-37077 Göttingen, Germany

Manuscript in preparation

3.2.1 Abstract

Mutually exclusive splicing is an important mechanism to expand the protein repertoire in a wide range of eukaryotic branches. Here, we present the first genome-wide analysis of mutually exclusive splicing in *Drosophila melanogaster* in comparison to eleven related *Drosophila* species. Our computational approach reports two times more mutually exclusive exons (MXEs) than known so far. We assessed the predictive power of the new method by also applying it to the human, *C.elegans* and *A.thaliana* genomes, and manually inspecting each gene, which includes predicted exon candidates. MXE candidates were evaluated by evolutionary conservation, trancriptome data analysis and identification of competing RNA secondary structural elements. The comparison of the mutually exclusive spliced exomes within the *Drosophila* clade showed that there is a continuous gain and loss of exons. These data substantially expand the number of known clusters of MXEs and provide a straightforward way to analyse future sequenced genomes.

3.2.2 Introduction

Since the first classical genetic experiments with fruit flies by Thomas Hunt Morgan in 1908, *Drosophila melanogaster* is one of the best-analyzed model organisms for genetic studies. The annotation of the *D.melanogaster* genome is in an advanced state, due to protein purification, EST sequencing, whole genome sequencing of other *Drosophila* species, transcriptome sequencing and additional computational methods. Even in this age of genome sequencing, we could show that an important mechanism of alternative splicing, mutually exclusive splicing, is underrepresented in the annotation. We found many new cases of mutually exclusive exons (MXEs) with an algorithm derived from biological knowledge.

The goal of this study is to systematically reconstruct the mutually exclusive spliced exome of a complete genome. We predicted 44 new exons and found multiple evidence for 92% of them. Our study is based on computational methods and public available experimental data only, what makes it reproducible for many additional species. We hypothesize that our approach outreaches experimental approaches, like transcriptome sequencing, in the special case of mutually exclusive splicing. So, this study is an important step in completing the gene annotation of the model organism *D.melanogaster*. In addition, this systematical approach produced a set of hints for the general involvement of the RNA secondary structure in regulating the biophysical mechanism of mutually exclusive splicing.

3.2.3 Results

Discovery of mutually exclusive exons

Here, we present a new approach to determine the mutually exclusive spliced exome of *Drosophila melanogaster* with the help of an *in silico* prediction pipeline [66]. MXEs have to fulfill the following essential preconditions: They need to be arranged next to each other in clusters, the reading-frames must be preserved and splice site patterns like GT---AG, GC---AG or AT---AC must be compatible for flanking constitutive exons and the MXEs. In addition, we expect these exons to have a similar length, if they code for the same region in the tertiary structure of the encoded protein. Thus, length differences are only possible in some loop regions to not disturb the overall protein structure. For the same reason and because MXEs likely evolved from exon duplication events, we expect high sequence similarity between those exons, especially in the slower evolving protein sequences.

Assessing search parameters and annotated mutually exclusive exons

To assess the predictive power of these criteria we analyzed all annotated internal MXEs of *Drosophila melanogaster* (Figure 3.2-1). The number of MXEs was evaluated as a function of sequence similarity and maximal length difference, while the minimal length of the exons was set to 15 aa (Figure 3.2-1A). The *Drosophila* genome contains 60 genes with 261 annotated internal MXEs of which 251 exons (96.2%) in 55 genes (92%) have length differences of less than 25 aa (239 have length differences of less than 10 aa), and 234 exons (89.7%) have a similarity scores of more than 1% within the respective clusters. Using these parameters we would predict 744 genes to encode 3,583 internal MXEs. However, already at more stringent values false positive candidates are predicted like an additional exon candidate for the first cluster of MXEs in the well-studied muscle myosin heavy chain gene (length difference of 1 aa and score of 10.4%). Therefore, we decided to use relatively stringent parameters for the analysis, a maximum length difference of 20 aa and a similarity score of 15%, in order to avoid the incorporation of many false positives while being aware to miss some of the most divergent cases. Under these criteria, 43 genes (71.7%) encode 218 annotated internal MXEs

(83.5% sensitivity), 201 MXE candidates are predicted, of which 44 are completely new exons in 40 genes (Supplementary figure 1, p. 187 and Supplementary figure 2, p. 189). Of the annotated MXEs, which we could not reconstruct, four pairs of exons do not show any sequence similarity, three have length differences of more than 50 aa, three are annotated as differentially included in the latest release (*Dm* r5.48), one pair does not consist of neighboring exons, and two pairs of exons have completely been removed from the latest annotation (Supplementary figure 3, p. 199). Thus, the sensitivity of our method is considerably higher than 83.5%. To exclude that the determined characteristics are *Drosophila* specific we also analysed the annotated mutually exclusive exomes of *Homo sapiens* (NCBI release 37.3), *Caenorhabditis elegans* (WormBase release WS230), and *Arabidopsis thaliana* (TAIR release 167; Supplementary information, p. 171). At a length difference of 20 aa and a similarity score of 15%, 58% (84 of 144) of human MXEs and 54% (19 of 35) of worm MXEs could be reconstructed. The *Arabidopsis* annotation does not contain MXEs matching our criteria.

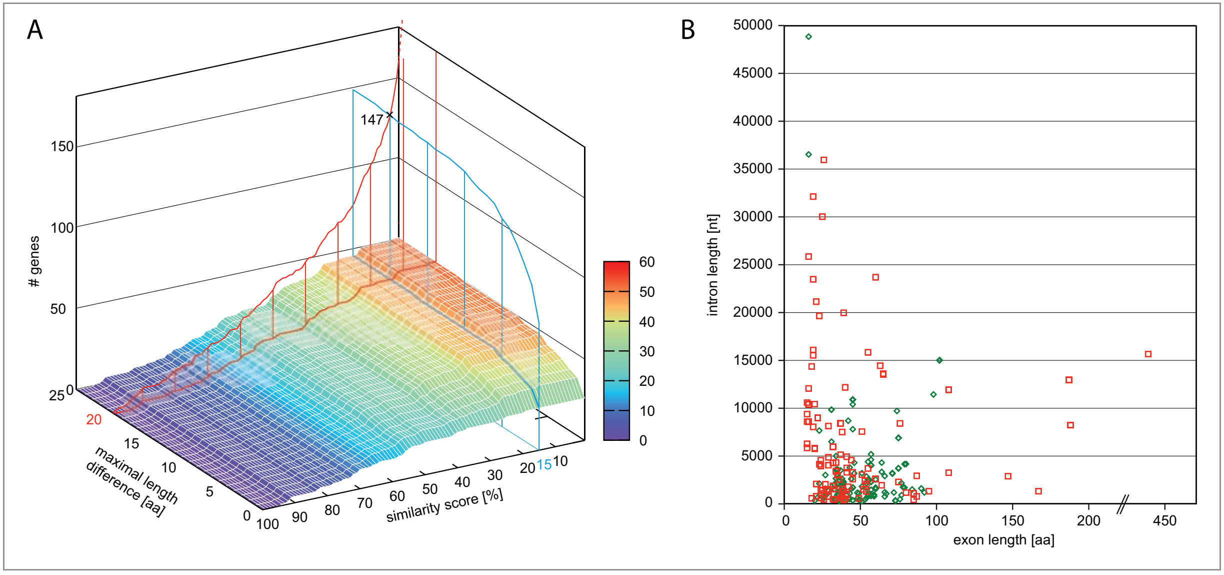


Figure 3.2-1 | Assessment of the search parameters. A) Dependence of the number of genes containing internal MXEs on the maximal length difference and similarity between search exon and MXE candidate. The colored grid denotes the number of genes with MXEs as annotated in FlyBase r5.36 that were also predicted by WebScipio. The red and blue lines mark the number of genes containing predicted MXE candidates at the maximal length difference of 20 amino acids and at the minimal similarity score of 15%, respectively. B) Scatter plot of the internal MXE candidates. Green, annotated in r5.36; red, predicted MXEs.

If exons are short the complexity of the translations will be low and chances will thus be high to predict false positive candidates, especially if the surrounding introns are long. In order to exclude such mis-predictions we analysed the exon lengths of the annotated MXEs (Supplementary information, p. 171). This showed that the lengths of the annotated and reconstructed MXEs are at least 15 residues. A similar value is found for human and *Caenorhabditis* MXEs, and has therefore been applied to the analysis. The introns surrounding annotated MXEs vary from 50 to 50,000 nucleotides (Figure 3.2-1B). Although most introns range up to 15,000 nucleotides we therefore cannot assume that potential MXE candidates in longer introns are false predictions. MXE candidates, which are conserved in other arthropods, were found for example in very long introns of the nAcRalpha-80B and bruno-3 genes (Supplementary figure 4, p. 201). In the case of long exons, it is very unlikely that by chance the translation of intronic region shows sequence similarity to neighbouring exons. However, if long exon candidates are found in long introns these could also, instead of being part of a cluster of MXEs, belong for example to mis-annotated tandemly arrayed gene duplicates or belong to the very rare cases of clusters of exons, which share sequence homology and are spliced as cluster. Here, we also found false positive MXE candidates, that are annotated in the latest FlyBase release as belonging to different tandemly arrayed gene duplicates (CG33243 gene region; FlyBase r5.48), and that were derived from isoforms containing different, mutually exclusive clusters of exons (CG30427 gene; pipe gene [223]; Supplementary figure 4, p. 201 and Supplementary figure 5, p. 202) and from some isoforms of the gigantic dumpy gene that displays a complex pattern of alternative splicing [224].

Estimating the number of false positives and false negatives

The number of false positives and false negatives could only be determined if an absolutely correct annotation of all genes were available. While such a dataset is missing we tried to estimate these numbers by searching for constitutive and differentially included exons that match the criteria of MXEs. Of the 60,401 exons annotated as constitutive or differentially included exons in the *Drosophila melanogaster* genome only 169 exons (0.28%) in 46 genes match these criteria. Several of these exons are even annotated as MXEs in the latest FlyBase release based on RNA-Seq evidence, including a cluster of MXEs in the βTub97EF gene, the Lipophorin receptor 1 gene, and the nicotinic Acetylcholine Receptor α 30D gene (Supplementary figure 6, p. 203). Another gene is now split into two tandemly arrayed gene duplicates (CG10039 is now CG43773 and CG43774). The putative constitutive exons in 15 other genes are now annotated as differentially included or as other types of alternative splice forms. This demonstrates that only a minor part of all internal exons matching the characteristics of MXEs is spliced constitutively. We conclude that most of the new MXE candidates will be spliced mutually exclusive and a minor part of them probably as differentially included.

Exploring the characteristics of mutually exclusive exons

To identify further parameters characterizing MXEs and to ensure that the predicted MXEs have the same features as the already annotated MXEs we analyzed these exons in comparison with all exons and constitutive exons matching the criteria of MXEs (Supplementary information, p. 171). The comparison of the exon and intron lengths did not reveal any distinctive features. The annotated MXEs, that we could not reconstruct, and the constitutive exons, that match the criteria for MXEs, have higher GC contents then the MXEs that we could reconstruct and that we predict. However, the distribution of the GC content is very broad for construct and that we predict. However, the distribution of the GC content is very broad for all types of exons ranging from 30 to 65% so that this cannot be taken as criterion for exclusion. Based on the annotation, MXEs are found in longer genes, and this is also true for the predicted MXE candidates. The codon usage is almost identical in all types of exons except for a considerably higher content of alanines (GCC codon) and glutamines (CAA and CAG) in the MXEs, which are annotated in FlyBase but that we could not reconstruct. The 5' splice junctions of constitutive and mutually exclusive exons are also slightly different, the latter having a higher priority for G in the -1 and a lower priority for GT in the +5 and +6 positions. Analysis of the start and end phases of the exons showed that the percentage of symmetric exons is a bit higher for the predicted and not annotated MXEs (51%) compared to the already annotated, but not predicted MXEs (26%), pointing that some of the predicted MXEs might rather be spliced constitutively or differentially included.

The mutually exclusive spliced exome of *Drosophila melanogaster*

To characterize the mutually exclusive spliced exome, we identified 1,297 MXEs in the annotated *D.melanogaster* genome of which 291 had similar length and sequence. 218 of them were internal MXEs and could be spliced by the competing intron RNA secondary structure mechanism. We predicted 539 exons of similar length and sequence that could be spliced mutually exclusive (two times the annotated exons; Figure 3.2-2). 419 of the MXE candidates were internal including 218 of the already annotated MXEs. Evidence for the predicted MXE candidates was obtained through additional data (Figure 3.2-2A, Supplementary table 1, p. 212): A) Mapping of EST and RNA-Seq data. B) Conservation of the MXE candidates in other arthropods. For this purpose we identified the homologs to the *D.melanogaster* genes in eleven sequenced *Drosophila* species, as well as in *Anopheles gambiae*, *Aedes aegypti*, *Atta cephalotes*, *Apis mellifera*, *Tribolium castaneum*, *Pediculus humanus corporis* and *Daphnia pulex*, and predicted MXE candidates in the homologs using the same pipeline as for *D.melanogaster*. C) *Ab initio* prediction of exonic regions in the respective introns using AUGUSTUS [225]. D) Identification of competing RNA secondary structures. Of the internal MXEs 92% were supported by multiple data types, 21% were supported by EST data. Of the 44 newly predicted internal MXEs eight were supported by EST and/or RNA-Seq data. 94.5% of the annotated and reconstructed internal MXEs and 76.6% of the total predicted internal MXEs are evolutionarily conserved in at least one of the eighteen further analyzed species. In total, only 120 cases of terminal mutually exclusive exons have been identified with similar length and sequence. These exons are, however, spliced by different mechanisms than the internal MXEs and represent only 73 (7.0%) of the annotated 1,036 terminal mutually exclusive exons. As many of these terminal MXE candidates belong to predicted genes that are not supported by full-length cDNA or functional studies yet some might turn to internal exons if further 5' and 3' exons are identified.

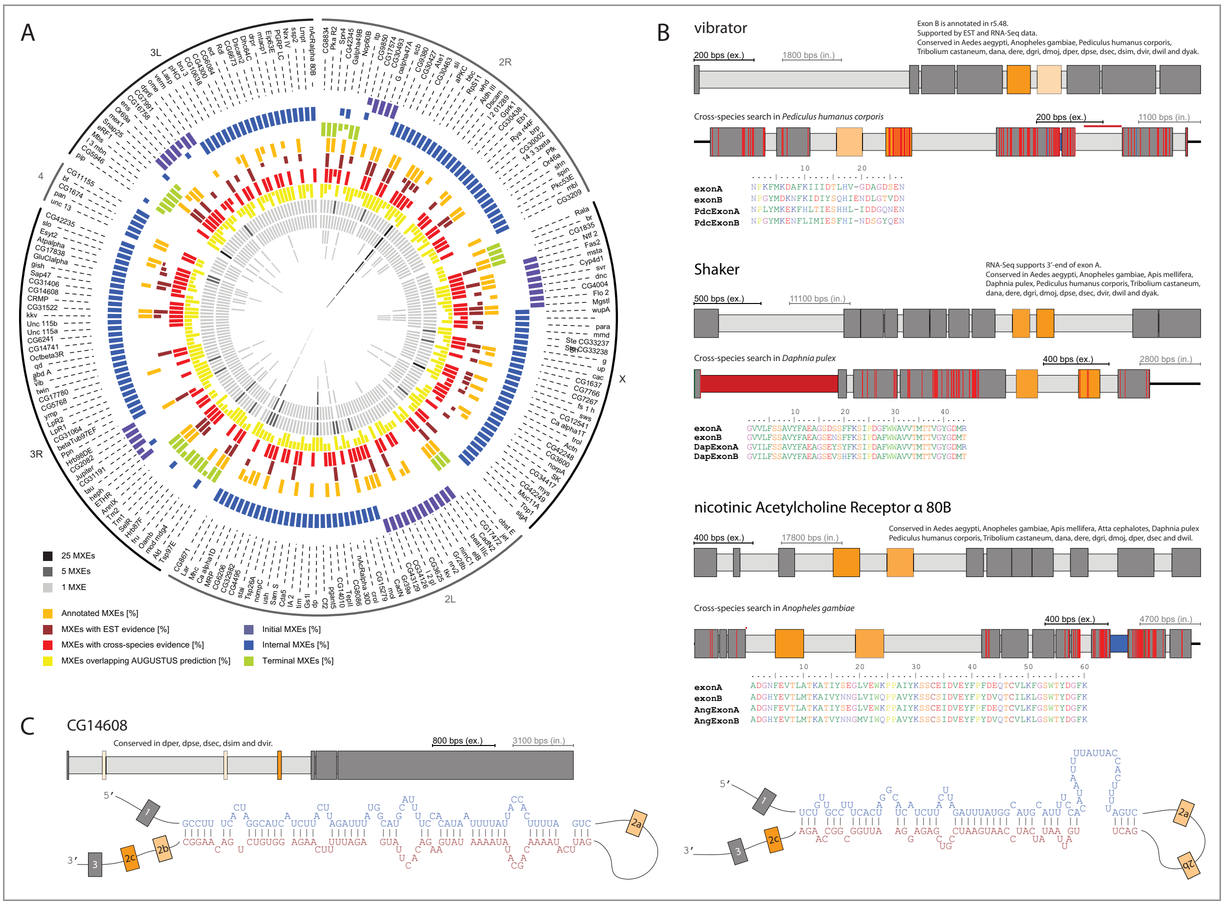


Figure 3.2-2 | Mutually exclusive exons predicted in *Drosophila melanogaster*. A) All genes containing predicted MXEs are listed. The gray bars show how many MXEs are predicted for each gene and the coloured bars show how many of those match specific criteria. B) Exon-intron gene structures of example genes, which contain newly predicted internal MXEs. C) Exon-intron gene structure of a gene, which contains two newly predicted internal MXEs and evidence for the RNA secondary structure splicing mechanism.

The genes containing MXEs are almost evenly spread on all chromosomes. 75% of them are named and have at least one functional study linked in FlyBase. Although these genes were studied at least to some extent in detail, 75% of the new internal MXE candidates were identified in named genes. As an example, the new gene model of the vibrator (vib) gene coding for a phosphatidylinositol transfer protein contains a cluster of MXEs, which was not known in r5.36 but is supported by EST and RNA-Seq data and is included in the latest release r5.48 (Figure 3.2-2B; for the complete list of new MXE candidates not included in r5.36 but in r5.48 see Supplementary figure 1, p. 187). Examples of new clusters of MXEs in well-known genes that are not included in r5.48 include the Shaker (Sh) gene, in which the cluster is conserved in all arthropod species analyzed and of which the 3’-end of the new MXE candidate is supported by RNA-Seq data, and the nicotinic Acetylcholine Receptor α 80B (nAcRalpha-80B) gene, in which the cluster is conserved from *Daphnia* to mosquitoes and *Drosophila* but not yet supported by experimental data (Figure 3.2-2B; for the complete list of new MXE candidates not included in r5.48 see Supplementary figure 2, p. 189).

Alternative splicing of some of the most extensively alternatively spliced genes like the muscle myosin heavy chain genes and the Dscam genes has been shown to be regulated by RNA structures [68, 69]. Docking sites (acceptor sequences) have been identified in the introns before or after the cluster of MXEs to which only one of the selector sequences downstream or upstream of each MXE, respectively, can bind at a time forming conserved base-pairing interactions. Although such sites have only been found for some of the MXE clusters in the 14-3-3ζ, the muscle myosin heavy chain, and the Dscam genes, the mechanism is supposed to also regulate the splicing of other MXE clusters [218]. We searched for complementing sequences in all predicted clusters of MXEs and found favorable sites in many of the annotated clusters. The CG14608 gene exemplifies a predicted cluster of MXEs, for which RNA-Seq evidence is not available but which is supported by cross-species evidence and by competing RNA secondary structure prediction (Figure 3.2-2C, Supplementary figure 7, p. 207).

In order to analyze the conservation pattern of the genes containing MXEs with respect to their involvement in the categories biological process, molecular function and cellular component we performed a GO analysis [226]. Surprisingly, the genes with annotated and reconstructed MXEs as well as the genes with predicted but not annotated MXEs both display strong enrichment in transmembrane transporters, ion channels activity and plasma membrane localization (Supplementary figure 8, p. 208 and Supplementary figure 9, p. 209). However, the clusters of MXEs never encode transmembrane regions.

Evolution of the mutually exclusive spliced exome in 12 *Drosophila* species

It is well known that the clusters of MXEs are highly conserved for example in the *Drosophila* muscle myosin heavy chain genes [67] while some variability has been observed for the DSCAM genes [66, 68]. In order to determine the extent of conservation within the *Drosophila* mutually exclusive spliced exomes we compared the data from *D.melanogaster* (dmel) with the reconstructed corresponding exomes of 11 further *Drosophila* species: *D.simulans* (dsim), *D.sechellia* (dsec), *D.yakuba* (dyak), *D.erecta* (dere), *D.ananassae* (dana), *D.pseudoobscura* (dpse), *D.persimilis* (dper), *D.willistoni* (dwil), *D.virilis* (dvir), *D.mojavensis* (dmoj), and *D.grimshawi* (dgrim). In total, 2,640 clusters were identified most of which are shared among several species, resulting in 770 unique clusters. The genomes of dsim, dsec, and dper are less complete than the other assemblies and were therefore analyzed as group. Overall, the grouping resulted in seven *Drosophila* species or species groups (Figure 3.2-3A). Surprisingly, many of the clusters are unique to one of these groups like 164 clusters within the Drosophila subgenus group (dvir, dmoj, dgri) or 95 clusters within the pseudoobscura group (dpse, dper). Only 68 clusters are conserved in all twelve species (115 in the seven groups). 36 clusters are missing in only one of the species and 16 clusters are absent in any two species. The alternative exons of these clusters could have been lost in these species due to single independent exon loss events or have not been detected. Potential reasons for the latter can be gaps in the assemblies leading to the absence of entire and partial genes or single exons, and exon sequence divergence leading to their exclusion under the given cutoff values. Most clusters are shared by at least two species or species groups and it is very unlikely that assembly gaps are present in independent genomes at exactly the same region in all these cases. Examples are the cacophony gene, for which an additional conserved cluster of MXEs was identified in dwil, dgri, dvir, and all other arthropods analyzed that has, however, been lost in dmel and the other *Drosophila* species (Figure 3.2-3C; Supplementary figure 10, p. 210), and the Ras opposite (Rop) gene, which is a single-exon gene in dmel, dere, dsec and dyak, but a multi-exon gene containing a conserved cluster of MXEs in dwil, dgri, dper, dvir and the other arthropods (Supplementary figure 11, p. 211). These predicted clusters of MXEs therefore represent MXEs of which the alternative exons have been lost in certain species, or exons that have been gained at a certain step in *Drosophila* evolution. To determine the exon gain and loss rates during the evolution of the *Drosophila* species we counted these events based on maximum parsimony requiring the least exon loss events (Figure 3.2-3B). The last common ancestor of the Drosophila species contained at least 186 clusters of mutually exclusive spliced genes (24.2% of all unique clusters). 456 clusters (59.2%) are unique to any of the *Drosophila* species and 111 clusters (14.4%) have been gained in certain branches.

3.2.4 Discussion

Our analysis of the mutually exclusive exome of *D.melanogaster* considerably increased the number of MXE events and thus the fly's ability to vary their gene repertoire. Specifically, we have identified two times more internal MXE candidates than already annotated of which almost 80% are supported by evolutionary conservation and experimental transcript data. This number is surprising given the enormous and long-standing efforts in annotating the *D.melanogaster* genome. However, annotation is a continuous process and even a recent exhaustive exploration of the developmental transcriptome of *D.melanogaster* using RNA-Seq, tilling microarrays and cDNA sequencing failed to detect expression of 12% of the genes although the coverage of the genome and transcriptome were 1,200- and 5,900-fold, respectively [74]. Due to the tight cut-offs of our analysis we are sure that many more MXE events can be identified through manual investigation of the unexplored data. Here, we provide an important step in completing the *D.melanogaster* genome annotation.

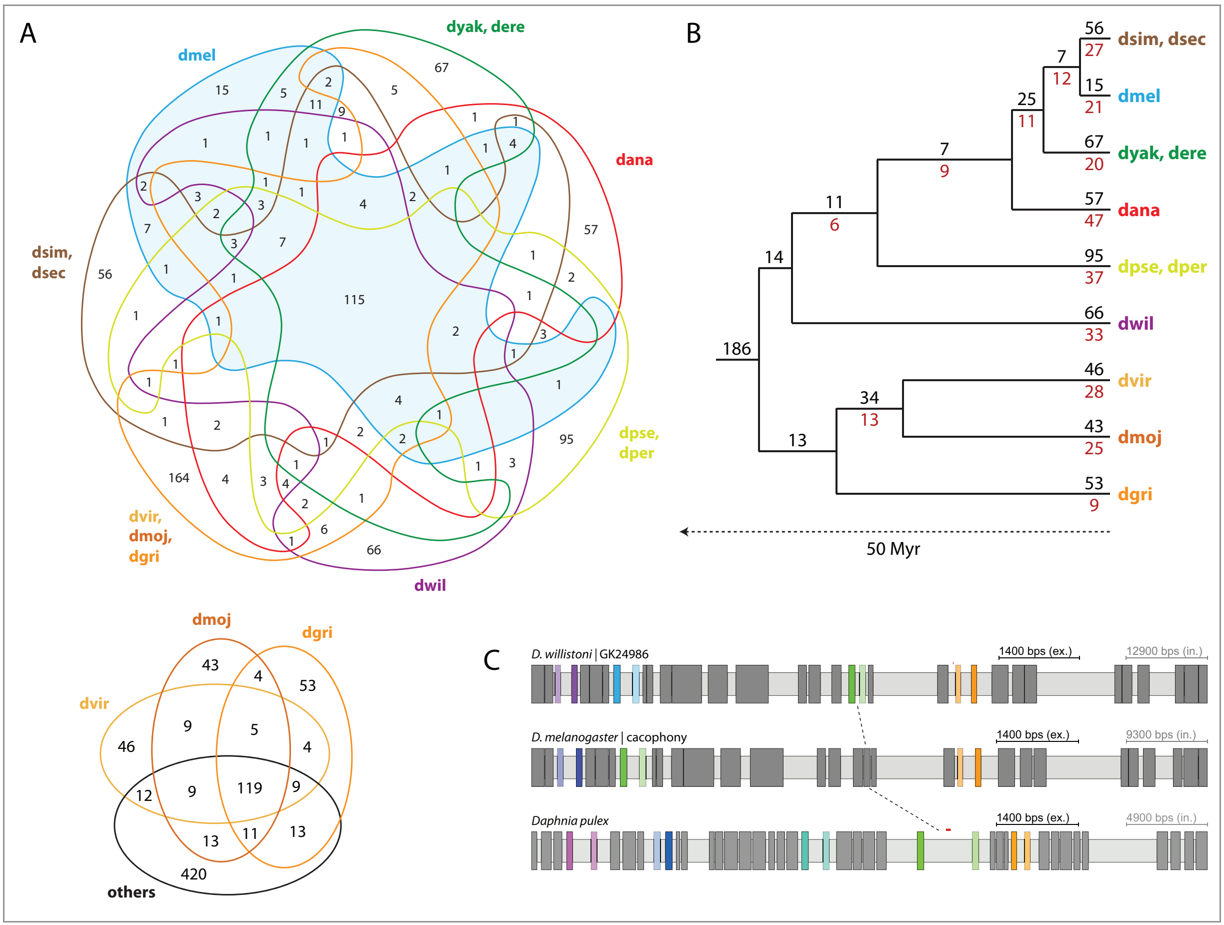


Figure 3.2-3 | Evolution of mutually exclusive splicing clusters. A) The Venn diagrams [227] show how many clusters of MXEs are shared between subsets of species groups. B) The phylogenetic analysis shows how many clusters were probably gained or lost during evolution (black and red numbers, respectively). C) The exon-intron gene structures illustrate an example of an MXE cluster, which has been lost in *Drosophila melanogaster*.

This is, to our knowledge, the first exhaustive genome-wide analysis of a specific alternative splice type purely based on computational methods. The methods have been applied to the human, *C.elegans* and *A.thaliana* genomes for comparison, and to eleven sequenced *Drosophila* species to assess their value in the analysis of less-annotated genomes. The usage of alternative splice types is very different across the species with vertebrates preferring cassette exons while intron retention is very common in fungi and plants. The water flea *Daphnia magna* has more MXEs than *D.melanogaster* and we are sure that this alternative splice type is even more prevalent in other species. Our methods provide a straightforward way to analyze other genomes in the future including resolving artificial fusions of tandemly arrayed gene duplicates and candidates for *trans*-splicing.

3.2.5 Methods

Genome assemblies and annotated proteins for the *Drosophila* species were obtained from FlyBase [140] (r 5.36 for *D.melanogaster*, r 1.2 for *D.virilis*, r 2.25 for *D.pseudoobscura*, and r 1.3 for all other *Drosophila* species), for *Caenorhabditis elegans* from WormBase [228] (WS 230), for *Arabidopsis thaliana* from TAIR [229] (v. 10) and for human from GenBank (v. 37.3). EST data were downloaded from GenBank. The gene structures for the annotated proteins were reconstructed with Scipio [36]. MXEs were predicted in the reconstructed genes using the algorithm implemented in WebScipio [66] with a minor modification favouring GT---AG splice junctions over the other possible splice sites (GC---AG and GG---AG) if several overlapping candidates existed. *Ab initio* exon prediction was done with AUGUSTUS using parameters to find alternative splice forms and the feature set for flies. Cross-species searches and mapping of EST data were done with WebScipio [23]. Binding windows for competing intron RNA secondary structures were predicted for all candidate clusters of MXEs using the SeqAn [230] package. The identified binding windows of all homologous genes were aligned using MUSCLE [215] and the RNA secondary structures predicted by RNAalifold (ViennaRNA package) [219]. The Gene Ontology enrichment analysis was done with AmiGO [231]. All data is available from Kassiopeia (www.motorprotein.de/kassiopeia).

3.2.6 Acknowledgements

We would like to thank Björn Hammesfahr, Christian Voss and Prof. Stephan Waack for helpful suggestions and discussions, and Prof. Christian Griesinger for continuous generous support. MK has been funded by grant KO 2251/6-1 of the Deutsche Forschungsgemeinschaft.

3.2.7 Authors' contributions

KH wrote software and scripts. KH and MK performed all data analyses and wrote the manuscript.

3.2.8 Supplementary information

Supplementary information, supplementary figures and supplementary tables are available in Appendix A1 (p. 171), Appendix A2 (p. 187) and Appendix A3 (p. 212), respectively.

4 Conclusions

This work incorporates six studies that investigate two different topics: gene annotations and phylogenetics. All studies have in common that they are based on sequencing data. They address biological questions that arose or got feasible due to the exponential growth of the sequencing data amount. The annotation of genes is the first step to derive meaningful biological knowledge from this data. The identification of protein-coding genes is important to understand life on the molecular level, because the DNA preserves the design of the proteins, and proteins are the molecular machines working in the cell.

The genetic information that is read out during expression is examined in transcriptome sequencing studies. They allow a better understanding of the regulation of gene transcription and alternative splicing in different states of development and in different cell types. To determine the cellular RNA in each state and type is challenging, but necessary to get a complete picture. We propose that complementary computational studies are important to gain this complete picture, especially since the huge amount of different species cannot be analysed manually in the lab.

Ab initio gene prediction is the first step to get an idea of the gene inventory of a newly sequenced genome. Those predictions are error-prone and incomplete, and therefore they are complemented by homology-based approaches as well as EST and RNA-Seq data, if those data are available.

In protein family analyses the viewpoint of the scientist is different to the whole genome perspective. Here, not the inventory of all genes in a genome is of interest, but whether the genes of a specific protein family are included in several genomes. In such analyses the conservation of protein sequences in combination with conserved intron positions constitute a very reliable validation for the correctness of the collected protein sequences. Tools like Scipio are important to reconstruct the exon-intron gene structure, if a gene annotation is not available or has very low quality. In contrast to other homology-based gene reconstruction tools, the new version of Scipio is able to reconstruct very small exons. In a protein family analysis their existence can be validated directly by the conservation of those small exons in related organisms.

The major limitation of a homology-based approach in contrast to *ab initio* gene prediction is that it is based on already known homologous annotations. In the case of protein family analyses this is not a disadvantage, because some starting sequences may result from *ab initio* predictions and are evaluated and refined during the collecting process. The two approaches complement each other well.

Sequencing approaches must be complemented by computer prediction methods to determine all exons that are alternatively spliced. The reason is that transcriptome studies cannot include each type of cell in each developmental state, and therefore may not include each possible splicing event. Our approach to predict mutually exclusive exons (MXEs) provides a general pattern to find this specific type of alternative splicing, because it depends only on the genomic sequence and an initial annotation, and not on further experimental data. The approach to use sequence information and derive biological meaningful transcripts should also be applicable to other alternative splicing types. The next step would be to search the genomic sequences for specific patterns that define these types. The determination would imply to resolve the splicing code, which is a challenging research field [232].

Our study is based on computational methods and public available experimental data and therefore reproducible for many additional species. We hypothesize that our approach outreaches experimental approaches, like transcriptome sequencing, in the special case of mutually exclusive splicing. That is why, this study is an important step in completing the gene annotation of the model organism *Drosophila melanogaster*. In addition, this systematical approach produced a set of hints for the general involvement of the RNA secondary structure in regulating the biophysical mechanism of mutually exclusive splicing.

One goal of the genome-wide analyses was to find those complementary elements that form a RNA secondary structure in the reconstructed clusters of MXEs (Figure 1.5-4, p. 18). Strong evidence for those interactions could be found for example in clusters of four MXEs in the wings up A gene and the Na pump α subunit gene, and in cluster of three MXEs in the CG14608 gene, the GluClα gene and the slowpoke gene. The MXEs of the gene CG14608 were so far unknown.

The application of the prediction pipeline to the whole *Drosophila melanogaster* genome and the detailed analysis of each predicted case resulted in a precise evaluation of our approach. This makes it a promising tool for every organism whose genome sequence is available and, in which mutually exclusive splicing is not explored so far. There is a lot of interest in the annotation of model organisms like *Arabidopsis thaliana*, *Caenorhabditis elegans* or mouse, in which mutually exclusive splicing was so far not analysed on the genome scale.

In the *Drosophila melanogaster* genome we could find several promising predictions of mutually exclusive splicing clusters. In the genes of Shaker, Topoisomerase 1, moladietz, and nicotinic Acetylcholine Receptor alpha 80B, we found reliable evidence that the newly predicted exons are real. We expect a high impact of these results, because *Drosophila melanogaster* is a model organism with a huge community and a long history. The Drosophila community is very active in improving the gene annotations in the Flybase[68] database [73] resulting in frequent updates.

[68] http://flybase.org

Our method determines all mutually exclusively spliced clusters in a whole genome up to a certain sensitivity. It allows learning how this mechanism evolves in the evolution. We expect that this holds also for the prediction of tandem gene duplicates. A whole genome analysis of tandem gene duplicates in different organisms would be of interest. In addition, it would show if our reconstruction method generates gene annotations that are not known so far.

The limitation of the methods that predict MXEs and tandem gene duplicates is that only exons that share similarity could be found. In the case of mutually exclusive splicing we showed that 90% of those exons in *Drosophila melanogaster* share some similarity.

Kassiopeia is the first database application that makes it possible to analyse whole mutually exclusive exomes of several organisms. The gene annotations of closely related species are linked to each other as demonstrated for the *Drosophila* species. That allows to compare the mutually exclusive splicing clusters directly. The web interface makes the data accessible to the community. To get reliable insight from the predictions it is important to visualise the data in an informative and comprehensive way. The database is not limited to the constraints of the prediction algorithm, because already annotated mutually exclusive splicing events are stored in addition to the predicted ones.

At the moment Kassiopeia is limited to mutually exclusive splicing, but the general system can be extended to more alternative splicing forms. The visualisation part is kept universal and therefore it is already prepared for this purpose. It would be necessary to extend the internal database structure for additional splicing forms. A promising approach to integrate those is implemented in the SpliceGrapher tool [233]. The goal of SpliceGrapher is to find alternative splice forms by combining RNA-Seq data with exon-intron gene structures and EST data. Our predictions would fit very well into this concept. SpliceGrapher is a standalone tool that could be integrated in the Kassiopeia back end. A motivation to develop SpliceGrapher was the assumption that RNA-Seq data would rarely support the prediction of novel splice forms unambiguously [233]. This reinforces our expectation that the prediction of alternative splice forms is necessary, especially in the era of sequencing.

In general, all presented methods show that the information derived from homology is important for reconstructing exon-intron gene structures including alternative splicing, because this data is very reliable, and available for a broad range of species. This holds especially as the number of whole sequenced genomes increases exponentially.

All predictions need further evidence to be plausible. Supportive data must be accessible in combination with the predicted data. This is only possible if the predictions are accessible in every detail. The Kassiopeia web interface is designed to give this precise information. These details are important to have impact on scientific communities in associated biological or medical research fields.

Beside alternative splicing, Kassiopeia is also a promising tool to be used as a general database tool for gene annotations, if uncoupled from functions, which are specific for the mutually exclusive splicing prediction. Research groups, which are interested in sequencing eukaryotic organisms might not have the resources to develop their own databases and user interfaces to store and provide the genome data. Kassiopeia in combination with Scipio would allow an easy way to provide a homology-based annotation of genes.

Nowadays, the reconstruction of the tree of life depends on sequencing data, even though whole genomic sequences are not incorporated in alignment-based approaches, because it is not possible to align the whole genomes of distant species to determine the evolutionary distances. An alignment-free method that reflects the distances between close and distant species would achieve this goal. On the way to develop such a tool it is important to evaluate and improve the available techniques. We evaluated the performance of the CGR method in a specific branch of the plants and with respect to different kinds of data: Whole genome, mitochondrial genome and EST data. In addition, a bootstrap re-sampling method was used the first time to get information about the reliability of the branching points. The evaluation should be a good reference for further alignment-free method evaluations.

We expect this method to be applicable to every branch in the tree of life, because as it was shown, it is able to handle very divergent sequencing data. It benefits from the whole genomic sequence information and is independent of specific protein families. Therefore, alignment-free methods should be used in addition to alignment-based methods.

References

1. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors**. *Proc. Natl. Acad. Sci. U.S.A.* 1977, **74**:5463–5467.

2. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG: **Life with 6000 Genes**. *Science* 1996, **274**:546–567.

3. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**:860 – 921.

4. Pertea M, Salzberg SL: **Between a chicken and a grape: estimating the number of human genes**. *Genome Biol* 2010, **11**:206.

5. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I, Van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guigó R, Hubbard TJ: **GENCODE: the reference human genome annotation for The ENCODE Project**. *Genome Res* 2012, **22**:1760–1774.

6. Metzker ML: **Sequencing technologies - the next generation**. *Nat Rev Genet* 2010, **11**:31–46.

7. Bennett S: **Solexa Ltd**. *Pharmacogenomics* 2004, **5**:433–438.

8. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J-B, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM: **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature* 2005, **437**:376–380.

9. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nat Rev Genet* 2009, **10**:57–63.

10. Ozsolak F, Milos PM: **RNA sequencing: advances, challenges and opportunities**. *Nat Rev Genet* 2011, **12**:87–98.

11. Consortium TEP: **The ENCODE (ENCyclopedia Of DNA Elements) Project**. *Science* 2004, **306**:636–640.

12. ENCODE Project Consortium: **A user's guide to the encyclopedia of DNA elements (ENCODE)**. *PLoS Biol* 2011, **9**:e1001046.

13. Dunham I, Kundaje A, Aldred SF, et al.: **An integrated encyclopedia of DNA elements in the human genome**. *Nature* 2012, **489**:57–74.

14. Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, Gilbert DM, Groudine M, Bender M, Kaul R, Canfield T, Giste E, Johnson A, Zhang M, Balasundaram G, Byron R, Roach V, Sabo PJ, Sandstrom R, Stehling AS, Thurman RE, Weissman SM, Cayting P, Hariharan M, Lian J, Cheng Y, Landt SG, Ma Z, Wold BJ, Dekker J, Crawford GE, Keller CA, Wu W, Morrissey C, Kumar SA, Mishra T, Jain D, Byrska-Bishop M, Blankenberg D, Lajoie1 BR, Jain G, Sanyal A, Chen K-B, Denas O, Taylor J, Blobel GA, Weiss MJ, Pimkin M, Deng W, Marinov GK, Williams BA, Fisher-Aylor KI, Desalvo G, Kiralusha A, Trout D, Amrhein H, Mortazavi A, Edsall L, McCleary D, Kuan S, Shen Y, Yue F, Ye Z, Davis CA, Zaleski C, Jha S, Xue C, Dobin A, Lin W, Fastuca M, Wang H, Guigo R, Djebali S, Lagarde J, Ryba T, Sasaki T, Malladi VS, Cline MS, Kirkup VM, Learned K, Rosenbloom KR, Kent WJ, Feingold EA, Good PJ, Pazin M, Lowdon RF, Adams LB: **An encyclopedia of mouse DNA elements (Mouse ENCODE)**. *Genome Biol* 2012, **13**:418.

15. Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF, Washietl S, Arshinoff BI, Ay F, Meyer PE, Robine N, Washington NL, Di Stefano L, Berezikov E, Brown CD, Candeias R, Carlson JW, Carr A, Jungreis I, Marbach D, Sealfon R, Tolstorukov MY, Will S, Alekseyenko AA, Artieri C, Booth BW, Brooks AN, Dai Q, Davis CA, Duff MO, Feng X, Gorchakov AA, Gu T, Henikoff JG, Kapranov P, Li R, MacAlpine HK, Malone J, Minoda A, Nordman J, Okamura K, Perry M, Powell SK, Riddle NC, Sakai A, Samsonova A, Sandler JE, Schwartz YB, Sher N, Spokony R, Sturgill D, Van Baren M, Wan KH, Yang L, Yu C, Feingold E, Good P, Guyer M, Lowdon R, Ahmad K, Andrews J, Berger B, Brenner SE, Brent MR, Cherbas L, Elgin SCR, Gingeras TR, Grossman R, Hoskins RA, Kaufman TC, Kent W, Kuroda MI, Orr-Weaver T, Perrimon N, Pirrotta V, Posakony JW, Ren B, Russell S, Cherbas P, Graveley BR, Lewis S, Micklem G, Oliver B, Park PJ, Celniker SE, Henikoff S, Karpen GH, Lai EC, MacAlpine DM, Stein LD, White KP, Kellis M: **Identification of functional elements and regulatory circuits by Drosophila modENCODE**. *Science* 2010, **330**:1787–1797.

16. Gerstein MB, Lu ZJ, Van Nostrand EL, et al.: **Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project**. *Science* 2010, **330**:1775–1787.

17. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Falconnet E, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena H, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Luo OJ, Park E, Persaud K, Preall JB, Ribeca P, Risk B, Robyr D, Sammeth M, Schaffer L, See L-H, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Ruan X, Hayashizaki Y, Harrow J, Gerstein M, Hubbard T, Reymond A, Antonarakis SE, Hannon G, Giddings MC, Ruan Y, Wold B, Carninci P, Guigó R, Gingeras TR: **Landscape of transcription in human cells**. *Nature* 2012, **489**:101–108.

18. Holste D, Ohler U: **Strategies for identifying RNA splicing regulatory motifs and predicting alternative splicing events**. *PLoS Comput Biol* 2008, **4**:e21.

19. Baek D, Green P: **Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing**. *Proc Natl Acad Sci USA* 2005, **102**:12813–12818.

20. Stein L: **Genome annotation: from sequence to biology**. *Nat Rev Genet* 2001, **2**:493–503.

21. Guigó R, Flicek P, Abril J, Reymond A, Lagarde J, Denoeud F, Antonarakis S, Ashburner M, Bajic V, Birney E, Castelo R, Eyras E, Ucla C, Gingeras T, Harrow J, Hubbard T, Lewis S, Reese M: **EGASP: the human ENCODE Genome Annotation Assessment Project**. *Genome Biol* 2006, **7**:S2.

22. Zhang MQ: **Computational prediction of eukaryotic protein-coding genes**. *Nat Rev Genet* 2002, **3**:698–709.

23. Hatje K, Keller O, Hammesfahr B, Pillmann H, Waack S, Kollmar M: **Cross-species protein sequence and gene structure prediction with fine-tuned Webscipio 2.0 and Scipio**. *BMC Res. Notes* 2011, **4**:265.

24. Eckert C, Hammesfahr B, Kollmar M: **A holistic phylogeny of the coronin gene family reveals an ancient origin of the tandem-coronin, defines a new subfamily, and predicts protein function**. *BMC Evol. Biol.* 2011, **11**:268.

25. Li Z, Zhang Z, Yan P, Huang S, Fei Z, Lin K: **RNA-Seq improves annotation of protein-coding genes in the cucumber genome**. *BMC Genomics* 2011, **12**:540.

26. Saravanaperumal SA, Pediconi D, Renieri C, La Terza A: **Skipping of exons by premature termination of transcription and alternative splicing within intron-5 of the sheep SCF gene: a novel splice variant**. *PLoS ONE* 2012, **7**:e38657.

27. Arias MC, Danchin EGJ, Coutinho P, Henrissat B, Ball S: **Eukaryote to gut bacteria transfer of a glycoside hydrolase gene essential for starch breakdown in plants**. *Mob Genet Elements* 2012, **2**:81–87.

28. Hammesfahr B, Kollmar M: **Evolution of the eukaryotic dynactin complex, the activator of cytoplasmic dynein**. *BMC Evol Biol* 2012, **12**:95.

29. Kent WJ: **BLAT--the BLAST-like alignment tool**. *Genome Res* 2002, **12**:656 – 664.

30. Needleman SB, Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins**. *J Mol Biol* 1970, **48**:443 – 453.

31. Hammesfahr B, Odronitz F, Hellkamp M, Kollmar M: **diArk 2.0 provides detailed analyses of the ever increasing eukaryotic genome sequencing data**. *BMC Res Notes* 2011, **4**:338.

32. Odronitz F, Hellkamp M, Kollmar M: **diArk--a resource for eukaryotic genome research**. *BMC Genomics* 2007, **8**:103.

33. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S, Feolo M, Fingerman IM, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Mizrachi I, Ostell J, Panchenko A, Phan L, Pruitt KD, Schuler GD, Se-

queira E: **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Res* 2011, **39**:D38 – 51.

34. Slater GS, Birney E: **Automated generation of heuristics for biological sequence comparison**. *BMC Bioinformatics* 2005, **6**:31.

35. Solovyev V, Kosarev P, Seledsov I, Vorobyev D: **Automatic annotation of eukaryotic genes, pseudogenes and promoters**. *Genome Biol* 2006, **7**:S10 11 – 12.

36. Keller O, Odronitz F, Stanke M, Kollmar M, Waack S: **Scipio: using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species**. *BMC Bioinformatics* 2008, **9**:278.

37. Gelfand MS, Mironov AA, Pevzner PA: **Gene recognition via spliced sequence alignment**. *Proc Natl Acad Sci USA* 1996, **93**:9061–9066.

38. Zhou L, Pertea M, Delcher AL, Florea L: **Sim4cc: a cross-species spliced alignment program**. *Nucleic Acids Res* 2009, **37**:e80.

39. Lu DV, Brown RH, Arumugam M, Brent MR: **Pairagon: a highly accurate, HMM-based cDNA-to-genome aligner**. *Bioinformatics* 2009, **25**:1587–1593.

40. Koralewski TE, Krutovsky KV: **Evolution of exon-intron structure and alternative splicing**. *PLoS ONE* 2011, **6**:e18055.

41. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ: **Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing**. *Nat Genet* 2008, **40**:1413 – 1415.

42. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes**. *Nature* 2008, **456**:470–476.

43. Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD: **Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays**. *Science* 2003, **302**:2141–2144.

44. Keren H, Lev-Maor G, Ast G: **Alternative splicing and evolution: diversification, exon definition and function**. *Nat Rev Genet* 2010, **11**:345 – 355.

45. Wahl MC, Will CL, Lührmann R: **The spliceosome: design principles of a dynamic RNP machine**. *Cell* 2009, **136**:701–718.

46. Tarn WY, Steitz JA: **Pre-mRNA splicing: the discovery of a new spliceosome doubles the challenge**. *Trends Biochem Sci* 1997, **22**:132–137.

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

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

49. Kim E, Goren A, Ast G: **Alternative splicing: current perspectives**. *BioEssays* 2008, **30**:38–47.

50. Nagasaki H, Arita M, Nishizawa T, Suwa M, Gotoh O: **Species-specific variation of alternative splicing and transcriptional initiation in six eukaryotes**. *Gene* 2005, **364**:53–62.

51. Allen MA, Hillier LW, Waterston RH, Blumenthal T: **A global analysis of C. elegans trans-splicing**. *Genome Res* 2011, **21**:255–264.

52. Shao W, Zhao Q-Y, Wang X-Y, Xu X-Y, Tang Q, Li M, Li X, Xu Y-Z: **Alternative splicing and trans-splicing events revealed by analysis of the Bombyx mori transcriptome**. *RNA* 2012, **18**:1395–1407.

53. McManus CJ, Duff MO, Eipper-Mains J, Graveley BR: **Global analysis of trans-splicing in Drosophila**. *Proc Natl Acad Sci USA* 2010, **107**:12975 – 12979.

54. Harrow J, Nagy A, Reymond A, Alioto T, Patthy L, Antonarakis SE, Guigó R: **Identifying protein-coding genes in genomic sequences**. *Genome Biol* 2009, **10**:201.

55. Frankish A, Mudge JM, Thomas M, Harrow J: **The importance of identifying alternative splicing in vertebrate genome annotation**. *Database (Oxford)* 2012, **2012**:bas014.

56. Foissac S, Schiex T: **Integrating alternative splicing detection into gene prediction**. *BMC Bioinformatics* 2005, **6**:25.

57. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B: **AUGUSTUS: ab initio prediction of alternative transcripts**. *Nucleic Acids Res* 2006, **34**:W435–439.

58. Barbazuk WB, Fu Y, McGinnis KM: **Genome-wide analyses of alternative splicing in plants: Opportunities and challenges**. *Genome Res* 2008, **18**:1381–1392.

59. Smith CWJ: **Alternative splicing--when two's a crowd**. *Cell* 2005, **123**:1–3.

60. Splawski I, Timothy KW, Sharpe LM, Decher N, Kumar P, Bloise R, Napolitano C, Schwartz PJ, Joseph RM, Condouris K, Tager-Flusberg H, Priori SG, Sanguinetti MC, Keating MT: **Ca(V)1.2 calcium channel dysfunction causes a multisystem disorder including arrhythmia and autism**. *Cell* 2004, **119**:19–31.

61. Mayr JA, Zimmermann FA, Horváth R, Schneider H-C, Schoser B, Holinski-Feder E, Czermin B, Freisinger P, Sperl W: **Deficiency of the mitochondrial phosphate carrier presenting as myopathy and cardiomyopathy in a family with three affected children**. *Neuromuscul Disord* 2011, **21**:803–808.

62. Christofk HR, Vander Heiden MG, Harris MH, Ramanathan A, Gerszten RE, Wei R, Fleming MD, Schreiber SL, Cantley LC: **The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth**. *Nature* 2008, **452**:230–233.

63. David CJ, Chen M, Assanah M, Canoll P, Manley JL: **HnRNP proteins controlled by c-Myc deregulate pyruvate kinase mRNA splicing in cancer**. *Nature* 2010, **463**:364–368.

64. Chen M, Zhang J, Manley JL: **Turning on a fuel switch of cancer: hnRNP proteins regulate alternative splicing of pyruvate kinase mRNA**. *Cancer Res* 2010, **70**:8977–8980.

65. Bluemlein K, Grüning N-M, Feichtinger RG, Lehrach H, Kofler B, Ralser M: **No evidence for a shift in pyruvate kinase PKM1 to PKM2 expression during tumorigenesis**. *Oncotarget* 2011, **2**:393–400.

66. Pillmann H, Hatje K, Odronitz F, Hammesfahr B, Kollmar M: **Predicting mutually exclusive spliced exons based on exon length, splice site and reading frame conservation, and exon sequence homology**. *BMC Bioinformatics* 2011, **12**:270.

67. Odronitz F, Kollmar M: **Comparative genomic analysis of the arthropod muscle myosin heavy chain genes allows ancestral gene reconstruction and reveals a new type of "partially" processed pseudogene**. *BMC Mol Biol* 2008, **9**:21.

68. Graveley BR: **Mutually Exclusive Splicing of the Insect Dscam Pre-mRNA Directed by Competing Intronic RNA Secondary Structures**. *Cell* 2005, **123**:65–73.

69. Anastassiou D, Liu H, Varadan V: **Variable window binding for mutually exclusive alternative splicing**. *Genome Biol* 2006, **7**:R2.

70. Lee C, Kim N, Roy M, Graveley BR: **Massive expansions of Dscam splicing diversity via staggered homologous recombination during arthropod evolution**. *RNA* 2010, **16**:91 – 105.

71. Kuhn RM, Haussler D, Kent WJ: **The UCSC genome browser and associated tools**. *Brief Bioinformatics* 2012.

72. Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, Raney BJ, Pohl A, Malladi VS, Li CH, Lee BT, Learned K, Kirkup V, Hsu F, Heitner S, Harte RA, Haeussler M, Guruvadoo L, Goldman M, Giardine BM, Fujita PA, Dreszer TR, Diekhans M, Cline MS, Clawson H, Barber GP, Haussler D, Kent WJ: **The UCSC Genome Browser database: extensions and updates 2013**. *Nucleic Acids Res* 2012.

73. Drysdale R: **FlyBase : a database for the Drosophila research community**. *Methods Mol Biol* 2008, **420**:45–59.

74. Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, Van Baren MJ, Boley N, Booth BW, Brown JB, Cherbas L, Davis CA, Dobin A, Li R, Lin W, Malone JH, Mattiuzzo NR, Miller D, Sturgill D, Tuch BB, Zaleski C, Zhang D, Blanchette M, Dudoit S, Eads B, Green RE, Hammonds A, Jiang L, Kapranov P, Langton L, Perrimon N, Sandler JE, Wan KH, Willingham A, Zhang Y, Zou Y, Andrews J, Bickel PJ, Brenner SE, Brent MR, Cherbas P, Gingeras TR, Hoskins RA, Kaufman TC, Oliver B, Celniker SE: **The developmental transcriptome of Drosophila melanogaster**. *Nature* 2011, **471**:473–479.

75. Ramani AK, Calarco JA, Pan Q, Mavandadi S, Wang Y, Nelson AC, Lee LJ, Morris Q, Blencowe BJ, Zhen M, Fraser AG: **Genome-wide analysis of alternative splicing in Caenorhabditis elegans**. *Genome Res* 2011, **21**:342–348.

76. Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, Wong W-K, Mockler TC: **Genome-wide mapping of alternative splicing in Arabidopsis thaliana**. *Genome Res* 2010, **20**:45–58.

77. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, Karthikeyan AS, Lee CH, Nelson WD, Ploetz L, Singh S, Wensel A, Huala E: **The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools**. *Nucleic Acids Res* 2011, **40**:D1202–D1210.

78. Smith CW, Nadal-Ginard B: **Mutually exclusive splicing of alpha-tropomyosin exons enforced by an unusual lariat branch point location: implications for constitutive splicing**. *Cell* 1989, **56**:749–758.

79. Letunic I, Copley RR, Bork P: **Common exon duplication in animals and its role in alternative splicing**. *Hum Mol Genet* 2002, **11**:1561–1567.

80. Jones RB, Wang F, Luo Y, Yu C, Jin C, Suzuki T, Kan M, McKeehan WL: **The Nonsense-mediated Decay Pathway and Mutually Exclusive Expression of Alternatively Spliced FGFR2IIIb and -IIIc mRNAs**. *J Biol Chem* 2001, **276**:4158–4167.

81. Yang Y, Zhan L, Zhang W, Sun F, Wang W, Tian N, Bi J, Wang H, Shi D, Jiang Y, Zhang Y, Jin Y: **RNA secondary structure in mutually exclusive splicing**. *Nat. Struct. Mol. Biol.* 2011, **18**:159–168.

82. Pervouchine DD, Khrameeva EE, Pichugina MY, Nikolaienko OV, Gelfand MS, Rubtsov PM, Mironov AA: **Evidence for widespread association of mammalian splicing and conserved long-range RNA structures**. *RNA* 2012, **18**:1–15.

83. Olson S, Blanchette M, Park J, Savva Y, Yeo GW, Yeakley JM, Rio DC, Graveley BR: **A regulator of Dscam mutually exclusive splicing fidelity**. *Nat Struct Mol Biol* 2007, **14**:1134–1140.

84. May GE, Olson S, McManus CJ, Graveley BR: **Competing RNA Secondary Structures Are Required for Mutually Exclusive Splicing of the Dscam Exon 6 Cluster**. *RNA* 2011, **17**:222–229.

85. McManus CJ, Graveley BR: **RNA structure and the mechanisms of alternative splicing**. *Curr Opin Genet Dev* 2011, **21**:373–379.

86. Yang Y, Sun F, Wang X, Yue Y, Wang W, Zhang W, Zhan L, Tian N, Shi F, Jin Y: **Conservation and regulation of alternative splicing by dynamic inter- and intra-intron base pairings in Lepidoptera 14-3-3z pre-mRNAs**. *RNA Biol* 2012, **9**:691–700.

87. Hatje K, Kollmar M: **Predicting Tandemly Arrayed Gene Duplicates with WebScipio**. In *Gene Duplication*. edited by Friedberg F InTech; 2011.

88. Odronitz F, Kollmar M: **Drawing the tree of eukaryotic life based on the analysis of 2,269 manually annotated myosins from 328 species**. *Genome Biol* 2007, **8**:R196.

89. Whelan S: **Inferring trees**. *Methods Mol Biol* 2008, **452**:287–309.

90. Hatje K, Kollmar M: **A phylogenetic analysis of the brassicales clade based on an alignment-free sequence comparison method**. *Front Plant Sci* 2012, **3**:192.

91. Edwards SV, Fertil B, Giron A, Deschavanne PJ: **A genomic schism in birds revealed by phylogenetic analysis of DNA strings**. *Syst Biol* 2002, **51**:599–613.

92. Wang Y, Hill K, Singh S, Kari L: **The spectrum of genomic signatures: from dinucleotides to chaos game representation**. *Gene* 2005, **346**:173–185.

93. Pandit A, Sinha S: **Using genomic signatures for HIV-1 sub-typing**. *BMC Bioinformatics* 2010, **11 Suppl 1**:S26.

94. Vinga S, Almeida J: **Alignment-free sequence comparison-a review**. *Bioinformatics* 2003, **19**:513–523.

95. Goto N, Prins P, Nakao M, Bonnal R, Aerts J, Katayama T: **BioRuby: Bioinformatics Software for the Ruby Programming Language**. *Bioinformatics* 2010, **26**:2617–2619.

96. Doring A, Weese D, Rausch T, Reinert K: **SeqAn an efficient, generic C++ library for sequence analysis**. *BMC Bioinformatics* 2008, **9**:11.

97. Mardis ER: **A decade's perspective on DNA sequencing technology**. *Nature* 2011, **470**:198 – 203.

98. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S: **Genome sequence of the human malaria parasite Plasmodium falciparum**. *Nature* 2002, **419**:498 – 511.

99. Tyler BM, Tripathy S, Zhang X, Dehal P, Jiang RH, Aerts A, Arredondo FD, Baxter L, Bensasson D, Beynon JL: **Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis**. *Science* 2006, **313**:1261 – 1266.

100. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, Iyer VN: **Evolution of genes and genomes on the Drosophila phylogeny**. *Nature* 2007, **450**:203 – 218.

101. Butler G, Rasmussen MD, Lin MF, Santos MA, Sakthikumar S, Munro CA, Rheinbay E, Grabherr M, Forche A, Reedy JL: **Evolution of pathogenicity and sexual reproduction in eight Candida genomes**. *Nature* 2009, **459**:657 – 662.

102. Srivastava M, Begovic E, Chapman J, Putnam NH, Hellsten U, Kawashima T, Kuo A, Mitros T, Salamov A, Carpenter ML: **The Trichoplax genome and the nature of placozoans**. *Nature* 2008, **454**:955 – 960.

103. Prochnik SE, Umen J, Nedelcu AM, Hallmann A, Miller SM, Nishii I, Ferris P, Kuo A, Mitros T, Fritz-Laylin LK: **Genomic analysis of organismal complexity in the multicellular green alga Volvox carteri**. *Science* 2010, **329**:223 – 226.

104. Picardi E, Pesole G: **Computational methods for ab initio and comparative gene finding**. *Methods Mol Biol* 2010, **609**:269 – 284.

105. Wei C, Brent MR: **Using ESTs to improve the accuracy of de novo gene prediction**. *BMC Bioinformatics* 2006, **7**:327.

106. Stanke M, Tzvetkova A, Morgenstern B: **AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome**. *Genome Biol* 2006, **7**:S11 11 – 18.

107. Babenko VN, Rogozin IB, Mekhedov SL, Koonin EV: **Prevalence of intron gain over intron loss in the evolution of paralogous gene families**. *Nucleic Acids Res* 2004, **32**:3724 – 3733.

108. Roy SW, Gilbert W: **Rates of intron loss and gain: implications for early eukaryotic evolution**. *Proc Natl Acad Sci USA* 2005, **102**:5773 – 5778.

109. Odronitz F, Pillmann H, Keller O, Waack S, Kollmar M: **WebScipio: an online tool for the determination of gene structures using protein sequences**. *BMC Genomics* 2008, **9**:422.

110. Van Nimwegen E, Paul N, Sheridan R, Zavolan M: **SPA: a probabilistic algorithm for spliced alignment**. *PLoS Genet* 2006, **2**:e24.

111. Odronitz F, Becker S, Kollmar M: **Reconstructing the phylogeny of 21 completely sequenced arthropod species based on their motor proteins**. *BMC Genomics* 2009, **10**:173.

112. Yoon SJ, Seiler SH, Kucherlapati R, Leinwand L: **Organization of the human skeletal myosin heavy chain gene cluster**. *Proc Natl Acad Sci USA* 1992, **89**:12078 – 12082.

113. Deutsch M, Long M: **Intron-exon structures of eukaryotic model organisms**. *Nucleic Acids Res* 1999, **27**:3219 – 3228.

114. Benton MJ, Donoghue PC: **Paleontological evidence to date the tree of life**. *Mol Biol Evol* 2007, **24**:26 – 53.

115. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S: **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Res* 2011:D38 – 51.

116. Salamov AA, Solovyev VV: **Ab initio gene finding in Drosophila genomic DNA**. *Genome Res* 2000, **10**:516 – 522.

117. Birney E, Clamp M, Durbin R: **GeneWise and Genomewise**. *Genome Res* 2004, **14**:988 – 995.

118. Yeh RF, Lim LP, Burge CB: **Computational inference of homologous gene structures in the human genome**. *Genome Res* 2001, **11**:803 – 816.

119. Stanke M, Waack S: **Gene prediction with a hidden Markov model and a new intron submodel**. *Bioinformatics* 2003, **19**:ii215 – 225.

120. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA**. *J Mol Biol* 1997, **268**:78–94.

121. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**:3389 – 3402.

122. Nilsen TW, Graveley BR: **Expansion of the eukaryotic proteome by alternative splicing**. *Nature* 2010, **463**:457 – 463.

123. Early P, Rogers J, Davis M, Calame K, Bond M, Wall R, Hood L: **Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways**. *Cell* 1980, **20**:313 – 319.

124. Alt FW, Bothwell AL, Knapp M, Siden E, Mather E, Koshland M, Baltimore D: **Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3' ends**. *Cell* 1980, **20**:293 – 301.

125. Mendes Soares LM, Valcarcel J: **The expanding transcriptome: the genome as the "Book of Sand"**. *EMBO J* 2006, **25**:923 – 931.

126. Black DL: **Mechanisms of alternative pre-messenger RNA splicing**. *Annu Rev Biochem* 2003, **72**:291 – 336.

127. Zavolan M, Van Nimwegen E: **The types and prevalence of alternative splice forms**. *Curr Opin Struct Biol* 2006, **16**:362 – 367.

128. Blencowe BJ: **Alternative splicing: new insights from global analyses**. *Cell* 2006, **126**:37 – 47.

129. Alekseyenko AV, Kim N, Lee CJ: **Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes**. *RNA* 2007, **13**:661 – 670.

130. Sugnet CW, Kent WJ, Ares M, Haussler D: **Transcriptome and genome conservation of alternative splicing events in humans and mice**. *Pac Symp Biocomput* 2004:66 – 77.

131. Matlin AJ, Clark F, Smith CW: **Understanding alternative splicing: towards a cellular code**. *Nat Rev Mol Cell Biol* 2005, **6**:386 – 398.

132. Stephan M, Moller F, Wiehe T, Kleffe J: **Self-alignments to detect mutually exclusive exon usage**. *Silico Biol* 2007, **7**:613 – 621.

133. Geeves MA, Holmes KC: **The molecular mechanism of muscle contraction**. *Adv Protein Chem* 2005, **71**:161 – 193.

134. Gotoh O: **An improved algorithm for matching biological sequences**. *J Mol Biol* 1982, **162**:705 – 708.

135. Eddy SR: **Where did the BLOSUM62 alignment score matrix come from?** *Nat Biotechnol* 2004, **22**:1035 – 1036.

136. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks**. *Proc Natl Acad Sci USA* 1992, **89**:10915 – 10919.

137. George EL, Ober MB, Emerson CP: **Functional domains of the Drosophila melanogaster muscle myosin heavy-chain gene are encoded by alternatively spliced exons**. *Mol Cell Biol* 1989, **9**:2957 – 2974.

138. Graveley BR, Kaur A, Gunning D, Zipursky SL, Rowen L, Clemens JC: **The organization and evolution of the dipteran and hymenopteran Down syndrome cell adhesion molecule (Dscam) genes**. *RNA* 2004, **10**:1499 – 1506.

139. Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, Dixon JE, Zipursky SL: **Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity**. *Cell* 2000, **101**:671 – 684.

140. Tweedie S, Ashburner M, Falls K, Leyland P, McQuilton P, Marygold S, Millburn G, Osumi-Sutherland D, Schroeder A, Seal R, Zhang H: **FlyBase: enhancing Drosophila Gene Ontology annotations**. *Nucleic Acids Res* 2009, **37**:D555 – 559.

141. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR: **The genome sequence of Drosophila melanogaster**. *Science* 2000, **287**:2185 – 2195.

142. Zhan XL, Clemens JC, Neves G, Hattori D, Flanagan JJ, Hummel T, Vasconcelos ML, Chess A, Zipursky SL: **Analysis of Dscam diversity in regulating axon guidance in Drosophila mushroom bodies**. *Neuron* 2004, **43**:673 – 686.

143. Neves G, Zucker J, Daly M, Chess A: **Stochastic yet biased expression of multiple Dscam splice variants by individual cells**. *Nat Genet* 2004, **36**:240 – 246.

144. Hummel T, Vasconcelos ML, Clemens JC, Fishilevich Y, Vosshall LB, Zipursky SL: **Axonal targeting of olfactory receptor neurons in Drosophila is controlled by Dscam**. *Neuron* 2003, **37**:221 – 231.

145. Labrador M, Mongelard F, Plata-Rengifo P, Baxter EM, Corces VG, Gerasimova TI: **Protein encoding by both DNA strands**. *Nature* 2001, **409**:1000.

146. Dorn R, Reuter G, Loewendorf A: **Transgene analysis proves mRNA trans-splicing at the complex mod(mdg4) locus in Drosophila**. *Proc Natl Acad Sci USA* 2001, **98**:9724 – 9729.

147. Horiuchi T, Giniger E, Aigaki T: **Alternative trans-splicing of constant and variable exons of a Drosophila axon guidance gene, lola**. *Genes Dev* 2003, **17**:2496 – 2501.

148. Shoja V, Zhang L: **A roadmap of tandemly arrayed genes in the genomes of human, mouse, and rat**. *Mol Biol Evol* 2006, **23**:2134–2141.

149. Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W: **On the origin of new genes in Drosophila**. *Genome Res* 2008, **18**:1446–1455.

150. Long M, Betrán E, Thornton K, Wang W: **The origin of new genes: glimpses from the young and old**. *Nat Rev Genet* 2003, **4**:865–875.

151. Nei M, Roychoudhury AK: **Probability of fixation and mean fixation time of an overdominant mutation**. *Genetics* 1973, **74**:371–380.

152. Ohno S: *Evolution by gene duplication*. Springer-Verlag; 1970.

153. Hahn MW: **Distinguishing among evolutionary models for the maintenance of gene duplicates**. *J Hered* 2009, **100**:605–617.

154. Li W-H, Yang J, Gu X: **Expression divergence between duplicate genes**. *Trends Genet* 2005, **21**:602–607.

155. Massingham T, Davies LJ, Liò P: **Analysing gene function after duplication**. *Bioessays* 2001, **23**:873–876.

156. Babushok DV, Ostertag EM, Kazazian HH Jr: **Current topics in genome evolution: molecular mechanisms of new gene formation**. *Cell Mol Life Sci* 2007, **64**:542–554.

157. Zhang J: **Evolution by gene duplication: an update**. *Trends Ecol Evol* 2003, **18**:292–298.

158. Quijano C, Tomancak P, Lopez-Marti J, Suyama M, Bork P, Milan M, Torrents D, Manzanares M: **Selective maintenance of Drosophila tandemly arranged duplicated genes during evolution**. *Genome Biol* 2008, **9**:R176.

159. Aloni R, Olender T, Lancet D: **Ancient genomic architecture for mammalian olfactory receptor clusters**. *Genome Biol* 2006, **7**:R88.

160. Garcia-Fernàndez J: **The genesis and evolution of homeobox gene clusters**. *Nat Rev Genet* 2005, **6**:881–892.

161. Zhang J, Nei M: **Evolution of Antennapedia-class homeobox genes**. *Genetics* 1996, **142**:295–303.

162. Bertrand D, Lajoie M, El-Mabrouk N: **Inferring ancestral gene orders for a family of tandemly arrayed genes**. *J Comput Biol* 2008, **15**:1063–1077.

163. Elemento O, Gascuel O, Lefranc M-P: **Reconstructing the duplication history of tandemly repeated genes**. *Mol Biol Evol* 2002, **19**:278–288.

164. Saez LJ, Gianola KM, McNally EM, Feghali R, Eddy R, Shows TB, Leinwand LA: **Human cardiac myosin heavy chain genes and their linkage in the genome**. *Nucleic Acids Res* 1987, **15**:5443–5459.

165. Weydert A, Daubas P, Lazaridis I, Barton P, Garner I, Leader DP, Bonhomme F, Catalan J, Simon D, Guénet JL: **Genes for skeletal muscle myosin heavy chains are clustered and are not located on the same mouse chromosome as a cardiac myosin heavy chain gene**. *Proc Natl Acad Sci U S A* 1985, **82**:7183–7187.

166. Sun YM, Da Costa N, Chang KC: **Cluster characterisation and temporal expression of porcine sarcomeric myosin heavy chain genes**. *J Muscle Res Cell Motil* 2003, **24**:561–570.

167. Blair C, Murphy RW: **Recent trends in molecular phylogenetic analysis: where to next?** *J Hered* 2011, **102**:130–138.

168. Dewey CN: **Whole-genome alignment**. *Methods Mol Biol* 2012, **855**:237–257.

169. Jeffrey HJ: **Chaos game representation of gene structure**. *Nucleic Acids Res* 1990, **18**:2163–2170.

170. Li M, Badger JH, Chen X, Kwong S, Kearney P, Zhang H: **An information-based sequence distance and its application to whole mitochondrial genome phylogeny**. *Bioinformatics* 2001, **17**:149–154.

171. Domazet-Loso M, Haubold B: **Efficient estimation of pairwise distances between genomes**. *Bioinformatics* 2009, **25**:3221–3227.

172. Sims GE, Jun S-R, Wu GA, Kim S-H: **Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions**. *Proc Natl Acad Sci U S A* 2009, **106**:2677–2682.

173. Basu S, Pan A, Dutta C, Das J: **Chaos game representation of proteins**. *J Mol Graph Model* 1997, **15**:279–289.

174. Pleissner KP, Wernisch L, Oswald H, Fleck E: **Representation of amino acid sequences as two-dimensional point patterns**. *Electrophoresis* 1997, **18**:2709–2713.

175. Deschavanne PJ, Giron A, Vilain J, Fagot G, Fertil B: **Genomic signature: characterization and classification of species assessed by chaos game representation of sequences**. *Mol Biol Evol* 1999, **16**:1391–1399.

176. Almeida JS, Carriço JA, Maretzek A, Noble PA, Fletcher M: **Analysis of genomic sequences by Chaos Game Representation**. *Bioinformatics* 2001, **17**:429–437.

177. Joseph J, Sasikumar R: **Chaos game representation for comparison of whole genomes**. *BMC Bioinformatics* 2006, **7**:243.

178. Cooper JA, Sept D: **New insights into mechanism and regulation of actin capping protein**. *Int. Rev. Cell Mol. Biol.* 2008, **267**:183–206.

179. Goley ED, Welch MD: **The ARP2/3 complex: an actin nucleator comes of age**. *Nat Rev Mol Cell Biol* 2006, **7**:713–726.

180. Campbell A, Mrázek J, Karlin S: **Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA**. *Proc Natl Acad Sci U S A* 1999, **96**:9184–9189.

181. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD: **Multiple sequence alignment with the Clustal series of programs**. *Nucleic Acids Res* 2003, **31**:3497–3500.

182. Jones DT, Taylor WR, Thornton JM: **The rapid generation of mutation data matrices from protein sequences**. *Comput. Appl. Biosci.* 1992, **8**:275–82.

183. Darriba D, Taboada GL, Doallo R, Posada D: **ProtTest 3: fast selection of best-fit models of protein evolution**. *Bioinformatics* 2011, **27**:1164–1165.

184. Stamatakis A, Hoover P, Rougemont J: **A rapid bootstrap algorithm for the RAxML Web servers**. *Syst Biol* 2008, **57**:758 – 771.

185. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng J-F, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, Haberer G, Hollister JD, Ossowski S, Ottilar RP, Salamov AA, Schneeberger K, Spannagl M, Wang X, Yang L, Nasrallah ME, Bergelson J, Carrington JC, Gaut BS, Schmutz J, Mayer KFX, Van de Peer Y, Grigoriev IV, Nordborg M, Weigel D, Guo

Y-L: **The Arabidopsis lyrata genome sequence and the basis of rapid genome size change**. *Nat Genet* 2011, **43**:476–481.

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

187. Wang X, Wang H, Wang J, et al.: **The genome of the mesopolyploid crop species Brassica rapa**. *Nat Genet* 2011, **43**:1035–1039.

188. Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KLT, Salzberg SL, Feng L, Jones MR, Skelton RL, Murray JE, Chen C, Qian W, Shen J, Du P, Eustice M, Tong E, Tang H, Lyons E, Paull RE, Michael TP, Wall K, Rice DW, Albert H, Wang M-L, Zhu YJ, Schatz M, Nagarajan N, Acob RA, Guan P, Blas A, Wai CM, Ackerman CM, Ren Y, Liu C, Wang J, Wang J, Na J-K, Shakirov EV, Haas B, Thimmapuram J, Nelson D, Wang X, Bowers JE, Gschwend AR, Delcher AL, Singh R, Suzuki JY, Tripathi S, Neupane K, Wei H, Irikura B, Paidi M, Jiang N, Zhang W, Presting G, Windsor A, Navajas-Pérez R, Torres MJ, Feltus FA, Porter B, Li Y, Burroughs AM, Luo M-C, Liu L, Christopher DA, Mount SM, Moore PH, Sugimura T, Jiang J, Schuler MA, Friedman V, Mitchell-Olds T, Shippen DE, dePamphilis CW, Palmer JD, Freeling M, Paterson AH, Gonsalves D, Wang L, Alam M: **The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus)**. *Nature* 2008, **452**:991–996.

189. Dassanayake M, Oh D-H, Haas JS, Hernandez A, Hong H, Ali S, Yun D-J, Bressan RA, Zhu J-K, Bohnert HJ, Cheeseman JM: **The genome of the extremophile crucifer Thellungiella parvula**. *Nat Genet* 2011, **43**:913–918.

190. Argout X, Salse J, Aury J-M, Guiltinan MJ, Droc G, Gouzy J, Allegre M, Chaparro C, Legavre T, Maximova SN, Abrouk M, Murat F, Fouet O, Poulain J, Ruiz M, Roguet Y, Rodier-Goud M, Barbosa-Neto JF, Sabot F, Kudrna D, Ammiraju JSS, Schuster SC, Carlson JE, Sallet E, Schiex T, Dievart A, Kramer M, Gelley L, Shi Z, Bérard A, Viot C, Boccara M, Risterucci AM, Guignon V, Sabau X, Axtell MJ, Ma Z, Zhang Y, Brown S, Bourge M, Golser W, Song X, Clement D, Rivallan R, Tahi M, Akaza JM, Pitollat B, Gramacho K, D'Hont A, Brunel D, Infante D, Kebe I, Costet P, Wing R, McCombie WR, Guiderdoni E, Quetier F, Panaud O, Wincker P, Bocs S, Lanaud C: **The genome of Theobroma cacao**. *Nat Genet* 2011, **43**:101–108.

191. Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyère C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, Felice N, Paillard S, Juman I, Moroldo M, Scalabrin S, Canaguier A, Le Clainche I, Malacrida G, Durand E, Pesole G, Laucou V, Chatelet P, Merdinoglu D, Delledonne M, Pezzotti M, Lecharny A, Scarpelli C, Artiguenave F, Pè ME, Valle G, Morgante M, Caboche M, Adam-Blondon A-F, Weissenbach J, Quétier F, Wincker P: **The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla**. *Nature* 2007, **449**:463–467.

192. Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, Fitzgerald LM, Vezzulli S, Reid J, Malacarne G, Iliev D, Coppola G, Wardell B, Micheletti D, Macalma T, Facci M, Mitchell JT, Perazzolli M, Eldredge G, Gatto P, Oyzerski R, Moretto M, Gutin N, Stefanini M, Chen Y, Segala C, Davenport C, Demattè L, Mraz A, Battilana J, Stormo K, Costa F, Tao Q, Si-Ammour A, Harkins T, Lackey A, Perbost C, Taillon B, Stella A, Solovyev V, Fawcett JA, Sterck L, Vandepoele K, Grando SM, Toppo S, Moser C, Lanch-

bury J, Bogden R, Skolnick M, Sgaramella V, Bhatnagar SK, Fontana P, Gutin A, Van de Peer Y, Salamini F, Viola R: **A high quality draft consensus sequence of the genome of a heterozygous grapevine variety**. *PLoS ONE* 2007, **2**:e1326.

193. Zhu X-Y, Chase MW, Qiu Y-L, Kong H-Z, Dilcher DL, Li J-H, Chen Z-D: **Mitochondrial matR sequences help to resolve deep phylogenetic relationships in rosids**. *BMC Evol Biol* 2007, **7**:217.

194. Bausher MG, Singh ND, Lee S-B, Jansen RK, Daniell H: **The complete chloroplast genome sequence of Citrus sinensis (L.) Osbeck var "Ridge Pineapple": organization and phylogenetic relationships to other angiosperms**. *BMC Plant Biol* 2006, **6**:21.

195. Cantino PD, Doyle JA, Graham SW, Judd WS, Olmstead RG, Soltis DE, Soltis PS, Donoghue MJ: **Towards a phylogenetic nomenclature of Tracheophyta**. *Taxon* 2007, **56**:1E–44E.

196. Angiuoli SV, Salzberg SL: **Mugsy: fast multiple alignment of closely related whole genomes**. *Bioinformatics* 2011, **27**:334–342.

197. Zhang W-J, Zhou J, Li Z-F, Wang L, Gu X, Zhong Y: **Comparative Analysis of Codon Usage Patterns Among Mitochondrion, Chloroplast and Nuclear Genes in Triticum aestivum L.** *J Integr Plant Biol* 2007, **49**:246–254.

198. Van de Peer Y: **A mystery unveiled**. *Genome Biol* 2011, **12**:113.

199. Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH: **Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps**. *Genome Res* 2008, **18**:1944–1954.

200. Wang H, Moore MJ, Soltis PS, Bell CD, Brockington SF, Alexandre R, Davis CC, Latvis M, Manchester SR, Soltis DE: **Rosid radiation and the rapid rise of angiosperm-dominated forests**. *Proc Natl Acad Sci U S A* 2009, **106**:3853–3858.

201. Velasco L, Fernández-Martínez JM: **Other Brassicas**. In *Oil Crops*. edited by Vollmann J, Rajcan I Springer New York; 2010, **4**:127–153.

202. Felsenstein J: **Confidence limits on phylogenies: An approach using the bootstrap.** *Evolution* 1985, **39**:783–791.

203. Tang ZZ, Sharma S, Zheng S, Chawla G, Nikolic J, Black DL: **Regulation of the mutually exclusive exons 8a and 8 in the CaV1.2 calcium channel transcript by polypyrimidine tract-binding protein**. *J. Biol. Chem.* 2011, **286**:10007–10016.

204. Warf MB, Berglund JA: **Role of RNA structure in regulating pre-mRNA splicing**. *Trends Biochem. Sci.* 2010, **35**:169–178.

205. Richard H, Schulz MH, Sultan M, Nürnberger A, Schrinner S, Balzereit D, Dagand E, Rasche A, Lehrach H, Vingron M, Haas SA, Yaspo M-L: **Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments**. *Nucleic Acids Res.* 2010, **38**:e112.

206. Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, Baren MJ van, Boley N, Booth BW, Brown JB, Cherbas L, Davis CA, Dobin A, Li R, Lin W, Malone JH, Mattiuzzo NR, Miller D, Sturgill D, Tuch BB, Zaleski C, Zhang D, Blanchette M,

Dudoit S, Eads B, Green RE, Hammonds A, Jiang L, Kapranov P, Langton L, Perrimon N, Sandler JE, Wan KH, Willingham A, Zhang Y, Zou Y, Andrews J, Bickel PJ, Brenner SE, Brent MR, Cherbas P, Gingeras TR, Hoskins RA, Kaufman TC, Oliver B, Celniker SE: **The developmental transcriptome of Drosophila melanogaster**. *Nature* 2010, **471**:473–479.

207. Castellana N, Bafna V: **Proteogenomics to discover the full coding content of genomes: a computational perspective**. *J. Proteomics* 2010, **73**:2124–2135.

208. Xia J, Caragea D, Brown SJ: **Prediction of alternatively spliced exons using support vector machines**. *Int. J. Data Min. Bioinform.* 2010, **4**:411–430.

209. Sinha R, Nikolajewa S, Szafranski K, Hiller M, Jahn N, Huse K, Platzer M, Backofen R: **Accurate prediction of NAGNAG alternative splicing**. *Nucleic Acids Res.* 2009, **37**:3569–3579.

210. Vukusic I, Grellscheid SN, Wiehe T: **Applying genetic programming to the prediction of alternative mRNA splice variants**. *Genomics* 2007, **89**:471–479.

211. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B: **AUGUSTUS: ab initio prediction of alternative transcripts**. *Nucleic Acids Res.* 2006, **34**:W435–439.

212. Ceccarelli M, Maratea A: **Virtual genetic coding and time series analysis for alternative splicing prediction in C. elegans**. *Artif. Intell. Med.* 2009, **45**:109–115.

213. Meijers R, Puettmann-Holgado R, Skiniotis G, Liu J, Walz T, Wang J, Schmucker D: **Structural basis of Dscam isoform specificity**. *Nature* 2007, **449**:487–491.

214. Graveley BR: **Mutually exclusive splicing of the insect Dscam pre-mRNA directed by competing intronic RNA secondary structures**. *Cell* 2005, **123**:65 – 73.

215. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity**. *BMC Bioinformatics* 2004, **5**:113.

216. Edgar RC: **MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput**. *Nucleic Acids Res.* 2004, **32**:1792–1797.

217. Jones DT: **Protein secondary structure prediction based on position-specific scoring matrices**. *J. Mol. Biol.* 1999, **292**:195–202.

218. Yang Y, Zhan L, Zhang W, Sun F, Wang W, Tian N, Bi J, Wang H, Shi D, Jiang Y, Zhang Y, Jin Y: **RNA secondary structure in mutually exclusive splicing**. *Nat Struct Mol Biol* 2011, **18**:159 – 168.

219. Lorenz R, Bernhart SH, Siederdissen CH zu, Tafer H, Flamm C, Stadler PF, Hofacker IL: **ViennaRNA Package 2.0**. *Algorithms Mol. Biol.* 2011, **6**:26.

220. Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF: **RNAalifold: improved consensus structure prediction for RNA alignments**. *BMC Bioinformatics* 2008, **9**:474.

221. Clark AG, Eisen MB, Smith DR, et al.: **Evolution of genes and genomes on the Drosophila phylogeny**. *Nature* 2007, **450**:203–218.

222. English AC, Patel KS, Loraine AE: **Prevalence of alternative splicing choices in Arabidopsis thaliana**. *BMC Plant Biol.* 2010, **10**:102.

223. Zhang Z, Zhu X, Stevens LM, Stein D: **Distinct functional specificities are associated with protein isoforms encoded by the Drosophila dorsal-ventral patterning gene pipe**. *Development* 2009, **136**:2779–2789.

224. Wilkin MB, Becker MN, Mulvey D, Phan I, Chao A, Cooper K, Chung HJ, Campbell ID, Baron M, MacIntyre R: **Drosophila dumpy is a gigantic extracellular protein required to maintain tension at epidermal-cuticle attachment sites**. *Curr. Biol.* 2000, **10**:559–567.

225. Stanke M, Waack S: **Gene prediction with a hidden Markov model and a new intron submodel**. *Bioinformatics* 2003, **19 Suppl 2**:ii215–225.

226. Rhee SY, Wood V, Dolinski K, Draghici S: **Use and misuse of the gene ontology annotations**. *Nat. Rev. Genet.* 2008, **9**:509–515.

227. McCandless D: *Information is Beautiful.* (Reissue). Collins; 2010.

228. Yook K, Harris TW, Bieri T, Cabunoc A, Chan J, Chen WJ, Davis P, De la Cruz N, Duong A, Fang R, Ganesan U, Grove C, Howe K, Kadam S, Kishore R, Lee R, Li Y, Muller H-M, Nakamura C, Nash B, Ozersky P, Paulini M, Raciti D, Rangarajan A, Schindelman G, Shi X, Schwarz EM, Ann Tuli M, Van Auken K, Wang D, Wang X, Williams G, Hodgkin J, Berriman M, Durbin R, Kersey P, Spieth J, Stein L, Sternberg PW: **WormBase 2012: more genomes, more data, new website**. *Nucleic Acids Res.* 2012, **40**:D735–741.

229. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, Karthikeyan AS, Lee CH, Nelson WD, Ploetz L, Singh S, Wensel A, Huala E: **The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools**. *Nucleic Acids Res.* 2012, **40**:D1202–1210.

230. Döring A, Weese D, Rausch T, Reinert K: **SeqAn An efficient, generic C++ library for sequence analysis**. *BMC Bioinformatics* 2008, **9**:11.

231. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S: **AmiGO: online access to ontology and annotation data**. *Bioinformatics* 2009, **25**:288–289.

232. Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ, Frey BJ: **Deciphering the splicing code**. *Nature* 2010, **465**:53–59.

233. Rogers MF, Thomas J, Reddy AS, Ben-Hur A: **SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data**. *Genome Biol* 2012, **13**:R4.

Acknowledgements

First of all I would like to thank my supervisor PD Dr. Martin Kollmar for trusting in me, for providing a lot of inspiration and for the critical reflection of my work.

I thank Prof. Christian Griesinger for providing such a great research environment.

I want to thank Prof. Burkhard Morgenstern and Prof. Bert de Groot for fruitful discussions and hints in the thesis committee meetings.

I thank all people of the motor proteins group for inspiring suggestions and discussions. Especially, I want to thank Björn Hammesfahr and Peggy Findeisen, it was a pleasure to work with them.

I would like to thank Gesa for her interest in my work and helpful ideas.

Appendix

A1 Supplementary information

Candidates for mutually exclusive spliced exons in dependency of the prediction criteria

To explore the parameters for predicting mutually exclusive exons (MXEs) we analysed all annotated exons in clusters of MXEs in the *Drosophila melanogaster* genome (*Dm*, Flybase release 5.36). To exclude that the determined characteristics are *Drosophila* specific we also analysed the annotated mutually exclusive exomes of *Homo sapiens* (*Hs*, NCBI release 37.3), *Caenorhabditis elegans* (*Ce*, WormBase release WS230), and *Arabidopsis thaliana* (*At*, TAIR release 167). These species have been chosen because of their widespread taxonomic distribution and their advanced and detailed annotations.

Maximal length difference of annotated internal mutually exclusive exons

To determine a suitable cut-off for the length difference in the search we analysed all internal clusters of annotated MXEs (Fig. 1). For all species analysed the curves look very similar. 64%, 20%, 48% and 0% of the annotated MXEs of *Dm*, *Hs*, *Ce*, and *At*, respectively, have no length difference (86%, 71%, 57% and 43% have length difference of less than five residues). A cut-off for the length difference of 20 residues should be appropriate to reconstruct almost all annotated cases and to not include too many mispredictions (95%, 82%, 77% and 100% have length difference of less than 20 residues).

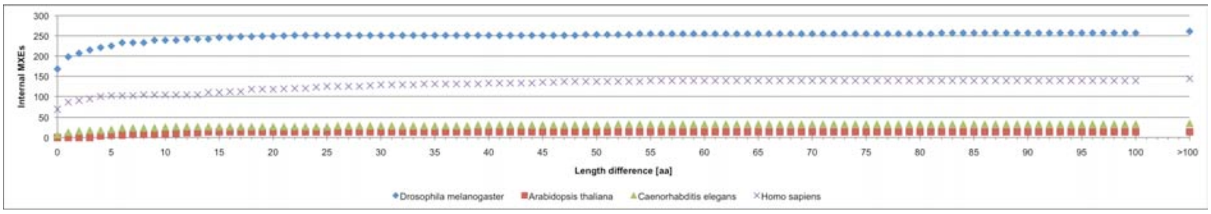


Fig. 1: Number of annotated internal MXEs as function of the minimal length difference to another MXE of the same cluster.

Sequence similarity of annotated internal mutually exclusive exons

In this project, we were supposing that the MXEs of a cluster code for identical secondary structural elements of the protein like in the *Dm* muscle myosin heavy chain. If this conditions holds true the MXEs should show a certain degree of sequence similarity. Analysis of the MXEs of *Dm* shows that 94.9% of the MXEs, which show any sequence similarity, have a sequence similarity of more than 15% (Fig. 2). In *Hs* and *Ce*, 98% and 86% of the MXEs, which show any sequence similarity, have higher sequence similarities than 15%. Therefore, we decided to use 15% sequence similarity as cut-off for further predictions.

However, a few cases of annotated MXEs do not show any sequence similarity and can not be reconstructed with our method (see difference of the two rightmost numbers).

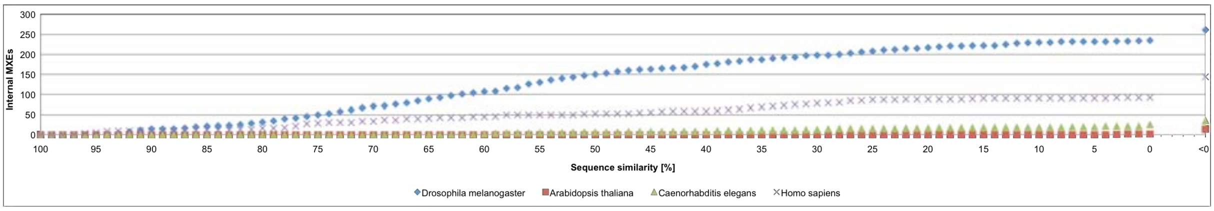


Fig. 2: Number of annotated internal MXEs as function of the sequence similarity to another MXE of the same cluster. In the case of similarity, two slightly different similarity scores can be calculated for a pair of MXEs dependent of which has been used as reference. Here, we included the respective higher scores.

Minimal exon length of annotated internal mutually exclusive exons

The shorter the exons are the more probable it becomes that their sequences are featureless and that false positive candidates will be predicted. Therefore, we introduced a parameter "minimal exon length". Based on the analysis of all annotated MXEs we set this parameter to 15 residues (Fig. 3).

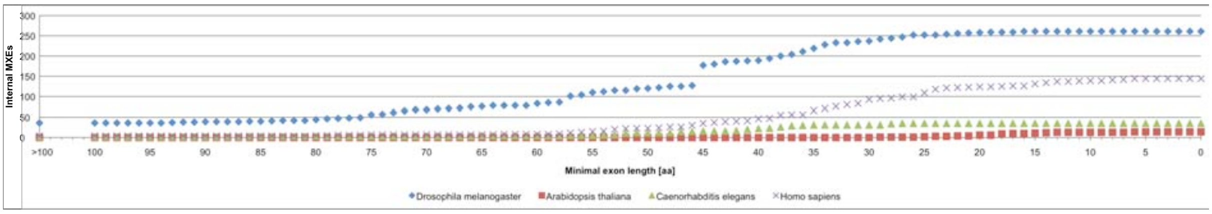


Fig. 3: Number of annotated internal MXEs as function of the respective length of the MXE. The two noticeable jumps in the scatter plot of the *Dm* MXEs are due to the MXEs in the large clusters of the DSCAM gene.

Reconstructed and predicted internal mutually exclusive exons at a similarity score cut-off of 15%

Apart from the MXEs that we cannot reconstruct because they are out of the scope of our preconditions (no sequence similarity, huge length difference), we assessed the sensitivity of our method when using a length difference of 20 residues and a similarity score of 15% as standard cut-offs. Given a similarity score of at least 15%, the analysis of the reconstructed MXEs shows that all annotated MXEs have length differences of less than 20 residues (Figs. 4 and 5). A similar distribution is found for the length difference of the internal MXEs that we predict newly (Figs. 6 and 7).

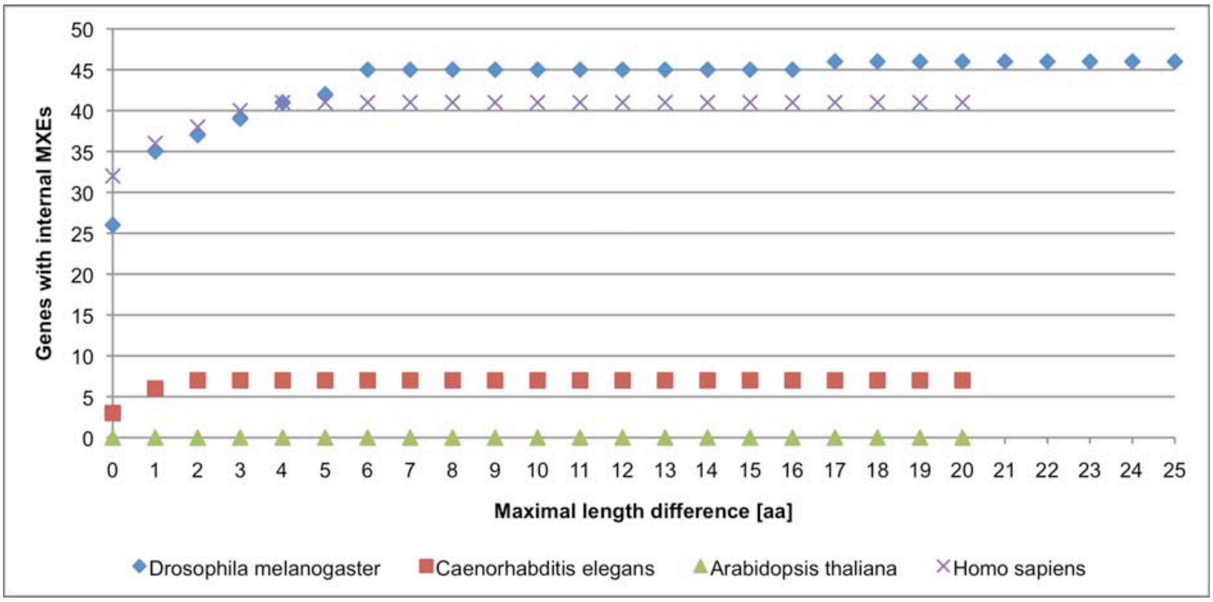


Fig. 4: Number of genes containing annotated internal MXEs that could be reconstructed at a given length difference cut-off having a similarity score of at least 15%.

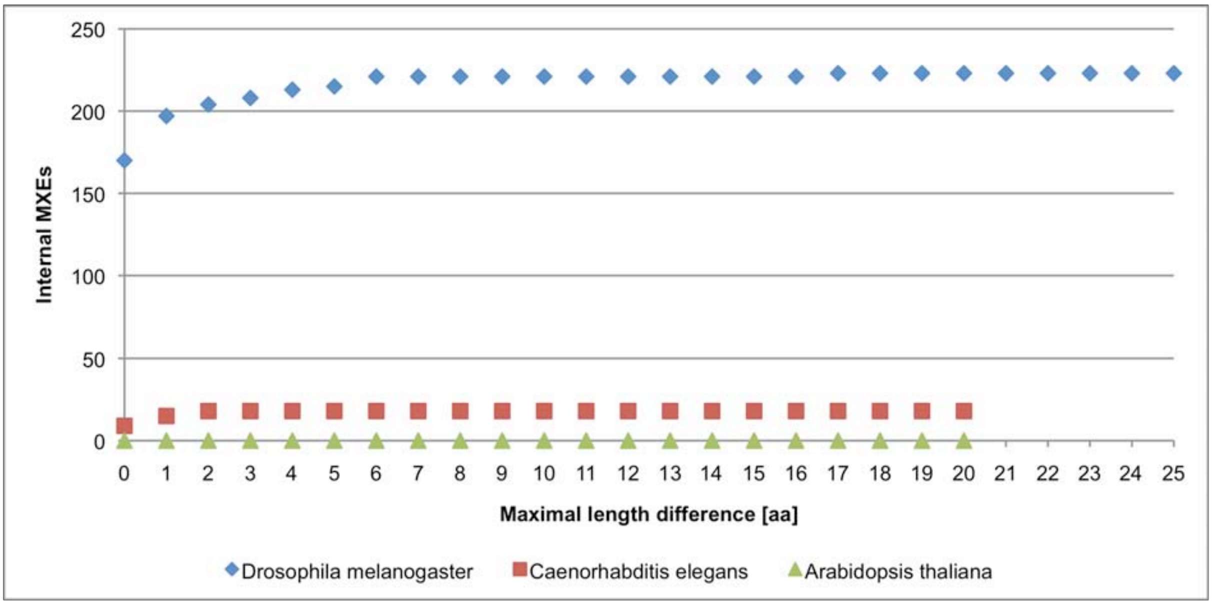


Fig. 5: Number of annotated internal MXEs that could be reconstructed at a given length difference cut-off having a similarity score of at least 15%.

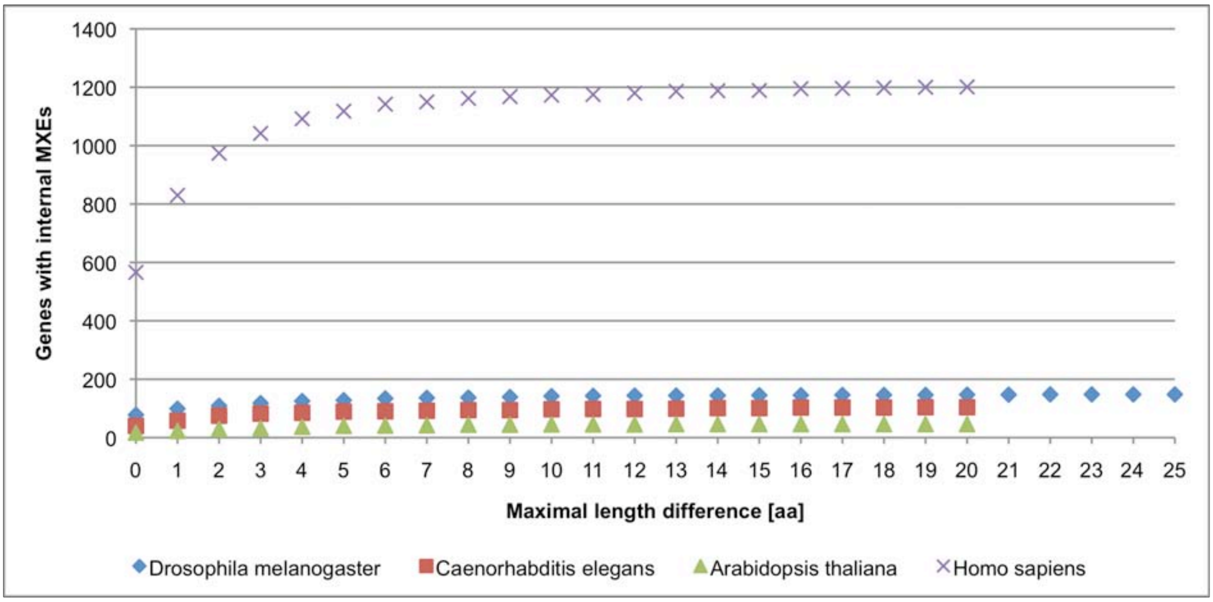


Fig. 6: Number of genes containing predicted internal MXEs (including annotated MXEs that could be reconstructed) with a similarity score of at least 15% at a given length difference.

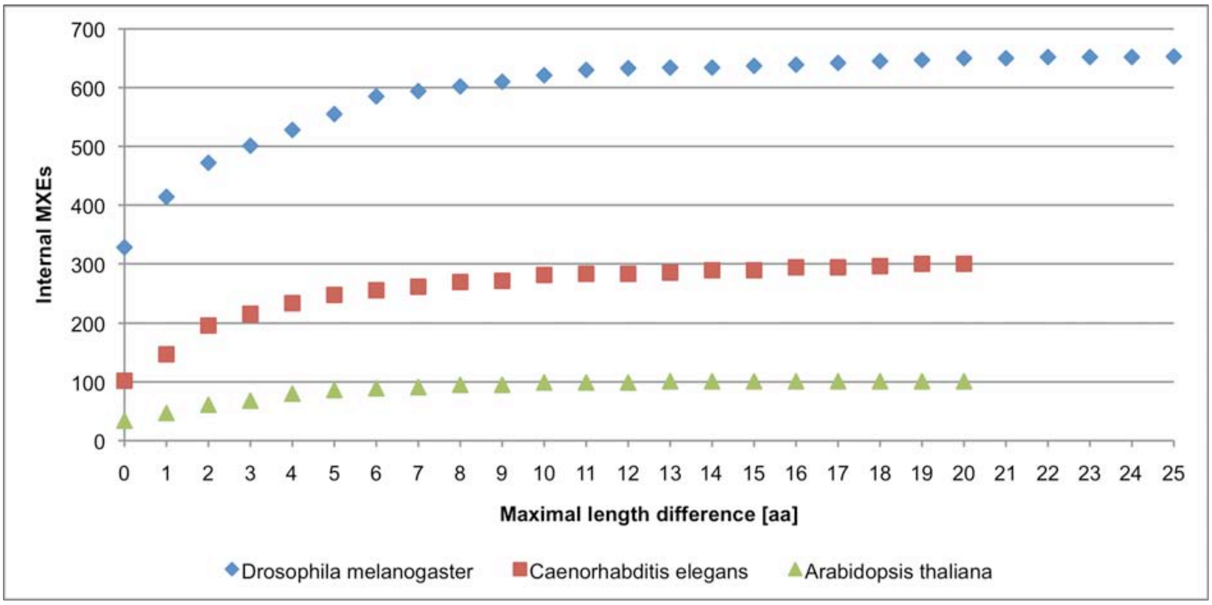


Fig. 7: Number of predicted internal MXE candidates (including annotated MXEs that could be reconstructed) with a similarity score of at least 15% at a given length difference.

Reconstructed and predicted internal mutually exclusive exons at a length difference cut-off of 20 residues

To assess the suitability of the sequence similarity cut-off of 15% within the preconditions of our prediction method, we analysed the distribution of the annotated exons with a length difference of less than 20 residues (Figs. 8 and 9). In contrast to the MXEs of *Hs* and *Ce*, the MXEs of *Dm* do not show a pronounced plateau. The number of predicted MXE candidates even shows an exponential increase below a similarity score of 10% (*Dm*) and 15% (*Hs*), respectively (Figs. 10 and 11).

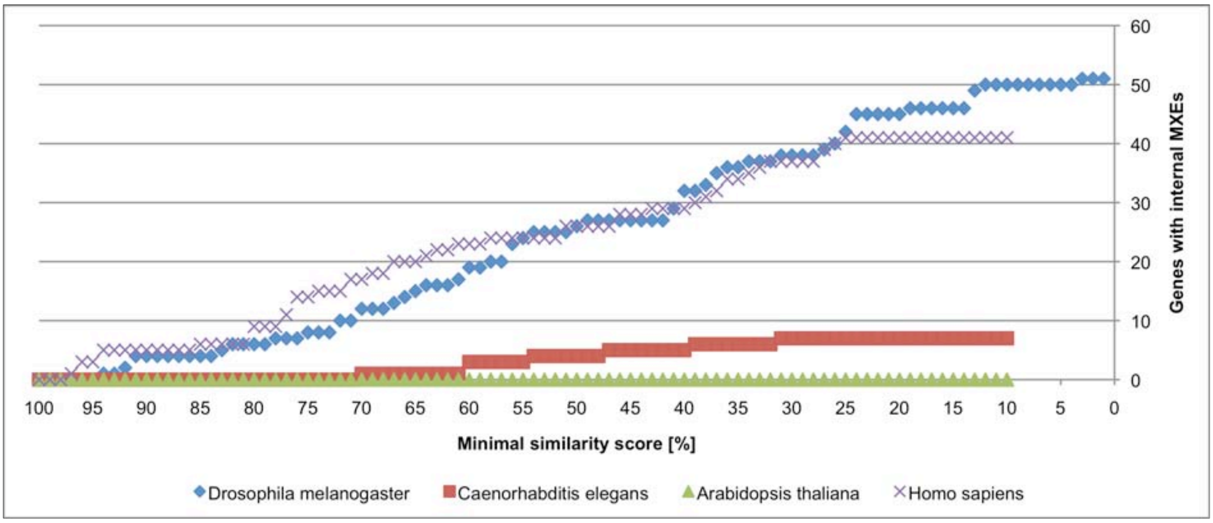


Fig. 8: Number of genes containing annotated internal MXEs that could be reconstructed at a given sequence similarity score cut-off and having a length difference of less than 20 aa.

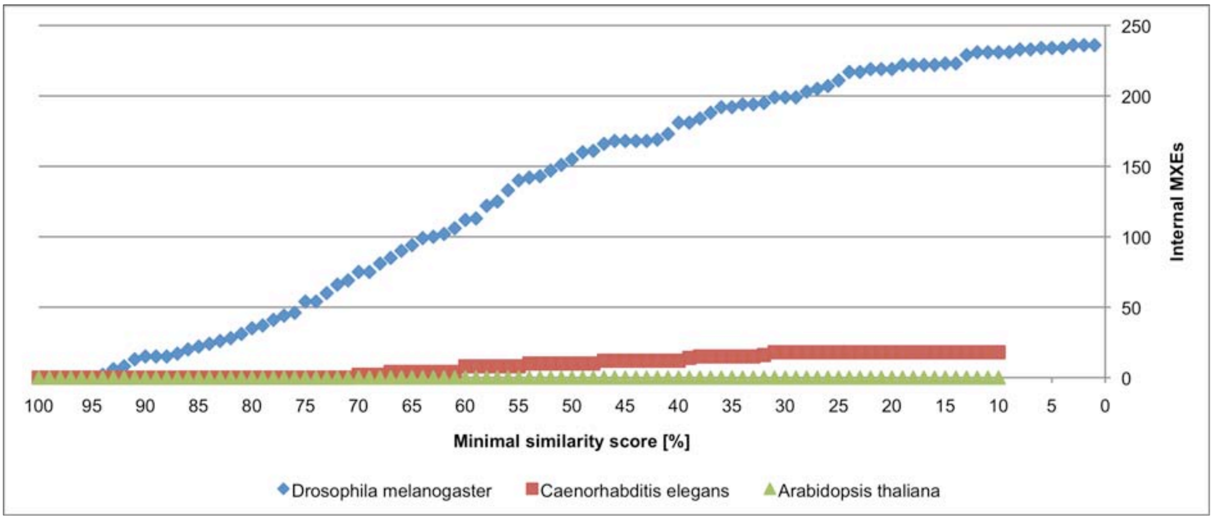


Fig. 9: Number of internal MXEs that could be reconstructed at a given sequence similarity score cut-off and having a length difference of less than 20 aa.

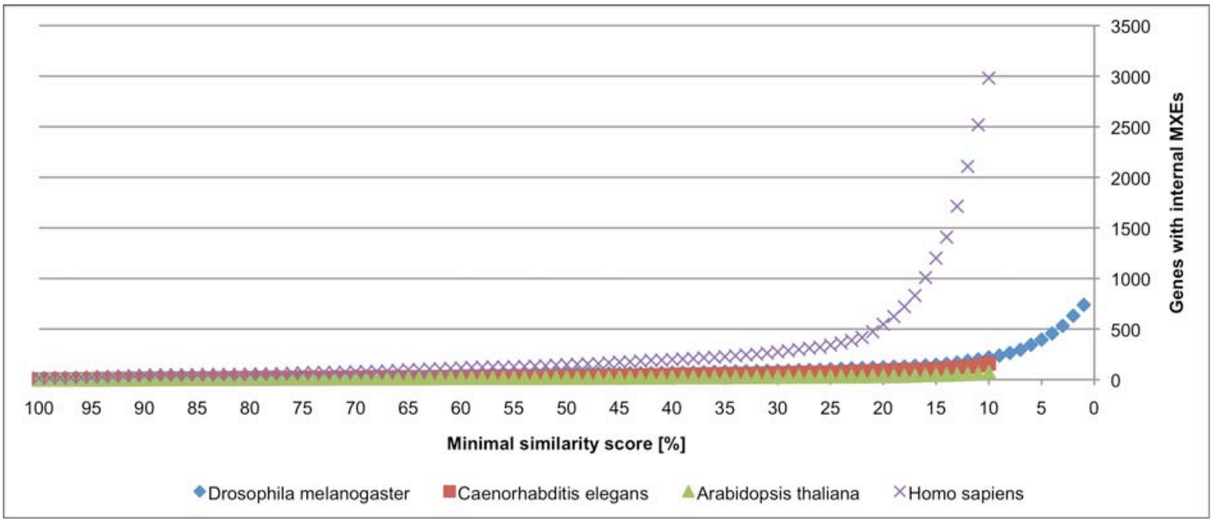


Fig. 10: Number of genes containing internal MXE candidates (including annotated MXEs that could be reconstructed) predicted at a given sequence similarity score cut-off and having a length difference of less than 20 aa.

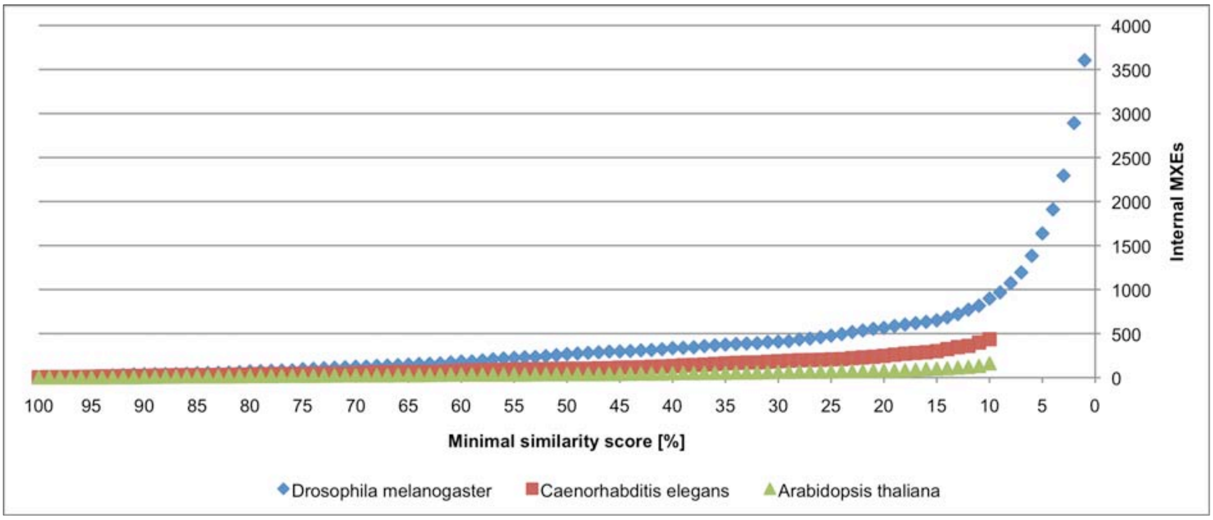


Fig. 11: Number of internal MXE candidates (including annotated MXEs that could be reconstructed) predicted at a given sequence similarity score cut-off and having a length difference of less than 20 aa.

Reconstructed and predicted internal mutually exclusive exons in dependence of a minimal original exon length

The sequences of very short exons do not contain enough complexity to exclude the identification of "similar" exon, especially if they are surrounded by long introns. Luckily, short exons within genes are rather rare and are predominantly found at gene borders. In order to avoid the inclusion of many false positives we introduced the parameter "minimal original exon length". Annotated MXEs, which we can reconstruct with a length difference cut-off of 20 residues and a similarity score cut-off of 15%, are all longer than ten residues (Figs. 12 and 13). For the initial search for MXE candidates in *Drosophila* we set this parameter to one residue (Figs 14 and 15). However, only a few candidates were found for exons shorter than 15 residues. Therefore, we set the minimal original exon length parameter to 15 residues for the analysis of the *Drosophila* genome and for the search for MXE candidates in the other model organisms (Figs. 14 and 15). The value seems appropriate for *Caenorhabditis* and *Ara-*

bidopsis while the number of MXE candidates is increasing exponentially in dependence of the search exon length in human. This is most probably due to the much longer introns in human compared to the other species analysed.

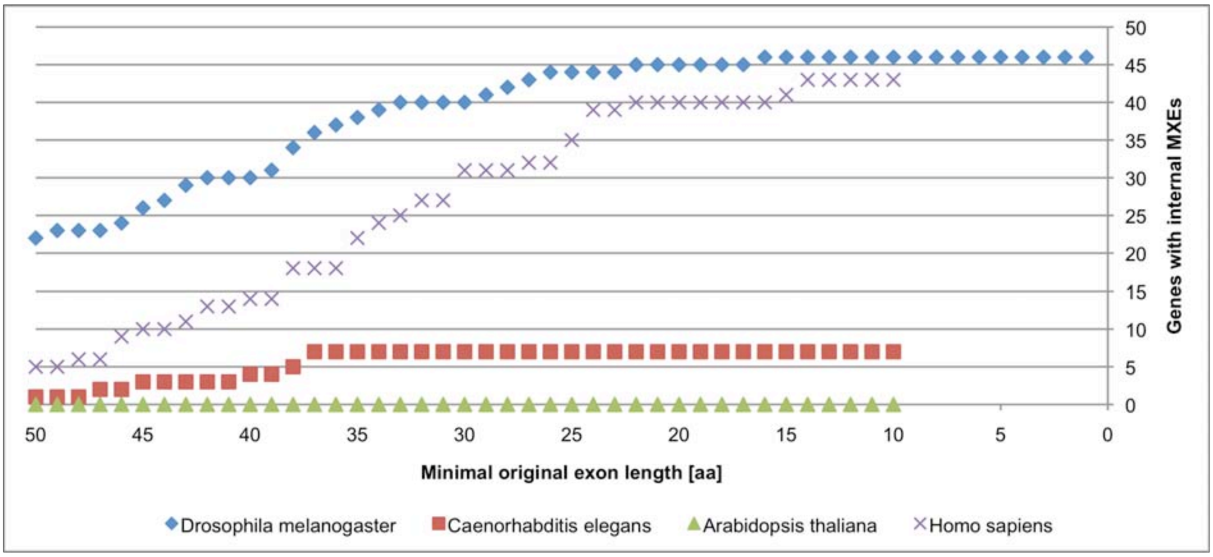


Fig. 12: Number of genes containing annotated internal MXEs in dependency of the length of the MXEs that could be reconstructed at a sequence similarity score cut-off of 15% and a length difference of less than 20 aa.

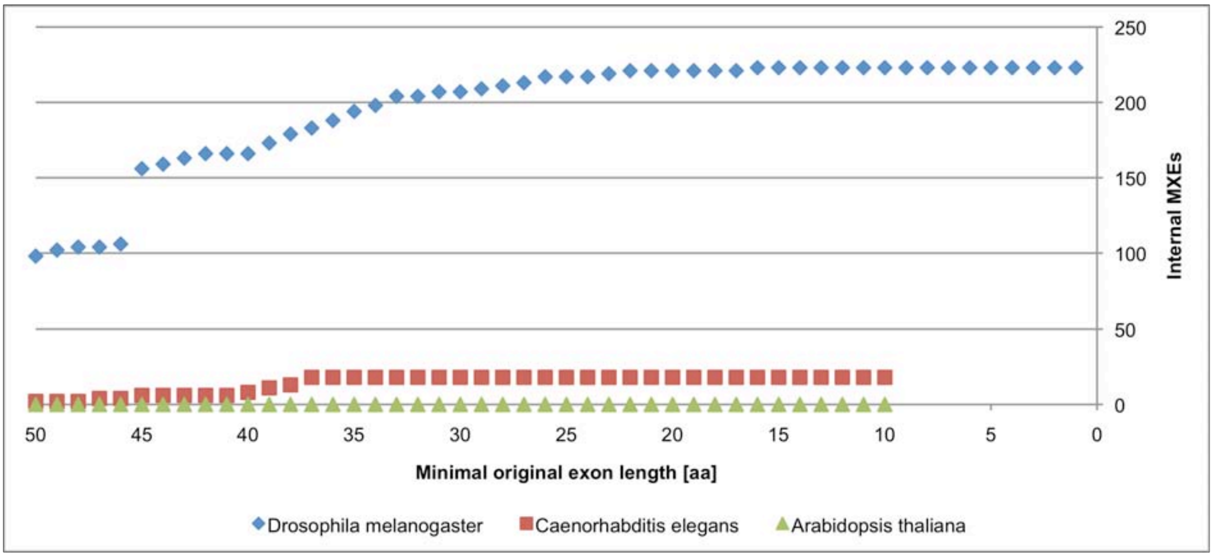


Fig. 13: Number of annotated internal MXEs in dependency of the length of the MXEs that could be reconstructed at a sequence similarity score cut-off of 15% and a length difference of less than 20 aa.

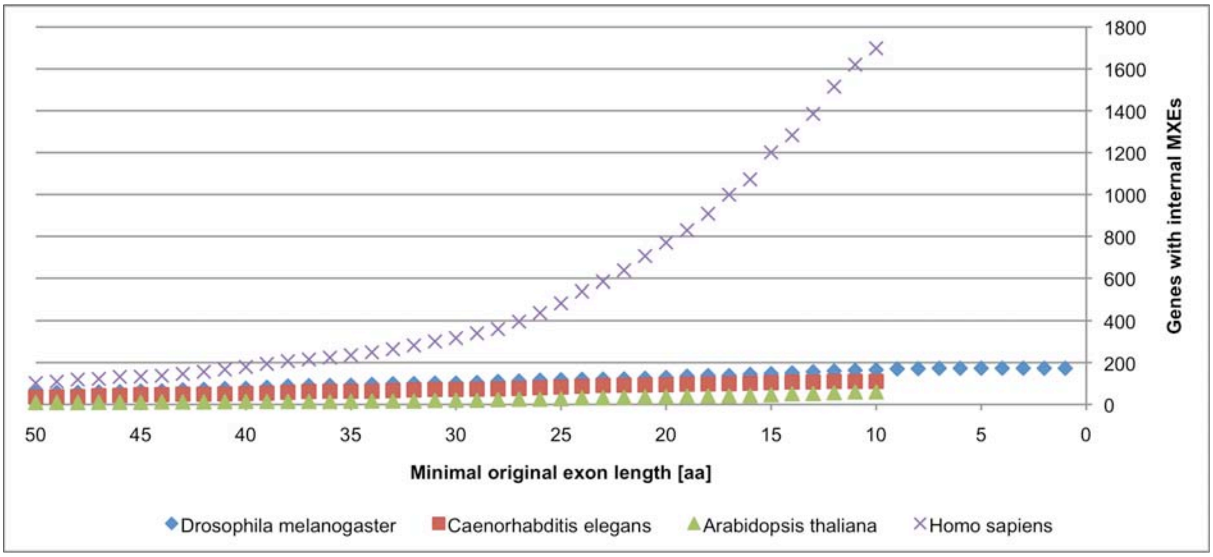


Fig. 14: Number of genes containing internal MXE candidatess in dependency of the length of the MXEs that were predicted at a sequence similarity score cut-off of 15% and a length difference of less than 20 aa.

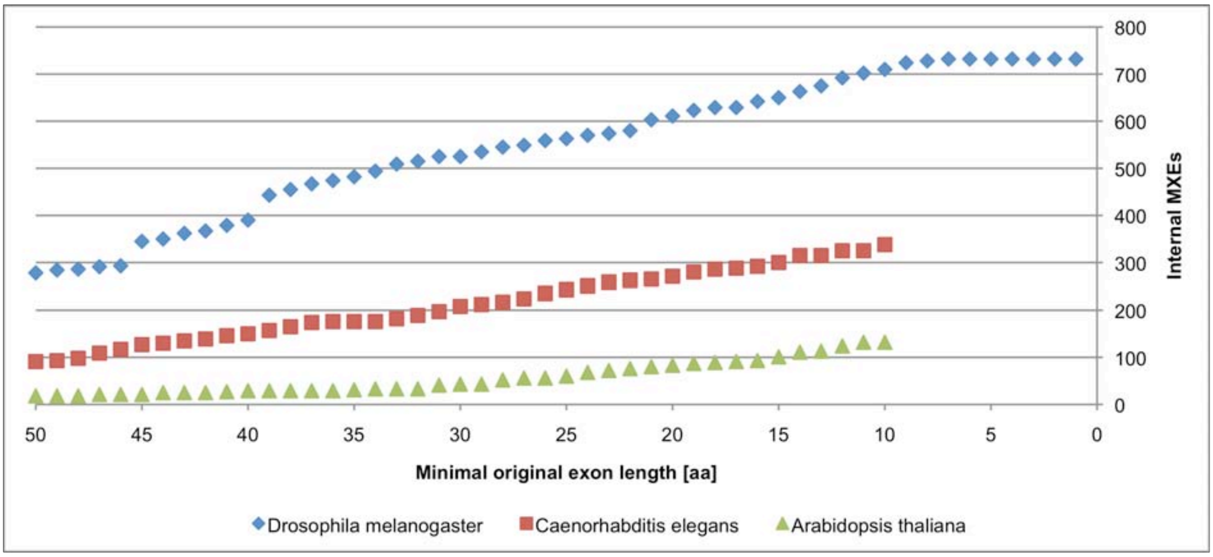


Fig. 15: Number of internal MXE candidates in dependency of the length of the MXEs that were predicted at a sequence similarity score cut-off of 15% and a length difference of less than 20 aa.

Statistics

In order to assess potential systematic features in mutually exclusive exons, and to identify potential outliers within the predicted MXE candidates we analysed all annotated and predicted exons of *Drosophila* with respect to exon and intron length and splice site patterns.

Exon lengths

The exon lengths of the annotated and predicted MXEs show almost the same distribution like all exons of *Drosophila* with a broad peak around 140 residues (Fig. 16). Interestingly, there is a second smaller peak for the length of MXEs at 300 amino acids. The comparison of the annotated MXEs to the predicted MXE candidates shows similar distributions meaning that the predictions represent normal MXEs. The internal MXEs that are annotated and that we cannot reconstruct also display a similar distribution but in addition tend to represent larger exons as compared to the other sets. Surprisingly, the constitutive exons sharing our criteria for MXEs show three striking peaks at 80, 320 and 340 residues but show a local minimum at 140 residues. This supports the notion that the predicted MXEs rather represent MXEs than potential constitutively spliced exons.

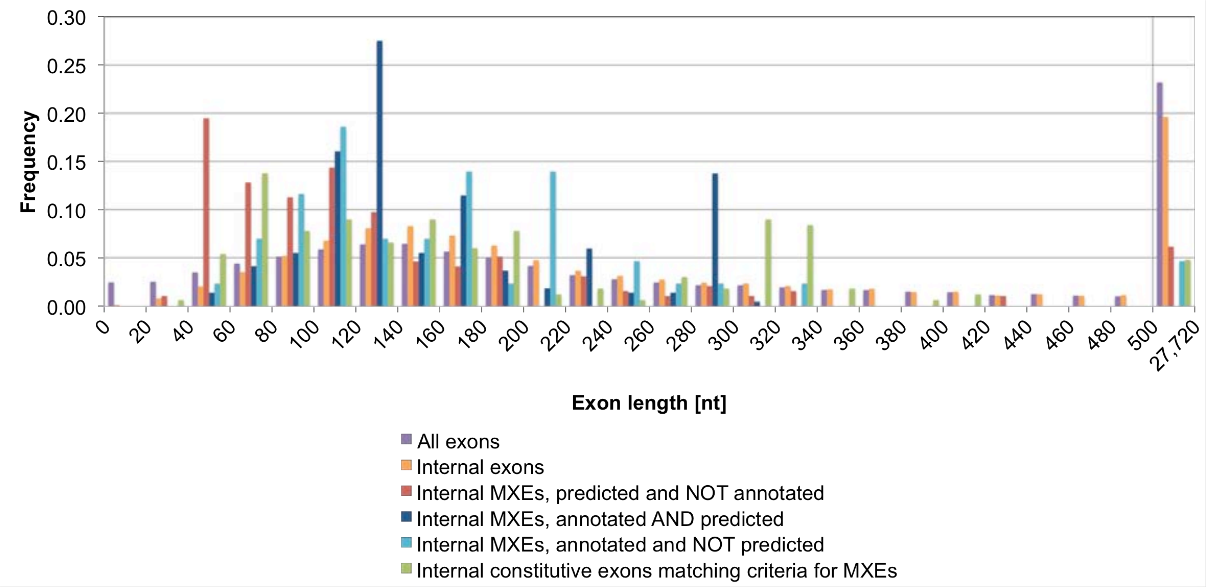


Fig. 16: Comparison of exon lengths. Various subsets of annotated and predicted MXEs are compared to all exons and internal constitutive exons sharing our criteria for MXEs.

Intron lengths

Comparison of the intron lengths also shows a broad distribution with a tendency to rather short introns (< 300 bp; Fig. 17).

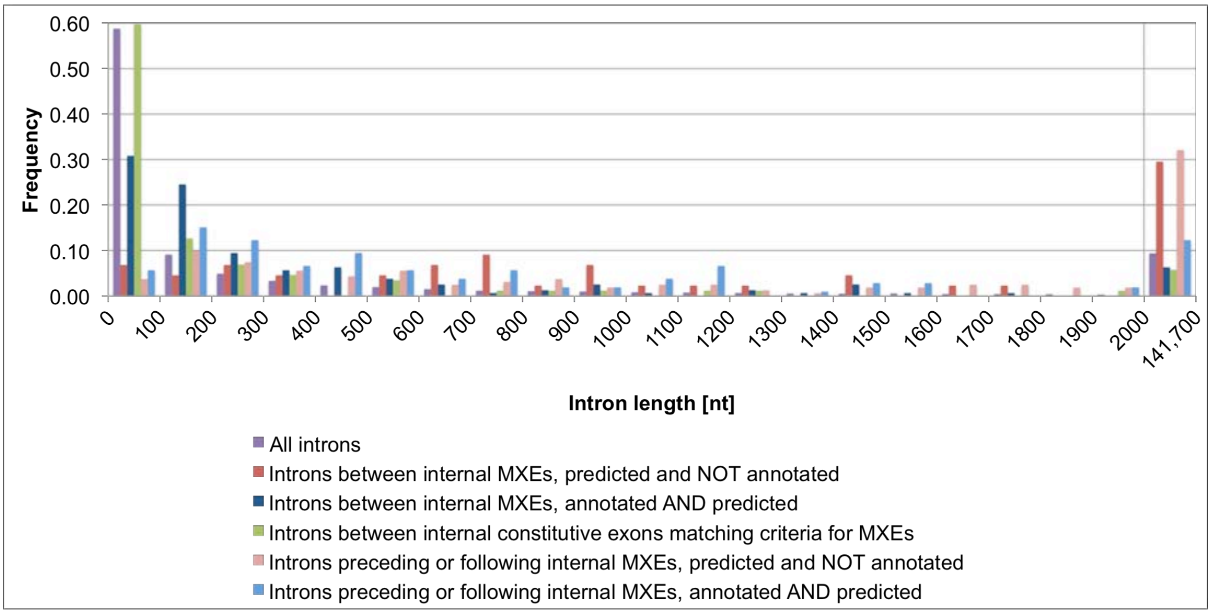


Fig. 17: Comparison of intron lengths. Introns next to various subsets of annotated and predicted MXEs are compared to all introns and introns next to internal constitutive exons sharing our criteria for MXEs.

Exon lengths of initial and terminal exons in multi-exon genes

Because the algorithm is based on protein coding sequence it could be possible that the initial and terminal exons of the coding region are not the initial and terminal exons of the transcripts. In this case, these exons would be regarded as internal exons. Therefore, we also analysed candidate exons of initial and terminal exons that share the criteria of MXEs. In general, initial and terminal exons of multi-exon genes are considerably shorter than internal exons (Figs. 18 and 19). Some of these match the criteria of MXEs. Of those, almost all code for at least 40 residues. In these cases it is unlikely that pseudo-duplicates of low-complexity exons were found.

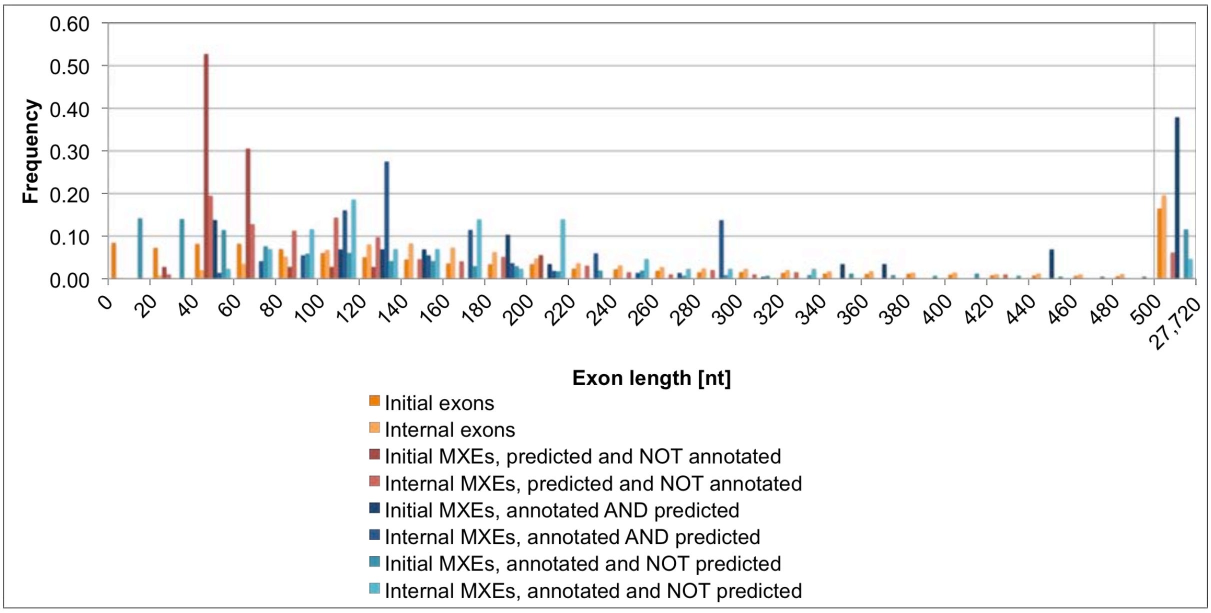


Fig. 18: Comparison of exon lengths of initial exons of multi-exon genes. Various subsets of annotated and predicted initial exons matching the criteria for MXEs are compared to all exons and internal MXEs.

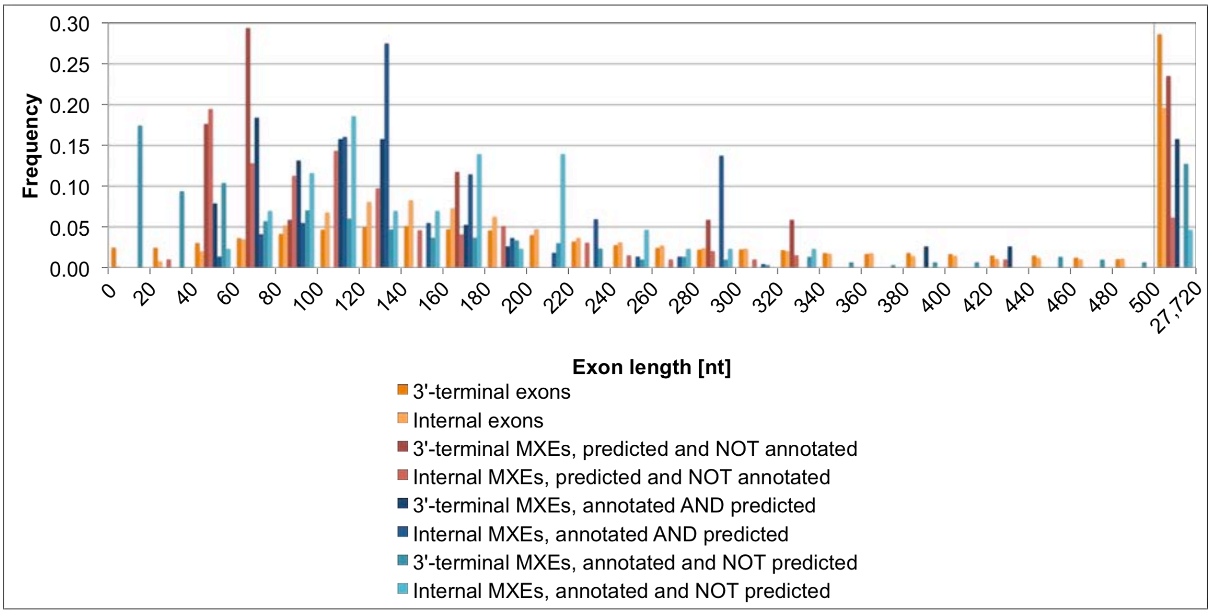


Fig. 19: Comparison of exon lengths of terminal exons of multi-exon genes. Various subsets of annotated and predicted terminal exons matching the criteria for MXEs are compared to all exons and internal MXEs.

GC content

The GC content of all exons shows a broad distribution around 55% (Fig. 20). The MXEs, which we cannot reconstruct, and the constitutive exons sharing our criteria of MXEs have a broader GC content distribution with a remarkably higher percentage of exons with GC contents of 60 to 75%. The distribution of the GC content of the predicted MXEs is similar to the distribution of the annotated MXEs except for a slight increase of exons with GC contents of 40 to 45%.

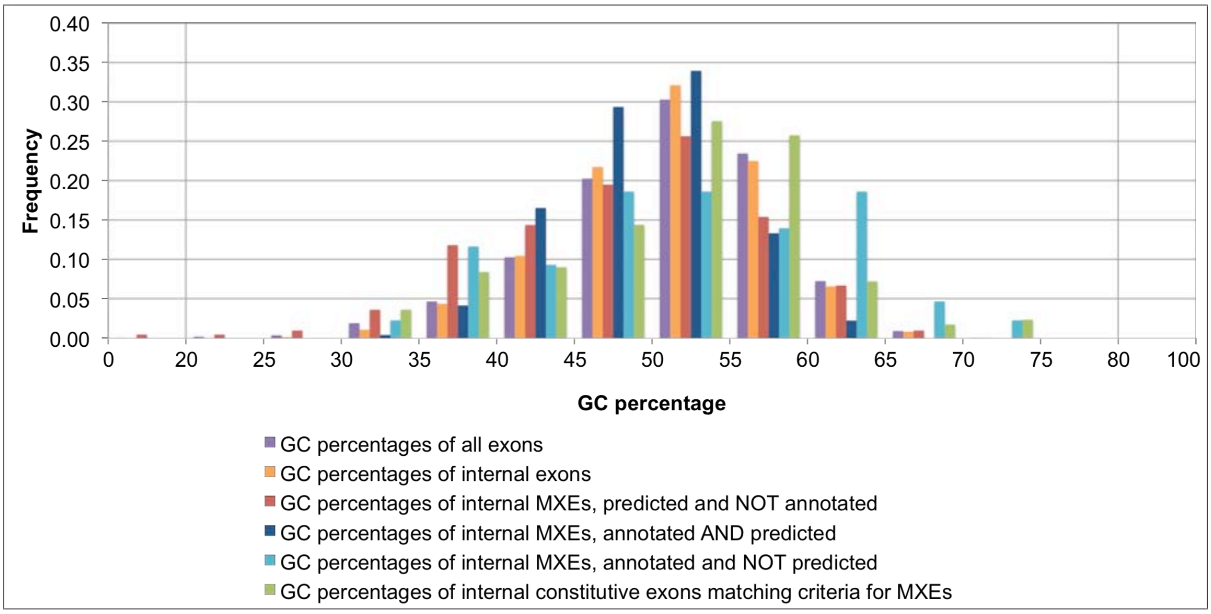


Fig. 20: Comparison of GC content of exons. The GC content of all exons (reference) is compared to the GC content of annotated and predicted internal MXEs and to internal constitutive exons sharing our criteria for MXEs.

Protein translation

To assess whether MXEs are predominantly found in proteins of a certain size, we analysed the lengths of the translations (Fig. 21). Here, from each alternatively spliced gene (independently of alternative splicing type) only one transcript and the corresponding translation were considered. Proteins built with MXEs are relatively longer than the average proteins. The distribution of the proteins with annotated MXEs and with predicted MXE candidates is very similar.

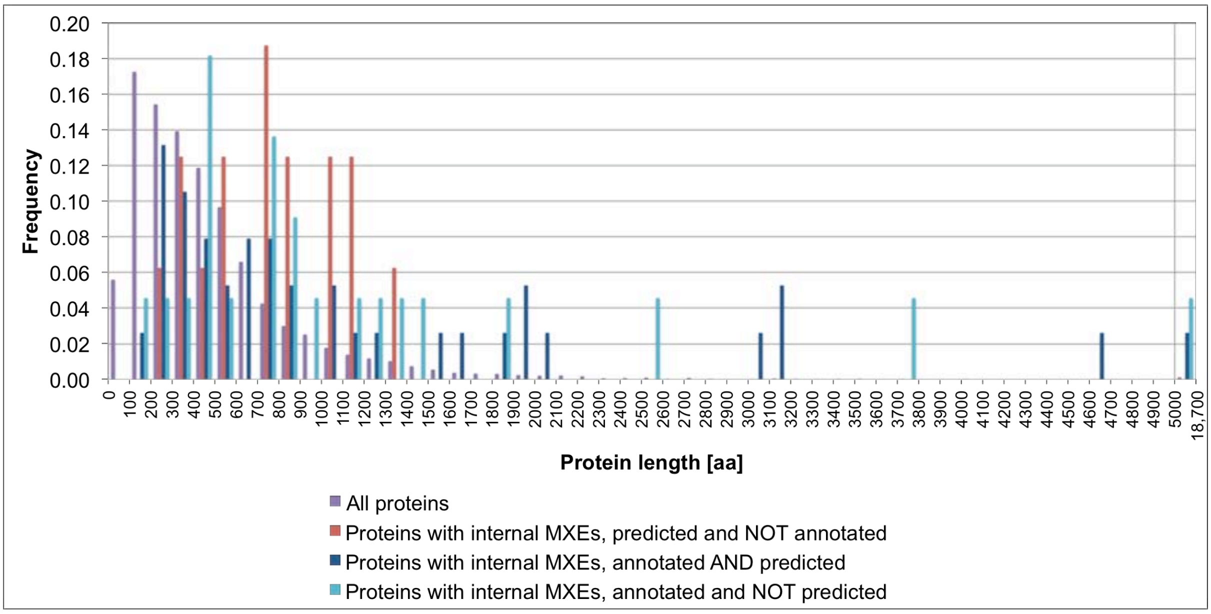


Fig. 21: Comparison of the lengths of the translations of one isoform per gene. For the reconstruction of the translations of the genes containing MXEs only one isoform has been chosen and only one exon of each cluster. For the protein lengths of all proteins, only the isoforms “A” were considered.

Codon usage

The codon usage of the MXEs (annotated and predicted) is very similar to the codon usage of all or all internal exons except for the codons AAG, AGC CAG and CTG that are slightly less represented in MXEs. Strikingly, the percentage of cysteine-coding codons (TGT and TGC) is five times higher in constitutive exons sharing our criteria of MXEs compared to all exons, and the MXEs, that are annotated in FlyBase but that we cannot reconstruct, have a considerably higher content of alanines (GCC codon) and glutamines (CAA and CAG codons).

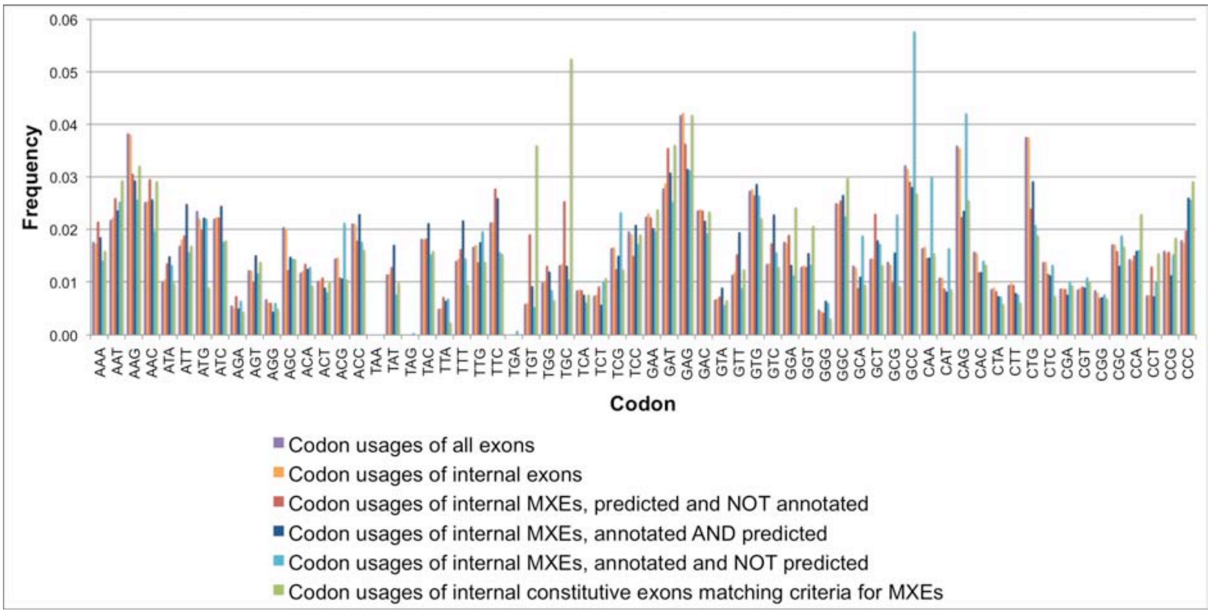


Fig. 22: Comparison of the codon usage. Codon usage in all exons is compared to that of genes containing annotated or predicted MXEs and to that of internal constitutive exons sharing our criteria for MXEs.

Start/end phases of exons

A strong indication for mutually exclusive splicing is the impossibility to incorporate more than one of the MXEs of a cluster into the final transcript because of the incompatibility of the splice site phases. Exons can be classified based on the phase of the flanking intron: symmetric exons are 0-0 (intron interrupts the reading frame between two consecutive codons), 1-1 (intron interrupts the reading frame between the first and second base of a codon) and 2-2, and asymmetric exons are 0-1, 0-2, 1-0, 1-2, etc. Symmetric exons are the only ones that can be spliced in succession without changing the reading frame. Thus, constitutive exons sharing our criteria of MXEs comprise only symmetric exons (Fig. 23). Compared to the annotated MXEs, the predicted MXEs show a slightly higher percentage of symmetric exons. Therefore, these potential exon candidates could also be spliced constitutively or they could be incorporated in a differentially included manner.

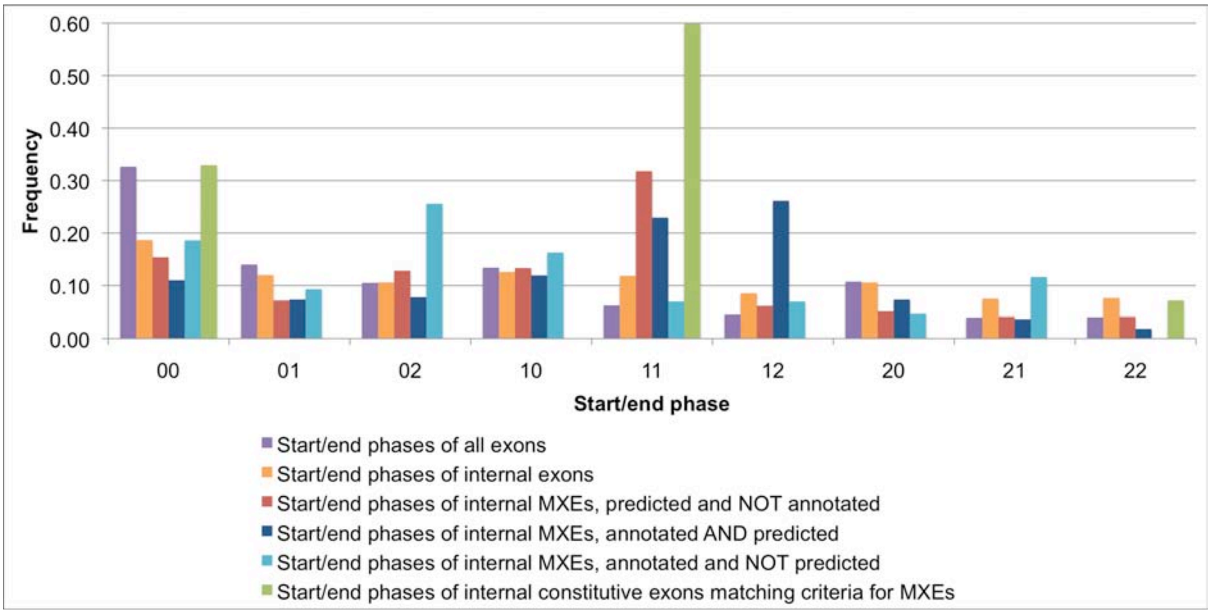


Fig. 23: Comparison of start/end phases of exons.

Splice junctions

As known, by far most introns have the splice junctions GT---AG followed by the GC---AG splice junctions (Fig. 24). Only a few of the annotated introns have other splice junctions. The percentage of the GC---AG splice junction in introns surrounding MXEs is slightly higher than that of all introns (Fig. 24). These numbers are, however, hard to interpret because the total number of MXEs spliced by GC---AG is very low.

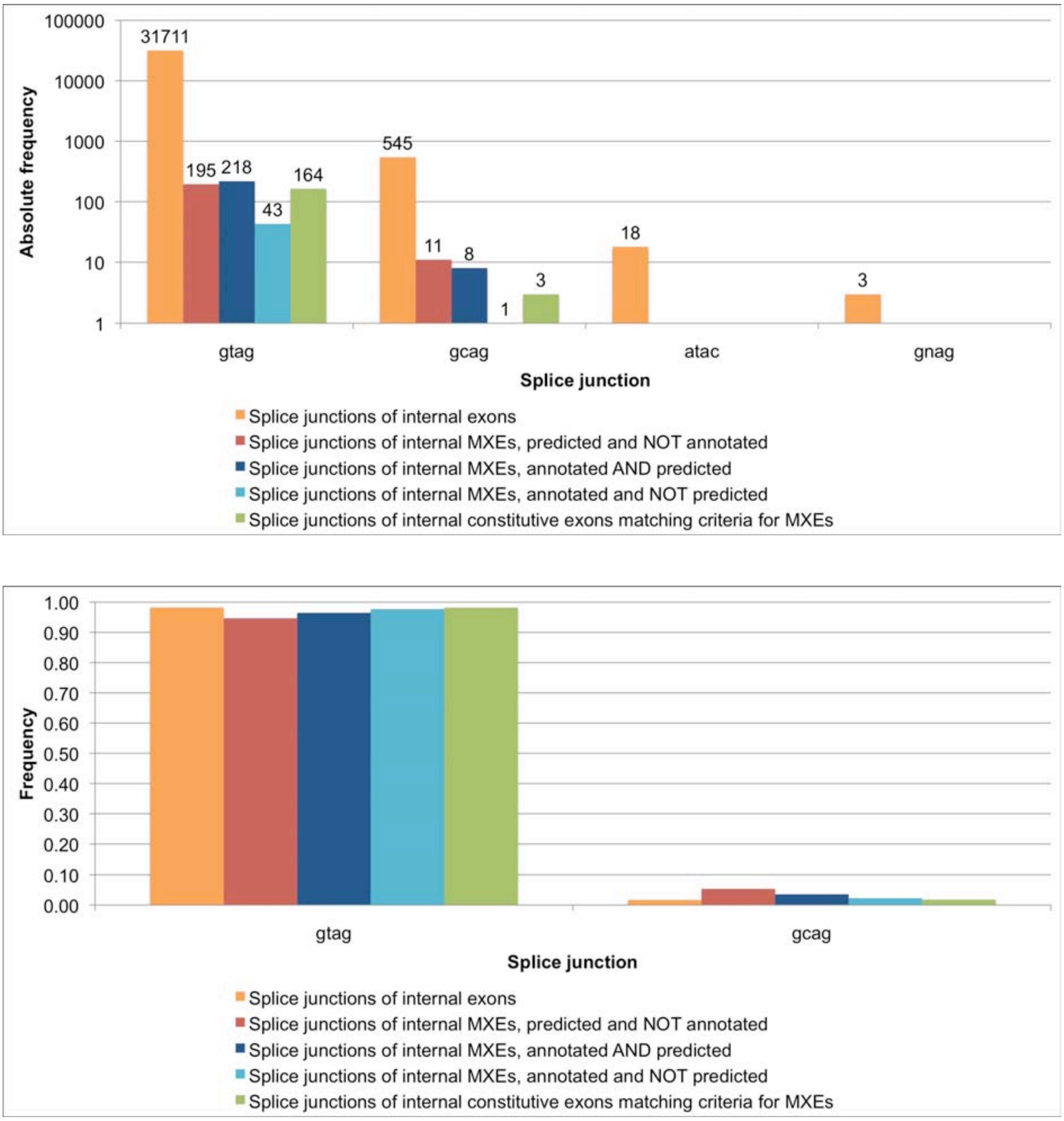


Fig. 24: Comparison of splice junctions. The splice junctions of all introns are compared to those of the putative introns between an MXE and the next constitutive exon before and after a cluster of MXEs. MXEs are separated in annotated or predicted MXEs and compared to internal constitutive exons sharing our criteria for MXEs.

Patterns of splice junctions

Splice junctions display sequence conservation beyond the two-base splice site (Fig. 25). Characteristic to all internal exons (pattern strongly dominated by constitutive exons) and the constitutive exons sharing our criteria of MXEs are the considerably stronger conservation of the bases AGT in positions +4, +5 and +6 of the intron. In contrast, the introns following the MXEs (annotated and predicted) have a stronger conserved G in position -1. The 3’ ends of the introns before the MXEs have similar patterns as compared to all introns.

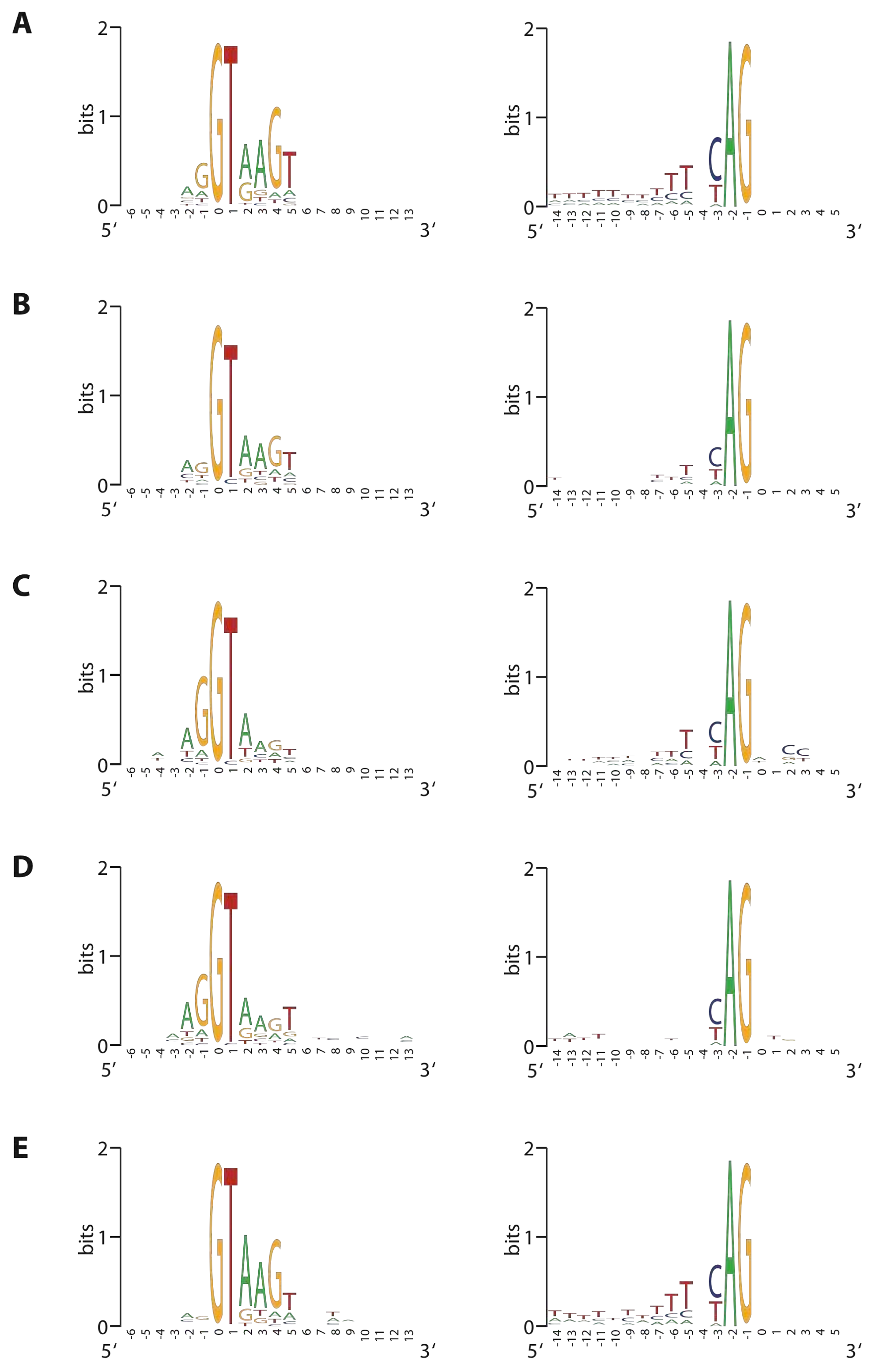


Fig. 25: Conservation of intron splice junctions. The weblogos were generated from the aligned 14 nucleotides of the intron and six nucleotides of the exon of both the 5’- and 3’-splice sites. The height of the letters represents the degree of conservation. A) All internal introns. B) Predicted internal MXEs that were not annotated. C) Annotated and reconstructed internal MXEs. D) Annotated but not reconstructed internal MXEs. E) Internal constitutive exons matching our criteria of MXEs.

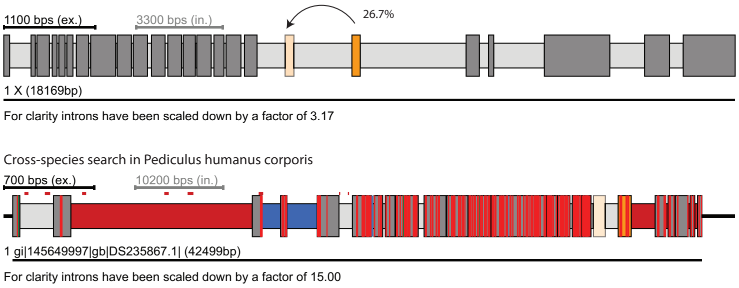
A2 Supplementary figures

Supplementary figure 1

Figure 1: Genes containing newly predicted mutually exclusive exons which were not annotated in Flybase release 5.36, but are annotated in Flybase release 5.48.

Gene: mmd, mind-meld, FBgn0259110
Polypeptide: mmd-PD, FBpp0288810

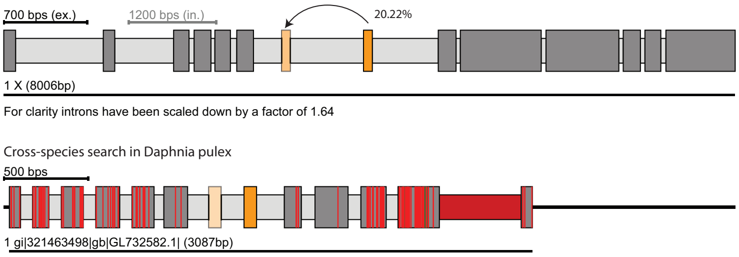
Exon A is annotated in r5.48.
Supported by RNA-Seq data.
Conserved in Anopheles gambiae, Pediculus humanus corporis, dana, dere, dgri, dmoj, dper, dsec, dsim, dvir, dwil and dyak.
RNA-Seq: Exon B has an alternative splice site at the 3'-end.


1100 bps (ex.)
3300 bps (in.)
26.7%
1 X (18169bp)
For clarity introns have been scaled down by a factor of 3.17
Cross-species search in Pediculus humanus corporis
700 bps (ex.)
10200 bps (in.)
1 gi|145649997|gb|DS235867.1| (42499bp)
For clarity introns have been scaled down by a factor of 15.00

```
                  10        20        30
          ....|....|....|....|....|....|....|..
exonA     GTSDQNRISTLTMVIILTVIVKCVFISFATLAVCYR-
exonB     --ENYHGSNTVFLVGVLMSVVGFVFITFTLMALCYR-
AngExonA  -DHDQGKVSTLAMVIMLVVIVKCVFLCFALMAVCYR-
AngExonB  --ANYHGSNTVFLVGVLMSVVGGVFITFALMALCYRS
PdcExonA  -GGNNNNLSTLAMVFILVGVVKGVFICFTLMAVCYR-
PdcExonB  --ENYHSTNTGFLVGVLMSVVGGVFILFALMALCYR-
```

Gene: g, garnet, FBgn0001087
Polypeptide: g-PB, FBpp0073673

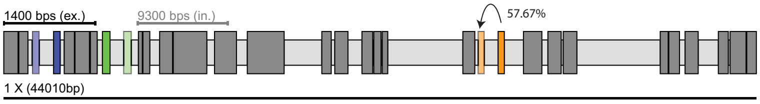
Exon A is annotaed in r5.48.
Supported by EST and RNA-Seq data.
Conserved in Aedes aegypti, Anopheles gambiae, Apis mellifera, Atta cephalotes, Daphnia pulex dana, dere, dgri, dmoj, dper, dpse, dsec, dvir, dwil and dyak.
RNA-Seq: Exon A has an alternative splice site at the 5'-end.


700 bps (ex.)
1200 bps (in.)
20.22%
1 X (8006bp)
For clarity introns have been scaled down by a factor of 1.64
Cross-species search in Daphnia pulex
500 bps
1 gi|321463498|gb|GL732582.1| (3087bp)

```
                   10        20
           ....|....|....|....|....|
exonA      FASLTTIEPALGRKLTQPLIEIISS
exonB      FGALTPLEPRLGKKLIEPLTNLIHS
DapExonA   LNAMTQCDSRLSKCISQPLIAIIKS
DapExonB   FGALTPLEPRLGKKLIEPLTNLIHS
```

Gene: cac, cacophony, FBgn0263111
Polypeptide: cac-PA, FBpp0298319

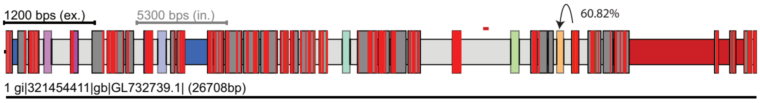
Exon A is annotaed in r5.48.
Supported by RNA-Seq data.
Conserved in Aedes aegypti, Anopheles gambiae, Apis mellifera, Atta cephalotes, Daphnia pulex, Pediculus humanus corporis dana, dere, dgri, dmoj, dper, dpse, dsec, dvir, dwil and dyak.



For clarity introns have been scaled down by a factor of 6.73

```
                    10        20        30
           ....|....|....|....|....|....|..
exonA      VFGNIRYDP-DTQLNRHNNFQSFSGGIMLLFR
exonB      VFGNIKLGTVENSITRHNNFQSFIQGVMLLFR
DapExonA   VFGNLHLDP-DSSVNRHNNFQSFIGGLLLLFR
DapExonB   VFGNILLEPGTTHIHRHNNFRSFIQGLMLLFR
```

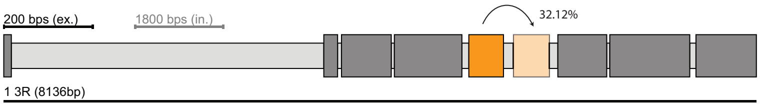
Cross-species search in Daphnia pulex



For clarity introns have been scaled down by a factor of 4.46

Gene: vib, vibrator, FBgn0262468
Polypeptide: vib-PA, FBpp0083159

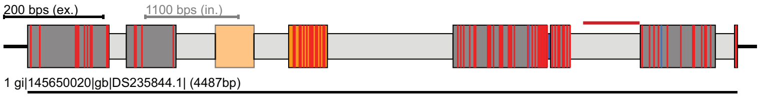
Exon B is annotated in r5.48.
Supported by EST and RNA-Seq data.
Conserved in Aedes aegypti, Anopheles gambiae, Pediculus humanus corporis, Tribolium castaneum, dana, dere, dgri, dmoj, dper, dpse, dsec, dsim, dvir, dwil and dyak.



For clarity introns have been scaled down by a factor of 9.08

```
                    10        20
           ....|....|....|....|....|..
exonA      NPKFMKDAFKIIIDTLHV-GDAGDSEN
exonB      NPGYMDKNFKIDIYSQHIENDLGTVDN
PdcExonA   NPLYMKEKFHLTIESHHL-IDDGQNEN
PdcExonB   NPGYMKENFLIMIESFHI-NDSGYQEN
```

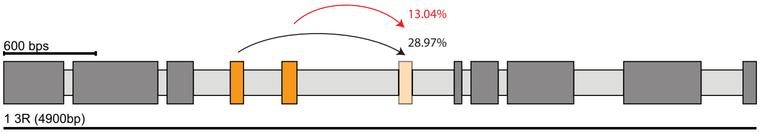
Cross-species search in Pediculus humanus corporis



For clarity introns have been scaled down by a factor of 5.90

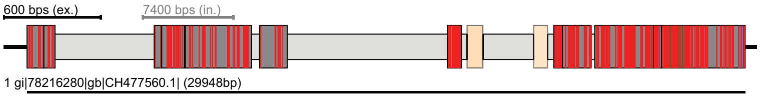
Gene: Esyt2, FBgn0039208
Polypeptide: Esyt2-PA, FBpp0084031

Exon C is annotated in r5.48.
Supported by RNA-Seq data.
Conserved in Aedes aegypti, Anopheles gambiae, dana, dere, dgri, dmoj, dper, dpse, dsec, dsim, dvir, dwil and dyak.



```
                    10        20        30
           ....|....|....|....|....|....|...
exonA      ATVFIEMGQFVEIQLKDSDDS----KKDENLGR
exonB      ACIFTTIGHYIGFSLWDYDQTMPGVQSDDVLGR
exonC      AVVEVSQHAILVLRLFDWDRT----SDDESLGR
AeaExonA   AFIHAESGQQLQIVLNDKDAG----GDDELLGR
AeaExonB   AEVNATLGQEIELNLWDWDPGFPGVQNDDYLGR
AeaExonC   ACVDVSHQTLIGIKLFDWDRT----GDHDPLGR
```

Cross-species search in Aedes aegypti



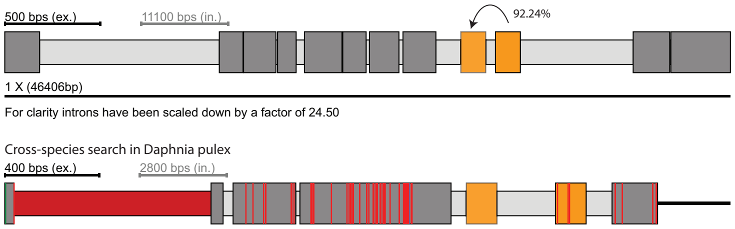
For clarity introns have been scaled down by a factor of 13.19

Supplementary figure 2

Figure 2: Genes containing newly predicted mutually exclusive exons which were not annotated in Flybase release 5.36 nor in release 5.48.

Gene: Sh, Shaker, FBgn0003380
Polypeptide: Sh-PB, FBpp0088600

RNA-Seq supports 3'-end of exon A.
Conserved in Aedes aegypti, Anopheles gambiae, Apis mellifera, Daphnia pulex, Pediculus humanus corporis, Tribolium castaneum, dana, dere, dgri, dmoj, dpse, dsec, dvir, dwil and dyak.



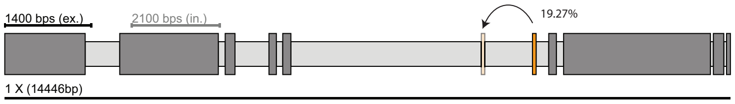
500 bps (ex.)
11100 bps (in.)
92.24%
1 X (46406bp)
Cross-species search in Daphnia pulex
400 bps (ex.)
2800 bps (in.)
1 gi|321463984|gb|GL732578.1| (11474bp)

For clarity introns have been scaled down by a factor of 24.50

For clarity introns have been scaled down by a factor of 7.60

```
                   10        20        30        40
           ....|....|....|....|....|....|....|....|....
exonA      GVVLFSSAVYFAEAGSDSSFFKSIPDGFWWAVVTMTTVGYGDMR
exonB      GVVLFSSAVYFAEAGSENSFFKSIPDAFWWAVVTMTTVGYGDMT
DapExonA   GVILFSSAVYFAEAGSEYSYFKSIPDAFWWAVVTMTTVGYGDMR
DapExonB   GVILFSSAVYFAEAGSEVSHFKSIPDAFWWAVVTMTTVGYGDMT
```

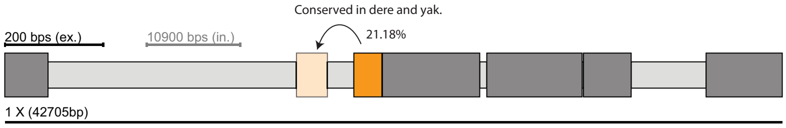
Gene: fs(1)h, female sterile (1) homeotic, FBgn0004656
Polypeptide: fs(1)h-PB, FBpp0071074



For clarity introns have been scaled down by a factor of 1.48

```
              10        20
       ....|....|....|....|
exonA  GRGNKKKGRKKSGR-RELRN
exonB  GDGDERPPRKKKSRDSNGSN
```

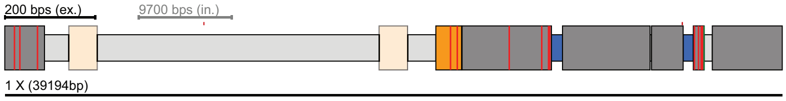
Gene: CG12541, FBgn0029930
Polypeptide: CG12541-PD, FBpp0289984



For clarity introns have been scaled down by a factor of 57.31

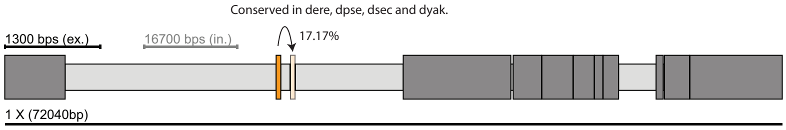
```
                    10        20
           ....|....|....|....|.
exonA      VRFMRSLMIAERASTKASLKY
exonB      V--VRLEVFAEEVTTAASLSE
dyakExonA  VIVAHTFAFEICVVTLAMCSS
dyakExonB  VRFMGSQVFAVRLSAKASLKY
dyakExonC  V--VRLEVFAEELTTAAALSE
```

Cross-species search in dyak



For clarity introns have been scaled down by a factor of 47.22

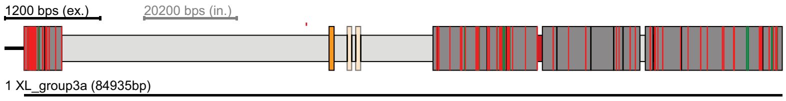
Gene (r5.36): CG42248
Polypeptide(r5.36): CG42248-PD, FBpp0288785
Gene (r5.48): CG43867, FBgn0264449
Polypeptide (r5.48): CG43867-PD, FBpp0304858



For clarity introns have been scaled down by a factor of 13.14

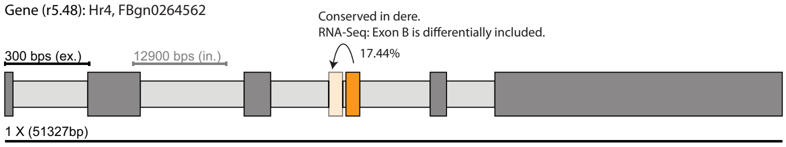
```
                    10        20
           ....|....|....|....|.
exonA      LTELEQRVIEAEERAEEAEDK
exonB      ASTWQLAVLESVENAGKSARK
dpseExonA  LTELEQRVIEAEERAEEAEDK
dpseExonB  LRGIERN--TARERESDVEER
dpseExonC  ATAREQRSCAACERESAARTC
```

Cross-species search in dpse



For clarity introns have been scaled down by a factor of 17.28

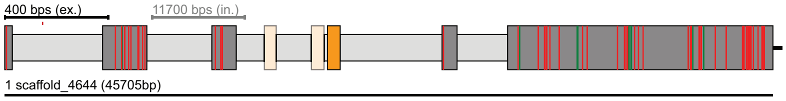
Gene (r5.36): CG3600
Polypeptide(r5.36): CG3600-PC, FBpp0288868
Gene (r5.48): Hr4, FBgn0264562



For clarity introns have been scaled down by a factor of 39.21

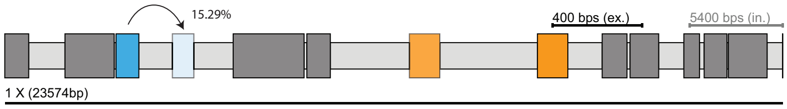
```
                     10
            ....|....|....|.
exonA       RARSAVGQRPVGGRFI
exonB       TCQAEEGQSSAGSHYT
dereExonA   RC-HELGERSSTSTWN
dereExonB   SERSAVGQRPVGGRFI
dereExonC   TCQAEEGQSSAGSHYT
```

Cross-species search in dpse



For clarity introns have been scaled down by a factor of 32.62

Gene: SK, small conductance calcium-activated potassium channel, FBgn0029761
Polypeptide: SK-PH, FBpp0289694

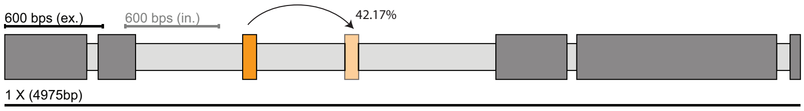


For clarity introns have been scaled down by a factor of 13.00

```
               10        20        30
       ....|....|....|....|....|....|....
exonA  ASFYSTALKTLISVSTVILLGLIVAYHALEVQVR
exonB  KSHNSYSLHTICSLSLSII--IITPNQCLPPQIN
```

Gene: mys, myospheroid, FBgn0004657
Polypeptide: mys-PA, FBpp0071061

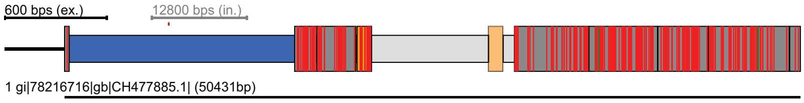
RNA-Seq data supports exon B.
Conserved in Aedes aegypti, Anopheles gambiae, dere, dgri, dmoj, dper, dpse, dsec, dsim, dvir, dwil and dyak.
RNA-Seq: Exons are differentially included.



For clarity introns have been scaled down by a factor of 1.05

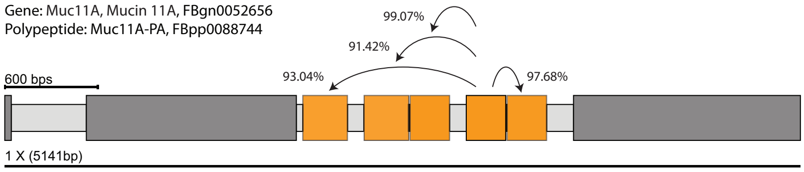
```
                   10        20
           ....|....|....|....|....|...
exonA      LEHPCENCKAPYGYQNHMPLNNNTESFS
exonB      LVEPCANCTATYGFHHQMVLDKNITQFT
AeaExonA   LEHPCDGCEAPYGYKNHMSLSVDTSRFS
AeaExonB   LREPCPQCAAPYGYHNLMPLSVDTHRFT
```

Cross-species search in Aedes Aegypti



For clarity introns have been scaled down by a factor of 22.94

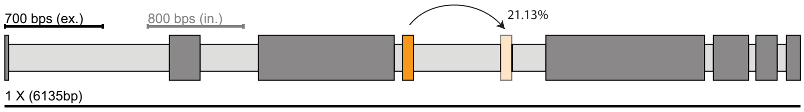
Gene: Muc11A, Mucin 11A, FBgn0052656
Polypeptide: Muc11A-PA, FBpp0088744



```
               10        20        30        40        50        60        70        80        90
       ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|.
exonA  ADGSSAAPGSPADVTTAAPGAPADGSSAAPGAPADGSSAAPGSPADVTTAAPGAPADGSSAAPGSPAEESSAAPGSPADVTTAAPGAPADGSSAAP
exonB  AEGSSAAPGAPADVTTAAPGAPADGSSAAPGAPADGSSAAPGSPADVTTAAPGAPADGSSAAPGAPADGSSAAPGSPADVTTAAPGAPADGSSAAP
exonC  AEGSSAAPGSPADVTTAAPGAPADGSSAAPGAPADGSSAAPGSPADVTTAAPGAPADGSSAAPGSPADVTTAAPGAPADGSSAAP
exonD  ADGSSAAPGSPADVTTAAPGAPADGSSAAPGAPADGSSAAPGSPADVTTAAPGAPADGSSAAPGSPADVTTAAPGAPADGSSAAP
exonE  AEGSSAAPGAPADVTTAAPGAPADGSSAAPGAPADGSSAAPGSPADVTTAAPGAPADGSSAAPGAPADVTTAAPGAPADGSSAAP
```

Gene: Top1, Topoisomerase 1, FBgn0004924
Polypeptide: Top1-PA, FBpp0073822

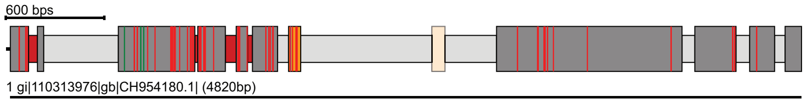
RNA-Seq data supports 3'-end of exon B.
Conserved in dere, dsec, dsim and dyak.
RNA-Seq: Exon A has an alternative splice site at 5'-end.



For clarity introns have been scaled down by a factor of 1.17

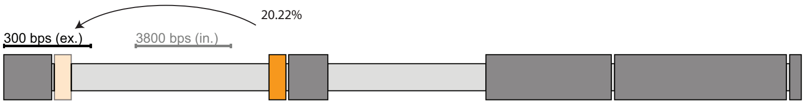
```
                     10        20
             ....|....|....|....|....|..
exonA        EP-EPAVSPGKRQKAKAKVEEEEVWRW
exonB        RMLAVATVAGKRRRVRRKSVQEEQIRW
dereExonA    EP-E-VVSPTKRQKAKVKEEEEEVWRW
dereExonB    RMVAKVTNDGKKRRVRRKSVQEEQVRW
```

Cross-species search in dere



Gene: mol, moladietz, FBgn0086711
Polypeptide: mol-PA , FBpp0080238

Conserved in Aedes aegypti, dsec, dsim and dwil.

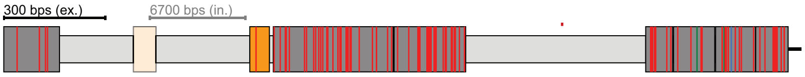


1 2L (16153bp)

For clarity introns have been scaled down by a factor of 11.57

```
                10        20
          ....|....|....|....|..
exonA     KARWFKLINLYYFKNATLL
exonB     KFTTFSTVTLSLFVGLVIL
AeaExonA  QFNWFSTRSVVRFSSLVHLKIQ
AeaExonB  KFTTFTTVTLSLFVGLVIL
```

Cross-species search in Aedes aegypti

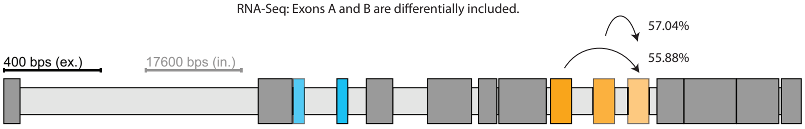


1 gi|78216149|gb|CH477448.1| (26216bp)

For clarity introns have been scaled down by a factor of 23.76

Gene: nAcRalpha-30D, nicotinic acetylcholine receptor alpha 30D, FBgn0032151

Conserved in Aedes aegypti, Anopheles gambiae, Apis mellifera, Atta cephalotes, Pediculus humanus corporis, Tribolium castaneum, dana, dere, dgri, dmoj, dper, dpse, dsim, dvir, dwil and dyak.
RNA-Seq: Exons A and B are differentially included.

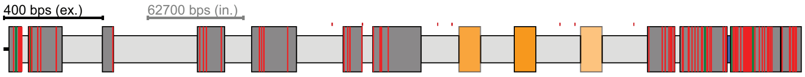


1 2L (72488bp)

For clarity introns have been scaled down by a factor of 44.59

```
                 10        20
          ....|....|....|....|....|....
exonA     GVTILLSLTVFLNLVAESMPTTSDAVPLI
exonB     GVTILLSLTVFLNLVAETLPQVSDAIPLL
exonC     GVTILLSQTVFSLLVGNVITKTSEAVPLL
AmExonA   GVTILLSLTVFLNLVAESMPTTSDAVPLI
AmExonB   GVTILLSLTVFLNLVAETLPQVSDAIPLL
AmExonC   GVTILLSQTVFSLLVNHVLTRTSEAVPLI
```

Cross-species search in Apis mellifera

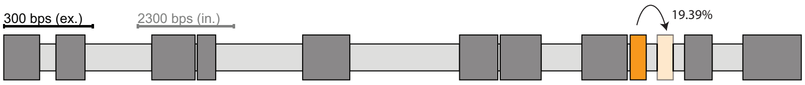


1 gi|318087635|gb|CM000055.5| (246068bp)

For clarity introns have been scaled down by a factor of 162.56

Gene: CG14010, FBgn0031725
Polypeptide: CG14010-PB , FBpp0292059

Conserved in dere, dwil and dyak.

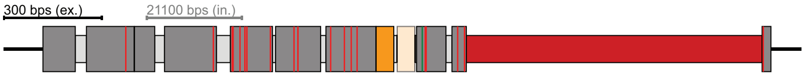


1 2L (10518bp)

For clarity introns have been scaled down by a factor of 7.08

```
                  10
           ....|....|....|...
exonA      RIPPNAVNYVENFEARHK
exonB      RIMGNIKLVANEWKARKK
dwilExonA  RIPPNAVNYVENFEARHK
dwilExonB  KAANISLIFVFVYQTRHK
```

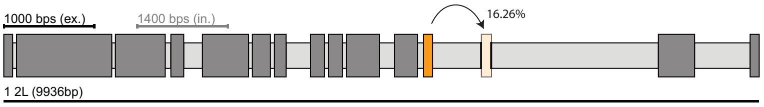
Cross-species search in dwil



1 scf2_1100000004521 (78736bp)

For clarity introns have been scaled down by a factor of 72.18

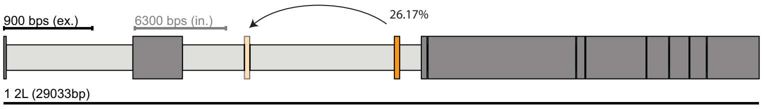
Gene: tim, timeless, FBgn0014396
Polypeptide: tim-PB , FBpp0077256



For clarity introns have been scaled down by a factor of 1.40

```
               10        20        30
       ....|....|....|....|....|....|....|..
exonA  Y--TPDPTP-PVPNWLQLVMRSKCNHRTGPSGDPSDC
exonB  FGPTPSPTPSPTPSTSQDPTRSDAAHPLAELAAPSIF
```

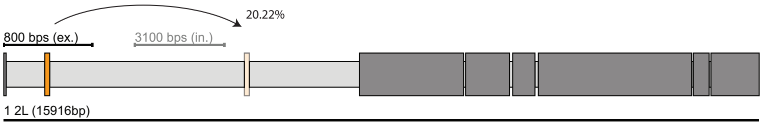
Gene: IA-2, IA-2 ortholog, FBgn0031294
Polypeptide: IA-2-PC , FBpp0290630



For clarity introns have been scaled down by a factor of 6.81

```
               10
       ....|....|....|....
exonA  ATEIIFLLC-PYSHVCCFD
exonB  GCQFVRTLCIPHSEV-CYD
```

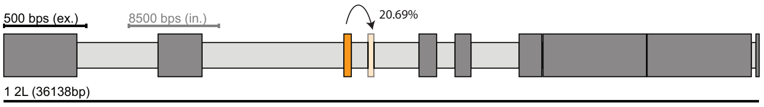
Gene: ush, u-shaped, FBgn0003963
Polypeptide: ush-PA, FBpp0077723



For clarity introns have been scaled down by a factor of 3.82

```
               10
       ....|....|....|
exonA  GDCSDTAEEMTVDSR
exonB  GWTTETVEVHIIELQ
```

Gene: CG32982, FBgn0052982
Polypeptide: CG32982-PE, FBpp0290262

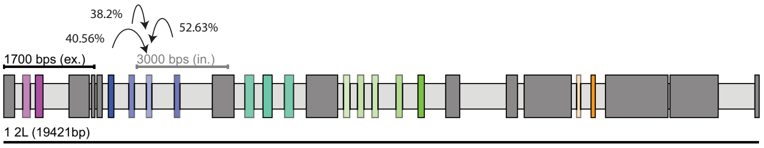


For clarity introns have been scaled down by a factor of 15.52

```
               10
       ....|....|....|
exonA  VSCNKQTNWLNFKQD
exonB  IEC---TMWLDRRES
```

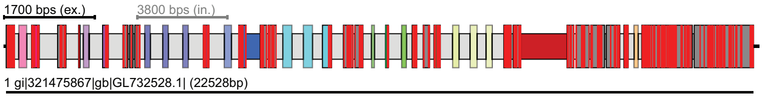
Gene: Mhc, Myosin heavy chain, FBgn0264695
Polypeptide: Mhc-PA, FBpp0080453

RNA-Seq data supports 3'-end of exon C.
Verified by literature.
Conserved in Aedes aegypti, Anopheles gambiae, Apis mellifera, Atta cephalotes, Daphnia pulex, Pediculus humanus corporis, Tribolium castaneum, dana, dere, dgri, dmoj, dper, dpse, dsec, dvir, dwil and dyak.



For clarity introns have been scaled down by a factor of 1.75

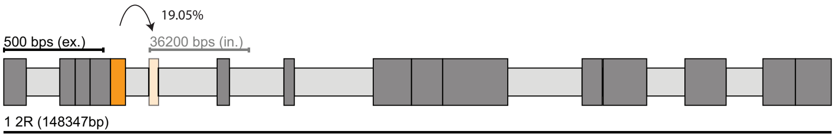
Cross-species search in Daphnia pulex



For clarity introns have been scaled down by a factor of 2.25

```
                   10        20        30
           ....|....|....|....|....|....|....|
exonA      DICLLTDNIYDYHIVSQGKVTVASIDDAEEFSLTD
exonB      EYCLLSNNIYDYRIVSQGKTTIPSVNDGEEWVAVD
exonC      EMVFLGQHIGDYPGICQGKTRIPGVNDGEEFELTD
exonD      EMCFLSDNIYDYYNVSQGKVTVPNMDDGEEFQLAD
DapExonA   ADCCLVDDIYQYNFVSQGKITIPSMDDSEEMALTD
DapExonB   ADCSLVDDIYTYNFVSQGKITIPSMDDSEEMGLTN
DapExonC   ADCRLVDDIYTYNYVSQGKITIPSMDDNEEMGLTD
DapExonD   AMCSLSDNIYDYPFVSQGKVTVPSIDDSEEMQMAD
```

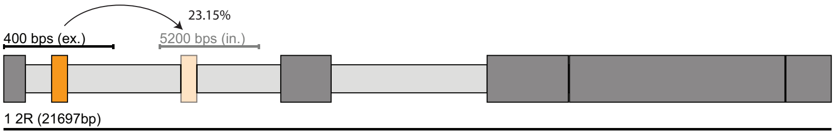
Gene: Gprk1, G protein-coupled receptor kinase 1, FBgn0260798
Polypeptide: Gprk1-PA, FBpp0110413



For clarity introns have been scaled down by a factor of 72.60

```
                10        20
         ....|....|....|....|....|
exonA    EYSKHAVASVQKYLLKNEVPVDLFE
exonB    ---------CKIFLLKNEVLVDLFE
```

Gene: CG30438, FBgn0050438
Polypeptide: CG30438-PB, FBpp0085404

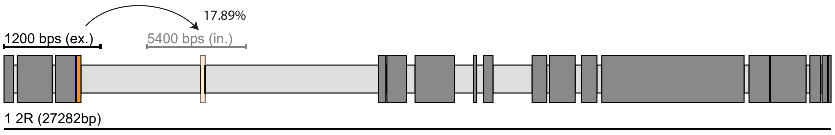


For clarity introns have been scaled down by a factor of 14.36

```
                10        20
         ....|....|....|....|
exonA    GGTKSHKIPFWELAKGLISR
exonB    GGLPEETTRKWRVQKGQWSQ
```

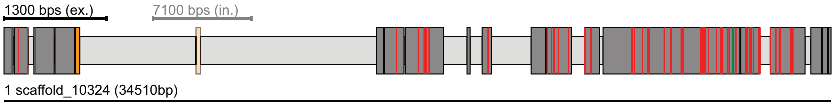
Gene: brp, bruchpilot, FBgn0259246
Polypeptide: brp-PD, FBpp0289193

Conserved in dgri, dmoj, dvir, dwil and dyak.



For clarity introns have been scaled down by a factor of 4.40

Cross-species search in dvir

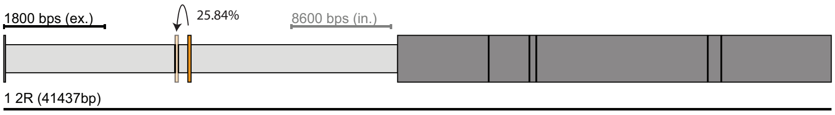


For clarity introns have been scaled down by a factor of 5.67

```
                    10        20
            ....|....|....|....|..
exonA       G--KEEERQMFQQMQAMA-QKQ
exonB       ---EQEQNRTFDSIQKSISQKA
dvirExonA   G--KEEERQMFQQMQAMA-QKQ
dvirExonB   GVKREKERRSRRQMQPCA--KQ
```

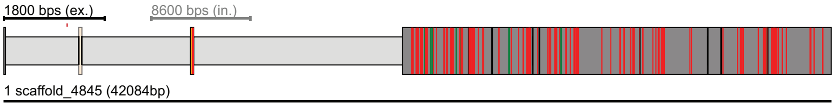
Gene: shn, schnurri, FBgn0003396
Polypeptide: shn-PD, FBpp0089118

Conserved in dere, dmoj, dsec, dsim and dwil.



For clarity introns have been scaled down by a factor of 4.86

Cross-species search in dvir

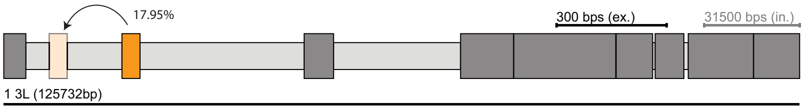


For clarity introns have been scaled down by a factor of 4.87

```
                    10        20
            ....|....|....|....|
exonA       KTTIVIKC-SKWVTSRHQEK
exonB       KSTVN-SRKSALETAREKTK
dereExonA   KQQIKATHK-ANRNKTQKIK
dereExonB   KSTVN-SRKSALESVREKPK
dmojExonA   KSTVN-SRKNTLESTREKLK
dmojExonB   KQQQRLSKKKCLSSALESSK
```

Gene: bru-3, bruno-3, FBgn0264001
Polypeptide: bru-3-PB, FBpp0303379

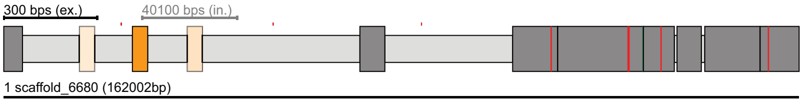
Conserved in dana, dere, dgri, dmoj, dper, dpse, dsec, dsim, dvir, dwil and dyak.



For clarity introns have been scaled down by a factor of 121.65

```
              10
         ....|....|....|.
exonA        IHKAGHSKPGNSSSFV
exonB        MNRALQLKPAENESRS
dmojExonA    SSQVLSVKCCSNIIES
dmojExonB    MNRALQLKPAENESRS
dmojExonC    MRAALDVLPISSLNSS
```

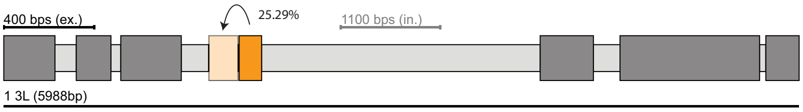
Cross-species search in dvir



For clarity introns have been scaled down by a factor of 131.27

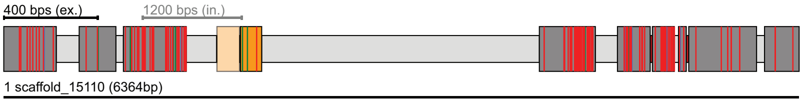
Gene: ect, ectodermal, FBgn0000451
Polypeptide: ect-PA, FBpp0076034

Conserved in dana, dere, dgri, dsim and dyak.



For clarity introns have been scaled down by a factor of 2.51

Cross-species search in dgri

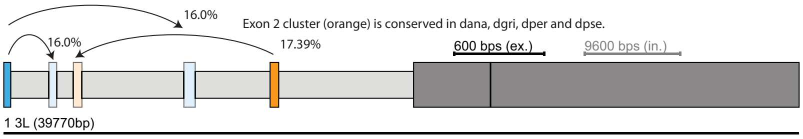


For clarity introns have been scaled down by a factor of 2.86

```
                  10        20        30        40
             ....|....|....|....|....|....|....|....|....
exonA        GAVAPGNVAAGADDTDNDDDDYDEDDETDDDDDDDIDDGVDEI-
exonB        G-------IAGDDDEEADDDD---DDDDDDIIGDDIIEARREA-
dgriExonA    ---------ASDDDDYDDEDD---EYDDDDYSDDDIDEGVDEIT
dgriExonB    ----------GDDEEADDDDD---DDDDDDIIGDDVIEARREA-
```

Gene: CG7991, FBgn0035260
Polypeptide: CG7991-PB, FBpp0292221

Exon 1 cluster (blue) is conserved in dere, dsec, dsim, dvir and dyak.

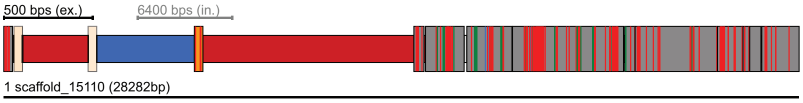


Exon 2 cluster (orange) is conserved in dana, dgri, dper and dpse.

For clarity introns have been scaled down by a factor of 15.12

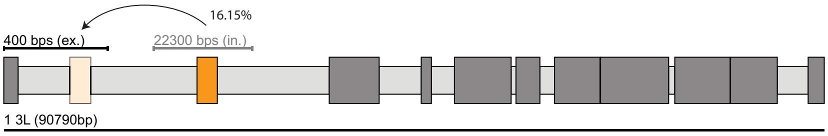
```
                        10        20
               ....|....|....|....|....|.
exon1A         MLF----------WKRRLQRSSSSS-
exon1B         MII---------DWRANMRKERSLST
exon1C         MAFILWRQGVASCWKNRRHKSSSLR-
exon2A         KRMQKRLFFSTFCHNMAKK
exon2B         SAIQERRFFGILRSAKRKD
dgriExon2A     ---QQQTEAVVATVGRRKM
dgriExon2B     ---QTIRFIDFRFAALRNN
dgriExon2C     ---QQRRFFGILRAGKRKD
```

Cross-species search in dgri



For clarity introns have been scaled down by a factor of 12.03

Gene: Eip63E, Ecdysone-induced protein 63E, FBgn0264001
Polypeptide: Eip63E-PD, FBpp0072990

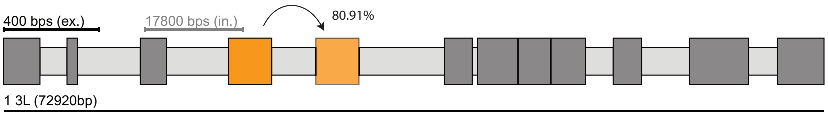


For clarity introns have been scaled down by a factor of 59.04

```
               10        20
       ....|....|....|....|....|..
exonA  GSTKIEKSDLKIQVIYMQMSNKYGQRG
exonB  GVTMREKKGGALQKLKKRLSHSFG-RL
```

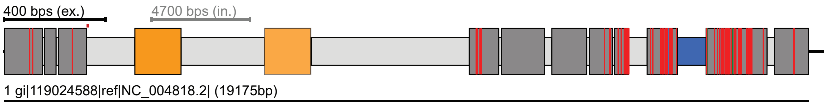
Gene: nAcRalpha-80B, nicotinic Acetylcholine Receptor alpha 80B, FBgn0037212
Polypeptide: nAcRalpha-80B-PC, FBpp0289395

Conserved in Aedes aegypti, Anopheles gambiae, Apis mellifera, Atta cephalotes, Daphnia pulex
Pediculus humanus corporis, Tribolium castaneum, dana, dere, dgri, dmoj, dper, dsec and dwil.



For clarity introns have been scaled down by a factor of 43.15

Cross-species search in Anopheles gambiae



For clarity introns have been scaled down by a factor of 12.19

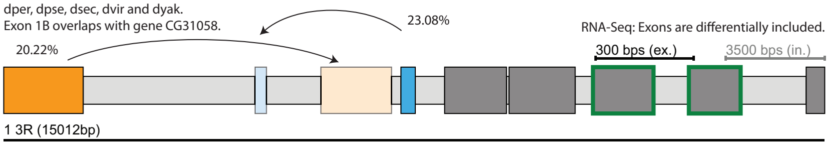
```
                  10        20        30        40        50        60
          ....|....|....|....|....|....|....|....|....|....|....|....|
exonA     ADGNFEVTLATKATIYSEGLVEWKPPAIYKSSCEIDVEYFPFDEQTCVLKFGSWTYDGFK
exonB     ADGHYEVTLMTKAIVYNNGLVIWQPPAVYKSSCSIDVEYFPYDVQTCILKLGSWTYDGFK
AngExonA  ADGNFEVTLATKATIYSEGLVEWKPPAIYKSSCEIDVEYFPFDEQTCVLKFGSWTYDGFK
AngExonB  ADGHYEVTLMTKATVYNNGMVIWQPPAVYKSSCSIDVEYFPYDVQTCVLKLGSWTYDGFK
```

Gene: tau, FBgn0051057
Polypeptide: tau-PA, FBpp0084567

Conserved in dana, dere, dgri,
dper, dpse, dsec, dvir and dyak.
Exon 1B overlaps with gene CG31058.

Conserved in dere, dper, dpse, dsec and dyak.

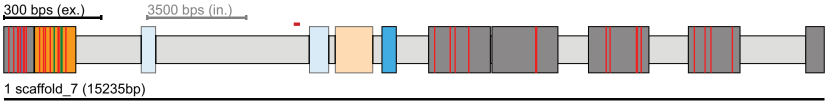
RNA-Seq: Exons are differentially included.



For clarity introns have been scaled down by a factor of 11.53

```
                   10        20
           ....|....|....|....|
exonA      VGDSDS---ESAQVA
exonB      EGDNDSGVDESTQEK
dperExonA  ELSNGFGPSQSQSQA
dperExonB  EQSDNGSAADEAGNAATAES
dperExonC  EGDNDSGVDESTQEK
```

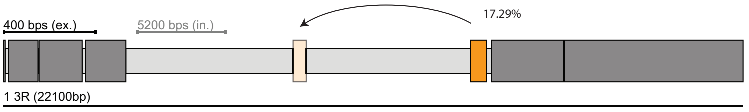
Cross-species search in dper



For clarity introns have been scaled down by a factor of 11.52

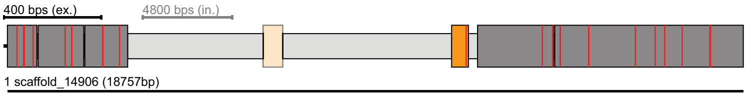
Gene: twin, FBgn0011725
Polypeptide: twin-PA, FBpp0083951

Conserved in dgri, dmoj and dper.



For clarity introns have been scaled down by a factor of 13.60

Cross-species search in dgri

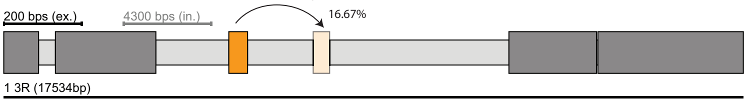


For clarity introns have been scaled down by a factor of 13.08

```
                     10        20
            ....|....|....|....|....|..
exonA       FFQAP----PPL--WVP--ENNPSEPW
exonB       FTVNP----PPQRPWLPLAKPNKTRPA
dgriExonA   FAVTPSLPTPPPLPSSPLSQAGHNRRP
dgriExonB   FTVNP----PPQRPWLPLAKPNKSRPA
dmojExonA   FFYLPVRSTRPIA-QLQMRKPNKSRLH
dmojExonB   FTVNP----PPQRPWLPLAKPNKSRPA
```

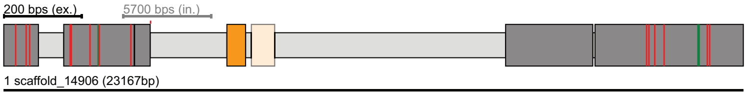
Gene: abd-A, abdominal A, FBgn0000014
Polypeptide: abd-A-PA, FBpp0082828

Conserved in dgri.



For clarity introns have been scaled down by a factor of 19.19

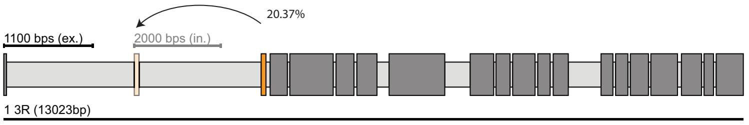
Cross-species search in dgri



For clarity introns have been scaled down by a factor of 25.25

```
                     10        20
            ....|....|....|....|
exonA       D----WMGSPFERVVCGDFN
exonB       D----W--RDFSSVVVGRQT
dgriExonA   D----WMGSPFERVVCGDFN
dgriExonB   DTSGNWQPMPFSSLIVDPCN
```

Gene: CG14741, FBgn0037989
Polypeptide: CG14741-PC, FBpp0297858

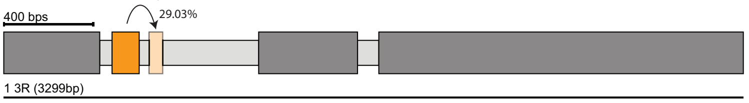


For clarity introns have been scaled down by a factor of 1.86

```
                 10        20
        ....|....|....|....|
exonA   EGEGPKRDHDDAFGTWHRKH
exonB   ENERRIRANDKEFNAQFKYH
```

Gene: CG6241, FBgn0037792
Polypeptide: CG6241-PA, FBpp0081663

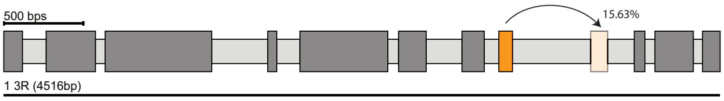
RNA-Seq data supports 3'-end of exon B.
Exon B overlaps with gene CG42759.



```
                 10        20        30        40
        ....|....|....|....|....|....|....|....|
exonA   DARIIYNHKTFKKGKKGKKSTLTGDPNDERAKFRLWNRTK
exonB   --------------------TLTGDPNDERAKFRLWNRTK
```

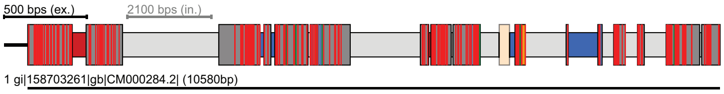
Gene: Unc-115a, FBgn0051352
Polypeptide: Unc-115a-PB, FBpp0289791

Conserved in Aedes aegypti, Anopheles gambiae, Tribolium castaneum, dana, dere, dgri, dmoj, dper, dpse, dsec, dsim, dvir, dwil and dyak.
RNA-Seq: Exon A is differentially included.



```
                    10        20        30        40
            ....|....|....|....|....|....|....|....|..
exonA       ------AIHSYRSPPKPGYG-------FKTTTLPYIRNGFSS
exonB       ------VVSALRHVPKPGYGLAPRSHTFSSTTSAAATMHGAT
TicExonA    ------VVSSLRSVPRPGYGLK------SSTLP---------
TicExonB    ALSLCAVVSALRHVPKPGYGL---------------------
```

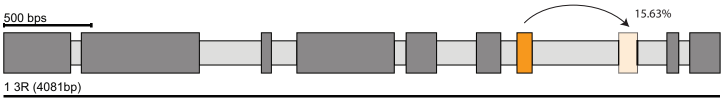
Cross-species search in Tribolium castaneum



For clarity introns have been scaled down by a factor of 4.17

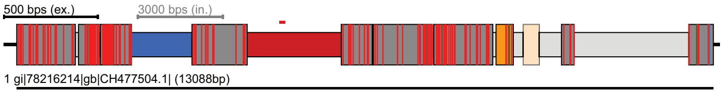
Gene: Unc-115b, FBgn0260463
Polypeptide: Unc-115b-PA, FBpp0081573

Conserved in Aedes aegypti, Anopheles gambiae dana, dere, dgri, dmoj, dper, dpse, dsec, dsim, dvir, dwil, dyak.
RNA-Seq: Exon A is differentially included.



```
                  10        20        30
            ....|....|....|....|....|....|....|.
exonA       AIHSYRSPPKPGYGFKTTTLP--YIRNGFSS-----
exonB       VVSALRHVPKPGYGLAPRSHTFSSTTSAAATMHGAT
AeaExonA    AIHSYRSPPKPGYGFKTSTLPPSSLRNGYSS-----
AeaExonB    VVSSLRQVPKPGYGLAPRSHTFSSATSGS-------
```

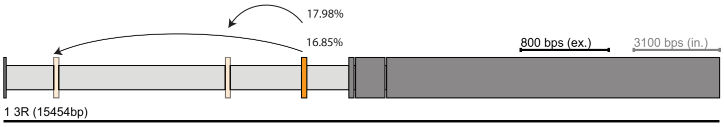
Cross-species search in Aedes aegytpi



For clarity introns have been scaled down by a factor of 6.65

Gene: CG14608, FBgn0037487
Polypeptide: CG14608-PC, FBpp0292311

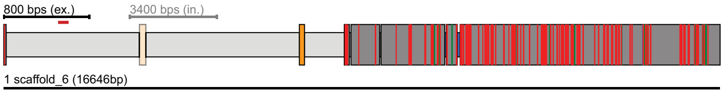
Conserved in dper, dpse, dsec, dsim and dvir.



For clarity introns have been scaled down by a factor of 3.98

```
                      10        20
            ....|....|....|....|
exonA       ----GGFPYSNRGSVVSNSC
exonB       ----ANKKWQS-SFRFSLVC
exonC       ----GNSKYKMRCTQLNLLC
dperExonA   DTCGVNSNYLLIATRLALVC
dperExonB   ----GNSKYKMRCTQLNLLC
```

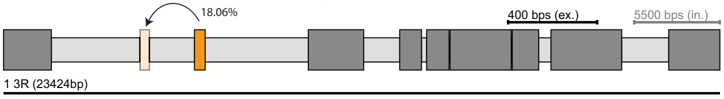
Cross-species search in dper



For clarity introns have been scaled down by a factor of 4.19

Gene: Sap47, Synapse-associated protein 47kD, FBgn0013334
Polypeptide: Sap47-PA, FBpp0082658

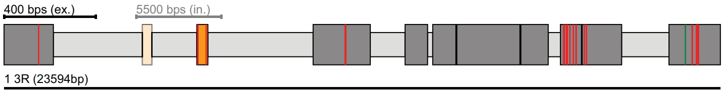
Conserved in dere and dyak.
EST and RNA-Seq: Exon B is differentially included.



For clarity introns have been scaled down by a factor of 14.24

```
                      10
             ....|....|....|.
exonA        IQFSVS--DAATREAT
exonB        LPKSASLVDSLVSEAT
dyakExonA    IQFSVS--DAATREAT
dyakExonB    IPKSASLVDSLVTEAT
```

Cross-species search in dyak



For clarity introns have been scaled down by a factor of 14.77

Supplementary figure 3

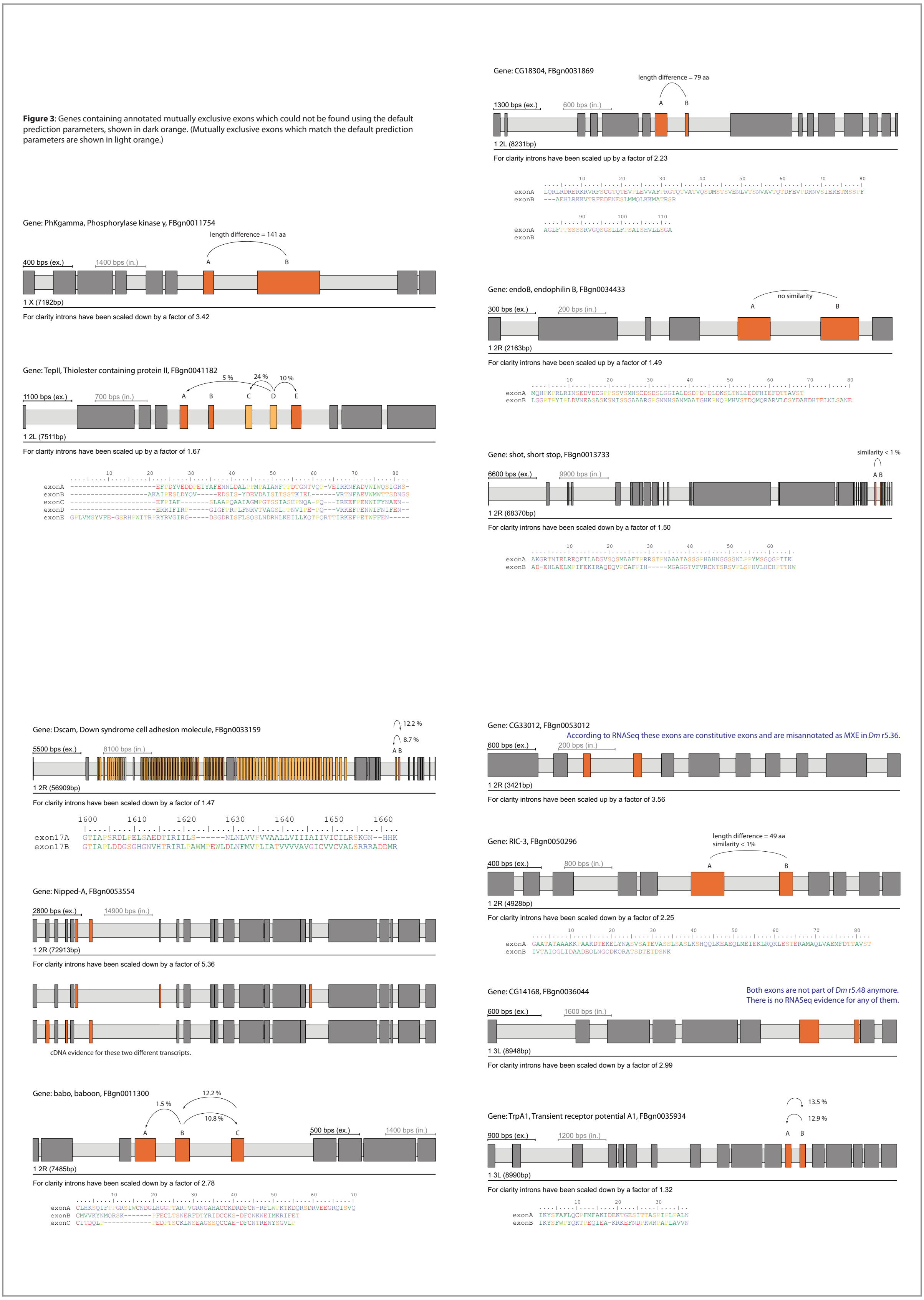


Figure 3: Genes containing annotated mutually exclusive exons which could not be found using the default prediction parameters, shown in dark orange. (Mutually exclusive exons which match the default prediction parameters are shown in light orange.)

Gene: PhKgamma, Phosphorylase kinase γ, FBgn0011754
length difference = 141 aa
400 bps (ex.)
1400 bps (in.)
A
B
1 X (7192bp)
For clarity introns have been scaled down by a factor of 3.42
Gene: TepII, Thiolester containing protein II, FBgn0041182
5 %
24 %
10 %
1100 bps (ex.)
700 bps (in.)
A
B
C
D
E
1 2L (7511bp)
For clarity introns have been scaled up by a factor of 1.67
Gene: CG18304, FBgn0031869
length difference = 79 aa
1300 bps (ex.)
600 bps (in.)
A
B
1 2L (8231bp)
For clarity introns have been scaled up by a factor of 2.23
Gene: endoB, endophilin B, FBgn0034433
no similarity
300 bps (ex.)
200 bps (in.)
A
B
1 2R (2163bp)
For clarity introns have been scaled up by a factor of 1.49
Gene: shot, short stop, FBgn0013733
similarity < 1 %
6600 bps (ex.)
9900 bps (in.)
A B
1 2R (68370bp)
For clarity introns have been scaled down by a factor of 1.50
Gene: Dscam, Down syndrome cell adhesion molecule, FBgn0033159
12.2 %
8.7 %
5500 bps (ex.)
8100 bps (in.)
A B
1 2R (56909bp)
For clarity introns have been scaled down by a factor of 1.47
Gene: CG33012, FBgn0053012
According to RNASeq these exons are constitutive exons and are misannotated as MXE in Dm r5.36.
600 bps (ex.)
200 bps (in.)
1 2R (3421bp)
For clarity introns have been scaled up by a factor of 3.56
Gene: Nipped-A, FBgn0053554
2800 bps (ex.)
14900 bps (in.)
1 2R (72913bp)
For clarity introns have been scaled down by a factor of 5.36
cDNA evidence for these two different transcripts.
Gene: RIC-3, FBgn0050296
length difference = 49 aa
similarity < 1%
400 bps (ex.)
800 bps (in.)
A
B
1 2R (4928bp)
For clarity introns have been scaled down by a factor of 2.25
Gene: CG14168, FBgn0036044
Both exons are not part of Dm r5.48 anymore.
There is no RNASeq evidence for any of them.
600 bps (ex.)
1600 bps (in.)
1 3L (8948bp)
For clarity introns have been scaled down by a factor of 2.99
Gene: babo, baboon, FBgn0011300
1.5 %
12.2 %
10.8 %
A
B
C
500 bps (ex.)
1400 bps (in.)
1 2R (7485bp)
For clarity introns have been scaled down by a factor of 2.78
Gene: TrpA1, Transient receptor potential A1, FBgn0035934
13.5 %
12.9 %
900 bps (ex.)
1200 bps (in.)
A
B
1 3L (8990bp)
For clarity introns have been scaled down by a factor of 1.32

```
                10        20        30        40        50        60        70        80
       ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|...
exonA  --------------------EFPDYVEDDPEIYAFENNLDALPPMPAIANFPPDTGNTVQP-VEIRKNFADVWIWQSIGRS-
exonB  ------------------AKAIPESLDYQV-----EDSIS-YDEVDAISITSSTKIEL------VRTNFAEVWMWTTSDNGS
exonC  --------------------EFPIAF-------SLAAPQAAIAGMPGTSSIASHPNQA-PQ---IRKEFPENWIFYNAEN--
exonD  --------------------ERRIFIRP-----GIGFPRPLFNRVTVAGSLPPNVIPE-PQ---VRKEFPENWIFNIFEN--
exonE  GPLVMSYVFE-GSRHPWITRPRYRVGIRG-----DSGDRISFLSQSLNDRNLKEILLKQTPQRTTIRKEFPETWFFEN-----
```

```
                10        20        30        40        50        60        70        80
       ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
exonA  LQRLRDRERKRVRFSCGTQTEVPLEVVAFPRGTQTVATVQSDMSTSVENLVTSNVAVTQTDFEVPDRNVSIERETMSSPF
exonB  ---AEHLRKKVTRFEDENESLMMQLKKMATRSR

                90       100       110
       ....|....|....|....|....|....|..
exonA  AGLFPPSSSSRVGQSGSLLFPSAISHVLLSGA
exonB
```

```
                10        20        30        40        50        60        70        80
       ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
exonA  MQHPKPRLRINSEDVDCGPPSSVSMHSCDSDSLGGIALDSDPDPDLDKSLTNLLEDFHIEFDTTAVST
exonB  LGGPTPYIPLDVNEASASKSNISSGAAARGPGNNHSANMAATGHKPNQPMHVSTDQMQRARVLCSYDAKDHTELNLSANE
```

```
                10        20        30        40        50        60
       ....|....|....|....|....|....|....|....|....|....|....|....|.
exonA  AKGRTNIELREQFILADGVSQSMAAFTPRRSTPNAAATASSSPHAHNGGSSNLPPYMSGQGPIIK
exonB  AD-EHLAELMPIFEKIRAQDQVPCAFPIH-----MGAGGTVFVRCNTSRSVPLSPHVLHCHPTTHW
```

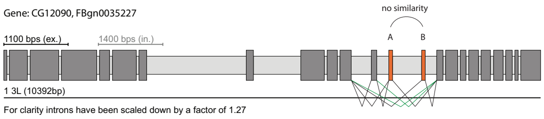
```
          1600      1610      1620      1630      1640      1650      1660
          |....|....|....|....|....|....|....|....|....|....|....|....|...
exon17A   GTIAPSRDLPELSAEDTIRIILS------NLNLVVPVVAALLVIIIAIIVICILRSKGN--HHK
exon17B   GTIAPLDDGSGHGNVHTRIRLPAWMPEWLDLNFMVPLIATVVVVAVGICVVCVALSRRRADDMR
```

```
                10        20        30        40        50        60        70        80
       ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|...
exonA  GAATATAAAKKPAAKDTEKELYNASVSATEVASSLSASLKSHQQLKEAEQLMEIEKLRQKLESTERAMAQLVAEMFDTTAVST
exonB  IVTAIQGLIDAADEQLNGQDKQRATSDTETDSNK
```

```
                10        20        30        40        50        60        70
       ....|....|....|....|....|....|....|....|....|....|....|....|....|....|
exonA  CLHKSQIFPPGRSIWCNDGLHGGPTARPVGRNGAHACCKDRDFCN-RFLWPKTKDQRSDRVEEGRQISVQ
exonB  CMVVKYNMQRSK-------PFECLTSNERFDTYRIDCCKS-DFCNKNEIMKRIFET
exonC  CITDQLP------------PEDPTSCKLNSEAGSSQCCAE-DFCNTRENYSGVLP
```

```
                10        20        30
       ....|....|....|....|....|....|....|..
exonA  IKYSFAFLQCPFMFAKIDEKTGESITTASPIPLPALN
exonB  IKYSFWPYQKTPEQIEA-KRKEFNDPKWRPAPLAVVN
```

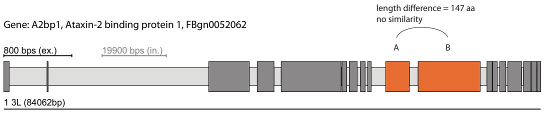
Gene: CG12090, FBgn0035227



For clarity introns have been scaled down by a factor of 1.27

```
                    10        20
            ....|....|....|....|..
exonA  1    GLTNSPFRERVGSNRLPEKRPS
exonB  1    QHDTPRITEKHHNQLNSPLQS
```

Gene: A2bp1, Ataxin-2 binding protein 1, FBgn0052062

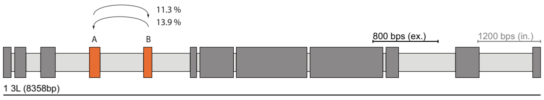


For clarity introns have been scaled down by a factor of 26.36

```
                      10        20        30        40        50        60        70        80
               ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
exonA    1     GYRLPVAAWPFLGAPVGAGAPTLTHLQLPVSMQMAQAPSAVAGGTAATSPATAAAAAAHAAAAAAATPIMLAPRPPHTAV
exonB    1     PAAAVAAAMRGVAIQRGHVGVVGATPYHHPHHPHHHPALLAASAAAAQQQQQRQLAAAAVATAAVAQQQQQQQQAVVQQQ
                      90       100       110       120       130       140       150       160
               ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
exonA   81     QAATPTAATPRRS
exonB   81     QQQVAAAAQQQHQQQQQQQQQAVQQQQAVQQQQQHQQQQQQQQQQQHAAVAAAAAAAASHPHMHAAHAHAHAHALGPQLAQ
                     170       180       190       200       210       220       230       240
               ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
exonA
exonB  161     LQAVAVPTAASNAAALQQSLAAAIQNPSGNPNAAAAAAAAYAARLSAATGATQSPQTAAAAAAAASMAASANAANNAAALH

               ....
exonA
exonB  241     GFAP
```

Gene: Nc73EF, Neural conserved at 73EF, FBgn0010352

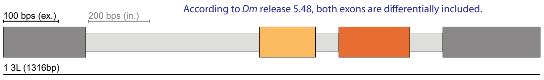


For clarity introns have been scaled down by a factor of 1.56

```
                    10        20        30        40
            ....|....|....|....|....|....|....|....|...
exonA  1    SRGHLASDLDPLGILTREKTVCKDGLARRANEDVLRQHSGFLF
exonB  1    IRGHNIAHLDPLEINTP---------ELPGNSSTKSIYANFSF
```

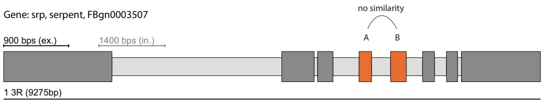
Gene: mtacp1, mitochondrial acyl carrier protein 1, FBgn0011361

According to *Dm* release 5.48, both exons are differentially included.



For clarity introns have been scaled down by a factor of 1.81

Gene: srp, serpent, FBgn0003507

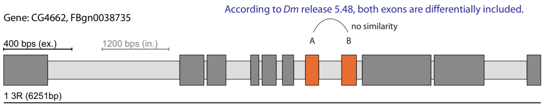


For clarity introns have been scaled down by a factor of 1.53

```
                    10        20        30        40        50        60        70
            ....|....|....|....|....|....|....|....|....|....|....|....|....|....|...
exonA  1    TLFDADYFTEGRECVNCG---------------AISTPLWRRDNTGHYLCNACGLYMKMNGMNRPLIKQPRRL
exonB  1    MAAESGGDFYKPNSFNVGGGGRSKANTSGAASSYSCPGSNATSAATSAVASGTAATAATTLDEHVSRANSRRL
```

Gene: CG4662, FBgn0038735

According to *Dm* release 5.48, both exons are differentially included.

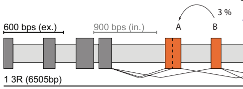


For clarity introns have been scaled down by a factor of 3.07

```
                    10        20
            ....|....|....|....|....|....
exonA  1    LSRLVSYDEQSQMTKPMAVPQAKRGI
exonB  1    MERIFSGAWKEKHGEQEPEEELATPTPLE
```

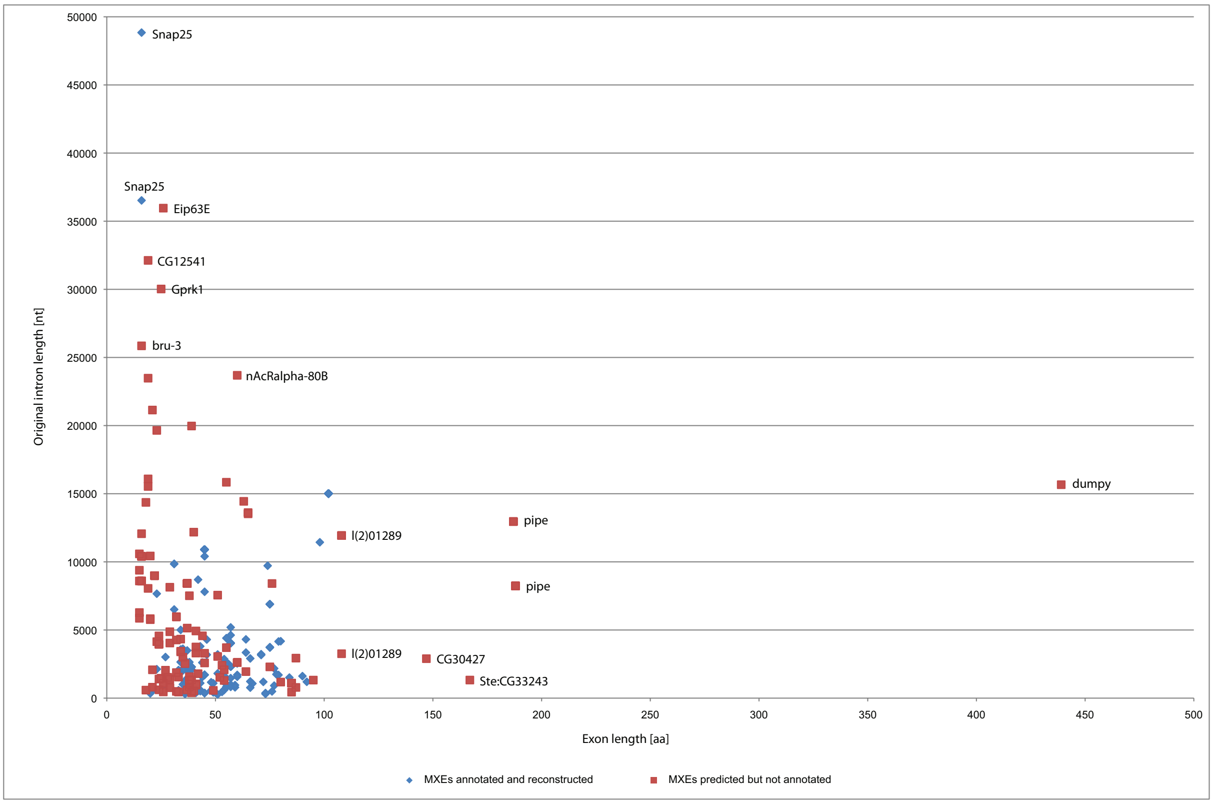
Gene: Vha100-1, Vacuolar H^+ ATPase subunit 100-1, FBgn0028671

According to *Dm* release 5.48, both exons are differentially included.



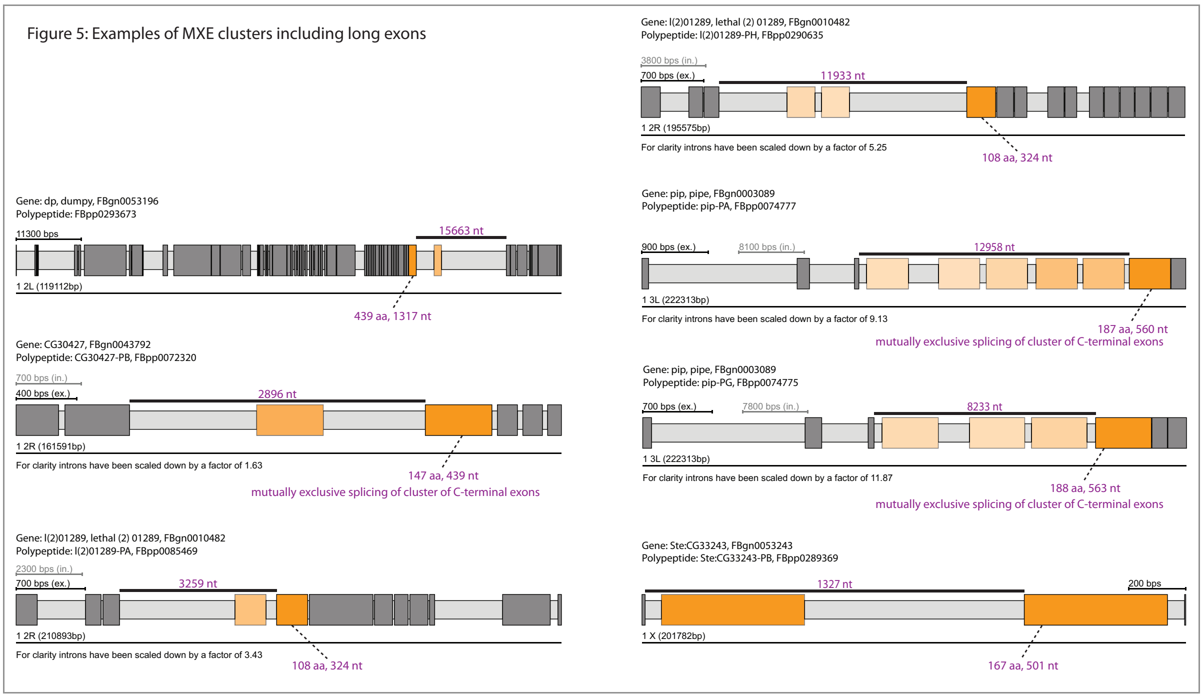
For clarity introns have been scaled down by a factor of 1.43

Supplementary figure 4



Supplementary figure 5

Figure 5: Examples of MXE clusters including long exons


Gene: l(2)01289, lethal (2) 01289, FBgn0010482
Polypeptide: l(2)01289-PH, FBpp0290635
3800 bps (in.)
700 bps (ex.)
11933 nt
1 2R (195575bp)
For clarity introns have been scaled down by a factor of 5.25
108 aa, 324 nt
Gene: dp, dumpy, FBgn0053196
Polypeptide: FBpp0293673
11300 bps
15663 nt
1 2L (119112bp)
439 aa, 1317 nt
Gene: pip, pipe, FBgn0003089
Polypeptide: pip-PA, FBpp0074777
900 bps (ex.)
8100 bps (in.)
12958 nt
1 3L (222313bp)
For clarity introns have been scaled down by a factor of 9.13
187 aa, 560 nt
mutually exclusive splicing of cluster of C-terminal exons
Gene: CG30427, FBgn0043792
Polypeptide: CG30427-PB, FBpp0072320
700 bps (in.)
400 bps (ex.)
2896 nt
1 2R (161591bp)
For clarity introns have been scaled down by a factor of 1.63
147 aa, 439 nt
mutually exclusive splicing of cluster of C-terminal exons
Gene: pip, pipe, FBgn0003089
Polypeptide: pip-PG, FBpp0074775
700 bps (ex.)
7800 bps (in.)
8233 nt
1 3L (222313bp)
For clarity introns have been scaled down by a factor of 11.87
188 aa, 563 nt
mutually exclusive splicing of cluster of C-terminal exons
Gene: l(2)01289, lethal (2) 01289, FBgn0010482
Polypeptide: l(2)01289-PA, FBpp0085469
2300 bps (in.)
700 bps (ex.)
3259 nt
1 2R (210893bp)
For clarity introns have been scaled down by a factor of 3.43
108 aa, 324 nt
Gene: Ste:CG33243, FBgn0053243
Polypeptide: Ste:CG33243-PB, FBpp0289369
1327 nt
200 bps
1 X (201782bp)
167 aa, 501 nt

Supplementary figure 6

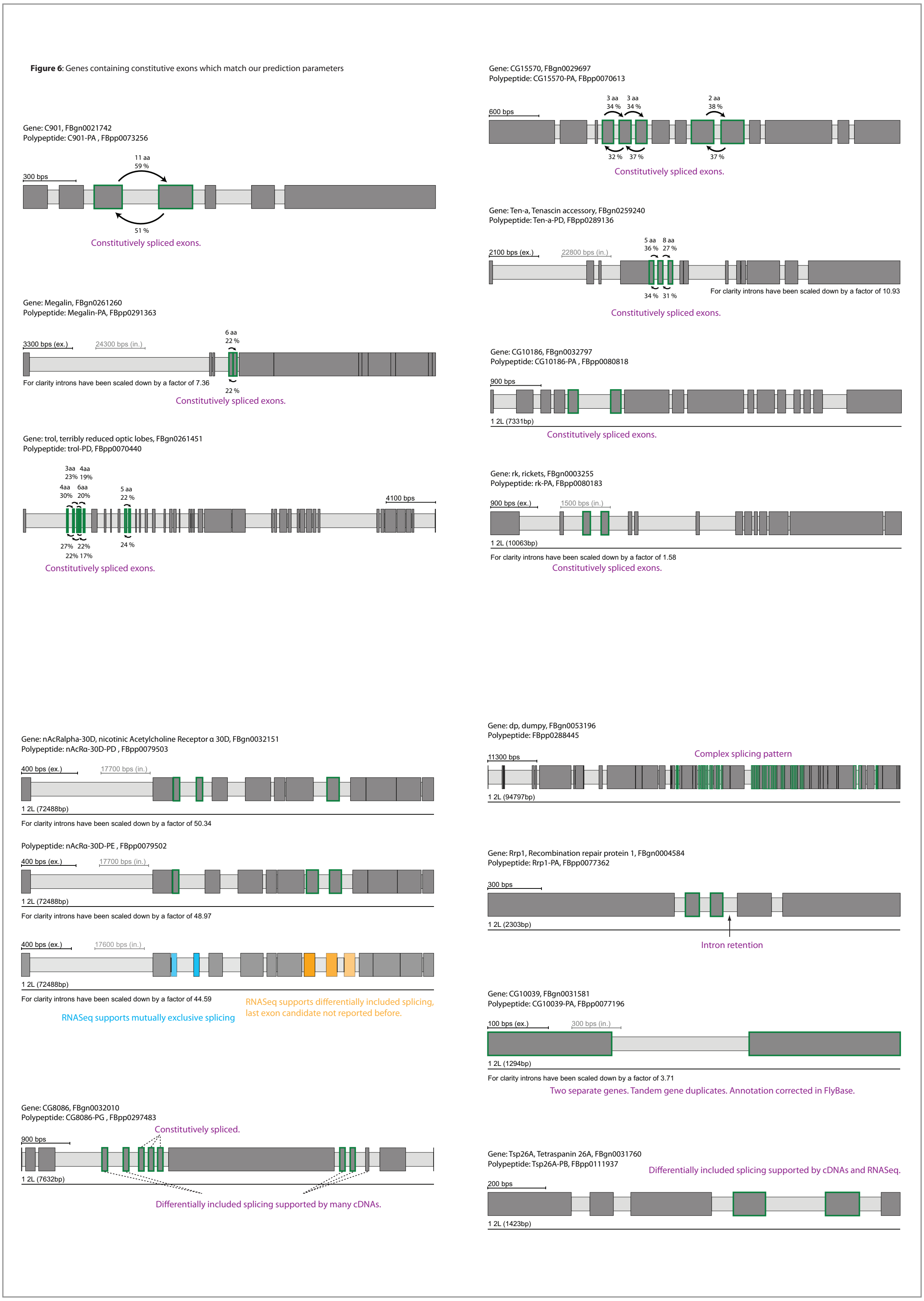
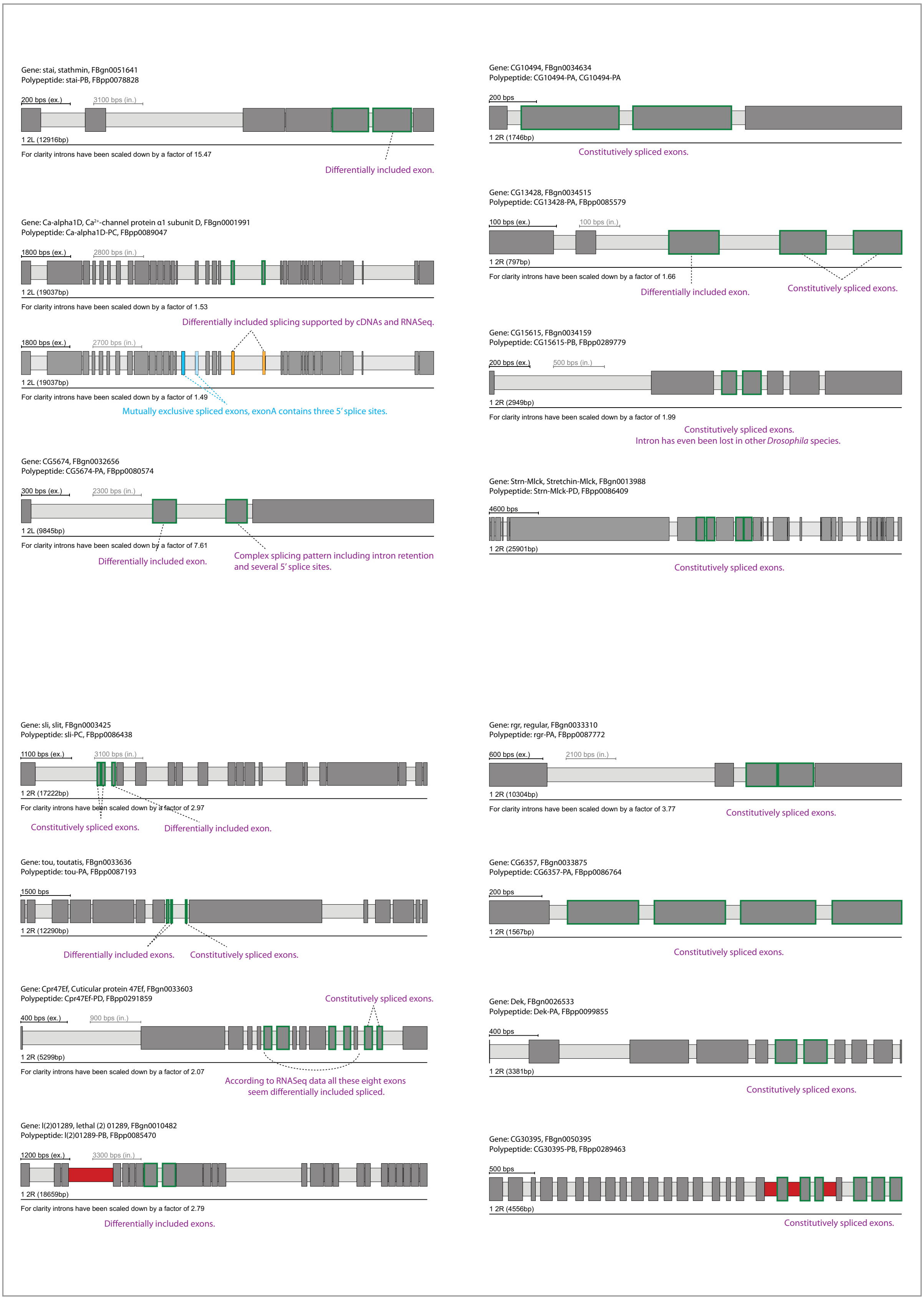
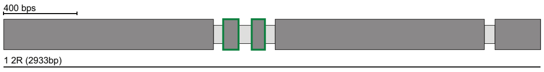


Figure 6: Genes containing constitutive exons which match our prediction parameters

Gene: C901, FBgn0021742
Polypeptide: C901-PA , FBpp0073256
300 bps
11 aa
59 %
51 %
Constitutively spliced exons.
Gene: Megalin, FBgn0261260
Polypeptide: Megalin-PA, FBpp0291363
3300 bps (ex.)
24300 bps (in.)
6 aa
22 %
22 %
For clarity introns have been scaled down by a factor of 7.36
Constitutively spliced exons.
Gene: trol, terribly reduced optic lobes, FBgn0261451
Polypeptide: trol-PD, FBpp0070440
3aa 4aa
23% 19%
4aa 6aa
30% 20%
5 aa
22 %
4100 bps
27% 22%
22% 17%
24 %
Constitutively spliced exons.
Gene: CG15570, FBgn0029697
Polypeptide: CG15570-PA, FBpp0070613
600 bps
3 aa 3 aa
34 % 34 %
2 aa
38 %
32 % 37 %
37 %
Constitutively spliced exons.
Gene: Ten-a, Tenascin accessory, FBgn0259240
Polypeptide: Ten-a-PD, FBpp0289136
2100 bps (ex.)
22800 bps (in.)
5 aa 8 aa
36 % 27 %
34 % 31 %
For clarity introns have been scaled down by a factor of 10.93
Constitutively spliced exons.
Gene: CG10186, FBgn0032797
Polypeptide: CG10186-PA , FBpp0080818
900 bps
1 2L (7331bp)
Constitutively spliced exons.
Gene: rk, rickets, FBgn0003255
Polypeptide: rk-PA, FBpp0080183
900 bps (ex.)
1500 bps (in.)
1 2L (10063bp)
For clarity introns have been scaled down by a factor of 1.58
Constitutively spliced exons.
Gene: nAcRalpha-30D, nicotinic Acetylcholine Receptor α 30D, FBgn0032151
Polypeptide: nAcRα-30D-PD , FBpp0079503
400 bps (ex.)
17700 bps (in.)
1 2L (72488bp)
For clarity introns have been scaled down by a factor of 50.34
Polypeptide: nAcRα-30D-PE , FBpp0079502
400 bps (ex.)
17700 bps (in.)
1 2L (72488bp)
For clarity introns have been scaled down by a factor of 48.97
400 bps (ex.)
17600 bps (in.)
1 2L (72488bp)
For clarity introns have been scaled down by a factor of 44.59
RNASeq supports mutually exclusive splicing
RNASeq supports differentially included splicing,
last exon candidate not reported before.
Gene: CG8086, FBgn0032010
Polypeptide: CG8086-PG , FBpp0297483
900 bps
Constitutively spliced.
1 2L (7632bp)
Differentially included splicing supported by many cDNAs.
Gene: dp, dumpy, FBgn0053196
Polypeptide: FBpp0288445
11300 bps
Complex splicing pattern
1 2L (94797bp)
Gene: Rrp1, Recombination repair protein 1, FBgn0004584
Polypeptide: Rrp1-PA, FBpp0077362
300 bps
1 2L (2303bp)
Intron retention
Gene: CG10039, FBgn0031581
Polypeptide: CG10039-PA, FBpp0077196
100 bps (ex.)
300 bps (in.)
1 2L (1294bp)
For clarity introns have been scaled down by a factor of 3.71
Two separate genes. Tandem gene duplicates. Annotation corrected in FlyBase.
Gene: Tsp26A, Tetraspanin 26A, FBgn0031760
Polypeptide: Tsp26A-PB, FBpp0111937
Differentially included splicing supported by cDNAs and RNASeq.
200 bps
1 2L (1423bp)

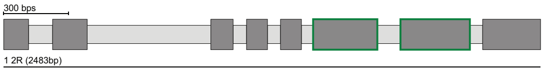

Gene: stai, stathmin, FBgn0051641
Polypeptide: stai-PB, FBpp0078828
200 bps (ex.)
3100 bps (in.)
1 2L (12916bp)
For clarity introns have been scaled down by a factor of 15.47
Differentially included exon.
Gene: Ca-alpha1D, Ca^{2+}-channel protein α1 subunit D, FBgn0001991
Polypeptide: Ca-alpha1D-PC, FBpp0089047
1800 bps (ex.)
2800 bps (in.)
1 2L (19037bp)
For clarity introns have been scaled down by a factor of 1.53
Differentially included splicing supported by cDNAs and RNASeq.
1800 bps (ex.)
2700 bps (in.)
1 2L (19037bp)
For clarity introns have been scaled down by a factor of 1.49
Mutually exclusive spliced exons, exonA contains three 5' splice sites.
Gene: CG5674, FBgn0032656
Polypeptide: CG5674-PA, FBpp0080574
300 bps (ex.)
2300 bps (in.)
1 2L (9845bp)
For clarity introns have been scaled down by a factor of 7.61
Differentially included exon.
Complex splicing pattern including intron retention and several 5' splice sites.
Gene: sli, slit, FBgn0003425
Polypeptide: sli-PC, FBpp0086438
1100 bps (ex.)
3100 bps (in.)
1 2R (17222bp)
For clarity introns have been scaled down by a factor of 2.97
Constitutively spliced exons.
Differentially included exon.
Gene: tou, toutatis, FBgn0033636
Polypeptide: tou-PA, FBpp0087193
1500 bps
1 2R (12290bp)
Differentially included exons.
Constitutively spliced exons.
Gene: Cpr47Ef, Cuticular protein 47Ef, FBgn0033603
Polypeptide: Cpr47Ef-PD, FBpp0291859
Constitutively spliced exons.
400 bps (ex.)
900 bps (in.)
1 2R (5299bp)
For clarity introns have been scaled down by a factor of 2.07
According to RNASeq data all these eight exons seem differentially included spliced.
Gene: l(2)01289, lethal (2) 01289, FBgn0010482
Polypeptide: l(2)01289-PB, FBpp0085470
1200 bps (ex.)
3300 bps (in.)
1 2R (18659bp)
For clarity introns have been scaled down by a factor of 2.79
Differentially included exons.
Gene: CG10494, FBgn0034634
Polypeptide: CG10494-PA, CG10494-PA
200 bps
1 2R (1746bp)
Constitutively spliced exons.
Gene: CG13428, FBgn0034515
Polypeptide: CG13428-PA, FBpp0085579
100 bps (ex.)
100 bps (in.)
1 2R (797bp)
For clarity introns have been scaled down by a factor of 1.66
Differentially included exon.
Constitutively spliced exons.
Gene: CG15615, FBgn0034159
Polypeptide: CG15615-PB, FBpp0289779
200 bps (ex.)
500 bps (in.)
1 2R (2949bp)
For clarity introns have been scaled down by a factor of 1.99
Constitutively spliced exons.
Intron has even been lost in other *Drosophila* species.
Gene: Strn-Mlck, Stretchin-Mlck, FBgn0013988
Polypeptide: Strn-Mlck-PD, FBpp0086409
4600 bps
1 2R (25901bp)
Constitutively spliced exons.
Gene: rgr, regular, FBgn0033310
Polypeptide: rgr-PA, FBpp0087772
600 bps (ex.)
2100 bps (in.)
1 2R (10304bp)
For clarity introns have been scaled down by a factor of 3.77
Constitutively spliced exons.
Gene: CG6357, FBgn0033875
Polypeptide: CG6357-PA, FBpp0086764
200 bps
1 2R (1567bp)
Constitutively spliced exons.
Gene: Dek, FBgn0026533
Polypeptide: Dek-PA, FBpp0099855
400 bps
1 2R (3381bp)
Constitutively spliced exons.
Gene: CG30395, FBgn0050395
Polypeptide: CG30395-PB, FBpp0289463
500 bps
1 2R (4556bp)
Constitutively spliced exons.

Gene: CG9861, FBgn0034844
Polypeptide: CG9861-PA, FBpp0071911



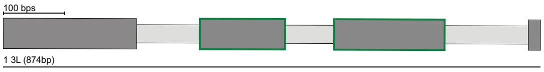
Constitutively spliced exons.

Gene: Mlp60A, Muscle LIM protein at 60A, FBgn0259209
Polypeptide: Mlp60A-PB, FBpp0288975



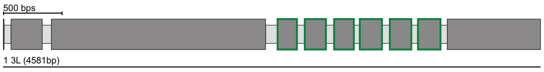
Constitutively spliced exons.

Gene: miple, FBgn0027111
Polypeptide: miple-PA, FBpp0072405



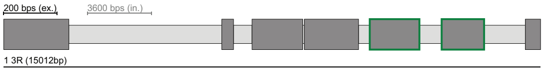
Constitutively spliced exons.

Gene: CG6947, FBgn0036233
Polypeptide: CG6947-PA, FBpp0075777



Constitutively spliced exons.

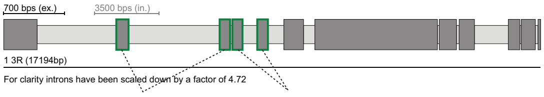
Gene: tau, FBgn0051057
Polypeptide: tau-PA, FBpp0084567



For clarity introns have been scaled down by a factor of 15.01

Differentially included exons.

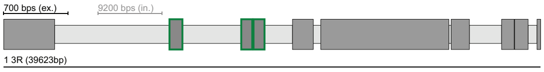
Gene: LpR1, Lipophorin receptor 1, FBgn0066101
Polypeptide: LpR1-PK, FBpp0290685



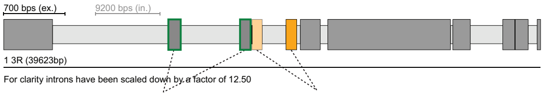
For clarity introns have been scaled down by a factor of 4.72

Constitutively spliced exons. Differentially included exons.

Gene: LpR2, Lipophorin receptor 2, FBgn0051092
Polypeptide: LpR2-PA, FBpp0084301



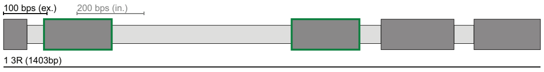
For clarity introns have been scaled down by a factor of 13.16



For clarity introns have been scaled down by a factor of 12.50

Constitutively spliced exons. Mutually exclusive exons

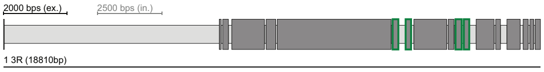
Gene: CG33483, FBgn0053483
Polypeptide: CG33483-PB, FBpp0292484



For clarity introns have been scaled down by a factor of 1.30

Constitutively spliced exons.

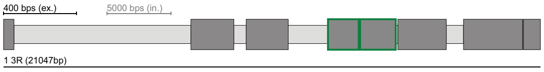
Gene: Ppn, Papilin, FBgn0003137
Polypeptide: Ppn-PE, FBpp0291051



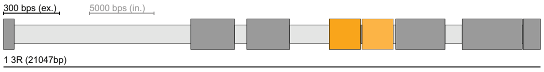
For clarity introns have been scaled down by a factor of 1.23

Differentially included exons.

Gene: betaTub97EF, β-Tubulin at 97EF, FBgn0003890
Polypeptide: βTub97EF-PA , FBpp0084630



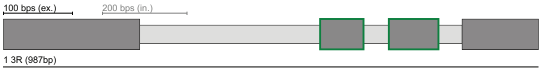
For clarity introns have been scaled down by a factor of 14.22



For clarity introns have been scaled down by a factor of 14.49

Mutually exclusive exons.
Misannotated as constitutive exons.

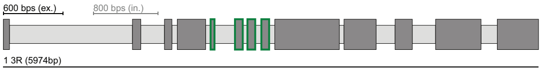
Gene: CG31406, FBgn0051406
Polypeptide: CG31406-PA, FBpp0081713



For clarity introns have been scaled down by a factor of 1.63

Constitutively spliced exons.

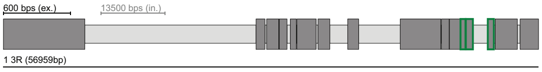
Gene: CG9297, FBgn0038181
Polypeptide: CG9297-PA, FBpp0082295



For clarity introns have been scaled down by a factor of 1.24

Constitutively spliced exons.

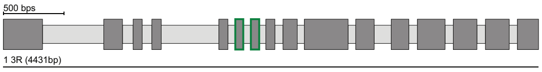
Gene: CG42342, FBgn0259244
Polypeptide: CG42342-PD, FBpp0289172



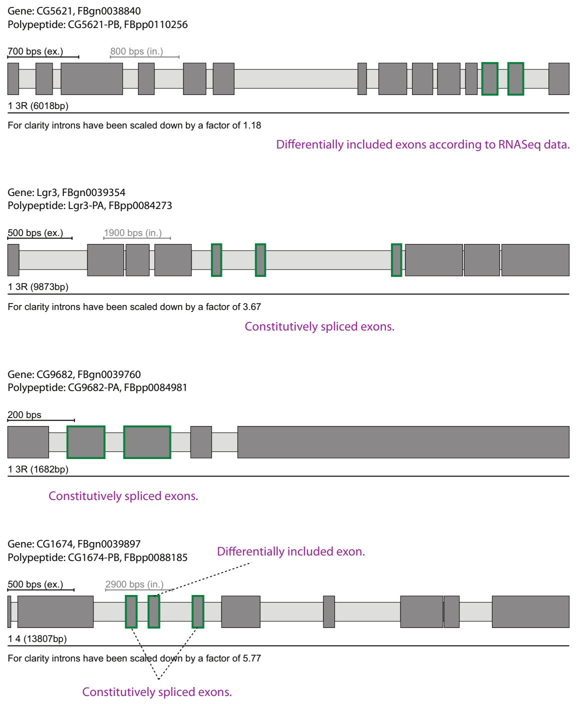
For clarity introns have been scaled down by a factor of 23.57

Constitutively spliced exons.

Gene: Fsh, Fsh-Tsh-like receptor, FBgn0016650
Polypeptide: Fsh-PA, FBpp0082933



Constitutively spliced exons.


Gene: CG5621, FBgn0038840
Polypeptide: CG5621-PB, FBpp0110256
700 bps (ex.)
800 bps (in.)
1 3R (6018bp)
For clarity introns have been scaled down by a factor of 1.18
Differentially included exons according to RNASeq data.
Gene: Lgr3, FBgn0039354
Polypeptide: Lgr3-PA, FBpp0084273
500 bps (ex.)
1900 bps (in.)
1 3R (9873bp)
For clarity introns have been scaled down by a factor of 3.67
Constitutively spliced exons.
Gene: CG9682, FBgn0039760
Polypeptide: CG9682-PA, FBpp0084981
200 bps
1 3R (1682bp)
Constitutively spliced exons.
Gene: CG1674, FBgn0039897
Polypeptide: CG1674-PB, FBpp0088185
Differentially included exon.
500 bps (ex.)
2900 bps (in.)
1 4 (13807bp)
For clarity introns have been scaled down by a factor of 5.77
Constitutively spliced exons.

Supplementary figure 7

Figure 7: RNA secondary structure prediction for gene CG14608 of *Drosophila melanogaster*

RNA secondary structure found in first intron and all internal introns (*Drosophila melanogaster* and related species):
(acceptor motif: TCTTTTATATTTGTTGCCTTGTCACCACATCCTCTTGAAAAATCGCATTCATTGGCTTCATGAAATTTTATCATGACGTTTTATTCAAAA)
(selector motif: TTTTGAATAAAACGTCATGATAAAATTTCATGAAGCCAATGAATGCGATTTTTCAAGAGGATGTGGTGACAAGGCAACAAATATAAAAGA)

Cross-species, intron 1 to 2a:

```
Species          Energy          Intron  RNA secondary structure prediction
                                         .......(((((((....((((((.((((....((((((....(((......((..((((.((((((......((((((((....
D. melanogaster  -19.2 Kcal/Mol  1:      ATCTGTTGCCTTTCACTAGGCATCATCTTACTTAGATTTATGGCATGCATTTTCCACATAATTTTATTACCACTTTTAGTC----
D. sechellia     -23.7 Kcal/Mol  1:      ATCTGTTGCCTTTCACTAGGCATCATCTTACTTAGATCTATTGCATGCATTTTCCACGTAATTTTTTTACCACTTTTGGTCAAGA
D. simulans      -20.2 Kcal/Mol  1:      ATCTGTTGCCTTTCACTAGGCATCATCTTACTTAGATTTATTGCATGCATTTTCCACGTAATTTTATTACCACTTTTGGTC----

Species          Intron  RNA secondary structure prediction
                         ....)))...)))))......))))))))))..)).....)))))))))...))))))))))...)).)))))..............
D. melanogaster  2a:     TTTTGATTCATAAAACAGCATATAAAATATGAAGACACTTATGAGATTTTTCAAGAGGTGTC--TGACAAGGCTAAAATATAAAAG
D. sechellia     2a:     TTTTGATTCATAAAACATCATATAAAATATGAAGACACTAATGGGATTTTTCAAGAGGTGTCGGGGACAAGGCTAAAATATAAAAG
D. simulans      2a:     TTTTGATTCATAAAACATCATATAAAATATGAAGACACTAATGGGATTTTTCAAGAGGTGTC--GGACAAGGCTAAAATATAAAAG
```

(Energy of ensemble: -21.38; Frequency of structure in ensemble: 0.00255816)

Cross-species, intron 1 to 2b:

```
Species          Energy          Intron  RNA secondary structure prediction
                                         .(((...(((..(((((.....((.((((.....(((((((((.(((.(((...((.....................((((....
D. melanogaster  -16.3 Kcal/Mol  1:      ATCTGTTGCCTTTCACTAGGCATCATCTTACTTAGATTTATGGCATGCATTTTCCACATAATTTTATTACCACTTTTAGTC----
D. sechellia     -17.9 Kcal/Mol  1:      ATCTGTTGCCTTTCACTAGGCATCATCTTACTTAGATCTATTGCATGCATTTTCCACGTAATTTTTTTACCACTTTTGGTCAAGA
D. simulans      -20.1 Kcal/Mol  1:      ATCTGTTGCCTTTCACTAGGCATCATCTTACTTAGATTTATTGCATGCATTTTCCACGTAATTTTATTACCACTTTTGGTC----

Species          Intron  RNA secondary structure prediction
                         ....))))))...)))..))).)))))))))....)))).))))))).)))..)))...........
D. melanogaster  2b:     AAACGACTTGATTAATATCATCCAATGAATCCGTCGAGAGGAATTGGCGGCCAAGAAATATAA----
D. sechellia     2b:     AAGCGACTTGATTAATCTCATCCAATGAATCCGACGAGAGGAAGTGGCGGCCAAGAAACATAATTTA
D. simulans      2b:     AAGCGACTTGATTAATCTCATCCAATGAATCCGACGAGAGGAAGTGGCGGCCAAGAAATATAA----
```

(Energy of ensemble: -17.52; Frequency of structure in ensemble: 0.0140401)

RNA secondary structure found in all internal introns and last intron (*Drosophila melanogaster* and related species):
(selector motif: TTTTGAATAAAACGTCATGATAAAATTTCATGAAGCCAATGAATGCGATTTTTCAAGAGGATGTGGTGACAAGGCAACAAATATAAAAGA)
(acceptor motif: TCTTTTATATTTGTTGCCTTGTCACCACATCCTCTTGAAAAATCGCATTCATTGGCTTCATGAAATTTTATCATGACGTTTTATTCAAAA)

Cross-species, intron 2a to 2c:

```
Species          Energy          Intron  RNA secondary structure prediction
                                         ...............................(((((......(((((.....((((((............................
D. melanogaster  -2.34 Kcal/Mol  2a:     TTTTGATTCATAAAACAGCATATAAAATATGAAGACACTTATGAGATTTTTCAAGAGGTGTC--TGACAAGGCTAAAATATAAAAG
D. sechellia     -5.54 Kcal/Mol  2a:     TTTTGATTCATAAAACATCATATAAAATATGAAGACACTAATGGGATTTTTCAAGAGGTGTCGGGGACAAGGCTAAAATATAAAAG
D. simulans      -5.64 Kcal/Mol  2a:     TTTTGATTCATAAAACATCATATAAAATATGAAGACACTAATGGGATTTTTCAAGAGGTGTC--GGACAAGGCTAAAATATAAAAG

Species          Intron  RNA secondary structure prediction
                         ..............)))))).)))))...)))))......................................................
D. melanogaster  2c:     --------------CTTTTTTATTTCTGTGTCTT-------TCTTTAATTGTTTTTATGCTTTAAAATTT---AATTATTGCATTT--
D. sechellia     2c:     TTTTAAAGTGTGCATCTCTTTATCCTCGCGGCTTAGAAAACTTTTAAATTGTATCA-------TAAATTTTTACATCAATATATTCGA
D. simulans      2c:     TTAGATATTAAGT-TCTCTTTATCCTTCCATCTTTAAAAACCTTTAATATGTATCAAGTATTTGATATTTTTACATCAATTTATTCGA
```

(Energy of ensemble: -6.0; Frequency of structure in ensemble: 0.214515)

Cross-species, intron 2b to 2c:

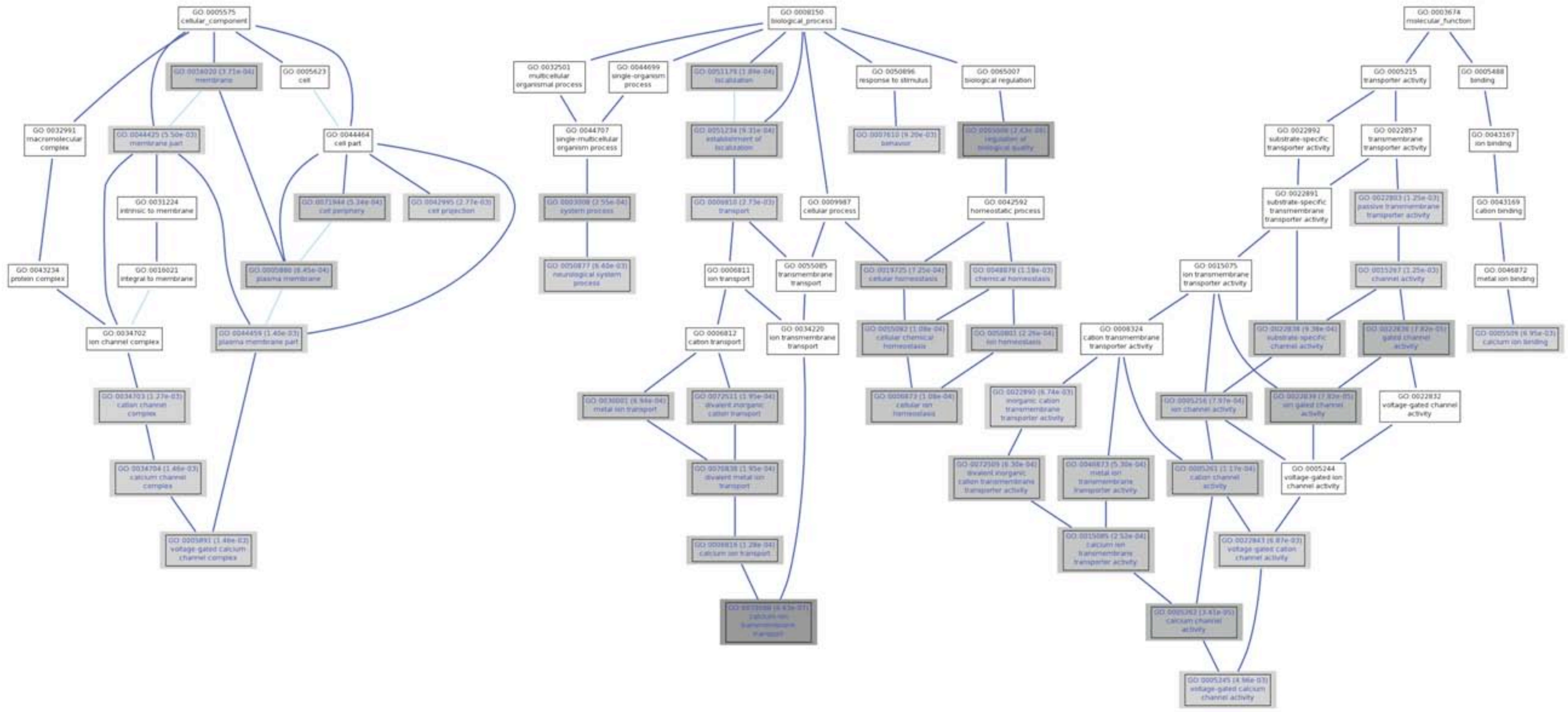
```
Species          Energy          Intron  RNA secondary structure prediction
                                         ..................................((.(((((...((.((.................
D. melanogaster  8.43 Kcal/Mol   2b:     AAACGACTTGATTAATATCATCCAATGAATCCGTCGAGAGGAATTGGCGGCCAAGAAATATAA----
D. sechellia     -7.27 Kcal/Mol  2b:     AAGCGACTTGATTAATCTCATCCAATGAATCCGACGAGAGGAAGTGGCGGCCAAGAAACATAATTTA
D. simulans      -2.77 Kcal/Mol  2b:     AAGCGACTTGATTAATCTCATCCAATGAATCCGACGAGAGGAAGTGGCGGCCAAGAAATATAA----

Species          Intron  RNA secondary structure prediction
                         ..............)).))...))))).))..........................................................
D. melanogaster  2c:     --------------CTTTTTTATTTCTGTGTCTT-------TCTTTAATTGTTTTTATGCTTTAAAATTT---AATTATTGCATTT--
D. sechellia     2c:     TTTTAAAGTGTGCATCTCTTTATCCTCGCGGCTTAGAAAACTTTTAAATTGTATCA-------TAAATTTTTACATCAATATATTCGA
D. simulans      2c:     TTAGATATTAAGT-TCTCTTTATCCTTCCATCTTTAAAAACCTTTAATATGTATCAAGTATTTGATATTTTTACATCAATTTATTCGA
```

(Energy of ensemble: -2.03; Frequency of structure in ensemble: 0.104356)

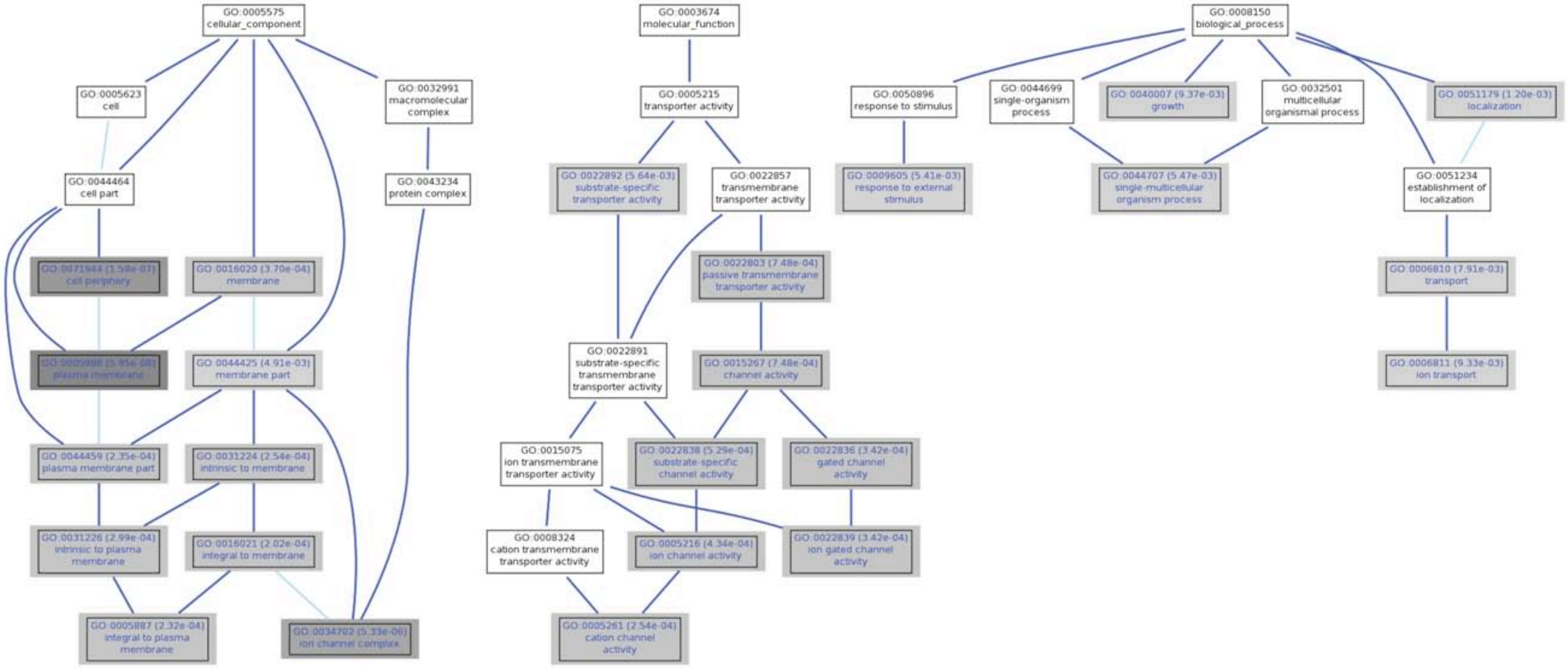
Supplementary figure 8

Figure 8: Gene Ontology (GO) term enrichment analysis of genes containing MXEs, which are annotated and reconstructed.

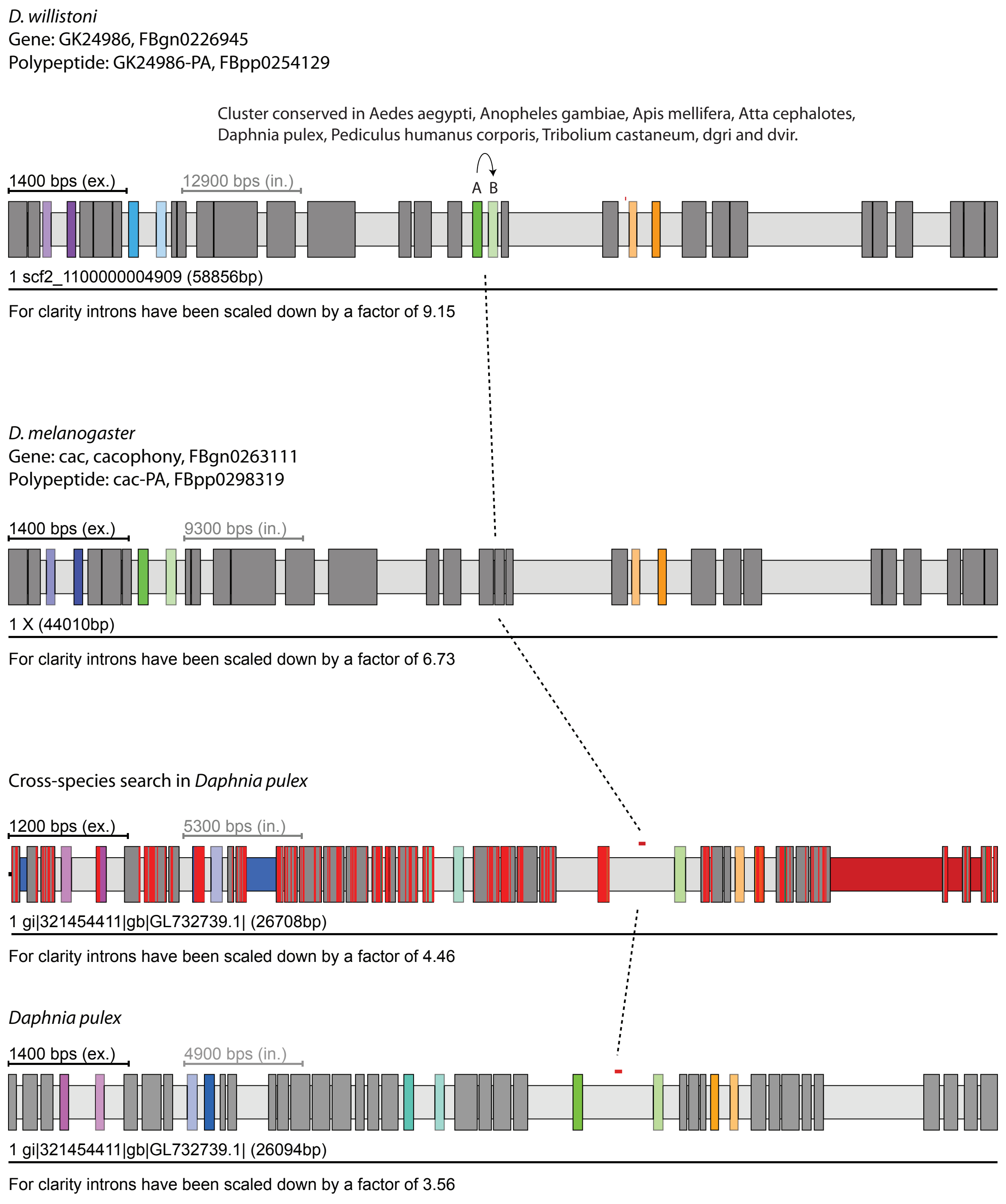

GO:0005575 cellular_component
GO:0016020 (3.71e-04) membrane
GO:0005623 cell
GO:0032991 macromolecular complex
GO:0044425 (5.50e-03) membrane part
GO:0044464 cell part
GO:0031224 intrinsic to membrane
GO:0071944 (5.34e-04) cell periphery
GO:0042995 (2.77e-03) cell projection
GO:0043234 protein complex
GO:0016021 integral to membrane
GO:0005886 (8.45e-04) plasma membrane
GO:0034702 ion channel complex
GO:0044459 (1.40e-03) plasma membrane part
GO:0034703 (1.27e-03) cation channel complex
GO:0034704 (1.46e-03) calcium channel complex
GO:0005891 (1.46e-03) voltage-gated calcium channel complex
GO:0008150 biological_process
GO:0032501 multicellular organismal process
GO:0044699 single-organism process
GO:0051179 (1.89e-04) localization
GO:0050896 response to stimulus
GO:0065007 biological regulation
GO:0044707 single-multicellular organism process
GO:0051234 (9.31e-04) establishment of localization
GO:0007610 (9.20e-03) behavior
GO:0065008 (2.43e-08) regulation of biological quality
GO:0003008 (2.55e-04) system process
GO:0006810 (2.73e-03) transport
GO:0009987 cellular process
GO:0042592 homeostatic process
GO:0050877 (6.40e-03) neurological system process
GO:0006811 ion transport
GO:0055085 transmembrane transport
GO:0019725 (7.25e-04) cellular homeostasis
GO:0048878 (1.18e-03) chemical homeostasis
GO:0006812 cation transport
GO:0034220 ion transmembrane transport
GO:0055082 (1.08e-04) cellular chemical homeostasis
GO:0050801 (2.29e-04) ion homeostasis
GO:0030001 (6.94e-04) metal ion transport
GO:0072511 (1.95e-04) divalent inorganic cation transport
GO:0006873 (1.08e-04) cellular ion homeostasis
GO:0070838 (1.95e-04) divalent metal ion transport
GO:0006816 (1.28e-04) calcium ion transport
GO:0070588 (6.43e-07) calcium ion transmembrane transport
GO:0003674 molecular_function
GO:0005215 transporter activity
GO:0005488 binding
GO:0022892 substrate-specific transporter activity
GO:0022857 transmembrane transporter activity
GO:0043167 ion binding
GO:0022891 substrate-specific transmembrane transporter activity
GO:0022803 (1.25e-03) passive transmembrane transporter activity
GO:0043169 cation binding
GO:0015075 ion transmembrane transporter activity
GO:0015267 (1.25e-03) channel activity
GO:0046872 metal ion binding
GO:0008324 cation transmembrane transporter activity
GO:0022838 (9.38e-04) substrate-specific channel activity
GO:0022836 (7.82e-05) gated channel activity
GO:0005509 (6.95e-03) calcium ion binding
GO:0022890 (6.74e-03) inorganic cation transmembrane transporter activity
GO:0005216 (7.97e-04) ion channel activity
GO:0022839 (7.82e-05) ion gated channel activity
GO:0022832 voltage-gated channel activity
GO:0072509 (6.30e-04) divalent inorganic cation transmembrane transporter activity
GO:0046873 (5.30e-04) metal ion transmembrane transporter activity
GO:0005261 (1.17e-04) cation channel activity
GO:0005244 voltage-gated ion channel activity
GO:0015085 (2.52e-04) calcium ion transmembrane transporter activity
GO:0022843 (6.87e-03) voltage-gated cation channel activity
GO:0005262 (3.41e-05) calcium channel activity
GO:0005245 (4.96e-03) voltage-gated calcium channel activity

Supplementary figure 9

Figure 9: Gene Ontology (GO) term enrichment analysis of genes containing MXEs, which were predicted but not annotated.


GO:0005575 cellular_component
GO:0005623 cell
GO:0044464 cell part
GO:0032991 macromolecular complex
GO:0043234 protein complex
GO:0071944 (1.58e-07) cell periphery
GO:0016020 (3.70e-04) membrane
GO:0005886 (5.95e-08) plasma membrane
GO:0044425 (4.91e-03) membrane part
GO:0044459 (2.35e-04) plasma membrane part
GO:0031224 (2.54e-04) intrinsic to membrane
GO:0031226 (2.99e-04) intrinsic to plasma membrane
GO:0016021 (2.02e-04) integral to membrane
GO:0005887 (2.32e-04) integral to plasma membrane
GO:0034702 (5.33e-06) ion channel complex
GO:0003674 molecular_function
GO:0005215 transporter activity
GO:0022892 (5.64e-03) substrate-specific transporter activity
GO:0022857 transmembrane transporter activity
GO:0022803 (7.48e-04) passive transmembrane transporter activity
GO:0022891 substrate-specific transmembrane transporter activity
GO:0015267 (7.48e-04) channel activity
GO:0015075 ion transmembrane transporter activity
GO:0022838 (5.29e-04) substrate-specific channel activity
GO:0022836 (3.42e-04) gated channel activity
GO:0008324 cation transmembrane transporter activity
GO:0005216 (4.34e-04) ion channel activity
GO:0022839 (3.42e-04) ion gated channel activity
GO:0005261 (2.54e-04) cation channel activity
GO:0008150 biological_process
GO:0050896 response to stimulus
GO:0044699 single-organism process
GO:0040007 (9.37e-03) growth
GO:0032501 multicellular organismal process
GO:0051179 (1.20e-03) localization
GO:0009605 (5.41e-03) response to external stimulus
GO:0044707 (5.47e-03) single-multicellular organism process
GO:0051234 establishment of localization
GO:0006810 (7.91e-03) transport
GO:0006811 (9.33e-03) ion transport

Supplementary figure 10

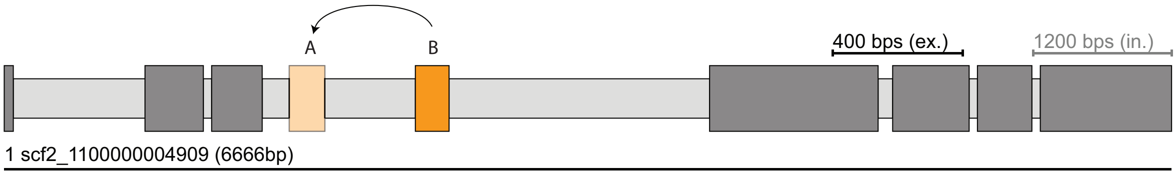

D. willistoni
Gene: GK24986, FBgn0226945
Polypeptide: GK24986-PA, FBpp0254129
Cluster conserved in Aedes aegypti, Anopheles gambiae, Apis mellifera, Atta cephalotes, Daphnia pulex, Pediculus humanus corporis, Tribolium castaneum, dgri and dvir.
1400 bps (ex.)
12900 bps (in.)
A
B
1 scf2_1100000004909 (58856bp)
For clarity introns have been scaled down by a factor of 9.15
D. melanogaster
Gene: cac, cacophony, FBgn0263111
Polypeptide: cac-PA, FBpp0298319
1400 bps (ex.)
9300 bps (in.)
1 X (44010bp)
For clarity introns have been scaled down by a factor of 6.73
Cross-species search in *Daphnia pulex*
1200 bps (ex.)
5300 bps (in.)
1 gi|321454411|gb|GL732739.1| (26708bp)
For clarity introns have been scaled down by a factor of 4.46
Daphnia pulex
1400 bps (ex.)
4900 bps (in.)
1 gi|321454411|gb|GL732739.1| (26094bp)
For clarity introns have been scaled down by a factor of 3.56

```
                   10        20        30
           ....|....|....|....|....|....|....|..
dwilExonA  YHNQSEMYEKSLKYINMGFTGMFSVETVLKIIGFGVK
dwilExonB  FHDAPPTHSDILTYMNLVFTLFFLMETILKLIAFGCT
dmelExon   YHNQGDMYEKSLKYINMGFTGMFSVETVLKIIGFGVK
DapExonA   YYKQSVLYKETLHYMNTAFTALFSIECMLKIISFGVR
DapExonB   YYEAPPALVDILAYMNLIFTMLFSLECILKLAAFGIK
```

Supplementary figure 11

D. willistoni
Gene: GK25120, FBgn0227079
Polypeptide: GK25120-PA, FBpp0254263

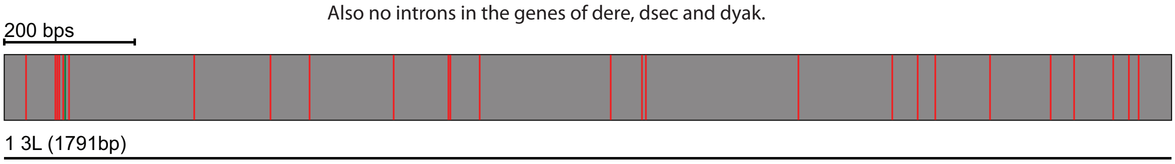
Cluster conserved in Anopheles gambiae, Apis mellifera, Atta cephalotes, Pediculus humanus corporis, Tribolium castaneum, dgri, dper and dvir.



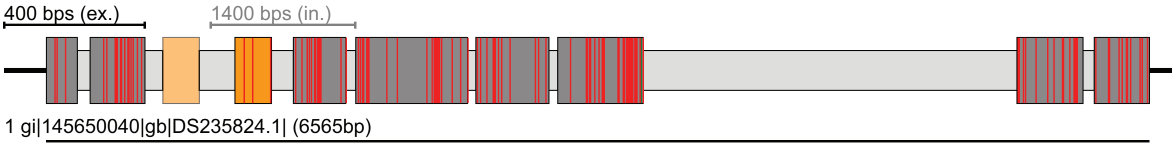
For clarity introns have been scaled down by a factor of 2.81

Cross-species search in *D. melanogaster*
Gene: Rop, Ras opposite, FBgn0004574
Polypeptide: Rop-PA, FBpp0073119

Also no introns in the genes of dere, dsec and dyak.



Cross-species search in *Pediculus humanus corporis*



For clarity introns have been scaled down by a factor of 3.41

```
                       10        20        30        40        50
               ....|....|....|....|....|....|....|....|....|....|....|.
dwilExonA      --------AIPPRIFEMLQSHKDICRRYVRTCKEINISFLAYEAQ-----------
dwilExonB      --------VCPEELFNDL--CKSCAARKIKTLKEINIAFLPYECQ-----------
dmelExon       YAHVFFTEVCPEELFNDL--CKSCAAGKIKTLKEINIAFLPYECQVFSLDSPDTFQ
PdcExonA       --------ACNDELFKEI--SHARVAKFIKTLKEINIAFIPFEEQ-----------
PdcExonB       --------VCPEELFNEL--CKSCAAKKIKTLKEINIAFLPYESQ-----------
```

A3 Supplementary tables

Supplementary table 1 | Mutually exclusive exons in *Drosophila melanogaster*

		Exons matching prediction criteria of MXEs					
	Annotated MXEs	**Annotated and reconstructed MXEs**		**Predicted MXEs**		**Exons annotated as constitutive or differentially included**	
			Cross / EST evidence		Cross / EST evidence		Annotated as MXEs in r5.48
Initial	660	31	28 / 17	65	47 / 20	2	0
3’-terminal	376	42	36 / 22	55	45 / 25	8	0
Internal	261	218	206 / 56	419	321 / 88	159	5
Sum	1297	291	270 / 95	539	413 / 133	169	5

A4 List of figures

A5 List of tables