

The genome of the model beetle and pest *Tribolium castaneum*

Tribolium Genome Sequencing Consortium*

Tribolium castaneum is a member of the most species-rich eukaryotic order, a powerful model organism for the study of generalized insect development, and an important pest of stored agricultural products. We describe its genome sequence here. This omnivorous beetle has evolved the ability to interact with a diverse chemical environment, as shown by large expansions in odorant and gustatory receptors, as well as P450 and other detoxification enzymes. Development in *Tribolium* is more representative of other insects than is *Drosophila*, a fact reflected in gene content and function. For example, *Tribolium* has retained more ancestral genes involved in cell–cell communication than *Drosophila*, some being expressed in the growth zone crucial for axial elongation in short-germ development. Systemic RNA interference in *T. castaneum* functions differently from that in *Caenorhabditis elegans*, but nevertheless offers similar power for the elucidation of gene function and identification of targets for selective insect control.

By far the most evolutionarily successful metazoans¹, beetles (Coleoptera) can luminesce (fireflies), spit defensive liquids (bombardier beetles), visually and behaviourally mimic bees and wasps, or chemically mimic ants that detect intruders by their foreign odour. Many beetles (for example, boll weevil, corn rootworm, Colorado potato beetle and Asian longhorn beetle) are associated with billions of dollars of agricultural and natural resource losses.

The red flour beetle, *Tribolium castaneum*, found wherever grains or other dried foods are stored, has a highly evolved kidney-like cryptonephridial organ to survive such extremely dry environments. It has demonstrated resistance to all classes of insecticides used against it. Like all beetles, *Tribolium* has elytra (wing covers) that coordinate precisely with folding wings, allowing flight while providing protection.

Tribolium facilitates genetic analysis with ease of culture, a short life cycle, high fecundity, and facility for genetic crosses (see ref. 2), allowing efficient genetic screens by means of chemical mutagens, radiation and binary transposon systems³. As in *Caenorhabditis elegans*, RNA interference (RNAi) is systemic in *Tribolium*, facilitating knockdown of specific gene products in any tissue, developmental stage or offspring of double-stranded (ds)RNA-injected females^{4,5}.

Particularly favoured for developmental studies, *Tribolium* is much more representative of other insects than is *Drosophila*⁶. In contrast to *Drosophila*, *Tribolium* larvae display eyes in a fully formed head and three pairs of thoracic legs (Supplementary Fig. 1). In addition, *Tribolium* develops via short-germ embryogenesis where additional segments are sequentially added from a posterior growth zone (Supplementary Fig. 1). This proliferative mechanism of segmentation differs from the *Drosophila* model, but resembles that of vertebrates and basal arthropods such as millipedes⁷.

Genome sequence and organization

Approximately 1.52 million sequence reads (7.3× coverage) were generated from the highly inbred Georgia 2 (GA2) strain and assembled into contigs totalling 152 megabases (Mb) and scaffolds spanning ~160 Mb of genomic sequence (Supplementary Tables 1–4 and Supplementary Information). Almost 90% of this sequence was mapped to the ten *Tribolium* linkage groups using a genetic map of

~500 markers generated from the GA2 strain⁸. Excluding heterochromatic regions dense in highly repetitive sequences, the genome is well represented and of high quality (see Supplementary Data for details).

G+C content. *Tribolium*, like *Apis*, has a very (A+T)-rich genome (33% and 34% G+C, respectively), but *Tribolium* G+C domains lack the extremes of G+C content present in *Apis mellifera* (Fig. 1 and Supplementary Fig. 3). Despite global G+C similarity to *Apis*, genes in *Tribolium*, as in *Anopheles* and *Drosophila* but not *Apis*, show a bias towards occurring in (G+C)-rich regions of the genome (Fig. 1). Whatever mechanism drives the accumulation of A+T nucleotides in *Tribolium*, it does not affect genes in the manner observed in the honeybee, where perhaps additional mechanisms are present.

Repetitive DNA. Fully one-third of the *Tribolium* genome assembly consists of repetitive DNA, which is also (A+T)-rich. Compared to other insects, there is a paucity of microsatellites (1–6-base-pair (bp) motifs) in *Tribolium*⁹. However, *Tribolium* contains a relative excess of larger satellites, including several with repeat units longer than 100 bp (2.5% of the *Tribolium* genome compared with 0.7% in *Drosophila*). Most (83%) of the microsatellites are found in intergenic regions (63%) or introns (20%), but there is strong overrepresentation of non-frameshift-causing repeats (3- and 6-bp motifs) due to a dearth of dinucleotide repeats (see Supplementary Information). Of 981 randomly chosen microsatellites, 509 (55.2%) are polymorphic in a sample of 11 *Tribolium* populations from around the world⁹, providing an extensive collection of markers for population studies. Preliminary efforts to assess global population structure show a shallow but significant correlation between geographic and genetic distance (Supplementary Fig. 4). This suggests that anthropogenic dispersal may maintain a modest level of gene flow across vast distances in this human commensal.

Transposable elements. Transposable elements and other repetitive DNA accumulate in regions along each linkage group that resemble the pericentric blocks of heterochromatin visible in *HpaII*-banded chromosomes¹⁰. These regions are probably composed largely of highly repetitive heterochromatic sequences, and represent most of the 44-Mb difference between the estimated genome size (0.2 pg or

*Lists of participants and affiliations appear at the end of the paper.

204 Mb¹¹) and the current assembly (160 Mb). Indeed, as much as 17% of the *Tribolium* genome is composed of a 360-bp satellite¹² that constitutes only 0.3% of the assembled genome sequence. Several families of DNA transposons, as well as long terminal repeat (LTR) and non-LTR retrotransposons, constituting approximately 6% of the genome, were identified via encoded protein sequence similarity to previously identified elements using TEPIPE or BLAST, and are listed in Supplementary Table 5.

Telomeres. *Tribolium* has a telomerase and telomeres containing TCAGG repeats¹³, a variant of the standard arthropod TTAGG telomeric repeat. Manual assembly of the proximal regions of multiple telomeres beyond the ends of the assembled scaffolds (Supplementary Information) reveals TCAGG repeats interrupted by full-length and 5'-truncated non-LTR retrotransposons belonging to the R1 clade, best known for insertions in the rDNA locus¹⁴. *Tribolium* telomeres range in length from 15 kilobases (kb) upwards and probably represent a stage intermediate to the loss of telomeres and telomerase in Diptera compared with the simple canonical structure of the honeybee¹⁵ or the more regular insertion of non-LTR retrotransposons into the simple repeats of the silkworm¹⁶.

Gene content and the proteome

Comparative gene content analysis. To understand the consensus set of 16,404 gene models in the context of other available insect and vertebrate genomes, all genes were classified according to their degree of similarity using systematic cross-species analysis. Five insects (*Drosophila melanogaster*, *Anopheles gambiae*, *Aedes aegypti*, *T. castaneum*, *A. mellifera*) and five vertebrates (*Homo sapiens*, *Mus musculus*, *Monodelphis domestica*, *Gallus gallus*, *Tetraodon nigroviridis*) with similar phylogenetic branching orders were chosen for the comparison. We found the fractions of universal and insect-specific orthologues in *Tribolium* similar to other insect genomes, as expected, whereas the number of genes without similarity is

considerably higher (Fig. 2), possibly attributable to less stringent gene prediction.

Over 47% of *Tribolium* genes (7,579) are ancient, with traceable orthologous relations between insects and vertebrates including 15% (2,403) universal single-copy orthologues. Another 1,462 *Tribolium* genes (9%) constitute the core of what are currently insect-specific orthologues. In comparison, 21% (4,937) of human genes have vertebrate-specific orthologues.

Several hundred ancient genes seem to be under limited evolutionary selection and were independently lost in several species studied (the patchy fraction, defined in Fig. 2). Each new genome uncovers previously invisible ancestral relations among genes—for example, as many as 126 orthologous gene groups shared between *Tribolium* and humans seem to be absent from the other sequenced insect genomes (Fig. 3 and Supplementary Table 10), 44 of which are single-copy genes present in all vertebrates.

The evolutionary emergence of many predicted *Tribolium* genes is not clear. Thousands of genes currently appear to be species-specific as either no sequence similarity to other genes is detectable, or homology but not orthology can be determined. Reassuringly, this fraction is similar in *Tribolium* and *Drosophila*.

We quantified the species phylogeny using a maximum likelihood approach with the concatenated multiple alignment of 1,150 universal single-copy orthologues present in all the organisms studied—an ideal genome-wide data set of essential genes evolving under similar constraints (Fig. 2 and Supplementary Fig. 6). This analysis confirmed previous analyses based on expressed sequence tag (EST) sequences that the Hymenoptera are basal within the Holometabola¹⁷. The shorter branch length for *Tribolium* implies that the elevated rate of evolution observed in *Drosophila* and *Anopheles* occurred more recently¹⁸.

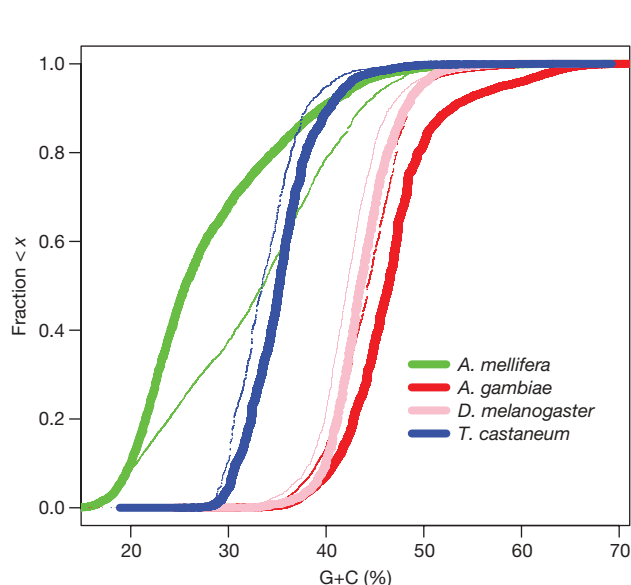


Figure 1 | Cumulative distribution of genic and genomic G+C-content domains in *Apis mellifera*, *Anopheles gambiae*, *Drosophila melanogaster* and *Tribolium castaneum*. Cumulative distributions show the fraction of genes (thick lines) or of the entire genome (thin lines) occurring in G+C-content domains less than a given percentage G+C (<X). The more (A+T)-rich half of the *T. castaneum* genome contains only 30.8% of all *T. castaneum* genes (31.4% and 33% of *A. gambiae* and *D. melanogaster* genes, respectively), whereas the more (A+T)-rich half of the *A. mellifera* genome contains 77.6% of its genes. At every point on the *T. castaneum*, *A. gambiae* and *D. melanogaster* curves there are fewer genes present in the fraction of the genome less than a given percentage G+C than would be expected if the genes were randomly distributed. In contrast, *A. mellifera* exhibits the opposite distribution.

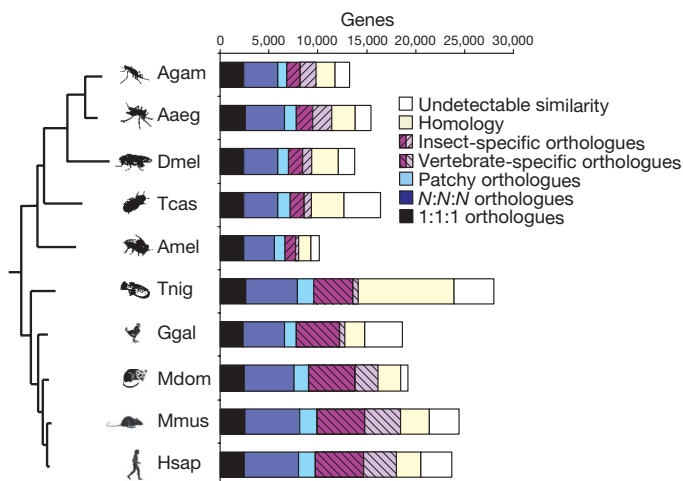


Figure 2 | Insect gene orthology. Comparison of the gene repertoire in five insect and five vertebrate genomes, ranging from the core of metazoan genes (dark blue fraction on the left) to the species-unique sequences (white band on the right). The striped boxes correspond to insect- and vertebrate-specific orthologous genes, where the darker bands correspond to all insects or vertebrates (allowing one loss). N:N indicates orthologues present in multiple copies in all species (allowing one loss); patchy indicates ancient orthologues (requiring at least one insect and one vertebrate gene) that have become differentially extinct in some lineages. The species tree on the left (shown in detail in Supplementary Fig. 6) was computed using the maximum-likelihood approach on concatenated sequences of 1,150 universal single-copy orthologues. It shows an accelerated rate of evolution in insects and confirms the basal position of the Hymenoptera within the Holometabola¹⁷. Aaeg, *Aedes aegypti*; Agam, *Anopheles gambiae*; Amel, *Apis mellifera*; Dmel, *Drosophila melanogaster*; Ggal, *Gallus gallus*; Hsap, *Homo sapiens*; Mdom, *Monodelphis domestica*; Mmus, *Mus musculus*; Tcas, *Tribolium castaneum*; Tnig, *Tetraodon nigroviridis*.

Gene family expansion, frequently associated with a particular adaptation pressure, might reveal physiologically and phenotypically unique features of beetles (Supplementary Table 9 and protein family discussions below). Many duplications shaped the gene content of *Tribolium*, most notably among odorant-binding proteins and the CYP450 subfamilies CYP6 and CYP9 (Supplementary Fig. 11), some of which are involved in the development of insecticide resistance in the Diptera¹⁹. Duplication of genes under copy-number selection in other species is indicative of species-specific neo-functionalization²⁰. At least 152 genes duplicated in *Tribolium* have single-copy status in all other insects studied, including sevenfold duplication of genes orthologous to *Drosophila* CG1625, encoding a putative structural constituent of cytoskeleton, and human ENSP00000269392, encoding centrosomal pre-acrosome localization protein 1.

We also analysed the phylogenetic distribution of orthologous gene group members to quantify evolutionary gene losses²¹. Although least affected, dozens of single-copy orthologues seem to be lost in each lineage. Thirty-eight such genes lost in *Tribolium* include rather unique genes, encoding phosphotriesterase-related protein and peroxisome assembly factor 1 (peroxin-2), compared to 59 such genes lost in *Drosophila*. Notably, for the less restricted fractions of orthologues (defined in Fig. 2), several hundred gene orthologues have been lost in each species.

Analysis of specific gene sets

In addition to a global automated analysis of the predicted *Tribolium* gene set, the consortium manually annotated and analysed ~2,000 genes (some additionally subjected to RNAi and expression analysis), focusing on developmental processes and genes of importance for agriculture and pest management.

Development

We identified and analysed homologues of known insect and vertebrate developmental genes to gain novel insights into the molecular basis of developmental differences between *Drosophila* and *Tribolium*. Supplementary Table 11 lists selected *Tribolium* developmental genes and their *Drosophila* and *Apis* orthologues.

Oogenesis. Despite profound differences in ovarian architecture—telotrophic versus polytrophic—we identified *Tribolium* orthologues

of most *Drosophila* genes required for stem cell maintenance, RNA localization and axis formation. Like *Apis*, however, *Tribolium* lacks a *bag of marbles* orthologue, which is essential for the differentiation of cystoblast versus germline stem cells in *Drosophila*²². Interestingly, an orthologue of the gene *gld-1*, which fulfils a similar function in *C. elegans*²³, is present in *Tribolium*.

Anterior–posterior patterning. Analysis of the genome sequence confirmed the absence of a *bcd* orthologue in *Tribolium*. Instead, anterior patterning is synergistically organized by *otd* and *hb* (ref. 23). However, it is still unclear how the posterior gradient of *Tribolium* Caudal is shaped in the absence of Bicoid. Notably, *Tribolium* contains an orthologue of *mex-3*, a factor that translationally represses the *C. elegans* *cad* homologue²⁴. Although the *Tribolium* genome contains orthologues of the *Drosophila* segmentation genes, their functions are not entirely conserved^{25–27}. Furthermore, the genome reveals the unexpected polycistronic organization of a novel gap gene, *mille-pattes*²⁸, the transcript of which encodes several short peptides.

In contrast to the classical protostomian model organisms *Drosophila* and *Caenorhabditis*, *Hox* genes in *Tribolium* map to a single cluster of ~750 kb on linkage group 2. Orthologues of all *Drosophila* *Hox* genes and the *Hox*-derived genes *ftz* and *zen* are transcribed from the same strand, and we find no evidence for inter-spersion of other protein-coding genes. Taken together, these results suggest that the evolutionary constraints preserving *Hox* cluster integrity still function in *Tribolium*.

Dorso-ventral patterning. As in *Drosophila*, the dorso-ventral axis of the *Tribolium* embryo depends on a nuclear gradient of Dl, an NF- κ B protein, which is established through ventral activation of a Tl receptor²⁹ (one of four in *Tribolium*). Factors required for localized Tl activation are also present in *Tribolium* (potential Tl ligands: six *spz*-like genes; extracellular proteases: one *gd*, six *snk* and four tandem *ea* genes), suggesting that, as in *Drosophila*, an extra-embryonic signal induces the embryonic dorso-ventral axis.

Tribolium *sog* inhibition of Dpp/BMP generates a patterning gradient along the dorso-ventral axis³⁰. Similar *chordin/sog* function in spiders and a hemichordate suggest that this may represent the ancestral bilaterian condition³⁰. Like *Apis*, *Tribolium* lacks an orthologue of *Drosophila* *scw*, but knockdown of another ligand, *Tribolium* *gbb1*, affected the embryonic Dpp/BMP gradient. *Tribolium* contains orthologues of all five *Drosophila* TGF- β receptors; however, Dpp signalling moderators that have duplicated and diverged in *Drosophila*, such as *Tol/tok* and *Cv/tsg*, occur as a single copy in *Tribolium*. Most strikingly, *Tribolium* contains homologues of *BMP10* as well as *bambi*, *Dan* and *gremlin* BMP inhibitors, which are all known from vertebrates, but are not found in *Drosophila*.

The growth zone. We identified several members of the Fgf and Wnt signalling pathways. The expression patterns of *Tribolium* *Fgf8*, *Wnt1*, *Wnt5* and *WntD/8* (refs 31, 32) highlight the dynamic organization of the growth zone and underline its role in axis elongation.

Head patterning. Orthologues of 25 out of 30 key regulators of the vertebrate anterior neural plate are specifically expressed in the *Tribolium* embryonic head (Supplementary Table 12). Two orthologues are not expressed in the head neuroectoderm (*barH*, *arx*) and three do not have *Tribolium* or *Drosophila* orthologues (*vax*, *hex1*, *atx*). Of the canonical *Drosophila* head gap genes, only the late head-patterning function of *otd* is conserved. *ems* function is restricted to parts of the antennal and ocular segments, and knockdown of *btd* seems to have no phenotypic consequences. Thus, analysis of *Tribolium* genes defines a set of genes that is highly conserved in bilaterian head development, and underscores the derived mode of *Drosophila* head patterning.

Leg and wing development. In contrast to *Drosophila*, ventral appendages in *Tribolium* develop during embryogenesis from buds that grow continuously along the proximo-distal axis³³. Nonetheless, we identified *Tribolium* orthologues for most of a core set of *Drosophila* appendage genes (Supplementary Table 13). On the other

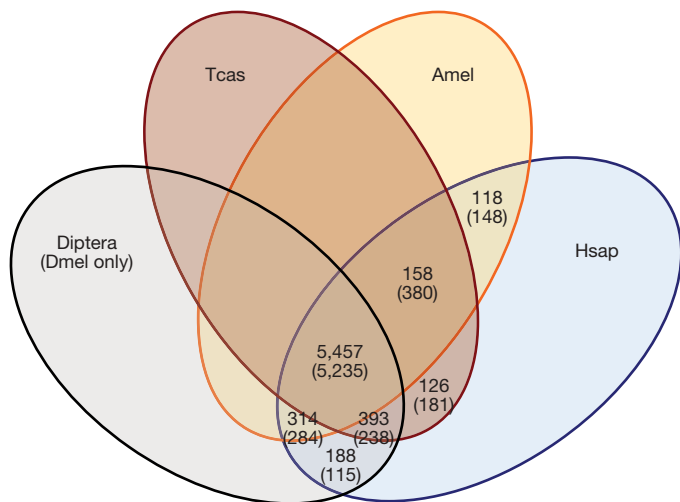


Figure 3 | Orthologous genes shared between insect and human genomes. The Venn diagram shows the number of orthologous groups of genes shared between the insect and human genomes. In addition to the majority of Urbilateria (last common ancestor of the Bilateria) genes shared by all the organisms, there are hundreds of genes that have been lost in some lineages (for example, only retained between human and *Tribolium* or human and honeybee, but lost in Diptera). Diptera is represented here by *Anopheles gambiae*, *Aedes aegypti* and *Drosophila melanogaster* (with numbers considering only *D. melanogaster* shown in parentheses).

hand, orthologues of genes not found in *Drosophila*, such as *Wnt11*, *gremlin*, *Fgf8* and an F-Box gene, are expressed in the embryonic legs^{21,31}. Although their exact function in *Tribolium* appendages is not known, *Fgf8* is essential to vertebrate limb development.

A major innovation driving the radiation of beetles was the evolution of a highly modified protective forewing. Expression analysis and RNAi experiments revealed a high degree of conservation between *Tribolium* and *Drosophila* wing gene networks (Supplementary Table 13), supporting the hypothesis that sclerotized elytra evolved from ancestral membranous wings mainly through new interactions between conserved patterning modules and as yet unknown downstream effector genes.

Eye development. *Tribolium* has orthologues of nearly all genes currently known to regulate specification and differentiation in the *Drosophila* retina (Supplementary Table 14). Exceptions are the linker protein Phyllopod and the lens crystallin protein Drosocrystallin, which are restricted to Diptera. Eight of fifty-seven investigated eye developmental genes are duplicated in the *Drosophila* genome but not *Tribolium*, and in four cases the *Drosophila* paralogues have similar function. This suggests a more dynamic evolution of *Drosophila* retina genes and higher genetic complexity, highlighting the value of *Tribolium* as a more ancestral and simply organized model of insect eye development.

Genes relevant to pest and *Tribolium* biology

Tribolium castaneum is a notorious invader of stored grains and grain products. Resultantly, much effort and expense is directed to find better ways to control this and other grain pests. Here we describe established and possible future pesticide targets, as well as genes underlying vision and taste. Finally, we describe genes forming the basis of systemic RNAi in *Tribolium*.

Established insecticide targets

Cys-loop ligand-gated ion channels. Members of this superfamily mediate chemical synaptic transmission in insects and are targets of successful pest control chemicals with animal health and crop protection applications³⁴. The *Tribolium* Cys-loop ligand-gated ion channel (Cys-loop LGIC) superfamily contains 24 genes, the largest known so far for insects (*Drosophila* and *Apis* superfamilies comprise 23 and 21 genes, respectively), due in part to the additional nicotinic acetylcholine receptor (nAChR) subunits in *Tribolium*. We also found genes for ion channels gated by γ -aminobutyric acid (γ -aminobutyric acid receptors (GABARs)), glutamate (GluCls) and histamine, as well as orthologues of the *Drosophila* pH-sensitive chloride channel³⁵. The molecular diversity of the *Tribolium* Cys-loop LGIC superfamily is broadened by alternative splicing and RNA A-to-I editing, which in some cases generates species-specific receptor isoforms³⁵. The *Tribolium* Cys-loop LGIC superfamily is the first complete set of genes encoding molecular targets of several insecticides—imidacloprid and other neonicotinoids (nAChRs), fipronil (GABARs) and avermectins (GluCls)—described for an agricultural pest species.

Cytochrome P450 proteins. Most insect cytochrome P450 proteins (CYPs) are thought to be involved in metabolic detoxification of host plant allelochemicals and toxicants, and several are insecticide resistance genes³⁶. Other CYPs act in the synthesis and degradation of lipid signalling molecules, such as ecdysteroids³⁷. Similarly to mosquitoes, especially *Aedes*, *Tribolium* has an independently expanded CYP gene family, particularly those involved in environmental response (Supplementary Table 16).

Within the *Tribolium* P450s, the CYP2 and mitochondrial clans have undergone relatively little gene expansion, lack pseudogenes, and are probably reserved for essential endogenous functions in ecdysteroid metabolism and development. In contrast, expansions via tandem duplication produced 85% of *Tribolium* P450s clustered in groups of 2–16 genes, with large expansions of CYP3 and CYP4 clans involved in environmental response. In comparison, *Apis* has

only four CYP4 genes, whereas *Aedes* has relatively similarly sized expansions of CYP3 and CYP4 clans (Supplementary Table 16). We speculate that both mosquito larvae (which are omnivorous scavengers) and *Tribolium* have adapted to diverse chemical environments in part by expansion of CYP gene families involved in detoxification.

Possible future insect control targets

C1 cysteine peptidase genes. *Tribolium castaneum* has successfully exploited cereal grains in spite of the arsenal of defensive allelochemicals, including inhibitors of serine peptidase digestive enzymes. In tenebrionid beetles, cathepsins B, L and serine peptidases such as trypsins and chymotrypsins are part of the digestive peptidase complex in the larval gut³⁸.

Comparing potential digestive peptidase genes in *Tribolium* with those in other sequenced insects (Supplementary Fig. 12) we found more C1 cysteine peptidase genes in *T. castaneum*. The proliferation of *Tribolium* C1 cysteine peptidase genes reflects expansions into five gene families, corresponding to four major clusters. This expansion is consistent with a trend seen in some beetles relative to other insects: a shift to a more acidic gut, conducive to cysteine peptidase activity.

Tribolium castaneum C1 cysteine peptidase genes encode B and L cathepsins, and include the first-known insect genes similar to O and K cathepsins (Supplementary Table 17). Most of the cathepsin-B-like peptidases lack conserved residues in functional regions and thus may lack peptidase activity, whereas all but two *Tribolium* cathepsin L peptidase genes encode potentially functional enzymes. In vertebrates, O and K cathepsins are lysosomal cysteine peptidases, involved in bone remodelling and resorption. Analysis of *Tribolium* cathepsins may provide insight into this family of proteins whose elevated expression is associated with a significant fraction of human breast cancers and tumour invasiveness.

Neurohormones and G-protein-coupled receptors. Insect neurohormones (neuropeptides, protein hormones and biogenic amines) control development, reproduction, behaviour, feeding and many other physiological processes, often by signalling through G-protein-coupled receptors (GPCRs). We found 20 genes encoding biogenic amine GPCRs in *Tribolium* (compared to 21 in *Drosophila* and 19 in *Apis*) and 52 genes encoding neuropeptide or protein hormone GPCRs (49 in *Drosophila*, 37 in *Apis*³⁹). Moreover, we identified the likely ligands for 45 of these 72 *Tribolium* GPCRs. Furthermore, we annotated 39 neuropeptide and protein hormone genes. We found excellent agreement (95%) between the proposed ligands for the *Tribolium* neurohormone GPCRs and the independently annotated neuropeptide and protein hormone genes. Interestingly, the *Tribolium* genome contains a vasopressin-like neuropeptide (TC06626) and a vasopressin-like GPCR gene (TC16363; Supplementary Fig. 14), neither of which has been detected in any other sequenced insect³⁹. Vasopressin in mammals is the major neurohormone stimulating water reabsorption in the kidneys⁴⁰. Its presence in *Tribolium* may help the beetle to survive in very dry habitats.

Genes relevant to *Tribolium* biology

Vision. Most of the 21 investigated genes that participate in the *Drosophila* photo-transduction network are conserved in *Tribolium* (Supplementary Table 14). Most notable is the lack of *ninaG* and *inaC*, which may be functionally replaced by closely related paralogues in *Tribolium*.

Tribolium contains only two opsin genes, representing members of the long-wavelength and ultraviolet-sensitivity-facilitating opsin subgroups. In contrast, *Drosophila* contains seven, and there is evidence for minimally three in most other insects. The lack of a blue-light-sensitive opsin gene in *Tribolium* is consistent with the unusual expression of long-wavelength opsin in all photoreceptor cells in this species⁴¹. The implied reduction in colour discrimination in *Tribolium* is probably a consequence of the widespread cryptic life-style of this species group.

Odorant and gustatory receptors. Odorant and gustatory receptors form the insect chemoreceptor superfamily. *Tribolium* has a major expansion of both odorant and gustatory receptors relative to *Drosophila*, *Anopheles* and *Aedes* mosquitoes, silkworm and honeybee (Supplementary Table 19). We identified and annotated 265 apparently functional odorant receptors, 42 full-length pseudogenes and 34 pseudogene fragments. Most of these *T. castaneum* odorant receptors are in seven species-specific subfamilies, including one containing 150 genes, and most are in tandem gene arrays, created by gene duplication within the *Tribolium* lineage in the last 300 million years⁴².

We annotated 220 apparently functional gustatory receptors and 25 pseudogenes (gustatory receptor gene fragments were not assessed). The gustatory receptor families in fruitflies and mosquitoes, but not honeybee, contain several genes that are alternatively spliced, with multiple alternative long first exons encoding at least the amino-terminal 50% of the gustatory receptor spliced into a set of short shared exons encoding the carboxy terminus^{43–45}. Most *Tribolium* gustatory receptors are encoded by single genes; however, *T. castaneum* Gr214 is a massive alternatively spliced locus with 30 alternative long 5' exons (six of which are pseudogenic) spliced into three shared 3' exons encoding the C terminus. Three *T. castaneum* gustatory receptors are orthologues of highly conserved gustatory receptors in other insects^{43–45}, two of which form a heterodimeric carbon dioxide receptor⁴⁶. The remainder form many species-specific subfamilies, one of which is expanded to 88 genes (Supplementary Information and Supplementary Fig. 16).

Systemic RNAi. In *Tribolium*, as in *C. elegans* but not *Drosophila*, the RNAi effect spreads systemically from the site of injection to other tissues⁵ and from injected females to their offspring⁴. Surprisingly, our survey of genes involved in systemic RNA did not reveal much conservation between *Tribolium* and *C. elegans*.

The SID-1 multi-transmembrane protein, essential for double-stranded RNA (dsRNA) uptake in *C. elegans*, is not found in *Drosophila*, suggesting that the presence or absence of a *sid-1* gene is the primary determinant of whether or not systemic RNAi occurs in an organism. We found three genes in *Tribolium* that encode proteins similar to SID-1. However, their sequences are more similar to another *C. elegans* protein, TAG-130 (also known as ZK721.1), which is not required for systemic RNAi in *C. elegans*⁴⁷. Additionally, the secondary argonaute proteins and RNA-dependent RNA polymerase (RdRP)^{48,49}, essential for the amplification of the initial dsRNA trigger in *C. elegans*, are absent in *Tribolium*. Therefore, the molecular basis for systemic RNAi in *Tribolium* and other insects might differ from that in *C. elegans* and remains to be elucidated.

Concluding remarks

We observe three trends when comparing *Tribolium* and other insect genomes. First, phylogenetic trees show shorter branch lengths for *Tribolium* (and *Apis*) than *Drosophila*. The accelerated evolution of the *Drosophila* lineage in some cases rendered *Drosophila* atypical for the Insecta. Second, *Tribolium* retains a different set of ancestral genes that have evolved at a moderate rate (for example, *gremlin* and cathepsins), and these may provide insights into the function of their vertebrate orthologues. Third, its own evolutionary path has led to beetle- and perhaps *Tribolium*-specific gene changes (for example, a large increase in odorant receptors).

Expansions of CYP proteins, proteinases, diuretic hormones, a vasopressin hormone and receptor, and chemoreceptors all indicate adaptation to a dry, chemically diverse and toxin-rich microenvironment. Whereas the flour beetle's drought tolerance probably explains the presence of vasopressin, it is more difficult to rationalize a need for such an unprecedented diversity of chemoreceptors. Functions stemming from the diversity of angiosperm-derived chemicals such as distant detection of food sources and avoidance of toxic host plant defence chemicals suggest that this expansion may be common to the Coleoptera. The expansion of odorant receptors is more intriguing

when considered in combination with the reduction of opsin genes. Both trends may reflect the long-term consequences of adaptation to low light biota by *Tribolium*, enforcing selection for increased discrimination of odour reception but not colour perception⁴¹.

Given the chemo-sensing and detoxifying genes described above, it is perhaps no surprise that *Tribolium* has demonstrated resistance to all insecticides used for its control. Given the association of *Tribolium* with human food, knowledge of all possible insecticide targets will aid greater selectivity in pesticide design, thereby mitigating possible side effects. Finally, the true value of this sequence may be the entry it provides into the many and richly diverse facets of beetle biology, physiology and behaviour.

METHODS SUMMARY

Detailed Methods are described in the Supplementary Information. Resources generated by this project can be found at the following locations: genome assemblies, sequences, and automated and manually curated gene model sequences are available from the BCM-HGSC website and ftp site (<http://www.hgsc.bcm.tmc.edu/projects/tribolium/>). Browser display of the genome sequence, all gene predictions and available tiling array data are available via <http://www.genboree.org> and Beetle Base (<http://www.bioinformatics.ksu.edu/BeetleBase/>), a long-term repository for *Tribolium* data.

Received 16 July 2007; accepted 6 February 2008.

Published online 23 March 2008.

- Hunt, T. *et al.* A comprehensive phylogeny of beetles reveals the evolutionary origins of a superradiation. *Science* **318**, 1913–1916 (2007).
- Sokoloff, A. *The Biology of Tribolium with Special Emphasis on Genetic Aspects I–III* (Clarendon Press and Oxford Univ. Press, Oxford, 1972, 1974, 1977).
- Lorenzen, M. D. *et al.* piggyBac-based insertional mutagenesis in *Tribolium castaneum* using donor/helper hybrids. *Insect Mol. Biol.* **16**, 265–275 (2007).
- Bucher, G., Scholten, J. & Klingler, M. Parental RNAi in *Tribolium* (Coleoptera). *Curr. Biol.* **12**, R85–R86 (2002).
- Tomoyasu, Y. & Denell, R. E. Larval RNAi in *Tribolium* (Coleoptera) for analyzing adult development. *Dev. Genes Evol.* **214**, 575–578 (2004).
- Tautz, D., Friedrich, M. & Schröder, R. in *Development 1994 Supplement* (eds Akam, M., Holland, P., Ingham, P. & Wray, G.) 193–199 (The Company of Biologists Limited, Cambridge, 1994).
- Tautz, D. Segmentation. *Dev. Cell* **7**, 301–312 (2004).
- Lorenzen, M. D. *et al.* Genetic linkage maps of the red flour beetle, *Tribolium castaneum*, based on bacterial artificial chromosomes and expressed sequence tags. *Genetics* **170**, 741–747 (2005).
- Demuth, J. P. *et al.* Genome-wide survey of *Tribolium castaneum* microsatellites and description of 509 polymorphic markers. *Mol. Ecol. Notes* **7**, 1189–1195 (2007).
- Wang, S. & Brown, S. J. Analysis of repetitive DNA distribution patterns in the *Tribolium castaneum* genome. *Genome Biol.* (in the press).
- Brown, S. J., Henry, J. K., Black, W. C. IV & Denell, R. Molecular Genetic manipulation of the red flour beetle: Genome organization and cloning of a ribosomal protein gene. *Insect Biochem.* **20**, 185–193 (1990).
- Ugarkovic, D., Podnar, M. & Plohl, M. Satellite DNA of the red flour beetle *Tribolium castaneum*—comparative study of satellites from the genus *Tribolium*. *Mol. Biol. Evol.* **13**, 1059–1066 (1996).
- Osanai, M., Kojima, K. K., Futahashi, R., Yaguchi, S. & Fujiwara, H. Identification and characterization of the telomerase reverse transcriptase of *Bombyx mori* (silkworm) and *Tribolium castaneum* (flour beetle). *Gene* **376**, 281–289 (2006).
- Xiong, Y. & Eickbush, T. H. The site-specific ribosomal DNA insertion element R1Bm belongs to a class of non-long-terminal-repeat retrotransposons. *Mol. Cell. Biol.* **8**, 114–123 (1988).
- Robertson, H. M. & Gordon, K. H. Canonical TTAGG-repeat telomeres and telomerase in the honey bee, *Apis mellifera*. *Genome Res.* **16**, 1345–1351 (2006).
- Fujiwara, H., Osanai, M., Matsumoto, T. & Kojima, K. K. Telomere-specific non-LTR retrotransposons and telomere maintenance in the silkworm, *Bombyx mori*. *Chromosome Res.* **13**, 455–467 (2005).
- Savard, J. *et al.* Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects. *Genome Res.* **16**, 1334–1338 (2006).
- Savard, J., Tautz, D. & Lercher, M. J. Genome-wide acceleration of protein evolution in flies (Diptera). *BMC Evol. Biol.* **6**, 7 (2006).
- Daborn, P. J. *et al.* A single p450 allele associated with insecticide resistance in *Drosophila*. *Science* **297**, 2253–2256 (2002).
- Ciccarelli, F. D. *et al.* Complex genomic rearrangements lead to novel primate gene function. *Genome Res.* **15**, 343–351 (2005).
- Wyder, S., Kriventseva, E. V., Schröder, R., Kadowaki, T. & Zdobnov, E. M. Quantification of ortholog losses in insects and vertebrates. *Genome Biol.* **8**, R242 (2007).
- Wong, M. D., Jin, Z. & Xie, T. Molecular mechanisms of germline stem cell regulation. *Annu. Rev. Genet.* **39**, 173–195 (2005).

23. Schröder, R. The genes *orthodenticle* and *hunchback* substitute for *bicoid* in the beetle *Tribolium*. *Nature* **422**, 621–625 (2003).
24. Draper, B. W., Mello, C. C., Bowerman, B., Hardin, J. & Priess, J. R. MEX-3 is a KH domain protein that regulates blastomere identity in early *C. elegans* embryos. *Cell* **87**, 205–216 (1996).
25. Bucher, G. & Klingler, M. Divergent segmentation mechanism in the short germ insect *Tribolium* revealed by giant expression and function. *Development* **131**, 1729–1740 (2004).
26. Cerny, A. C., Bucher, G., Schröder, R. & Klingler, M. Breakdown of abdominal patterning in the *Tribolium* *Kruppel* mutant *jaws*. *Development* **132**, 5353–5363 (2005).
27. Choe, C. P., Miller, S. C. & Brown, S. J. A pair-rule gene circuit defines segments sequentially in the short-germ insect *Tribolium castaneum*. *Proc. Natl Acad. Sci. USA* **103**, 6560–6564 (2006).
28. Savard, J., Marques-Souza, H., Aranda, M. & Tautz, D. A segmentation gene in *Tribolium* produces a polycistronic mRNA that codes for multiple conserved peptides. *Cell* **126**, 559–569 (2006).
29. Fonseca, R. N. *et al.* Self-regulatory circuits in dorsoventral axis formation of the short germ beetle *Tribolium castaneum*. *Dev. Cell* (in the press).
30. van der Zee, M., Stockhammer, O., von Levetzow, C., Nunes da Fonseca, R. & Roth, S. *Sog/Chordin* is required for ventral-to-dorsal *Dpp/BMP* transport and head formation in a short germ insect. *Proc. Natl Acad. Sci. USA* **103**, 16307–16312 (2006).
31. Beermann, A. & Schröder, R. Sites of FGF signalling and perception during embryogenesis of the beetle *Tribolium castaneum*. *Dev. Genes Evol.* (in the press).
32. Bolognesi, R. *et al.* *Tribolium* Wnts: evidence for a larger repertoire in insects with overlapping expression patterns that suggest multiple redundant functions in embryogenesis. *Dev. Genes Evol.* (in the press).
33. Beermann, A. *et al.* The *Short antennae* gene of *Tribolium* is required for limb development and encodes the orthologue of the *Drosophila* Distal-less protein. *Development* **128**, 287–297 (2001).
34. Raymond-Delpech, V., Matsuda, K., Sattelle, B. M., Rauh, J. J. & Sattelle, D. B. Ion channels: molecular targets of neuroactive insecticides. *Invert. Neurosci.* **5**, 119–133 (2005).
35. Jones, A. K. & Sattelle, D. B. The cys-loop ligand-gated ion channel gene superfamily of the red flour beetle, *Tribolium castaneum*. *BMC Genomics* **8**, 327 (2007).
36. Ffrench-Constant, R. H., Daborn, P. J. & Le Goff, G. The genetics and genomics of insecticide resistance. *Trends Genet.* **20**, 163–170 (2004).
37. Ono, H. *et al.* *Spook* and *Spookier* code for stage-specific components of the ecdysone biosynthetic pathway in Diptera. *Dev. Biol.* **298**, 555–570 (2006).
38. Vinokurov, K. S. *et al.* Diversity of digestive proteinases in *Tenebrio molitor* (Coleoptera: Tenebrionidae) larvae. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **145**, 126–137 (2006).
39. Hauser, F., Cazzamali, G., Williamson, M., Blenau, W. & Grimmelikhuijzen, C. J. A review of neurohormone GPCRs present in the fruitfly *Drosophila melanogaster* and the honey bee *Apis mellifera*. *Prog. Neurobiol.* **80**, 1–19 (2006).
40. Bankir, L. Antidiuretic action of vasopressin: quantitative aspects and interaction between V1a and V2 receptor-mediated effects. *Cardiovasc. Res.* **51**, 372–390 (2001).
41. Jackowska, M. *et al.* Genomic and gene regulatory signatures of cryptozoic adaptation: loss of blue sensitive photoreceptors through expansion of long wavelength-opsin expression in the red flour beetle *Tribolium castaneum*. *Front. Zool.* **4**, 24 (2007).
42. Engstontia, P. *et al.* The red flour beetle's large nose: an expanded odorant receptor gene family in *Tribolium castaneum*. *Insect Biochem. Mol. Biol.* (in the press).
43. Clyne, P. J., Warr, C. G. & Carlson, J. R. Candidate taste receptors in *Drosophila*. *Science* **287**, 1830–1834 (2000).
44. Hill, C. A. *et al.* G protein-coupled receptors in *Anopheles gambiae*. *Science* **298**, 176–178 (2002).
45. Robertson, H. M., Warr, C. G. & Carlson, J. R. Molecular evolution of the insect chemoreceptor gene superfamily in *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **100** (suppl. 2), 14537–14542 (2003).
46. Jones, W. D., Cayirlioglu, P., Kadow, I. G. & Vosshall, L. B. Two chemosensory receptors together mediate carbon dioxide detection in *Drosophila*. *Nature* **445**, 86–90 (2007).
47. Tomoyasu, Y. *et al.* Exploring systemic RNA interference in insects; a genome-wide survey for RNAi genes in *Tribolium*. *Genome Biol.* (in the press).
48. Lipardi, C., Wei, Q. & Paterson, B. M. RNAi as random degradative PCR: siRNA primers convert mRNA into dsRNAs that are degraded to generate new siRNAs. *Cell* **107**, 297–307 (2001).
49. Sijen, T. *et al.* On the role of RNA amplification in dsRNA-triggered gene silencing. *Cell* **107**, 465–476 (2001).

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

Acknowledgements Work at the BCM-HGSC was funded by grants from the NHGRI and USDA. FgenesH and FgenesH++ analysis was donated by Softberry Inc. This research was additionally supported in part by the Intramural Research Program of the NIH, National Library of Medicine.

Author Information The *Tribolium* genome sequence, at the NCBI, has project accession AAJJ00000000. Reprints and permissions information is available at

www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and is freely available to all readers at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.R. (stephenr@bcm.edu).

The *Tribolium* Genome Sequencing Consortium

Project leader: Stephen Richards^{1,2}

Principal investigators: Richard A. Gibbs^{1,2}, George M. Weinstock^{1,2}

White paper: Susan J. Brown³, Robin Denell³, Richard W. Beeman⁴, Richard Gibbs^{1,2}

Analysis leaders: Richard W. Beeman⁴, Susan J. Brown³, Gregor Bucher⁵, Markus Friedrich⁶, Cornelis J. P. Grimmelikhuijzen⁷, Martin Klingler⁸, Marce Lorenzen³, Stephen Richards^{1,2}, Siegfried Roth⁹, Reinhard Schröder^{10,†}, Diethard Tautz¹¹, Evgeny M. Zdobnov^{12,13,14}

DNA sequence and global analysis: DNA sequencing Donna Muzny (leader)^{1,2}, Richard A. Gibbs^{1,2}, George M. Weinstock^{1,2}, Tony Attaway^{1,2}, Stephanie Bell^{1,2}, Christian J. Buhay^{1,2}, Mimi N. Chandrabose^{1,2}, Dean Chavez^{1,2}, Kerstin P. Clerk-Blankenburg^{1,2}, Andrew Cree^{1,2}, Marvin Dao^{1,2}, Clay Davis^{1,2}, Joseph Chacko^{1,2}, Huyen Dinh^{1,2}, Shannon Dugan-Rocha^{1,2}, Gerald Fowler^{1,2}, Toni T. Garner^{1,2}, Jeffrey Ganes^{1,2}, Andreas Gnirke^{1,2}, Alica Hawes^{1,2}, Judith Hernandez^{1,2}, Sandra Hines^{1,2}, Michael Holder^{1,2}, Jennifer Hume^{1,2}, Shalini N. Jhangiani^{1,2}, Vandita Joshi^{1,2}, Ziad Mohid Khan^{1,2}, LaRonda Jackson^{1,2}, Christie Kovar^{1,2}, Andrea Kowis^{1,2}, Sandra Lee^{1,2}, Lora R. Lewis^{1,2}, Jon Margolis^{1,2}, Margaret Morgan^{1,2}, Lynne V. Nazareth (leader)^{1,2}, Ngoc Nguyen^{1,2}, Geoffrey Okwuonu^{1,2}, David Parker^{1,2}, Stephen Richards^{1,2}, San-Juana Ruiz^{1,2}, Jireh Santibanez^{1,2}, Joël Savard¹¹, Steven E. Scherer^{1,2}, Brian Schneider^{1,2}, Erica Sodergren^{1,2}, Diethard Tautz¹¹, Selina Vattahil^{1,2}, Donna Villasana^{1,2}, Courtney S. White^{1,2}, Rita Wright^{1,2}; **EST sequencing** Yoonseong Park¹⁸, Richard W. Beeman⁴, Jeff Lord⁴, Brenda Oppert⁴, Marce Lorenzen⁴, Susan Brown³, Liangjiang Wang³, Joël Savard¹¹, Diethard Tautz¹¹, Stephen Richards¹, George Weinstock^{1,2}, Richard A. Gibbs^{1,2}; **genome assembly** Yue Liu^{1,2}, Kim Worley^{1,2}, George Weinstock^{1,2}; **G+C content** Christine G. Elsik¹⁹, Justin T. Reese¹⁹, Eran Elhaik²⁰, Giddy Landan²⁰, Dan Graur²⁰; **repetitive DNA, transposons and telomeres** Peter Arensburger²¹, Peter Atkinson²¹, Richard W. Beeman⁴, Jim Beidler²², Susan J. Brown³, Jeffery P. Demuth²³, Douglas W. Drury²⁴, Yu-Zhou Du²⁵, Haruhiko Fujiwara²⁶, Marce Lorenzen³, Vincenza Maselli²⁷, Mizuko Osana²⁶, Yoonseong Park¹⁸, Hugh M. Robertson²⁸, Zhijian Tu²², Jian-jun Wang²⁵, Suzhi Wang³; **gene prediction and consensus gene set** Stephen Richards^{1,2}, Henry Song^{1,2}, Lan Zhang^{1,2}, Erica Sodergren^{1,2}, Doreen Werner²⁹, Mario Stanke²⁹, Burkhard Morgenstern²⁹, Victor Solovyev³⁰, Peter Kosarev³¹, Garth Brown³², Hsiu-Chuan Chen³², Olga Ermolaeva³², Wratko Hlavina³², Yuri Kapustin³², Boris Kyrutin³², Paul Kitts³², Donna Maglott³², Kim Pruitt³², Victor Sapojnikov³², Alexandre Souvorov³², Aaron J. Mackey³³, Robert M. Waterhouse¹⁴, Stefan Wyder¹², Evgeny M. Zdobnov^{12,13,14}; **global gene content analysis** Evgeny M. Zdobnov^{12,13,14}, Stefan Wyder¹², Evgenia V. Kriventseva^{12,34}, Tatsuhiko Kadowaki³⁵, Peer Bork^{36,37}

Developmental processes and signalling pathways: Manuel Aranda¹¹, Riyue Bao⁶, Anke Beermann¹⁰, Nicola Berns¹⁰, Renata Bolognesi³, François Bonneton³⁸, Daniel Bopp³⁹, Susan J. Brown³, Gregor Bucher⁵, Thomas Butts⁴⁰, Arnaud Chaumont⁴¹, Robin E. Denell³, David E. K. Ferrier⁴⁰, Markus Friedrich⁶, Cassandra M. Gordon³, Marek Jindra⁴², Martin Klingler⁸, Que Lan⁴³, H. Michael G. Lattorf⁴⁴, Vincent Laudet³⁸, Cornelia von Levetsow⁹, Zhenyi Liu⁴⁵, Rebekka Lutz¹⁰, Jeremy A. Lynch⁹, Rodrigo Nunes da Fonseca³, Nico Posnien⁵, Rolf Reuter¹⁰, Siegfried Roth⁹, Joël Savard¹¹, Johannes B. Schinko⁵, Christian Schmitt⁸, Michael Schoppmeier⁸, Reinhard Schröder¹⁰, Teresa D. Shippy³, Franck Simonnet⁵, Henrique Marques-Souza¹¹, Diethard Tautz¹¹, Yoshinori Tomoyasu³, Jochen Trauner⁸, Maurijn Van der Zee¹¹, Michel Vervoot⁴⁶, Nadine Wittkopp¹⁰, Ernst A. Wimmer⁵, Xiaoyun Yang⁶

Pest biology, senses, Medea and RNAi: ligand gated ion channels Andrew K. Jones⁴⁷, David B. Sattelle⁴⁷; **oxidative phosphorylation** Paul R. Ebert⁴⁸; **P450 genes** David Nelson⁴⁹, Jeffrey G. Scott⁵⁰, Richard W. Beeman⁴; **chitin and cuticular proteins** Subbaratnam Muthukrishnan⁵¹, Karl J. Kramer^{4,51}, Yasuyuki Arakane^{4,51}, Richard W. Beeman⁴, Qingsong Zhu⁵¹, David Hogenkamp⁵¹, Radhika Dixit⁵¹; **digestive proteinases** Brenda Oppert⁴, Haobo Jiang⁵², Zhen Zou⁵², Jeremy Marshall³, Elena Elpidina⁵³, Konstantin Vinokurov⁵³, Cris Oppert⁴; **immunity** Zhen Zou⁵², Jay Evans⁵⁴, Zhiqiang Lu⁵², Picheng Zhao⁵², Niranji Sumathipala⁵², Boran Altincicek⁵⁵, Andreas Vilcinskis⁵⁵, Michael Williams⁵⁶, Dan Hultmark⁵⁶, Charles Hetru⁵⁷, Haobo Jiang⁵²; **neurohormones and GPCRs** Cornelis J. P. Grimmelikhuijzen⁷, Frank Hauser⁷, Giuseppe Cazzamali⁷, Michael Williamson⁷, Yoonseong Park¹⁸, Bin Li¹⁸, Yoshiaki Tanaka⁵⁸, Reinhard Predel⁵⁹, Susanne Neupert⁵⁹, Joachim Schachtner⁶⁰, Peter Verleyen⁶¹; **neuropeptide processing enzymes** Florian Raible³⁶, Peer Bork^{36,37}; **opsins** Markus Friedrich⁶; **odorant receptors and gustatory receptors** Kimberly K. O. Walden²⁸, Hugh M. Robertson²⁸; **odorant binding and chemosensory proteins** Sergio Angeli⁶², Sylvain Forêt⁶³, Gregor Bucher⁵, Stefan Schuetz⁵, Ryszard Maleszka⁶³, Ernst A. Wimmer⁵; **Medea** Richard W. Beeman⁴, Marce Lorenzen⁴; **systemic RNAi** Yoshinori Tomoyasu³, Sherry C. Miller³, Daniela Grossmann⁵ & Gregor Bucher⁵

Affiliations for participants: ¹Human Genome Sequencing Center, and ²Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA. ³Division of Biology, Ackert Hall, Kansas State University, Manhattan, Kansas 66506, USA. ⁴Grain Marketing and Production Research Center, Agricultural Research Service, United States Department of Agriculture, 1515 College Avenue, Manhattan, Kansas 66502, USA. ⁵Johann Friedrich Blumenbach Institute, Department of Developmental Biology, Georg August University, von-Liebig-Weg-11, 37077 Göttingen, Germany. ⁶Department of Biological Sciences, Wayne State University, Detroit, Michigan 48202, USA. ⁷Center for Functional and Comparative Insect Genomics, and Department of Cell Biology and Comparative Zoology, Institute of Biology, University of Copenhagen, Universitetsparken 15, DK-2100 Copenhagen, Denmark. ⁸Institute for Biology, Department of Developmental Biology, Friedrich-Alexander-University Erlangen, Staudtstrasse 5, 91058 Erlangen, Germany. ⁹Institute for Developmental Biology, University of Cologne, 50674 Cologne, Germany. ¹⁰Animal Genetics, Interfaculty Institute for Cell Biology, University of Tübingen, Auf der Morgenstelle 28, 72076 Tübingen, Germany. ¹¹Department of Genetics, University of Cologne, 50674 Cologne, Germany. ¹²Department of Genetic Medicine and Development, University of Geneva Medical School, 1 rue Michel-Servet, 1211 Geneva, Switzerland. ¹³Swiss Institute of Bioinformatics, 1 rue Michel-Servet, 1211 Geneva, Switzerland. ¹⁴Imperial College London, South Kensington Campus, SW7 2AZ London, UK. ¹⁵Children's Hospital Oakland Research Institute, BACPAC Resources, 747 52nd Street, Oakland, California 94609, USA. ¹⁶Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. ¹⁷AgraQuest, Inc., 1530 Drew Avenue, Davis, California 95616, USA. ¹⁸Department of Entomology, Waters Hall, Kansas State University, Manhattan, Kansas 66506, USA. ¹⁹Department of Animal Science, Texas A&M University, College Station, Texas 77843, USA. ²⁰Department of Biology and Biochemistry, University of Houston, Houston, Texas 77204, USA. ²¹Department of Entomology, University of California, 900 University Avenue, Riverside, California 92521, USA. ²²Department of Biochemistry, Virginia Tech, Blacksburg, Virginia 24061, USA. ²³Department of Biology, University of Texas at Arlington, Arlington, Texas 76019, USA. ²⁴Department of Biology, Indiana University, Bloomington, Indiana 47405, USA. ²⁵Department of Plant Protection, Yangzhou University, Yangzhou 225009, China. ²⁶Department of Integrated Biosciences, Graduate School of Frontier Sciences, University of Tokyo, Bioscience Building 501, Kashiwa, Chiba 277-8562, Japan. ²⁷European School of Molecular Medicine and Telethon Institute of Genetics and Medicine, Via Pietro Castellino 111, 80131 Napoli, Italy. ²⁸Department of Entomology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA. ²⁹Institute for Microbiology and Genetics, Department of Bioinformatics, University of Göttingen, Goldschmidtstraße 1, 37077 Göttingen, Germany. ³⁰Department of Computer Science, Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK. ³¹Softberry Inc., 116 Radio Circle, Suite 400, Mount Kisco, New York 10549, USA. ³²National Center for Biotechnology Information, National Library of Medicine, Bethesda, Maryland 20894, USA. ³³GlaxoSmithKline, Collegeville, Pennsylvania 19426, USA. ³⁴Department of Structural Biology and Bioinformatics, University of Geneva Medical School, 1 rue Michel-Servet, 1211 Geneva, Switzerland. ³⁵Graduate School of Bioagricultural Sciences, Nagoya University, Chikusa, Nagoya 464-8601, Japan. ³⁶European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69117 Heidelberg, Germany. ³⁷Max-Delbrück-Centre for Molecular Medicine, Berlin-Buch, Robert-Roessle-Strasse 10, 13092 Berlin, Germany. ³⁸Institut de Genomique Fonctionnelle de Lyon, Equipe de Zoologie Moléculaire, ENS Lyon, Université Lyon 1, CNRS UMR5242, INRA, IFR128, 46 Allée d'Italie, 69364 Lyon cedex 07, France. ³⁹Zoological Institute of the University Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland. ⁴⁰Department of Zoology, University of Oxford, Tinbergen Building, South Parks Road, Oxford OX1 3PS, UK. ⁴¹CEMAGREF, Laboratoire d'écotoxicologie, 3bis quai Chauvea, CP220 69336 Lyon cedex 09, France. ⁴²Institute of Entomology ASCR, Branisovská 31, České Budejovice 370 05, Czech Republic. ⁴³Department of Entomology, University of Wisconsin-Madison, 1630 Linden Drive, Madison, Wisconsin 53706, USA. ⁴⁴Institute of Biology, Molecular Ecology, Martin-Luther-University Halle-Wittenberg Hoher Weg 4, 06099 Halle (Saale), Germany. ⁴⁵Department of Molecular Biology and Pharmacology, Washington University in St Louis School of Medicine, 3600 Cancer Research Building, 660 South Euclid Avenue, St Louis, Missouri 63110, USA. ⁴⁶Université Paris 7 – Denis Diderot, Centre de genétique moléculaire – CNRS UPR 2167, 1 Avenue de la Terrasse, 91198 Gif-sur-Yvette cedex, France. ⁴⁷MRC Functional Genetics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, South Parks Road, Oxford OX1 3QX, UK. ⁴⁸School of Integrative Biology & School of Molecular and Microbial Sciences, University of Queensland, St Lucia, Queensland 4072, Australia. ⁴⁹Department of Molecular Sciences and Center of Excellence in Genomics and Bioinformatics, University of Tennessee, Memphis, Tennessee 38163, USA. ⁵⁰Department of Entomology, Daljit and Elaine Sarkaria Professor of Insect Physiology and Toxicology, Cornell University, Ithaca, New York 14853, USA. ⁵¹Department of Biochemistry, Kansas State University, Manhattan, Kansas 66506, USA. ⁵²Department of Entomology and Plant Pathology, Oklahoma State University, Stillwater, Oklahoma 74078, USA. ⁵³A. N. Belozersky Institute of Physico-Chemical Biology, Moscow State University, Leninskie Gory, Moscow 119992, Russia. ⁵⁴USDA-ARS Bee Research Laboratory, Beltsville, Maryland 20705, USA. ⁵⁵Institute of Phytopathology and Applied Zoology, Interdisciplinary Research Center, Justus-Liebig-University of Giessen, Heinrich-Buff-Ring 26-32, D-35392 Giessen, Germany. ⁵⁶Umea Centre for Molecular Pathogenesis, Umea University, Umea SE-90187, Sweden. ⁵⁷Institut Biol Moléc Cell, CNRS, Strasbourg 67084, France. ⁵⁸National Institute of Agrobiological Science, Division of Insect Science, Tsukuba, Ibaraki 305-8634, Japan. ⁵⁹Institute of General Zoology, University of Jena, Erbertstrasse 1, D-07743 Jena, Germany. ⁶⁰Department of Animal Physiology, University of Marburg, Karl-von-Frisch Strasse 8, D-35032 Marburg, Germany. ⁶¹Department of Animal Physiology and Neurobiology, University of Leuven, Naamsestraat 59, BE-3000 Leuven, Belgium. ⁶²Institute for Forest Zoology and Forest Conservation, Büsgenweg 3 D-37077 Göttingen, Germany. ⁶³Visual Sciences and ARC Centre for the Molecular Genetics of Development, Research School of Biological Sciences, The Australian National University, Canberra, ACT 0200, Australia. †Present address: Bioscience Institute, University of Rostock, Albert-Einstein-Strasse 3, 18059 Rostock, Germany.

Supplementary Materials

Strain selection. The GA-2 strain used for sequencing is a highly inbred, near-homozygous version of GA-1, the latter having been collected in a farmer's corn bin in Georgia in 1983. GA-2 was created by Scott Thomson (University of Wisconsin, Parkside, Kenosha) by 20 consecutive generations of virgin single-pair, full-sib inbreeding. The near-homozygous inbred condition of GA-2 was confirmed by Southern hybridization analysis of hypervariable “snapback” loci¹. Such loci have become monomorphic in the GA-2 strain.

BAC library. A BAC library was prepared from DNA isolated from the GA2 sequencing strain at Exelixis (South San Francisco, CA). The Tc_Ba library is available from Clemson University Genomics Institute for a small distribution fee via the website <https://www.genome.clemson.edu/>.

Sequencing and assembly. A pure WGS approach was taken, similar to that used for the sequencing and assembly of *Drosophila pseudoobscura*². Short insert libraries were prepared according to a double adaptor strategy³. A fosmid library was prepared using the Epi-FOS vector (Epicentre Biotechnologies, Madison, WI). Paired end sequences were generated using Applied Biosystems (Foster City, CA) 3730 sequencing machines. The pure WGS Assembly was performed as previously described² using the Atlas suite of assembly tools⁴.

Known assembly issues. The chromosomes are not all uniform in their coverage as a mixture of male and female embryos was used for DNA isolation (isolation of nuclei from washed embryos avoids both mitochondrial and other contaminants such as gut contents and food source contamination). Because of this the X chromosome is sequenced to $\frac{3}{4}$ the coverage of the autosomes, and the (very small – perhaps 2% of the total genome) Y chromosome at only $\frac{1}{4}$ sequence coverage. Additionally, whole genome assembly software often has problems assembling highly repetitive sequences such as centromeres and telomeres, and these are expected to be under-represented in draft genome sequences, although manual efforts did identify some *Tribolium* telomeres – see below.

Assembly QC. 96% of ~41,000 EST sequences could be aligned to the whole genome sequence (WGS) assembly suggesting transcribed sequences are well represented in the genome. To assess sequence quality, we compared the WGS assembly to 795 kb of finished sequence derived from BAC clones from the GA2 strain. We found 99.33% of the finished BAC sequence present within the assembly. Of the aligned bases, only 0.19% had overlap within the WGS assembly; otherwise there was linear alignment between the WGS assembly and the finished BAC sequences, suggesting that mis-assembly is rare. The quality of the aligned sequence was generally extremely high, where except for a single reptig, a total of 5 substitutions and 38 indels were found in the aligned 795 kb (an error rate of ~ 1 in 18,000 bp – note that several years elapsed between isolating DNA for BAC library preparation and WGS sequencing from the same strain). The erroneous reptig sequence increased the total number of errors to 3,314, however we believe such reptigs errors are rare as other reptigs in the aligned region for this and other projects had zero errors.

Mapping. Scaffolds containing 70% of the genome sequence were initially pinned to a sequence-based genetic map⁵ by BLASTN of the mapped fragments to the genome scaffolds. After this process, the 40 largest, unpinned scaffolds were genetically mapped by PCR-SSCP in an attempt to increase the genome coverage of pinned scaffolds to 90%. For each scaffold, PCR primers that amplify unique fragments 200-300 nt in length were designed. PCR products were subjected to single-strand conformational polymorphism (SSCP) analysis, and those that identified a dimorphism between the mapping strains (GA-2 and another near-homozygous inbred strain, ab-2) were mapped using the same set of 179 backcross progeny DNAs used to generate the original map, giving a nominal map resolution of ~0.6%. After determining map order of scaffolds on the chromosomes, the sequence of each chromosome was defined by linking individual scaffold sequences with arbitrary, 300 kb spacer segments. When unknown, scaffold orientation was assigned randomly.

A note on sequence nomenclature. The relationship between linkage groups and chromosomes is difficult to determine. In particular, only the Y and one other chromosome corresponding to linkage group three can be distinguished. The Y can be

distinguished because of its small size (in addition, the X can be distinguished in males due to a “parachute” pairing in metaphase squashes, but not females), and the other chromosome because it is twice as large as the others, and likely corresponds to the largest linkage group: linkage group three.

The remaining eight chromosomes are indistinguishable in length or other characteristics at the current time (note, there is no equivalent of polytene chromosomes as seen in some diptera in *T. castaneum*). Given this state of affairs, we have named the chromosome length sequences generated using the term linkage group to link with the previously published linkage map⁵, and to underscore the difficulties of chromosome identification.

GC content. The *Tribolium* genome, like other animal genomes, is a mosaic of sequence stretches of variable length and GC composition. A comparison of the distributions of GC-content domain length among the *T. castaneum*, *A. mellifera*, *A. gambiae*, *D. pseudoobscura*, *D. melanogaster*, *D. simulans*, and *D. yakuba* genomes is shown in Fig. S2. The G-test of goodness-of-fit was used to determine that none of the segment length distributions are similar (see GC analysis methods below). Interestingly, *Tribolium* has the highest abundance of small-to-medium size GC content domains (15 Kb - 160 Kb) relative to other sequenced insect genomes. The GC content of the long homogeneous segments in the red flour beetle is 33.1% and does not differ significantly from the mean GC content for the entire genome. In contrast, in *Drosophila* and humans long homogeneous segments have lower GC content than their respective genomes.

GC analysis methods: Genomic sequences were partitioned into segments by the binary recursive segmentation procedure, D_{JS}, proposed by Bernaola-Galván, Róman-Roldán, and Oliver⁶. In this procedure, the sequencing scaffolds were recursively segmented by maximizing the difference in GC content between adjacent subsequences. The process of segmentation was terminated when the difference in GC content between two neighboring segments was no longer statistically significant⁷.

Repetitive DNA and transposable elements.

Among microsatellites, trinucleotide repeats are the most abundant, and dinucleotide repeats, which predominate in other arthropods⁸⁻¹¹ are relatively rare. For example, the longest microsatellite is an AT-rich trinucleotide repeat (ATT₁₉₅) on LG3. The majority (83%) of microsatellites are found in intergenic regions (63%) or introns (20%), but there is a strong overrepresentation of non-frameshift causing repeats (3 and 6 bp motifs), which may represent functional amino acid repeats, in exons (2x3 contingency table $\chi^2=363$, $P<0.001$)¹². 545 tri-nucleotide repeats are found in the coding regions of 504 genes. These genes were analyzed by gene ontology category to identify over and under-represented gene categories (Fig. S5).

Transposons. Several families of DNA transposons, as well as LTR and non-LTR retrotransposons, constituting approximately 6% of the genome¹³ were identified via encoded protein sequence similarity to previously identified elements using TEPIPE¹⁴ or BLAST¹⁵, and are listed in Supplementary Table 5^{13, 16, 17}. DNA transposons of the IS630-*Tc1-mariner* superfamily¹⁸ are the most plentiful (30 sub-families identified). In comparison, only a few *hAT* superfamily transposons were found, including one *hermit*¹⁹, two *Herves*²⁰ elements and a single intact copy of *TcBuster* that appears to be very active in *in vivo* assays¹⁶. Surprisingly, 14 families of *piggyBac*-like elements were identified; most copies of which were found to be defective, having lost one or both inverted terminal repeats (ITR) or suffered internal deletions¹⁷. The remaining ITRs differ in sequence from those of the *T. ni piggyBac* transposon used for transformation²¹, suggesting the *Tc piggyBac* elements are not likely to be mobilized by the foreign element. One *Tribolium piggyBac* element encodes an intact transposase that is more similar to a human *piggyBac* transposase than to the *T. ni* transposase and is therefore unlikely to remobilize copies of the *T. ni* transposon transformed into the *Tribolium* genome¹⁷. One helitron²² and two polintons²³, extremely long (13-17 Kb) self-replicating DNA transposons, have also been identified in the *Tribolium* genome.

Nine different non-LTR and six different LTR retrotransposon clades (Table S5) are represented in the *Tribolium* genome¹³. One currently active member of the Osvaldo family, Woot, was previously discovered in the analysis of a spontaneous *Tcabdominal-A* mutant²⁴. *Tribolium* appears to be replete with mobile elements and is likely to harbor additional elements that may be identified in future analysis of the genome sequence.

Telomeres. Due to the difficulty of assembling regions rich in highly repetitive sequences, none of the 20 telomeres are fully represented in the current assembly. Candidate telomeres were identified by searching the 70,000 fosmid end reads with 1000 bp of TCAGG repeats. All of the top 50 matches, and most of the remainder, are in the “plus/minus” orientation, indicating that long stretches of continuous TCAGG repeats constitute the extreme termini of telomeres. Most of the mate pairs of the top 50 matches represent transposons and other common repeat sequences in the genome, indicating that most telomeres are longer than the 30-40 kb fosmid insert size. However, ten match seven unique terminal sequences on long scaffolds that might be inside telomeres of less than 30-40 kb in length. Manual assembly of the proximal regions of these seven candidate telomeres beyond the ends of the assembled scaffolds reveals TCAGG repeats interrupted by full-length and 5’ truncated non-LTR retrotransposons belonging to the R1 clade, best known for insertions in the rDNA locus²⁵. We named these non-LTR retrotransposons SART-Tcas for their sequence similarity to the *Bombyx* telomere-specific SART1 element. All of the insertions are in the same orientation with their “poly-A” tails distal, and the insertions are almost always between the TCA and GG of a telomeric repeat. Using this information we performed a second search of the fosmid reads using the query sequence:

AAAAAAAAAAAAAAAAAAGGTCAGGTCAGGTCAGGTCAGGTCAGGTCAG,
which represents the junction of a SART-Tcas element and telomeric repeats. In addition to re-identifying some of the seven candidate telomeric scaffolds described above, we found one more long internal scaffold. We also identified two more candidate telomeric scaffolds as reasonably long assemblies that have multiple copies of the above sequence, for a total of 10 candidate telomeric ends. The two longest of these were already mapped to ends of chromosomes 8 and 9 confirming that these sequences are indeed telomeric. We mapped two more of these scaffolds, one from the seven above and one from these last three, and located them at the two extreme ends of chromosome 10. Thus, we have identified 10 of the 20 telomeres. The remaining 10 telomeres differ only in being so long that we cannot confidently identify unique sequences flanking them using these fosmid mate pairs. The ten telomeric scaffolds are listed in Table S6, with the first being on

chromosome 8, the second on chromosome 9, and the third and fourth newly mapped at the ends of chromosome 10.

EST sequencing. At the time of analysis, a total of 61,228 expressed sequence tag (EST) sequences have been generated. The sequences include 32,544 *de novo* clones (~70% were sequenced at both ends) and 10,704 available from the NCBI. A large portion of the EST data (47% of the total) was incorporated into the computerized gene annotations, the additional EST sequences were produced more recently. Five different tissue- or stage-enriched cDNA libraries have been used to generate EST sequences, including adult hindgut and Malpighian tubules (23,236 sequences from 14,654 clones), ovary (1,742 sequences from 1,082 clones), adult head (1,448 sequences from 855 clones), larval carcass (2,270 sequences from 1,818 clones), and mixed-stage, whole larvae (21,828 sequences from 14,135 clones). Almost half of the sequenced clones are from the cDNA libraries for excretory organ (hindgut and Malpighian tubules) and mixed-stage, whole larvae, while more recent EST sequences derive from neural tissue (adult head), fatbody and epidermis (larval carcass). The 61,228 sequences were contiged into 12,351 clots (UniESTs) after assembly of paired reads and redundant sequences. 10,134 UniEST clots mapped onto 6,463 of the 16,422 genes in the predicted Glean gene set (39%), while more than 1,200 UniEST remain as novel transcripts. We conservatively estimate that the current EST set covers more than 7,500 transcription units, including the ESTs that were not presented in the Glean set.

Automated annotation. The automated phase of the annotation involved 2 stages, the running of a number of automated gene prediction and annotation pipelines and programs, and the production of a consensus gene model set using the GLEAN program. The automated gene prediction and annotation pipelines are described first.

AUGUSTUS. The eukaryotic gene prediction program AUGUSTUS^{26, 27} is based on a hidden Markov model that probabilistically models the DNA sequence and its gene structure. As existing cross-trained versions of AUGUSTUS (for example, the *Drosophila melanogaster* or *Aedes aegypti* versions) do not perform optimally on *T. castaneum* due to differences in intron length distributions and base compositions between these species, we trained a specific *T. castaneum* version for optimal

performance. For that purpose the parameters of AUGUSTUS needed to be estimated on a training set of bona fide genes. To compile a training gene set we constructed spliced alignments of all available *T. castaneum* ESTs to the Repeat-Masked genome assembly using BLAT²⁸ and sim4²⁹. Only those ESTs that could be aligned over at least 90% of their length with at least 95% sequence identity were used. Overlapping spliced alignments with compatible splicing were clustered to partial transcripts using PASA³⁰. Those partial transcripts that contained an open reading frame of at least 300bp and an in-frame stop codon upstream of the first ATG in that reading frame were used as training genes. The training set used to estimate the AUGUSTUS parameters comprised 85 complete genes and 510 partial genes incomplete at the 3' end. Meta parameters of the model such as the window size of the splice site models, smoothing parameters and the order of the Markov chain models for coding and non-coding regions were iteratively optimized doing a tenfold cross-validation on the training gene set. The final gene predictions on the complete assembly of *T. castaneum* with AUGUSTUS were performed *ab initio* restricting AUGUSTUS to predict only a single transcript per gene.

Fgenesh³¹ / Fgenesh++³² Gene finding parameters were trained on 1,185 genes of Endopterygota (*Drosophila* genes were removed to avoid bias) from Genbank (33 genes of *Tribolium* or 77 genes of Coleoptera from Genbank were not enough for the training procedure). These parameters were then used to produce preliminary gene predictions on the *Tribolium* genome assembly. Predicted genes with protein sequence similarity to known proteins from the non redundant database at NCBI (NR) were used to retrain gene finding parameters for *Tribolium*. Using Fgenesh, 24,097 genes were predicted (including incomplete gene structures, i.e. those with initial and/or terminal exon(s) absent; genes with score < 6 were filtered out). When using the Fgenesh++ pipeline, we did not map known mRNAs, as there were few mRNAs known for *Tribolium*. Instead, gene prediction was based on the similarity of predicted proteins to known eukaryotic proteins from the NCBI NR database identified using Blast¹⁵. Overall, 23,448 genes were predicted (including incomplete gene structures, i.e. those with initial or/and terminal exon(s) absent). Among these there are 1,877 genes with protein support and 21,571 *ab initio* gene predictions.

NCBI gene predictions. NCBI gene prediction is a combination of homology searching and *ab initio* modeling. cDNAs and ESTs were aligned to the genomic sequences using Splign³³. Proteins were aligned to the genomic sequences using ProSplign³⁴. The best scoring CDS was identified for all cDNA alignments using the same scoring system used by Gnomon³⁵, the NCBI *ab initio* prediction tool. All cDNAs with CDS scores above a certain threshold were marked as coding cDNAs, and all others were marked as UTRs. Some of the CDS were incomplete, meaning they lacked a translation initiation or termination signal. Protein alignments were scored the same way and CDS that did not satisfy the threshold criterion for a valid CDS were removed. After determining the UTR/CDS nature of each alignment, the alignments were assembled using a modification of the Maximal Transcript Alignment algorithm³⁰, taking into account not only exon-intron structure compatibility but also the compatibility of the reading frames. Two coding alignments were connected only if both had open and compatible CDS. UTRs were connected to coding alignments only if the necessary translation initiation or termination signals were present. There were no restrictions on the connection of UTRs other than exon-intron structure compatibility. All assembled models with a complete CDS, including the translation initiation and termination signals, were combined into alternatively spliced isoform groups. Incomplete models were directed to Gnomon³⁵ for extension by *ab initio* prediction. Gnomon³⁵ was also used to predict pure *ab initio* models in regions of the genome that lacked any cDNA, EST or protein alignments.

HGSC-Ensembl annotation pipeline, Genscan and Geneid. The BCM-HGSC has imported a version of the Ensembl genome annotation pipeline³⁶. The pipeline was run using standard conditions³⁷, with ~40,000 *Tribolium* ESTs as input as well as protein sequences from the honeybee, fruitfly, human and mouse genomes. Genscan³⁸ and Geneid³⁹, both pure *ab initio* predictors, were also run at the BCM-HGSC by S. Richards.

A consensus gene set. Results from two automated annotation pipelines (an HGSC import of the Ensembl^{36 37} gene annotation pipeline and NCBI-Gnomon³⁵) and four *ab initio* prediction programs (FgenesH++^{31, 32}, Augustus²⁶, Genscan³⁸ and Geneid³⁹) were combined into a consensus set of 16,404 gene models using GLEAN⁴⁰. GLEAN uses Latent Class Analysis to estimate accuracy and error rates of intron-exon boundaries for

each source of gene evidence. A dynamic programming method is then used to compute the highest probability path of intron-exon boundaries amongst the input data, thus producing a consensus gene model set. The resultant gene models from all of these are described in Table S7. Comparison of this consensus gene set to a gold standard gene set of manually annotated genes not used as an input to the automated programs confirmed the GLEAN consensus sets higher quality than all single gene sets by multiple metrics (Table S8). Note that the automated annotation pipelines, producing only gene models supported by EST or homology evidence, constructed ~9,500 gene models of considerably longer length and higher quality than the *ab initio* gene models. The consensus gene set provides an appropriate balance between the quality of the evidence-based methods and the gene discovery potential of the *ab initio* methods.

Global analysis of genes orthologous between *Tribolium*, other insects and vertebrates revealed 138 shared genes that have been either artificially fused with other genes in *Tribolium* or missed in the consensus gene set. Erroneous gene models were corrected manually, selecting the best homologous gene model for FgenesH++ -assisted gene model calling and quality controlled by multi-species protein alignment.

Global comparison of the gene set to other organisms – Methods:

Orthology. Groups of orthologous genes were automatically identified using a variant of a strategy employed previously⁴¹⁻⁴³, based on all-against-all protein comparisons using the Smith-Waterman algorithm, followed by clustering of reciprocally best matching triangles between each set of three species that overlap by at least 30aa to avoid the domain walking effect. Furthermore, orthologous groups were expanded by including genes that are more similar to each other within a genome than to any gene in any of the other genomes. All orthologous classifications and the corresponding species copy-number distribution are available from

http://cegg.unige.ch/Insecta/Tribolium/Tribolium_analysis.html

Phylogeny. Multiple alignments of protein sequences were produced using Muscle⁴⁴ and the well aligned regions of these alignments were extracted using Gblocks⁴⁵ with default parameters for further phylogenetic analysis using Maximum-Likelihood method as implemented in PHYML⁴⁶ and TREE-PUZZLE⁴⁷ using the JTT⁴⁸ model for amino acid

substitutions with a gamma correction using four discrete classes, an estimated alpha parameter and proportion of invariable sites. The values of statistical support were obtained from 500 replicates of bootstrap analyses. The phylogeny of the five insect and five vertebrate species (Fig. 2) was quantified using 1150 orthologous genes found in exactly one copy, which were aligned separately, and well aligned regions of which were then concatenated into a 336,069 aa long alignment. The species names are abbreviated from Latin names.

Methods: RNAi experiments: RNA interference (RNAi) was used to knock down gene function following previously established methods⁴⁹⁻⁵¹. Injection of dsRNA (100pg/ul up to 2ug/ul) into freshly laid eggs leads to phenotypes in the injected individual⁴⁹. In order to generate large amounts of knock-down embryos parental RNAi was performed, in which female pupae or adults are injected (1 to 6 ug/ul) and consecutive egg-lays are collected, fixed and stained using standard procedures⁵⁰. dsRNA injections into late larval stages (larval RNAi) leads to subsequent gene knockdown e.g. in metamorphosis⁵¹. The strength of the knockdown depends of the amount of injected dsRNA, the time between injection and phenotype assessment, but varies from gene to gene. Genetic null phenotypes are phenocopied in all cases studied so far - the portion of null phenocopies was up to 70% (e.g. *Tc-Krüppel*⁵² and *Tc-knirps*, not published) but is expected to be lower for some genes. Embryonic RNAi usually induces stronger phenotypes. Expression analysis of the RNAi knockdowns was performed using standard methods as described previously⁵³.

Examples of micro-synteny. We did not perform a systematic search for gene synteny, and given the minimum of ~300My evolutionary separation between *Tribolium* and other sequenced insect genomes one would expect little conservation of gene order. In our survey of *Tribolium* genes we did, however, find some examples of conserved gene order. In addition to the HOMC (described in the main text), and Wnt clusters (described in the signal transduction section), we observed conserved gene orders around Runx, GATA factor and NK-homeobox genes.

A cluster of genes for GATA factors in *T. castaneum* and other insects. The five genes for GATA factors known from higher dipterans are also present in the *Tribolium*

genome: *serpent* (*srp*), *GATAe*, *pannier* (*pnr*), *GATAd* and *grain* (*grn*, aka *dGATAc*). Interestingly, the former three genes, *srp*, *GATAe* and *pnr* are organized in a cluster spanning approximately 54 kb which is a highly conserved feature of insect genomes. Such a cluster of about the same size and with the same uniform transcriptional orientation is also seen in dipterans from *Drosophila melanogaster* to *Anopheles gambiae* and in the hymenopteran *Apis mellifera*. The degree of conservation is striking, however, the functional constraints that keep the cluster together remains elusive. For one member of the cluster, *Tcpnr* (formerly known as *TcGATAx*) we have observed an expression pattern homologous to the expression of *Drosophila pnr*, i.e. in the dorsal ectoderm and mesoderm. Functional analysis reveals that *Tcpnr* is in fact essential for the formation of dorsal epidermis in the embryo, illustrating that at least the function of individual members of the cluster is highly conserved in insects.

The NK homeobox gene cluster. Clustering of ANTP-class homeobox genes is not restricted to the Hox gene cluster. The NK cluster was first described in *Drosophila* and the component genes are all involved in mesoderm development⁵⁴. The *Tribolium* NK cluster consists of 2 *Msx/Drop* genes, followed by *tinman*, *bagpipe*, *Lbx* and *C15*. It is not clear from the phylogenetic trees whether the ancestral insect had a single *Msx/Drop* gene or a pair of genes as in *Tribolium* and *Apis*. *Slouch* is also a member of the NK cluster in *Drosophila*. *Tribolium slouch* is on the same chromosome as the NK cluster (linkage group 9), but has been separated away from the cluster. *Tribolium slouch* is neighboured by *Hmx/NK5*, as it is in mosquito⁵⁵, which is consistent with the hypothesis that the NK cluster in the ancestral insect was *Msx/Drop* – *tinman*/NK4 – *bagpipe*/NK3 – *Lbx* – *C15*/Tlx – *slou*/NK1 – *Hmx*/NK5.

Lipid/sterol transport proteins. Lipid/sterol transport proteins are vital for the survival of insects, because storage and mobilization of lipid/sterols is integral to development⁵⁶, growth⁵⁷⁻⁵⁹, and reproduction^{60, 61}. The *Tribolium* genome encodes two independent copies of the fatty acid binding protein and homologs of lipophorin genes ApoL-I, -II, and -III, similar to other insects with sequenced genomes except *Drosophila* which lacks ApoL-III. Two families of intracellular sterol transport proteins are found in insects:

sterol carrier protein-2^{62, 63} and steroidogenic acute regulatory protein domain (START-related) proteins⁶⁴. The *Tribolium* genome encodes two START-related genes: *start1* and *start10*. *Tribolium start1* is similar to its honey bee ortholog (GB11881-PA), which lacks the long amino acid sequence inserted in the START domain seen in the dipteran genes (⁶⁴; EAA03945; AAX85201). There are four sterol carrier protein-2 proteins in *Tribolium* but only three in *Apis*. Interestingly, the increase in SCP-2 domain proteins in *Tribolium* is achieved via an alternative transcription start site of the 17- β -hydroxysteroid dehydrogenase-4 gene generating a transcript containing the SCP-2 domain only. In contrast, expansion of the SCP-2 protein family in dipteran genomes (7-8 members) occurred by gene duplication.

Immune pathway components. *Tribolium* harbors a range of natural pathogens and parasites, from bacteria to fungi, microsporidians and tapeworms⁶⁵⁻⁶⁷. Its genome reveals probable orthologs for nearly all members of the Toll, IMD and JAK/STAT immune pathways. Paralog counts for candidates for these pathways are roughly equivalent to those found for *D. melanogaster* or *A. gambiae*, but are substantially higher than those observed for the honeybee⁶⁸⁻⁷⁰. We have identified ~300 immunity-related genes based on sequence homology. More clip-domain serine proteinases and serpins exist in the *Tribolium* genome than in the other insects sequenced to date. In line with the increase in clip-domain serine proteinases, gene duplication resulted in a cluster of 16 serpin genes within a 50 kb region. Four of the nine Toll-like proteins are grouped in the clade containing *Drosophila Toll*. As in other insect genomes, some immunity-related gene families show a high frequency of lineage-specific expansions at the expense of 1:1 orthology. Real time PCR analyses support the up-regulation of antimicrobial proteins upon bacterial and fungal challenge, as well as up-regulation of signaling molecules in the IMD pathway. Immune responses toward the opportunistic fungal pathogen *Candida albicans* are much greater than those toward *Saccharomyces cerevisiae*, an environmental non-pathogen added to the diet.

Besides identification of candidate immunity-related genes in *Tribolium* based on sequence homology, subtractive hybridization experiments have expanded the spectrum of immune-inducible genes to include a thaumatin-like peptide (representing an ancient

antifungal peptide originally reported from plants) absent from the genomes of *Drosophila*, *Anopheles*, and *Apis*. Additionally, septic injury induces expression of genes involved in stress adaptation (e.g. heat shock proteins and hypoxia-inducible genes), or insecticide resistance (e.g. cytochrome P450s, and ABC transporters), suggesting there is crosstalk between the immune and stress responses in *Tribolium*.

Signal transduction pathways. Signalling pathways regulate numerous developmental processes, with functional diversity reflected in specialized pathway components that often vary among taxa. Among insects examined to date, *Tribolium* contains the largest complement of *Wnt* genes, including orthologs of all *Drosophila Wnts* (*Wnt1*, 5, 6, 7, 9, 10 and *Wnt8/D*) and, in addition, *Tribolium* has orthologs of the vertebrate *Wnt11* gene (Table 1) and *WntA*, an ancestral *Wnt* gene not found in vertebrates. As in *Anopheles* and *Bombyx*, *Wnt9* is linked to the evolutionarily conserved *Wnt* gene cluster comprising *Wnt1*, 6 and 10, confirming that the ancestral insect cluster contained four *Wnt* genes⁷¹.

Four **FGF** genes were found in the *Tribolium* genome, including an ortholog of *Drosophila bnl* and two genomically linked *Tribolium* FGF genes that appear similar to vertebrate *fgf1*. The fourth FGF groups phylogenetically with the *Drosophila* and vertebrate FGF8s and is expressed in the growth zone, segment primordia, limbs, anlagen of fore- and hindgut, and the Malpighian tubules⁷². The single FGF-receptor gene in *Tribolium* is orthologous to *Drosophila htl*.

Most components of the **EGF pathway** are conserved between *Tribolium* and *Drosophila*, with the notable exception of the EGF ligand Vein. In addition, a single TGF- α -like ligand was found in *Tribolium*, compared with three (Gurken, Spitz and Keren) in *Drosophila* (Fig. S7). Of seven **Rhomboid** family proteases identified (RhoA-RhoG), three, RhoA-RhoC), cluster phylogenetically with the four *Drosophila* EGF Rhomboids⁷³, but lack clear orthologous relationships. This suggests that independent duplication events produced the multiple EGF Rhomboids in each species.

Single homologs of all 25 genes involved in *Drosophila Notch* signalling were identified in *Tribolium*, except *groucho* for which we found two (Table S11). However, no component of this pathway was expressed in *Tribolium* in a pattern suggesting involvement in segment formation, while we found that *Notch* signalling was involved in

appendage formation and nervous system specification. Two *Enhancer of split* and three *achaete scute* class bHLH *Notch* target genes were found in *Tribolium* (two clear orthologs⁷⁴, and a more diverged member at another chromosomal location, TC07826), reflecting the basal arrangement also found in *Apis* and *Anopheles*⁷⁵. Finally, all components of the **Jak/Stat pathway** were identified in *Tribolium* except the fast evolving ligand Unpaired, which also has not been detected in *Apis* or *Anopheles*.

Transcription factor families. Most of the prominent TF-families are remarkably similar in size and composition to those of *Drosophila*, for example the **T-box**, **Pax** and **Runx** genes, as well as the **basic HLH** transcription factors (Note, however, that *NeuroD* has been lost in *Drosophila*). *Tribolium* has 103 **homeobox**-containing genes (Table S15) with representatives of all homeobox gene families except for the Pou1 and Tcf1/2-Hnf families (Figs. S8-10) that were present in the last common ancestor of the Bilateria, unlike *Apis* and *Drosophila* where significant homeobox gene losses were observed.

Nuclear receptors. We identified 21 nuclear receptors in *Tribolium*, similar to other insects with a sequenced genome⁴². A comprehensive phylogenetic analysis revealed that most nuclear receptors that act as early ecdysone-response genes during metamorphosis experienced an accelerated sequence divergence in the Diptera and Lepidoptera⁷⁶. It seems likely that the upstream part of the ecdysone cascade has evolved rapidly during the diversification of holometabolous insects, which may affect the design of specific insecticides targeting the ecdysone pathway.

Sex determination. The exact mechanism of sex determination in *Tribolium* is unknown. *Tribolium* and other insects have well-conserved homologues of *Sex-lethal*, the top-most switch in the *Drosophila* pathway, however *Sxl* is not involved in sex determination except in *Drosophila*^{77, 78}. Instead, homologs of *transformer* (*tra*) play a pivotal role in sex determination in the Mediterranean fruit fly *Ceratitis capitata*, the housefly *Musca domestica* and *Apis*^{79, 80, 81}. BLAST searches do not identify *tra* in the *Tribolium* genome but this may well be due to the high sequence evolution of Tra proteins. However, orthologs of the *transformer* targets *doublesex* and *fruitless* are present. Sex-specific splicing of the *Tribolium doublesex* homolog corresponds to the male and female variants

in *Drosophila* suggesting a *transformer* activity in *Tribolium*. In addition, we have identified *Tribolium* homologs of *transformer2*, an essential cofactor for *doublesex* regulation in *Drosophila*⁸².

Cuticle and chitin biosynthesis and metabolism.

The insect cuticle is a strong and lightweight biomaterial consisting of a laminar array of fibrils of the polysaccharide chitin embedded in a protein matrix. Cuticle serves as both a skin and skeleton, and a waterproof coat-of-armor that is sufficiently flexible to accommodate both growth and motion. Several families of structural proteins and enzymes contribute to the formation, stabilization and turnover of the cuticle during the molting cycle. These include the structural cuticle proteins (CPs), as well as enzymes involved in chitin biosynthesis and reutilization such as the chitin synthases, chitinases, chitin deacetylases (CDA), and N-acetylglucosaminidases. The cuticle proteins are cross-linked (sclerotized) by oxidized catecholic intermediates derived from N-acetyldopamine and N- β -alanyldopamine. Enzymes involved in the tanning and pigmentation of the newly-formed cuticle include the laccases (Lac) and phenoloxidases, and others^{83, 84}. Because the cuticle is so critical for insect survival and so intricately regulated, many of these genes/proteins may be viable targets for general or selective biopesticides.

Results of annotations of five families of genes involved in cuticle biosynthesis and turnover are shown (Table S20). The CDA expansion in *Tribolium* is associated with a tandem array of five *CDA* genes on chromosome 5. Among all cuticle-associated genes, those encoding structural cuticle proteins (CPs) are by far the most numerous. Several families of CPs are now recognized, the most numerous being the RR proteins that bear a cuticle-binding domain. *CP* gene numbers vary widely among species, only 28 being identified in *Apis*, ~100 each in *Tribolium* and *Drosophila* (Table S20) and >150 in *Anopheles*. There are two major subfamilies of RR proteins, with the RR-1 form attributed to “soft” and the RR-2 form to “hard” cuticle⁸⁵. In *Tribolium* there are 57 very small genes, each encoding a single copy of the hard-cuticle-associated RR-2 motif. Twenty-five of these are tightly clustered on chromosome 5. Interestingly, the RR-2 form predominates in both *Anopheles* and *Tribolium* (65% and 56% respectively), while

in *Drosophila* RR-2 genes comprise only 35% of the total⁸⁶. All three species have an ortholog with three RR consensus regions. Two far smaller families (CPF, CPFL) are represented by several genes each, with species differences in the gene number of each family (Table S20)⁸⁷.

Neuropeptide processing enzymes. Neuropeptide genes encode precursors of neuropeptides, which in turn are activated by post-translational cleavage and modification. Two major classes of proteases have been implicated in prohormone processing. These are the kex2/subtilisin-like prohormone convertases (PC1/3, PC2), and an endocrine expressed cathepsin L. Defects in PC1/3 or PC2 cause endocrine diseases in mammals⁸⁸⁻⁹⁰.

Tribolium, like vertebrates but unlike the Diptera, has a clear PC1/3 ortholog, other kex2/subtilisin-like prohormone convertases and an endocrine cathepsin L (Fig. S15). *Tribolium* also possesses peptidylglycine- α -hydroxylating monooxygenase (PHM) and peptidyl- α -hydroxyglycine α -amidating lyase (PAL) enzymes required for C-terminal α -amidation of neuropeptides. As in the other insects⁹¹, these enzymes are encoded by monofunctional genes, whereas nematodes, mollusks and chordates possess single peptidylglycine α -amidating monooxygenase (PAM) genes that encode both enzymatic activities. This full complement of processing machinery implies that *Tribolium* is able to generate a significantly more complex repertoire of active neuropeptides than *Drosophila*.

Odorant-Binding and Chemosensory Proteins. The dendrites of insects chemosensory neurons are surrounded by sensillar lymph, containing high concentrations of Odorant-Binding Proteins (OBPs)⁹² and Chemosensory proteins (CSPs)^{93, 94}. These proteins are believed to shuttle hydrophobic odorants from the cuticle pores of the sensilla, to the dendritic olfactory receptors. Forty-seven OBP genes were identified in the *Tribolium* genome. This is considerably more than the honeybee (21 OBPs)⁹⁵, but is similar to *Drosophila* (51 OBPs) and less than *Anopheles* (70 OBPs)^{96, 97}. Interestingly, three members of the classic *Tribolium* OBPs (TcasOBP-6, -7 and -8) possess a unique

additional cysteine residue, three amino acids after the third conserved cysteine. These proteins are similar to the few previously described Coleopteran pheromone-binding proteins (PBPs)⁹⁸, suggesting a role in chemical communication.

The *Tribolium* gustatory receptor family. The gustatory receptor (Gr) gene family in *Tribolium* exhibits considerable expansion relative to the available fly, moth, and honey bee Gr families. Indeed the *Apis mellifera* Gr complement is remarkably small at just 10 intact genes (and approximately 50 highly degraded pseudogenes that form a unique lineage⁹⁹). TcGr gene models were built manually using previously described methods⁹⁹. A phylogenetic tree was built using almost all the TcGrs, the 10 AmGrS, three available HvCrS from *Heliothis virescens*¹⁰⁰ (named chemoreceptors (Crss), a convention maintained here), and representative DmGrS and AgGrS from *Drosophila* and *Anopheles* (Fig. S16). Like the TcOrs, the TcGrS range in divergence from extremely conserved orthologs to divergent singletons to lineage-specific gene subfamily expansions.

Tribolium has single orthologs (TcGr1-3) for the carbon dioxide heterodimer identified in the Diptera, DmGr21a/AgGr22 and DmGr63a/AgGr24^{101, 102}, and a relative that was lost from *Drosophila* (AgGr23/TcGr2), implying that it can sense carbon dioxide concentrations. The function of the AgGr23/TcGr2 protein is not yet known, but like the carbon dioxide heterodimer, it is shared with *Bombyx mori* (HMR unpublished). Remarkably this entire three gene lineage, which is the most conserved amongst all the insect chemoreceptors, is missing from the honey bee *Apis mellifera* genome⁹⁹. The high conservation of the individual proteins indicates that this is an old gene lineage, so we infer it was lost from bees, which nevertheless sense carbon dioxide as an indicator of hive aeration. Bees presumably utilize other receptors for this purpose.

Tribolium has a considerable expansion of the candidate sugar receptor subfamily. This lineage consists of eight genes in each of the fly genomes, although they are not all strictly orthologous lineages. Two genes have been described from the moth *Heliothis virescens*¹⁰⁰, and *Bombyx mori* has five genes in this lineage (Robertson unpublished), while *Apis mellifera* has just two⁹⁹. *Tribolium* has 16 genes in this lineage (TcGr4-19).

Their phylogenetic relationships are not well resolved in Figure S16b, but in more detailed analyses of this candidate sugar receptor subfamily they form a distinct *Tribolium*-specific expansion. They presumably mediate perception of diverse carbohydrates.

There is only one more Gr lineage that is conserved across the available insect genomes, that of DmGr43a, AgGr25, HvCr4, and AmGr3⁹⁹, although it does not have bootstrap support in the tree. Again, *Tribolium* exhibits an expansion of this lineage to 10 genes (TcGr20-28 and 183), but the function of this lineage is unknown so it is not possible to speculate about the importance of this expansion to *Tribolium* sensory biology.

The tree reveals several major expansions of TcGr lineages into large subfamilies, including one of 88 genes (Fig. S16c). While some of these are in related clusters on particular linkage groups, many are in small groups or are singletons spread around the genome. Like the Gr lineage expansions in *Drosophila melanogaster*¹⁰³, these are probably bitter receptors that recognize diverse plant secondary metabolites that function as plant defensive compounds.

A fourth highly divergent and expanded lineage is unusual and contains the two TcGr loci that are alternatively-spliced, compared with three such Gr loci in *D. melanogaster*¹⁰⁴ and four loci in *A. gambiae*¹⁰⁵. TcGr212 encodes a single protein, but immediately downstream of it, TcGr213 has two long exons encoding N-termini hypothesized to be alternatively-spliced to a shared two-exon C-terminus, yielding proteins TcGr213a and b. About 9Mb along linkage group 3 is a massively alternatively-spliced locus, TcGr214, the largest known amongst the insect Grs (AgGr9 encodes 14 different proteins¹⁰⁵). This locus has 30 potential long N-terminal exons, all apparently alternatively-spliced into a shared 3-exon C-terminus. Six of these N-terminal-coding exons are pseudogenic, leaving 24 intact and fairly divergent Grs encoded by this single locus (there are also another three fragmentary N-terminal exonic regions in this complex locus). The shortest of the “alternatively-spliced introns” between two of these N-terminal exons is just 63 base pairs (from the intron donor splice site to the ATG start codon of the next exon).

There are no obvious features of a promoter in this or any of the other “introns”, leading us to propose that this locus and the other insect alternatively-spliced Gr loci has a single promoter at the 5’ end that directs expression of the entire locus and all the alternative protein isoforms in a single set of gustatory sensory neurons. Exactly how the alternative splicing works is unknown, indeed there might be a novel mechanism for association of the various N-terminal exons with the three C-terminal exons. This huge locus is immediately followed by TcGr215 encoding a single protein, which is nevertheless so highly divergent it does not cluster with the TcGr212-214 proteins in the tree.

Finally, many TcGrs are idiosyncratic singleton or doublet lineages scattered around the genome, and phylogenetically amongst the fly Grs with no confident association with any of them. These are all remarkably divergent proteins, and like many of the fly Grs might be involved in detection of diverse bitter compounds or cuticular hydrocarbons involved in sex and species recognition^{106, 107}. Altogether the impression gained, similar to that of for the highly expanded TcOr gene family, is that *Tribolium* has a remarkably expanded gustatory receptor repertoire, presumably reflecting diverse interactions with diverse arrays of attractive and repellent chemicals in the environment.

References for supplementary data (including figures and tables).

1. Stuart, J. J. et al. Useful DNA polymorphisms are identified by snapback, a midrepetitive element in *Tribolium castaneum*. *Genome* 39, 568-78 (1996).
2. Richards, S. et al. Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. *Genome Res* 15, 1-18 (2005).
3. Andersson, B., Wentland, M. A., Ricafrente, J. Y., Liu, W. & Gibbs, R. A. A "double adaptor" method for improved shotgun library construction. *Anal Biochem* 236, 107-13 (1996).
4. Havlak, P. et al. The Atlas genome assembly system. *Genome Res* 14, 721-32 (2004).
5. Lorenzen, M. D. et al. Genetic linkage maps of the red flour beetle, *Tribolium castaneum*, based on bacterial artificial chromosomes and expressed sequence tags. *Genetics* 170, 741-7 (2005).
6. Bernaola-Galvan, P., Roman-Roldan, R. & Oliver, J. L. Compositional segmentation and long-range fractal correlations in DNA sequences. *Physical Review. E. Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics* 53, 5181-5189 (1996).
7. Cohen, N., Dagan, T., Stone, L. & Graur, D. GC composition of the human genome: in search of isochores. *Mol Biol Evol* 22, 1260-72 (2005).
8. Colbourne, J. K., Robison, B., Bogart, K. & Lynch, M. Five hundred and twenty-eight microsatellite markers for ecological genomic investigations using *Daphnia*. *Mol Ecol Notes* 4, 485-490 (2004).
9. Prasad, M. D. et al. Survey and analysis of microsatellites in the silkworm, *Bombyx mori*: frequency, distribution, mutations, marker potential and their conservation in heterologous species. *Genetics* 169, 197-214 (2005).
10. Ross, C. L. et al. Rapid divergence of microsatellite abundance among species of *Drosophila*. *Mol Biol Evol* 20, 1143-57 (2003).
11. Solignac, M. et al. Five hundred and fifty microsatellite markers for the study of the honeybee (*Apis mellifera* L.) genome. *Molecular Ecology Notes* 3, 307 - 311 (2003).
12. Demuth, J. P. et al. Genome-wide survey of *Tribolium castaneum* microsatellites and description of 509 polymorphic markers. (in Press). *Molecular Ecology Notes* (2007).
13. Wang, S., Brown, S. J. & Tu, Z. Transposable elements in the *Tribolium* genome, in prep. . (2007).
14. Biedler, J. et al. Transposable element (TE) display and rapid detection of TE insertion polymorphism in the *Anopheles gambiae* species complex. *Insect Mol Biol* 12, 211-6 (2003).
15. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25, 3389-402 (1997).
16. Arensburger, P. et al. TcBuster1 from *Tribolium castaneum* is a member of the hAT superfamily and is an active transposable element. (2007).

17. Wang, J.-j., Du, Y.-Z., Wang, S., Brown, S. J. & Park, Y. Large diversity of the piggyBac-like element PLE families in the genome of *Tribolium castaneum*. *Insect Biochemistry and Molecular Biology* (2007).
18. Coy, M. R. & Tu, Z. Gambol and Tc1 are two distinct families of DD34E transposons: analysis of the *Anopheles gambiae* genome expands the diversity of the IS630-Tc1-mariner superfamily. *Insect Mol Biol* 14, 537-46 (2005).
19. Coates, C. J. et al. The hermit transposable element of the Australian sheep blowfly, *Lucilia cuprina*, belongs to the hAT family of transposable elements. *Genetica* 97, 23-31 (1996).
20. Arensburger, P. et al. An active transposable element, Herves, from the African malaria mosquito *Anopheles gambiae*. *Genetics* 169, 697-708 (2005).
21. Berghammer, A. J., Klingler, M. & Wimmer, E. A. A universal marker for transgenic insects. *Nature* 402, 370-1 (1999).
22. Pritham, E. J., Putliwala, T. & Feschotte, C. Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. *Gene* (2006).
23. Kapitonov, V. V. & Jurka, J. Self-synthesizing DNA transposons in eukaryotes. *Proc Natl Acad Sci U S A* 103, 4540-5 (2006).
24. Beeman, R. W. et al. Woot, an active gypsy-class retrotransposon in the flour beetle, *Tribolium castaneum*, is associated with a recent mutation. *Genetics* 143, 417-26 (1996).
25. Xiong, Y. & Eickbush, T. H. The site-specific ribosomal DNA insertion element R1Bm belongs to a class of non-long-terminal-repeat retrotransposons. *Mol Cell Biol* 8, 114-23 (1988).
26. Curwen, V. et al. The Ensembl automatic gene annotation system. *Genome Res* 14, 942-50 (2004).
27. Sodergren, E. et al. The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science* 314, 941-52 (2006).
28. Souvorov, A., Tatusova, T. & Lipman, D. Eukariotic Genome Annotation with Gnomon - a Multi-step Combined Gene Prediction Tool. *ISMB 2004*, 125 (2004).
29. Salamov, A. A. & Solovyev, V. V. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res* 10, 516-22 (2000).
30. Solovyev, V., Kosarev, P., Seledsov, I. & Vorobyev, D. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol* 7 Suppl 1, S10 1-12 (2006).
31. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res* 32, W309-12 (2004).
32. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* 268, 78-94 (1997).
33. Guigo, R., Knudsen, S., Drake, N. & Smith, T. Prediction of gene structure. *J Mol Biol* 226, 141-57 (1992).
34. Elsik, C. G. et al. Creating a honey bee consensus gene set. *Genome Biol* 8, R13 (2007).
35. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res* 33, W465-7 (2005).

36. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* 12, 656-64 (2002).
37. Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M. & Miller, W. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res* 8, 967-74 (1998).
38. Haas, B. J. et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* 31, 5654-66 (2003).
39. Kapustin, Y., Souvorov, A. & Tatusova, T. Splign - a Hybrid Approach To Spliced Alignments. RECOMB 2004 - Currents in Computational Molecular Biology, 174 (2004).
40. Kiryutin, B. & Souvorov, A. in ISMB 2005. (2005).
41. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432, 695-716 (2004).
42. Velarde, R. A., Robinson, G. E. & Fahrbach, S. E. Nuclear receptors of the honey bee: annotation and expression in the adult brain. *Insect Mol Biol* 15, 583-95 (2006).
43. Zdobnov, E. M. et al. Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science* 298, 149-59 (2002).
44. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32, 1792-7 (2004).
45. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 17, 540-52 (2000).
46. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52, 696-704 (2003).
47. Schmidt, H. A., Strimmer, K., Vingron, M. & von Haeseler, A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18, 502-4 (2002).
48. Jones, D. T., Taylor, W. R. & Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 8, 275-82 (1992).
49. Brown, S. J., Mahaffey, J. P., Lorenzen, M. D., Denell, R. E. & Mahaffey, J. W. Using RNAi to investigate orthologous homeotic gene function during development of distantly related insects. *Evol Dev* 1, 11-5 (1999).
50. Bucher, G., Scholten, J. & Klingler, M. Parental RNAi in *Tribolium* (Coleoptera). *Curr Biol* 12, R85-6 (2002).
51. Tomoyasu, Y. & Denell, R. E. Larval RNAi in *Tribolium* (Coleoptera) for analyzing adult development. *Dev Genes Evol* 214, 575-8 (2004).
52. Cerny, A. C., Bucher, G., Schroder, R. & Klingler, M. Breakdown of abdominal patterning in the *Tribolium Kruppel* mutant *jaws*. *Development* 132, 5353-63 (2005).
53. Tautz, D. & Pfeifle, C. A non-radioactive in situ hybridization method for the localization of specific RNAs in *Drosophila* embryos reveals translational control of the segmentation gene *hunchback*. *Chromosoma* 98, 81-5 (1989).
54. Jagla, K., Bellard, M. & Frasch, M. A cluster of *Drosophila* homeobox genes involved in mesoderm differentiation programs. *Bioessays* 23, 125-33 (2001).
55. Garcia-Fernandez, J. The genesis and evolution of homeobox gene clusters. *Nat Rev Genet* 6, 881-92 (2005).

56. Panakova, D., Sprong, H., Marois, E., Thiele, C. & Eaton, S. Lipoprotein particles are required for Hedgehog and Wingless signalling. *Nature* 435, 58-65 (2005).
57. Prasad, S. V., Ryan, R. O., Law, J. H. & Wells, M. A. Changes in lipoprotein composition during larval-pupal metamorphosis of an insect, *Manduca sexta*. *J Biol Chem* 261, 558-62 (1986).
58. Smith, A. F., Tsuchida, K., Hanneman, E., Suzuki, T. C. & Wells, M. A. Isolation, characterization, and cDNA sequence of two fatty acid-binding proteins from the midgut of *Manduca sexta* larvae. *J Biol Chem* 267, 380-4 (1992).
59. Ziegler, R., Willingham, L. A., Sanders, S. J., Tamen-Smith, L. & Tsuchida, K. Apolipophorin-III and adipokinetic hormone in lipid metabolism of larval *Manduca sexta*. *Insect Biochem Mol Biol* 25, 101-8 (1995).
60. Jouni, Z. E. et al. Transfer of cholesterol and diacylglycerol from lipophorin to *Bombyx mori* ovarioles in vitro: role of the lipid transfer particle. *Insect Biochem Mol Biol* 33, 145-53 (2003).
61. Blitzer, E. J., Vyazunova, I. & Lan, Q. Functional analysis of AeSCP-2 using gene expression knockdown in the yellow fever mosquito, *Aedes aegypti*. *Insect Mol Biol* 14, 301-7 (2005).
62. Krebs, K. C. & Lan, Q. Isolation and expression of a sterol carrier protein-2 gene from the yellow fever mosquito, *Aedes aegypti*. *Insect Mol Biol* 12, 51-60 (2003).
63. Takeuchi, H. et al. Characterization of a sterol carrier protein 2/3-oxoacyl-CoA thiolase from the cotton leafworm (*Spodoptera littoralis*): a lepidopteran mechanism closer to that in mammals than that in dipterans. *Biochem J* 382, 93-100 (2004).
64. Roth, G. E. et al. The *Drosophila* gene *Start1*: a putative cholesterol transporter and key regulator of ecdysteroid synthesis. *Proc Natl Acad Sci U S A* 101, 1601-6 (2004).
65. Blaser, M. & Schmid-Hempel, P. Determinants of virulence for the parasite *Nosema whitei* in its host *Tribolium castaneum*. *J Invertebr Pathol* 89, 251-7 (2005).
66. Wade, M. J. & Chang, N. W. Increased male fertility in *Tribolium confusum* beetles after infection with the intracellular parasite *Wolbachia*. *Nature* 373, 72-4 (1995).
67. Zhong, D., Pai, A. & Yan, G. Costly resistance to parasitism: evidence from simultaneous quantitative trait loci mapping for resistance and fitness in *Tribolium castaneum*. *Genetics* 169, 2127-35 (2005).
68. Adams, M. D. et al. The genome sequence of *Drosophila melanogaster*. *Science* 287, 2185-95 (2000).
69. Christophides, G. K., Vlachou, D. & Kafatos, F. C. Comparative and functional genomics of the innate immune system in the malaria vector *Anopheles gambiae*. *Immunol Rev* 198, 127-48 (2004).
70. Evans, J. D. et al. Immune pathways and defence mechanisms in honey bees *Apis mellifera*. *Insect Mol Biol* 15, 645-56 (2006).
71. Bolognesi, R. et al. *Tribolium* Wnts: evidence for a larger repertoire in insects with overlapping expression patterns that suggest multiple redundant functions in embryogenesis. - In press. *Development, Genes and Evolution* (2007).

72. Beermann, A. & Schröder, R. Sites of FGF signalling and perception during embryogenesis of the beetle *Tribolium castaneum* - *In press*. Dev. Genes. Evol. (2007).
73. Urban, S., Lee, J. R. & Freeman, M. A family of Rhomboid intramembrane proteases activates all *Drosophila* membrane-tethered EGF ligands. *Embo J* 21, 4277-86 (2002).
74. Wheeler, S. R., Carrico, M. L., Wilson, B. A., Brown, S. J. & Skeath, J. B. The expression and function of the *achaete-scute* genes in *Tribolium castaneum* reveals conservation and variation in neural pattern formation and cell fate specification. *Development* 130, 4373-81 (2003).
75. Schlatter, R. & Maier, D. The Enhancer of split and Achaete-Scute complexes of *Drosophilids* derived from simple ur-complexes preserved in mosquito and honeybee. *BMC Evol Biol* 5, 67 (2005).
76. Annotation of *Tribolium* nuclear receptors reveals an increase in evolutionary rate of a network controlling the ecdysone cascade (in Press). *Insect Biochemistry and Molecular Biology* (2008).
77. Bopp, D., Calhoun, G., Horabin, J. I., Samuels, M. & Schedl, P. Sex-specific control of *Sex-lethal* is a conserved mechanism for sex determination in the genus *Drosophila*. *Development* 122, 971-82 (1996).
78. Traut, W., Niimi, T., Ikeo, K. & Sahara, K. Phylogeny of the sex-determining gene *Sex-lethal* in insects. *Genome* 49, 254-62 (2006).
79. Pane, A., Salvemini, M., Delli Bovi, P., Polito, C. & Saccone, G. The *transformer* gene in *Ceratitis capitata* provides a genetic basis for selecting and remembering the sexual fate. *Development* 129, 3715-25 (2002).
80. Beye, M., Hasselmann, M., Fondrk, M. K., Page, R. E. & Omholt, S. W. The gene *csd* is the primary signal for sexual development in the honeybee and encodes an SR-type protein. *Cell* 114, 419-29 (2003).
81. Boop, D. Unpublished results.
82. Tian, M. & Maniatis, T. Positive control of pre-mRNA splicing in vitro. *Science* 256, 237-40 (1992).
83. Andersen, S. O. in *Comprehensive Molecular Insect Science* (eds. Gilbert, L. I., Iatrou, K. & Gill, S. S.) 145-170. (Elsevier, New York, 2005).
84. Kramer, K. J. & Muthukrishnan, S. in *Comprehensive Molecular Insect Science* (eds. Gilbert, L. I., Iatrou, K. & Gill, S. S.) 111-144 (Elsevier, New York, 2005).
85. Willis, J. H., Iconomidou, V. A., Smith, R. F. & Hamodrakas, S. J. in *Comprehensive Molecular Insect Science* (eds. Gillbert, L. I., Iatrou, K. & Gill, S. S.) 79 - 110 (Elsevier, Oxford, 2005).
86. Karouzou, M. V. et al. *Drosophila* cuticular proteins with the R&R consensus: Annotation and classification with a new tool for discriminating RR-1 and RR-2 sequences - *In press*. *Insect Biochem Mol Biol* 10 (2007).
87. Togawa, T., Dunn, W. A., Emmons, A. C. & Willis, J. H. CPF and CPFL, two related gene families encoding cuticular proteins of *Anopheles gambiae* and other insects - *In Press*. *Insect Biochem Mol Biol* (2007).
88. Jackson, R. S. et al. Obesity and impaired prohormone processing associated with mutations in the human prohormone convertase 1 gene. *Nat Genet* 16, 303-306 (1997).

89. Zhu, X. et al. Disruption of PC1/3 expression in mice causes dwarfism and multiple neuroendocrine peptide processing defects. *Proc Natl Acad Sci U S A* 99, 10293-10298 (2002).
90. Furuta, M. et al. Defective prohormone processing and altered pancreatic islet morphology in mice lacking active SPC2. *Proc Natl Acad Sci U S A* 94, 6646-6651 (1997).
91. Han, M. et al. *Drosophila* uses two distinct neuropeptide amidating enzymes, dPAL1 and dPAL2. *J Neurochem* 90, 129-141 (2004).
92. Pelosi, P., Zhou, J. J., Ban, L. P. & Calvello, M. Soluble proteins in insect chemical communication. *Cell Mol Life Sci* 63, 1658-76 (2006).
93. Angeli, S. et al. Purification, structural characterization, cloning and immunocytochemical localization of chemoreception proteins from *Schistocerca gregaria*. *Eur J Biochem* 262, 745-54 (1999).
94. Tomaselli, S. et al. Solution structure of a chemosensory protein from the desert locust *Schistocerca gregaria*. *Biochemistry* 45, 10606-13 (2006).
95. Foret, S. & Maleszka, R. Function and evolution of a gene family encoding odorant binding-like proteins in a social insect, the honey bee (*Apis mellifera*). *Genome Res* 16, 1404-13 (2006).
96. Biessmann, H., Nguyen, Q. K., Le, D. & Walter, M. F. Microarray-based survey of a subset of putative olfactory genes in the mosquito *Anopheles gambiae*. *Insect Mol Biol* 14, 575-89 (2005).
97. Hekmat-Scafe, D. S., Scafe, C. R., McKinney, A. J. & Tanouye, M. A. Genome-wide analysis of the odorant-binding protein gene family in *Drosophila melanogaster*. *Genome Res* 12, 1357-69 (2002).
98. Nikonov, A. A., Peng, G., Tsurupa, G. & Leal, W. S. Unisex pheromone detectors and pheromone-binding proteins in scarab beetles. *Chem Senses* 27, 495-504 (2002).
99. Robertson, H. M. & Wanner, K. W. The chemoreceptor superfamily in the honey bee, *Apis mellifera*: expansion of the odorant, but not gustatory, receptor family. *Genome Res* 16, 1395-403 (2006).
100. Krieger, J. et al. A divergent gene family encoding candidate olfactory receptors of the moth *Heliothis virescens*. *Eur J Neurosci* 16, 619-28 (2002).
101. Jones, W. D., Cayirlioglu, P., Kadow, I. G. & Vosshall, L. B. Two chemosensory receptors together mediate carbon dioxide detection in *Drosophila*. *Nature* 445, 86-90 (2007).
102. Kwon, J. Y., Dahanukar, A., Weiss, L. A. & Carlson, J. R. The molecular basis of CO₂ reception in *Drosophila*. *Proc Natl Acad Sci U S A* 104, 3574-8 (2007).
103. Marella, S. et al. Imaging taste responses in the fly brain reveals a functional map of taste category and behavior. *Neuron* 49, 285-95 (2006).
104. Robertson, H. M., Warr, C. G. & Carlson, J. R. Molecular evolution of the insect chemoreceptor gene superfamily in *Drosophila melanogaster*. *Proc Natl Acad Sci U S A* 100 Suppl 2, 14537-42 (2003).
105. Hill, C. A. et al. G protein-coupled receptors in *Anopheles gambiae*. *Science* 298, 176-8 (2002).
106. Amrein, H. & Thorne, N. Gustatory perception and behavior in *Drosophila melanogaster*. *Curr Biol* 15, R673-84 (2005).

107. Bray, S. & Amrein, H. A putative *Drosophila* pheromone receptor expressed in male-specific taste neurons is required for efficient courtship. *Neuron* 39, 1019-29 (2003).
108. Wang, S., Brown, S. J. & Tu, Z. Transposable elements in the *Tribolium* genome, in prep. (2007).
109. Robertson, H. M. The mariner transposable element is widespread in insects. *Nature* 362, 241-5 (1993).
110. Tudor, M., Lobočka, M., Goodell, M., Pettitt, J. & O'Hare, K. The pogo transposable element family of *Drosophila melanogaster*. *Mol Gen Genet* 232, 126-34 (1992).
111. Sarkar, A. et al. Molecular evolutionary analysis of the widespread piggyBac transposon family and related "domesticated" sequences. *Mol Genet Genomics* 270, 173-80 (2003).
112. Wang, J.-j., Du, Y.-Z., Wang, S., Brown, S. J. & Park, Y. Large diversity of the piggyBac-like element PLE families in the genome of *Tribolium castaneum* (in Press). *Insect Biochemistry and Molecular Biology* (2008).
113. Fingerman, E. G., Dombrowski, P. G., Francis, C. A. & Sniegowski, P. D. Distribution and sequence analysis of a novel Ty3-like element in natural *Saccharomyces paradoxus* isolates. *Yeast* 20, 761-70 (2003).
114. Wang, S. & Brown, S. J. Analysis of Repetitive DNA Distribution Patterns in the *Tribolium castaneum* Genome, in prep. (2007).
115. Avedisov, S. N., Kuzin, A. B. & Il'in Iu, V. [Molecular analysis of full-sized and shortened copies of *Drosophila* MDGZ retrotransposons]. *Mol Biol (Mosk)* 31, 950-5 (1997).
116. Marlbor, R. L., Parkhurst, S. M. & Corces, V. G. The *Drosophila melanogaster* gypsy transposable element encodes putative gene products homologous to retroviral proteins. *Mol Cell Biol* 6, 1129-34 (1986).
117. Inouye, S., Yuki, S. & Saigo, K. Complete nucleotide sequence and genome organization of a *Drosophila* transposable genetic element, 297. *Eur J Biochem* 154, 417-25 (1986).
118. Labrador, M. & Fontdevila, A. High transposition rates of Osvaldo, a new *Drosophila buzzatii* retrotransposon. *Mol Gen Genet* 245, 661-74 (1994).
119. Michaille, J. J., Mathavan, S., Gaillard, J. & Garel, A. The complete sequence of mag, a new retrotransposon in *Bombyx mori*. *Nucleic Acids Res* 18, 674 (1990).
120. Mount, S. M. & Rubin, G. M. Complete nucleotide sequence of the *Drosophila* transposable element copia: homology between copia and retroviral proteins. *Mol Cell Biol* 5, 1630-8 (1985).
121. Besansky, N. J. Evolution of the T1 retroposon family in the *Anopheles gambiae* complex. *Mol Biol Evol* 7, 229-46 (1990).
122. Lovsin, N., Gubensek, F. & Kordi, D. Evolutionary dynamics in a novel L2 clade of non-LTR retrotransposons in Deuterostomia. *Mol Biol Evol* 18, 2213-24 (2001).
123. Jakubczak, J. L., Xiong, Y. & Eickbush, T. H. Type I (R1) and type II (R2) ribosomal DNA insertions of *Drosophila melanogaster* are retrotransposable elements closely related to those of *Bombyx mori*. *J Mol Biol* 212, 37-52 (1990).

124. Priimagi, A. F., Mizrokhi, L. J. & Ilyin, Y. V. The *Drosophila* mobile element *jockey* belongs to LINEs and contains coding sequences homologous to some retroviral proteins. *Gene* 70, 253-62 (1988).
125. Sassaman, D. M. et al. Many human L1 elements are capable of retrotransposition. *Nat Genet* 16, 37-43 (1997).
126. Warren, A. M., Hughes, M. A. & Crampton, J. M. Zebedee: a novel copia-Ty1 family of transposable elements in the genome of the medically important mosquito *Aedes aegypti*. *Mol Gen Genet* 254, 505-13 (1997).
127. Burke, W. D., Muller, F. & Eickbush, T. H. R4, a non-LTR retrotransposon specific to the large subunit rRNA genes of nematodes. *Nucleic Acids Res* 23, 4628-34 (1995).
128. Abad, P. et al. A long interspersed repetitive element--the I factor of *Drosophila teissieri*--is able to transpose in different *Drosophila* species. *Proc Natl Acad Sci U S A* 86, 8887-91 (1989).
129. Fawcett, D. H., Lister, C. K., Kellett, E. & Finnegan, D. J. Transposable elements controlling I-R hybrid dysgenesis in *D. melanogaster* are similar to mammalian LINEs. *Cell* 47, 1007-15 (1986).
130. Zdobnov, E. M., Campillos, M., Harrington, E. D., Torrents, D. & Bork, P. Protein coding potential of retroviruses and other transposable elements in vertebrate genomes. *Nucleic Acids Res* 33, 946-54 (2005).
131. Beermann, A. & Schröder, R. Functional stability of the *aristaless* gene in appendage tip formation during evolution. *Dev Genes Evol* 214, 303-308 (2004).
132. Beermann, A. et al. The *Short antennae* gene of *Tribolium* is required for limb development and encodes the orthologue of the *Drosophila* Distal-less protein. *Development* 128, 287-297 (2001).
133. Nagy, L. M. & Carroll, S. Conservation of wingless patterning functions in the short-germ embryos of *Tribolium castaneum*. *Nature* 367, 460-3 (1994).
134. Ober, K. A. & Jockusch, E. L. The roles of *wingless* and *decapentaplegic* in axis and appendage development in the red flour beetle, *Tribolium castaneum*. *Dev Biol* 294, 391-405 (2006).
135. Peel, A. D., Telford, M. J. & Akam, M. The evolution of hexapod engrailed-family genes: evidence for conservation and concerted evolution. *Proc Biol Sci* 273, 1733-42 (2006).
136. Park, Y. et al. Analysis of transcriptome data in the red flour beetle, *Tribolium castaneum*. *Insect Biochem Mol Biol* (submitted).
137. Meloun, B., Baudys, M., Pohl, J., Pavlik, M. & Kostka, V. Amino acid sequence of bovine spleen cathepsin B. *J Biol Chem* 263, 9087-93 (1988).
138. Ray, C. & McKerrow, J. H. Gut-specific and developmental expression of a *Caenorhabditis elegans* cysteine protease gene. *Mol Biochem Parasitol* 51, 239-49 (1992).
139. Zhu-Salzman, K., Koiwa, H., Salzman, R. A., Shade, R. E. & Ahn, J. E. Cowpea bruchid *Callosobruchus maculatus* uses a three-component strategy to overcome a plant defensive cysteine protease inhibitor. *Insect Mol Biol* 12, 135-45 (2003).
140. Tryselius, Y. & Hultmark, D. Cysteine proteinase 1 (CP1), a cathepsin L-like enzyme expressed in the *Drosophila melanogaster* haemocyte cell line mbn-2. *Insect Mol Biol* 6, 173-81 (1997).

141. Bown, D. P., Wilkinson, H. S., Jongsma, M. A. & Gatehouse, J. A. Characterisation of cysteine proteinases responsible for digestive proteolysis in guts of larval western corn rootworm (*Diabrotica virgifera*) by expression in the yeast *Pichia pastoris*. *Insect Biochem Mol Biol* 34, 305-20 (2004).
142. Koiwa, H. et al. A plant defensive cystatin (soyacystatin) targets cathepsin L-like digestive cysteine proteinases (DvCALs) in the larval midgut of western corn rootworm (*Diabrotica virgifera virgifera*). *FEBS Lett* 471, 67-70 (2000).
143. McArthur, A. G. et al. The *Giardia* genome project database. *FEMS Microbiol Lett* 189, 271-3 (2000).
144. Skuce, P. J. et al. Molecular cloning and characterization of gut-derived cysteine proteinases associated with a host protective extract from *Haemonchus contortus*. *Parasitology* 119 (Pt 4), 405-12 (1999).
145. Chan, S. J., San Segundo, B., McCormick, M. B. & Steiner, D. F. Nucleotide and predicted amino acid sequences of cloned human and mouse preprocathepsin B cDNAs. *Proc Natl Acad Sci U S A* 83, 7721-5 (1986).
146. Hu, K. J. & Leung, P. C. Shrimp cathepsin L encoded by an intronless gene has predominant expression in hepatopancreas, and occurs in the nucleus of oocyte. *Comp Biochem Physiol B Biochem Mol Biol* 137, 21-33 (2004).
147. Mitchel, R. E., Chaiken, I. M. & Smith, E. L. The complete amino acid sequence of papain. Additions and corrections. *J Biol Chem* 245, 3485-92 (1970).
148. Girard, C. & Jouanin, L. Molecular cloning of cDNAs encoding a range of digestive enzymes from a phytophagous beetle, *Phaedon cochleariae*. *Insect Biochem Mol Biol* 29, 1129-42 (1999).
149. Merckelbach, A., Hasse, S., Dell, R., Eschlbeck, A. & Ruppel, A. cDNA sequences of *Schistosoma japonicum* coding for two cathepsin B-like proteins and Sj32. *Trop Med Parasitol* 45, 193-8 (1994).
150. Klinkert, M. Q., Felleisen, R., Link, G., Ruppel, A. & Beck, E. Primary structures of Sm31/32 diagnostic proteins of *Schistosoma mansoni* and their identification as proteases. *Mol Biochem Parasitol* 33, 113-22 (1989).
151. Butler, R., Michel, A., Kunz, W. & Klinkert, M.-Q. Sequence of *Schistosoma mansoni* cathepsin C and its structural comparison with papain and cathepsins B and L of the parasite. *Protein Pept. Lett.* 2, 313-320 (1995).
152. Cristofolletti, P. T., Ribeiro, A. F. & Terra, W. R. The cathepsin L-like proteinases from the midgut of *Tenebrio molitor* larvae: sequence, properties, immunocytochemical localization and function. *Insect Biochem Mol Biol* 35, 883-901 (2005).
153. Lynch, M. & Milligan, B. G. Analysis of population genetic structure with RAPD markers. *Mol Ecol* 3, 91-9 (1994).
154. Felsenstein, J. Phylogenies and the comparative method. *American Naturalist* 125, 1-15 (1985).
155. Vekemans, X., Beauwens, T., Lemaire, M. & Roldan-Ruiz, I. Data from amplified fragment length polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between degree of homoplasy and fragment size. *Mol Ecol* 11, 139-51 (2002).

156. Savard, J. et al. Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects. *Genome Res* 16, 1334-8 (2006).
157. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* 23, 127-8 (2007).

Supplementary Table List

Table S1. Sequence reads generated and contained in the genome assembly .

Table S2. Scaffold and Contig statistics for the *Tribolium* genome assembly.

Table S3. Quality statistics for the assembled contigs.

Table S4. Assembly gap statistics.

Table S5. Transposable elements in the *Tribolium* genome

Table S6. *T. castaneum* sequence scaffolds containing telomere sequences.

Table S7. *Tribolium* gene model statistics.

Table S8. *Tribolium* gene model overlap with 1,650 gold standard control exons.

Table S9. List of top 50 InterPro families in insects.

Table S10. Genes present in *Tribolium* and Human but not *Drosophila*.

Table S11. Developmental genes table.

Table S12. A core of highly conserved head developmental genes.

Table S13a, b. Surveys of *Tribolium* candidate ventral limb (a) and wing (b) genes.

Table S14. Survey of *Tribolium* Eye gene orthologs.

Table S15. The 103 Homeobox genes of *Tribolium castaneum*

Table S16. Cytochrome P450s in insects by P450 clan.

Table S17. Predicted cysteine proteinases in the *T. castaneum* genome.

Table S18. Identification of sequences used in the Fig. S13. phylogenetic analysis.

Table S19. Comparison of the chemoreceptor superfamilies of various insects

Table S20. Gene families in *Tribolium* and *Drosophila* involved in cuticle metabolism.

Table S1. Sequence reads generated and contained in the genome assembly

Insert size	Raw reads	Passed reads	Assembled reads	Clone
2-3 kb	5,700	5,289	4,165	Plasmid
4-6 kb	2,105,766	1,799,988	1,454,662	Plasmid
36 kb	70,059	50,404	39,223	Fosmid
130 kb	53,181	33,574	28,823	BAC

Table S2. Scaffold and contig statistics for the Tribolium genome assembly

Scaffolds/Contigs	Number	N50(kb)	Total (Mb)
Anchored Scaffolds	173	1,135	137.8
Unanchored Scaffolds	309	153	18.5
All Scaffolds	482	992	156.3
All Contigs	9,708	41	152.1

Table S3. Quality statistics for the assembled contigs*

Quality	# base pairs	% of assembly (151,885,229 total bp)
Low (< Phred 20)	256,024	0.17
Medium (Phred 20-39)	1,074,828	0.71
High (> Phred 39)	150,554,377	99.12

* The average quality score was 87. Low quality regions are generally found in low coverage regions at the ends of contigs.

Table S4. Assembly gap statistics*

Sequence	Captured gaps	Uncaptured gaps	Estimated captured gap size (bp)
All sequence scaffolds	6,493	-	9,132,397
Chromosome linear sequences	5,239	166	6,831,147

* Captured gap sizes are estimated by reference to the average clone size of clones spanning the gap. Uncaptured gaps are those between large scaffolds adjacently placed onto linkage groups using genetic map data. In FASTA files we have used a gap size of 300Kb to designate this fact. A theoretical maximum size of the uncaptured gaps can be calculated by dividing 44Mb (204Mb estimated genome size – 160Mb assembly size) by the 166 gaps to give ~265kb – smaller than the resolution of the genetic map. In reality, much of the 44Mb of uncaptured genome is repetitive sequences near the ends of chromosomes, or in pericentric heterochromatin, and uncaptured gap sizes are likely between 0 and 100Kb.

Table S5: Transposable elements in the *Tribolium* genome

Name	Number of elements	Element ref*	Tc ref [†]
DNA transposons	48		
<i>IS630/Tc1/mariner</i> (ITm)	30		
Tc1	10	18	108
Mariner	8	109	108
Pogo	4	110	108
other ITms	8	18	108
<i>hAT</i>	4		
Herves	2	20	108
Hermit	1	19	108
Buster	1	16	108, 16
<i>piggyBac</i>	14	111	108, 112
Helitron	1	22	22
Polinton	2	23	23
LTR retrotransposon	49		
Ty3	4	113	108, 114
Mdg3	15	115	15
Gypsy	16	116, 117	15
Osvaldo (Tcwoot)	4	118, 24	15, 24
Mag	7	119	15
Copia	3	120	15
Non-LTR retrotransposon	69		
CR1	15	121	15
L2	16	122	15
R1	19	123	15
Jockey	11	124	15
L1	1	125	15
RTE	1	126	15
R4	2	127	15
R2	3	123	15
I	1	128, 129	15

*Element ref = reference sequence

[†] Tc ref = *Tribolium castaneum* sequence

Pale blue = class, orange = superfamily, green = family, yellow = clade

Table S6. *T. castaneum* sequence scaffolds containing telomere sequences

Original scaffold name	Length	Other scaffold name	Accession number (scaffold co-ordinates)	Linkage group
Contig1125_Contig2371	1,394kb	TcaLG8_WGA116_1	gb CM000283.1 (14379095 - 15773733)	8
Contig3139_Contig4836	976kb	TcaLG9_WGA130_1	gb CM000284.1 (14245876-15222296)	9
Contig4667_Contig4672	283kb	TcaLGUn_WGA217_1	gb CH476329.1	10
Contig7743_Contig1961	88kb	TcaLGUn_WGA242_1	gb CH476354.1	10
Contig2034_Contig8422	64kb	TcaLGUn_WGA247_1	gb CH476359.1	-
Contig3439_Contig360	299kb	TcaLGUn_WGA171_1	gb CH476283.1	-
Contig4765_Contig152	252kb	TcaLGUn_WGA229_1	gb CH476341.1	-
Contig4892_Contig8074	281kb	TcaLGUn_WGA170_1	gb CH476282.1	-
Contig4939_Contig7689	494kb	TcaLGUn_WGA161_1	gb CH476273.1	-
Contig4370	2kb	TcaLGUn_WGA826_1	gb AAJJ01006840.1	-
Reptig2323_Reptig218	36kb		gb CH476575.1	-

Table S7. Tribolium gene model statistics

Program	Total mRNAs	Total Exons	Mean exons/mRNA	Total bp	Mean bp/mRNA	Mean bp /exon
Augustus	12,945	57,726	4.5	18,386,878	1420	319
Fgenesh	23,448	88,285	3.8	23,163,707	988	262
Geneid	16,404	52,413	3.2	19,864,281	1211	379
Genscan	14,244	61,314	4.3	21,118,930	1483	344
Glean	16,365	71,357	4.4	23,133,621	1414	324
Ensembl (HGSC)*	23,815	152,181	6.4	37,826,255	1588	249
NCBI abinitio	13,963	69,086	4.9	20,155,905	1444	292
NCBI supported	9,427	53,348	5.7	15,472,947	1641	290

*Overlapping alternate transcripts, vs 1 transcript per gene for the other predictions means that numbers for the Ensembl (HGSC) and the other gene prediction programs are not directly comparable. The Ensembl (HGSC) run produced 9,159 gene models.

Table S8. *Tribolium* gene model overlap with 1,650 gold standard control exons

	Missed bp	Overlapping bp	False positive bp	Any overlap	Correct splices	Within 6bp
Glean	88,630	356,970	36,856	1,281	905	749
RefSeq supported	69,982	273,127	33,343	970	657	564
RefSeq abinitio	79,500	309,887	36,085	1,091	732	620
Augustus	87,534	332,802	33,832	1,121	713	581
Ensembl(HGSC)	45,314	297,614	92,125	1,117	755	607
Fgenesh	85,858	334,021	51,970	1,231	674	483
Geneid	96,666	281,663	44,857	872	506	362
Genscan	90,476	317,585	33,556	1,046	725	608

Statistics describing the overlap between gold standard gene models and various gene model sets. Overlaps were detected using blat, and parsed using custom perl scripts. **Missed bp** is the number of base pairs in the gold standard gene models not represented in the automated gene model set. **Overlapping bp** is the number of basepairs of overlap between the gold standard gene models and the automated gene model set. **False positive bp** is the number of base pairs in the automated gene model set that do not overlap the gold standard gene models in cases where at least part of the gene model does overlap a gold standard gene model. **Any overlap** is the number of gene models which have >1 bp overlap with the gold standard set. **Correct splices** is the number of exactly correct splices found when automated gene models are compared to the gold standard set. **Within 6bp** is the number of automated gene models splice sites found within +/-6bp of a gold standard gene model splice site.

Table S9. List of top 50 InterPro families in insects

The count of genes per genome is shown together with their rank in each of the genomes in brackets. Families expanded over 2-fold in comparison with honeybee and bigger than in fruitfly (two gene families, IPRO005135 and IPRO000595 do not meet these criteria, they are 2-fold expansion compared to Honey Bee, but NOT bigger than in fruitfly) are marked in bold. Several domain families were omitted from the table, e.g. Reverse transcriptase (IPR000477), Integrase (IPR001584), and various Zinc finger protein domains that are frequently found in transposon or viral proteins (see also¹³⁰). Note that olfactory receptors are underrepresented in the automated gene model sets, as this class of genes is problematic for automated methods.

<i>Tribolium castaneum</i>	<i>Apis mellifera</i>	<i>Drosophila melanogaster</i>	<i>Aedes aegypti</i>	<i>Anopheles gambiae</i>	Family
223 (1)	201 (1)	207 (2)	253 (2)	216 (2)	IPR000719: Protein kinase
190 (2)	184 (2)	184 (3)	207 (4)	174 (3)	IPR001680: WD-40 repeat
176 (3)	107 (5)	107 (7)	188 (5)	158 (5)	IPR001611: Leucine-rich repeat
168 (4)	98 (6)	130 (4)	129 (11)	102 (10)	IPR011701: Major facilitator superfamily MFS_1
163 (5)	58 (21)	243 (1)	387 (1)	315 (1)	IPR001254: Peptidase S1 and S6, chymotrypsin/Hap
145 (6)	116 (3)	122 (6)	140 (9)	113 (8)	IPR003593: AAA ATPase
141 (7)	97 (7)	102 (8)	161 (8)	163 (4)	IPR007110: Immunoglobulin-like
139 (8)	97 (8)	82 (13)	117 (12)	91 (13)	IPR002110: Ankyrin
127 (9)	47 (27)	86 (11)	173 (6)	106 (9)	IPR001128: Cytochrome P450
123 (10)	114 (4)	128 (5)	167 (7)	127 (7)	IPR000504: RNA-binding region RNP-1 (RNA recognition motif)
107 (11)	31 (34)	100 (9)	237 (3)	146 (6)	IPR000618: Insect cuticle protein
100 (12)	89 (9)	99 (10)	96 (16)	88 (14)	IPR001356: Homeobox
96 (13)	75 (13)	76 (15)	98 (15)	77 (20)	IPR001440: Tetratricopeptide repeat TPR
96 (14)	69 (16)	84 (12)	110 (13)	77 (21)	IPR002048: Calcium-binding EF-hand
89 (15)	78 (11)	67 (20)	83 (20)	70 (22)	IPR001849: Pleckstrin-like
86 (16)	55 (23)	71 (17)	88 (18)	99 (11)	IPR000276: Rhodopsin-like GPCR superfamily
81 (17)	78 (12)	74 (16)	94 (17)	82 (17)	IPR011545: DEAD/DEAH box helicase, N-terminal
80 (18)	43 (29)	56 (26)	86 (19)	54 (28)	IPR002198: Short-chain dehydrogenase/reductase SDR
77 (19)	71 (15)	68 (19)	104 (14)	88 (15)	IPR000210: BTB/POZ

76 (20)	73 (14)	71 (18)	79 (22)	62 (24)	IPR001452: SH3
74 (21)	66 (19)	63 (21)	81 (21)	65 (23)	IPR001478: PDZ/DHR/GLGF
72 (22)	58 (20)	61 (22)	69 (24)	61 (25)	IPR003961: Fibronectin, type III
71 (23)	67 (18)	56 (27)	63 (26)	55 (27)	IPR000357: HEAT
62 (24)	57 (22)	61 (24)	64 (25)	85 (16)	IPR006210: Type I EGF/EGF-like
60 (25)	51 (25)	55 (28)	73 (23)	61 (26)	IPR001806: Ras GTPase
55 (26)	38 (30)	81 (14)	131 (10)	82 (18)	IPR002557: Chitin binding Peritrophin-A
52 (27)	83 (10)	61 (24)	56 (29)	82 (19)	IPR004117: Olfactory receptor, Drosophila
51 (28)	48 (26)	40 (32)	42 (38)	43 (33)	IPR000008: C2 calcium/lipid-binding region, CaLB
51 (29)	53 (24)	58 (25)	63 (27)	44 (31)	IPR001092: Basic helix-loop-helix dimerisation region bHLH
51 (30)	27 (40)	35 (38)	60 (28)	44 (32)	IPR002018: Carboxylesterase, type B
49 (31)	10 (195)	12 (219)	11 (323)	11 (252)	IPR005135: Endonuclease/exonuclease/phosphatase
47 (32)	20 (67)	30 (43)	54 (30)	45 (30)	IPR001251: Cellular retinaldehyde-binding/triple function, C-terminal
47 (33)	33 (31)	38 (36)	47 (33)	37 (39)	IPR005821: Ion transport protein
46 (34)	33 (32)	43 (31)	44 (36)	43 (34)	IPR002172: Low density lipoprotein-receptor, class A
43 (35)	19 (74)	28 (55)	46 (34)	28 (63)	IPR001509: NAD-dependent epimerase/dehydratase
42 (36)	13 (140)	28 (56)	21 (149)	17 (148)	IPR004272: Odorant binding protein
38 (37)	19 (73)	18 (131)	26 (104)	24 (77)	IPR000595: Cyclic nucleotide-binding
38 (38)	13 (134)	22 (84)	25 (111)	17 (149)	IPR001140: ABC transporter, transmembrane region
38 (39)	28 (37)	27 (58)	35 (43)	38 (35)	IPR001214: Nuclear protein SET
38 (40)	32 (33)	46 (30)	46 (35)	35 (40)	IPR001993: Mitochondrial substrate carrier
37 (41)	13 (141)	54 (29)	50 (31)	38 (36)	IPR004119: Protein of unknown function DUF227
36 (42)	27 (41)	40 (34)	44 (37)	32 (42)	IPR001623: Heat shock protein DnaJ, N-terminal
35 (43)	18 (82)	34 (39)	33 (72)	16 (155)	IPR000301: CD9/CD37/CD63 antigen
35 (44)	31 (35)	34 (40)	41 (40)	33 (41)	IPR001781: LIM, zinc-binding
35 (45)	26 (42)	26 (65)	30 (86)	29 (44)	IPR002219: Protein kinase C, phorbol ester/diacylglycerol binding
34 (46)	31 (36)	32 (41)	36 (42)	32 (43)	IPR000980: SH2 motif
34 (47)	28 (39)	36 (37)	49 (32)	38 (37)	IPR001810: Cyclin-like F-box
34 (48)	8 (243)	39 (35)	29 (92)	38 (38)	IPR004045: Glutathione S-transferase, N-terminal
29 (49)	11 (163)	20 (113)	16 (214)	10 (259)	IPR000953: Chromo
29 (50)	23 (50)	22 (81)	33 (73)	21 (97)	IPR000910: HMG1/2 (high mobility group) box

Table S10. Genes present in *Tribolium* and human, but not diptera. Complete proteomes of 5 insect and 5 vertebrate species were classified into orthologous groups (see supplementary methods). The human orthologs are given together with the best blast hit in the *H. sapiens* (hugo name) and *Tribolium* proteomes (*Tribolium* ortholog). The respective E-values are shown (blast human; blast *Tribolium*). Orthologous groups are sorted according to their uniqueness in *Tribolium*, i.e. depending on the similarity with the next best hit in the *Tribolium* genome (blast *Tribolium*, *paralog*). Human genes are color coded according to biological or biochemical function.

Present in <i>Tribolium</i> & human but not in Diptera				E values blastp to proteomes		
ensembl description	hugo name	human ortholog	<i>Tribolium</i> ortholog	blast human	blast <i>Drosophila</i>	blast <i>Tribolium</i> (paralog)
similar to BETA- GALACTOSIDASE	NM_138342.2	ENSP00000344659	TC00170	0.00E+000	1.00E-081	no hit
Motile sperm domain-containing protein 1	MSPD1	ENSP00000359819	TC02877	1.00E-035	no hit	no hit
PARP-12 (Poly [ADP-ribose] polymerase 12)	PAR12	ENSP00000263549	TC08702	1.00E-023	3.00E-003	no hit
INTER ALPHA TRYPSIN INHIBITOR HEAVY CHAIN	ITIH4	ENSP00000266041	TC14153	5.00E-088	5.90E-001	no hit
Hypothetical protein FLJ32549	NP_689653.3	ENSP00000311486	TC07818	8.00E-062	3.00E-053	no hit
F-box only protein 21	FBX21	ENSP00000328187	TC02740	2.50E-002	no hit	no hit
Frat-2 (GSK-3-binding protein; PROTO ONCOGENE)	FRAT2	ENSP00000360058	TC13599	8.00E-004	1.10E+000	no hit
Basophilic leukemia expressed protein Bles03	CK068	ENSP00000307933	TC09881	1.00E-005	5.70E+000	no hit
no hit	no hit	ENSP00000316016	TC13761	8.90E-002	no hit	no hit
Gemin-6 (Gem-associated protein 6)	GEMI6	ENSP00000281950	TC02690	2.00E-007	no hit	no hit
APCDD1 precursor (Adenomatosis polyposis coli down-regulated 1 protein)	APCD1	ENSP00000347433	TC02518	3.00E-009	no hit	no hit
PRPK (p53-related-protein-kinase binding protein)	Q8IWR7	ENSP00000325398	TC06547	1.00E-021	no hit	no hit
Inositol-tetrakisphosphate 1-kinase	Q13572-2	ENSP00000308468	TC07773	1.00E-066	6.40E+000	no hit

Present in <i>Tribolium</i> & human but not in Diptera				E values blastp to proteomes		
ensembl description	hugo name	human ortholog	<i>Tribolium</i> ortholog	blast human	blast <i>Drosophila</i>	blast <i>Tribolium</i> (paralog)
Meteorin precursor	METR1	ENSP00000219542	TC03221	2.00E-024	no hit	no hit
Dual specificity protein phosphatase 23	DUS23	ENSP00000357089	TC09574	3.00E-007	no hit	no hit
H2A.2 (Histone H2A type 1-E)	H2A1E	ENSP00000259791	TC04100	no hit	no hit	no hit
Kaptein (Actin-associated protein 2E4)	KPTN	ENSP00000337850	TC09674	1.00E-047	5.90E+000	no hit
MAD4 (Max-interacting transcriptional repressor)	MAD4	ENSP00000346191	TC00785	2.00E-004	7.40E-001	no hit
GM2-AP (Ganglioside GM2 activator precursor)	SAP3	ENSP00000349687	TC08068	2.00E-015	no hit	no hit
Calsequestrin-2 precursor	CASQ2	ENSP00000261448	TC16118	4.00E-026	1.00E-018	no hit
Immediate early response gene 5 protein	IER5	ENSP00000294850	TC13386	5.00E-008	3.00E+000	no hit
SID1 transmembrane family member 1 precursor	SIDT1	ENSP00000264852	TC15033	0.00E+000	1.30E+000	no hit
FLJ20571	NM_001033549.1	ENSP00000300965	TC04528	9.00E-017	no hit	no hit
C9orf80 protein	NP_067041.1	ENSP00000363360	TC15263	no hit	no hit	no hit
Galactoside 2-alpha-L-fucosyltransferase 2 (EC 2.4.1.69)	FUT2	ENSP00000349071	TC04858	6.00E-029	no hit	no hit
Caveolin-2	CAV2	ENSP00000222693	TC01628	2.00E-014	6.50E-001	no hit
Acyloxyacyl hydrolase precursor	AOAH	ENSP00000258749	TC04431	0.00E+000	9.10E+000	no hit
STEREOCILIN PRECURSOR		ENSP00000371102	TC13972	5.00E-007	5.50E+000	no hit
ADULT MALE TESTIS CDNA		ENSP00000332875	TC15905	2.00E-011	3.00E-003	no hit
FAM45A	FA45A	ENSP00000354688	TC12182	2.00E-038	6.70E+000	no hit
Q96NH3 Isoform 3	Q96NH3-2	ENSP00000321539	TC08283	7.80E+000	no hit	no hit
DNA-3-methyladenine glycosylase	3MG	ENSP00000219431	TC03671	5.00E-050	no hit	9.80E+000
HORMA domain containing 2	NP_689723.1	ENSP00000336984	TC00115	4.00E-009	3.10E-001	9.70E+000
C17orf53 protein	NP_076937.2	ENSP00000313500	TC06118	3.00E-016	8.80E-001	9.10E+000
Maspardin (Spastic paraplegia 21 autosomal recessive Mast syndrome protein)	SPG21	ENSP00000204566	TC09913	0.00E+000	4.10E+000	8.80E+000
Platelet-activating factor	PAFA	ENSP00000274793	TC06223	5.00E-054	2.60E+000	8.70E+000

Present in <i>Tribolium</i> & human but not in Diptera				E values blastp to proteomes		
ensembl description	hugo name	human ortholog	<i>Tribolium</i> ortholog	blast human	blast <i>Drosophila</i>	blast <i>Tribolium</i> (paralog)
acetylhydrolase precursor (PAF acetylhydrolase)						
C20orf85	CT085	ENSP00000360210	TC03295	1.00E-004	2.60E+000	8.20E+000
FAM51A1	F51A1	ENSP00000369898	TC07179	4.00E-007	1.30E+000	7.90E+000
YIPF3 (YIP1 family member 3; Natural killer cell-specific antigen KLIP1)	YIPF3	ENSP00000259737	TC12103	9.00E-018	1.50E+000	7.30E+000
NP_001014979.1		ENSP00000347050	TC13540	8.00E-008	no hit	7.20E+000
BRE (Brain and reproductive organ-expressed protein; BRCA1/BRCA2-containing complex subunit 45)	Q9NXR7-3	ENSP00000368953	TC12831	8.00E-032	1.60E+000	7.10E+000
UPF0287	CP061	ENSP00000219400	TC04699	4.00E-015	1.70E+000	6.90E+000
transmembrane protein 136 (TMEM136)	NP_777586.1	ENSP00000312672	TC07603	1.00E-004	9.50E+000	6.80E+000
Transmembrane protein 98 (Protein TADA1)	TMM98	ENSP00000261713	TC13919	2.00E-042	8.90E+000	6.40E+000
citrate lyase beta like	NP_996531.1	ENSP00000365538	TC06816	7.00E-047	no hit	6.30E+000
coiled-coil domain containing 108 isoform 1	NP_919278.2	ENSP00000340776	TC11839	5.00E-012	2.00E-003	6.30E+000
Growth-arrest-specific protein 1 precursor (GAS-1)	GAS1	ENSP00000298743	TC09285	4.00E-021	2.20E+000	5.90E+000
Centromere protein S (CENP-S) (Apoptosis-inducing TAF9-like domain- containing protein 1)	Q8N2Z9-2	ENSP00000317110	TC05212	9.00E-009	2.80E-001	5.80E+000
chromosome X open reading frame 59	NP_775966.1	ENSP00000367929	TC13400	2.00E-010	2.40E+000	5.40E+000
Ecto-ADP-ribosyltransferase 5 precursor	NAR5	ENSP00000352992	TC02417	6.00E-004	no hit	4.50E+000
integrin alpha FG-GAP repeat containing 2	NP_060933.2	ENSP00000228799	TC03087	9.00E-068	no hit	4.00E+000
Cob(I)yrinic acid a,c-diamide adenosyltransferase, mitochondrial	MMAB	ENSP00000266839	TC12709	1.00E-033	5.80E+000	3.80E+000

Present in <i>Tribolium</i> & human but not in Diptera				E values blastp to proteomes		
ensembl description	hugo name	human ortholog	<i>Tribolium</i> ortholog	blast human	blast <i>Drosophila</i>	blast <i>Tribolium</i> (paralog)
precursor						
Synaptoporin	SYNPR	ENSP00000295894	TC04580	1.00E-035	no hit	3.60E+000
NP_001025034.1		ENSP00000346931	TC15036	4.00E-015	4.80E+000	3.60E+000
Telomerase reverse transcriptase	TERT	ENSP00000324616	TC10963	1.00E-011	4.70E+000	3.50E+000
Transcription cofactor vestigial-like protein 4 (Vgl-4)	VGLL4	ENSP00000273038	TC10976	4.00E-009	3.00E-008	3.40E+000
nucleoredoxin	NP_071908.2	ENSP00000349978	TC14856	3.00E-057	3.90E-001	2.90E+000
Ribonuclease P protein subunit p38	RPP38	ENSP00000367445	TC06167	1.00E-005	7.10E-001	2.60E+000
Deleted in lung and esophageal cancer protein 1 (DLC-1)	DLEC1	ENSP00000308597	TC12646	2.00E-016	9.20E+000	2.40E+000
family with sequence similarity 79 (FAM79A)	NP_877429.2	ENSP00000367595	TC12549	1.00E-029	6.10E+000	2.40E+000
Protein FAM100A	F100A	ENSP00000283474	TC10338	2.00E-026	3.50E-001	2.30E+000
FLJ16237	NM_001004320.1	ENSP00000341662	TC04241	0.00E+000	2.30E+000	2.30E+000
Tumor protein p53-inducible nuclear protein 1 (p53DINP1)	T5311	ENSP00000344215	TC04030	6.00E-014	2.00E-003	2.20E+000
cytokine receptor-like factor 3	NM_015986.2	ENSP00000318804	TC00209	1.00E-046	8.40E-002	2.20E+000
K0232		ENSP00000303928	TC04527	2.00E-019	3.00E+000	2.00E+000
BMP and activin membrane-bound inhibitor homolog precursor (Putative transmembrane protein NMA)	BAMBI	ENSP00000364683	TC12274	1.00E-021	6.80E-001	1.80E+000
F-box only protein 7	FBX7	ENSP00000266087	TC04309	2.00E-012	2.00E-001	1.70E+000
Brain-specific membrane-anchored protein precursor	BSMAP	ENSP00000262817	TC16273	5.00E-011	no hit	1.40E+000
RP11-506B15.1 protein isoform 1	NP_001012978.1	ENSP00000340375	TC04798	0.00E+000	7.20E+000	1.20E+000
NP_620129.2		ENSP00000355385	TC10768	5.00E-018	4.10E+000	8.50E-001
FLJ11132	NP_060805.2	ENSP00000262236	TC10432	3.60E-001	2.00E-005	8.30E-001
Ly-6 antigen/uPA receptor-like domain-containing protein	NP_808879.2	ENSP00000280115	TC15648	5.00E-015	1.30E-002	6.90E-001
Histidine ammonia-lyase(Histidase)	HUTH	ENSP00000261208	TC10072	0.00E+000	no hit	6.70E-001
NF-kappa-B-repressing factor (NFKB-	NKRF	ENSP00000304803	TC04678	6.00E-015	8.00E-024	6.10E-001

Present in <i>Tribolium</i> & human but not in Diptera				E values blastp to proteomes		
ensembl description	hugo name	human ortholog	<i>Tribolium</i> ortholog	blast human	blast <i>Drosophila</i>	blast <i>Tribolium</i> (paralog)
repressing factor)						
FLJ21062		ENSP00000334655	TC13877	8.00E-008	2.60E-001	5.80E-001
thymidylate kinase family LPS-inducible member	NP_997198.2	ENSP00000256722	TC03748	4.00E-029	1.50E+000	4.50E-001
C4orf13 protein	NP_001025169.1	ENSP00000334594	TC06130	9.00E-082	2.60E+000	3.80E-001
		ENSP00000333752	TC03217	2.00E-007	9.40E+000	3.30E-001
F-box only protein 6 (F-box/G-domain protein 2)	FBX6	ENSP00000365944	TC00371	1.00E-024	4.00E-002	1.80E-001
FLJ14480 (KIAA1706 protein)	NP_085139.2	ENSP00000242108	TC14464	5.00E-067	6.00E-003	1.30E-001
7,8-dihydro-8-oxoguanine triphosphatase (8-oxo-dGTPase)	P36639-4	ENSP00000349148	TC06761	2.00E-028	1.70E+000	8.80E-002
Fibroblast growth factor 3 (FGF-3)	FGF3	ENSP00000334122	TC06602	2.00E-022	2.00E-007	8.20E-002
Proline-rich nuclear receptor coactivator 2	PNRC2	ENSP00000334840	TC04991	5.00E-006	4.00E-004	6.70E-002
Selenoprotein S (VCP-interacting membrane protein)	SELS	ENSP00000254188	TC02967	1.00E-015	3.80E-001	6.60E-002
betaGal beta-1,3-N-acetylglucosaminyltransferase-like 1	NP_001009905.1	ENSP00000319979	TC02589	2.00E-096	3.80E-001	6.40E-002
AMINO ACID PERMEASE 3, FLJ90709	NP_775785.1	ENSP00000316596	TC15635	9.00E-060	9.90E-003	6.30E-002
Tetratricopeptide repeat protein 5 (TPR repeat protein 5)	TTC5	ENSP00000258821	TC07407	0.00E+000	2.00E-003	6.30E-002
Alkylated repair protein alkB homolog 2 (Oxy DC1)	ALKB2	ENSP00000343021	TC12881	5.00E-057	9.80E-003	5.20E-002
ubiquitin-binding protein homolog	NP_061989.2	ENSP00000219638	TC03520	7.00E-082	1.80E-001	5.00E-002
Transmembrane prostate androgen-induced protein (Solid tumor-associated 1 protein)	NP_954640.1	ENSP00000265626	TC03302	8.00E-010	7.00E-005	4.20E-002
KIAA0556	NP_056017.1	ENSP00000261588	TC06497	0.00E+000	8.40E-002	3.20E-002
MORN repeat-containing protein 2	MORN2	ENSP00000344551	TC05783	4.00E-004	5.00E-004	3.20E-002
leucine zipper and CTNNBIP1 domain containing	NP_115744.2	ENSP00000366430	TC13935	1.00E-045	1.00E-003	2.30E-002

Present in <i>Tribolium</i> & human but not in Diptera				E values blastp to proteomes		
ensembl description	hugo name	human ortholog	<i>Tribolium</i> ortholog	blast human	blast <i>Drosophila</i>	blast <i>Tribolium</i> (paralog)
FLJ30976 (weakly similar to ADENYLATE KINASE)	Q5TCS8	ENSP00000357944	TC11699	2.00E-046	2.60E-002	2.30E-002
Ankyrin repeat and MYND domain-containing protein 1	ANKY1	ENSP00000272972	TC02220	5.00E-020	6.20E-002	2.30E-002
calcium binding and coiled-coil domain 2		ENSP00000365863	TC05737	5.00E-014	7.00E-003	9.00E-003
SPFH domain-containing protein 2 precursor	O94905-2	ENSP00000335220	TC06180	4.00E-054	7.00E-003	7.00E-003
KIAA0240	K0240	ENSP00000313933	TC15821	4.00E-015	2.00E-015	6.00E-003
C-C chemokine receptor type 6 (C-C CKR-6; Chemokine receptor-like 3, G-protein coupled receptor 29)	FR1OP	ENSP00000355812	TC00133	3.00E-025	4.00E-003	5.00E-003
RAB3A-interacting protein (Rabin-3)	Q96QF0-2	ENSP00000247833	TC07982	1.00E-065	2.00E-003	3.00E-003
Retinoblastoma-binding protein 8 (RBBP-8; CtBP-interacting protein)	RBBP8	ENSP00000323050	TC11796	4.00E-010	3.00E+000	3.00E-003
CREB-regulated transcription coactivator 1	Q6UUUV9-2	ENSP00000345001	TC09383	5.00E-023	2.00E-007	2.00E-003
KIAA1468	Q9P260	ENSP00000256858	TC03006	0.00E+000	8.00E-004	2.00E-003
MST101 protein	Q96PV3	ENSP00000305151	TC16110	4.00E-028	5.00E-004	2.00E-003
abhydrolase domain containing 8	NP_078803.3	ENSP00000247706	TC04663	2.00E-037	4.00E-003	2.00E-003
coiled-coil domain containing 74A	NP_620125.1	ENSP00000295171	TC15222	3.00E-003	4.00E-003	2.00E-003
JAW1-related protein isoform b	NP_569056.2	ENSP00000307885	TC07394	2.00E-016	3.50E-002	2.00E-003
Acylamino-acid-releasing enzyme	ACPH	ENSP00000296456	TC12101	0.00E+000	1.00E-005	1.00E-003
coiled-coil domain containing 112 isoform 1	NP_001035530.1	ENSP00000368931	TC08924	3.00E-014	1.00E-006	7.00E-004
Thioredoxin-like selenoprotein M precursor (Protein SelM)	SELM	ENSP00000355008	TC12041	2.00E-010	2.00E-003	5.00E-004
PHD finger protein 23	NP_077273.2	ENSP00000322579	TC14182	4.00E-021	4.00E-004	5.00E-004
formin binding protein 4	NM_015308.1	ENSP00000263773	TC05008	6.00E-031	1.00E-004	4.00E-004
Breast cancer type 1 susceptibility protein (RING finger protein 53)	BRCA1	ENSP00000350283	TC14390	3.00E-008	1.80E-002	3.00E-004
Meiotic recombination protein REC8-	REC8L	ENSP00000308699	TC11436	3.00E-008	2.90E-002	2.00E-004

Present in <i>Tribolium</i> & human but not in Diptera				E values blastp to proteomes		
ensembl description	hugo name	human ortholog	<i>Tribolium</i> ortholog	blast human	blast <i>Drosophila</i>	blast <i>Tribolium</i> (paralog)
like 1 (Cohesin Rec8p)						
CDNA FLJ11811	Q9HAC4	ENSP00000206466	TC01428	1.00E-003	8.50E+000	2.00E-004
ELG protein	NP_061023.1	ENSP00000158149	TC11615	2.00E-003	2.00E-007	1.00E-004
Neurexin-1-beta precursor (Neurexin I-beta)	Q08AH0	ENSP00000332400	TC09678	1.00E-041	0.00E+000	1.00E-004
WAP four-disulfide core domain protein 2 precursor (Putative protease inhibitor WAP5)	WFDC2	ENSP00000361761	TC11324	6.00E-016	3.00E-004	1.00E-004
BRCA1/BRCA2-containing complex subunit 3	P46736-2	ENSP00000328641	TC12858	4.00E-047	8.00E-006	7.00E-005
Carbohydrate kinase-like protein	CARKL	ENSP00000225519	TC12506	0.00E+000	3.00E-006	6.00E-005
PHD finger protein 21A (BRAF35-HDAC complex protein BHC80)	Q96BD5-2	ENSP00000323152	TC08103	3.00E-022	1.00E-003	6.00E-005
radical S-adenosyl methionine domain containing 2	NP_542388.2	ENSP00000371471	TC11614	0.00E+000	1.00E-003	3.00E-005
HEAT-like repeat-containing protein isoform 1	NP_478144.1	ENSP00000305924	TC02629	3.00E-054	3.00E-005	2.00E-005
zinc finger CCCH-type containing 10	NP_116175.1	ENSP00000257940	TC13984	6.00E-030	6.00E-006	2.00E-005
CDNA FLJ26619	Q6ZP31	ENSP00000364953	TC10776	6.00E-012	6.00E-010	6.00E-006
Serine/threonine-protein kinase Haspin (Haploid germ cell-specific nuclear protein kinase)	HASP	ENSP00000325290	TC11145	6.00E-068	3.00E-007	5.00E-006
Prothymosin alpha	Q9NYD3	ENSP00000322133	TC04988	1.00E-003	1.00E-003	4.00E-006
Ancient ubiquitous protein 1 precursor	AUP1	ENSP00000258081	TC06447	1.00E-047	8.10E-002	4.00E-006
Ubiquitin-protein ligase CHFR	NM_018223.1	ENSP00000320557	TC15080	1.00E-023	7.00E-006	2.00E-006
Putative adenylate kinase 7	KAD7	ENSP00000267584	TC08636	0.00E+000	4.00E-004	1.00E-006
Leucine-rich repeat-containing protein 51	LRC51	ENSP00000289488	TC06119	3.00E-027	5.00E-005	5.00E-007
DPY30 domain-containing protein 1	DYDC1	ENSP00000361278	TC09616	2.00E-011	4.00E-006	3.00E-007
Alstrom syndrome protein 1	ALMS1	ENSP00000264448	TC12053	3.00E-006	3.00E-012	3.00E-007
Protein TFG (TRK-fused gene)	TFG	ENSP00000240851	TC02949	7.00E-053	2.00E-007	2.00E-007

Present in <i>Tribolium</i> & human but not in Diptera				E values blastp to proteomes		
ensembl description	hugo name	human ortholog	<i>Tribolium</i> ortholog	blast human	blast <i>Drosophila</i>	blast <i>Tribolium</i> (paralog)
protein)						
Proline-rich protein 11	PRR11	ENSP00000262293	TC15562	3.00E-010	1.00E-006	2.00E-007
pleckstrin homology domain containing	NP_060519.1	ENSP00000318075	TC01772	4.00E-013	3.00E-004	1.00E-007
NP_777612.1		ENSP00000295268	TC13065	3.00E-010	3.00E-004	1.00E-007
KIAA1018	NP_055782.2	ENSP00000354497	TC10140	1.00E-048	4.00E-003	9.00E-008
C3 and PZP-like, alpha-2-macroglobulin domain containing 8	Q8NC09	ENSP00000291440	TC00808	0.00E+000	7.00E-009	8.00E-008
ataxin 3 isoform 1	P54252-2	ENSP00000352872	TC03940	1.00E-072	2.00E-005	5.00E-008
Ankyrin repeat and zinc finger domain-containing protein 1	ANKZ1	ENSP00000321617	TC06259	5.00E-085	2.00E-007	3.00E-008
Leucine-rich repeat-containing protein 34	LRC34	ENSP00000326150	TC14596	6.00E-021	2.00E-008	2.00E-008
SH2 domain containing 4B	Q5SQS7	ENSP00000361223	TC06771	8.00E-059	4.00E-008	2.00E-008
Ankyrin repeat domain-containing protein 40	ANR40	ENSP00000285243	TC05611	3.00E-030	1.00E-005	1.00E-008
protein tyrosine phosphatase domain containing 1 protein	NP_818931.1	ENSP00000364509	TC14187	2.00E-069	5.00E-007	1.00E-008
Fatty acid-binding protein, liver (L-FABP)	FABPL	ENSP00000295834	TC01310	2.00E-009	9.00E-010	8.00E-009
proline rich protein	Q5T870	ENSP00000357733	TC09703	6.00E-009	5.00E-009	8.00E-009
Hepatocyte growth factor precursor (Scatter factor; Hepatopoeitin A)	HGF	ENSP00000222390	TC08647	3.00E-026	4.00E-037	8.00E-009
SH2 domain-containing leukocyte protein (SLP-76 tyrosine phosphoprotein)	LCP2	ENSP00000046794	TC00433	8.00E-007	5.00E-005	4.00E-009
Golgi-associated PDZ and coiled-coil motif-containing protein	Q9HD26-2	ENSP00000357485	TC14370	2.00E-092	2.00E-011	1.00E-009
HEAT SHOCK; Caseinolytic peptidase B protein homolog	CLPB	ENSP00000294053	TC07578	0.00E+000	1.00E-008	1.00E-009
Protein C4orf8	CD008	ENSP00000324587	TC01604	5.00E-009	4.00E-011	7.00E-010
Lathosterol oxidase (Lathosterol 5-	SC5D	ENSP00000264027	TC00699	9.00E-014	1.00E-014	7.00E-010

Present in <i>Tribolium</i> & human but not in Diptera				E values blastp to proteomes		
ensembl description	hugo name	human ortholog	<i>Tribolium</i> ortholog	blast human	blast <i>Drosophila</i>	blast <i>Tribolium</i> (paralog)
desaturase)						
Fatty acid-binding protein (FABPI)	FABPI	ENSP00000274024	TC12473	3.00E-010	4.00E-012	7.00E-010
MGC11332	NP_116107.2	ENSP00000258436	TC16221	3.00E-047	1.00E-010	6.00E-010
leucine-rich B7 protein isoform 1	NP_964013.1	ENSP00000007969	TC16127	5.00E-036	8.00E-009	5.00E-010
Cytokine-inducible SH2-containing protein (CIS; Suppressor of cytokine signaling)	CISH	ENSP00000294173	TC05844	7.00E-023	2.00E-004	4.00E-010
coiled-coil domain containing 34	NP_110398.1	ENSP00000330240	TC12845	5.00E-019	2.00E-014	3.00E-010
Growth factor receptor-bound protein 10 (Insulin receptor-binding protein GRB-IR).	GRB10	ENSP00000338543	TC13230	1.00E-055	3.00E-010	2.00E-010
Microtubule-associated proteins 1A/1B light chain 3B precursor	MLP3B	ENSP00000268607	TC15312	7.00E-032	2.00E-017	1.00E-010
KRAB-A domain containing 2	NP_998762.1	ENSP00000328017	TC11789	2.00E-042	1.00E+000	7.00E-011
AN1-type zinc finger protein 1	ZFAN1	ENSP00000220669	TC13251	1.00E-032	2.00E-016	3.00E-011
Lipoma HMGIC fusion partner-like 1 protein	LHPL1	ENSP00000361036	TC01038	3.00E-037	3.00E-008	2.00E-011
Carbonyl reductase [NADPH] (Prostaglandin-E(2) 9-reductase)	DHCA	ENSP00000290349	TC14539	2.00E-061	1.00E-017	8.00E-012
Ubiquitin-like PHD and RING finger domain-containing protein 1	UHRF1	ENSP00000262952	TC01240	0.00E+000	4.00E-008	7.00E-012
Death-associated protein kinase 3 (DAP kinase 3)	DAPK3	ENSP00000301264	TC06299	0.00E+000	3.00E-010	3.00E-012
Nucleoside diphosphate kinase homolog 5 (NDK-H 5; (Testis-specific nm23 homolog)	NDK5	ENSP00000265191	TC01111	5.00E-037	1.00E-011	3.00E-012
Protein-L-isoaspartate O-methyltransferase domain-containing protein 1	PCMD1	ENSP00000353739	TC01944	2.00E-080	3.00E-008	1.00E-012
B-cell lymphoma 3-encoded protein (Bcl-3)	BCL3	ENSP00000164227	TC04188	no hit	no hit	1.00E-012
Novel protein	Q5TGS4	ENSP00000367192	TC15003	3.00E-011	7.00E-013	7.00E-013

Present in <i>Tribolium</i> & human but not in Diptera				E values blastp to proteomes		
ensembl description	hugo name	human ortholog	<i>Tribolium</i> ortholog	blast human	blast <i>Drosophila</i>	blast <i>Tribolium</i> (paralog)
Leucine-rich repeat-containing protein 27		ENSP00000342641	TC11723	1.00E-021	5.00E-009	3.00E-013
Ubiquitin ligase protein DZIP3 (DAZ-interacting protein 3)	DZIP3	ENSP00000355028	TC08244	2.00E-012	8.00E-009	3.00E-013
KPL2 protein isoform 2	NP_079143.2	ENSP00000348314	TC05735	2.00E-073	6.00E-023	1.00E-013
Receptor-interacting serine/threonine-protein kinase 5 (Dusty protein kinase)	RIPK5	ENSP00000356130	TC14189	0.00E+000	9.00E-015	9.00E-014
Zinc finger CCHC domain-containing protein 9	ZCHC9	ENSP00000369549	TC05967	2.00E-035	3.00E-012	8.00E-015
leucine rich repeat containing 43	NP_689972.2	ENSP00000289014	TC14131	4.00E-013	5.00E-013	7.00E-015
Hypoxia-inducible factor 1 alpha inhibitor (FIH-1)	HIF1N	ENSP00000299163	TC02888	0.00E+000	2.00E-016	4.00E-015
Follistatin-related protein 5 precursor (Follistatin-like 5)	Q4W5K3	ENSP00000368462	TC10347	0.00E+000	3.00E-014	3.00E-015
Vascular endothelial growth factor C precursor (VEGF-C)	VEGFC	ENSP00000280193	TC08148	6.00E-006	1.00E-017	3.00E-015
BTG3 protein (Tob5 protein)	Q14201-2	ENSP00000344609	TC01568	1.00E-022	2.00E-016	1.00E-015
Pleckstrin homology domain-containing family A member 3 (Phosphoinositol 4-phosphate adaptor protein 1) (FAPP-1)	PKHA3	ENSP00000234453	TC01680	2.00E-055	1.00E-017	1.00E-015
RUN and FYVE domain-containing protein 2 (Rab4-interacting protein related)	Q5TC48	ENSP00000265865	TC00322	3.00E-049	1.00E-015	8.00E-016
Keratin-associated protein 4-15	KR415	ENSP00000328270	TC15145	3.00E-022	3.00E-010	7.00E-016
Tetraspanin-8 (Tspan-8) (Tumor-associated antigen CO-029)	TSN8	ENSP00000247829	TC12641	9.00E-024	8.00E-031	1.00E-016
hydrocephalus inducing	Q8N3H8	ENSP00000288168	TC14490	0.00E+000	2.00E-013	1.00E-016
Sperm-associated antigen 6 (PF16 protein homolog; Sperm flagellar protein)	O75602-3	ENSP00000365788	TC14890	0.00E+000	1.00E-014	3.00E-017

Present in <i>Tribolium</i> & human but not in Diptera				E values blastp to proteomes		
ensembl description	hugo name	human ortholog	<i>Tribolium</i> ortholog	blast human	blast <i>Drosophila</i>	blast <i>Tribolium</i> (paralog)
collagen and calcium binding EGF domains 1	NP_597716.1	ENSP00000331473	TC04465	8.00E-058	4.00E-010	1.00E-017
Protein SMG7 (SMG-7 homolog)	SMG7	ENSP00000340766	TC00162	8.00E-096	1.00E-008	1.00E-017
zinc finger protein 532	NP_060651.2	ENSP00000262716	TC00404	3.00E-039	1.00E-012	9.00E-018
Serine/threonine/tyrosine-interacting protein	STYX	ENSP00000346599	TC07281	1.00E-052	1.00E-016	8.00E-018
DNA mismatch repair protein Mlh3 (MutL protein homolog 3)	Q9UHC1-2	ENSP00000238662	TC09474	2.00E-056	2.00E-024	9.00E-019
Jouberin (Abelson helper integration site 1 protein homolog)	Q8N157-3	ENSP00000265602	TC02135	0.00E+000	9.00E-020	7.00E-019
NP_775836.2		ENSP00000297186	TC13368	2.00E-052	2.00E-017	5.00E-020
ETS homologous factor (hEHF)	EHF	ENSP00000257831	TC07077	2.00E-026	2.00E-019	4.00E-020
polycomb group ring finger 1	NP_116062.2	ENSP00000233630	TC04601	2.00E-026	3.00E-039	3.00E-020
G-protein coupled receptor 120 (G-protein coupled receptor PGR4)	GP120	ENSP00000360538	TC02068	4.00E-026	6.00E-022	2.00E-020
no hit	no hit	ENSP00000301953	TC06117	4.00E-035	6.00E-018	2.00E-020
Sorting nexin family member 30	Q5VWJ9	ENSP00000363349	TC12068	1.00E-065	4.00E-012	3.00E-021
Sorting nexin-4	SNX4	ENSP00000251775	TC00603	0.00E+000	2.00E-012	3.00E-021
enoyl Coenzyme A hydratase domain containing 1	Q9NZ30	ENSP00000357289	TC01618	1.00E-040	6.00E-023	3.00E-021
RING finger and WD repeat domain protein 2 (Ubiquitin- protein ligase COP1; Constitutive photomorphogenesis protein 1 homolog)	RFWD2	ENSP00000356641	TC00377	0.00E+000	4.00E-023	7.00E-022
sodium channel associated protein 1	NP_653244.1	ENSP00000281142	TC08020	3.00E-029	5.00E-019	6.00E-022
WD repeat, SAM and U-box domain containing 1	Q8N6N8	ENSP00000350866	TC08907	6.00E-045	1.00E-023	3.00E-022
TBC1 domain family member 12	TBC12	ENSP00000225235	TC03153	0.00E+000	2.00E-022	3.00E-022
NAD-dependent deacetylase sirtuin-5 (SIR2-like protein 5)	Q9NXA8-2	ENSP00000368564	TC05187	2.00E-076	5.00E-019	3.00E-022
protogenin (NEURONAL CELL)	Q8N7D8	ENSP00000299577	TC04237	4.00E-029	6.00E-018	1.00E-022

Present in <i>Tribolium</i> & human but not in Diptera				E values blastp to proteomes		
ensembl description	hugo name	human ortholog	<i>Tribolium</i> ortholog	blast human	blast <i>Drosophila</i>	blast <i>Tribolium</i> (paralog)
ADHESION MOLECULE PRECURSOR)						
solute carrier family 25, member 38 (PROBABLE MITOCHONDRIAL CARRIER)	NP_060345.1	ENSP00000273158	TC05358	2.00E-087	5.00E-025	9.00E-025
LRP16	NP_001028258.1	ENSP00000217246	TC12114	4.00E-061	7.00E-022	3.00E-025
Laminin alpha-4 chain	LAMA4	ENSP00000230538	TC03461	5.00E-087	0.00E+000	2.00E-025
Ras-related protein Rab-33B	RB33B	ENSP00000306496	TC08069	6.00E-043	2.00E-024	4.00E-026
Synaptotagmin-15 (Synaptotagmin XV)	Q9BQS2-4	ENSP00000363450	TC05622	1.00E-038	9.00E-019	3.00E-026
Tubulin delta chain (Delta tubulin)	TBD	ENSP00000320797	TC03644	5.00E-046	8.00E-026	2.00E-026
guanosine monophosphate reductase 2 (GMPR2)	NM_001002000.1	ENSP00000334409	TC09423	0.00E+000	4.00E-024	1.00E-026
two pore segment channel 1	NP_060371.2	ENSP00000335300	TC15674	0.00E+000	5.00E-026	7.00E-028
poly(A)-specific ribonuclease (PARN)-like domain containing 1	NP_775787.1	ENSP00000275275	TC11836	2.00E-047	no hit	4.00E-029
Ankyrin repeat domain-containing protein 16	ANR16	ENSP00000352361	TC06097	5.00E-048	2.00E-031	1.00E-029
Ubiquitin-conjugating enzyme E2 T	UBE2T	ENSP00000356243	TC08968	5.00E-043	4.00E-031	2.00E-031
Ubiquitin-conjugating enzyme E2 J1 (Non-canonical ubiquitin-conjugating enzyme 1)	UB2J1	ENSP00000354684	TC02588	1.00E-071	1.00E-030	1.00E-031
Bone morphogenetic protein 10 precursor (BMP-10)	BMP10	ENSP00000295379	TC06506	6.00E-043	7.00E-036	2.00E-032
pleiomorphic adenoma gene 1 (Zinc finger)	NP_002646.1	ENSP00000325546	TC08868	4.00E-067	3.00E-033	2.00E-034
DnaJ (Hsp40) homolog, subfamily C, member 10	NP_061854.1	ENSP00000264065	TC00309	0.00E+000	8.00E-041	1.00E-034
Transmembrane BAX inhibitor motif-containing protein 4 (Z-protein)	Q9HC19	ENSP00000286424	TC13429	1.00E-054	5.00E-035	5.00E-035
O60290		ENSP00000223210	TC01388	1.00E-005	no hit	2.00E-035
Krüppel-like factor 13 (Transcription	KLF13	ENSP00000302456	TC00837	2.00E-049	8.00E-036	9.00E-036

Present in <i>Tribolium</i> & human but not in Diptera				E values blastp to proteomes		
ensembl description	hugo name	human ortholog	<i>Tribolium</i> ortholog	blast human	blast <i>Drosophila</i>	blast <i>Tribolium</i> (paralog)
factor BTEB3)						
Serine/threonine-protein kinase 33		ENSP00000351743	TC07794	2.00E-040	4.00E-034	9.00E-036
Serine/threonine-protein kinase LMTK3	LMTK3	ENSP00000270238	TC02138	3.00E-056	5.00E-033	6.00E-036
PR domain zinc finger protein 10	PRD10	ENSP00000363948	TC12726	0.00E+000	3.00E-040	5.00E-037
Parathyroid hormone/parathyroid hormone-related peptide receptor precursor (PTH/PTHr receptor)	PTHR1	ENSP00000321999	TC08110	4.00E-055	8.00E-039	4.00E-037
Dentin matrix acidic phosphoprotein 1 precursor (DMP-1)	DMP1	ENSP00000340935	TC11330	1.00E-047	5.00E-057	3.00E-041
GLI pathogenesis-related 1 like 1 (CYSTEINE RICH SECRETORY LCCL DOMAIN CONTAINING PRECURSOR)	Q6UWM5	ENSP00000367967	TC00595	8.00E-017	1.00E-029	1.00E-041
Orexigenic neuropeptide QRFP receptor (G-protein coupled receptor 103)	QRFP	ENSP00000335610	TC14211	2.00E-040	3.00E-039	2.00E-042
Kunitz-type protease inhibitor 2 precursor (Hepatocyte growth factor activator inhibitor type 2)	SPIT2	ENSP00000301244	TC08976	2.00E-026	2.00E-036	2.00E-043
Zinc finger protein 143 (SPH-binding factor)	ZN143	ENSP00000299606	TC07234	2.00E-084	1.00E-037	2.00E-044
Transcription factor Sp5	SP5	ENSP00000364430	TC11696	9.00E-052	4.00E-047	2.00E-047
Protein p25-beta	P25B	ENSP00000317595	TC02134	1.00E-029	3.00E-052	9.00E-049
Tubulin epsilon chain (Epsilon tubulin)	TBE	ENSP00000357651	TC04947	2.00E-082	1.00E-049	5.00E-050
Tubulin-tyrosine ligase-like protein 2 (Testis-specific protein NYD- TSPG)	TTLL2	ENSP00000239587	TC14642	2.00E-089	4.00E-048	1.00E-052
Propionyl-CoA carboxylase beta chain, mitochondrial precursor	PCCB	ENSP00000251654	TC13669	0.00E+000	5.00E-053	9.00E-055
Ras-related protein M-Ras (Ras-related protein R-Ras3)	RASM	ENSP00000289104	TC11829	7.00E-075	3.00E-053	2.00E-055
Ankyrin repeat and IBR domain-	AKIB1	ENSP00000265742	TC09974	0.00E+000	1.00E-058	2.00E-056

Present in <i>Tribolium</i> & human but not in Diptera				E values blastp to proteomes		
ensembl description	hugo name	human ortholog	<i>Tribolium</i> ortholog	blast human	blast <i>Drosophila</i>	blast <i>Tribolium</i> (paralog)
containing protein 1						
Mastermind-like protein 2 (Mam-2)	MAML2	ENSP00000327563	TC12051	1.00E-019	2.00E-016	8.00E-057
Sulfotransferase 1A2 (Aryl sulfotransferase 2)	ST1A2	ENSP00000338742	TC05277	2.00E-045	6.00E-060	2.00E-057
Protein Wnt-8a precursor (Wnt-8d)	WNT8A	ENSP00000354726	TC02386	2.00E-070	3.00E-052	8.00E-058
Meiotic recombination protein DMC1/LIM15 homolog (DNA REPAIR RAD51 HOMOLOG)	DMC1	ENSP00000216024	TC03146	4.00E-099	1.00E-057	5.00E-062
Protein Wnt-11 precursor	WNT11	ENSP00000325526	TC14270	3.00E-073	2.00E-056	6.00E-063
zinc finger, BED-type containing 5	NP_067034.2	ENSP00000250524	TC06077	0.00E+000	no hit	2.00E-064
Oxysterol-binding protein-related protein 6 (OSBP-related protein 6)	Q9BZF3-2	ENSP00000352713	TC09085	0.00E+000	7.00E-076	3.00E-069
Maternal embryonic leucine zipper kinase (hMELK) (Protein kinase PK38)	MELK	ENSP00000298048	TC03567	0.00E+000	2.00E-068	2.00E-069
Serpin B7 (Megsin)	SPB7	ENSP00000337212	TC05751	1.00E-053	1.00E-056	1.00E-069
Carboxypeptidase E precursor (Enkephalin convertase)	CBPE	ENSP00000352733	TC05137	0.00E+000	4.00E-068	4.00E-070
Platelet glycoprotein 4 (SCAVENGER RECEPTOR CLASS B MEMBER)	CD36	ENSP00000308165	TC10353	6.00E-049	5.00E-062	6.00E-071
C9orf90	CI090	ENSP00000362170	TC10286	5.00E-008	1.80E-002	7.00E-072
Myb protein P42POP	NP_001012661.1	ENSP00000325402	TC11219	4.00E-012	2.00E-008	7.00E-072
FLJ90238 (weakly similar to EXCISION REPAIR PROTEIN ERCC-6)	NP_060139.2	ENSP00000334675	TC09972	0.00E+000	2.00E-074	5.00E-074
K1718 (JMJC DOMAIN CONTAINING HISTONE DEMETHYLATION)	K1718	ENSP00000006967	TC10820	0.00E+000	4.00E-077	3.00E-076
Lysosomal alpha-glucosidase precursor (Acid maltase)	NM_000152.2	ENSP00000305692	TC02741	0.00E+000	1.00E-060	3.00E-077
WD repeat domain phosphoinositide-interacting protein 4 (WIPI-4)	Q9Y484-3	ENSP00000348848	TC01220	0.00E+000	7.00E-075	2.00E-079

Present in <i>Tribolium</i> & human but not in Diptera				E values blastp to proteomes		
ensembl description	hugo name	human ortholog	<i>Tribolium</i> ortholog	blast human	blast <i>Drosophila</i>	blast <i>Tribolium</i> (paralog)
Cyclin-dependent kinase-like 2 (Serine/threonine- protein kinase KKIAMRE)	CDKL2	ENSP00000306340	TC05369	4.00E-092	7.00E-083	3.00E-083
CN155		ENSP00000344579	TC06721	2.00E-046	2.00E-081	1.00E-085
ATPase family AAA domain-containing protein 2	ATAD2	ENSP00000287394	TC04983	0.00E+000	2.00E-051	7.00E-088
ADAMTS-2 precursor (A disintegrin and metalloproteinase with thrombospondin motifs 2)	ATS2	ENSP00000251582	TC04822	0.00E+000	8.00E-083	5.00E-093
Macrophage migration inhibitory factor (MIF)	MIF	ENSP00000215754	TC15450	7.00E-015	3.00E-034	4.00E-094
DNA polymerase beta	DPOLB	ENSP00000265421	TC15815	0.00E+000	9.00E-016	4.00E-098
Protein FAM44A	NP_683692.2	ENSP00000040738	TC09815	1.00E-037	1.00E-082	7.00E-099
tryptophan/serine protease	NP_940866.2	ENSP00000333003	TC01300	5.00E-042	1.00E-092	0.00E+000
lysocardiolipin acyltransferase isoform 1	Q8N1Q7	ENSP00000368826	TC15335	1.00E-078	0.00E+000	0.00E+000
dehydrogenase/reductase (SDR family) member 13	NP_653284.1	ENSP00000368173	TC12772	4.00E-058	4.00E-059	0.00E+000
Solute carrier family 2 (Glucose transporter type 5, small intestine; Fructose transporter)	GTR5	ENSP00000366641	TC13486	0.00E+000	0.00E+000	0.00E+000
C1orf112	NP_060656.2	ENSP00000356746	TC00013	2.90E-002	8.00E-015	0.00E+000
sperm-specific sodium proton exchanger	NP_898884.1	ENSP00000306627	TC09070	1.00E-065	7.00E-006	0.00E+000
transmembrane protein 132B	NP_443139.2	ENSP00000266765	TC01411	2.00E-052	0.00E+000	0.00E+000
DnaJ homolog subfamily A member 2 (Cell cycle progression restoration gene 3 protein)	DNJA2	ENSP00000314030	TC13913	0.00E+000	0.00E+000	0.00E+000
Neuroendocrine convertase 1 precursor (Prohormone convertase 1)	NEC1	ENSP00000369295	TC04402	0.00E+000	0.00E+000	0.00E+000
Ankyrin repeat domain-containing protein 28	ANR28	ENSP00000373287	TC15680	0.00E+000	6.00E-076	0.00E+000

Present in <i>Tribolium</i> & human but not in Diptera				E values blastp to proteomes		
ensembl description	hugo name	human ortholog	<i>Tribolium</i> ortholog	blast human	blast <i>Drosophila</i>	blast <i>Tribolium</i> (paralog)
guanine nucleotide exchange factor p532 (HECT DOMAIN AND RCC1 DOMAIN CONTAINING 2)	NP_003913.2	ENSP00000261887	TC00971	0.00E+000	0.00E+000	0.00E+000
Putative ATP-dependent RNA helicase DHX30 (DEAH box protein 30)	DHX30	ENSP00000343442	TC09437	0.00E+000	1.00E-097	0.00E+000
Bestrophin-3 (Vitelliform macular dystrophy 2-like protein 3)	BEST3	ENSP00000332413	TC07875	0.00E+000	0.00E+000	0.00E+000
Propionyl-CoA carboxylase alpha chain, mitochondrial precursor (PYRUVATE CARBOXYLASE)	Q5VXU2	ENSP00000365463	TC13459	0.00E+000	0.00E+000	0.00E+000
THAP domain containing 9	NP_078948.3	ENSP00000305533	TC02001	4.00E-031	9.50E+000	0.00E+000
polycystin 1-like 2 isoform a	NP_001070248.1	ENSP00000299598	TC05805	5.00E-033	8.00E-025	0.00E+000
Cytochrome P450 4F2 (Leukotriene-B(4) omega- hydroxylase)	CP4F2	ENSP00000221700	TC12662	1.00E-090	5.00E-097	0.00E+000
alpha 1 type XIII collagen isoform 1	NP_005194.3	ENSP00000348695	TC11335	0.00E+000	0.00E+000	0.00E+000
Metabotropic glutamate receptor 5 precursor (mGluR5)	MGR5	ENSP00000306138	TC01106	0.00E+000	0.00E+000	0.00E+000
Thymic stromal cotransporter homolog	TSCOT	ENSP00000363345	TC08783	4.00E-016	0.00E+000	0.00E+000
unknown, ambiguous						
chromatin, DNA						
sperm / testis						
transcription / signalling						
kinases / phosphatases						
nervous system						

Table S11. Selected developmental genes. *Tribolium*, *Apis* and *D. melanogaster* gene identifiers for a selection of developmental genes.

Gene Name	<i>D. mel</i>	<i>A. mel</i>	<i>T. cas</i>
<i>Ab</i>	CG4807	GB15811	TC13099
<i>abd-A</i>	CG10325	GB19738	TC00894
<i>Abd-B</i>	CG11648	GB10341	TC00889
<i>activin-beta</i>	CG11062	GB13169	TC15808
<i>Actn</i>	CG4376	GB11028	TC07894
<i>Ago</i>	CG15010	GB17249	TC06451
<i>Al</i>	CG3935	GB11341	TC13331
<i>Alk</i>	CG8250	GB14602	TC02114
<i>AlkB</i>	CG33250	GB11535	TC07602
<i>Alp23B</i>	CG16987	GB11204	TC04297
<i>alpha-Adaptin</i>	CG4260	GB16637	TC10798
<i>alpha-Cat</i>	CG17947	GB12545	TC04609
<i>alpha-Spec</i>	CG1977	GB18557	TC00749
<i>alphaTub84B</i>	CG1913	GB10514	TC04873
<i>Aly</i>	CG1101	GB18402	TC11974
<i>Amos</i>	CG10393	GB15725	TC03170
<i>Antp</i>	CG1028	GB18813	TC00912
<i>Aop</i>	CG3166	GB17935	TC07831
<i>AP-2</i>	CG7807	GB17109	TC03974
<i>Apc</i>	CG1451	GB11953	TC01543
<i>aPKC</i>	CG10261	GB19525	TC05980
<i>Aret</i>	CG31762	GB18240	TC12080
<i>Argos</i>	CG4531	GB13926	TC13607
<i>Arm</i>	CG11579	GB12463	miniclust: TC12388, TC12389
<i>Armi</i>	CG11513	GB15508	TC10546
<i>Arr</i>	CG5912	GB11226	TC08151
<i>Arr1</i>	CG5711	GB16006	TC13804
<i>Arr2</i>	CG5962	GB12766	TC09551
<i>Ase</i>	CG3258	GB18627	TC08437
<i>Asx</i>	CG8787	GB11002	TC13912
<i>Ato</i>	CG7508	GB13095	TC11336
<i>Aub</i>	CG6137	GB10293	TC08711
<i>Aur</i>	CG3068	GB14418	TC01817
<i>Awd</i>	CG2210	GB17251	TC02492
<i>Axn</i>	CG7926	GB11539	TC06314
<i>bab1</i>	CG9097	GB13762	TC03627
<i>bab2</i>	CG9102	GB15064	TC03621
<i>Bap</i>	CG7902	GB13498	TC12743
<i>Baz</i>	CG5055	GB10346	TC12086
<i>beat-IIIc</i>	CG15138	GB17449	TC07050

Gene Name	<i>D. mel</i>	<i>A. mel</i>	<i>T. cas</i>
<i>Bel</i>	CG9748	GB19873	TC13328
<i>Ben</i>	CG18319	GB19498	TC01755
<i>beta-Spec</i>	CG5870	GB11407	TC07173
<i>Bgb</i>	CG7959	GB19853	TC13723
<i>BgcN</i>	CG30170	GB15771	TC11100
<i>B-H1</i>	CG5529	GB10569	TC16195
<i>Bi</i>	CG3578	GB15082	TC15795
<i>Bib</i>	CG4722	GB12287	TC10832
<i>Bic</i>	CG3644	GB13433	TC12217
<i>BicC</i>	CG4824	GB14069	TC05315
<i>BicD</i>	CG6605	GB16687	TC09111
<i>Bif</i>	CG1822	GB16223	TC01382
<i>bip2</i>	CG2009	GB16554	TC14217
<i>Blimp-1</i>	CG5249	GB11750	TC14741
<i>Blue</i>	CG6451	GB20006	TC14417
<i>Bol</i>	CG4760	GB13749	TC15063
<i>Botv</i>	CG15110	GB14142	TC13377
<i>Bowl</i>	CG10021	GB18696	TC05784
<i>Br</i>	CG11491	GB14070	TC05474
<i>Brat</i>	CG10719	GB12558	TC08260
<i>Brk</i>	CG9653	GB10994	TC00748
<i>Brm</i>	CG5942	GB13381	TC11073
<i>Bs</i>	CG3411	GB15081	TC04911
<i>Bsh</i>	CG10604	GB15683	TC15394
<i>Bsk</i>	CG5680	GB16401	TC06810
<i>Btz</i>	CG12878	GB17731	TC10061
<i>Bub3</i>	CG7581	GB15882	TC11049
<i>Bun</i>	CG5461	GB10878	TC10592
<i>Bx</i>	CG6500	GB11268	TC07525
<i>Byn</i>	CG7260	GB17086	TC14076
<i>C15</i>	CG7937	GB18034	TC11749
<i>Cact</i>	CG5848	GB10655	TC02003
<i>Cactin</i>	CG1676	GB13677	TC08782
<i>Cad</i>	CG1759	GB10821	miniclust TC07576, TC07577
<i>Cad87A</i>	CG6977	GB18254	miniclust TC00221, TC00222
<i>Cad88C</i>	CG3389	GB17702	TC01129
<i>Cad89D</i>	CG14900	GB13624	TC11155
<i>Cad96Ca</i>	CG10244	GB11488	TC04976
<i>Cad96Cb</i>	CG10421	GB18252	TC10411
<i>Cad99C</i>	CG31009	GB16616	TC11374
<i>CadN</i>	CG7100	GB12853	miniclust: TC13220, TC13221, TC13226
<i>Cam</i>	CG8472	GB15633	TC01251
<i>capt</i>	CG5061	GB12447	TC01635
<i>capu</i>	CG3399	GB10982	TC12258
<i>caup</i>	CG10605	GB18111	TC03632
<i>cbt</i>	CG4427	GB10114	TC14124

Gene Name	<i>D. mel</i>	<i>A. mel</i>	<i>T. cas</i>
<i>cdc2</i>	CG5363	GB19434	TC15375
<i>cher</i>	CG3937	GB13109	TC05186
<i>Chi</i>	CG3924	GB14041	TC00244
<i>chic</i>	CG9553	GB13380	TC14115
<i>chp</i>	CG1744	GB11660	TC00643
<i>ci</i>	CG2125	GB11331	TC03000
<i>cic</i>	CG5067	GB10773	TC04697
<i>cnc</i>	CG17894	GB11981	TC04149
<i>cno</i>	CG2534	GB11919	TC14012
<i>Con</i>	CG7503	GB11011	TC08134
<i>cos</i>	CG1708	GB18262	TC08613
<i>crb</i>	CG6383	GB14525	TC04424
<i>crl</i>	CG4443	GB18477	TC03477
<i>croc</i>	CG5069	GB10529	TC02813
<i>crol</i>	CG14938	GB18263	TC06693
<i>Csk</i>	CG17309	GB16210	TC10831
<i>csw</i>	CG3954	GB10063	TC12910
<i>ct</i>	CG11387	GB17945	TC15699
<i>CtBP</i>	CG7583	GB19314	TC12453
<i>cv</i>	CG12410	GB13066	TC03620
<i>cv-2</i>	CG15671	GB10648	TC12674
<i>D</i>	CG5893	GB16191	TC13163
<i>d</i>	CG10595	GB18229	TC10547
<i>da</i>	CG5102	GB19677	TC09743
<i>Dab</i>	CG9695	GB15573	TC09426
<i>dac</i>	CG4952	GB17219	TC07637
<i>Dad</i>	CG5201	GB19187	TC04840
<i>dally</i>	CG4974	GB11050	TC14566
<i>dan</i>	CG11849	GB13750	TC16383
<i>Dfd</i>	CG2189	GB13409	TC00920
<i>Dhc64C</i>	CG7507	GB10654	TC08801
<i>disco-r</i>	CG32577	GB14651	TC01693
<i>disp</i>	CG2019	GB13340	TC10878
<i>DI</i>	CG3619	GB12464	TC04114
<i>dl</i>	CG6667	GB19537	TC07697
<i>dlg1</i>	CG1725	GB14011	TC00855
<i>Dll</i>	CG3629	GB14516	TC09351
<i>dnc</i>	CG32498	GB15311	TC12593
<i>dom</i>	CG9696	GB10524	TC12058
<i>dome</i>	CG14226	GB12159	TC01874
<i>dos</i>	CG1044	GB17687	TC08021
<i>dpn</i>	CG8704	GB12076	TC05224
<i>dpp</i>	CG9885	GB17971	TC08466
<i>Dr</i>	CG1897	GB13830	TC11744
<i>drpr</i>	CG2086	GB14962	TC00689
<i>ds</i>	CG17941	GB16221	TC07181

Gene Name	<i>D. mel</i>	<i>A. mel</i>	<i>T. cas</i>
<i>Dscam</i>	CG17800	GB15141	TC12539
<i>dsh</i>	CG18361	GB14219	TC14903
<i>dve</i>	CG5799	GB19998	TC01741
<i>dx</i>	CG3929	GB10770	TC05760
<i>ea</i>	CG4920	GB14247	TC13277
<i>EcR</i>	CG1765	GB15434	miniclust TC12112, TC12113
<i>ed</i>	CG12676	GB13261	TC12257
<i>Egfr</i>	CG10079	GB12207	TC03986
<i>egh</i>	CG9659	GB18841	TC08154
<i>egl</i>	CG4051	GB10691	TC15348
<i>elB</i>	CG4220	GB11405	TC00868
<i>emc</i>	CG1007	GB19457	TC00024
<i>en</i>	CG9015	GB15566	TC08952
<i>ena</i>	CG15112	GB17061	TC02504
<i>enc</i>	CG10847	GB10730	TC09582
<i>Eph</i>	CG1511	GB12585	TC06032
<i>epsin-like</i>	CG31170	GB15239	TC12168
<i>esg</i>	CG3758; CG3956	GB12880	TC14474
<i>Ets97D</i>	CG6338	GB12455	TC14932
<i>eve</i>	CG2328	GB10623	TC09469
<i>exd</i>	CG8933	GB15837	TC11311
<i>exu</i>	CG8994	GB19360	TC09494
<i>eya</i>	CG9554	GB11435	TC08985
<i>eyg</i>	CG10488	GB15698	TC07194
<i>eyes</i>	CG7245	GB19577	TC10461
<i>f</i>	CG5424	GB14006	TC05627
<i>faf</i>	CG1945	GB10029	TC10455
<i>fas</i>	CG17716	GB10494	TC05058
<i>Fas1</i>	CG6588	GB15085	TC11300
<i>Fas2</i>	CG3665	GB14520	TC07253
<i>Fas3</i>	CG5803	GB17084	TC14942
<i>fat2</i>	CG7749	GB16822	TC00401
<i>fbf</i>	CG5725	GB11006	TC04491
<i>Fim</i>	CG8649	GB11573	TC01769
<i>flkh</i>	CG10002	GB14416	TC13245
<i>fng</i>	CG10580	GB17604	TC11785
<i>Fpps</i>	CG12389	GB12385	TC09257
<i>fra</i>	CG8581	GB10232	TC09930
<i>frc</i>	CG3874	GB19692	TC03250
<i>fru</i>	CG14307	GB17617	TC00589
<i>fry</i>	CG32045	GB19491	TC00108
<i>fs(1)N</i>	CG11411	GB18228	TC07081
<i>ft</i>	CG3352	GB10152	miniclust TC07877, TC07878
<i>ftz-f1</i>	CG4059	GB16873	TC02550
<i>fu</i>	CG6551	GB10754	TC06825
<i>fus</i>	CG8205	GB16152	TC09268

Gene Name	<i>D. mel</i>	<i>A. mel</i>	<i>T. cas</i>
<i>futsch</i>	CG3064	GB11509	TC01001
<i>fw</i>	CG1500	GB11792	TC02811
<i>fwd</i>	CG7004	GB19870	TC11124
<i>fy</i>	CG13396	GB16647	TC08255
<i>fz</i>	CG17697, CG3646?	GB18517	TC14055
<i>fz2</i>	CG9739	GB12765	TC03407
<i>Gap1</i>	CG6721	GB12604	TC14250
<i>GATAd</i>	CG5034	GB18471	TC06488
<i>gbb</i>	CG5562	GB18733	miniclust TC14017, TC14018
<i>gcl</i>	CG8411	GB20151	TC01571
<i>gcm</i>	CG12245	GB19906	TC14730
<i>gkt</i>	CG8825	GB19412	TC01393
<i>Gl</i>	CG9206	GB10667	TC12455
<i>gl</i>	CG7672	GB15041	TC12565
<i>Gli</i>	CG3903	GB12309	TC10824
<i>glu</i>	CG11397	GB15940	TC00075
<i>grh</i>	CG5058	GB13030	TC04589
<i>grn</i>	CG9656	GB11761	TC02315
<i>gro</i>	CG8384	GB11858	TC01206, TC01371
<i>grp</i>	CG17161	GB16086	TC01409
<i>gsb</i>	CG3388	GB15632	TC06788
<i>gsb-n</i>	CG2692	GB14483	TC05342
<i>Gsc</i>	CG2851	GB12726	TC11819
<i>gt</i>	CG7952	GB16015	TC07492
<i>Gug</i>	CG6964	GB18685	TC14949
<i>h</i>	CG6494	GB14857	TC12851
<i>H</i>	CG5460	GB17995	TC08831
<i>Hand</i>	CG18144	GB19031	TC04726
<i>hb</i>	CG9786	GB19977	TC13553
<i>hbn</i>	CG33152	GB13412	TC08926
<i>hdc</i>	CG15532	GB11853	TC01081
<i>Hem</i>	CG5837	GB13021	TC01541
<i>hep</i>	CG4353	GB17167	TC00385
<i>hh</i>	CG4637	GB14574	TC01364
<i>hkb</i>	CG9768	GB14090	TC10992
<i>HLHmgamma</i>	CG8333	GB19475	TC06580
<i>homer</i>	CG11324	GB14479	TC15941
<i>hop</i>	CG1594	GB16422	TC08648
<i>how</i>	CG10293	GB13678	TC00827
<i>hpo</i>	CG11228	GB18142	TC04606
<i>hth</i>	CG17117	GB18348	TC08629
<i>htl</i>	CG7223	GB19884	TC04713
<i>hts</i>	CG9325	GB15113	TC04497
<i>if</i>	CG9623	GB13598	TC01667
<i>in</i>	CG16993	GB10716	TC01193
<i>ind</i>	CG11551	GB14802	TC06888

Gene Name	<i>D. mel</i>	<i>A. mel</i>	<i>T. cas</i>
<i>InR</i>	CG18402	GB18331	TC10784
<i>insc</i>	CG11312	GB14948	TC01320
<i>ix</i>	CG13201	GB19364	TC10345
<i>jar</i>	CG5695	GB10158	TC00685
<i>Jra</i>	CG2275	GB12004	TC06814
<i>kay</i>	CG15509	GB12212	TC11870
<i>kek1</i>	CG12283	GB17490	TC07055
<i>kek2</i>	CG4977	GB12232	TC08448
<i>kek5</i>	CG12199	GB19774	TC07110
<i>kek6</i>	CG1804	GB11036	TC08070
<i>ken</i>	CG5575	GB18560	TC01442
<i>Khc</i>	CG7765	GB10827	TC11608
<i>kirre</i>	CG3653	GB11991	TC02914
<i>kkv</i>	CG2666	GB17253	TC14634
<i>klar</i>	CG17046	GB19116	TC01444
<i>Klp3A</i>	CG8590	GB13714	TC15915
<i>klu</i>	CG12296	GB19470	TC02783
<i>kn</i>	CG10197	GB14092	TC01270
<i>knk</i>	CG6217	GB13189	TC10653
<i>knrl</i>	CG4761	GB13710	TC03413
<i>Kr</i>	CG3340	GB16053	TC11460
<i>Krn</i>	CG32179	GB19294	TC03429
<i>krz</i>	CG1487	GB13683	TC01639
<i>ksr</i>	CG2899	GB12129	TC05910
<i>kst</i>	CG12008	GB15664	TC01109
<i>kuz</i>	CG7147	GB13192	TC01512
<i>l(2)gl</i>	CG2671	GB20098	TC15986
<i>l(2)tid</i>	CG5504	GB10850	TC08059
<i>l(3)mbt</i>	CG5954	GB18742	TC01922
<i>lab</i>	CG1264	GB14027	TC00926
<i>lbe</i>	CG6545	GB10613	TC11748
<i>lgs</i>	CG2041	GB13227	TC10773
<i>lic</i>	CG12244	GB16739	TC05618
<i>lilli</i>	CG8817	GB12566	TC07363
<i>Lim1</i>	CG11354	GB12408	TC14939
<i>lin</i>	CG11770	GB16309	TC11514
<i>lkb1</i>	CG9374	GB10693	TC12166
<i>loco</i>	CG5248	GB15675	TC09818
<i>lola</i>	CG12052	GB12094	TC03097
<i>lqf</i>	CG8532	GB16241	TC05393
<i>lwr</i>	CG3018	GB16281	TC06191
<i>lz</i>	CG1689	GB16431	TC05796
<i>Mad</i>	CG12399	GB11582	TC14921
<i>mad2</i>	CG17498	GB12183	TC13206
<i>mael</i>	CG11254	GB17844	TC08172
<i>mago</i>	CG9401	GB10361	TC16112

Gene Name	<i>D. mel</i>	<i>A. mel</i>	<i>T. cas</i>
<i>mam</i>	CG8118	GB11946	TC00809
<i>mav</i>	CG1901	GB13629	TC04299
<i>mbc</i>	CG10379	GB16498	TC12454
<i>mbl</i>	CG14477	GB13919	TC14149
<i>Med</i>	CG1775	GB18981	TC10848
<i>Mef2</i>	CG1429	GB14174	TC10850
<i>mew</i>	CG1771	GB16257	TC06750
<i>Mhc</i>	CG17927	GB11965	TC05924
<i>mib1</i>	CG5841	GB13756	TC14445
<i>mib2</i>	CG17492	GB19770	TC13531
<i>mio</i>	CG7074	GB17477	miniclust TC12249, TC12250
<i>mirr</i>	CG10601	GB11441	TC03634
<i>Mkk4</i>	CG9738	GB13132	TC05515
<i>mle</i>	CG11680	GB14139	TC03184
<i>mnb</i>	CG7826	GB20129	TC07717
<i>Moe</i>	CG10701	GB11282	TC00998
<i>msl-2</i>	CG3241	GB18291	TC01753
<i>msl-3</i>	CG8631	GB19559	TC11005
<i>msps</i>	CG5000	GB10660	TC04968
<i>Myb</i>	CG9045	GB12498	TC10032
<i>Myd88</i>	CG2078	GB12344	TC03185
<i>mys</i>	CG1560	GB19541	TC11707
<i>N</i>	CG3936	GB10567	TC04393
<i>nau</i>	CG10250	GB13572	TC15855
<i>ndl</i>	CG10129	GB19590	TC00870
<i>neb</i>	CG10718	GB19627	TC13493
<i>nej</i>	CG15319	GB12228	TC08222
<i>NetB</i>	CG10521	GB15820	TC02285
<i>neur</i>	CG11988	GB14273	TC00216
<i>nkd</i>	CG11614	GB11962	TC01226
<i>Nle</i>	CG2863	GB10500	TC04394
<i>nmo</i>	CG7892	GB12339	TC12666
<i>noc</i>	CG4491	GB17714	TC00693
<i>Nrg</i>	CG1634	GB11846	TC01889
<i>Ntf-2</i>	CG1740	GB19311	TC12876
<i>nub</i>	CG6246	GB16262	miniclust TC07645, TC07646
<i>numb</i>	CG3779	GB18756	TC12074
<i>oaf</i>	CG9884	GB18014	TC08462
<i>oc</i>	CG12154	GB16866	miniclust TC03354, TC03355
<i>okr</i>	CG3736	GB16633	TC15104
<i>opa</i>	CG1133	GB12480	TC10234
<i>orb</i>	CG10868	GB12560	TC11262
<i>orb2</i>	CG5735	GB15835	TC14191
<i>org-1</i>	CG11202	GB12301	TC15327
<i>par-1</i>	CG8201	GB15281	TC02567
<i>pav</i>	CG1258	GB14780	TC13058

Gene Name	<i>D. mel</i>	<i>A. mel</i>	<i>T. cas</i>
<i>Pax</i>	CG31794	GB19612	TC10609
<i>pb</i>	CG31481	GB11988	TC00925
<i>Pc</i>	CG32443	GB12523	TC12316
<i>peb</i>	CG12212	GB18410	miniclust TC09560, TC09561
<i>pelo</i>	CG3959	GB10750	TC01682
<i>pho</i>	CG17743	GB11924	TC15577
<i>pip</i>	CG9614	GB11396	TC02293
<i>Pka-C1</i>	CG4379	GB17175	TC05012
<i>plexA</i>	CG11081	GB16227	TC01765
<i>plexB</i>	CG17245	GB15035	TC04144
<i>pll</i>	CG5974	GB16397	TC15365
<i>pnr</i>	CG3978	GB19895	TC10407
<i>pnt</i>	CG17077	GB18613	TC14512
<i>polo</i>	CG12306	GB10053	TC14023
<i>POSH</i>	CG4909	GB10700	TC07357
<i>prd</i>	CG6716	GB15469	TC15804
<i>pros</i>	CG17228	GB14533	TC10596
<i>Psn</i>	CG18803	GB15051	TC10178
<i>ptc</i>	CG2411	GB16349	TC04745
<i>Ptx1</i>	CG1447	GB15295	TC01113
<i>pum</i>	CG9755	GB10504	TC05073
<i>put</i>	CG7904	GB18110	TC11357
<i>Pvf3</i>	CG31629	GB12742	TC08417
<i>qkr54B</i>	CG4816	GB10438	TC08871
<i>Rab11</i>	CG5771	GB17764	TC04925
<i>Rab5</i>	CG3664	GB15021	TC14786
<i>Rac1</i>	CG2248	GB11373	TC02141
<i>repo</i>	CG31240	GB14165	TC13309
<i>ret</i>	CG14396	GB19007	TC12783
<i>retn</i>	CG5403	GB18541	TC08720
<i>Rho1</i>	CG8416	GB13135	TC09158
<i>rho-4</i>	CG1697	GB16638	TC06133
<i>robo</i>	CG13521	GB17658	TC02775
<i>Rop</i>	CG15811	GB12540	TC11120
<i>run</i>	CG1849	GB11654	TC06542
<i>Rx</i>	CG10052	GB19717	TC09912
<i>S</i>	CG4385	GB13389	TC12408
<i>salm</i>	CG6464	GB19037	TC13501
<i>sax</i>	CG1891	GB19039	TC15948
<i>sca</i>	CG17579	GB11902	TC03194
<i>Scr</i>	CG1030	GB13491	TC00917
<i>scrt</i>	CG1130	GB18548	TC16391, TC16394
<i>Sema-1a</i>	CG18405	GB11468	TC10143, TC14179
<i>Sema-2a</i>	CG4700	GB16014	TC01219
<i>Sema-5c</i>	CG5661	GB11625	TC01449
<i>sev</i>	CG18085	GB12743	TC01239

Gene Name	<i>D. mel</i>	<i>A. mel</i>	<i>T. cas</i>
<i>sgg</i>	CG2621	GB15424	TC08141
<i>shf</i>	CG3135	GB10372	TC01979
<i>shg</i>	CG3722	GB17989	TC13570
<i>shn</i>	CG7734	GB19562	TC09542
<i>Six4</i>	CG3871	GB10752	TC03853
<i>sli</i>	CG8355	GB19929	TC00214
<i>sll</i>	CG7623	GB18325	TC15899
<i>slmb</i>	CG3412	GB10096	TC01086
<i>slmo</i>	CG9131	GB17877	TC14470
<i>slp2</i>	CG2939	GB12972	TC08062
<i>smo</i>	CG11561	GB10379	TC05545
<i>Smox</i>	CG2262	GB19607	TC10162
<i>so</i>	CG11121	GB15213	TC13834
<i>sob</i>	CG3242	GB16145	TC05788
<i>sog</i>	CG9224	GB16025	TC12650
<i>SoxN</i>	CG18024	GB19937	TC14065
<i>Sp1</i>	CG1343	GB15089	TC11697
<i>spir</i>	CG10076	GB14715	TC14290
<i>Spred</i>	CG10155	GB13001	TC01559
<i>spz</i>	CG6134	GB15688	TC01054
<i>Spz3</i>	CG7104	GB17772	TC05940
<i>ss</i>	CG6993	GB11901	TC11105
<i>Stat92E</i>	CG4257	GB18923	TC13218
<i>stau</i>	CG5753	GB13840	TC04615
<i>stumps</i>	CG31317	GB10564	TC11323
<i>sty</i>	CG1921	GB12262	TC07446
<i>Su(H)</i>	CG3497	GB11411	TC14468
<i>su(Hw)</i>	CG8573	GB15778	TC08904
<i>sub</i>	CG12298	GB18655	TC13546
<i>sv</i>	CG11049	GB18397	TC03569, TC03570?
<i>svp</i>	CG11502	GB17100	TC01722
<i>Taf2</i>	CG6711	GB16704	TC11774
<i>Taf4</i>	CG5444	GB13892	TC00268
<i>Taf5</i>	CG7704	GB15901	TC13143
<i>Taf6</i>	CG32211	GB15269	TC13033
<i>Taf7</i>	CG2670	GB17187	TC03817
<i>Taf8</i>	CG7128	GB14552	TC09938
<i>tafazzin</i>	CG8766	GB11956	TC09822
<i>Tak1</i>	CG18492	GB14664	TC05572
<i>Tehao</i>	CG7121	GB18520	TC04438
<i>Ten-m</i>	CG5723	GB12554	TC08116
<i>TepII</i>	CG7052	GB12605	TC09667
<i>Tequila</i>	CG4821	GB12538	TC15110
<i>TER94</i>	CG2331	GB20017	TC09174
<i>tkv</i>	CG14026	GB15083	TC06474
<i>tlk</i>	CG32782	GB15719	TC08538

Gene Name	<i>D. mel</i>	<i>A. mel</i>	<i>T. cas</i>
<i>tl</i>	CG1378	GB20053	TC00441
<i>Toll-6</i>	CG7250	GB17781	TC04895
<i>Toll-7</i>	CG8595	GB15177	TC04474
<i>Tollo</i>	CG6890	GB10640	TC04898
<i>Tor</i>	CG5092	GB11213	TC05546
<i>torp4a</i>	CG3024	GB13575	TC00824
<i>tou</i>	CG10897	GB16601	TC09937
<i>toy</i>	CG11186	GB11714	TC07409
<i>Tpi</i>	CG2171	GB17473	TC07346
<i>TpnC25D</i>	CG6514	GB19642	TC06493
<i>TpnC41C</i>	CG2981	GB13594	TC12196
<i>TpnC73F</i>	CG7930	GB10545	TC12704
<i>Tppl</i>	CG3991	GB13954	TC04702
<i>Tpr2</i>	CG4599	GB19952	TC14273
<i>Tps1</i>	CG4104	GB12797	TC07883
<i>tra2</i>	CG10128	GB11130	TC12340
<i>trh</i>	CG6883	GB17871	TC01448
<i>trn</i>	CG11280	GB19945	TC01975
<i>trx</i>	CG8651	GB16330	TC04768
<i>tsl</i>	CG6705	GB18663	TC08090
<i>tud</i>	CG9450	GB17525	TC03753
<i>twi</i>	CG2956	GB18475	TC14598
<i>Ubx</i>	CG10388	GB11524	TC00906
<i>ush</i>	CG2762	GB16457	TC13689
<i>usp</i>	CG4380	GB16648	miniclust TC14027, TC14028
<i>Vang</i>	CG8075	GB17442	TC01197
<i>vas</i>	CG3506	GB14804	TC10103
<i>wg</i>	CG4889	GB19984	TC14084
<i>wit</i>	CG10776	GB15265	TC09314
<i>wkd</i>	CG5344	GB17762	TC15650
<i>Wnt10</i>	CG4971	GB13356	TC14086
<i>Wnt2</i>	CG1916	GB16102	TC09318
<i>Wnt6</i>	CG4969	GB14164	TC13707
<i>Zw</i>	CG12529	GB15779	TC13648

Table S12. A core of highly conserved head developmental genes

category	vertebrate name	synonyms	Tribolium name	Tc number	remark
expressed in the anterior head Anlage (25)	<i>Otx</i>		<i>Tc-othodenticle-1</i>	03354	Li, Y. et al, 1996
			<i>Tc-othodenticle-2</i>	03355	
	<i>Tlx</i>		<i>Tc-tailless</i>	00441	Schröder, R. et al; 2000
	<i>gsc</i>		<i>Tc-goosecoid</i>	12684	
	<i>rx</i>		<i>Tc-rx</i>	09911	
	<i>fez</i>		<i>Tc-fez</i>	04673	
	<i>Pax6</i>		<i>Tc-twin of eyeless</i>	07409	
			<i>Tc-eyeless/Pax6</i>	08176	
	<i>irx</i>		<i>Tc-mirror</i>	03634	paralog <i>Tc-iroquois</i> not expressed in the head
	<i>emx</i>		<i>Tc-empty-spiracles</i>	11763	
	<i>nkx2.1</i>		<i>Tc-scarecrow</i>	08996	
	<i>Gbx</i>		<i>Tc-unplugged</i>	09309	
	<i>Dbx</i>	<i>hlx</i>	<i>Tc-dbx</i>	15146	
	<i>FoxG1</i>	<i>brain-factor1</i>	<i>Tc-sloppy paired1</i>	08064	Choe, CP. et al, 2007
			<i>Tc-sloppy paired 2</i>	08062	
	<i>six1</i>		<i>Tc-sine-oculis</i>	13834	
	<i>six3</i>		<i>Tc-optix/six3</i>	00361	
	<i>six4</i>		<i>Tc-six4</i>	03853	
	<i>lim-1</i>		<i>Tc-lim-1</i>	14939	
	<i>shh</i>		<i>Tc-hedgehog</i>	01364	
	<i>eya</i>		<i>Tc-eyes absent</i>	08985	
	<i>otp</i>		<i>Tc-orthopedia</i>	08928	
	<i>Wnt1</i>		<i>Tc-wingless</i>	14084	Nagy, LM. et al, 1994
	<i>Gli3</i>		<i>Tc-cubitus interruptus</i>	03000	
	<i>engrailed</i>		<i>Tc-engrailed</i>	08952	Brown, S. et al, 1994
	<i>Pitx</i>		<i>Tc-ptx</i>	01112	

	<i>Dlx</i>		<i>Tc-Distal-less</i>	09351	Beermann, A. et al, 2001
	<i>SP-8</i>		<i>Tc-SP8</i>	11697	Beermann, A. et al, 2003
not expressed in the anterior head (2)	<i>barH</i>		<i>Tc-barH</i>	16195	
	<i>arx</i>		<i>Tc-munster</i>	06110	<i>Drosophila</i> : expression in the larval eyes
not found in the genomes of <i>Tribolium</i> and <i>Drosophila</i> (3)	<i>vax</i>				
	<i>hesx1</i>	<i>Ganf, Anf</i>			
	<i>atx</i>	<i>Dmbx, otx3</i>			

Table S13a. A survey of *Tribolium* candidate ventral appendage genes.

Gene	Tribolium gene	Leg Expression			Gnathal expression	pRNAi effect seen	matches Drosophila phenotype	remarks/references
		early	middle	late				
AP-2	TC09922	√	√	√	√	no		
Apterous	TC03973	√				no		
	TC13329							
	TC13331							
Aristaless	TC13332		√	√	√	yes	yes	Beermann <i>et Al.</i> (2004) ¹³¹
arm1	TC12388					no		Ubiquitous
arm2	TC12389					no		Ubiquitous
								No embr.expression
Awh	TC03238							
Bar H1/H2	TC16195	√	√	√	√			
Ci	TC03000	√				yes	yes	
Dachsous	TC07180	√	√	√	√	yes	yes	
Disco	TC01693							Ubiquitous
dLIM	TC14939	√	√	√	√	no		
Dll	TC09351	√	√	√	√	yes	yes	Beermann <i>et Al.</i> (2001) ¹³²
Hdc	TC01076							No embr.expression
Hh	TC01364					yes	yes	hh-pathway
LHX 9	TC03974	√	√	√	√	no		
Numb	TC12074			√				
Odd	TC05785							hh-pathway
Omb	TC15795	√	√	√	√	yes	yes	
Patched	TC04745	√	√	√	√			hh-pathway
Sp5	TC11696	√	√	√	√			
Sp8	TC11697	√	√	√	√	yes	yes	Beermann <i>et Al.</i> (2004) ¹³¹
Spineless	TC11105				√antenna	no		
supernumerary limbs(slmb)	TC01086							Ubiquitous
tipsy/C15	TC11749		√	√	√	yes	yes	
Wnt1 = wg	TC14084	√	√	√	√	yes	yes	Nagy <i>et Al.</i> (1999) ¹³³ , Ober <i>et Al.</i> (2006) ¹³⁴
Wnt 10	TC14086	√	√	√	√	no		
Wnt 11	TC14270	√	√	√	√	no		
Wnt 5	TC09318	√	√	√	√	no		
Wnt 6	TC13707	√	√	√	√	no		
Wnt 7	TC10155	√	√	√	√	no		

Table S13b. A survey of *Tribolium* candidate wing genes

	gene	Tc gene#	expression		RNAi		ref	remarks
			elytron	hindwing	elytron	hindwing		
anterior/posterior	en/inv		!	!	n.d.	n.d.	1	Glean#: 08952 (en), 09896+09897 (inv)
	hh	01364	!	!	Con	n.d.		
	ci	03000	!	!	n.d.	n.d.		
	ptc	04745	!	!	n.d.	n.d.		
	dpp	08466	!	!	Abs	Abs		
	dad	04840	!	!	Abs	Abs		
	brk	00748	n.d.	n.d.	Abs	Abs		
	omb	15796*	!	!	Con	Con	1	
dorsal/ventral	sal	13501	"	!	n.d.	n.d.	1	solo ortholog for Dm-salm and salr
	apA	03973	!	!	Con+	Con		apA+apB double RNAi
	apB	03974*	!	!				
	N	04393	n.d.	n.d.	Con+	Con		phenotype at lower dsRNA concentrations (LL at 1ug/ul)
	Ser	10113	n.d.	n.d.	Con	Con		
	wg	14084	!	!	Abs	Abs		
	Dll	09351	!	!	Con	Con		
vein and sensory	srf	04911	!	!	Con	Con		
	rho	06329	n.d.	n.d.	Abs	Abs		several paralogs in Tribolium
	kni	03413	!	!	n.d.	n.d.		
	knot	01270	"	!	n.d.	n.d.		
	iro	03632	!#	!	n.d.	n.d.	1	solo ortholog for Dm-ara and caup
	ab	13098	+	!	n.d.	n.d.		
	net	05579	n.d.	n.d.	Abs	Abs		
	DI	04114	!	!	n.d.	n.d.		
	emc	00024	n.d.	n.d.	Con	Con		
	ASH	08433	!	!	Con+	Con	1,2	solo ortholog for Dm-ac and sc
bodywall/wing	ase	08437	n.d.	n.d.	Abs	Abs		
	tsh	12322	"	"	Abs-Con	Abs-Con		solo ortholog for Dm-tsh and tio
	hth	08631*	"	"	Abs-Con	Abs-Con		
	nub	07646	!	!	Con+	Con		solo ortholog for Dm-nub and pdm2
	vvl	14350*	!	!	LL	LL		
	vg	10897	!	!	Con	Con		
Hox	Cx/Scr	00917*	"	"	Abs-Con	Abs-Con	1	
	ptl/Antp	00921	"	"	Con/Incon?	Abs-Con?		
	Utx/Ubx	00903*	"	!	Abs-Con?	Con/Incon?		

RNAi phenotypes***Con: consistent with the *Drosophila* forewing phenotypeCon+: consistent with the *Drosophila* forewing phenotype + additional phenotypeINcon: inconsistent with the *Drosophila* forewing phenotypeAbs-Con: Absence of the phenotype is consistent with *Drosophila* results. Other phenotypes are observed in other tissues, indicating successful RNAi.

Abs: absence of any visible phenotype****

LL: unable to assess the phenotype due to larval lethality (1ug/ul dsRNA). Might be visible at lower concentrations.

References

1 Tomoyasu et al. 2005

2 Tomoyasu and Denell 2004

Notes

*The gene spans more than one GLEAN prediction

**Expression in sensory organs but not in pre-vein regions

***Phenotypes observed in tissues other than dorsal appendages (such as body wall) are excluded

****The absence of phenotype might be due to ineffective RNAi.

Table S14. Survey of *Tribolium* eye gene orthologs

Gene	Acronym	Functional context	Beetle	Fruit fly
distal antenna-related (hernandez)	danr	eye development	XP_969154	NP_651343
distal antenna (fernandez)	dan	eye development		NP_651346
eyegone	eyg	eye development	XP_972576	NP_001014582
twin of eyegone	toe			NP_524041
eyeless	ey	eye development	XP_975543	NP_524628
twin of eyeless	toy			NP_524638
eyes absent	eya	eye development	XP_974387	NP_723188
sine oculis	so	eye development	XP_972167	NP_476733
dachshund	dac	eye development	XP_969771	NP_723969
optix	optix	eye development	XP_975128	NP_524695
teashirt	tsh	eye development	XP_975699	NP_523615
tip-top	tio			NP_524733
optix binding protein	obp	eye development	XP_968302 (1-554aa)	NP_724479
sine oculis binding protein	sbp	eye development	XP_967801	NP_610703
microphthalmia associated transcription factor	mitf	eye development	TC14225	NP_001015077
hairy	h	eye development	XP_971935	NP_523977
extramacrochaete	emc	eye development	TC00024	NP_523876
daughterless	da	eye development	XP_973272	NP_477189

atonal	ato	eye development	XP_970709	NP_731223
scabrous	sca	eye development	XP_972571	NP_476710
glass	gl	eye development	NP_001034508	NP_476854
Photoreceptor-cell-specific nuclear receptor	PNR	eye development	XP_970391	NP_611032
pebbled	peb	eye development	XP_973372	NP_476674
bunched	bun	eye development	XP_972199	NP_525103
senseless	sens	eye development	XP_974438	NP_524818
rough	ro	eye development	XP_968945	NP_524521
lozenge	lz	eye development	XP_971415	NP_511099
runt	run	eye development	XP_969277	NP_523424
big brother	bgb	eye development	XP_966458	NP_477065
brother	bro	eye development		NP_477066
seven up	svp	eye development	XP_967537	NP_524325
BarH1	B-H1	eye development	XP_969286	NP_523387
BarH2	B-H2	eye development		NP_523386
prospero	pros	eye development	XP_971664	NP_731565
sevenless	sev	eye development	XP_970953	NP_511114
bride of sevenless	boss	eye development	gb CH476256.1 _39 geneid_v1.2_predicted_protein_39	NP_542440
phyllopod	phyl	eye development	-	NP_725394
seven in absentia	sina	eye development	XP_971492	NP_476725

anterior open (yan)	aop (yan)	eye development	XP_975017	NP_722766
shaven (sparkling)	sv (spa)	eye development	XP_968041	NP_524633
orthodenticle	otd	eye development	NP_001034513 NP_001034526	NP_511091
cut	ct	eye development	XP_970668	NP_524764
tramtrack	ttk	eye development	XP_971335	NP_733443
spalt-major	salm	eye development		NP_723670
spalt-related	salr	eye development	XP_973229	NP_523548s
homothorax	hth		NP_001034489	NP_476578
spineless	ss	eye development	XP_967876	NP_476748
embryonic lethal, abnormal vision	elav	eye development	-	NP_525033(elav) NP_572842(fne) NP_476937(Rbp9)
drosocrystallin	dcry	eye development	-	NP_476906
klignon	klg	eye development	?	NP_524454
chaoptic	chp	eye development	XP_975453	NP_524605
SoxN	SoxN	eye development	XP_974496	NP_524735
oncut	oncut	eye development	XP_624996	NP_524842
prominin CG14955	prom	eye development	XP_001122309	NP_647770 CG14955
Munster (PvuII-PstI homology 13)	munster(Pph13)	eye development	XP_001121339	NP_477330
eyes shut	eyes	eye development	XP_001122168	NP_001027571
warts	wt	eye development	XP_973217	NP_733403

melted	mlt	eye development	XP_968590+ XP_968433	NP_523953
neither inactivation nor afterpotential G	ninaG	phototransduction	-	NP_650070
neither inactivation nor afterpotential A	ninaA	phototransduction	XP_973192	NP_476656
Arrestin 1	Arr1	phototransduction	XP_966595	NP_476681
Arrestin 2	Arr2	phototransduction	XP_972592	NP_523976
retinal degeneration C	rdgC	phototransduction	XP_974915	NP_536738
G protein-coupled receptor kinase 1	Gprk 1	phototransduction	XP_966480	NP_001036438
G protein-coupled receptor kinase 2	Gprk 2	phototransduction	TC11652	NP_476867
Rhodopsin 1	Rh1	phototransduction	XP_973147	NP_524407
Rhodopsin 2	Rh2			NP_524398
Rhodopsin 6	Rh6			NP_524368
Rhodopsin 3	Rh3	phototransduction	XP_970344	NP_524411
Rhodopsin 4	Rh4			NP_476701
Rhodopsin 5	Rh5	phototransduction	-	NP_477096
Rhodopsin 7	Rh7	phototransduction	-	NP_524035
G γ 30A	G protein γ 30A	phototransduction	TC15232	NP_524807
G β 76C	G protein β 76C	phototransduction	XP_973851	NP_523720
G α 49B	G protein α 49B	phototransduction	XP_966311	NP_523718
no receptor potential A	norpA	phototransduction	TC08027	NP_525069
inactivation nor afterpotential D	inaD	phototransduction	TC09802	NP_726260
no inactivation nor afterpotential C	ninaC	phototransduction	XP_968286	NP_723271
transient receptor potential	Trp	phototransduction	XP_968670	NP_476768
transient receptor potential-like	Trpl	phototransduction	XP_968598	NP_476895
transient receptor potential γ	Trpg	phototransduction	TC07028	NP_609802
inactivation nor afterpotential F	inaF	phototransduction	-	NP_572744

retinal degeneration B	rdgB	phototransduction	TC03397	NP_727733
retinal degeneration A	rdgA	phototransduction	XP_972412	NP_511092
CDP diclyceride synthetase	CdsA	phototransduction	XP_975257 XP_968133	NP_524661
lazaro	Laza	phototransduction	-	NP_649391
inactivation nor afterpotential C	inaC	phototransduction	-	NP_476863
Calphotin	Cpn	phototransduction	-	NP_731673
Calx	Calx	phototransduction	XP_974130	NP_524423
pinta	retinoid binding (retina localization?)	phototransduction	XP_974921	NP_651042
stunted	sun	phototransduction	XP_970169	NP_524682
<i>Drosophila</i> phosphatidylinositol synthase	dpis	phototransduction	XP_967177	NP_573055
Phospholipase D	Pld	phototransduction	XP_969697	NP_523627

Table S15. The 103 Homeobox genes of *Tribolium castaneum* This includes the incomplete homeobox sequence of the *Otp* gene found in the present assembly, but doesn't include Pax2/5/8 or Pax 1/9, which lack complete homeoboxes but are derived from homeobox-containing genes. Two genes are present that proved impossible to classify and presumably represent Coleoptera-specific rapidly evolving genes (*BeetleBox1* and 2). In cases where duplication history was ambiguous due to lack of signal in the homeodomain, full-length protein alignments were used to classify genes (see Comments column). * mistake present in the gene model (ie. incorrect gene structure). - no gene model present in the version 2.0 assembly. ^a Butts *et al.* in prep.

Class	Family	Name	Protein model	Comments
Antp	BarH	<i>Bh</i>	TC16195	
	Bsx	<i>Bsh</i>	TC15394*	
	Cdx	<i>cad1</i>	TC07577	
		<i>cad2</i>	TC07576	The caudal genes are a beetle-specific duplication.
	Dbx	<i>Dbx</i>	TC15146	
	Dlx	<i>Dll</i>	TC09351	
	Emx	<i>Ems</i>	TC11763	
	En	<i>en</i>	TC09897	Both genes were present in the last common ancestor of Endopterygota. Gene conversion events have lead to scrambled phylogenetic signal within the homeobox (Peel <i>et al.</i> 2006). Assignment was made based upon full protein sequence alignment.
		<i>inv</i>	TC08952	
	Evx	<i>eve</i>	TC09469	
	Gbx	<i>unpg</i>	TC09309	
	Gsx	<i>Ind</i>	TC06888	
	Hex	<i>Hex</i>	TC04555*	
	Hlx	<i>Hlx</i>	TC08368	
	Hmx	<i>Hmx</i>	TC12136	
	Hox	<i>lab</i>	TC00926	
		<i>mvp/pb</i>	TC00925	
		<i>zen1</i>	TC00922	
		<i>zen2</i>	TC00921	
		<i>Dfd</i>	TC00920	
		<i>Cx/Scr</i>	TC00917	
		<i>ftz</i>	TC00916	

		<i>ptl/antp</i>	TC00912	
		<i>Ubx</i>	TC00903	
		<i>abd-A</i>	TC00894	
		<i>Abd-B</i>	TC00889	
	Lbx	<i>Lbx</i>	TC11748	
	Mnx	<i>Exex</i>	TC09461	
	Mox	<i>Btn</i>	-	
	Msx	<i>Dr1</i>	TC12748	Probably a gene pair in the insect ancestor, with gene conversion events leading to similar homeobox sequences in different insect lineages ¹³⁵ .
		<i>Dr2</i>	TC11744	
	Msx-like	<i>Msx-like</i>	TC15928	
	NK1	<i>Slou</i>	TC12332	
	NK2	<i>Vnd</i>	TC07014	
		<i>Scro</i>	TC08996	
	NK3	<i>Bap</i>	TC12743	
	NK4	<i>Tin</i>	TC11745	
	NK6	<i>Hgtx</i>	TC14200	
	NK7	<i>NK7</i>	TC13614*	
	Not	<i>Not</i>	TC00483	
	Ro	<i>Ro</i>	-	
	Tlx	<i>Cl5</i>	TC11749	
	Named elsewhere ^a	<i>TcaCg11085</i>	TC00424	
	Named elsewhere ^a	<i>TcaCg13424</i>	TC09463	
	Named elsewhere ^a	<i>TcaCg34031</i>	TC01164	
Prd	Al	<i>Al</i>	TC13331*	
	Drgx	<i>Drgx</i>	TC05600	
	Eyg	<i>Eyg</i>	TC07194	
	Gsc	<i>Gsc</i>	TC11819	
	Hbn	<i>Hbn</i>	TC08926	
	Pax3/7	<i>Gsb</i>	TC06788	
		<i>Gsbn</i>	TC05342	
		<i>Prd</i>	TC15804	
	Pax6	<i>Ey</i>	TC08176	Inferred duplication at the base of the Endopterygota based upon full protein alignment.
		<i>Toy</i>	TC07409	
	Phox	<i>Phox</i>	-	
	Pitx	<i>Ptx</i>	TC01112	
	Prop	<i>Prop</i>	TC07335	
	Prrx	<i>Prrx</i>	TC00527	

	Otp	<i>Otp</i>	TC08928*	Inferred duplication at the base of the Endopterygota based upon full protein alignment.
	Otx	<i>otd2</i>	TC03355	
		<i>otd1</i>	TC03354	
	Repo	<i>repo</i>	TC13309	
	Rx	<i>Rx</i>	TC09911	
	Shox	<i>Shox</i>	TC01726	
	Unc4	<i>Unc4</i>	TC05661	
	Vsx	<i>Vsx</i>	TC07654	
	CG2819	<i>TcaCg2819</i>	TC06110	
	CG11294	<i>TcaCg11294</i>	TC11539	
Cut	Cmp	<i>Dve</i>	TC01741	
	Cux	<i>Ct</i>	TC15699	
	Onecut	<i>Onecut</i>	TC04129	
Lass	Lag 1	<i>Lag1a</i>	TC06052	
		<i>Lag1b</i>	TC06053	
Lim	Islet	<i>Tup</i>	TC09339	
	Lim1/5	<i>Lim1</i>	TC14939	
	Lhx2/9	<i>Ap1</i>	-	
		<i>Ap2</i>	TC03975	
	Lhx6/8	<i>Awh</i>	TC03238	
	Lhx3/4	<i>Lim3</i>	TC14400	
	Lmx	<i>Lmxa</i>	TC01289	
		<i>Lmxb</i>	TC01291	From full length alignment, Lmxa and Lmxb are orthologous to fly genes <i>CG4328</i> and <i>CG32105</i> . Thus Lmx duplicated before the divergence of Coleoptera and Diptera.
POU	Pou2	<i>Pdm</i>	TC07646	
	Pou3	<i>Vvl</i>	TC14350	
	Pou4	<i>Acj6</i>	TC03196	
	Pou6	<i>POU6a</i>	TC06824	
		<i>POU6b</i>	-	
Pros	Pros	<i>Pros</i>	TC10596	
Six	Six1/2	<i>So</i>	TC13834	
	Six3/6	<i>Optix</i>	TC00361	
	Six4/5	<i>Six4</i>	TC03853	
TALE	Irx	<i>Ara</i>	TC03632	
		<i>Mir</i>	TC03634	
	Meis	<i>Hth</i>	TC08633	
	Pbx	<i>Exd</i>	TC11311	
	Prep	<i>Prep</i>	TC06040	
	Tgif	<i>Tgif1</i>	TC09623	

		<i>Tgif2</i>	TC13909	Duplications in beetle and fly are independent based upon full alignment which is consistent with the homeodomain phylogeny.
	CG11617	<i>TcaCg11617</i>	TC13021	No evidence of orthology with the Mohawk family of deuterostomes from synteny; phylogenetic trees do not group this gene with the deuterostome Mohawks robustly, but don't exclude the grouping unequivocally either.
ZF	ZFH1	<i>ZFH1</i>	TC11114	
	ZFH2	<i>ZFH2</i>	TC03891	
Novel	-	<i>BeetleBox1</i>	-	
	-	<i>BeetleBox2</i>	TC15038	

Table S16. Cytochrome P450s in insects by P450 clan *genes(pseudogenes)*

	<i>Apis</i>	<i>Pediculus humanus</i>	<i>Drosophila</i>	<i>Tribolium</i>	<i>Anopheles</i>	<i>Aedes</i>
CYP2 Clan	8(0)	8(0)	7(0)	8(0)	10(-)	11(0)
CYP3 Clan	28(2)	12(0)	36(4)	72(7)	41(-)	80(4)
CYP4 Clan	4(0)	9(0)	32(0)	45(3)	45(-)	58(2)
Mito Clan	6(0)	7(1)	12(0)	9(0)	9(-)	9(0)
Total	46(2)	36(1)	87(4)	134(10)	105(7)	158(6)

Note that the *T. castaneum* and mosquito CYP expansions are species-specific independent events as shown by phylogenetic analysis relative to the last common ancestor.

Table S17. Predicted cysteine proteinases in the *T. castaneum* genome

Members of the C1 cysteine peptidase family, clan CA, are listed with features of the active site and/or critical residues for activity, and a putative identification of functional activity. Sequential Tribolium gene model numbers indicate clustering of the genes, likely due to tandem duplication.

Tribolium Gene Model	Linkage Group	Drosophila Ortholog	Active Site Residues ¹	Critical Residues ²	EST Support ³	Putative Functional Activity
TC01950	10	CG6692-PC	QCHN	n/a	partial	cathepsin L
TC02843 ⁴	3	CG6692-PC	QCHN	n/a	partial	cathepsin L
TC02952	3	CG10992-PA	QCHN	HH	full	cathepsin B
TC02953	3	CG10992-PA	QCHN	HH	partial	cathepsin B
TC02954	3	CG10992-PA	QCHN	KG	full	cathepsin B ⁵
TC02955	3	CG10992-PA	QCHN	HH	full	cathepsin B
TC05431 ⁶	8	CG10992-PA	QCHN	NS	partial	cathepsin B ⁵
TC05432 ⁶	8	CG10992-PA	QCHN	DS	none	cathepsin B ⁵
TC05953	8	CG10992-PA	QCHN	DG	none	cathepsin B ⁵
TC05954	8	CG10992-PA	QSTN	R-	none	cathepsin B ⁵
TC05955	8	CG10992-PA	QCSN	YA	none	cathepsin B ⁵
TC05956	8	CG6692-PC	QCHN	n/a	none	cathepsin L
TC07214	4	CG12163-PB	QCHN	n/a	none	cathepsin O
TC09217	7	CG3074-PB	QSHN	CR	full	cathepsin B ⁵
TC09362	7	CG6692-PC	QCHN	n/a	full	cathepsin L
TC09363	7	CG4847-PA	-CHN	n/a	none	cathepsin L ⁵
TC09364	7	CG6692-PC	QCHN	n/a	none	cathepsin L
TC09365	7	CG6692-PC	QCHN	n/a	full	cathepsin L
TC09448	7	CG6692-PC	QCHN	n/a	full	cathepsin L
TC10999	10	CG6692-PC	QCHN	n/a	one	cathepsin L
TC11000	10	CG4847-PD	QCHN	n/a	full	cathepsin L
TC11001	10	CG6692-PC	QCHN	n/a	full	cathepsin L
TC11002	10	CG6692-PC	ESHN	n/a	full	cathepsin L ⁵
TC11003	10	CG6692-PC	QCHN	n/a	full	cathepsin L
TC13582 ⁴	5	CG5367-PA	QCHN	n/a	none	cathepsin K

¹Conserved diad residues Cys25 and His159 (papain numbering), and Gln19 and Asn/Asp175 (Rawlings and Barrett, 1993).

²Two His residues (His110/111) in the occluding loop region of cathepsin B are critical for activity in cathepsin B proteinases, because they block the C-terminal end of the active site cleft and cause the enzyme to act as a dipeptidase (Musil *et al.*, 1991).

³Park *et al.*¹³⁶

⁴Expression noted with NimbleScan Chip.

⁵These chemical homologs carry polymorphisms in the predicted active site making them unlikely to function as proteases and are possible pseudogenes.

⁶One gene with two splice variants.

Table S18. Identification of sequences used in the phylogenetic analysis, Fig. S13

Abbreviation	Organism	Accession	Location	Predicted Function	Reference
AaCathB	<i>Aedes aegypti</i>	AY626233	Lysosome	Cathepsin B	Isoe et al., unpublished
BtCathB	<i>Bos taurus</i>	P07688	Lysosome	Cathepsin B	Meloun et al., 1988 ¹³⁷
CeCathB	<i>Caenorhabditis elegans</i>	P25807	Larval gut cells	Digestive cathepsin B	Ray and McKerrow, 1992 ¹³⁸
CmCathL	<i>Callosobruchus maculatus</i>	AF544836	Gut	Cathepsin L	Zhu-Salzman et al., 2003 ¹³⁹
DmCathB	<i>Drosophila melanogaster</i>	Q9VY87	Salivary gland	Cathepsin B	Adams et al., 2000 ⁶⁸
DmCathL		Q95029	Embyonic/larval midgut	Fertility, maybe digestive cathepsin L	Tryselius and Hultmark, 1997 ¹⁴⁰
DvCathBa	<i>Diabrotica virgifera</i>	AJ583513	Gut	Digestive cathepsin B	Bown et al., 2004 ¹⁴¹
DvCathBb		AJ583509			
DvCathL		AF190653	Larval midgut	Digestive cathepsin L	Koiwa et al., 2000 ¹⁴²
GlCathB	<i>Giardia lamblia</i>	XP_771222	-	Cathepsin B	McArthur et al., 2000 ¹⁴³
HcCathB	<i>Haemonchus contortus</i>	Z69343	Gut	Digestive cathepsin B	Skuce et al., 1999 ¹⁴⁴
HsCathB	<i>Homo sapiens</i>	P07858	Liver lysosome	Cathepsin B	Chan et al, 1986 ¹⁴⁵
McCathL	<i>Metapenaeus ensis</i>	AY126712	Hepatopancreas	Digestive cathepsin L	Hu and Leung, 2004 ¹⁴⁶
Papain	<i>Carica papaya</i>	P00784	-	Thiol protease	Mitchel et al., 1970 ¹⁴⁷
PcCathL	<i>Phaedon cochleariae</i>	O97397	Gut	Digestive cathepsin L	Girard and Jouanin, 1999 ¹⁴⁸
SjCathB	<i>Schistosoma japonicum</i>	P43157	Intestine (gut)	Digestive cathepsin B	Merckelbach et al., 1994 ¹⁴⁹
SmCathB	<i>Schistosoma mansoni</i>	P25792	Intestine (gut)	Digestive cathepsin B	Klinkert et al., 1989 ¹⁵⁰
SmCathC		Q26563	Lysosome	Digestive cathpesin C	Butler et al., 1995 ¹⁵¹
TcGLEAN#	<i>Tribolium castaneum</i>	Same as glean #	-	-	This paper
TiCathB	<i>Triatoma infestans</i>	AY363262	Digestive tract	Cathepsin B	Kollien et al., unpublished
TiCathL		AY363263		Cathepsin L	
TmCathBa	<i>Tenebrio molitor</i>	DQ356052	Anterior midgut	Cathepsin B	Prabhakar et al., 2007
TmCathBb		DQ356051			
TmCathLa		DQ356055		Cathepsin L	
TmCathLb		DQ356054			
TmCathLc	<i>T. molitor</i>	AY332270	Midgut, hemolymph, fat body, malpighian tubules	Cathepsin L	Cristofolletti et al., 2005 ¹⁵²
TmCathLd		AY332271			
TmCathLe		AY33273			
TmCathLf		AY33272	Migut, hemolymph	Putative digestive cathepsin L	
TmCathLg		AY337517		Digestive cathepsin L	

Table S19. Comparison of the chemoreceptor superfamilies of various insects*

Species	Odorant receptors			Gustatory receptors		
	Genes	Pseudo	Proteins	Genes	Pseudo	Proteins
<i>Drosophila melanogaster</i> ¹⁰⁴	60	2	60	62	0	68
<i>Anopheles gambiae</i> ¹⁰⁵	79	0	79	60	0	90
<i>Aedes aegypti</i>	120	15	105	79	23	88
<i>Bombyx mori</i>	>48	0	>48	63	3	60
<i>Apis mellifera</i>	170	7	163	>60	>50	10
<i>Tribolium castaneum</i>	307	42	265	215	25	220

*Numbers do not add up to 100%, especially for the Grs, because of the alternatively-spliced genes. The numbers of proteins are best estimates of different functional chemoreceptors encoded by these genomes. In *Apis* there are a large number of pseudogenic remnants of Grs of unclear evolutionary origin⁹⁹, and annotation of the *Bombyx* Grs is ongoing. Gene fragments encoding less than 50% of a typical chemoreceptor (roughly 200 amino acids) are excluded.

Table S20. Gene families in *Tribolium* and *Drosophila* involved in cuticle metabolism

Gene family	# genes in <i>Tribolium</i>	# genes in <i>Drosophila</i>	Function	Terminal RNAi phenotype
CDA	9	6	Deacetylation of chitin & chitin oligosaccharides	Ecdysis failure
Lac	3	4	Tanning	White cuticle (Lac 2)
CPs (RR family)	102	101	Structure of cuticle	Unknown
CPs (CPF family)	5	4	Structure of cuticle	Unknown
CPs (CPFL family)	3	7	Structure of cuticle	Unknown

Supplementary Figures List

Figure S1. Developmental stages of *T. castaneum*.

Figure S2. The frequency of GC-content domain lengths in *Tribolium castaneum*.

Figure S3. Comparison of GC-content domains in *Apis mellifera* (green), *Anopheles gambiae* (red), *Drosophila melanogaster* (pink), *Tribolium castaneum* (blue).

Figure S4. *Tribolium* population structure: Correlation between geographic and genetic distance.

Figure S5. Gene Ontology summary of triplet repeat containing proteins relative to the *Tribolium* proteome.

Figure S6. Species phylogeny, based on 1,150 universal single copy orthologs.

Figure S7. Phylogenetic tree of the FGF-receptor family.

Figure S8. The Homeobox Genes of *Tribolium castaneum*

Figure S9. The ANTP Class of Homeobox Genes in *Tribolium castaneum*.

Figure S10. The PRD Class of Homeobox Genes in *Tribolium castaneum*.

Figure S11a,b. The insect P450 gene family (b – fewer species for clarity).

Figure S12. Total number of aspartic, cysteine, and serine peptidase genes found in several insect species.

Figure S13. Phylogenetic analysis of predicted *T. castaneum* cysteine cathepsins and related sequences in other species

Figure S14. A *Tribolium* vasopressin receptor.

Figure S15. *Tribolium* possesses both classes of endocrine/ neuroendocrine-specific prohormone convertases, PC1/3 and PC2, providing the molecular basis for a more complex (neuro)endocrine system.

Figure S16a-c. Phylogenetic tree relating the TcGr proteins to the 10 AmGr, 3 HvGr, and representative DmGr and AgGr.

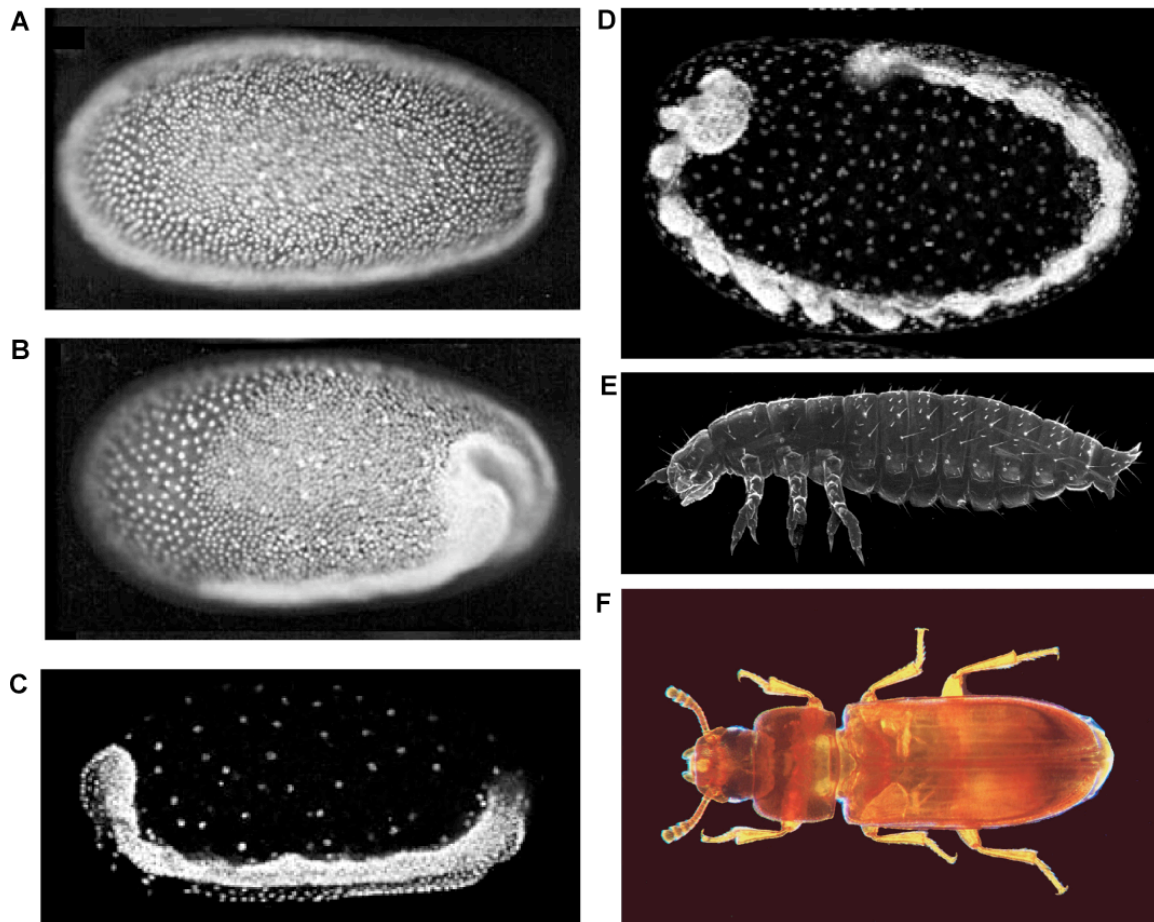


Figure S1. Developmental stages of *T. castaneum*. A, early embryo nuclei staining, prior to formation of germ band growth zone, B, initial germ band formation, C, early germ band with approximately 4 segments developing growth zone is visible at the posterior germ band, D, full germ band extension, E, larvae, F, adult.

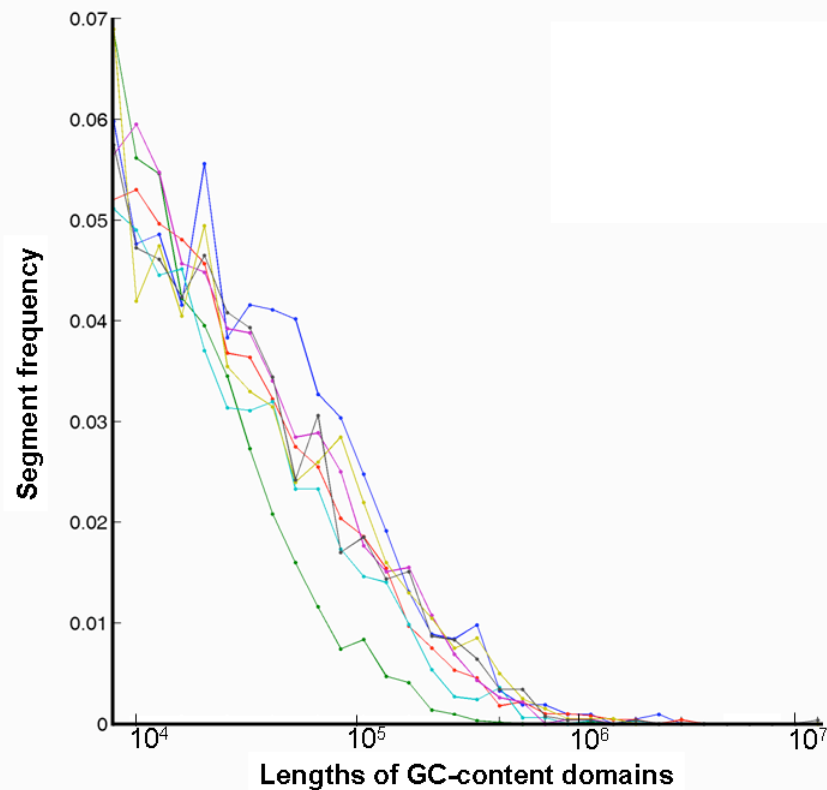


Figure S2. GC-content domain length frequency in sequenced insect species. *Tribolium castaneum* (blue), *Apis mellifera* (green), *Anopheles gambiae* (red), *Drosophila pseudoobscura* (turquoise), *Drosophila melanogaster* (pink), *Drosophila simulans* (yellow), and *Drosophila yakuba* (gray). GC analysis methods are described in the supplementary data.

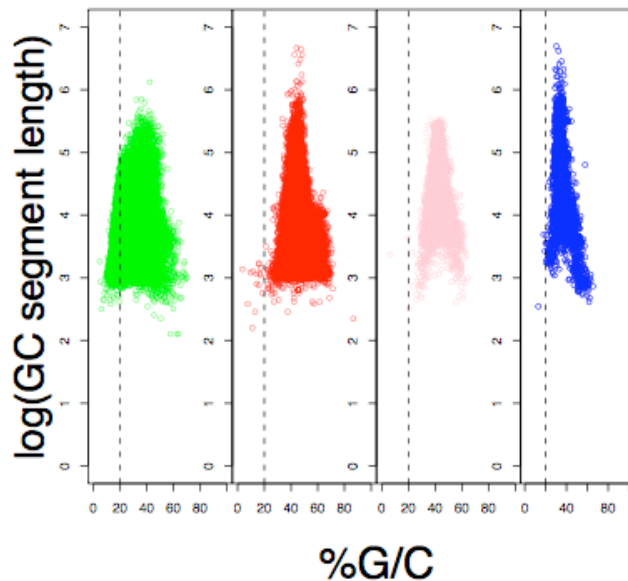


Figure S3, Comparison of GC-content domains in *Apis mellifera* (green), *Anopheles gambiae* (red), *Drosophila melanogaster* (pink), *Tribolium castaneum* (blue). GC-content domain lengths versus GC percentage. Hatched line at 20% shown for comparison. *T. castaneum* domains lack the extremes of GC content present in *A. mellifera*. 0.08% of *T. castaneum* GC-content domains have a GC-composition < 20% (23% for *A. mellifera*) and 99.3% of GC-content domains in the *T. castaneum* genome have a GC content between 20% and 60% (76.7% in *A. mellifera*).

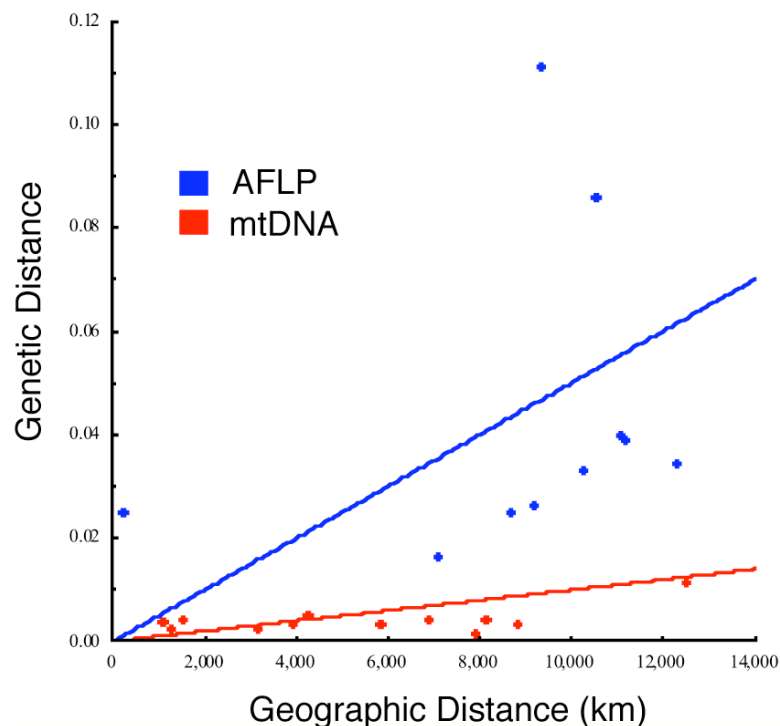


Figure S4. Tribolium population structure: Correlation between geographic and genetic distance. Blue, 133 individuals from 12 populations were genotyped for 434 polymorphic AFLPs ($r^2 = 0.693$, $p = 0.001$, mean $F_{st}=0.133$). Red, 1,423bp mtDNA control region sequenced in 35 individuals from 10 populations (24 polymorphic sites $r^2 = 0.758$, $p < 0.001$) Genetic distances for AFLP data (Nei's D , cf. Lynch and Milligan¹⁵³) and mtDNA sequences (substitutions per site) are corrected for non-independence¹⁵⁴. The intercept was forced to pass through the origin for regression analysis. F_{st} values were computed using AFLP-SURV v1.0¹⁵⁵

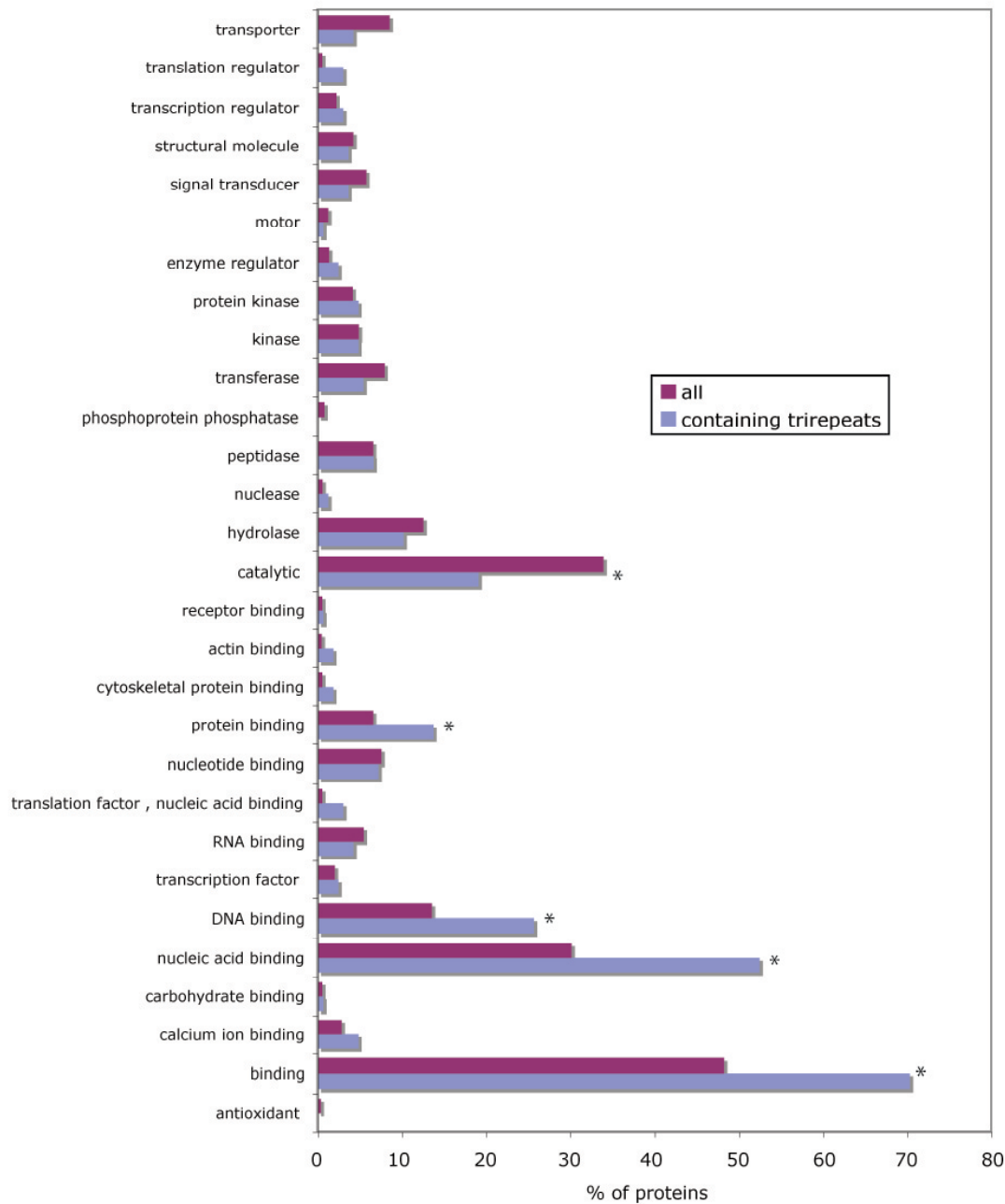


Figure S5. Gene Ontology summary of proteins containing tri-nucleotide repeats relative to the *Tribolium* proteome. For each GOslim category, the percentage of proteins placed in that category

was normalized by dividing it by the total number of proteins that could be matched to any term in the ontology. The values sum to more than 100% because some proteins were placed into multiple categories. Only the Molecular Process ontology is shown as the other ontologies did not contain statistically significant differences.

*Statistically significant differences $FDR < 5\%$ (two-sided Fisher's exact test adjusted for multiple testing).

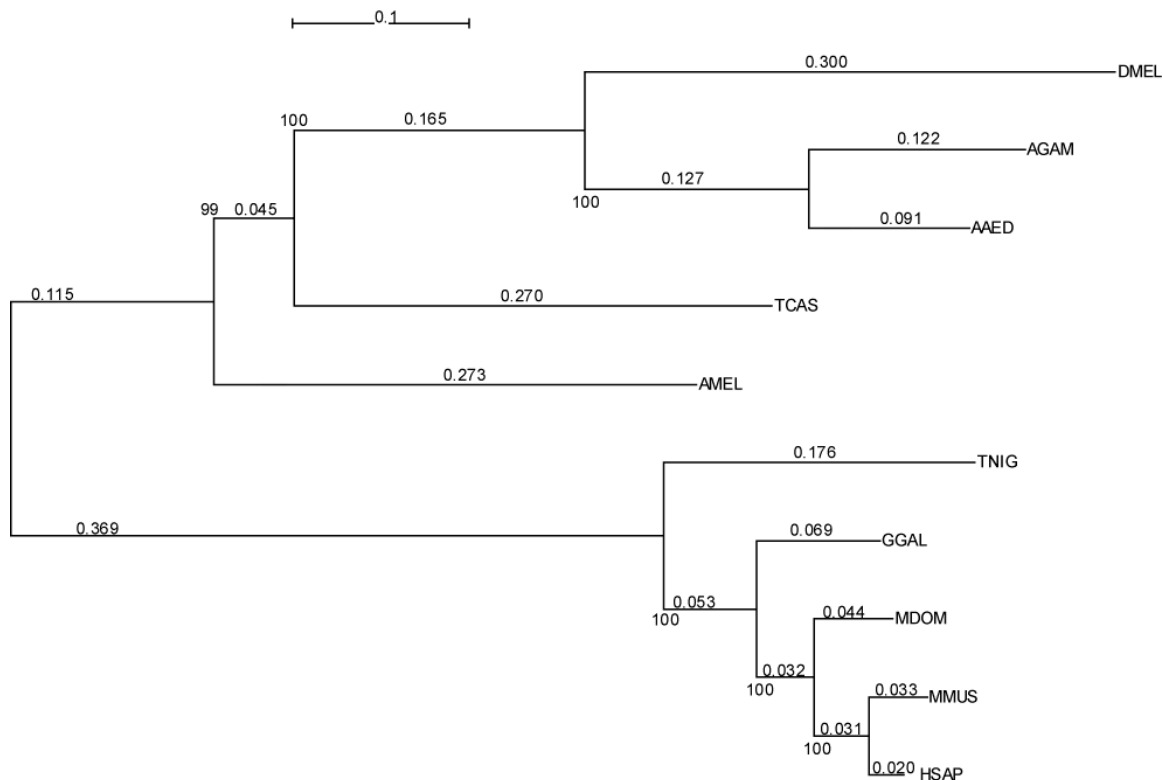


Fig. S6. Expanded species tree from Fig 2. showing scale, measures of the branch length and bootstrap support. It was computed using maximum-likelihood approach on concatenated sequences of 1150 universal single-copy orthologs. It shows accelerated rate of evolution in insects and confirms the basal position of the hymenoptera within the holometabola¹⁵⁶.

Abbreviations: Agam: *Anopheles gambiae*, Aaed: *Aedes aegypti*, Dmel: *Drosophila melanogaster*, Tcas: *Tribolium castaneum*, Amel: *Apis melifera*, Hsap: *Homo sapiens*, Mmus: *Mus musculus*, Mdom: *Monodelphis domestica*, Ggal: *Gallus gallus*, Tnig: *Tetraodon nigroviridis*.

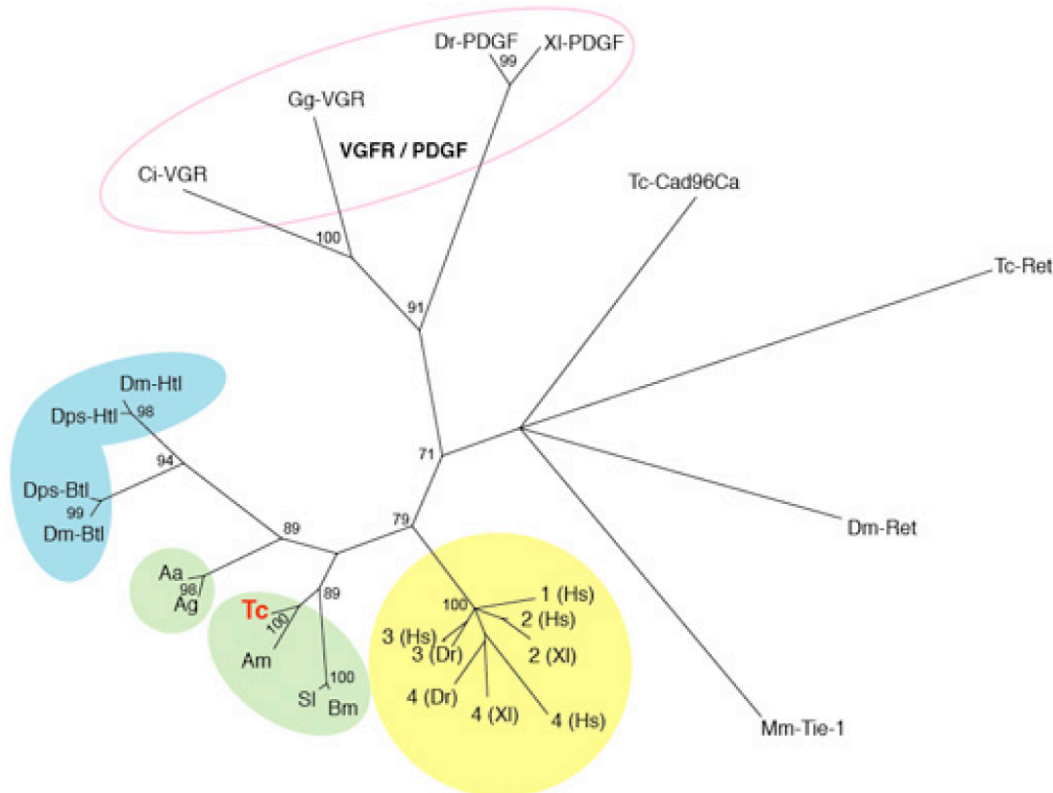
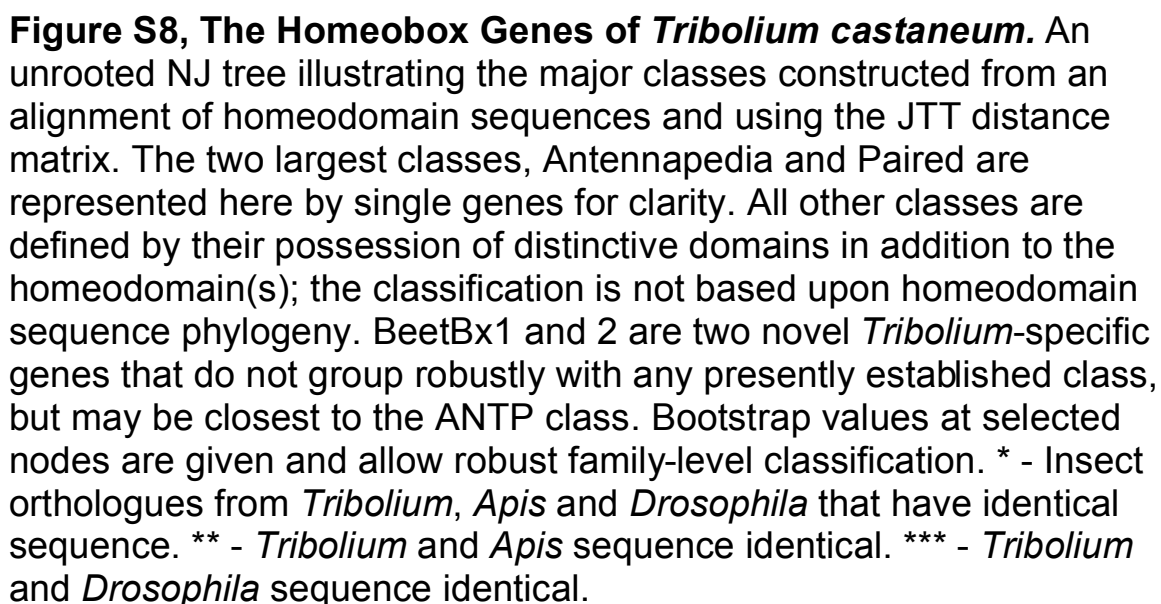
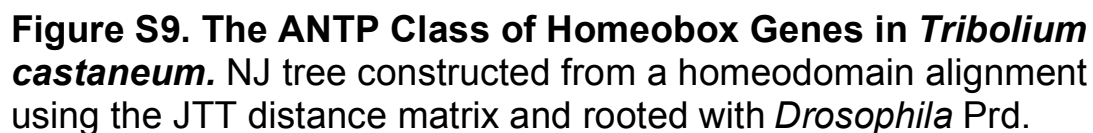


Figure S7. Phylogenetic tree of the FGF-receptor family. Within the insects a duplication of FGF-receptor took place in the line leading to the higher dipterans. The analysis is based on the comparison of the tyrosine kinase domains; tree is generated by neighbour-joining method and quartet sampling, 10 000 puzzling steps (Strimmer, K. and von Haeseler, A. 1996. *Mol. Biol. Evol.* 13: 964-969). The Ret family of tyrosine kinases was used as an outgroup.

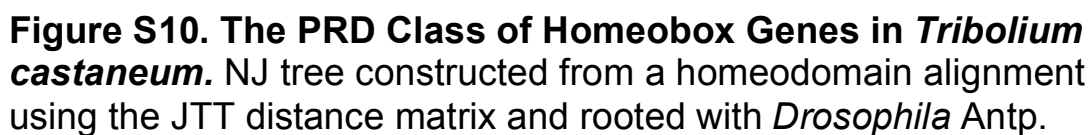
Aa (*Aedes aegyptii*); Ag (*Anopheles gambiae*); Am (*Apis mellifera*); Bm (*Bombyx mori*); Dm (*Drosophila melanogaster*); Dps (*Drosophila pseudoobscura*); Ci (*Ciona intestinalis*); Dr (*Danio rerio*); Gg (*Gallus gallus*); Hs (*Homo sapiens*); Mm (*Mus musculus*); Sl (*Spodoptera littoralis*); Tc (*Tribolium castaneum*); XI (*Xenopus laevis*).

blue: FGFR of higher dipterans, green: insects with only one FGFR in the genome, yellow: vertebrate FGFRs, pink: VGFR (Vascular endothelial growth factor) and PDGF (Platelet derived growth factor) group of tyrosine kinases.





Bootstrap values are given that illustrate robust family-level classification. Note the assignment of insect Not orthologues, which is not well supported in this tree, receives strong support with the use of *Branchiostoma floridae* sequence. In addition, the naming of the families containing CG11085, CG13424 and CG34031 will be presented elsewhere (Butts *et al.* in prep.).



Bootstrap values are given that illustrate robust family-level classification. Two novel insect-specific families are present (represented by *Drosophila* genes CG2819 and CG11294).

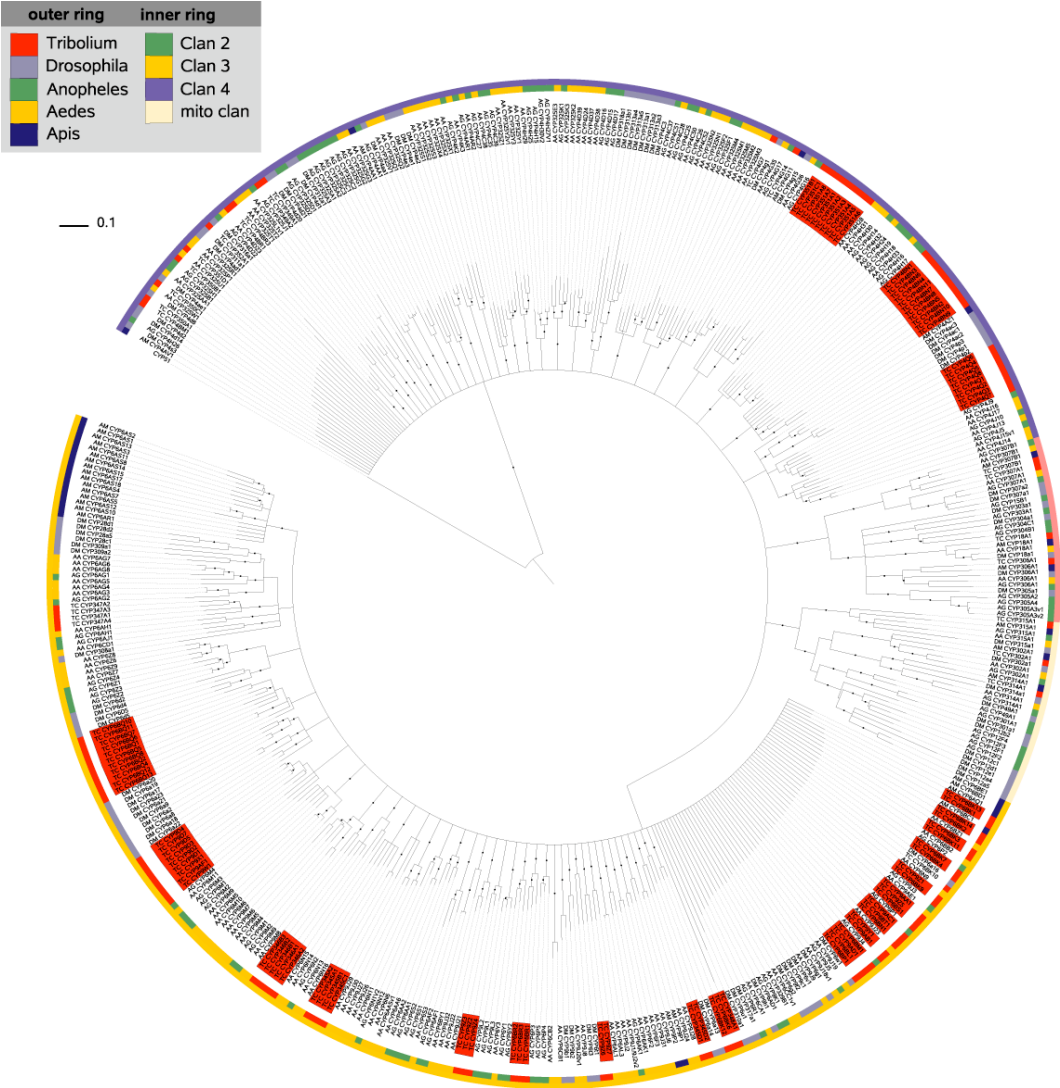


Figure S11. Insect P450 gene family. All orthologs of CYP6s and CYP9 genes annotated in the fruitfly have been subjected to phylogenetic analysis (phymI JTT+G+I,100 bootstraps) that clearly shows Tribolium expansions in both gene families, colored in red. The inner color ring denotes the clan, the outer color ring the species. The tree rooted with sea urchin CYP51 was visualized using iTOL¹⁵⁷. Nodes with bootstraps > 70 are marked with a dot.

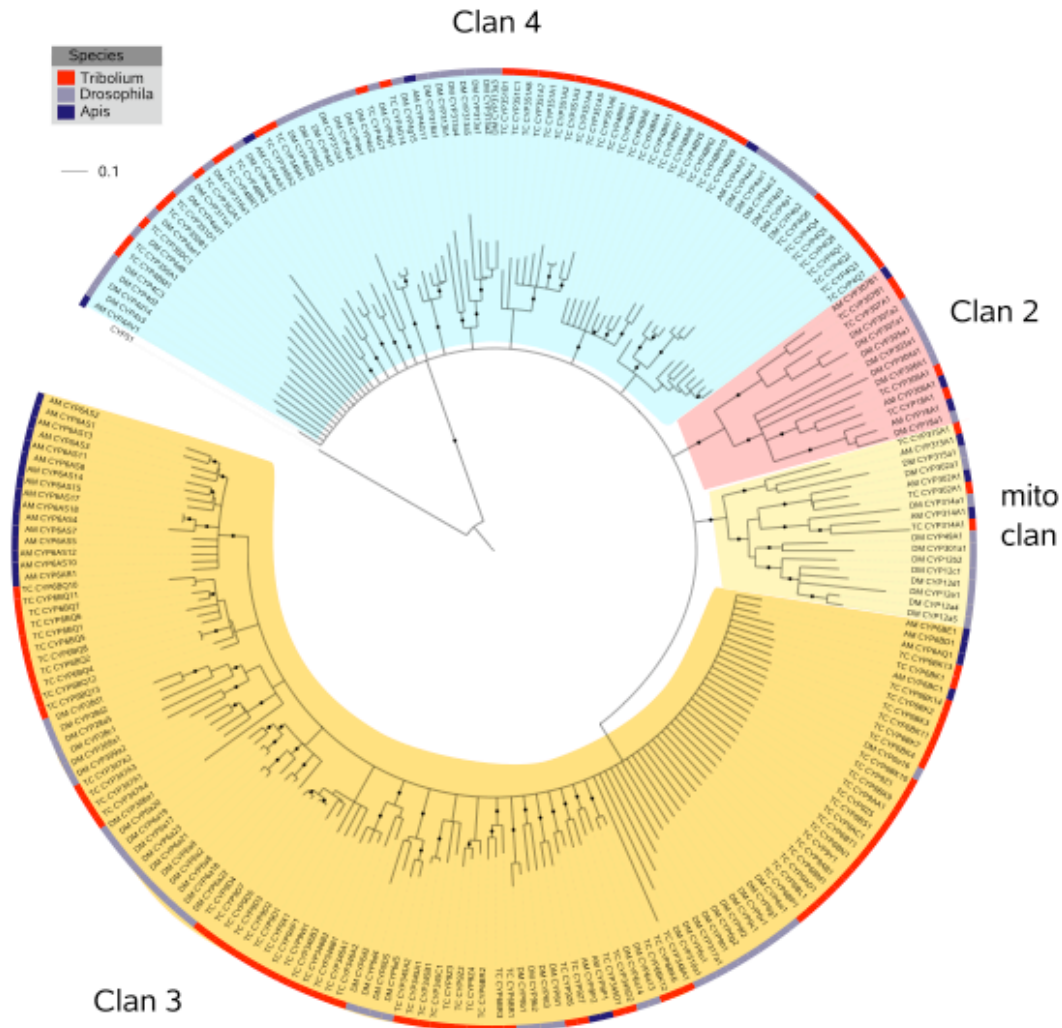


Figure 11b | P450 gene family (fewer species for clarity). CYP genes from *Tribolium* (red), *Drosophila* (grey) and *Apis* (dark blue) have been subjected to phylogenetic analysis (phymI JTT+G+I, 100 bootstraps) that clearly shows *Tribolium* expansions in CYP3 and CYP4 clans, colored in red. The majority-rule tree rooted with sea urchin CYP51 was visualized using iTOL. Nodes with bootstraps > 70% are marked with a dot. Figure S11a shows the same tree with the addition of *Aedes* and *Anopheles*.

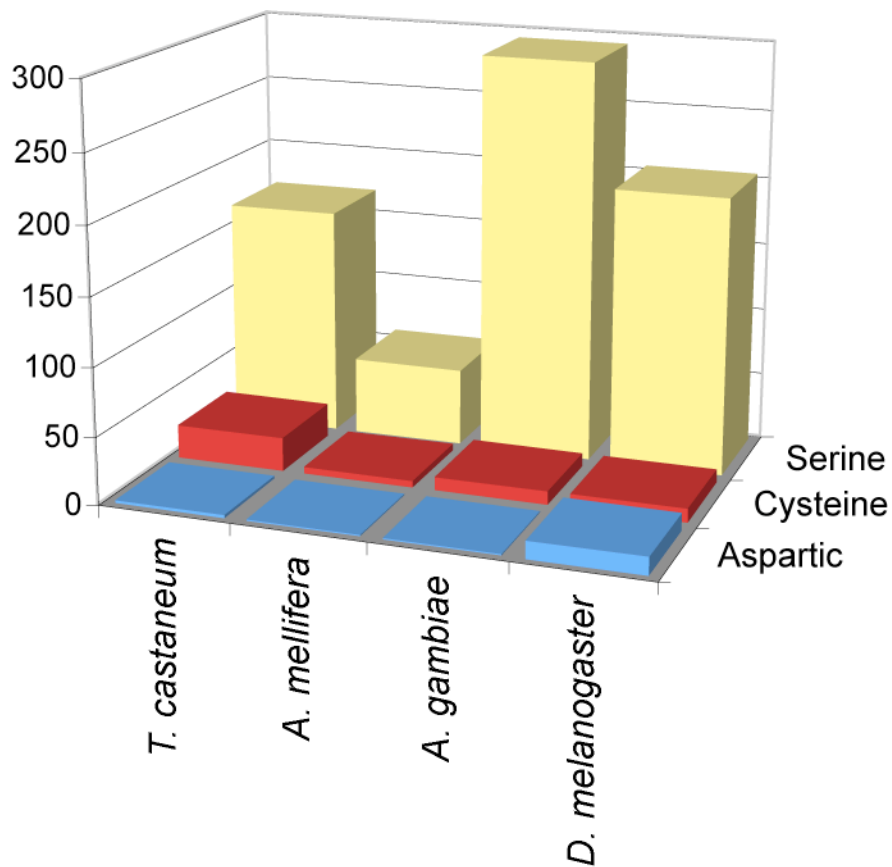


Figure S12. Total number of aspartic, cysteine, and serine peptidase genes found in several insect species.

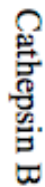


Fig. S13. Phylogenetic analysis of predicted *T. castaneum* cysteine cathepsins and related sequences in other species, as indicated in Table S18. A heuristic search via maximum parsimony was conducted in PAUP (Swofford 2002) with gaps counted as missing data, 10 random taxon addition replicates, and tree-bisection and reconnection (TBR) branch swapping. Sixty

most parsimonious trees were found and the phylogeny represents a strict consensus of those trees (values above braches are from the consensus analysis). Bootstrap analysis was conducted via a fast-sequence addition (10,000 replicates) approach (values below branches). Sequences encoding cathepsins B and L were nested within separate clades, each with reasonably strong parsimony bootstrap support. Outside these clades, cathepsin C (SmCathC), cathepsin O (Tc07214), and cathepsin K (Tc13582) genes were found. Within the cathepsin L clade, three separate clades contained *T. molitor* orthologs and other invertebrate cathepsin L genes (clades 1, 2, and 3). Clade 1 contained a putative digestive cathepsin L (TmCathLf), and the expression patterns of TmCathLa and b suggested that these enzymes may be involved in digestion (Prabhakar et al., in press). Clade 2 of cathepsin L contained invertebrate genes speculated to be involved in protein digestion (DmCathL, MeCathL, DvCathL, and PcCathL), and the lower clade contained the gene encoding digestive cathepsin L from *T. molitor* (TmCathLg, Cristofolleti et al., 2005). Therefore, this entire clade may consist of digestive cathepsin L enzymes. Within the cathepsin B clade, clade 4 consisted of enzymes from vertebrates and invertebrates, but clade 5 contained all (except Tc05956) cathepsin B homologs, and all were clustered on linkage group 8. Experimental data are needed to confirm the function and location of these gene products.

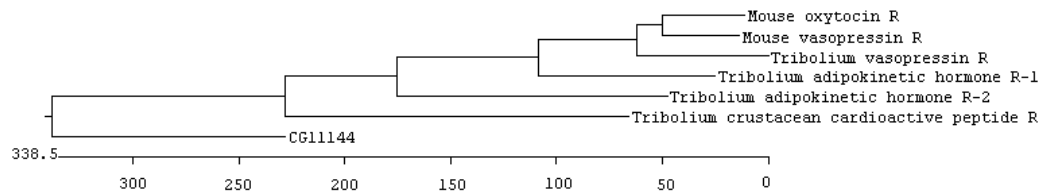


Figure S14. A *Tribolium* vasopressin receptor. Phylogenetic tree analysis of the protein encoded by TC16363. This neuropeptide GPCR belongs to a cluster of closely related *Tribolium* neuropeptide GPCRs consisting of two adipokinetic hormone (AKH) and one crustacean cardioactive peptide (CCAP) receptor. However, the TC16363 receptor is more closely related to mammalian vasopressin (V1) and oxytocin receptors than to its most closely related *Tribolium* AKH and CCAP receptors, indicating that it is a vasopressin receptor. This is supported by our finding of a vasopressin peptide (structure CLITNCPRGamide) in *Tribolium* encoded by the gene TC06626. This is the first time that a vasopressin receptor has been identified in arthropods and the first time that that a vasopressin peptide has been found in a holometabolous insect.

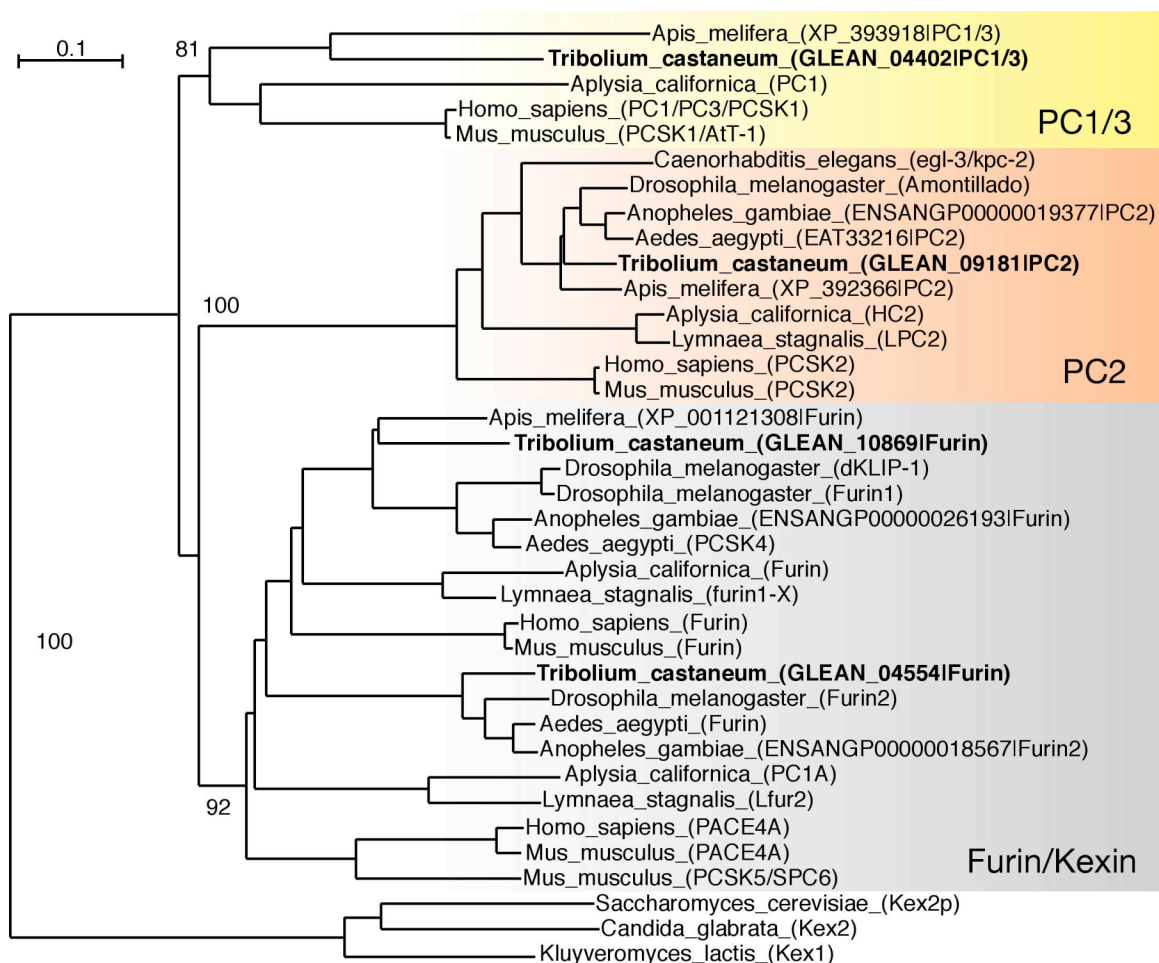


Figure S15. *Tribolium* possesses both classes of endocrine/ neuroendocrine-specific prohormone convertases, PC1/3 and PC2, providing the molecular basis for a more complex (neuro)endocrine system. Phylogenetic analysis with other kex2/subtilisin-like proteases, bootstrap support (in %) indicated for major branches.

Figure S16 a-c. Phylogenetic tree relating the TcGr proteins to the 10 AmGr_s, 3 HvCr_s, and representative DmGr_s and AgGr_s. Gr_s from particular species are colored as are the branches that lead to them, purple for TcGr_s, red for AmGr_s, green for HvCr_s, blue for AgGr_s, and cyan for DmGr_s. The tree is rooted at the midpoint, with Fig. S16a being the basal third including many highly-divergent lineages in all insects, Fig. S16b the middle third of the tree including the heterodimeric carbon dioxide receptor, the candidate sugar receptors and the Tc214 proteins, and Fig. S16c an entirely *Tribolium*-specific expansion at the top of the tree. It is a corrected distance tree built as in Robertson and Wanner⁹⁹, with bootstrap support from 1000 replications of uncorrected distance analysis indicated for major branches. Major lineages mentioned in the Supplementary Information are indicated by vertical lines on the right, as are the locations of most clusters of TcGr_s on linkage groups (LG) and rough position in Mbp according to the NCBI MapViewer. Suffixes after protein names indicate details of partial gene models (PAR – usually resulting from inter-contig gaps), pseudogenes (PSE – involving various problems like in-frame stop codons or frameshifting insertions or deletions in otherwise alignable exons), and corrected gene models using information from the Trace Archive (FIX – commonly involving extensions into inter-contig gaps). For the phylogenetic tree, three sequences that cause extremely long branch length problems were removed, specifically Gr169PSE which is similar to Gr170, and is missing only the C-terminus, and Gr192PSE and Gr194PSE which have their entire C-terminus missing, but are otherwise similar to Gr190-197.

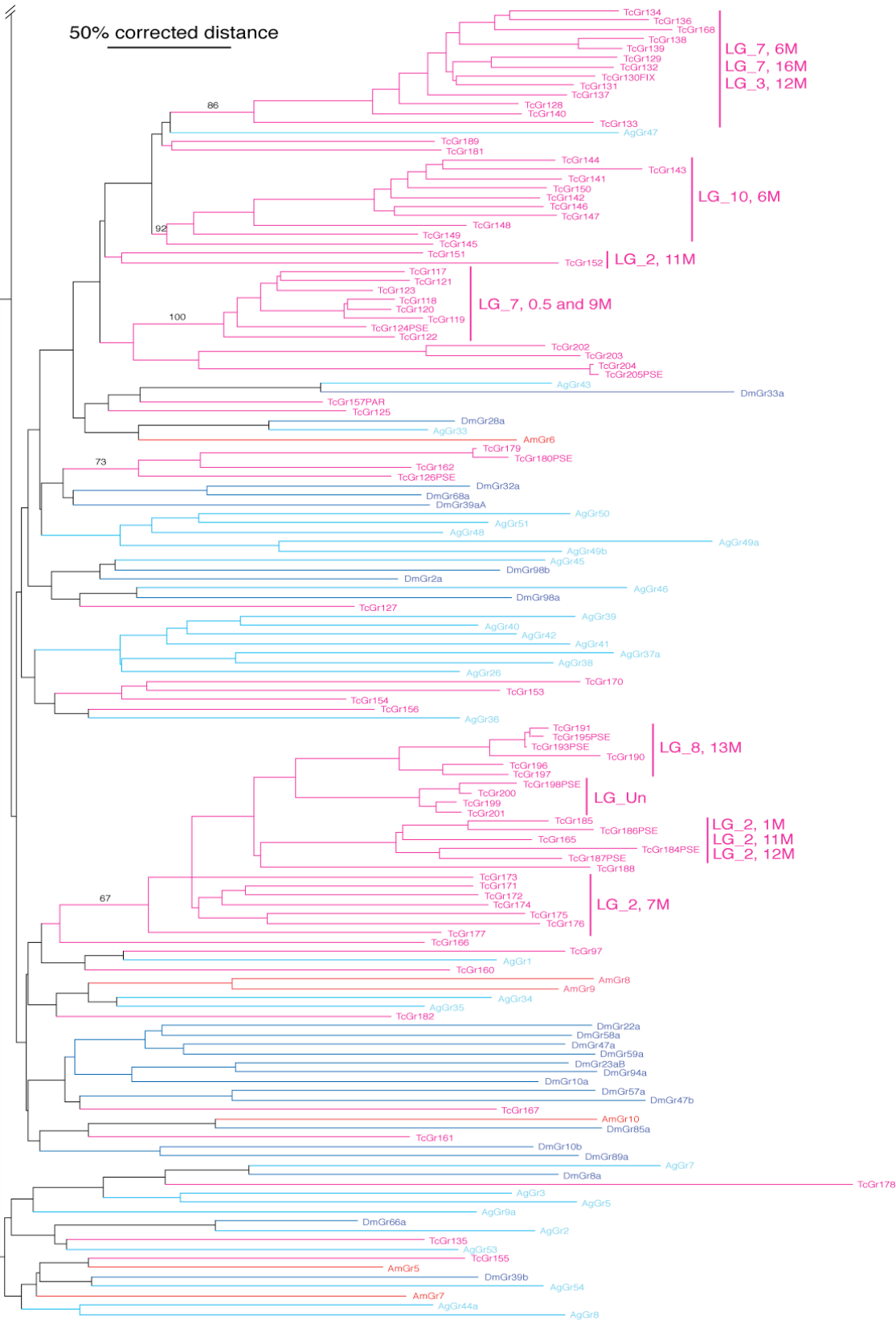


Fig. S16a.

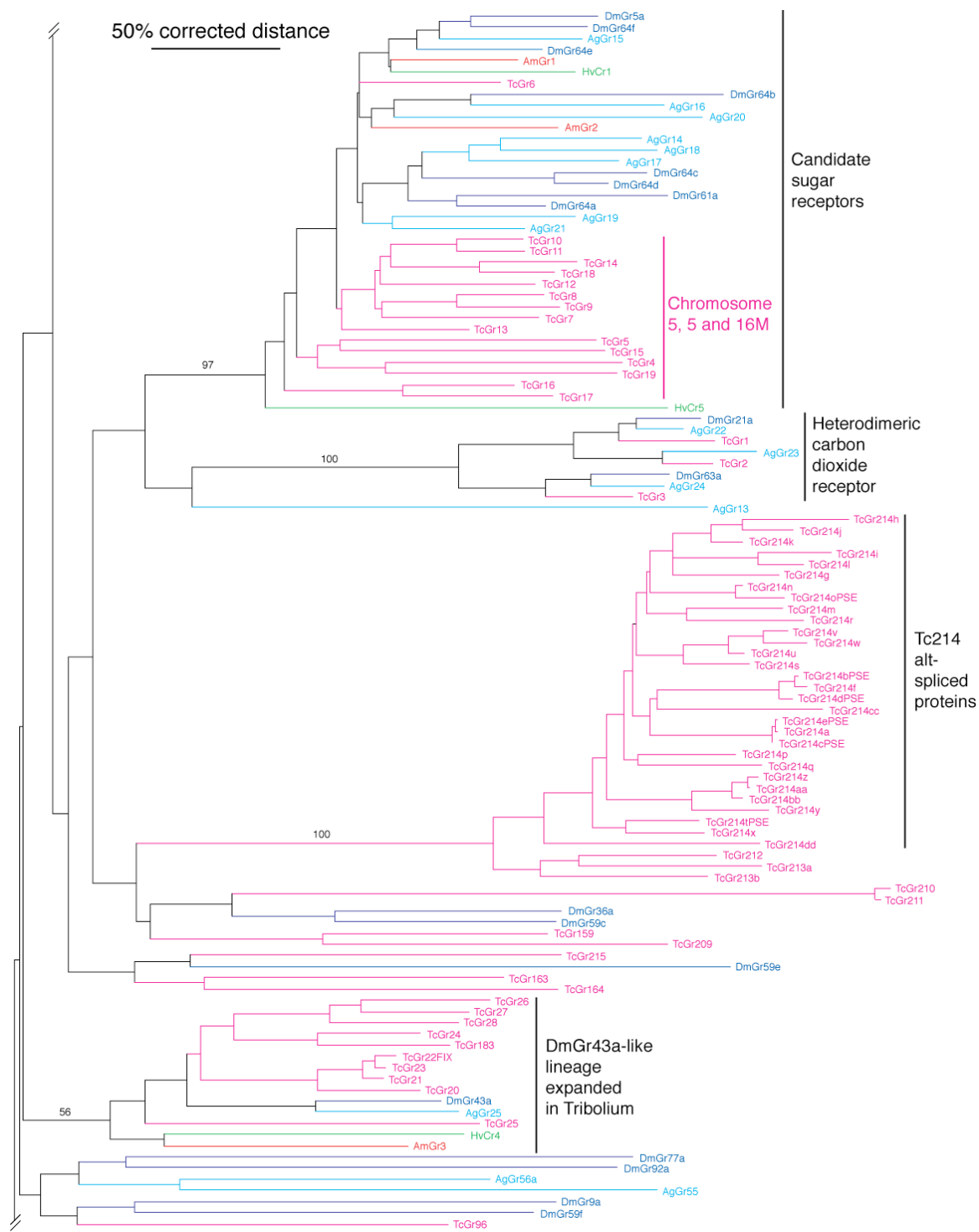


Fig. S16b.

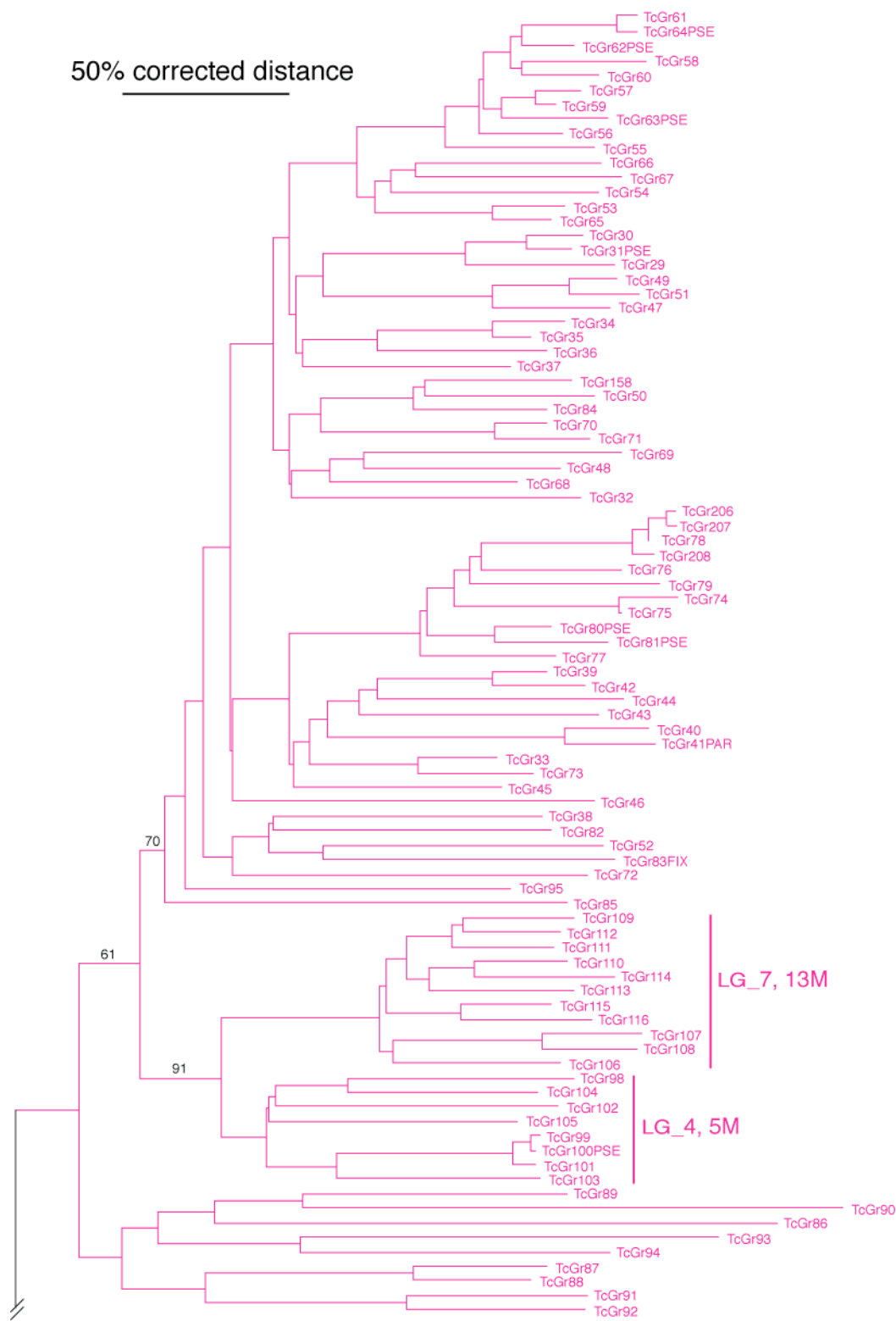


Fig. S16c.