Enhanced genome assembly and a new official gene set for Tribolium castaneum - from a draft to a reference genome

Nicolae Herndon

Department of Computer Science, East Carolina University Greenville, NC 27858

Jennifer Shelton

Division of Biology, Kansas State University, Manhattan, KS 66506

Lizzy Gerischer

Institut für Mathematik und Informatik, Universität Greifswald, Greifswald, Germany

Panos Ioannidis

Department of Genetic Medicine and Development, University of Geneva Medical School and Swiss Institute of Bioinformatics, Geneva, 1211 Switzerland

Maria Ninova

Faculty of Biology, Medicine and Health, University of Manchester, Michael Smith Building, Oxford Road, Manchester, M13 9PT, UK

Jürgen Dönitz

Department of Evolutionary Developmental Genetics, GZMB, University of Göttingen, Justus-von-Liebig-Weg 11, 37077 Göttingen, Germany

Robert M. Waterhouse

Department of Ecology and Evolution, University of Lausanne and Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

Chun Liang

Department of Biology, Miami University, Oxford, OH 45056, USA

Carsten Damm

Institut für Informatik, Fakultät für Mathematik und Informatik, Georg-August-Universität Göttingen, Goldschmidtstr. 7, D-37077 Göttingen, Germany

Janna Siemanowski

Department of Evolutionary Developmental Genetics, GZMB, University of Göttingen, Justus-von-Liebig-Weg 11, 37077 Göttingen, Germany

Peter Kitzmann

Department of Evolutionary Developmental Genetics, GZMB, University of Göttingen, Justus-von-Liebig-Weg 11, 37077 Göttingen, Germany

Julia Ulrich

Department of Evolutionary Developmental Genetics, GZMB, University of Göttingen, Justus-von-Liebig-Weg 11, 37077 Göttingen, Germany

Stefan Dippel

Georg-August-Universitat Gottingen Gottinger Graduiertenschule fur Neurowissenschaften Biophysik und Molekulare Biowissenschaften

Georg Oberhofer

Department of Evolutionary Developmental Genetics, GZMB, University of Göttingen, Justus-von-Liebig-Weg 11, 37077 Göttingen

Yonggang Hu

Department of Evolutionary Developmental Genetics, GZMB, University of Göttingen, Justus-von-Liebig-Weg 11, 37077 Göttingen, Germany

Jonas Schwirz

Department of Evolutionary Developmental Genetics, GZMB, University of Göttingen, Justus-von-Liebig-Weg 11, 37077 Göttingen, Germany

Magdalena Schacht

Department of Evolutionary Developmental Genetics, GZMB, University of Göttingen, Justus-von-Liebig-Weg 11, 37077 Göttingen, Germany

Sabrina Lehmann

Department of Evolutionary Developmental Genetics, GZMB, University of Göttingen, Justus-von-Liebig-Weg 11, 37077 Göttingen, Germany

Alice Montino

Department of Evolutionary Developmental Genetics, University of Göttingen, Justus-von-Liebig-Weg 11, 37077 Göttingen, Germany

Nico Posnien

Department of Developmental Biology, GZMB, University of Göttingen, Justus-von-Liebig-Weg 11, 37077 Göttingen, Germany

Daniela Gurska

Institute for Zoology: Developmental Biology, University of Cologne, Zülpicher Str. 47b, 50674 Cologne, Germany

Thorsten Horn

Institute for Zoology: Developmental Biology, University of Cologne, Zülpicher Str. 47b, 50674 Cologne, Germany

Jan Seibert

Institute for Zoology: Developmental Biology, University of Cologne, Zülpicher Str. 47b, 50674 Cologne, Germany

Iris M. Vargas Jentzsch

Institute for Zoology: Developmental Biology, University of Cologne, Zülpicher Str. 47b, 50674 Cologne, Germany

Kristen A. Panfilio

School of Life Sciences, University of Warwick, Gibbet Hill Campus, Coventry CV4 7AL, UK

Jianwei Li

Department Developmental Biology, GZMB, University of Göttingen, Justus-von-Liebig-Weg 11, 37077 Göttingen, Germany

Ernst A. Wimmer

Department of Developmental Biology, University of Göttingen, Justus-von-Liebig-Weg 11, 37077 Göttingen, Germany

Dominik Stappert

Institute of Zoology: Developmental Biology, University of Cologne, Zülpicher Weg 47b, 50674 Cologne, Germany

Siegfried Roth

Institute of Zoology: Developmental Biology, University of Cologne, Zülpicher Str. 47b, 50674 Cologne, Germany

Reinhard Schröder

Institut für Biowissenschaften, Universität Rostock, Albert-Einstein-Str. 3, 18059 Rostock, Germany

Yoonseong Park

Department of Entomology, Kansas State University, Manhattan, KS, 66506, United States

Michael Schoppmeier

Department of Biology, Divison of Developmental Biology, Friedrich-Alexander-University of Erlangen-Nürnberg, Staudtstr. 5, 91058 Erlangen, Germany

Martin Klingler

Department of Biology, Division of Developmental Biology, Friedrich-Alexander-University of Erlangen-Nürnberg, Staudtstr. 5, 91058 Erlangen, Germany

Ho-Ryun Chung

Max-Planck-Institute for Molecular Genetics, Department of Computational Molecular Biology, Ihnenstraße 63-73, 14195 Berlin, Germany

Sebastian Kittelmann

Oxford Brookes University, Centre for Functional Genomics, Gipsy Lane, Oxford, OX3 0BP, UK

Markus Friedrich

Department of Anatomy and Cell Biology, Wayne State University, Detroit, MI 48202, USA,

Rui Chen

Baylor College of Medicine

Boran Altincicek

Institute of Crop Science and Resource Conservation (INRES-Phytomedicine), Rheinische Friedrich-Wilhelms-University of Bonn, Bonn, Germany

Andreas Vilcinskas

Institute for Insect Biotechnology, Justus-Liebig University of Giessen, Heinrich-Buff-Ring 26-32, 35392 Giessen, Germany

Evgeny Zdobnov

Department of Genetic Medicine and Development, University of Geneva Medical School and Swiss Institute of Bioinformatics, Geneva, 1211, Switzerland

Sam Griffiths-Jones

Faculty of Biology, Medicine and Health, University of Manchester, Michael Smith Building, Oxford Road, Manchester, M13 9PT, UK

Matthew Ronshaugen

Faculty of Biology, Medicine and Health, University of Manchester, Michael Smith Building, Oxford Road, Manchester, M13 9PT, UK

Mario Stanke

Institut für Mathematik und Informatik, Universität Greifswald, Greifswald, Germany

Sue J. Brown

Division of Biology, Kansas State University, Manhattan, KS 66506

Gregor Bucher (✉ gbucher1@uni-goettingen.de)

https://orcid.org/0000-0002-4615-6401

Research article

Keywords: Tribolium castaneum; genome assembly Tcas5.2; official gene set OGS3, miRNA, splice variants, RefSeq genome

Posted Date: September 11th, 2019

DOI: https://doi.org/10.21203/rs.2.14298/v1

Version of Record: A version of this preprint was published on January 14th, 2020. See the published version at https://doi.org/10.1186/s12864-019-6394-6.

Abstract

Background: The red flour beetle Tribolium castaneum has emerged as an important model organism for the study of gene function in development and physiology, for ecological and evolutionary genomics, for pest control and a plethora of other topics. RNA interference (RNAi), transgenesis and genome editing are well established and the resources for genome-wide RNAi screening have become available in this model. All these techniques depend on a high quality genome assembly and precise gene models. However, the first version of the genome assembly was generated before the advent of next generation sequencing and with a small set of RNA sequence data limiting annotation quality.

Results: Here, we present a new genome assembly (Tcas5.2) and an enhanced genome annotation resulting in a new official gene set (OGS3) for Tribolium castaneum , which significantly increase the quality of the genomic resources. By adding large-distance jumping library DNA sequencing and using an improved assembly pipeline, the gaps in the genome assembly were reduced and the N50 increased to 4,753kbp. The precision of the gene models was enhanced by the use of a large body of RNA-Seq reads of different life history stages and tissue types, leading to the discovery of 1,452 novel gene sequences. We also added new features such as alternative splicing, well defined UTRs and miRNA target predictions. For quality control, 399 gene models were evaluated by manual inspection. The current gene set was submitted to Genbank and accepted as a RefSeq genome by NCBI.

Conclusions: The new genome assembly (Tcas5.2) and the official gene set (OGS3) provide enhanced genomic resources for genetic work in Tribolium castaneum . The much improved information on transcription start sites supports transgenic and gene editing approaches. Further, novel types of information such as splice variants and microRNA target genes open additional possibilities for analysis.

Background

The red flour beetle *Tribolium castaneum* is an excellent insect model system for functional genetics. In many respects the biology of *Tribolium* is more representative of insects than that of the fly *Drosophila melanogaster* [1–3]. This is especially true with respect to embryonic development: The *Tribolium* embryo is enveloped by extraembryonic membranes like most insects [4], develops embryonic legs, displays an everted head [5] and its posterior segments are formed sequentially from a posterior segment addition zone [6, 7]. With respect to postembryonic development, the *Tribolium* larval epidermal cells build most of the adult epidermis while in *Drosophila* they are replaced by imaginal cells [8]. In the telotrophic ovary type of *Tribolium* the biology of somatic stem cells can be studied independent of germline stem cells, which cease to divide prior to hatching [9]. *Tribolium* is also studied with respect to beetle specific evolutionary novelties such as elytra [10] and gin traps [11]. With regard to physiology, the formation of the extremely hard cuticle is studied [12] as well as the cryptonephridial system [13], which is a model for unique adaptation to dry habitats. Odoriferous glands are studied to understand the production of toxic secretions without harming the animal [14]. Finally, *Tribolium* is a representative of the Coleoptera, which is the most species rich taxon on earth [15] including many economically important pests such as leaf and snout beetles. Hence, it has been used as a model for pest control [16, 17]. In summary, *Tribolium* is useful for evolutionary comparisons of gene function among insects, for studying processes that are not represented in *Drosophila* and for pest control.

Research on gene function in *Tribolium* is fostered by an extensive toolkit. Transposon mediated transgenesis has led to the development of imaging and misexpression tools, and has facilitated a large scale insertional mutagenesis screen [18–24]. However, the main strength of the model system lies in its reverse genetics via RNAi. First, the RNAi response is very strong, reaching the null phenotype in those cases where a genetic mutant was available for comparison [25–28]. In addition, RNAi is environmental, i.e. cells very efficiently take up dsRNA from the hemolymph and the RNAi effect is transmitted from injected mothers to their offspring [29–31]. Based on this strength, a genome wide RNAi screen has been performed (iBeetle screen), where embryonic and other phenotypes were documented and made available via the iBeetle-Base [32–34]. Importantly, the genome wide collection of templates generated by iBeetle can be used for future screens directed at other processes. Recently, CRISPR/Cas9 mediated genome editing has been shown to work efficiently [35, 36].

An essential requirement for studying gene function is a high quality genome assembly and a well annotated gene set. Indeed, the first genomic sequence published in 2008 and subsequent updates [37, 38] contributed significantly to the growth of the community and increased the diversity of research topics studied in *Tribolium.* However, in the first published *Tribolium* genome assembly a substantial number of scaffolds had not been mapped to any chromosome. Further, the first gene annotations were mainly based on the detection of sequence features by bioinformatics tools and comparably few gene predictions were supported by RNA data. Hence, precision in the coding parts was limited and the non-coding UTR sequences, transcription start sites were usually not defined and splice variants were usually not predicted.

Here, we made use of new sequencing and mapping techniques in order to significantly enhance the genomic resources of *Tribolium.* In the new *Tribolium* assembly, Tcas5.2, scaffold length has been increased fivefold (scaffold N50: 4,753kbp). With the inclusion of RNA-Seq data, the precision of gene models was improved and additional features such as UTRs and alternative splice variants were added to 1,335 gene models. 1,452 newly predicted genes replaced a similar number of previously false positively predicted small genes. The current set of gene models (OGS3) is the first NCBI RefSeq annotation for *Tribolium castaneum.* Based on the enhanced annotation we compared the degree of conservation of protein sequences of a number of model systems revealing a higher conservation of *Tribolium* sequences compared to other Ecdysozoa. Moreover, based on the identification of UTRs we mapped for the first time in a beetle potential target genes of the miRNA complement and identified a conserved target gene set for a conserved miRNA.

Results

Re-sequencing of the Genome

The first published *Tribolium* reference genome sequence (NCBI Tcas3.0) was based on a Sanger 7x draft assembly [38] totaling 160 Mb, 90% of which was anchored to pseudomolecules or Linkage Groups (LGs) representing linkage groups in the molecular recombination map [39]. However, several large scaffolds (up to 1.17 Mb) were not included in these LGs. To improve this draft assembly, we sequenced the paired ends of three large-insert jumping libraries (appr. 3,200 bp, 6,800 bp, and 34,800 bp inserts, respectively). These sequences were used to link scaffolds in the Sanger assembly and fill gaps. Further, whole genome physical maps produced from images of ultra-long individual molecules of *Tribolium* DNA labeled at restriction sites (BioNano Genomics) were used to validate the assembly and merge scaffolds. The entire workflow and key steps are described below.

Using the long-insert jumping libraries, Atlas-Link (Baylor College of Medicine; www.hgsc.bcm.edu/software/atlas-link) joined unplaced scaffolds with the mapped scaffolds and reduced the total number of scaffolds from 2,320 to 2,236. Of these, three were manually split because Atlas-Link had erroneously joined scaffolds known to be on different linkage groups based on the molecular genetic recombination map, leading to a total of 2,240 scaffolds. The number of scaffolds in and total length of each LG is listed in Table 1. This analysis added formerly unplaced scaffolds to all LGs except LG4. In addition, 16 unplaced scaffolds were linked together.

We also took advantage of the new Illumina sequence information gained from the long insert jumping libraries to fill gaps and extend contigs. GapFiller [40] added 1,103,253 nucleotides and closed 2,232 gaps. Specifically, the number of gaps of assigned length 50, which actually included gaps less than 50 nucleotides long or potentially overlapping contigs, was reduced to 34.3% (from 1,793 to 615).

Finally, BioNano Genomics consensus maps were used to validate and further improve the assembly (for details, see [41]). More than 81% of Tcas 5.2 was directly validated by direct alignment with BioNano Genomics Consensus maps and the number of scaffolds was reduced by 4% to 2,148 and the N50 increased 3-fold to 4,753.0 kb (Table 1). In total, the N50 was increased almost 5-fold where superscaffolding with BioNano Genomics optical maps improved the contiguity of the assembly the most. Table 1 shows the impact of the single steps of the workflow to the quality of the genome assembly.

Re-Annotation of the *Tribolium* genome assembly

Re-annotation was performed using the gene finder AUGUSTUS [42]. For the current release, new data were available and incorporated as extrinsic evidence including RNA-Seq, ESTs (Expressed Sequence Tags) and protein sequences. The most important additional information was provided by the inclusion of extensive RNA-Seq data (approximately 6.66 billion reads) covering different stages and tissues (Table 2). This allowed us to determine UTRs and alternative splice variants, which were not annotated in the previous official gene set, and increased the accuracy of the predicted gene features. The parameters of automated annotation were adjusted based on manual quality control of more than 500 annotations of previously published genes. The new gene set, OGS3, consists of 16,593 genes with a total of 18,536 transcripts. 15,258 (92%) genes have one isoform, 944 (5.7%) genes have two, 270 (1.6%) have three and 121 (0.7%) genes have more than three isoforms. During the re-annotation of the *Tribolium* gene set a basic parameter set for AUGUSTUS was developed and is now delivered with AUGUSTUS as parameter set “tribolium2012” (link for download: see Materials and Methods).

Major Changes in the OGS3

We compared the previous official gene set OGS2 [37], which was ‘lifted’ to the new assembly, Tcas5.2, with the new OGS3 and found that 9,294 genes have identical protein sequences, 3,039 genes have almost identical protein sequences (95% minimum identity and 95% minimum coverage). 1,452 genes were completely new, meaning that they did not overlap any lifted OGS2 gene above the given thresholds. A similar number of predicted genes from OGS2 do not exist anymore in OGS3. Based on the lack of a Blast hit in invertebrates (e-value cutoff: e–05), GO annotation or RNA-Seq coverage (see Methods) we assume that these OGS2 genes do not correspond to real genes. Table 3 compares important characteristics between the previous and the current OGS.

When examining the newly found genes, we observe that 528 of 1,452 (36%) genes have significant Blast hits in other insect species (see below). Further, 690 of 997 (69.2%) spliced genes among the new genes have at least one intron supported by RNA-Seq. The new single exon genes have an average read coverage of about 550,000 reads per gene with minimum coverage of 11 reads per gene.

We further examined gene structure changes (not including the identification of splice variants). For this, we counted both, gene *join* and *split* events that occurred in the new gene set. We speak of a *join* event, if the CDS of an OGS3 gene overlaps the CDS of two or more genes from the previous gene set on the same strand. In total, we observe 949 such *join* events. In 485 (51%) of these events, an OGS3 intron supported by spliced read alignments spanned the gap between two neighboring OGS2 genes. We detected gene *split* events by counting gene *join* events where an old OGS2 gene joins multiple OGS3 genes. We observe 424 such events. In 45 cases (10%) the joining OGS2 intron has RNA-Seq support. Possibly, these genes were erroneously split, but this can have other reasons, too, such as spliced non-coding transcripts.

RNA-Seq support for the gene sets

Analysis of differential gene expression has become an essential tool in studying the genetic basis of biological processes. Such analyses profit from a better gene model where a higher number of reads can be mapped. To test whether the new gene set performed better in such analyses, we mapped our collection of RNA-Seq reads to both. In this analysis 6.66 billion RNA-Seq reads from *Tribolium* where mapped against the two gene sets (transcriptome) OGS3 and, for comparison, OGS2 with the alignment tool BLAT [43]. Alignments with less than 90% identity were discarded and only the best alignment was kept for each read (Table 4). The CDS of OGS2 was hit by about 70% of the reads whereas the CDS of the OGS3 was hit by 81% of the reads.

To evaluate the splice sites in the new gene set we compiled a set of splices suggested by gaps in RNA-Seq read alignments compared to the genomic sequence (intron candidates). These RNA-Seq read alignments where filtered by a range of criteria (see Methods). In total this set contained 65,274 intron candidates. We will refer to the term multiplicity of an intron candidate as the number of read alignments where deletions in the read alignment and intron boundaries are identical. Some intron candidates can be assumed not to be introns of coding genes, e.g. from alignment errors or from spliced noncoding genes. The intron candidates have an average multiplicity of 7898. The set contains 1403 intron candidates with multiplicity of one and 3362 with multiplicity smaller or equal to five. OGS3 contains about 30% more RNA-Seq supported introns than OGS2: 41,921 of 54,909 introns in the OGS2 (76.3%) and 54,513 of 63,211 in the OGS3 (86.2%) are identical to an intron suggested by RNA-Seq spliced read alignments.

BUSCO analysis reveals very high accuracy of the gene set

The completeness of OGS3 it was assessed using BUSCO (Benchmarking Universal Single-Copy Orthologs) and compared to the value for OGS2 [44] and to those of other sequenced genomes [45–47]. The genome of *Drosophila melanogaster* can be assumed to represent the maximal quality, the genome of *Apis mellifera* was recently re-annotated and is therefore comparable to the OGS3 from *Tribolium* and for *Parasteatoda tepidariorum* the first genome version was just published with the peculiarity of large duplication events. Nearly all of the conserved genes from the BUSCO Arthropoda set where found in OGS2 and OGS3 (Table 5). With 99.6% OGS3 was slightly better than OGS2 (99.3%). The completeness is close to *Drosophila* (99.8%) but better than in *Apis* (97.9%) or *Parasteatoda* (94.4%) (Table 5).

Official gene set and NCBI RefSeq genome

The genome assembly as well as the gene models have been submitted to Genbank (NCBI) as the RefSeq genome (GCF_000002335.3) and official gene set for Tribolium (OGS3) (GCA_000002335.3). It has also been accepted as RefSeq genome for the red flour beetle [48]. Genome assembly 5.2 and gene set OGS3 are available on the NCBI website (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/002/335/GCF_000002335.3_Tcas5.2) and are available as a preselection in several NCBI services, such as the BLAST search.

Protein sequence conservation

Drosophila melanogaster and *Caenorhabditis elegans* are the main invertebrate models for functional genetics and have contributed tremendously to the understanding of cellular and molecular processes relevant for vertebrate biology. However, their protein sequences are quite diverged compared to *Apis mellifera* or the annelid *Platynereis dumerilii [*49]. The extent to which diverged sequences correlate with diverged function remains unclear, which in turn calls into question the transferability of invertebrate findings to vertebrate biology. In *Tribolium,* the genetic toolkit is more developed compared to other insects or annelids. Unbiased genome-wide screening has been established making *Tribolium* an excellent alternative model for studying basic biological processes. We therefore asked how the protein sequences of the red flour beetle compare to other species.

We identified 1,263 single-copy orthologs across five species, made an alignment and calculated a phylogenetic tree (Fig. 1A). This analysis suggested that the *Tribolium* proteome is indeed more conserved compared to vertebrates than the proteomes of *Drosophila* and *Caenorhabditis,* but that it is more diverged than the annelid proteome. In such alignment-based sequence comparisons, the less onserved non-aligneable parts of the proteins are not considered. Therefore, we used an alignment-free method for measuring sequence distances [50, 51] on the same dataset and found it to basically reflect the tree albeit with less resolution (Fig. 1B). We conclude that the *Tribolium* proteome is more conserved than that of the above mentioned genetic model organisms while annelids appear to have a higher degree of conservation but have a much more restricted toolkit available.

Prediction of miRNA binding sites

MicroRNAs are short non-coding RNAs that regulate gene expression by guiding the RNA-induced silencing complex (RISC) to complementary sites in the 3’UTR regions of target mRNAs (reviewed in [52]). The principal interaction between microRNAs and their targets occurs through the so-called “seed” region, corresponding to the 2nd and 8th position of the mature microRNA sequence [53], and this complementarity can be used for computational predictions of microRNA-target pairs. Previous studies experimentally identified 347 microRNA genes in the *Tribolium castaneum* genome, each of which can generate two mature microRNAs derived from the two arms (5p and 3p) of the microRNA precursor hairpin (Supplementary Table 1)[54, 55]. We extracted the 3’UTR sequences of *Tribolium* protein-coding genes and annotated potential microRNA binding sites in these regions using an algorithm based on the microRNA target recognition principles described in [53]. In addition, we generated an alternative set of computational microRNA target predictions using an algorithm based on the thermodynamic properties of miRNA-mRNA duplexes irrespective of seed complementarity [56]. The two algorithms identified 309,675 and 340,393 unique putative microRNA-target pairs, with approximately 60% overlap. In each prediction set, 13,136 and 13,057 genes had at least one microRNA target site.

Comparison of miRNA target gene sets

MicroRNAs are recognized as important players in animal development, and their role in insects is best understood in the classical model organism *Drosophila melanogaster.* Comparative genomic analyses showed that 83 *Tribolium castaneum* miRNAs have one or more homologs in *Drosophila* [54, 55]. To assess whether conserved microRNAs also have a conserved target repertoire, we sought to assess the number of orthologous genes targeted by each conserved microRNA pair. To this end, we used an identical target prediction approach to determine microRNA-target pairs in *Drosophila melanogaster,* and calculated the numbers of homologous and non-homologous targets for each conserved microRNA pair in the two species (Supplementary Table 1). Results indicated that even though the majority of homologous microRNAs have conserved seed sequences for at least one mature product, their target repertoires diverged.

Nonetheless, a subset of well-conserved microRNAs had higher numbers of common predicted targets than expected by chance, especially based on seed complementarity. These included members of the bantam, mir–184, mir–279/286, mir–2/11/13/2944/6, mir–9, mir–14, mir–1, mir–7, mir–34 seed families, which have been previously identified for their roles in key developmental processed in *Drosophila,* and are highly expressed in both fruit fly and beetle embryos.

Given the large number of target predictions identified for individual microRNAs we examined the specific conserved targets for one of the microRNAs that both exhibited significant target conservation and had well characterized targets in *Drosophila.* The mir–279/286 family has been extensively characterized for its role in regulating the emergence of CO2 sensing neurons and in circadian rhythms. We find that in Tribolium of the nine characterized targets identified in *Drosophila,* one had no clear ortholog (upd), four did not have conserved targeted sequences in their UTRs (*STAT, Rho1, boss,* and *gcm)*,, but four targets (*nerfin–1, esg, ru,* and *neur)* had strongly conserved predicted target sites. microRNA regulation of all these four targets has clear functional importance in these developmental processes and two of them (*nerfin–1* and *esg)* work together as key players in the formation of CO–2 sensing neurons [57].

These findings suggest conserved miRNA regulate developmental pathways between the two taxa. The predicted miRNA binding sites are now available as tracks in the genome browser at iBeetle-Base (https://ibeetle-base.uni-goettingen.de/gb2/gbrowse/tribolium/)

Discussion

With respect to the toolkit for functional genetics in insects, the red flour beetle *Tribolium castaneum* is second only to *Drosophila melanogaster.* With this work, we aimed at enhancing the respective genomic resources in order to support the functional genetic work in *Tribolium castaneum.* To that end we quantitatively increased the accuracy of both the genome assembly and the OGS. In addition, we qualitatively enhanced the resource by adding novel information such as splice variants and miRNA target sites.

In order to close gaps and place more contigs on scaffolds, we added data from long-insert jumping libraries and BioNano Genomics optical mapping. It turned out that the latter contributed much more to enhance the previous assembly based on Sanger sequencing: While the first approach increased the N50 by 20% the BioNano Genomics consensus mapping led to another 3-fold increase of the N50. Hence, data from large single molecules is best suited to overcome the limits of sequencing-based assemblies. Compared to the recently re-sequenced genome assembly of the honey bee [46] our scaffold N50 is significant higher (4,753kb compared to 997kb). This is also true for the number of placed contigs (2149 compared to 5645). However, compared to *Drosophila,* the most thoroughly sequenced insect genome (contig N50 19,478kb), our improved assembly still lags behind.

The improved genome assembly and extensive RNA-Seq data provided the basis for an enhanced gene prediction. The BUSCO values indicate a more complete OGS, closer to *Drosophila* than to other emerging model insects. Further, 11% more RNA-Seq reads could be mapped to the gene predictions of OGS3 compared to OGS2, which is a relevant increase e.g. for differential gene expression analyses. The overall number of genes did not increase much. On one hand, 1,452 genes without sequence similarity to OGS2 were newly added to the gene set. On the other hand, a similar number of genes from OGS2 was not represented in OGS3 anymore. These were mostly very short genes not supported by RNA-Seq data. Hence, most of them were probably false predictions in the former gene set.

A qualitative enhancement is the detection and annotation of alternative splice variants. Since RNAi is splice variant specific in *Tribolium* [58], this opens the possibility to systematically check for the differences of function of isoforms. Further, the inclusion of UTR regions for many more genes enabled us for the first time to comprehensively map candidate miRNA binding sites to our gene set. Indeed, we have identified a large number of microRNA target sites in orthologs of both *Drosophila* and *Tribolium.* The miRNAs that we identified to have conserved targets belong mostly to miRNA families where obvious loss of function phenotypes have previously been characterized in other animals. One example is the miR–279/miR–996 family that share a common seed and have been found to play a key role in Drosophila CO2 sensing neurons and ovarian border cell development [57]. A number of the key microRNA targets identified in *Drosophila,* such as *nerfin, escargot,* and *neuralized* are predicted to be targets of *Tribolium* miR–279. This striking example of conservation illustrates that further comparative approaches have the potential to identify conserved regulatory networks involving miRNAs within insects based on the resources provided here. The enhanced coverage with RNA data revealed the transcription start sites of most genes, which helps in the design of genome editing approaches and of transgenic constructs based on endogenous enhancers and promoters [22, 23, 35, 59].

Finally, we show that the proteome of *Tribolium* is more conserved than that of *Drosophila,* which is an argument for using *Tribolium* as alternative model system when the biochemical function of proteins with relevance to human biology is studied.

Conclusions

The new genome assembly for *Tribolium castaneum* and the respective gene prediction resulted in a RefSeq genome and a new official gene set (OGS3). This promotes functional genetics studies with respect to a plethora of topics in *Tribolium,* opens the way for further comparative genomics, e.g. with respect to miRNAs and positions *Tribolium* as a central model organism within insects.

Methods

Genome resequencing and assembly

Reference genome files

The *T. castaneum* reference genome assembly (Tcas_3.0, NCBI accession number AAJJ01000000) was downloaded from NCBI. The following 23 contigs, which had been marked by NCBI as contaminants were removed: AAJJ01000455, AAJJ01001129, AAJJ01001336, AAJJ01001886, AAJJ01003084, AAJJ01003125, AAJJ01003874, AAJJ01004029, AAJJ01004493, AAJJ01004617, AAJJ01005150, AAJJ01005727, AAJJ01005755, AAJJ01006305, AAJJ01006331, AAJJ01007110, AAJJ01007612, AAJJ01007893, AAJJ01008452, AAJJ01009546, AAJJ01009593, AAJJ01009648, and AAJJ01009654. In addition, the first 411 nucleotides from AAJJ01009651, and the first 1,846 and last 46 nucleotides from AAJJ01005383 were removed after being identified as contaminants. The remaining 8,815 contigs (N50 = 43 Kb) had been used to construct the 481 scaffolds (N50 = 975 Kb) included in Tcas 3.0. Information from a genetic recombination map based on molecular markers [39], was used to anchor 176 scaffolds in 10 superscaffolds (often referred to as pseudomolecules or chromosome builds). In Tcas 3.0 these are referred to as ChLGX and ChLG2–10, representing the linkage groups in the recombination map. The remaining 305 scaffolds and 1,839 contigs that did not contribute to the superscaffolds were grouped together in Beetlebase (http://beetlebase.org or ftp://ftp.bioinformatics.ksu.edu/pub/BeetleBase/3.0/Tcas_3.0_BeetleBase3.0.agp) (unknown placement).

Description of Illumina libraries

The DNA used to construct three long-insert jumping libraries (3, 8, and 20 kb target size) was isolated at the Baylor Human Genome Sequencing Center in 2004 for Sanger-based sequencing. Thus, the source of DNA for these data is the same as for the original reference genome. The insert sizes for the three libraries are 3,173 bp, 6,775 bp, and 34,825 bp, respectively, with 10–15% standard deviation. Library construction, Illumina sequencing and cleaning were performed by MWGOperon (Europe). For all libraries, reads of minimum length 30 bp and maximum 100 bp were retained after cleaning and removal of the internal spacer. The “_1” files contain the forward reads while the “_2” files contain the reverse reads. Reads lacking the spacer or containing insert sequence only on one side of the spacer were not used. Table 4 lists the number of reads and their length for the jumping libraries.

Scaffolds linked with Atlas-Link v0.01

Atlas-Link is a software tool that links and orients scaffolds using mate pair libraries (www.hgsc.bcm.edu/software/atlas-link). Scaffolds were indexed using the IS algorithm in BWA (bwa index -a is tcas3.contigs.fa), the SA coordinates of the input reads were determined (E.g., bwa aln -t 4 tcas3.contigs.fa 3kb_1.fastq > 3kb_1.fastq.aln.), and alignments generated given the single-end reads, respectively (E.g., bwa samse tcas3.contigs.fa 3kb_1.fastq.aln 3kb_1.fastq > 3kb_1.fastq.sam). The mapped reads were merged (We used merge_pair_in_sam.pl from ATLAS GapFill v2.2, downloaded from www.hgsc.bcm.edu/software/atlas-gapfill. E.g., merge_pair_in_sam.pl 3kb_1.fastq.sam 3kb_2.fastq.sam

> 3kb.merged.sam) and Atlas-Link was run on each long insert jumping library (Ran first ATLAS_link/generate_input_for_ATLAS-link.pl -l sam_file_list -f tcas3.contigs.fa to generate the input files for Atlas-Link. Then, ran ATLAS_link/do-Atlas-Link.pl -l sam_file_list.libfile -t sam_file_list.vScaf.contig -f atlas.link.configure.file -a tcas.agp -o Output), with the settings described in Supplementary file 2. Table 5 shows the improvements that where achieved by Atlas-Link. Scaffold order and placement within Chromosome LG builds was used to validate the Atlas -Link output. We found that using a low value (3–10) for "minimum number of links to merge scaffolds" tended to scramble the order and orientation of scaffold that had been placed on Chromosome LG builds in the original assembly Tcas_3.0. Using a value of 300 minimum links reproduced most of the original order, linking neighboring scaffolds and including scaffolds that were unplaced in Tcas_3.0. In addition, we also updated the output AGP file, because Atlas-Link modified the contig_start and contig_stop coordinates making them all start from one. This is not an issue for the contigs that actually start from one, but for the ones that do not, i.e., the ones which NCBI has with x and x+y contig_start and contig_stop coordinates, Atlas-Link changed their coordinates to 1 and 1+y, thus indicating a different fragment of the contig to be kept. We fixed these to reflect the NCBI coordinates.

Contigs extended and gaps closed with GapFiller v1.10

GapFiller was run with the following parameters (perl GapFiller_v1–10_linux-x86_64/GapFiller.pl -l libraries.txt -s scaffolds.fasta -m 29 -o 10 -r 0.7 -n 10 -d 100 -t 0 -T 10 -i 10 -b standard_out. The results of the GapFiller run are listed in Table 6, the contents of libraries.txt are listed and explained in Supplementary file 2): minimum number of overlapping bases with the edge of the gap 29, minimum number of reads needed to call a base during an extension 10, percentage of reads that should have a single nucleotide extension in order to close a gap in a scaffold 0.7, minimum overlap required between contigs to merge adjacent sequences in a scaffold 10, maximum difference between the gapsize and the number of gapclosed nucleotides (extension is stopped if it matches this parameter + gap size) 100, number of reads to trim off the start and begin of the sequence 0, number of threads to run 10, and number of iterations to fill the gaps 20.

Note that the output from Atlas-Link is an AGP file, while GapFiller uses a batch FASTA file with the scaffolds as input. To generate the input file for GapFiller we used the AGP file generated by Atlas-Link and the batch FASTA file with the contigs and generated the scaffolds using the make_scaffolds_fa.pl (make_scaffolds_fa.pl tcas3.contigs.fasta atlas_link_with_fixed_contig_coords.agp scaffolds.fasta) script listed in Supplementary file 2. In the process, we merged the overlapping, and the adjacent contigs with no gaps in between them. We split the scaffolds into contigs with minimum gap length of 10, using fasta_to_agp.pl listed in Supplementary file 2. An intermediate assembly (Tcas_4.0) reflected these changes but lacked the BioNano Genomics mapping (see below).

Scaffolds joined using BioNano Genomics consensus maps

Annotation

The reannotation of the protein-coding genes of *Tribolium castaneum* was done in three main steps: 1) automatic gene prediction based on an unpublished intermediate assembly 4.0 with AUGUSTUS [42] incorporating evidence from multiple sources, 2) merging the gene prediction with the previous official gene set OGS2 [37] and 3) a mapping of the new gene set to assembly 5.2 using liftover [60]. Additionally, manual curation and correction was done for 399 genes.

Protein-coding Genes

AUGUSTUS is a gene prediction tool based on a hidden Markov model that allows to incorporate extrinsic evidence such as from RNA-Seq or protein homology. Such extrinsic evidence is summarized in the form of so-called 'hints' that are input to AUGUSTUS and that represent mostly soft evidence on the location of exons, introns and other gene features.

RNA-Seq libraries of around 6.66 billion reads from the iBeetle consortium and 9 external contributors constitute the majority of evidence. All reads were aligned against the repeat masked genome assembly 4.0 with GSNAP [61]. Hits were filtered according to three criteria. A hit must reach a minimum identity threshold of 92%. Furthermore, a paired read filter was applied: Reads that are paired must not exceed a genomic distance of 200 Kbp and must be correctly oriented towards each other. Subsequently, reads that could not be unambiguously aligned to a single locus (the identities of the two highest-scoring alignments were within 4% of each other) were discarded in order to avoid false positives such as from pseudogenes.

It is often hard to correctly align spliced reads, especially when they are spliced near the beginning or end of the read. Therefore, an iterative mapping approach was applied. First a set of preliminary introns was generated by using the spliced alignments found by GSNAP and by predicting introns *ab initio* with AUGUSTUS. Removing sequences of these introns produced partial spliced transcripts to which all reads were aligned a second time. We obtained an improved spliced alignment set with additional spliced alignments via a coordinate change induced by the coordinates of the preliminary introns (http://bioinf.uni-greifswald.de/bioinf/wiki/pmwiki.php?n = IncorporatingRNAseq.GSNAP). From the gaps in the read alignments hints on the location of introns were compiled, including the number of reads that support each intron. Further, from the RNA-Seq genome coverage hints on the location of (parts of) exons were generated.

In addition, evidence from 64,571 expressed sequence tags (ESTs), 19,284 proteins of invertebrates (from uniprot/swissprot database), repetitive regions in the genome detected by RepeatMasker (Smit, AFA, Hubley, R & Green, P. *RepeatMasker Open−4.0.*2013–2015, http://www.repeatmasker.org), 387 published coding genes from NCBI, 69 odorant binding Proteins [62] and 60 "gold standard" sequences, that were manually produced from cDNA by different groups within the *Tribolium* community. The RNA-Seq reads are available at public databases in the Bioproject PRJNA275195.

Integration of the previous gene set

By and large the AUGUSTUS gene set is more accurate. However, unclear loci remain, in which the true annotation is yet unknown. In order to introduce some stability in the gene set update we kept the old genes when in doubt whether a newly predicted gene with another structure is indeed a correction of the old gene structure. We address the problem of finding such gene structures by introducing the concept of specifically supported genes. Consider a gene gOGS2 from the previous gene set and a set of overlapping genes GAUG from the AUGUSTUS prediction. gOGS2 is said to be specifically supported, if it has at least one intron supported by RNA-Seq, that none of the genes in GAUG have. Additionally, every supported intron of genes in GAUG is also in gOGS2. In OGS3 we kept all specifically supported OGS2 genes and discarded all AUGUSTUS genes overlapping them.

The set of supported intron candidates was compiled from spliced RNA-Seq reads with a number of restrictions. Each intron candidate had to have a length between 32 and 350,000 bp, all splice sites had to be biologically feasible and the number of hints supporting a contradicting gene structure had to be at most 9 times higher than the number of hints supporting the intron candidate itself.

Additionally, we kept an OGS2 gene that did not overlap any AUGUSTUS gene, if it had homologs in *Drosophila* or other invertebrates or an annotated function (GO term listed in the Gene Onthology database [63]) or was covered by RNA-Seq reads with FPKM ≥ 0.01 (calculated with eXpress [64]). In total we kept 3,087 OGS2 genes and 13,413 AUGUSTUS genes.

Liftover from assembly 4.0 to assembly 5.2

After a *Tribolium* community call many genes were manually reviewed and edited based on an intermediate assembly 4.0. To preserve manually curated gene structures, we decided to transfer the new gene set to assembly 5.2. We created an assembly map that assigns each base of assembly 4.0 to a base in the new assembly 5.2, if possible. This map file was used to ‘lift’ above gene set to the updated assembly 5.2 using liftOver taken from the UCSC Genome Toolbox (http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64.v287/). 337 genes could not be unambiguously and completely mapped. We applied our annotation pipeline to the new assembly and merged the result with the lifted gene set from the previous assembly. Consequently, we were able to identify gene structures for which the improved assembly allowed a better annotation. The new gene set was complemented by 469 gene structures, that could only be predicted based on the new assembly. Furthermore, we corrected 745 of the lifted gene structures according to the concept of specific supported genes as described above.

The standard Viterbi algorithm used in AUGUSTUS predicted 159 transcripts with an in-frame stop codon spliced by an intron. To replace them with alternative gene structures that do not contain in-frame stop codons we ran AUGUSTUS with the option –mea = 1 on the affected regions. MEA is an alternative algorithm that can prohibit spliced in-frame stop codons but needs more computational time.

During the GenBank submission process some gene models were revised and a small number of genes had to be manually edited or deleted.

Orthology assignment and proteome analyses

Orthologs and paralogs between *T. castaneum* and *D. melanogaster* were found using the OrthoDB database [65] and results were formatted accordingly using custom Perl scripts.

For the phylogenetic analysis, we compared *T. castaneum* (Insecta:Coleoptera) with three other invertebrates; *Drosophila melanogaster* (Insecta:Diptera), *Caenorhabditis elegans* (Nematoda) and *Capitella teleta* (Annelida). The mammalian *Mus musculus* was used as outgroup. More specifically, we used OrthoDB and obtained 1,263 single-copy orthologs, in order to perform a phylogenomics analysis with RAxML [66]. Briefly, a multiple sequence alignment was built for each orthologous group separately, using MUSCLE [67]. Then, the resulting alignments were trimmed using trimAl [68] with parameters "-w 3 -gt 0.95 -st 0.01" and concatenated using custom Perl scripts. The concatenated alignment was subsequently used to perform a phylogenomic analysis using RAxML 7.6.6 (PROTGAMMAJTT model of amino acid substitutions) with 100 bootstrap replicates. The final tree was edited with EvolView [69] and InkScape 0.91.

The same set of genes was analyzed separately in an alignment independent approach (see Supplementary file 2 for details). Two approaches were performed using six distance measures (d1, …, d6): In the first approach, we used 'gdist' to determine the pairwise distances between sequences inside the groups, then 'phylip neighbor' to compute corresponding phylogenetic trees, rooted by setting MMUSC as outgroup, and computing the consensus tree using 'phylip consense'. In the second approach, we concatenated sequences in the groups in random order to form five artificial "whole proteom" sequences (one for each of the species), determined their pairwise distances and computed a phylogenetic tree using 'phylip neighbor', again setting the MMUSC sequence as outgroup. To check for robustness of the approach and also the influence of sequence lengths we performed these experiments with different subsets: (1) with all 1263 groups and (2) with a subset of the all groups. The subsets we considered were: (2a) groups with a certain minimum sequence length, (2b) only groups whose sequence lengths differed by at most a certain percentage, and (2c - only for experiment (B)) a random selection of groups (for instance, randomly select 80% of all groups for concatenation). Concatenation experiment (B) produced phylogenies that turned out to be almost immune against changes in order of concatenation and considerably robust against restricting consideration to all groups or subsets of groups concatenation. Best signals where obtained by distance d6, which resulted in the phylogeny displayed in Fig. 1B.

miRNA prediction

Mature sequences of *T. castaneum* microRNAs (Supplementary file 1) were retrieved from previous annotations [54, 55], and *D. melanogaster* microRNAs were retrieved from miRBase v21 [70]. *D. melanogaster* transcript 3'UTR sequences were retrieved from Flybase r6.09 [71]. MicroRNA target predictions in the two species were performed using two independent approaches. First, we identified target transcripts having regions complementary to the microRNA 7A1, 7m8 and 8mer seed sequences as described in [53], using a custom script provided by Antonio Marco [54], and the miRanda algorithm [56], with default parameters. Previously established conserved microRNAs between *T. castaneum* and *D. melanogaster* [54, 55] were used to assess conserved microRNA-target pairs. For microRNAs with more than 1 homolog in the other species, we assessed all possible combinations of homologous pairs. The numbers of conserved microRNA-target interactions (homologous microRNAs targeting homologous genes) were calculated using a custom script. The significance of the conserved target pair numbers was assessed by comparison with the number of orthologous genes obtained by random sampling of equal size without replacement 1000 times.

Abbreviations

OGS3: official gene set version 3

Tcas5.2: official assembly of genomic sequence of *Tribolium castaneum* version 5.2

RNAi: RNA interference

RNA-Seq: next generation sequencing of mRNAs

mRNA: messenger RNA

Mb: megabases

bp: base pairs

LG: linkage group

EST: expressed sequence tag

BLAST: Basic local alignment search tool

BLAT: BLAST like alignment tool

CDS: coding sequence

UTR: untranslated region

BUSCO: Benchmarking Universal Single-Copy Orthologs

Declarations

All manuscripts must contain the following sections under the heading 'Declarations':

- Ethics approval and consent to participate

not relevant

- Consent for publication

not relevant

- Availability of data and material

The datasets generated and analyzed during the current study are available in the following repositories:

The RefSeq genome assembly 5.2 (GCF_000002335.3) and the official gene set for *Tribolium castaneum* (OGS3) (GCA_000002335.3) are available at Genbank (NCBI). (Genbank: https://www.ncbi.nlm.nih.gov/genome/?term = GCA_000002335.3; ftp download: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/002/335/GCF_000002335.3_Tcas5.2) and at iBeetle-Base: https://ibeetle-base.uni-goettingen.de/help/resources

The RNA-Seq reads are available at public databases in the Bioproject PRJNA275195 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA275195)

The data and software underlying the alignment free sequence comparison is found in the following repository https://hdl.handle.net/21.11101/0000–0007-D64E–1. It contains: sequence data of the single-copy orthologs; executables of the used software (along with the source code; a jupyter notebook to execute the analysis we have done and a README file.

- Competing interests
  - Robert M. Waterhouse is Editorial Board Member at BMC Genomics. Otherwise, there are no competing interests
- Funding

  - We thank the Max-Planck-Society for financial support via the department of Dr. Herbert Jäckle and for the help by Dr. Ulrike Löhr.
  - SJB, NH and JS were supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P20 GM103418. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health.

- GB, JD, MS, MK and EAW were supported by the iBeetle project (DFG research unit FOR1234; BU1443/6, BU1443/7, BU1443/8, STA 1009/10, KL 656/7, Scho1058/4, WI1797/5).
- DS was supported by a BIF PhD fellowship from Boehinger Ingelheim Fonds; SR was funded by the SFB680 (DFG).
- DG, TH, JS, IMVJ, and KAP were supported by DFG grants PA 2044/1–1 and SFB 680 Project A12 to KAP.
- RMW was supported by Swiss National Science Foundation grant PP00P3_170664
- The funders hat no role in design of the study, interpretation and publication of the data.

- Authors’ contributions

NH, JS and SJB performed the genome assembly, analyzed the data and drafted the manuscript. LG and MSt performed the annotation, analyzed the data, determined the new official gene set and drafted the manuscript. PI, RMW, EZ performed the orthology assignment of the new gene set and drafted the manuscript. MN, SGJ and MR annotated the miRNA binding sites and drafted the manuscript. JD analyzed data and drafted the manuscript. CL determined the termini of the gene annotations. CD performed the alignment free sequence comparison and drafted the manuscript. JS, PK, JU, SD, GO, YH, JS, MS, SL, AM, NP, DG, TH, JS, IMVJ, KAP performed the manual quality control of the new gene set. JL, EAW, DS, SR, RS, YP,MS, MK, HRC, SK MF, BA, AV contributed sequencing data. MSt, SJB and GB conceived of and coordinated the project and drafted the manuscript. All authors consent with the publication of this version of the manuscript.

- Acknowledgements.

- We thank Jiaxin Qu, Xing-Zhi Song, Xiang Qin, Kim Worley, and Stephen Richards at HGCS, BCM supplying the DNA and for instruction on running Atlas link.

- Authors’ information (optional)

References

1. Brown SJ, Shippy TD, Miller S, Bolognesi R, Beeman RW, Lorenzen MD, et al. The red flour beetle, Tribolium castaneum (Coleoptera): a model for studies of development and pest biology. Cold Spring Harb Protoc. 2009;2009:pdb.emo126.
2. Klingler M. Tribolium. Curr Biol. 2004;14:R639–40.
3. Schröder R, Beermann A, Wittkopp N, Lutz R. From development to biodiversity—Tribolium castaneum, an insect model organism for short germband development. Dev Genes Evol. 2008;218:119–26.
4. Panfilio KA. Extraembryonic development in insects and the acrobatics of blastokinesis. Dev Biol. 2008;313:471–91.

5. Posnien N, Schinko JB, Kittelmann S, Bucher G. Genetics, development and composition of the insect head - A beetle's view. Arthropod Struct Dev. 2010;39:399–410.
6. Tautz D. Segmentation. Dev Cell. 2004;7:301–12.
7. Davis GK, Patel NH. SHORT, LONG, AND BEYOND: Molecular and Embryological Approaches to Insect Segmentation. Annu Rev Entomol. 2002;47:669–99.
8. Snodgrass R. Insect Metamorphosis: Smithsonian Miscellaneous Collections, V122, No. 9. Washington: Literary Licensing; 1954.
9. Bäumer D, Strohlein NM, Schoppmeier M. Opposing effects of Notch-signaling in maintaining the proliferative state of follicle cells in the telotrophic ovary of the beetle Tribolium. Front Zool. 2012;9:15.
10. Tomoyasu Y, Wheeler SR, Denell RE. Ultrabithorax is required for membranous wing identity in the beetle Tribolium castaneum. Nature. 2005;433:643–7.
11. Hu Y, Schmitt-Engel C, Schwirz J, Stroehlein N, Richter T, Majumdar U, et al. A morphological novelty evolved by co-option of a reduced gene regulatory network and gene recruitment in a beetle. Proc R Soc B. 2018;285:20181373.
12. Noh MY, Muthukrishnan S, Kramer KJ, Arakane Y. Cuticle formation and pigmentation in beetles. Curr Opin Insect Sci. 2016;17:1–9.
13. King B, Denholm B. Malpighian tubule development in the red flour beetle (Tribolium castaneum). Arthropod Struct Dev. 2014;43:605–13.
14. Li J, Lehmann S, Weißbecker B, Ojeda Naharros I, Schütz S, Joop G, et al. Odoriferous Defensive Stink Gland Transcriptome to Identify Novel Genes Necessary for Quinone Synthesis in the Red Flour Beetle, Tribolium castaneum. PLoS Genet. 2013;9:e1003596.
15. Hunt T, Bergsten J, Levkanicova Z, Papadopoulou A, John OSt, Wild R, et al. A Comprehensive Phylogeny of Beetles Reveals the Evolutionary Origins of a Superradiation. Science. 2007;318:1913–6.
16. Merzendorfer H, Kim HS, Chaudhari SS, Kumari M, Specht CA, Butcher S, et al. Genomic and proteomic studies on the effects of the insect growth regulator diflubenzuron in the model beetle species Tribolium castaneum. Insect Biochem Mol Biol. 2012;42:264–76.
17. Ulrich J, Dao VA, Majumdar U, Schmitt-Engel C, Schwirz J, Schultheis D, et al. Large scale RNAi screen in Tribolium reveals novel target genes for pest control and the proteasome as prime target. BMC Genomics. 2015;16. doi:10.1186/s12864–015–1880-y.
18. Berghammer AJ, Klingler M, Wimmer EA. A universal marker for transgenic insects. Nature. 1999;402:370–1.
19. Koniszewski NDB, Kollmann M, Bigham M, Farnworth M, He B, Büscher M, et al. The insect central complex as model for heterochronic brain development-background, concepts, and tools. Dev Genes Evol. 2016;226:209–19.

20. Lorenzen MD, Berghammer AJ, Brown SJ, Denell RE, Klingler M, Beeman RW. piggyBac-mediated germline transformation in the beetle Tribolium castaneum. Insect Mol Biol. 2003;12:433–40.

21. Sarrazin AF, Peel AD, Averof M. A Segmentation Clock with Two-Segment Periodicity in Insects. Science. 2012. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd = Retrieve&db = PubMed&dopt = Citation&list_uids = 22403177.

22. Schinko JB, Weber M, Viktorinova I, Kiupakis A, Averof M, Klingler M, et al. Functionality of the GAL4/UAS system in Tribolium requires the use of endogenous core promoters. BMC Dev Biol. 2010;10:53.

23. Schinko JB, Hillebrand K, Bucher G. Heat shock-mediated misexpression of genes in the beetle Tribolium castaneum. Dev Genes Evol. 2012;222:287–98.

24. Trauner J, Schinko J, Lorenzen MD, Shippy TD, Wimmer EA, Beeman RW, et al. Large-scale insertional mutagenesis of a coleopteran stored grain pest, the red flour beetle Tribolium castaneum, identifies embryonic lethal mutations and enhancer traps. BMC Biol. 2009;7:73.

25. Beermann A, Jay DG, Beeman RW, Hulskamp M, Tautz D, Jürgens G. The Short antennae gene of Tribolium is required for limb development and encodes the orthologue of the Drosophila Distal-less protein. Development. 2001;128:287–97.

26. Brown SJ, Mahaffey JP, Lorenzen MD, Denell RE, Mahaffey JW. Using RNAi to investigate orthologous homeotic gene function during development of distantly related insects. Evol Dev. 1999;1:11–5.

27. Peel AD, Schanda J, Grossmann D, Ruge F, Oberhofer G, Gilles AF, et al. Tc-knirps plays different roles in the specification of antennal and mandibular parasegment boundaries and is regulated by a pair-rule gene in the beetle Tribolium castaneum. BMC Dev Biol. 2013;13:25.

28. Curtis CD, Brisson JA, DeCamillis MA, Shippy TD, Brown SJ, Denell RE. Molecular characterization of Cephalothorax, the Tribolium ortholog of Sex combs reduced. Genesis. 2001;30:12–20.

29. Bucher G, Scholten J, Klingler M. Parental RNAi in Tribolium (Coleoptera). Curr Biol. 2002;12:R85–6.

30. Tomoyasu Y, Denell RE. Larval RNAi in Tribolium (Coleoptera) for analyzing adult development. Dev Genes Evol. 2004;214:575–8.

31. Tomoyasu Y, Miller SC, Tomita S, Schoppmeier M, Grossmann D, Bucher G. Exploring systemic RNA interference in insects: a genome-wide survey for RNAi genes in Tribolium. Genome Biol. 2008;9:R10.

32. Schmitt-Engel C, Schultheis D, Schwirz J, Ströhlein N, Troelenberg N, Majumdar U, et al. The iBeetle large-scale RNAi screen reveals gene functions for insect development and physiology. Nat Commun. 2015;6:7822.

33. Dönitz J, Schmitt-Engel C, Grossmann D, Gerischer L, Tech M, Schoppmeier M, et al. iBeetle-Base: a database for RNAi phenotypes in the red flour beetle Tribolium castaneum. Nucleic Acids Res. 2015;43:D720–5.

34. Dönitz J, Gerischer L, Hahnke S, Pfeiffer S, Bucher G. Expanded and updated data and a query pipeline for iBeetle-Base. Nucleic Acids Res. 2018;46:D831–5.

35. Gilles AF, Schinko JB, Averof M. Efficient CRISPR-mediated gene targeting and transgene replacement in the beetle Tribolium castaneum. Dev Camb Engl. 2015;142:2832–9.

36. Gilles AF, Schinko JB, Schacht MI, Enjolras C, Averof M. Clonal analysis by tunable CRISPR-mediated excision. bioRxiv. 2018;:394221.

37. Kim HS, Murphy T, Xia J, Caragea D, Park Y, Beeman RW, et al. BeetleBase in 2010: revisions to provide comprehensive genomic information for Tribolium castaneum. Nucleic Acids Res. 2010;38:D437–42.

38. Richards S, Gibbs RA, Weinstock GM, Brown SJ, Denell R, Beeman RW, et al. The genome of the model beetle and pest Tribolium castaneum. Nature. 2008;452:949–55.

39. Lorenzen MD. Genetic Linkage Maps of the Red Flour Beetle, Tribolium castaneum, Based on Bacterial Artificial Chromosomes and Expressed Sequence Tags. Genetics. 2005;170:741–7.

40. Boetzer M, Pirovano W. Toward almost closed genomes with GapFiller. Genome Biol. 2012;13:R56.

41. Shelton JM, Coleman MC, Herndon N, Lu N, Lam ET, Anantharaman T, et al. Tools and pipelines for BioNano data: molecule assembly pipeline and FASTA super scaffolding tool. BMC Genomics. 2015;16. doi:10.1186/s12864–015–1911–8.

42. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24:637–44.

43. Kent WJ. BLAT—-The BLAST-Like Alignment Tool. Genome Res. 2002;12:656–64.

44. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.

45. Gramates LS, Marygold SJ, Santos G dos, Urbano J-M, Antonazzo G, Matthews BB, et al. FlyBase at 25: looking to the future. Nucleic Acids Res. 2017;45:D663–71.

46. Elsik CG, Worley KC, Bennett AK, Beye M, Camara F, Childers CP, et al. Finding the missing honey bee genes: lessons learned from a genome upgrade. BMC Genomics. 2014;15:86.

47. Schwager EE, Sharma PP, Clarke T, Leite DJ, Wierschin T, Pechmann M, et al. The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. BMC Biol. 2017;15:62.

48. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44:D733–45.

49. Raible F, Tessmar-Raible K, Osoegawa K, Wincker P, Jubin C, Balavoine G, et al. Vertebrate-type intron-rich genes in the marine annelid Platynereis dumerilii. Science. 2005;310:1325–6.

50. Otu HH, Sayood K. A new sequence distance measure for phylogenetic tree construction. Bioinformatics. 2003;19:2122–30.

51. Ulitsky I, Burstein D, Tuller T, Chor B. The average common substring approach to phylogenomic reconstruction. J Comput Biol J Comput Mol Cell Biol. 2006;13:336–50.

52. Ha M, Kim VN. Regulation of microRNA biogenesis. Nat Rev Mol Cell Biol. 2014;15:509–24.

53. Bartel DP. MicroRNAs: Target Recognition and Regulatory Functions. Cell. 2009;136:215–33.

54. Marco A, Hui JHL, Ronshaugen M, Griffiths-Jones S. Functional shifts in insect microRNA evolution. Genome Biol Evol. 2010. doi:10.1093/gbe/evq053.

55. Ninova M, Ronshaugen M, Griffiths-Jones S. MicroRNA evolution, expression, and function during short germband development in Tribolium castaneum. Genome Res. 2016;26:85–96.

56. Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. MicroRNA targets in Drosophila. Genome Biol. 2003;5:R1.

57. Sun K, Jee D, de Navas LF, Duan H, Lai EC. Multiple In Vivo Biological Processes Are Mediated by Functionally Redundant Activities of Drosophila mir–279 and mir–996. PLOS Genet. 2015;11:e1005245.

58. Arakane Y, Muthukrishnan S, Beeman RW, Kanost MR, Kramer KJ. Laccase 2 is the phenoloxidase gene required for beetle cuticle tanning. Proc Natl Acad Sci U A. 2005;102:11337–42.

59. Lai Y-T, Deem KD, Borràs-Castells F, Sambrani N, Rudolf H, Suryamohan K, et al. Enhancer identification and activity evaluation in the red flour beetle, Tribolium castaneum. Dev Camb Engl. 2018;145.

60. Hinrichs AS. The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. 2006;34:D590–8.

61. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010;26:873–81.

62. Dippel S, Oberhofer G, Kahnt J, Gerischer L, Opitz L, Schachtner J, et al. Tissue-specific transcriptomics, chromosomal localization, and phylogeny of chemosensory and odorant binding proteins from the red flour beetle Tribolium castaneum reveal subgroup specificities for olfaction or more general functions. BMC Genomics. 2014;15:1141.

63. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25:25–9.

64. Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Methods. 2013;10:71–3.

65. Kriventseva EV, Tegenfeldt F, Petty TJ, Waterhouse RM, Simão FA, Pozdnyakov IA, et al. OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. Nucleic Acids Res. 2015;43 Database issue:D250–256.

66. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006;22:2688–90.

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

68. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972–3.

69. Zhang H, Gao S, Lercher MJ, Hu S, Chen W-H. EvolView, an online tool for visualizing, annotating and managing phylogenetic trees. Nucleic Acids Res. 2012;40:W569–72.

70. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 2014;42:D68–73.

71. St Pierre SE, Ponting L, Stefancsik R, McQuilton P, FlyBase Consortium. FlyBase 102—advanced approaches to interrogating FlyBase. Nucleic Acids Res. 2014;42 Database issue:D780–788.

Tables

Table 1: Assembly improvement

Assembly	Length	Scaffolds	Scaffold N50 (kbp)
Tcas 3.0	160 445 652	2 320	976.4
After Atlas-Link	160 667 144	2 240	1 175.4
After GapFiller	160 744 700	2 240	1 176.7
After BioNano Genomics / Tcas 5.2	165 921 904	2 148	4 753.0

Table 2: Read coverage of OGS2 and OGS3

	OGS2	OGS3
Total number of reads matching	4 634 356 882	7 418 675 525
Number of reads per transcript	278 926	400 317
Number of reads per exon position	285.77	260.45
Number of reads per coding position	286.37	577.42

Table 3: Annotation improvement:

	OGS2	OGS3
Number of genes	16 561	16 593
Average coding length	1341 bp	1473 bp
Number of coding exons per transcript	4.32	5.02
GC content	0.4597%	0.4625%
Fraction of single exon genes	17.66%	17.74%
Number of introns (excluding UTR)	54 909 (54 875)	63 211 (58 837)
Average intron length	1167 bp	1362 bp

Table 4: Comparison of RNA-Seq read mapping to OGS2 and OGS3

	OGS2	OGS3
Number of …		
… reads matching the CDS	4 630 851 186	5 386 090 982
… reads per transcript	278 715	290 653
… reads per coding position	285.67	291.40

Table 5 BUSCO analysis

	Tcas OGS2	Tcas OGS3	Dmel r16.19	Amel 4.5	Ptep 2.0
Complete	1058 (99.3%)	1061 (99.6%)	1063 (99.8%)	1043 (97.9%)	1007 (94.4%)
Complete single copy	1054 (98.9%)	1056 (99.1%)	1055 (99%)	1038 (97.4%)	966 (90.6%)
Complete duplicated	4 (0.4%)	5 (0.5%)	8 (0.8%)	5 (0.5%)	41 (3.8%)
Fragmented	5 (0.5%)	2 (0.2%)	0 (0%)	15 (1.4%)	18 (1.7%)
Missing	3 (0.2%)	3 (0.2%)	3 (0.2%)	8 (0.7%)	41 (3.9%)
Genes in BUSCO profile	1066	1066	1066	1066	1066

Table 6: Mate pairs jumping library statistics

FastQ	Total reads	Total length
3kb_1	23 677 983	2 120 896 823
3kb_2	23 677 983	2 123 186 604
8kb_1	23 202 365	2 093 651 921
8kb_2	23 202 365	2 096 015 114
20kb_1	12 884 671	1 151 209 160
20kb_2	12 884 671	1 153 515 873

Table 7: Number of scaffolds and ungapped length before and after running Atlas-Link

Molecule	Scaffolds before	Ungapped length before	Scaffolds after	Ungapped length after	Unplaced scaffolds added	Unplaced ungapped length added
LG1=X	13	7 011 684	13	7 071 107	2	59 423
LG2	20	14 013 343	18	14 229 660	2	216 317
LG3	35	27 022 651	29	28 072 007	8	1 049 356
LG4	7	11 540 046	6	11 540 046	-	-
LG5	17	13 832 902	17	14 111 830	3	278 928
LG6	15	8 229 537	12	8 262 430	2	32 893
LG7	18	14 841 431	15	15 084 119	3	242 688
LG8	16	12 760 817	14	12 870 760	1	109 943
LG9	21	14 567 469	21	14 900 846	2	333 377
LG10	14	7 043 942	12	7 070 154	1	26 212
Unplaced multi-contig	305	16 272 476	263	14 079 574		
Unplaced single-contig	1 839	4 176 957	1 820	4 020 722		
Total	2 320	151 313 255	2 240	151 313 255		

Table 8: Ungapped length and spanned gaps before and after running GapFiller

Molecule	Ungapped length before	Spanned gaps before	Ungapped length after	Spanned gaps after
LG1=X	7 071 107	301	7 096 881	201
LG2	14 229 660	359	14 306 202	192
LG3	28 072 007	1 451	28 315 770	929
LG4	11 540 046	300	11 632 658	160
LG5	14 111 830	358	14 196 565	193
LG6	8 262 430	555	8 332 882	407
LG7	15 084 119	429	15 185 902	258
LG8	12 870 760	577	12 987 347	378
LG9	14 900 846	634	15 007 071	384
LG10	7 070 154	498	7 128 489	365
Unplaced multi-contig	14 079 574	1 111	14 205 681	874
Unplaced single-contig	4 020 722	-	4 021 060	-
Total	151 313 255	6 573	152 416 508	4 341

Table 9: Number of scaffolds, scaffolds’ length, and N50 before and after using BNG consensus maps

Molecule	Scaffolds before	Scaffolds after	Length (Mb) before	Length (Mb) after	N50 (kb) before	N50 (kb) after	Unplaced scaffolds added
LG1=X	13	4	7.34	8.92	1160.70	7264.05	2
LG2	18	8	14.78	15.034064	1207.76	9314.472	0
LG3	29	18	29.78	31.017975	1409.81	2672.697	3
LG4	6	3	12.11	12.24	2906.70	9484.15	0
LG5	17	7	14.64	15.36	1402.64	4484.65	1
LG6	12	9	9.02	9.25	956.12	2189.88	0
LG7	15	6	15.74	16.48	1333.70	8809.74	0
LG8	14	9	13.66	13.98	1312.85	4002.45	1
LG9	21	10	15.81	16.12	893.90	4920.63	0
LG10	12	11	7.54	8.84	1198.49	1224.30	3
Unplaced	2083	2072	20.33	17.35	150.43	104.32	2
Total	2240	2157	160.74	164.60	1160.70	4002.45	12

Figures

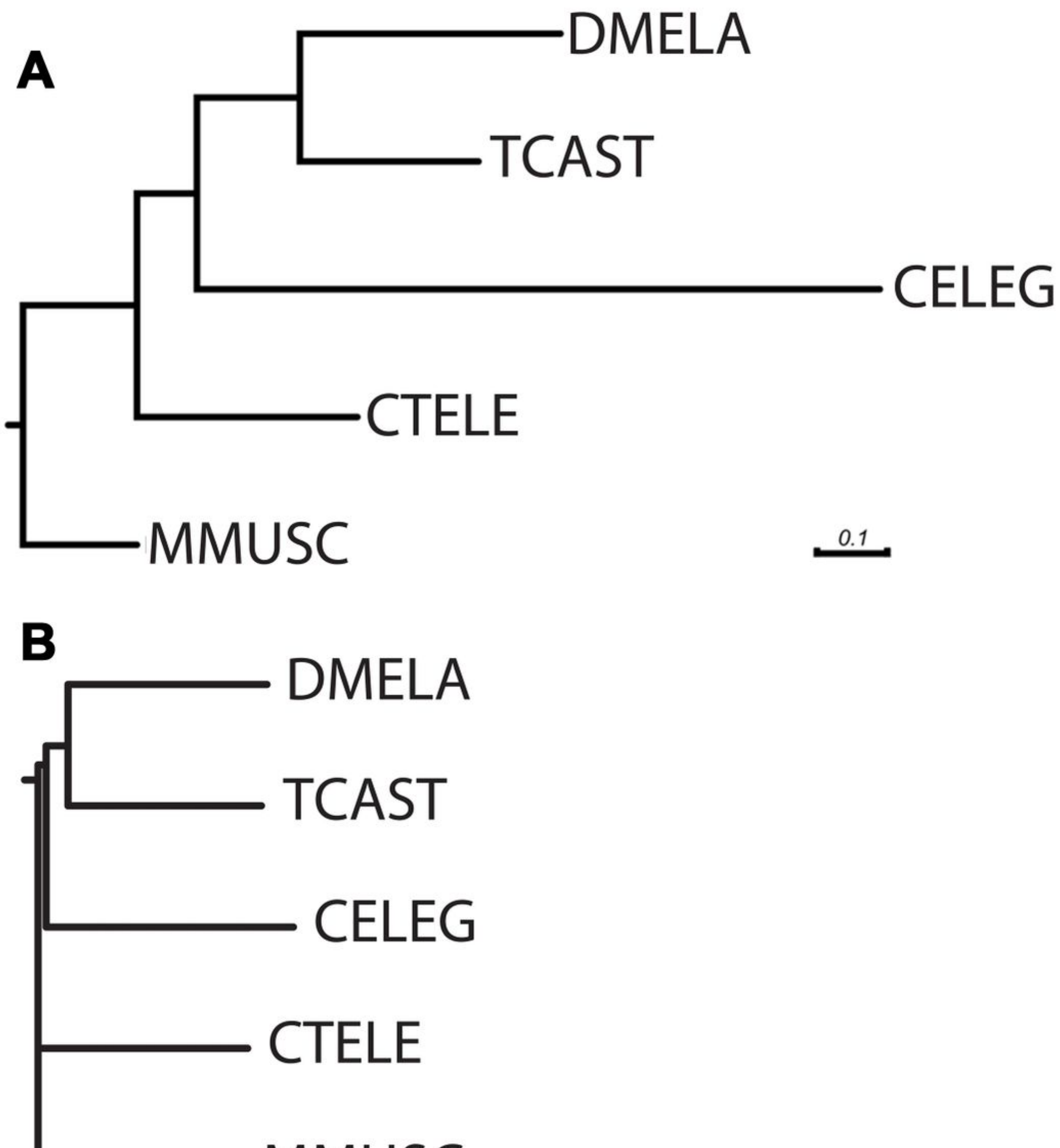

Figure 1

Protein evolution in selected model organisms. A) An alignment-based comparison of the protein sequences of 1,263 single-copy orthologs indicate that the proteome of Tribolium is more conserved than that of the main invertebrate models Drosophila melanogaster (DMELA) or Caenorhabditis elegans (CELEG). Sequences of annelids are more conserved. Shown is Capitella teleta - see Raible et al. 2005 for Platynereis dumerilii. The tree was rooted using the Mus musculus (Mammalia) as outgroup. The distances are shown as substitutions per site. B) An alignment-free comparison shows the same trend but with lower resolution. DMELA: Drosophila melanogaster; TCAST: Tribolium castaneum; CELEG: Caenorhabditis elegans; CTELE: Capitella telata ; MMUSC: Mus musculus

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplement1.xlsx
- supplement2.docx