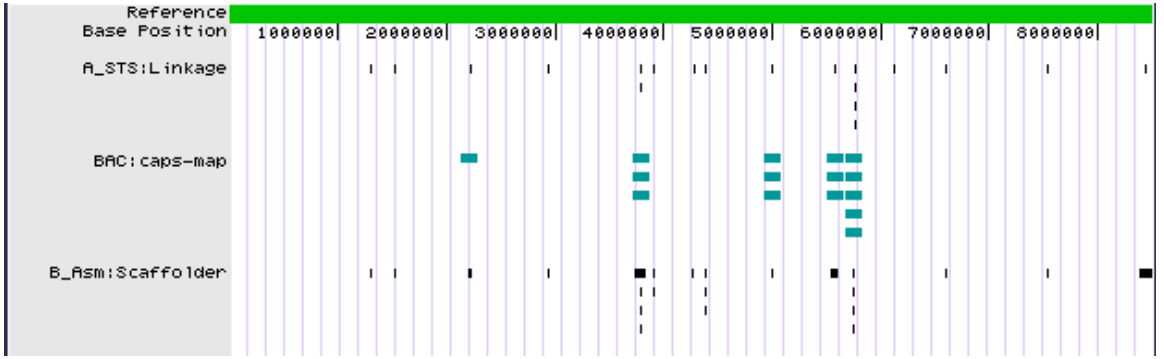
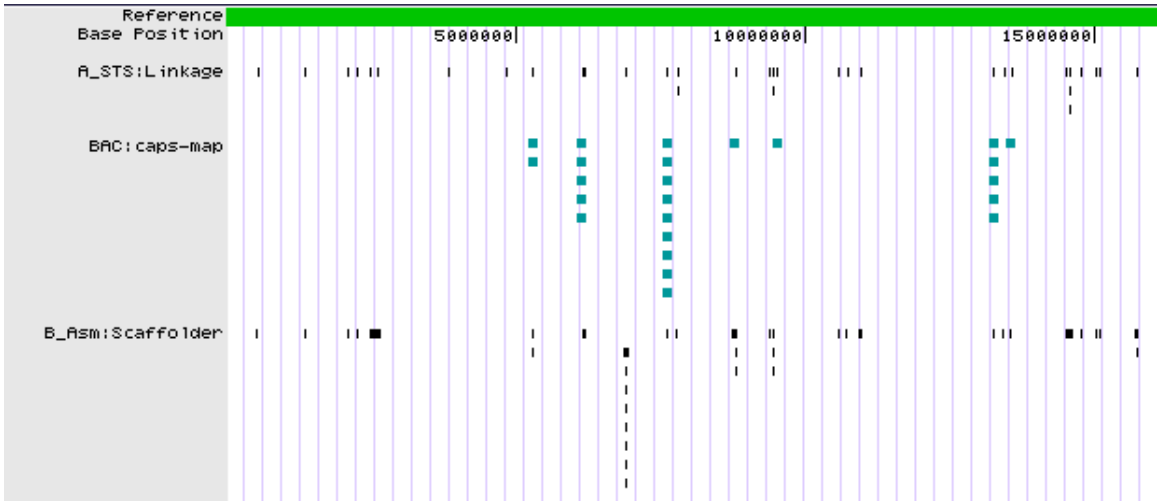
# Figure S1. Non-random distribution of BAC clones throughout the *Apis* genome

BACs representing several-fold clone coverage of the genome were sequenced to low coverage and then mapped to the genome based on microsatellite marker position. A portion of the genome (two of the linkage groups) is shown as displayed in the Genboree browser (www.genboree.org). The top track (A_STS:Linkage) shows the position of markers (at this stage of map development there were 854 mapped markers). The middle track (BAC:CAPS_MAP) shows the position of BACs mapping to this region. Note the 'pile-ups' of BACs over some markers and the absence of BACs over others, indicating a non-uniform distribution. The assembly of WGS reads (v1.0) was also mapped to the genome and is shown in the bottom track (B_Asm:Scaffolder). It is evident that the WGS assembly covers nearly all the markers in a much more uniform manner than the BACs.

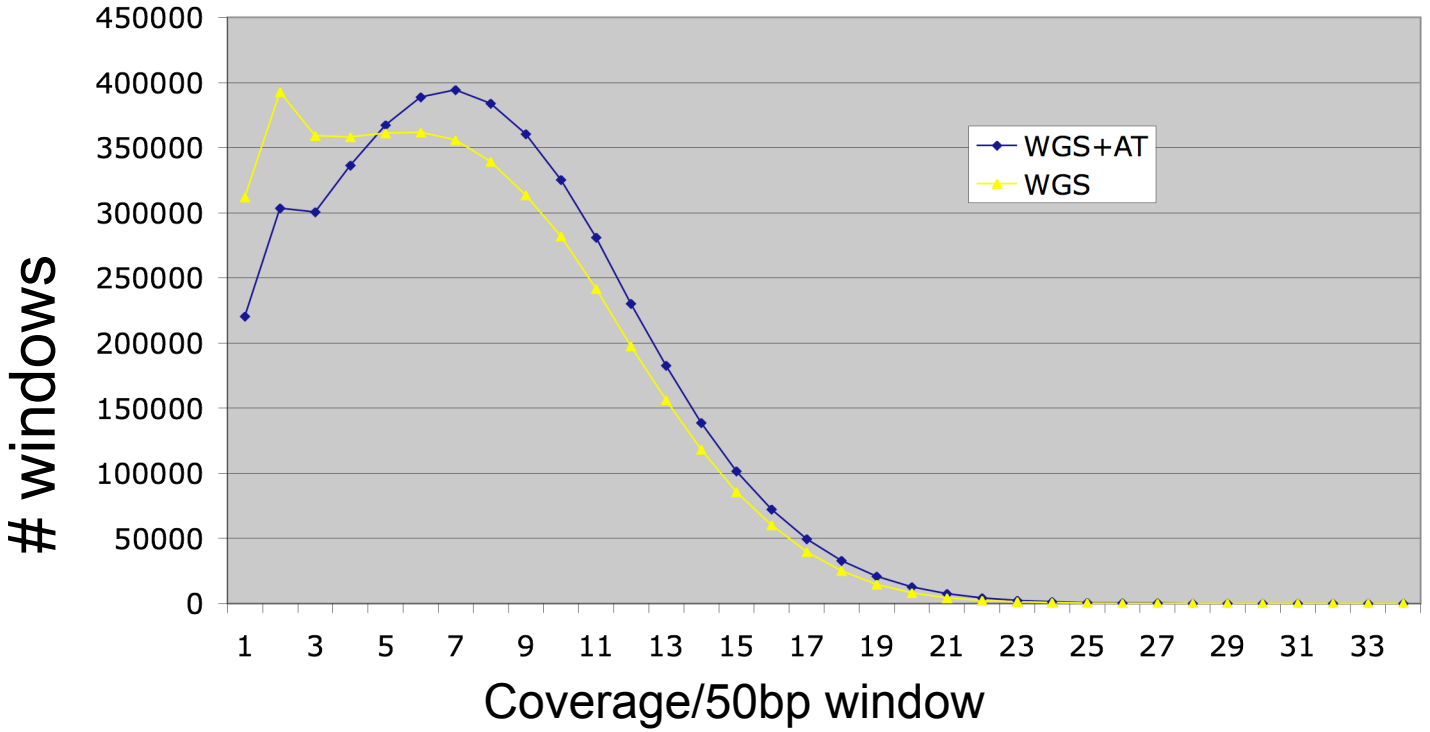
Linkage group 11




Linkage group 14



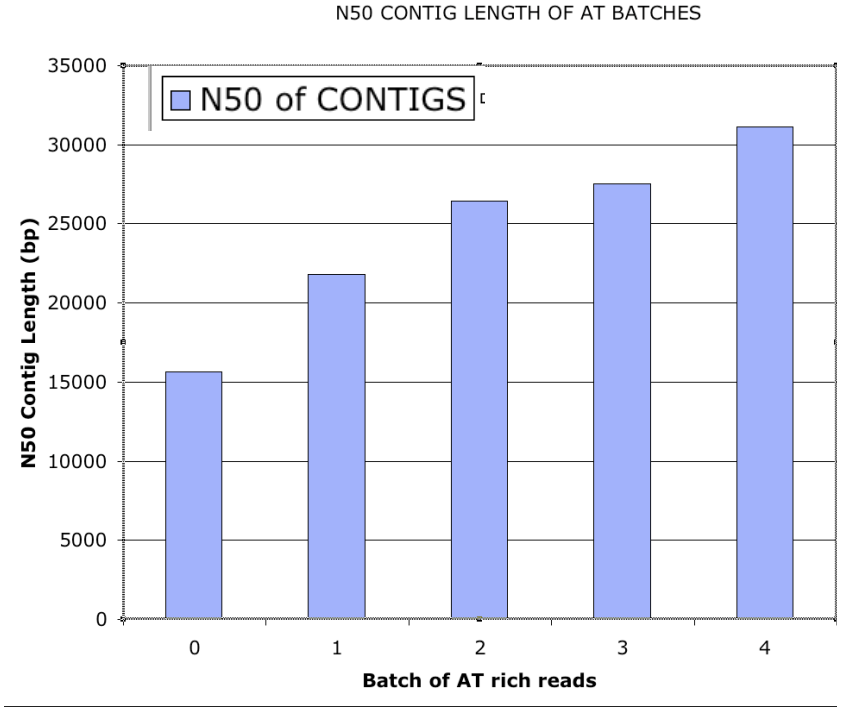
# Figure S2. Coverage distribution per 50bp windows



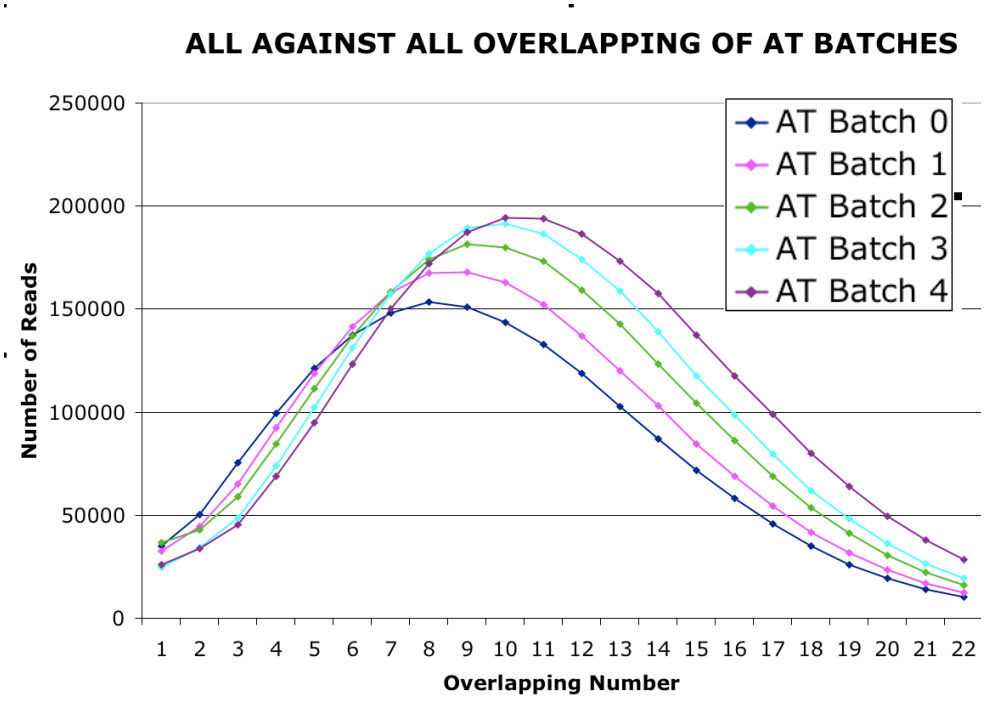

The sequence coverage was computed for 50 base wide windows over the whole genome for assembly v1.2 (yellow) and v2.0 (blue, after addition of AT-rich reads).

# Figure S3. Improvement of the *Apis mellifera* genome assembly by addition of AT-rich reads.

## A. Contig size



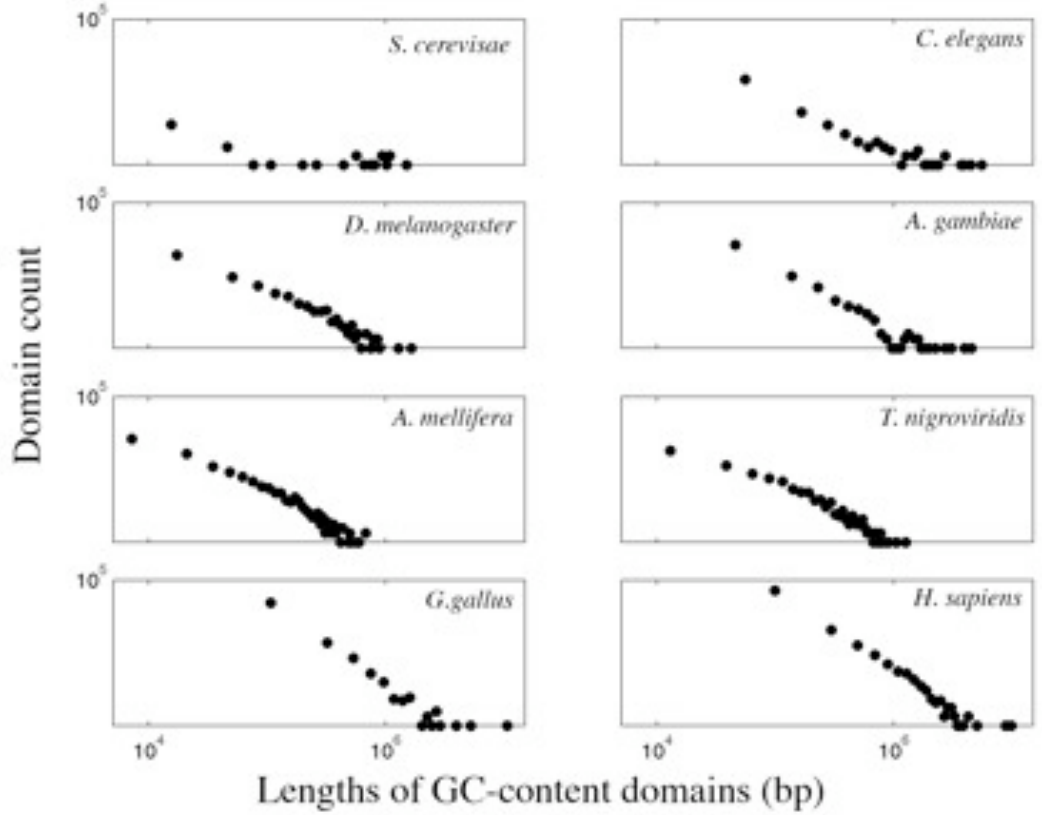

## B. Coverage




For assemblies after each of the four batches of 200,000 AT-rich reads, the N50 of the contigs (A) and the coverage (B, expressed as the overlap number of reads) is shown.

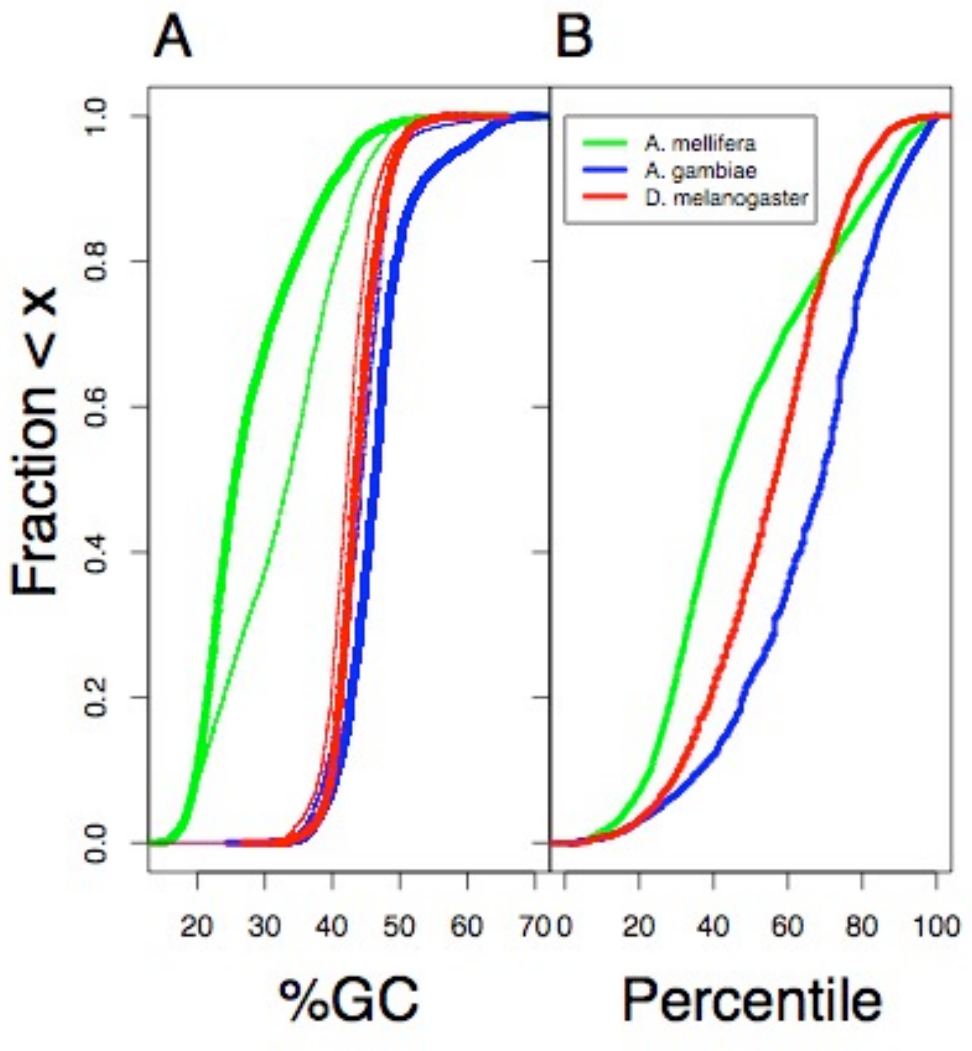
# Figure S4. GC content domains

Length distributions of GC-content domains in eukaryotes plotted on a log-log scale. Domains shorter than 10 kb are not shown. In contrast to the fungal genome, *S. cerevisiae*, the metazoan genomes show power law distributions of GC-content domain lengths, with many short domains and few large domains.



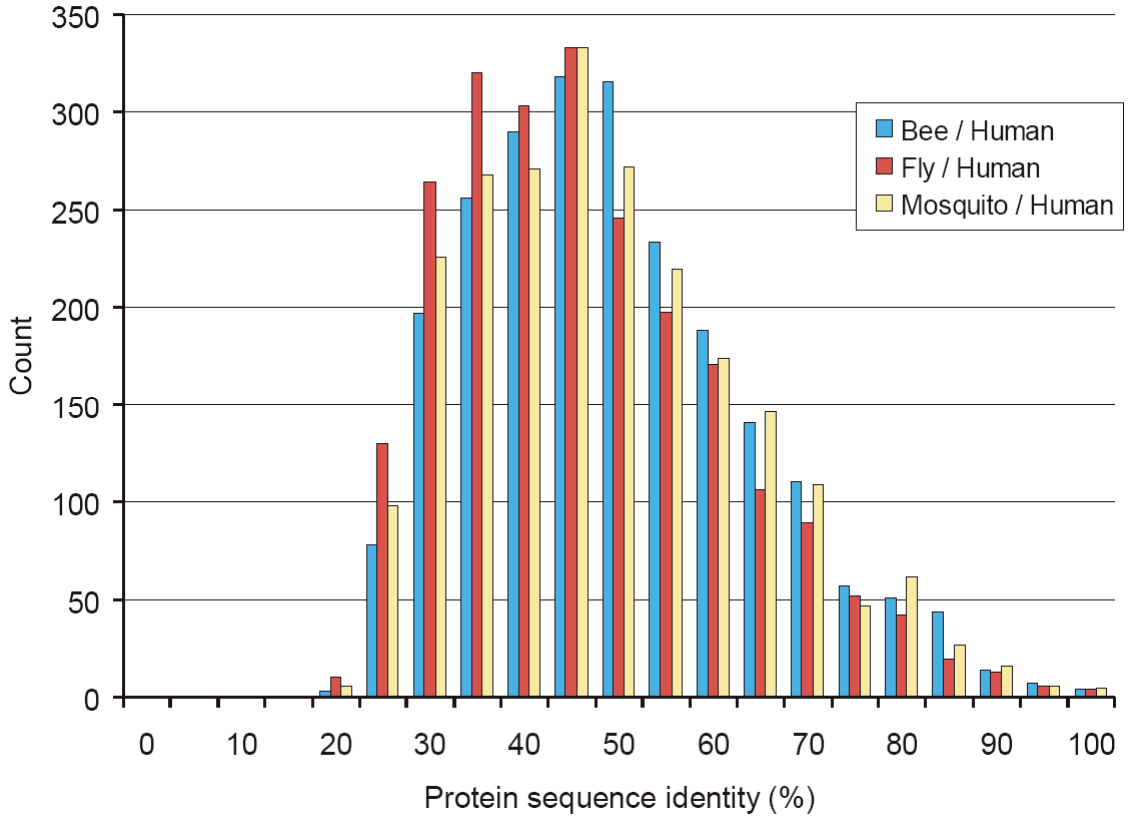
# Figure S5. Cumulative distribution of genes (thick lines) and nucleotides (thin lines) vs. GC content.

A) Cumulative distribution of genes (thick lines) and nucleotides (thin lines) plotted against %GC of GC-content domains in which the genes (nucleotides) are embedded for *A. mellifera* (green), *A. gambiae* (blue) and *D. melanogaster* (red). These plots indicate the low GC content of *A. mellifera* genes and genome compared to the other insects. The difference between the *A. mellifera* gene and nucleotide plots indicates the preference of genes for lower GC content regions of the genome, while the distribution of genes in *A. gambiae* and *D. melanogaster* more closely resembles genome composition. B) Cumulative distribution of genes plotted against GC-content percentile of GC-content domains in which the genes are embedded for *A. mellifera*, *A. gambiae* and *D. melanogaster*. This plot also indicates the tendency of *A. mellifera* genes to lower GC-content regions of the genome.



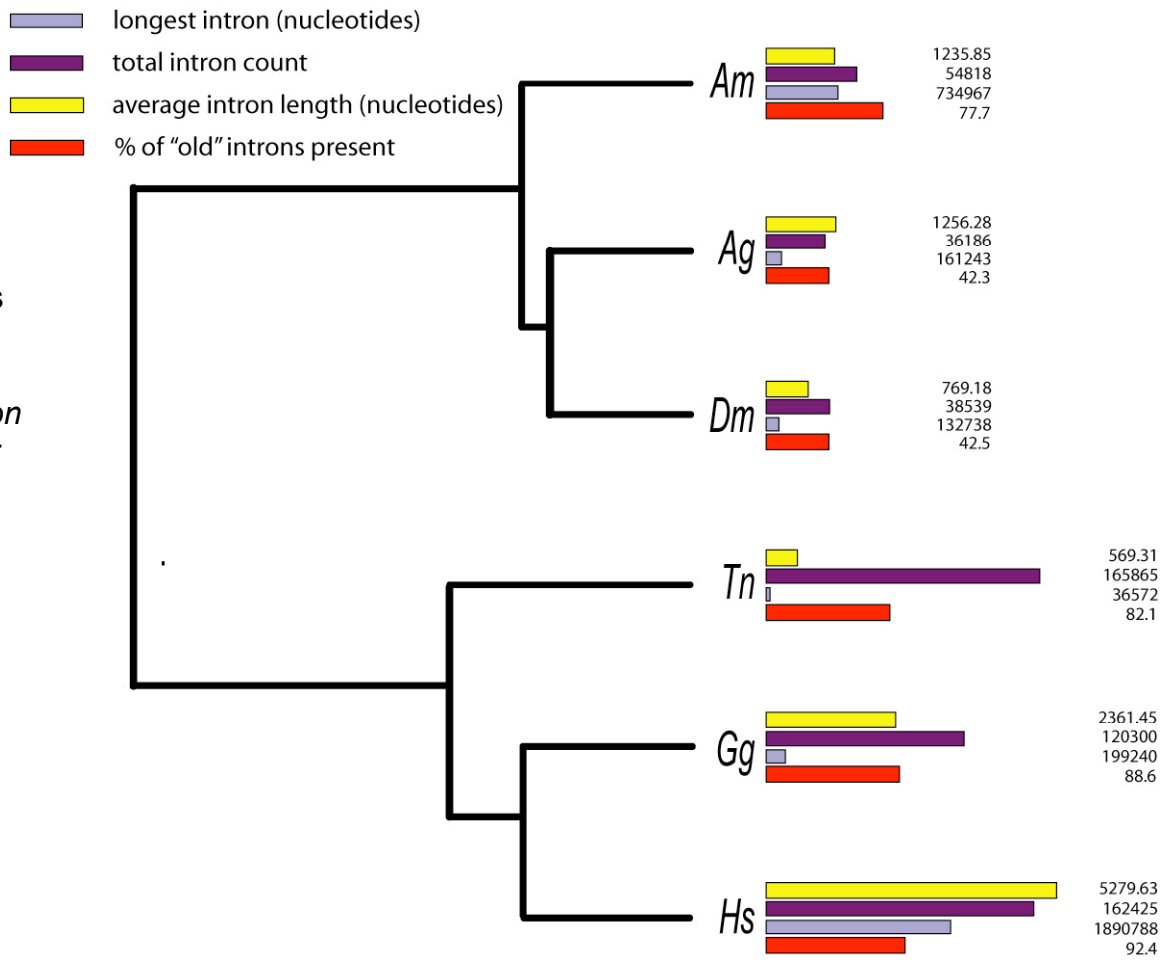
# Figure S6. Distribution of sequence identity between single-copy orthologs.

The histograms show sequence identity distributions of bee, fly and mosquito proteins in comparison to human orthologs.



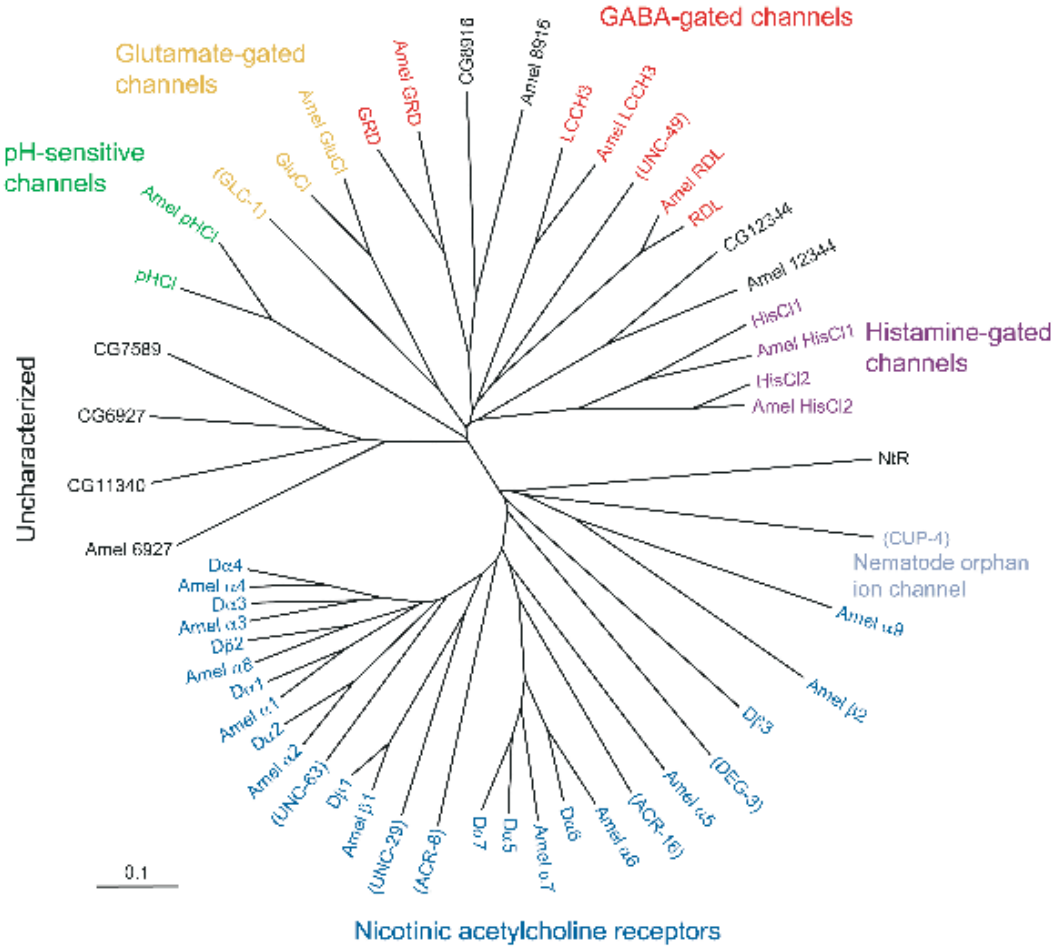
# Figure S7. Comparative introns

The fraction of retained “old” introns was calculated using 4,441 orthologous groups that have at least one representative in each of the six Metazoan analyzed (*Homo sapiens* (Hs), *Gallus gallus* (Gg), *Tetraodon nigroviridis* (Tn), *Drosophila melanogaster* (Dm), *Anopheles gambiae* (Ag), *Apis mellifera* (Am)). We identified 15,560 ancient introns by positional conservation between at least one vertebrate and one insect gene. In case of multiple co-orthologues we considered the longest ones.



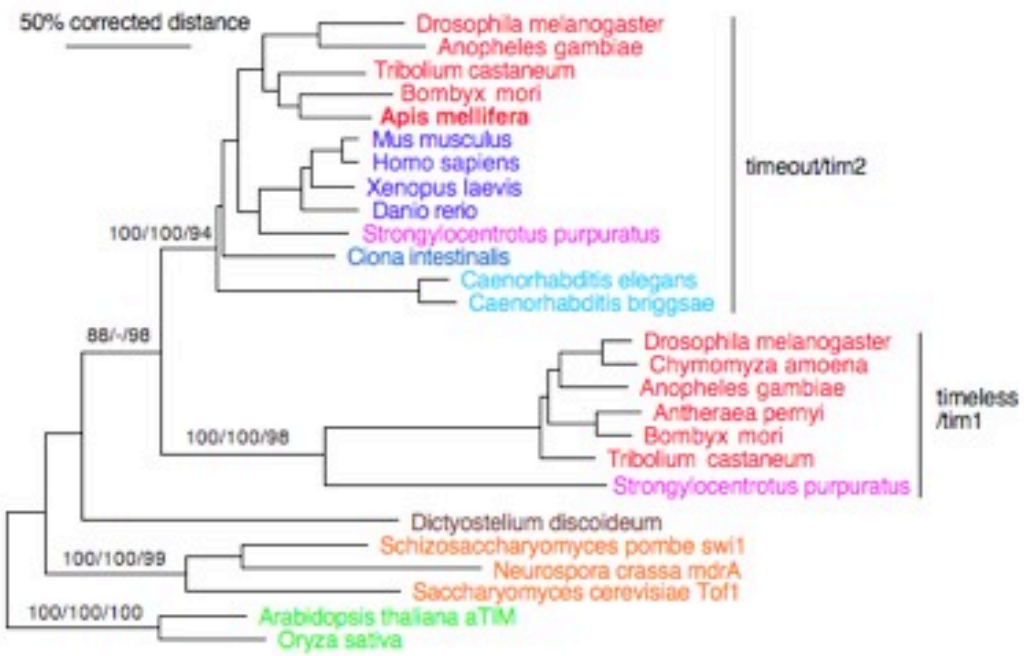
# Figure S8. Honey Bee Cys-loop ligand-gated ion channels

The accession numbers of the *Drosophila* sequences used in constructing the tree are: Da1 (CAA30172), Da2 (CAA36517), D a3 (CAA75688), Da4 (CAB77445), Da5 (AAM13390), Da6 (AAM13392), Da7 (AAK67257), Db1 (CAA27641), Db2 (CAA39211), Db3 (CAC48166), GluCl (AAG40735), GRD (Q24352), HisCl1 (AAL74413), HisCl2 (AAL74414), LCCH3 (AAB27090), Ntr (NP_651958), pHCl (NP_001034025), RDL (AAA28556). The GB identifiers for the honeybee sequences are: Ama1 (GB17133), Ama2 (GB18518), Ama3 (GB10583), Ama4 (GB19836), Ama5 (GB14283), Ama6 (GB17000), Ama7 (GB19257), Ama8 (GB15196), Ama9 (GB16984), Amb1 (GB17819), Amb2 (GB12006), AmGluCl (GB11639), AmGRD (GB11033), AmHisCl1 (GB19505), AmHisCl2 (GB15968), AmLCCH3 (GB12078), AmpHCl (GB11444), AmRdl (GB14080), AmCG7589 (GB11903), AmGC8916 (GB10798), AmCG12344 (GB18933).



# Figure S9. Phylogeny of timeless/timeout proteins in animals

Phylogenetic relationships of timeless/Tim1 and timeout/Tim2 proteins in animals are shown in a corrected distance tree. The animal proteins are rooted with the single orthologs in plants (green), yeasts (orange), and *Dictyostelium discoideum* (brown). Insect lineages are in red, and show that *Apis mellifera* has lost the ortholog of timeless in *Drosophila* and other insects. The timeless lineage was also lost from chordates (blue), but its antiquity is shown by the presence of an orthologue in the sea urchin, *Strongylocentrotus purpuratus* (purple). Numbers on major branches are percentage presence in 1000 replications of distance and parsimony bootstrap analysis, and 10,000 maximum likelihood quartet puzzling steps. Distances were corrected as described in references below.





Rubin, E. et al. Molecular and phylogenetic analyses reveal mammalian-like clockwork in the honey bee (*Apis mellifera*) and shed new light on the molecular evolution of the circadian clock. Genome Res (in press) (2006).

Velarde, R. A., Sauer, C. D., KK, O. W., Fahrbach, S. E. & Robertson, H. M. Pteropsin: A vertebrate-like non-visual opsin expressed in the honey bee brain. Insect Biochem Mol Biol 35, 1367-77 (2005).

Robertson, H. M. & Gordon, K. H. J. Canonical TTAGG repeat telomeres and telomerase in the honey bee, Apis mellifera. Genome Res (in press), (2006).

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).

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. (in press) (2006).

# Supplementary Notes and Methods

## Construction of draft genome assemblies

**BAC library construction.** Agarose-embedded genomic DNA prepared from *Apis mellifera* strain DH4 partially treated with *Eco*RI and *Eco*RI Methylase, was size-selected using pulsed-field gel electrophoresis, ligated into pTARBAC2.1 vector between *Eco*RI sites, and transformed into DH10B (T1 phage-resistant) electro-competent cells (Invitrogen) for the CHORI-224 BAC library. The library was arrayed in 384-well format.

**Sequencing strategy and assembly.** The original strategy approved by the NIH-NHGRI aimed to draw on the success of the combined whole genome shotgun (WGS) and BAC approach used in the rat genome project[1], and assemble this data into a high coverage draft sequence with the Atlas assembler[2]. However, the BAC library showed characteristics that made it unsuitable for this approach. BAC clones showed an insert size distribution that indicated significant deletion or instability, as there was a pronounced skew towards smaller inserts (library CHORI-224, size distribution shown at bacpac.chori.org/bee224.htm). This was not seen in BAC libraries of other organisms made by this method (as shown at bacpac.chori.org) but was conspicuously present in with smaller inserts showed the symmetrical size distribution normally found for large insert libraries (library CHORI-1224; bacpac.chori.org/bee1224.htm). When reads from sequenced BACs were co-assembled with WGS reads, and the result compared to a pure WGS assembly, the scaffold size was smaller indicating the BAC data was detrimental to assembly. Anecdotal observations suggested that deletions within BACs produced novel joints whose read pairs were inconsistent with the WGS read pairs and this led to limits on the length of the scaffolds that could be built. A small minority of novel read pairs could have a significant effect. We also observed that the BAC sequences did not cover the genome uniformly (Supplemental Figure S1) while WGS reads did span the genome appropriately. Thus there appeared to be biases in the BAC data set and consequently it was used sparingly. Some sequences produced from pooled BAC arrays[3] (CHORI-224 library) were included in the read set, omitting conflicting reads.

The project thus became a pure WGS approach, using small insert clones, fosmid end sequences (which did not show biased distribution), and limited BAC sequences. The Atlas assembler was used as this had successfully assembled the pure WGS data for the *Drosophila pseudoobscura* project[4]. Analysis of the initial assemblies (v1.0 through v1.2, Supplementary Table S1) showed that some AT-rich sequences had low coverage (Supplementary Figure S2), and notably this occurred in genic regions, preventing complete definition of gene content. To address this, we prepared additional shotgun libraries from AT-rich DNA prepared from density gradients[5]. We 'titrated' the amount of AT-rich sequence to add by producing increments of 200,000 reads from this library and reassembling the genome. Statistics of the new assembly were calculated for each batch of reads to determine if targeted improvements were occurring and whether diminishing return had been reached (Figure S3). After 4 rounds (800,000 reads), the contig N50 had increased two-fold to over 30 kb and the sequence coverage had increased from about 4x to 6x. Although both of these statistics were still improving with the final batch of reads, the rate of increase was significantly less than for earlier batches and it was judged not to be cost effective to continue with this mode of upgrading. In addition, the current statistics appeared to have improved the assembly sufficiently to allow accurate assessment of gene content.

**Atlas assembly of WGS reads.** The Atlas assembler[2] has been used in a number of genome projects including the rat, *D. pseudoobscura*, and microbial projects. The process involves initially finding overlaps between reads based on their content of oligonucleotide sequences. Reads that contain highly repeated sequences (e.g. represented at more than 5x the sequencing coverage, although this is tuned differently for different assemblies) are set aside. Overlaps are determined among the remaining reads and lower copy number repeats are identified by anomalous read layouts they cause (e.g. forking in a linear path). These forks are split and then the groups of overlapping reads (bins) are assembled into contigs (contiguous sequence stretches). Next the contigs are linked into scaffolds (contigs that are ordered and oriented with respect to each other but separated by gaps) based on read pairs. This was the v2 assembly for the honey bee. The repeated sequence reads are separately assembled into contigs (called reptigs) using much more stringent overlapping criteria and these reptigs are added to the assembly based on their read pair links with the scaffolded reads. Thus the content of highly repeated sequences that appear in the assembly depends on how well they assemble into reptigs and can be placed with read pairs, and is likely incomplete. This created the v3 assembly. Finally the scaffolds are aligned to the linkage map (AmelMap3[6]) based on sequence matching with genetic markers to create the final assembly. The v4 sequence was created using the latest set of markers for this purpose. Although gaps are present, a 'linearized' sequence is produced for each chromosome by including 'Ns' as place-holders. Scaffolds that do not map to any chromosome, i.e. that fall between markers and thus cannot be placed, are combined into an unmapped set called Chromosome or Group Un. Redundant contigs from a second haplotype were omitted from the linearized sequence but were accessioned in GenBank.

**Assessment of accuracy of assembly – comparison to directly sequenced BACs.** 187 BACs were randomly selected for direct sequencing. Because the BAC library posed some difficulties, a total of 27 of these (over 4 Mb) that were sequenced to at least 6x coverage were assembled with PHRAP and compared to the v4 honey bee assembly. Statistics for the 27 and 187 BACs are provided in Table S3 D and E. The dot plots are available for viewing at www.hgsc.bcm.tmc.edu/~lzhang5/images/hb-bac/

Each plot is one BAC (x-axis) against one scaffold (y-asix) from the v4 assembly. The results can be summarized as follows:

**1. Coverage:** all contigs in the 27 BACs were found in the latest assembly using BLAST, except one contig (size=1.1kb) which has only one read and did not hit BIN0 reads (these are reads that overlap few or no other reads; thus there 'bin' is called BIN0 and they are not included in the assembly) or the nr database at NCBI, and is thus likely bogus. All except for four of the 27 BACs showed >94% coverage by the assembly, the remaining four being BACs with high repeat contents that showed anomalies in their assemblies. At the BAC level, all BACs could be mapped to chromosomes 1-16, except for small reptig (repeated sequence contig) matches to the Unmapped group (GroupUn). Some contigs had one end hanging in the uncaptured gaps (gaps that are not spanned by a read pair) between scaffolds. Examples: BAC AMAX.Contig47 was found on scaffolds 4.17 and 4.18, with 14kb in the middle falling in the gap between them. Other cases are: AMEO.Contig12 has 5kb at one end in the gap between 4.16 and 4.17. AMEO.Contig15 has 26kb at one end in the gap between 4.16 and 4.17. These could be problems in the BAC assembly, rather than issues with the v4 assembly, but are relatively minor and not unexpected for draft sequences.

**2. Duplication:** the BLAST result showed no major duplication in the v4 assembly. All contigs had a unique genome location (except some small ones that matched only reptigs on GroupUn).

Based on this, each contig was assigned to a scaffold (or two only if both are partial) on Chromosome 1-16 for analysis by running crossmatch.

**3. Insertion/Deletion of BAC contigs:** no major insertion or deletion at the contig level were found in the crossmatch results. Almost all matches were contiguous, with allowance for some captured gaps in the assembly, within one scaffold. Only one exception was found, the problem of AMDL described below in detail.

**4. Tails of BAC contigs:** no large tails were found in the crossmatch results.

**5. Conflicts with assembly:** two BACs conflicted with the assembly, as detailed below.

Case 1: BAC AMDL matches Group6.27, 6.28 and 6.30 (shown in three plots at the above URL). The tail of Contig104 on 6.27 and the missing part on 6.28 indicate an insertion in the end of 6.27. It turned out to be a 69kb reptig at the end of 6.27. This big reptig might be misplaced. The matches of AMDL on Group6.30 are all small contigs (less than 2kb). These matches on 6.30 are on two big assembly contigs (no reptigs involved). If we link 6.27 and 6.28 based on the match of Contig104, we cannot explain the matches on 6.30 because these were beyond the size of the BAC. It is likely that these short low coverage BAC contigs belonged to some other BACs, not AMDL. Thus AMDL might be contaminated or mixed with some other BAC, which happens at this frequency.

Case 2: AMCI matches Group13.10 and Group4.17 (shown in two plots in the above URL). AMCI contigs can be separated into two non-overlapping sets, separately matching 13.10 and 4.17. Almost all the longer contigs go to 13.10. On both scaffolds, the matches are clustered in a BAC sized region. This indicates AMCI is a mixture of two BACs. The marker mapping on 13.10 and 4.17 confirmed each matched region belongs to the corresponding chromosome. No reptigs were involved in these regions. Again, we would expect infrequent BAC contamination in the production pipeline.

In general, the result looks very good in terms of agreement between directly sequenced BACs and the draft v4 assembly. The few very detailed problems affect a small portion (~0.1Mb) of the ~4.5Mb sampled by 27 BACs, and are expected for a draft sequence, while the vast majority of the sequence is in agreement. One extrapolates from this that there would be of the order of 100 such issues in the whole genome.

**Assessment of completeness of assembly.** Markers, cDNAs and EST contigs were searched against the linear scaffold sequences and unassembled BIN0 reads using BLAST. For marker matches, any hit with a minimum identity of 95%, a minimum score of 100 and an e-value of e-20 was considered as a good hit. For cDNAs, a minimum identity of 95%, a minimum score of 200 and an e-value of e-10 was required for a good match. For ESTs, a minimum identity of 95%, a minimum score of 200 and an e-value of e-10 was required for a good match. For cDNA and ESTs, 90% or more of the sequence length coverage was required.

**Assembly of superscaffolds for chromosomes 13, 14, 15, and 16.** Components of the superscaffolds are scaffolds (GenBank/EMBL/DDBJ accessions NW_001262656, NW_001262654, NW_001262552, NW_001262481, NW_001261703, NW_001261682, NW_001261680, NW_001261637, NW_001261604, NW_001261552, NW_001261486, NW_001261446, NW_001261395, NW_001261255, NW_001260980, NW_001260955, NW_001260913, NW_001260790, NW_001260755, NW_001260754, NW_001260720, NW_001260636, NW_001260546, NW_001260518, NW_001260341, NW_001260051, NW_001259944, NW_001259874, NW_001259640, NW_001259366, NW_001259332, NW_001259258, NW_001259142, NW_001258991, NW_001258569, NW_001258460, NW_001258104, NW_001258076, NW_001257940, NW_001257881, NW_001257763,

NW_001256909, NW_001256766, NW_001256757, NW_001256741, NW_001256690, NW_001256521, NW_001256508, NW_001256454, NW_001256441, NW_001256302, NW_001256206, NW_001256174, NW_001255979, NW_001255975, NW_001255918, NW_001255799, NW_001255727, NW_001255725, NW_001255715, NW_001255593, NW_001255580, NW_001255563, NW_001255538, NW_001255359, NW_001255287, NW_001255268, NW_001255176, NW_001255098, NW_001255078, NW_001254957, NW_001254953, NW_001254943, NW_001254754, NW_001254718, NW_001254713, NW_001254706, NW_001254653, NW_001254637, NW_001254631, NW_001254618, NW_001254516, NW_001254510, NW_001254495, NW_001254478, NW_001254477, NW_001254460, NW_001254453, NW_001254370, NW_001254293, NW_001254290, NW_001254259, NW_001254246, NW_001254223, NW_001254211, NW_001254191, NW_001254142, NW_001254087, NW_001254041, NW_001254002, NW_001253957, NW_001253952, NW_001253945, NW_001253937, NW_001253893, NW_001253861, NW_001253848, NW_001253842, NW_001253839, NW_001253838, NW_001253789, NW_001253782, NW_001253774, NW_001253716, NW_001253683, NW_001253670, NW_001253651, NW_001253608, NW_001262697, NW_001262290, NW_001262125, NW_001262070, NW_001261685, NW_001261553, NW_001261509, NW_001261245, NW_001260794, NW_001260574, NW_001260530, NW_001260376, NW_001260343, NW_001260233, NW_001260211, NW_001259947, NW_001259826, NW_001259496, NW_001259243, NW_001259232, NW_001253166, NW_001253165, NW_001253164, NW_001253163, NW_001253162, NW_001253161, NW_001253160, NW_001253159, NW_001253158, NW_001253157, NW_001253156, NW_001253155, NW_001253154, NW_001253153, NW_001253152, NW_001253151, NW_001253150, NW_001253149, NW_001253148, NW_001253147, NW_001253146, NW_001253145, NW_001253144, NW_001253143, NW_001253142, NW_001253141, NW_001253140, NW_001253139, NW_001253138, NW_001253137, NW_001253136, NW_001253135, NW_001253134, NW_001253133, NW_001253132, NW_001253131, NW_001253130, NW_001253129, NW_001253128, NW_001253127, NW_001253126, NW_001253125, NW_001253124, NW_001253123, NW_001253122, NW_001253121, NW_001253120, NW_001253119, NW_001253118, NW_001253117, NW_001253116, NW_001253115, NW_001253114, NW_001253113, NW_001253112, NW_001253111, NW_001253110, NW_001253109, NW_001253108, NW_001253107, NW_001253106, NW_001253105, NW_001253104, NW_001253103, NW_001253102, NW_001253101, NW_001253100, NW_001253099, NW_001253098, NW_001253097, NW_001253096, NW_001253095, NW_001253094, NW_001253093, NW_001253092, NW_001253091, NW_001253090, NW_001253089, NW_001253088, NW_001253087, NW_001253086, NW_001253085, NW_001253084, NW_001253083, NW_001253082, NW_001253081, NW_001253080, NW_001253079, NW_001253078, NW_001253077, NW_001253076, NW_001253075, NW_001253074, NW_001253073, NW_001253072, NW_001253071, NW_001253070, NW_001253069, NW_001253068, NW_001253067, NW_001253066, NW_001253065, NW_001253064, NW_001253063, NW_001253062, NW_001253061, NW_001253060, NW_001253059, NW_001253058, NW_001253188), PCR products (GenBank/EMBL/DDBJ accessions DQ833243-DQ833248), and unscaffolded contigs (GenBank/EMBL/DDBJ accessions ED552173-ED552188).

## Genome organization

**Partition of Genomic Sequences into Segments that Have Characteristic GC Contents and Differ Significantly from the GC Contents of Adjacent Segments.** Several methods have been proposed in the literature for identifying segments with characteristic GC content[7, 8]. In this study, we partitioned the genomic sequences into segments by the binary recursive segmentation procedure, DJS, as proposed[9]. In this procedure, the chromosomes are recursively segmented by maximizing the difference in GC content between adjacent subsequences. The process of segmentation is terminated when the difference in GC content between two neighboring segments is no longer statistically significant[10].

## Transposons

**Details of the *mariners* summarized in Table 2[11].** The *mellifera* subfamily of *mariners* was named for the first *mariner* discovered in bees[12], and this element, *AmMar1* (1287 bp), has been described from bee as a relatively recent horizontal transfer into this genome[13, 14]. RepeatScout[15] built separate consensus sequences for the many internally deleted copies of *AmMar1* described[13] and for the full-length version of *AmMar1*, which differs from a previously described sequence based on five copies (GenBank AY155490.1) by just six bases[14]. We found approximately 360 copies in genome assembly v4 that differ from the consensus by 4-5%; these represent an explosion of copies from a single relatively recent horizontal transfer, with a particular internally-deleted copy becoming common. There do not appear to be any intact putatively active copies of *AmMar1* left in the bee genome, unlike the closely related *Famar1* element in the earwig *Forficula auricularia* even though the bee's element is younger[16].

Like most other animal genomes, the bee genome has apparently been repeatedly invaded by different lineages of *mariners*. A second type of *mariner*, also in the *mellifera* subfamily, was originally identified as PCR clone "honey.bee.4.4"[12] and a full-length consensus sequence based on preliminary genome sequence was included as *AmMar2*[17]). RepeatScout[15] generated two slightly different overlapping consensus sequences for this element, which when combined yield a 1284 bp *mariner* encoding a full-length ORF. There are about 100 copies of *AmMar2* in the genome assembly, differing from the consensus by 6-10% and commonly affected by internal and terminal deletions, so this *mariner* resulted from a slightly older horizontal transfer into the bee genome, and again no intact putatively functional copies remain.

There are at least four more even older lineages of *mariner* family elements in the bee genome. *AmMar3* is an *irritans* subfamily element with 83 degraded copies. There are two overlapping RepeatScout consensi that can be extended to almost full-length but still does not encode an ORF. *AmMar4* is a *rosa* subfamily *mariner* (or ITmDD41D family[18]) with just 10 reasonably full-length copies and over 380 copies with an internal deletion of 386 bp removing most of the coding region for the N-terminus of the transposasewith over 390 degraded copies. The RepeatScout consensus fuses this with another low-copy repeat. No intact copies appear to remain in the bee genome, but a 1304 bp consensus of the 10 full-length copies does encoded a full-length transposase ORF.

*AmMar4* is an *irritans* subfamily element with about 80 degraded copies. There are two overlapping RepeatScout consensi that can be extended to almost full-length but it still does not encode an ORF. Several copies are nearly identical in the genome, but this appears to be because they are embedded within a longer recently duplicated sequence. *AmMar5* and *AmMar6* are short consensus sequences of fragments of highly divergent *mariners* with approximately 70 and 130

highly degraded copies, respectively.repeat and refinement of the *AmMar4* consensus allows recognition that it is a 918 bp element with an internal deletion removing the N-terminal coding region of the transposase gene. No intact copies appear to remain in the bee genome. *AmMar5* and *AmMar6* are short consensus sequences of fragments of highly divergent *mariners* with approximately 76 and 140 highly degraded copies, respectively.

The consensus sequences for these six *mariners* are:

**AmMar1**

TTGGGTTGGCAACTAAGTAATTGCGGATTTCACTCATAGATGGCTTCAGTTGAATTTTTAGGTT
TGCTGGCGTAGTCCAAATGTAAAACACATTTTGTTATTTGATAGTTGGCAATTCAGCTGTCAAT
CAGTAAAAAAAGTTTTTTGATCGGTTGCGTAGTTTTCGTTTGGCGTTCGTTGAAAAATGGAAAA
TCAAAAGGAACATTATCGTCATATTTTGCTTTTTTATTTTCGCAAAGGGAAAAACGCATCGCAA
GCTCACAAAAAGTTATGTGCTGTTTATGGCGACGAAGCCTTAAAAGAACGGCAGTGTCAAAATT
GGTTTGACAAATTTCGTTCTGGTGATTTTTCACTCAAAGAAAAAAAACGCTCTCGTCGTCCAGT
TGAAGTTGATGACGACCTAATCAAAGCAATAATCGATTCGGATCGTCACAGTACAACTCGTGAG
ATTGCAGAGAAGCTTCATGTATCACATACATGCATTGAAAACCACTTAAAACAACTTGGCTATG
TTCAAAAACTCGATACATGGGTTCCTCACGAACTGAAAGAAAAGCATTTAACGCAACGCATTAA
CAGCTGCGATTTGCTAAAGAAACGTAATGAAAATGATCCATTTTTAAAACGACTGATAACTGGC
GATGAAAAATGGGTTGTTTACAACAATATCAAGCGGAAAAGATCGTGGAGCAGGCCACGTGAAC
CAGCTCAAACAACATCAAAAGCTGGTATTCATCAAAAGAAGGTTTTGTTATCAGTTTGGTGGGA
TTACAAAGGAATTGTCTATTTTGAACTCTTACCACCCAACCGAACGATCAATTCTGTTGTCTAC
ATTGAACAACTAACGAAATTAAACAATGCAGTTGAAGAAAAGCGGCCCGAATTGACAAATCGAA
AAGGTGTTGTATTCCATCATGACAATGCAAGGCCACACACATCTTTGGTCACTCGGCAAAAATT
ATTGGAGCTTGGTTGGGATGTTTTGCCACATCCACCATATAGTCCTGACCTTGCACCATCTGAT
TACTTTTTGTTTCGATCTTTACAAAACTCCTTGAATGGTAAAAATTTCAATAATGATGATGATA
TCAAATCGTACCTGATTCAGTTTTTTGCTAATAAAAACCAGAAGTTTTATGAACGTGGGATTAT
GATGCTGCCTGAAAGATGGCAAAAGGTCATTGATCAAAATGGGCAACACATTACAGAATAAAGT
TATTTAGTTCCATGAAAAAATTGTCTTTGATTTTCTAAAAAAAATCCGCAATTACTTAGTTGCC
AACCCAA

**AmMar2**

TTAGGTCTACCGGAAAGTTCTGTCCGAATCTATGACATCATTTTCGCCACGTAAGCACATGTTT
ATTTATTGCATGTTCGGCTCTATATTTTTATCGCTTAATGTATACATACTGACGTAGCAAATAA
ACTATAATAAAGTTGATTCACATTAGTCTTAAGTGTGAAACGATAGTATACCCATGGCGACTGA
TAAAGTTCATTTACGCCACTGTATTTTATACGAATTTCAACAAGGAAGAAATGCTACAGAAGCA
TGTAGAAATTTATTGAAAGTGTTTGGTGAAGGTACAGTTTCTGATAGGACATGCAGAAGATGGT
ACGAAAAATTTGAAACAGGTGATTTCGACCTTTCTGATAAGCCACGTTCTGGGCGACCATCTTT
GATCGACGACGATGTTGTTAAGGCAATGTTGGAGCAAGATCCTTTTCTGACAACATCGGAGATC
GCAGAAAGGCTTAATTCAGCTCAACAAACCATTTCTGACCATATTCGGAAGATAGGATTGGTGT
GGAAGTATTCAAGATGGGTGCCACATGAATTAAGTCAGAAAAATTTGGATGATCGAGTTGTCAT
ATGCACATCTCTGCTTGCTCGGAACAAAATCGAGCCCTTTTTGAACCGGATGATAACTGGGGAT
GAAAAGTGGATTACATACAACAACATTGTAAGGAAAAGGGCATATTGTGAACCCGGAAAACCTA
GCCCTTCCACCTCTAAACCAAATTTGACTCTGAATAAGAGAATGTTGTGTATATGGTGGGACAT
TCGAGGACCAATATATTATGAGCTTTTAAAACCGAACGAAAAGCTCAATTCGGAGAAGTATTGT
CAGCAACTGGATAATTTAAAGACAGCGGTCCAAAAAAAGAGGCCGGCAATGTTCAATAGGAAGG
ACATGATACTGCACCACGATAACGCCAGACCACACGCTGCTTTAGGGACTCGTCAAAAAATTGC
AGAAGTAGGCTGGGAAATTCTGTCGCACCCACCATATTCCCCGGACATAGCACCCTCTGATTAT
CACTTGTTTTTATCCTTACAAAATTTTTTGACGGGCAAAAAATTCAAAAATGAAGAAGATGTAA
AATAATCATTATTTAAATTTTTCATATCAAAATATAAAATATTTTTAAAAAATGGAATATACAA

ATTGCCCTCACGCTGGCAAGAGATCATTAATAATAATGGCAATTATATTATTCAATAAAGTTAA
TTGGCGGTAAGAAAAAATTTGTATTTTGTTTTATTCCAAAAACGGACAGAACTTTCCGGTAGAC
CTAA

**AmMar3**

AAGGGTGTCCCAAAATTAACGCAAGATATGAATTTGCCGCTATTTTTGCATTAAGTTGTTGGCA
ACCCTGAAAAAAGAACAGTTTGACAGCTGAGAGTTTAGTGTTAGTAAAAATGGAGCGTTATACG
ATACAACAACGTGTCTTCATTATTGAACAATATTTTAAAAATAATGAAAGTTTGGCGGCCGCAG
TTCGAAAATTTTATACAAAATATGATAAAAATAGTGTTTTAACCTCGTCAACTGTGAAAAGATT
AATTGAAAAATTCGTGGAGACTGGATCAGTTGGAGACGCTAAACACACCGGTCGTCCAAAAACA
AGCCGTTCAAATGTCAATATTGAAGCAGTGCGTGAGAGTGTTGGTGAAAACCCAGGAACATCAA
TTCGGCGTCGTGGACAAGAATTGCAAATTTCAAGAAGCTCTCTACAGCGTATACTCACAAAAGA
TCTGTGTCTTCATGCTTACAAAATTCAATTAACACAACAACTGAAGCCTAATGACCATGAACAG
CGAAGAGAGTTCGTGGAATGGATTATTAATCATCAAAAAGTGGATGCTGGTTTTTCGAGCAAAA
TAATCCTAAGCAATGAAGCACATTTTCACCTCGATGGCTTTGTTAATCGCCAAAATTGCCGTGT
TTGGGGTTCGGAGAACCCACGTGTGATTAGCGAAAAACAAATGCATCCACAACGTGTCACTTTT
TGGTGCGGATTTTGGGCAGGAAGCATCATCGGACCATACTTTTTTGAGAATGAGGCTGGTCAAG
CAGCAACTGTTAATGGTGCTCGATATCGCGATATGATAACACAGTTCTTTCTGTCGAAATTGGA
TGATATTGATGTGGCCAATATGTGGTTTCAACAAGACGATGCCACATGCCATACAGCCAATGAA
ACAATTCAATTACTGCATGAGACATTTCCTGGTCGTGTACTCTCTCGTTTCGGTGATCAGAATT
GGCCCCCTAGATCATGTGATTTAACACCATTAGATTTCTTCTTATGGGATTATTTGAAATCAAA
GGTCTATGTCAACAATCCCACAACCACACGTGCATTACAAGAGGAAATTAAACGCTGCATCAAT
GAAATTCAACCACAATTATGCAGAAAGGTCATGAAAAATTTCGACGAAAGGGTGCGTATGTGCC
AGCAAAGCCGTGGAGGCCATTTGCCCGATGTGTTATTCCATAAATAACCCTATCCTATGTACTT
TATGATTCACTTAAAAAATAAATATCTAAAGAATAAAAACTCTCTTTTATATTTAATTCAAATC
TTGCGTTAATTTTGGGATACCCTT

**AmMar4**

TGCATCAGGTTGGAAAGAAGGTTTTCACGATTTTTATATTGATATATTGATTTTATATTGATTT
AAATGCTCTATTTTTGATTAATGCAAAATTCTATTGGTTTTGTATATAATTTTAATTTTGCATT
TTTCGCTTTATGAAAATATCATATTTTATTTTCAAATTAGTTTTTCTGATTTCTTAAGTTATGT
TAAAGACATAATGAACAAGAATTCGAACAATTTTCCTATTCAAACTAAATCGGAAAACTACGAA
AACAATTCGCGAAGCGTTTGGGAATACTAACAAACATACAATATCAACGAAGTGTTTGGGTTTA
TTAACAAACATACTGCATATTGGTGGTTTAAAAAATTTTGATGACGAAAGCCTTGAAGACAATC
AGCGCTGTAAAATTAGCTATCAGATATTGACAATAGTGACTTGAAGATTCTAGTTGAAGCTAAT
CCTCATACAACCGTACGGAAGTTTGTTTCTGAATTGAATGTAAAGCATATAACAATTTATAATC
ATTTAAGAAAATTGGAAAAACAAAAAAGCTTGATAAATTGGATGCTTGATTGGGTGCCGCGACC
CAATTAAATTAAAGAAAAAAAAAATCATTGTTTTGAAATATCATCTGCCCTTCTTTTGCGCAAT
AAAAATGATTCATTTTCTCGAAGGAATTGTAACGTGCGATGAAAAATGGATTCTTTATAATAAT
TGGCGACGATCGACTCATTGACTAGATCAAGACGAAGCTTCACAACATTTCCCAAAGTCAAAAT
TTCACCAAAAGAAGATCATAATGATAGTTTGGTGGTCTGTGGCCAGTTTGATTCATCACAACTT
CCTGAATTCTGGCGAAACTATTACAACTAAAATGTACTGTCAATTCGATGAAATGCACGAAAAA
CTTCGTTTGTGTCCAATATTGCTCAACAGAAATGTTCTATCCTTCTCCACGATAATGCTCCGTC
ACACGTCGCTCAACTGATCCTTTAAAAATTGAACGAATTGGCCTACAAAACTCTACTTTATCCA
TACTTGCTAAATCTCTTATCCACCGATTACCATTTTTTCAAGCATTTCGACAACTTTTTATATG
AGAAATGCTTCAAATCCCAGAAAGATATTGAAACAGCATTCAATGAATTTGTTGCCTCCAAGAT
TTCAGAATTTTATTCAACCGGAATAACAAAAGTTGTTCCTTATTGGGAAAAGTGCATTGATTAA

AATGATTTTTATTTTAATTAATAAAGTTCTGAATTGAAATATGTGTATTTAAATTTAATAGTTA
AAAACCGCAAGAACTTTCTTTCTAATCTAA

**AmMar5**

TATATAATATAAAATGAGTAAAAGTTGTTAAAAAATCACATCCTTTAAAACGTTTTTTACTATT
CAATATGAATAGCAAAAGGAAGCATTTGCGGCATGTTATACTTTATTTTTTTAAAAAAAGGTGA
TAATGAAAATGATACTGCAGATGAAATTTGTATTGTTTACAGGAATGATGGTATAACCATTACG
ACCATCCATAATTGGTTTGAGAGATATAGTGCTGGCAATTTTGACTTGAAAAATGAAGGACCCT
ACGGCCATCCAGCAACGATAAATATGGATGTTATCAAGACCATGCTTGCTGAAAATCCGCGATG
CACAGTGTGCAAGAGATAGTGAATGCCATTAATATTTCCAGGAAAGCTGTAGATAATTTTTGGA
GAATTTGGGTTCTACAGCTATTGATGAAAACCGACTTTATGAAATAAAGTCTCTATGTGCAATT
TCCTTCTTCAAAGACATGAAAGAGATCTTTTTTTAAAGAGGCTTATCACTGAAGAGTAGACTTG
GATTTTGTATCAAAATGTATCGAAAACGCACTTGATTTAAGAACGATAGACCTTCAACTGTCGC
GAAACCTAGACTTCGCGAAACCTGAAGAAAGTTTTTTTATCCATTTGTTGGGATTGGAAATGTA
TAATCTATTATGAGCTCCTTTCTCAAAATAAAGCTCCTTTCTCATCAATTTTAGAAAATATTCT
TTTCGTGATAAATGGCTCCAATTCGTAAGCGAAAAAAAACGCACTTCGAGGAATATTTTGTAAA
TAAACCCCAACAATTTTGGAAAAAGAGAGGCTTTCTGAGAAATAGAAGAACAAAACGATGATAG
AGCAGAATGATTCATATATAATACAATAAATTAATCTTAAACAAAAAATATCGTATATTCATTT
CGTA

**AmMar6**

ACTTGGATTTTATATCAAAATATACATTGAAAATGCATTTGATTTAAGAACAATAGATCTTCAA
TTGTCGCGAAGCTTGGACTTCACCCGAAGAAAGTTCTTTTGTTCATTTGATGGGATTGGAAAGA
AGTAGTTTATTATGAGTTCCTTTCTCAAGATGAAATCATCAATTCTAGAAAATATTGCAATCAG
CTTGATAAAATTAAAAAAGCCATAGCAGAAAAACGACCAAAATTGGCAAAATGATGAGGCCATC
ACGACAATGCAAAACCACATGTTGCATTGACTGTAAGAGAAAAGCTGTTACAGTTTGATTGGGA
TATTCTATTGTATCTTCTGTATTCTCCAGATCTTGCTCTATCCTACTATTATTTGTTCCTGTTA
TTAAAAAATTCTTTTCATGATAAACGATTCCAATTCGTAAGCGAAATAAAAACGCATCTCGAGG
AATATTTTGCAAATAAACTCCAATAATTTTGAAAAAAGAAAATAATGAAAATTTATAAAAGATA
AAAAAAAATAATAGAACAGATCATTTAATATATAATAAAATAAATTAATCTTAAACAACAAATA
TTATATATTTATTTCATATCAAAA

**Searches for other transposons:** No matches were found in searches of assembly v3 using TBLASTN, and of the official and *ab initio* protein sets using BLASTP, for transposons and transposases of the mori subfamily of *mariners* (or ITmDD37D/maT family), *gambol* (ItmDD34E), *Tc1* (ITmDD34E), and other families in the *IS630-Tc1-mariner* or ITm superfamily that are widespread in animal and other genomes (e.g. [18-21]). The bee genome appears similarly devoid of other major DNA transposon superfamilies such as piggyBac[22], P elements, Transibs, and helitrons[23]. We also examined all of the high and medium copy repeats identified by RepeatScout for ORFs in hopes of identifying additional transposons encoding little or no similarity to known transposases/integrases. This approach is based on the ability to reconstruct intact consensus sequences for transposons from many degraded copies in a genome (e.g. *AmMar1* and *2* above[14, 23]), and should allow identification of recently active transposons. No candidates beyond *AmMar1* and *2* were discovered this way. Additional more intensive searching of this genome as described[23] might eventually lead to discovery of new transposons in the bee genome.

## Gene Sets

**Gene Predictions.** Six gene prediction sets were independently generated. The NCBI gene prediction process included cDNA, EST and protein alignments, using Splign and ProSplign[24, 25]. The best scoring CDS was identified for all cDNA alignments, using a 3-periodic fifth-order Markov model for the coding propensity score and WMM models for the splice signals and translation initiation and termination signals. These are the same scores used with Gnomon[26], the NCBI *ab initio* prediction tool. All cDNAs with CDS scoring above a certain threshold were marked as coding cDNAs, and all others were marked as UTRs. Some of the CDS were incomplete, meaning that they lacked a translation initiation or termination signal. All 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, they were assembled using a modification of the Maximal Transcript Alignment algorithm[27], taking into account not only exon-intron structure compatibility but also the compatibility of the reading frames. Two coding alignments were connected only if they both had open and compatible CDS. UTRs were connected to coding alignments only if there were necessary translation initiation or termination signals. There were no restrictions on the connection of UTRs other than the exon-intron structure compatibility. All assemblies with a complete CDS, including the translation initiation and termination signals, were combined into alternatively spliced isoform groups. Incomplete assemblies were directed to Gnomon for extension[26].

The Ensembl gene predictions were built using an iterative strategy, based on both protein and EST evidence. The method was adapted from Curwen et al. 2004[28]. First ESTs from dbEST and Riken were aligned to the honey bee genome. The first-round gene predictions were made using alignments of Uniprot proteins to the genome, with gene structures created by Genewise, an evidence-based gene predictor[29]. These first-round protein-based gene predictions were combined with the spliced EST-based gene predictions to yield a non-redundant set of high-confidence transcripts. The second round of gene predictions concentrated on the gaps between the first-round genes of greater than 5 kb. These gaps were re-aligned to Uniprot using Blast, and the resulting hits were used to seed a second run of Genewise in these gaps. The resulting transcripts were again collapsed down to a non-redundant set.

The "Evolutionary Core set" used a homology-based gene prediction pipeline, which relies on similarity to known proteins to identify putative genes and uses orthology to identify evolutionary conserved core of genes (E. Zdobnov, unpublished). The pipeline consists of the following steps: 1) identification of all genomic regions with significant homology to known proteins by applying TBLASTN[30] searches to a comprehensive, non-redundant collection of protein sequences, e.g. Uniref50[31]; 2) identification of matches that are consecutive along the protein and genome to delineate gene loci; 3) selection of the most similar in sequence and well defined proteins for each of these genomic regions with protein-coding potential; 4) use the selected protein for homology-assisted gene prediction using Fgenesh+[32]; and 5) identification of orthologous gene relations[33, 34] among candidate gene predictions across several species through application of all-against-all Smith-Waterman comparisons to discriminate the conserved core genes. This lightweight schema targets to acquire a sensitive and unbiased view of evolutionary conserved core of genes in multiple genomes rather than produce a full catalogue of genes. However, it performed very well in comparison to more sophisticated pipelines described above and it seems particularly suited for studying divergent genomes where only limited EST or cDNA data are available.

Two gene prediction sets were generated at Softberry Inc using Fgenesh and the Fgenesh++ pipeline[32, 35]. Fgenesh is a HMM based *ab initio* gene prediction program. Fgenesh++ is a pipeline for automatic prediction of genes, which in addition to Fgenesh, includes sequence analysis software to incorporate information from full-length cDNA alignments and similar proteins from the eukaryotic part of the NCBI NR database (without predicted *Drosophila* proteins). Both Fgenesh and Fgenesh++ used bee-specific gene-finding parameters trained on known genes of organisms closely related to honey bee.

Another gene prediction set (the "Drosophila Ortholog"set) was developed by mapping *D. melanogaster* gene models onto the honey bee genome using the comparative gene finder GeneWise[29] to build gene models, and a modified reciprocal-BLAST approach to assign orthology/paralogy relationships (V.N.Iyer, D.A.Pollard and M.B.Eisen, unpublished). Candidate regions for building gene models were identified on the honey bee genome scaffolds by TBLASTN with *D. melanogaster* translations as query, followed by clustering of the HSPs. GeneWise was used to build gene models in these regions with the *D. melanogaster* translation as evidence, and the resulting gene models were compared back to the set of *D. melanogaster* translations using BLASTP. Orthology/Paralogy relationships were assigned by a heuristic algorithm that takes into account (a) the rank of the starting *D. melanogaster* translation in the BLASTP results, (b) the rank of alternative translations from the gene corresponding to the starting *D. melanogaster* translation, and (c) whether or not there were highly ranked hits to genes other than the gene corresponding to the starting *D. melanogaster* translation. Hits having e-values within one order of magnitude were considered to be equivalently ranked. "One-to-one orthology" was assigned when there was a single honey bee gene model and the only top-ranked hits in the BLASTP results for that model were translations from the gene corresponding to the starting *D. melanogaster* translation. The final *Drosophila* Ortholog set consisted of the one-to-one ortholog models.

The individual gene prediction sets were integrated using GLEAN[36]. GLEAN is a tool for creating consensus gene lists by integrating gene evidence (Mackey *et al*., personal communication). It uses Latent Class Analysis to estimate accuracy and error rates for each source of gene evidence, and then uses these estimates to reconstruct the consensus prediction based on patterns of agreement/disagreement observed between each evidence source. The GLEAN analysis integrated the following gene prediction lists: NCBI, Ensembl, Fgenesh, Evolutionary Core, and Drosophila Ortholog, as well as aligned proteins and ESTs. The Fgenesh set was selected instead of Fgenesh++ as an input data set, because Fgenesh is an *ab initio* program that does not use homologue evidence, and was expected to increase the overall yield of genes in the consensus set. The proteins were from metazoan SwissProt, aligned using EXONERATE[37] with a minimum score 50, using only the highest scoring of overlapping sequences. The ESTs were consensus dbEST and Riken ESTs aligned using TGICL[38] with a minimum 95% identity and 90% alignment coverage. GLEAN analysis labels each prediction with a confidence score reflecting the underlying support for that gene. Evaluations using manually curated gene models built from evidence that was not used in the training sets (gold standard gene models) showed that the GLEAN consensus gene models were superior to the individual input gene models[36]. These evaluations used FASTA[39] with 100% alignment coverage, at least 99% identity and no gaps for identity matches and at least 95% identity, not considering gaps or alignment coverage for presence (weighted by number of genes in a prediction set).

**Functional Predictions.** The functions for 70% of the GLEAN consensus gene models were reliably inferred as described[40]. A complete list of all proteins that show any annotation with features from InterPro, SwissProt keywords, Enzyme database, GO information, and family members present or absent in other insect genomes is available from ProtoBee (www.protobee.cs.huji.ac.il).

## Analysis by a consortium.

The annotation consortium used tools at BeeBase and elsewhere to manually annotate gene models using standard operating procedures developed by community members and BCM-HGSC. Gene models were submitted to the BCM-HGSC annotation database and included transcript and protein sequences, exon coordinates, homolog identifiers, functional descriptions, multiple alignments of gene families, phylogenetic trees, and comments about corrections to GLEAN gene models. Following transfer to BeeBase, each gene model was mapped to the assembly (version 2), then viewed using the Apollo annotation browser to verify splice sites, identify redundant submissions and assign "GB" identifiers[41]. New gene models and corrections to the GLEAN set were incorporated into a second release of the OGS.

**Annotation of chromosome 15 and 16 superscaffolds.** The chromosome 15 and16 superscaffolds containing approximately 6.5 % of all bee genome were selected as the most representative part of genome with respect to genomic landscape and genetic properties. They contain a conservative HOX locus, one of the odorant chemoreceptor clusters and dozens of bee genes -human orthologs- associated with human diseases. Also the euchromatic part of chromosome 16 is abundant with CpG islands associated with mRNA starts of specific genes that was unusual for the known insect genomes.

We generated gene prediction data sets, BLAST hits to known protein/ EST/mRNA. Using Apollo Genome Annotation Curation tool[42] we carefully inspected each gene model and gene evidence and then manually curated 720 and 337 gene variants for 15th and 16th chromosomes respectively, of which 5 and 7 are tRNAs, 5 and 14 are pseudogenes with multiple frame shifts, and 71 and 62 are splice variants for the chromosomes 15 and 16 respectively.

About 188 and 116 gene models of the OGS (15th and 16th chromosome respectively) were significantly corrected by merge/split transcripts, adding/removing exons, adjusting alternative splice sites. Using BlastX, PSI- , PHI-BLAST searches we assigned putative functions to 639 and 254 protein-coding genes (15th and 16th chromosome respectively). Sequence and annotation data on these superscaffolds are available through Genome Browser Web site: racerx00.tamu.edu/...

48 and 23 new protein-coding gene models (15th and 16th chromosomes respectively) were added to the OGS, including 40 and 15 that were previously supported by only Fgenesh *ab initio* gene models. 56 and 21 transcripts of two superscaffolds were assigned as problematic because their genomic sequences contained gaps between contigs, insertions/ deletions and caused open reading frame shifts or ORF truncations.

The deficiency of EST data was compensated by extensive using of protein information – orthologous sequences annotated from genomes *Drosophila, Anopheles , Homo sapiens*, mouse and rat. Even though the 35% of annotated gene models were not covered by at least one spliced EST sequence the extensive protein similarities to preliminary described protein sequences allowed us to assign ORF and known protein function for more than 90% of the gene models. It is not excluded that enhanced manual annotation of all chromosomes with a using of additional

EST data allows us to exceed our preliminary gene estimations, because even with a deficiency of EST data we could create about 1057 gene models for only 6.4% of honeybee genome.

**MicroRNA identification.** Candidate microRNAs were identified using three different methods. First, sequences with homology to mature miRNAs from other species were identified. BLAST of the assembly (v. 2.0) using known miRNAs (release 7.0 http://microrna.sanger.ac.uk/sequences/index.shtml) identified several hundred provisional candidate bee miRNAs with significant matches to miRNAs from other species (E value $\leq 0.01$, wordsize 11). Refined alignments of the identified genomic regions (+/-~100bp) surrounding each were generated using Water (EMBOSS). Second, the bee genome was exhaustively searched for microconserved sequence elements (MCEs)[43]with exact matches to *A. mellifera, D. melanogaster* and *A. gambiae.* Finally, additional putative miRNAs were predicted using a new algorithm designed to scan segments of the *A. mellifera* genome for structural and thermodynamic characteristics found in miRNA precursors – stem-loop scanning (SLS) , which shares some features of previously described procedures[44]. *Apis* sequences identified by SLS were compared to similar sequences generated by implementation of SLS on the *D. melanogaster* genome. This third class of miRNA candidates consisted of SLS sequences in *A. mellifera* that aligned well with SLS sequences in *D. melanogaster*. Each candidate miRNA from all 3 sets, homologs, MCE and SLS, was folded to verify the thermodynamic propensity of the precursor miRNA to adopt appropriate hairpin secondary structure – with the mature miRNA residing in the stem. Table S7 presents microRNAs that were found.

**Tiling array experiments.** Briefly, 6,503,344 36-mer probes were selected uniformly from both strands of the entire *A. mellifera* genome with an average spacing of 10 bases between the neighboring probes. The arrays were designed from the V3.0 assembly, in order to provide the best available sequence at that time, but then the oligonucleotides were mapped back to V2.0, because that was the assembly that the OGS was derived from and thus the sequence that the gene coordinates referred to.

The chosen probes were divided into 17 groups and synthesized on 17 glass-based arrays using a maskless array synthesizer[45]. The arrays were hybridized in a single measurement with poly-A RNA pooled from multiple bee tissues and life-stages. Standard procedures were used for RNA amplification, labeling, hybridization, scanning and data post-processing[45]. A gene was considered transcribed if the probes within it showed high signals. This was determined based on a probability-based score that reflected the likelihood that the probes showed high-signals not by chance alone.

In whole-genome tiling array experiments, single measurement has been sufficient to generate meaningful information, because (i) the purpose of the experiment is genome-wide transcript discovery rather than differential gene expression monitoring, (ii) sufficient redundancy is built into the probe selection procedure to provide multiple observations from several oligonucleotide probes for each long transcript, (iii) data from the oligonucleotide arrays used here had shown high level of reproducibility[45-47]. However, key results from the above experiment were further confirmed with additional measurements as described below:

**Confirmation array.** A new array containing 388,422 36-mer probes was designed. The probes represented OGS genes, *ab initio* predicted genes, novel intergenic transcripts and few predicted miRNAs. The array was hybridized with a new pooled RNA sample. Single measurement was performed.

**GC-rich region array.** A new array was designed with 189,000 50-mer probes tiling both strands of 15 longest GC-rich chromosomal regions without OGS genes. Two hybridization experiments were performed on the new array measuring the same pooled RNA sample.

**RT-PCR.** RT-PCR experiments were perfomed on 24 genomic segments from four categories: (a) five long OGS exons untranscribed in the tiling arrays, (b) five *ab initio* predicted genes with strong tiling array signals, (c) five novel intergenic transcripts from the AT-rich regions of the genome, and (d) ten novel intergenic transcripts from the GC-rich regions.

**Orthology mapping.** Orthologous relationships between genes of honey bee, fruit fly, mosquito, human, chicken and fish were inferred through all-against-all protein sequence similarity searches using the Smith–Waterman algorithm and retaining only the longest predicted transcript per locus. The orthologous groups were formed then by: grouping recently duplicated sequences with over 97% identity within genomes to be treated subsequently as single sequences, forming triangles and tuples of the mutually reciprocal best hits between genomes, and expanding the seed orthologous groups by inclusion of co-orthologous sequences that are more similar to the othologous gene than to any other gene in any other genome, requiring also that all members of the group have matches overlapping by at least 20 residues. All orthologous classifications and the corresponding species copy-number distribution are available from http://cegg.unige.ch/SUPPL/Bee/. Orthology relations were not significantly influenced by the particular gene set used as shown in Supplementary Tables S8 and S9.

**Circadian rhythms.** The honey bee genome encodes a single orthologue for each of the "clock genes" *period (per)*, *timeless (tim)*, *cryptochrome (cry), clock (clk), cycle (cyc*), *vrille (vri*), and *Par Domain Protein 1* (*pdp1*). There are no orthologues to *Cry*-d (*Drosophila*-type Cry) and *Timeless1(Tim1)* genes, which are essential for clock function in *Drosophila*. The honey bee genome encodes only the mammalian-type paralogues (Cry-m, and Timeout = Tim2) which are thought to have different clock function[48]. Honey bee AmCRY lacks C-terminal domains implicated in dCRY photoreceptor function in *Drosophila* but domains that are implicated in the function of CRY-m (mammalian-type) proteins are highly conserved[49]. The honey bee orthologue of *tim2* (*amTIM2*) does not contain domains implicated in the negative feedback function of *dTIM1* in *Drosophila*. These analyses suggest that AmCRY and AmTIM2 fill different clock functions than dCRY and dTIM1 proteins. *AmCYC* and *AmCLK* are similar to orthologues of other insects. However, a transactivation domain is found on the C-terminal end of *AmCYC* as in mammals. In contrast, *Drosophila* has a transactivation domain in dCLK but not in dCYC[49].

These findings raise two questions: first, how did honey bees and mammals end up with similar clock proteins and flies with different ones, and second, does the honey bee clock work like the mammalian clock? In terms of the first question, phylogenetic analyses show that the basal animal lineage had both the mammalian and *Drosophila* types of *Cry* and *Tim*[49] (Supplementary Figure S9). *Drosophila* specialized on using one set of orthologues; both mammals and honey bees lost these orthologues and specialized on the other set. In terms of the second question, the temporal pattern of clock gene expression in the honey bee brain is more similar to mammals than to *Drosophila*[49]. These findings challenge the distinction commonly made between insect and vertebrate clocks and raise critical questions concerning the evolution and functional significance of species-specific variation in the molecular clockwork.

**The honey bee cys-loop ligand-gated ion channel superfamily.** The honey bee genome has revealed a cys-loop neurotransmitter-gated ion channel[50] superfamily consisting of 21 subunit members, two less than Drosophila although the honey bee possesses an extra nicotinic

acetylcholine receptor subunit[51, 52] (deposited in GenBank/EMBL/DDBJ under accessions DQ026031-DQ026039 and DQ667181-DQ667195). Members of this superfamily are known to play roles in many aspects of honeybee behaviour including foraging, learning, memory, olfactory signal processing, mechanosensory antennal input and visual processing[53-58]. Uncharacterized subunits may well represent novel key components of the honey bee nervous system. The superfamily also contains targets for imidacloprid (nicotinic receptors) and fipronil (GABA receptors), which are widely-selling insecticides used in crop protection[59, 60]. Understanding how insecticides interact with target receptors will help in the development of improved compounds that selectively act on pest species and spare beneficial insects.

The accession numbers of the sequences used in constructing the tree are: *Apis mellifera* Amel α1 (DQ026031), Amel α2 (NM_001011625), Amel α3 (DQ026032), Amel α4 (DQ026033), Amel α5 (AY569781), Amel α6 (DQ026035), Amel α7 (NM_001011621), Amel α8 (AF514804), Amel α9 (DQ026037), Amel β1 (DQ026038), Amel β2 (DQ026039), Amel RDL (DQ667181), Amel GRD (DQ667183), Amel LCCH3 (DQ667184), Amel GluCl (DQ667185), Amel HisCl1 (DQ667187), Amel HisCl2 (DQ667188), Amel pHCl (DQ667189), Amel 8916 (DQ667193), Amel 12344 (DQ667194), Amel 6927 (DQ667195); *Drosophila melanogaster* Dα1 (CAA30172), Dα2 (CAA36517), Dα3 (CAA75688), Dα4 (CAB77445), Dα5 (AAM13390), Dα6 (AAM13392), Dα7 (AAK67257), Dβ1 (CAA27641), Dβ2 (CAA39211), Dβ3 (CAC48166), GluCl (AAG40735), GRD (Q24352), HisCl1 (AAL74413), HisCl2 (AAL74414), LCCH3 (AAB27090), Ntr (NP_651958), pHCl (NP_001034025), RDL (AAA28556), CG6927 (AAF45992), CG7589 (AAF49337), CG8916 (AAF48539), CG11340 (AAF57144), CG12344 (AAF58743); *Caenorhabditis elegans* UNC-49 (AAD42386), GLC-1 (NP_507090) and representatives of the five major *C. elegans* nicotinic acetylcholine receptor groups[61], ACR-16 (P48180), UNC-63 (AAK83056), ACR-8 (NP_509745), UNC-29 (P48181) and DEG-3 (P54244) as well as CUP-4 (AAT42012), which represents the "orphan" subunit group, are shown in brackets. The scale bar represents substitutions per site.

**Transcription factor binding motifs.** Genes with caste-specific patterns of expression were used to search for transcription factor binding motifs. AlignACE, MDscan and MEME algorithms[62-64] were used with 1,000 bp regions upstream of GLEAN3-predicted genes to search for motifs.

**Identification of regulatory motifs involved in the development of behavior.** We scanned a region of 2000 bp upstream of the translation start site for each gene in each bee gene set. We performed a comprehensive scan of promoters for transcription factor binding sites, modeling the binding specificity of each transcription factor by a position-specific weight matrix (PWM). We used the computer algorithm Stubb[65] to score a promoter for matches to PWMs. The Stubb algorithm was previously found to accurately predict *cis*-regulatory modules involved in the segmentation pathway in *Drosophila*[66].

**Venom.** Honey bee venom contains components that are non-allergenic to humans and other vertebrates, as well as at least six allergenic components (phospholipase A2, hyaluronidase, acid phosphatase, melittin, Api m 6 and CUB serine protease). Recently, three novel venom proteins were found: one with a platelet-derived and vascular endothelial growth factor domain, venom protein 2 (not related to any known protein) and a major royal jelly protein family member MRJP9[67]. The annotated genome points toward several other candidate venom allergens (Supplementary Table S13). At least 9 honey bee homologues of allergens from other insect species were found including antigen 5, a venom protein found in several wasps, hornets, fire ants and yellowjackets but has not yet been studied in honey bees. Several homologues of

scorpion and snake venom proteins and peptides were likewise identified: desintegrins, neurotoxins and anticoagulant peptides, all of which have promise in understanding allergic responses and improving prophylactic or therapeutic treatments. Having the honey bee genome sequence available has also help understand the phenomena of *Api m 6* heterogeneity, previously known mainly from the protein realm[68]. It was shown that substantial protein-level variation for *Api m 6* arises from genome-level polymorphism at a single locus.

**Heat Shock Proteins/Chaperones.** Heat-shock protein genes (*hsps*) are nearly universal in organisms, highly-conserved and assigned to families on the basis of sequence homology and typical molecular weight[69-71]. Despite being one of a very few endothermic insects[72] and having extraordinary levels of thermotolerance (they develop in and live in colonies whose temperatures are 33-35°C and even survive 50°C for up to 1 hour; Elekonich and Roberts, unpubl.) honey bees do not show an increase in the number of genes encoding hsp70 family members with 6 in the bee in comparison to 5 in the fly and 8 in humans.

Hsp70 proteins often function in protein complexes with Hsp90 and Hsp40[73]. In both the hsp40 and *hsp90* gene families the honey bee resembles the human more than the fly. Hsp90 family proteins are involved in signal transduction and ligand binding as well as responding during cellular stress. Although there are 4 *hsp90*s in honey bee, only one *hsp90* in *Drosophila,* and 2 in human, there are 5 *hsp90*s in *Anopheles* suggesting that *Drosophila* is unusual for an insect. In the honey bee the *hsp40/dnaj* family comprises 25 genes making it more similar to the human *hsp40* family which has 44[74] than the fly which has 5 family members. There are 39 putative *hsp40*s in *Anopheles* again suggesting that it is the fly that is unusual.

**Nectar and pollen utilization.** Annotation of 174 genes encoding carbohydrate-metabolizing enzymes and 28 genes encoding lipid-metabolizing enzymes, based on orthology to their counterparts in the fly and mosquito, shows the majority of genes have simple, 1:1:1 orthology (*Apis*: *Drosophila*: *Anopheles)*. The noticeable changes in one or more species are more common in enzymes of glycolysis and gluconeogenesis, suggesting the number of genes for carbohydrate metabolism is less conserved than for lipid metabolism[75]. Some enzyme types with particularly striking changes in gene number include acyl-CoA oxidase (2:6:6), acylphosphatase (2:6:2), and pyruvate kinase (2:6:1). Three glycolysis/gluconeogenesis genes have 2:1:1 orthology – pyruvate dehydrogenase, dihydrolipoamide dehydrogenase, and phosphopyruvate hydratase – representing either recent duplications in *Apis* or gene losses in the dipterans.

Two enzymes found in the dipteran species, glucose-6-phosphatase and the monomeric trehalose-6-phosphate phosphatase, appear to be completely missing in the current assembly of the *Apis* genome. As glucose-6-phosphatase activity has been described in honey bee flight muscle[76], it is possible that another phosphatase has shifted its specificity to fill this role. However, if these are true gene losses, bees would be left with a single functional pathway by which to convert gluconeogenic substrates to both of the primary carbohydrate energy stores used by insects, trehalose and glycogen.

Among other metabolic proteins are several unusual carbohydrate-metabolizing enzymes with important roles in honey bee biology. Glucose oxidases contribute antiseptic activity to honey by producing D-gluconic acid and hydrogen peroxide[77]. So far, this is the only known example of a glucose oxidase gene to be found in an animal genome. Phylogenetic analyses suggest that GLOX evolved from a gene encoding glucose dehydrogenase (GLD). There are three glucose dehydrogenase/glucose oxidase genes in the *Apis* genome. These genes are more distantly related to glucose-methanol-choline (GMC) oxido-reductases, a group of FAD

flavoproteins with diverse but poorly understood catalytic activities. There are 21 members of the GMC family in *Apis* compared to 16 in *Drosophila.*

Honey bees may also extract nutrients from pollen grains that are their primary protein source. Contrary to previous beliefs that bees do not have the enzymes needed to digest complex carbohydrates[78] the bee genome contains an active gene encoding cellulase belonging to the GHF9 family of glycoside hydrolases[79].

**Antioxidants.** Aerobic organisms have evolved an elaborated network of enzymatic and nonenzymatic antioxidant systems to prevent oxidative damage. A comparative analysis of honey bee with *Drosophila melanogaster* and *Anopheles gambiae* show that although the basic components of the antioxidant system are conserved, there are important species differences in the number of paralogs. These include the duplication of thioredoxin reductase and the expansion of the thioredoxin family in fly; lack of expansion of the Thetha, Delta and Omega Glutathione S-transferase classes in honey bee and the no expansion of the Sigma class in dipteran species. The increase in the number of Sigma class members in bees, seems to be involved with protection against oxidants produced by aerobic metabolism, rather than xenobiotics. In flies, members of this class are primarily located in the indirect flight muscles[80] and have been reported to have an important role in the detoxification of lipid peroxidation products[81]. Honey bees take foraging trips that may last up to one hour and they carry heavy loads of nectar and pollen during this time[82], so they likely produce a high level free radicals. Perhaps this aspect of their lifestyle exerted selection on these antioxidant genes.

## SNPs and Population Genetics

**Samples.** Sampling is described in detail in Whitfield et al[83]. In total, 175 *A. mellifera* were collected from 14 different geographical subspecies from their native ranges in Africa, Europe and Asia. The 10 subspecies represented in the current study include A. *m. mellifera* (N=20), *A. m. iberiensis* (11), *A. m. ligustica* (18), *A. m. carnica* (16), *A. m. anatoliaca* (18), *A. m. caucasica* (14), *A. m. syriaca* (9), *A. m. scutellata* (21), *A. m. lamarckii* (19) and *A. m. intermissa* (19). All individuals were workers (diploid females) collected from different colonies except for 6 of the *A. m. scutellata* (2 were collected from each of 3 colonies). Sampling of introduced New World bees is as described[83].

**SNP identification.** SNPs were identified using the honey bee genome assembly 3.08 and 2483 shotgun genome traces from North American Africanized bees (genome-derived) or ESTs, including 71861 from mixed domestic bees from Illinois, USA (21,408 from a brain library[84] and 50453 from a whole heads library (from Riken) and 4998 from African hybrid bees from Brazil[85]. ESTs were pre-clustered with assembly scaffolds using BLASTN ($e < 10^{-5}$). Genome-traces and (pre-clustered) ESTs were aligned with assembly scaffolds and scanned for SNPs using POLYBAYES v. 3.0[86], which assigns a probability score (P) that each putative SNP is a true polymorphism rather than a sequencing error (default settings; quality scores were used; prior polymorphism rate = 0.003). For genome-derived SNPs, assembly 3.08 was used as anchor and as template for sequence comparison, and SNPs were identified in both mapped and unmapped scaffolds. For EST-derived SNPs, assembly 3.08 was used as anchor only, and SNPs were identified only in scaffolds mapped to chromosomes. A subset of 1536 SNPs were selected for genotyping based on spacing criteria, SNP probability score (P), "designability" scores for genotyping oligonucleotides (provided by Illumina), and manual inspection of SNP flanking sequence aligned with genome traces (for removal of SNPs with immediately flanking SNPs or indels that might interfere with genotyping assay).

**Genotyping and quality filtering.** SNPs were typed with the Illumina BeadStation 500G using a custom Oligo Pool Assay (OPA) designed to detect the 1536 SNPs selected above[73, 87]. SNP genotypes were generated for a total of 369 *A. mellifera* and 13 related species (*A. cerana, A. dorsata* and *A. florea*). *A. mellifera* samples included 16 drones (haploid males) used in SNP quality assessment but not analyzed in the current study. SNPs were removed from the data set for the following reasons: 194 (12.6%) had good base calls in < 80% of *A. mellifera* samples; 191 (12.4%) appeared to be monomorphic (based on 0 or 1 occurrence of minor allele in all samples); 4 erroneously duplicate SNPs; and 11 that were typed as "heterozygote" in any of the 16 drones (possibly reflecting paralogous rather than polymorphic target sequences). *A. mellifera* samples with < 80% SNP calls were removed from the data set (other *Apis* species were retained irrespective of base call rate). These quality criteria resulted in a final data set consisting of 1136 SNPs analyzed in 328 *A. mellifera* (with mean and minimum call rates of 98.4 and 88.0%, respectively) and 13 individuals from related species (with call rates of 48.6 ± 0.5%, 46.2 ± 0.2%, and 37.6 ± 0.4% for *A. cerana*, *A. dorsata* and *A. florea*, respectively; mean±SE).

**Distance measures and phylogenetic analysis.** $F_{ST}$ values were calculated using Weir and Cockerham's unbiased estimator[88]. Bootstrap and phylogentic analyses were performed using the PHYLIP software package[89]. Population distances were calculated using Nei's genetic distance[90] implemented in the GENDIST function. Population distance tree was generated using the Neighbor-Joining algorithm[91] implemented in the NEIGHBOR function.

References


1. Gibbs, R. A. et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428, 493-521 (2004).
2. Havlak, P. et al. The Atlas genome assembly system. Genome Res 14, 721-32 (2004).
3. Cai, W. W., Chen, R., Gibbs, R. A. & Bradley, A. A clone-array pooled shotgun strategy for sequencing large genomes. Genome Res 11, 1619-23 (2001).
4. Richards, S. et al. Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. Genome Res 15, 1-18 (2005).
5. Beye, M. & Raeder, U. Rapid DNA preparation from bees and %GC fractionation. Biotechniques 14, 372-4 (1993).
6. Solignac, M. et al. The genome of *Apis mellifera*: dialog between mapping and sequencing. Genome Biol Submitted (2006).
7. Wen, S. Y. & Zhang, C. T. Identification of isochore boundaries in the human genome using the technique of wavelet multiresolution analysis. Biochem Biophys Res Commun 311, 215-22 (2003).
8. Zhang, C. T. & Zhang, R. An isochore map of the human genome based on the Z curve method. Gene 317, 127-35 (2003).
9. 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).
10. 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).
11. Consortium, H. B. G. S. Insights into social insects from the genome of the honey bee *Apis mellifera*. Nature In press (2006).
12. Robertson, H. M. & MacLeod, E. G. Five major subfamilies of *mariner* transposable elements in insects, including the Mediterranean fruit fly, and related arthropods. Insect Mol Biol 2, 125-39 (1993).
13. Ebert, P. R., Hileman, J. P. t. & Nguyen, H. T. Primary sequence, copy number, and distribution of *mariner* transposons in the honey bee. Insect Mol Biol 4, 69-78 (1995).
14. Lampe, D. J., Witherspoon, D. J., Soto-Adames, F. N. & Robertson, H. M. Recent horizontal transfer of *mellifera* subfamily *mariner* transposons into insect lineages representing four different orders shows that selection acts only during horizontal transfer. Mol Biol Evol 20, 554-62 (2003).
15. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. Bioinformatics 21 Suppl 1, i351-i358 (2005).
16. Barry, E. G., Witherspoon, D. J. & Lampe, D. J. A bacterial genetic screen identifies functional coding sequences of the insect *mariner* transposable element *Famar1* amplified from the genome of the earwig, *Forficula auricularia*. Genetics 166, 823-33 (2004).
17. Robertson, H. M. in Mobile DNA II (eds. Craig, N. L., Craigie, R., Gellert, M. & Lambowitz, A. M.) (ASM Press, Washington, DC, 2002).
18. Robertson, H. M. & Walden, K. K. *Bmmar6*, a second *mori* subfamily *mariner* transposon from the silkworm moth *Bombyx mori*. Insect Mol Biol 12, 167-71 (2003).

19. Avancini, R. M., Walden, K. K. & Robertson, H. M. The genomes of most animals have multiple members of the *Tc1* family of transposable elements. Genetica 98, 131-40 (1996).
20. Arkhipova, I. R. & Meselson, M. Diverse DNA transposons in rotifers of the class Bdelloidea. Proc Natl Acad Sci U S A 102, 11781-6 (2005).
21. 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).
22. Sarkar, A. et al. Molecular evolutionary analysis of the widespread *piggyBac* transposon family and related "domesticated" sequences. Mol Genet Genomics 270, 173-80 (2003).
23. Kapitonov, V. V. & Jurka, J. Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. Proc Natl Acad Sci U S A 100, 6569-74 (2003).
24. Kaspustin, Y., Souvorov, A. & Tatusova, T. in RECOMB 2004 741 (2004).
25. Kiryutin, B. & Souvorov, A. in ISMB (2005).
26. Souvorov, A., Tatusova, T. & Lipman, D. J. in ISMB 125 (2004).
27. Haas, B. J. et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res 31, 5654-66 (2003).
28. Curwen, V. et al. The Ensembl automatic gene annotation system. Genome Res 14, 942-50 (2004).
29. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. Genome Res 14, 988-95 (2004).
30. Altschul, S. F. & Koonin, E. V. Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases. Trends Biochem Sci 23, 444-7 (1998).
31. Bairoch, A. et al. The Universal Protein Resource (UniProt). Nucleic Acids Res 33, D154-9 (2005).
32. Salamov, A. A. & Solovyev, V. V. Ab initio gene finding in *Drosophila* genomic DNA. Genome Res 10, 516-22 (2000).
33. Hillier, L. W. et al. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432, 695-716 (2004).
34. Zdobnov, E. M. et al. Comparative genome and proteome analysis of *Anopheles* gambiae and *Drosophila melanogaster*. Science 298, 149-59 (2002).
35. Solovyev, V. V. in Handbook of Statistical Genetics (ed. Balding, D. e. a.) 83-127 (John Wiley & Sons, Ltd., 2001).
36. Elsik, C., Mackey, A. J., Reese, J., Milshina, N. & Weinstock, G. M. Creating a honey bee consensus gene set. Genome Res Submitted (2006).
37. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6, 31 (2005).
38. Pertea, G. et al. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics 19, 651-2 (2003).
39. Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 85, 2444-8 (1988).
40. Kaplan, N. & Linial, M. ProtoBee: Hierarchical classification and annotation of the honey bee proteome. Genome Res In press. (2006).
41. Elsik, C. et al. Community annotation: procedures, protocols, and supporting tools. Genome Res In the press (2006).

42. Lewis, S. E. et al. Apollo: a sequence annotation editor. Genome Biol 3, RESEARCH0082 (2002).
43. Tran, T., Havlak, P. & Miller, J. MicroRNA enrichment among short 'ultraconserved' sequences in insects. Nucleic Acids Res 34, e65 (2006).
44. Bentwich, I. et al. Identification of hundreds of conserved and nonconserved human microRNAs. Nat Genet 37, 766-70 (2005).
45. Nuwaysir, E. F. et al. Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. Genome Res 12, 1749-55 (2002).
46. Bertone, P. et al. Global identification of human transcribed sequences with genome tiling arrays. Science 306, 2242-6 (2004).
47. Stolc, V. et al. Identification of transcribed sequences in *Arabidopsis thaliana* by using high-resolution genome tiling arrays. Proc Natl Acad Sci U S A 102, 4453-8 (2005).
48. Panda, S., Hogenesch, J. B. & Kay, S. A. Circadian rhythms from flies to human. Nature 417, 329-35 (2002).
49. Rubin, E. et al. Molecular and phylogenetic analyses reveal mammalian-like clockwork in the honey bee (*Apis mellifera*) and shed new light on the molecular evolution of the circadian clock. Genome Res In press (2006).
50. Sine, S. M. & Engel, A. G. Recent advances in Cys-loop receptor structure and function. Nature 440, 448-55 (2006).
51. Jones, A. K., Raymond-Delpech, V., Thany, S. H., Gauthier, M. & Sattelle, D. B. The nicotinic acetylcholine receptor gene family of the honeybee, *Apis mellifera*. Genome Res In press (2006).
52. Jones, A. K. & Sattelle, D. B. The cys-loop ligand-gated ion channel superfamily of the honeybee, *Apis mellifera*. Invert Neurosci 6, 123-32 (2006).
53. Dacher, M., Lagarrigue, A. & Gauthier, M. Antennal tactile learning in the honeybee: effect of nicotinic antagonists on memory dynamics. Neuroscience 130, 37-50 (2005).
54. Thany, S. H., Crozatier, M., Raymond-Delpech, V., Gauthier, M. & Lenaers, G. *Apisalpha2*, *Apisalpha7-1* and *Apisalpha7-2*: three new neuronal nicotinic acetylcholine receptor alpha-subunits in the honeybee brain. Gene 344, 125-32 (2005).
55. Lozano, V. C., Armengaud, C. & Gauthier, M. Memory impairment induced by cholinergic antagonists injected into the mushroom bodies of the honeybee. J Comp Physiol [A] 187, 249-54 (2001).
56. Thany, S. H. & Gauthier, M. Nicotine injected into the antennal lobes induces a rapid modulation of sucrose threshold and improves short-term memory in the honeybee *Apis mellifera*. Brain Res 1039, 216-9 (2005).
57. El Hassani, A. K., Dacher, M., Gauthier, M. & Armengaud, C. Effects of sublethal doses of fipronil on the behavior of the honeybee (*Apis mellifera*). Pharmacol Biochem Behav 82, 30-9 (2005).
58. Decourtye, A. et al. Comparative sublethal toxicity of nine pesticides on olfactory learning performances of the honeybee *Apis mellifera*. Arch Environ Contam Toxicol 48, 242-50 (2005).
59. Buckingham, S. D., Biggin, P. C., Sattelle, B. M., Brown, L. A. & Sattelle, D. B. Insect GABA receptors: splicing, editing, and targeting by antiparasitics and insecticides. Mol Pharmacol 68, 942-51 (2005).
60. Tomizawa, M. & Casida, J. E. Neonicotinoid insecticide toxicology: mechanisms of selective action. Annu Rev Pharmacol Toxicol 45, 247-68 (2005).

61. Jones, A. K. & Sattelle, D. B. Functional genomics of the nicotinic acetylcholine receptor gene family of the nematode, *Caenorhabditis elegans.* Bioessays 26, 39-49 (2004).
62. Clarke, N. D. & Granek, J. A. Rank order metrics for quantifying the association of sequence features with gene regulation. Bioinformatics 19, 212-8 (2003).
63. Liu, X. S., Brutlag, D. L. & Liu, J. S. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nat Biotechnol 20, 835-9 (2002).
64. Roth, F. P., Hughes, J. D., Estep, P. W. & Church, G. M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotechnol 16, 939-45 (1998).
65. Sinha, S., van Nimwegen, E. & Siggia, E. D. A probabilistic method to detect regulatory modules. Bioinformatics 19 Suppl 1, i292-301 (2003).
66. Sinha, S., Schroeder, M. D., Unnerstall, U., Gaul, U. & Siggia, E. D. Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila. BMC Bioinformatics 5, 129 (2004).
67. Peiren, N. et al. The protein composition of honeybee venom reconsidered by a proteomic approach. Biochim Biophys Acta 1752, 1-5 (2005).
68. Peiren, N., de Graaf, D. C., Evans, J. D. & Jacobs, F. J. Genomic and transcriptional analysis of protein heterogeneity of the honeybee venom allergen Api m 6. Insect Mol Biol Submitted. (2006).
69. Gething, M. J. & Sambrook, J. Protein folding in the cell. Nature 355, 33-45 (1992).
70. Parsell, D. A. & Lindquist, S. The function of heat-shock proteins in stress tolerance: degradation and reactivation of damaged proteins. Annu Rev Genet 27, 437-96 (1993).
71. Morimoto, R. I., Tissieres, A. & Georgopoulos, C. (eds.) Heat shock proteins: structure, function and regulation (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1994).
72. Heinrich, B. The hot-blooded insects, strategies and mechanisms of thermoregulation (Springer, Berlin, 1993).
73. Fan, C. Y., Lee, S. & Cyr, D. M. Mechanisms for regulation of Hsp70 function by Hsp40. Cell Stress Chaperones 8, 309-16 (2003).
74. Venter, J. C. et al. The sequence of the human genome. Science 291, 1304-51 (2001).
75. Kunieda, T. et al. Carbohydrate metabolism genes and pathways in insects: insights from the honey bee genome. Insect Mol Biol In press (2006).
76. Surholt, B. & Newsholme, E. A. Maximum activities and properties of glucose 6-phosphatase in muscles from vertebrates and invertebrates. Biochem J 198, 621-9 (1981).
77. Ohashi, K., Natori, S. & Kubo, T. Expression of amylase and glucose oxidase in the hypopharyngeal gland with an age-dependent role change of the worker honeybee (Apis mellifera L.). Eur J Biochem 265, 127-33 (1999).
78. Grogan, D. E. & Hunt, J. H. Pollen proteases: their potential role in insect digestion. Insect Biochem 9, 309-313 (1979).
79. Tokuda, G. et al. Metazoan *cellulase* genes from termites: intron/exon structures and sites of expression. Biochim Biophys Acta 1447, 146-59 (1999).
80. Franciosa, H. & Berge, J. B. *Glutathione S-transferases* in housefly (*Musca domestica*): location of *GST-1* and *GST-2* families. Insect Biochem Mol Biol 25, 311-7 (1995).

81. Singh, S. P., Coronella, J. A., Benes, H., Cochrane, B. J. & Zimniak, P. Catalytic function of *Drosophila melanogaster glutathione S-transferase DmGSTS1-1* (*GST-2*) in conjugation of lipid peroxidation end products. Eur J Biochem 268, 2912-23 (2001).
82. Winston, M. L. The biology of the honey bee (Harvard University Press, Cambridge, 1987).
83. Whitfield, C. W. et al. Thrice out of Africa: ancient and modern expansions of the honey bee, *Apis mellifera*. Science In press (2006).
84. Whitfield, C. W. et al. Annotated expressed sequence tags and cDNA microarrays for studies of brain and behavior in the honey bee. Genome Res 12, 555-66 (2002).
85. Nunes, F. M. et al. The use of Open Reading frame ESTs (ORESTES) for analysis of the honey bee transcriptome. BMC Genomics 5, 84 (2004).
86. Marth, G. T. et al. A general approach to single-nucleotide polymorphism discovery. Nat Genet 23, 452-6 (1999).
87. Fan, J. B. et al. Highly parallel SNP genotyping. Cold Spring Harb Symp Quant Biol 68, 69-78 (2003).
88. Weir, B. S. & Cocherham, C. C. Estimating *f*-statistics for the analysis of population structure. Evolution 38, 1358-1370 (1984).
89. Felsenstein, J. (Department of Genetics, University of Washington, Seattle, WA, 1993).
90. Nei, M. Genetic distance among populations. American Naturalist 106, 283-292 (1972).
91. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4, 406-25 (1987).
92. Robertson, H. M. & Gordon, K. H. J. Canonical TTAGG repeat telomeres and telomerase in the honey bee, *Apis mellifera*. Genome Res In press (2006).
93. Robertson, H. M. & Wanner, K. The chemoreceptor superfamily in the honey bee *Apis mellifera*: expansion of the odorant, but not gustatory, receptor family. Genome Res In the press (2006).
94. 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).
95. Velarde, R. A., Sauer, C. D., KK, O. W., Fahrbach, S. E. & Robertson, H. M. Pteropsin: A vertebrate-like non-visual opsin expressed in the honey bee brain. Insect Biochem Mol Biol 35, 1367-77 (2005).

**Supplementary Table List**

Table S12: One to One to One Orthologs with Duplications in Honey Bee.

Table S13. Comparative Domains.

Table S14. Homeobox Genes.

Table S15. Candidate new bee venom components.

Table S16. Mean population differentiation ($F_{ST}$) for evolutionary lineages of *Apis mellifera*, based on 1136 SNPs.

Table S17. Access to the Genome Assemblies.

## Table S1. Genome Assemblies.

| Version | Date | Description |
|---|---|---|
| 1.0 | 12/2003 | The first version of the honey bee *A. mellifera* assembly. |
| 1.1 | 1/2004 | This release is an incremental change of the assembly. It assembled bin0 reads, used more markers, and corrected the problems in some files of the previous release. All of the BIN0 reads which clustered into groups of 2 or more reads were assembled and scaffolded with the previous assembly. About 100K BIN0 reads were added to the new assembly and about 30K are still in BIN0. More markers (1050 in this version vs. 854 in Amel_1.0) were used for placing sequences on linkage groups. Inconsistencies in agp and sequence files and mapping problems were corrected. Linkage group numbers were remapped so that they are from 1 to 16. |
| 1.2 | 7/2004 | This release added reads from shotgun sequencing of purified AT rich regions of the genome to the previous WGS reads. Of 171K high quality AT rich reads, 141k were added to the assembly. Addition of these reads increased both the contig and scaffold length, and contributed 10 Mb to the final assembly. This release used the Solignac map, with 1300 markers, to anchor parts of the genome to chromosomes. |
| 2.0 | 1/2005 | This release added reads from shotgun sequencing of purified AT rich regions of the genome, Fosmid clones ends and BAC reads to the previous WGS reads. 480k AT rich reads, 470k BAC reads, and 40k Fosmid clone end reads were added to the assembly. Addition of these reads increased both contig and scaffold length, and contributed 17 Mb to the final assembly. This release used the Solignac map, with 1634 markers, to anchor parts of the genome to chromosomes. |
| 3.0 | 3/2005 | Moderately repetitive sequences were assembled separately and placed using mate pair information and merged with sequence contigs from the version 2.0 assembly into new combined contigs. Identified haplotype contigs were omitted from this assembly. Contaminated |

| | | |
|---|---|---|
| | | regions identified in the version 2.0 assembly were omitted from this assembly. This release used the Solignac map, with 1634 markers, to anchor parts of the genome to chromosomes. |
| 4.0 | 3/2006 | Moderately repetitive sequences were assembled separately and placed using mate pair information and merged with sequence contigs from the version 2.0 assembly into new combined contigs. Highly repeated sequences, low coverage sequences, and contigs with length less than 1 kb were omitted from this assembly and are available as separate data sets on the FTP site. Identified haplotype contigs were omitted from this assembly and will be accessioned and presented with sequence and quality files. This release used the Solignac map with 2013 markers, to anchor parts of the genome to chromosomes. |

The Amel assemblies were produced by assembling whole genome shotgun reads with the Atlas genome assembly system at the Baylor College of Medicine Human Genome Sequencing Center. Several WGS libraries, with inserts of 2-4 kb, 4-6 kb, and about 35 kb, were used to produce the data. About 2.7 million reads were assembled, representing about 1.8 Gb of sequence and about 7.5x coverage of the (clonable) honey bee genome. The products of the Atlas assembler are a set of contigs (contiguous blocks of sequence) and scaffolds. Scaffolds include sequence contigs that can be ordered and oriented with respect to each other as well as isolated contigs that could not be linked (single contig scaffolds or singletons). Reads which did not overlap other sequences were not assembled and are found in the collection of reads called BIN0.

## Table S2. Read Statistics of the honey bee genome assembly, v4.0.

| Read type | WGS | WGS | WGS | AT-rich[1] | BAC[2] | |
|---|---|---|---|---|---|---|
| Insert Size (kb) | 2-4 | 4-6 | 35 | 1-2 | 1-2 | Total |
| Source/Vector | Plasmid | Plasmid | Fosmid | Plasmid | Plasmid | |
| Reads (million) | | | | | | |
| All | 1.12 | 1.23 | 0.11 | 0.85 | | 3.31 |
| Trimmed[3] | 0.95 | 1.01 | 0.08 | 0.73 | 0.47 | 3.23 |
| Paired ends | 0.90 | 0.95 | 0.06 | 0.55 | 0.21 | 2.67 |
| Assembled | 0.87 | 0.94 | 0.07 | 0.62 | 0.47 | 2.97 |
| Unassembled, Bin0[4] | 0.02 | 0.02 | 0.002 | 0.009 | | 0.04 |
| Unassembled, highly repetitive reads[5] | 0.06 | 0.05 | 0.007 | 0.09 | | 0.21 |
| Unassembled, other repetitive reads[6] | 0.009 | 0.004 | 0.0004 | 0.016 | | 0.03 |
| Bases (million) | | | | | | |
| Trimmed | 589.8 | 624.7 | 50.9 | 510.3 | 280.3 | 2056 |
| Assembled | 544.0 | 589.3 | 45.5 | 432.2 | 280.3 | 1891 |
| Unassembled, Bin0[4] | 6.3 | 6.0 | 0.9 | 4.2 | | 17.4 |
| Unassembled, highly repetitive reads[5] | 34.2 | 26.9 | 4.3 | 62.6 | | 128 |
| Unassembled, other repetitive reads[6] | 5.4 | 2.5 | 0.16 | 11.3 | | 19.4 |
| | | | | | | |
| Sequence Coverage[7] | 2.5x | 2.6x | 0.2x | 2.2x | | 7.6x |
| Clone Coverage[8] | 5.6x | 10.4x | 5.0x | 2.1x | | 23.1x |

[1] WGS reads prepared from AT-rich DNA isolated by density gradient centrifugation as described[1].

[2] WGS reads prepared from pooled BAC DNA. The reads were treated *en masse* as if they were genomic WGS reads and not from individual BAC clones.

[3] Removal of low base quality or very short reads.

[4] During assembly reads that share sequence overlaps are placed into "bins." Reads that have no overlaps are placed in Bin0 – these are often reads with poor sequence quality or contaminants. They are not assembled since they do not overlap other reads. A small fraction of the Bin0 reads represent regions that are simply low coverage due to sampling statistics or difficulty in cloning, such as the AT-rich regions of the honey bee genome. Other reads that are not assembled are highly repeated sequences or sequences giving inconsistent placement.

[5] Highly repetitive reads are those containing high copy number repeated sequences that are not included in the assembly either by read pair information or as part of the stringent assembly of repeats into reptigs.

[6] Other repeats are those that are not high copy number but fail to assemble into reptigs or by read pair information.

[7] Sequence coverage was calculated as total trimmed bases divided by calculated genome size. Genome size was calculated as total contig length divided by the fraction of markers hit by all contigs(231M/99%=233M).

[8] Clone coverage was calculated as sum of insert sizes for paired reads divided by the calculated genome size.

**Table S3. Scaffold, contig, and completeness statistics for assembly v4.**

## A. Scaffolds

| Scaffolds/Contigs | Number | N50(kb) | Bases+Gaps (Mb) | Bases(Mb) | Total Scaffold Length (%) |
|---|---|---|---|---|---|
| Anchored & Oriented | 320 | 621 | 152 | 150 | 64.7 |
| Anchored Not Oriented | 306 | 135 | 34 | 33 | 14.6 |
| Unanchored Scaffolds | 9,244 | 16 | 49 | 48 | 20.6 |
| All Scaffolds | 9,870 | 362 | 235 | 231 | 100 |
| All Contigs | 18,072 | 41 | 231 | 231 | 98.3 |

## B. Contigs

| | Mapped, Total | Mapped, oriented | Mapped, unoriented | Unmapped |
|---|---|---|---|---|
| Total Contig Number | 6,596 | 5,136 | 1,460 | 11,476 |
| Total Contig Length, Mb | 183.3 | 149.6 | 33.7 | 47.7 |
| Avg Contig Length, bp | 27,793 | 29,133 | 23,078 | 4,157 |
| N50 Contig Length, bp | 52,503 | 54,923 | 41,860 | 9,001 |
| % A+T | 65 | 64 | 70 | 76 |
| No. Repeat Regions* | 1,571 | 1,261 | 310 | 3,684 |
| Total Length Repeat Regions, Mb | 1.7 | 1.3 | 0.4 | 5.9 |
| Repeats as % of Total | 0.9 | 0.9 | 1.1 | 12.4 |

* either as reptig (a contig formed by stringent assembly of repeats) or merged in contig (e.g. due to read pairing for placement).

## C. Completeness

| | Assembly(%) | Bin0(%) | Assembly+Bin0(%) |
|---|---|---|---|
| Markers (2,013) | 99 | 23 | 99 |
| EST (3,136) | 98 | 8 | 98 |
| cDNA (57) | 96 | 30 | 96 |

## D. Completeness of 27 BACs

| **Clone name** | **HGSC project name** | **reads attempted** | total Q20 bases | **cov for 165kb** | **Accession** | cov of assembled contigs | # of ctgs | # of high quality ctgs | # of bases in high quality ctgs | # of bases covered in asm | base covered percentage |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CH224-57G9 | AMFD | 2280 | 1209724.6 | 7.331664242 | AC141723 | 8.183435931 | 4 | 4 | 147526 | 146210 | 99% |
| CH224-59E21 | AMEB | 2280 | 1299532.5 | 7.875954545 | AC141751 | 8.216101132 | 5 | 5 | 157769 | 153066 | 97% |
| CH224-63A6 | AMDD | 1520 | 886960.2 | 5.375516364 | AC141775 | 8.475167695 | 7 | 7 | 104054 | 101538 | 98% |
| CH224-58F16 | AMEP | 2280 | 1102015.2 | 6.67888 | AC141737 | 6.836068136 | 8 | 6 | 158327 | 157088 | 99% |
| CH224-62B14 | AMBB | 2280 | 1255890.4 | 7.61145697 | AC141822 | 7.033712119 | 9 | 9 | 177753 | 172041 | 97% |
| CH224-60D20 | AMDW | 2280 | 1179258.1 | 7.147018788 | AC141754 | 7.288774406 | 9 | 9 | 160991 | 154802 | 96% |
| CH224-58F1 | AMEF | 2280 | 1256070.6 | 7.612549091 | AC141747 | 8.711822722 | 9 | 9 | 143380 | 136677 | 95% |
| CH224-56H7 | AMFW | 1710 | 945769.5 | 5.731936364 | AC141706 | 5.391580586 | 10 | 10 | 174516 | 173059 | 99% |
| CH224-55I1 | AMGK | 2280 | 1217914 | 7.38129697 | AC141692 | 7.753315126 | 10 | 10 | 156183 | 148594 | 95% |
| CH224-58F11 | AMEK | 1710 | 965304 | 5.850327273 | AC141740 | 6.386693396 | 10 | 10 | 150243 | 144583 | 96% |
| CH224-58F15 | AMEO | 2090 | 1098812 | 6.659466667 | AC141736 | 6.594164456 | 11 | 11 | 165634 | 163711 | 99% |
| CH224-58F3 | AMEG | 2280 | 1227715.5 | 7.4407 | AC141744 | 7.193336419 | 12 | 12 | 169574 | 163230 | 96% |
| CH224-60D19 | AMDV | 2280 | 1205360.2 | 7.305213333 | AC141757 | 7.093650579 | 12 | 10 | 166385 | 156892 | 94% |
| CH224-59E23 | AMED | 2184 | 1182490.4 | 7.166608485 | AC141749 | 6.136909464 | 14 | 14 | 191385 | 184545 | 96% |
| CH224-63A10 | AMDH | 1520 | 877001 | 5.315157576 | AC141771 | 4.965918099 | 14 | 14 | 175304 | 170326 | 97% |
| CH224-60D5 | AMCJ | 1900 | 864328.5 | 5.238354545 | AC141792 | 4.895908033 | 15 | 15 | 175141 | 170321 | 97% |
| CH224-58F12 | AMEL | 1520 | 864309.6 | 5.23824 | AC141741 | 5.086208601 | 15 | 15 | 168532 | 163355 | 97% |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CH224-61C17 | AMBY | 2090 | 1018145.6 | 6.170579394 | AC141803 | 5.320465709 | 17 | 17 | 189764 | 180244 | 95% |
| CH224-58F6 | AMEH | 2090 | 1127941.5 | 6.836009091 | AC141745 | 6.119075907 | 17 | 17 | 182732 | 167743 | 92% |
| CH224-54J17 | AMHV | 1520 | 864939.6 | 5.242058182 | AC141662 | 4.753799733 | 17 | 17 | 180347 | 172356 | 96% |
| CH224-56H8 | AMFX | 2090 | 891854 | 5.405175758 | AC141707 | 5.318849223 | 24 | 20 | 160820 | 142529 | 89% |
| CH224-61C5 | AMBM | 2280 | 1165036 | 7.060824242 | AC141812 | 6.953819708 | 24 | 8 | 147252 | 144340 | 98% |
| CH224-61C10 | AMBR | 2660 | 1478875.2 | 8.96288 | AC141808 | 8.231567581 | 25 | 24 | 175827 | 149733 | 85% |
| CH224-62B10 | AMAX | 1900 | 917461.6 | 5.560373333 | AC141826 | 5.508460268 | 26 | 13 | 148154 | 143230 | 97% |
| CH224-57G4 | AMFA | 6080 | 3915801 | 23.73212727 | AC141724 | 30.75630906 | 45 | 45 | 122917 | 106793 | 87% |
| CH224-60D3 | AMCI | 1615 | 982131 | 5.952309091 | AC141795 | 4.954652313 | 50 | 50 | 193324 | 184996 | 96% |
| CH224-60D7 | AMDL | 2470 | 1061479.9 | 6.433211515 | AC141767 | 5.091226042 | 51 | 48 | 198566 | 159696 | 80% |

## E. Statistics of 187 BACs

| HGSC project name | reads attempted | pass_qual | pass_screen | avg_P20 | total Q20 bases | cov for 165kb | Accession | Clone name | # ctgs | # bases | cov of assembled contigs |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AMBL | 2850 | 2436 | 2240 | 577 | 1292480 | 7.833212121 | AC141815 | CH224-61C4 | 94 | 409619 | 3.155322385 |
| AMEU | 1900 | 1768 | 1587 | 621.2 | 985844.4 | 5.974814545 | AC141730 | CH224-58F21 | 90 | 274341 | 3.59350006 |
| AMCG | 2470 | 2179 | 1981 | 638.6 | 1265066.6 | 7.667070303 | AC141797 | CH224-60D1 | 58 | 268255 | 4.715910607 |
| AMFD | 2280 | 2128 | 2026 | 597.1 | 1209724.6 | 7.331664242 | AC141723 | CH224-57G9 | 4 | 147826 | 8.183435931 |
| AMEB | 2280 | 2145 | 2007 | 647.5 | 1299532.5 | 7.875954545 | AC141751 | CH224-59E21 | 5 | 158169 | 8.216101132 |
| AMDD | 1520 | 1433 | 1338 | 662.9 | 886960.2 | 5.375516364 | AC141775 | CH224-63A6 | 7 | 104654 | 8.475167695 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AMEP | 2280 | 1783 | 1716 | 642.2 | 1102015.2 | 6.67888 | AC141737 | CH224-58F16 | 8 | 161206 | 6.836068136 |
| AMBB | 2280 | 2112 | 2012 | 624.2 | 1255890.4 | 7.61145697 | AC141822 | CH224-62B14 | 9 | 178553 | 7.033712119 |
| AMDW | 2280 | 2127 | 1993 | 591.7 | 1179258.1 | 7.147018788 | AC141754 | CH224-60D20 | 9 | 161791 | 7.288774406 |
| AMEF | 2280 | 2129 | 2011 | 624.6 | 1256070.6 | 7.612549091 | AC141747 | CH224-58F1 | 9 | 144180 | 8.711822722 |
| AMFW | 1710 | 1646 | 1585 | 596.7 | 945769.5 | 5.731936364 | AC141706 | CH224-56H7 | 10 | 175416 | 5.391580586 |
| AMGK | 2280 | 2061 | 1921 | 634 | 1217914 | 7.38129697 | AC141692 | CH224-55I1 | 10 | 157083 | 7.753315126 |
| AMEK | 1710 | 1519 | 1476 | 654 | 965304 | 5.850327273 | AC141740 | CH224-58F11 | 10 | 151143 | 6.386693396 |
| AMEO | 2090 | 1957 | 1870 | 587.6 | 1098812 | 6.659466667 | AC141736 | CH224-58F15 | 11 | 166634 | 6.594164456 |
| AMEG | 2280 | 2142 | 2035 | 603.3 | 1227715.5 | 7.4407 | AC141744 | CH224-58F3 | 12 | 170674 | 7.193336419 |
| AMDV | 2280 | 2049 | 1903 | 633.4 | 1205360.2 | 7.305213333 | AC141757 | CH224-60D19 | 12 | 169921 | 7.093650579 |
| AMED | 2184 | 2039 | 1924 | 614.6 | 1182490.4 | 7.166608485 | AC141749 | CH224-59E23 | 14 | 192685 | 6.136909464 |
| AMDH | 1520 | 1459 | 1405 | 624.2 | 877001 | 5.315157576 | AC141771 | CH224-63A10 | 14 | 176604 | 4.965918099 |
| AMCJ | 1900 | 1528 | 1449 | 596.5 | 864328.5 | 5.238354545 | AC141792 | CH224-60D5 | 15 | 176541 | 4.895908033 |
| AMEL | 1520 | 1445 | 1386 | 623.6 | 864309.6 | 5.23824 | AC141741 | CH224-58F12 | 15 | 169932 | 5.086208601 |
| AMBY | 2090 | 1905 | 1679 | 606.4 | 1018145.6 | 6.170579394 | AC141803 | CH224-61C17 | 17 | 191364 | 5.320465709 |
| AMEH | 2090 | 2002 | 1869 | 603.5 | 1127941.5 | 6.836009091 | AC141745 | CH224-58F6 | 17 | 184332 | 6.119075907 |
| AMHV | 1520 | 1436 | 1346 | 642.6 | 864939.6 | 5.242058182 | AC141662 | CH224-54J17 | 17 | 181947 | 4.753799733 |
| AMFX | 2090 | 1746 | 1543 | 578 | 891854 | 5.405175758 | AC141707 | CH224-56H8 | 24 | 167678 | 5.318849223 |
| AMBM | 2280 | 2030 | 1880 | 619.7 | 1165036 | 7.060824242 | AC141812 | CH224-61C5 | 24 | 167539 | 6.953819708 |
| AMBR | 2660 | 2518 | 2271 | 651.2 | 1478875.2 | 8.96288 | AC141808 | CH224-61C10 | 25 | 179659 | 8.231567581 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AMAX | 1900 | 1580 | 1522 | 602.8 | 917461.6 | 5.560373333 | AC141826 | CH224-62B10 | 26 | 166555 | 5.508460268 |
| AMFA | 6080 | 5598 | 5598 | 699.5 | 3915801 | 23.73212727 | AC141724 | CH224-57G4 | 45 | 127317 | 30.75630906 |
| AMCI | 1615 | 1486 | 1329 | 739 | 982131 | 5.952309091 | AC141795 | CH224-60D3 | 50 | 198224 | 4.954652313 |
| AMDL | 2470 | 2090 | 1703 | 623.3 | 1061479.9 | 6.433211515 | AC141767 | CH224-60D7 | 51 | 208492 | 5.091226042 |
| AMHX | 1520 | 1415 | 1329 | 642.1 | 853350.9 | 5.171823636 | | | | | |
| AMCH | 1615 | 1486 | 1341 | 634.7 | 851132.7 | 5.15838 | | | | | |
| AMEY | 1520 | 1471 | 1293 | 649.3 | 839544.9 | 5.088150909 | | | | | |
| AMDM | 1520 | 1411 | 1316 | 637.4 | 838818.4 | 5.083747879 | | | | | |
| AMFU | 1520 | 1425 | 1331 | 628.2 | 836134.2 | 5.06748 | | | | | |
| AMBD | 1520 | 1432 | 1284 | 648.2 | 832288.8 | 5.044174545 | | | | | |
| AMDY | 1520 | 1446 | 1324 | 627.8 | 831207.2 | 5.037619394 | | | | | |
| AMDO | 1520 | 1388 | 1348 | 614.9 | 828885.2 | 5.023546667 | | | | | |
| AMDX | 1520 | 1414 | 1343 | 613.2 | 823527.6 | 4.991076364 | | | | | |
| AMAI | 1520 | 1424 | 1328 | 616.7 | 818977.6 | 4.963500606 | | | | | |
| AMEC | 1900 | 1673 | 1299 | 628.6 | 816551.4 | 4.948796364 | | | | | |
| AMCX | 1520 | 1404 | 1365 | 589.1 | 804121.5 | 4.873463636 | | | | | |
| AMBJ | 1520 | 1428 | 1233 | 649.6 | 800956.8 | 4.854283636 | | | | | |
| AMDJ | 1520 | 1403 | 1271 | 629.3 | 799840.3 | 4.84751697 | | | | | |
| AMGI | 1330 | 1235 | 1179 | 667.8 | 787336.2 | 4.771734545 | | | | | |
| AMDZ | 1520 | 1462 | 1130 | 696.4 | 786932 | 4.769284848 | | | | | |
| AMBH | 1520 | 1342 | 1296 | 606 | 785376 | 4.759854545 | | | | | |
| AMCO | 1520 | 1402 | 1263 | 620.9 | 784196.7 | 4.752707273 | | | | | |
| AMHN | 1520 | 1437 | 1306 | 599.6 | 783077.6 | 4.745924848 | | | | | |
| AMFP | 1520 | 1387 | 1294 | 601.7 | 778599.8 | 4.718786667 | | | | | |
| AMDT | 1520 | 1370 | 1186 | 653.2 | 774695.2 | 4.695122424 | | | | | |
| AMEJ | 1520 | 1394 | 1339 | 576.1 | 771397.9 | 4.675138788 | | | | | |
| AMAL | 1616 | 1377 | 1332 | 570 | 759240 | 4.601454545 | | | | | |
| AMCB | 1520 | 1304 | 1150 | 657.2 | 755780 | 4.580484848 | | | | | |
| AMDP | 1330 | 1241 | 1183 | 638.6 | 755463.8 | 4.578568485 | | | | | |
| AMDQ | 1710 | 1284 | 1218 | 618.9 | 753820.2 | 4.568607273 | | | | | |
| AMHS | 1520 | 1351 | 1270 | 589.1 | 748157 | 4.534284848 | | | | | |
| AMCL | 1330 | 1226 | 1168 | 636.8 | 743782.4 | 4.507772121 | | | | | |
| AMEE | 1520 | 1334 | 1285 | 576.4 | 740674 | 4.488933333 | | | | | |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AMFM | 1330 | 1235 | 1159 | 637.6 | 738978.4 | 4.47865697 | | | | | |
| AMFV | 1520 | 1382 | 1262 | 579.5 | 731329 | 4.43229697 | | | | | |
| AMGA | 1330 | 1267 | 1130 | 644.8 | 728624 | 4.41590303 | | | | | |
| AMBU | 1520 | 1347 | 1228 | 592.6 | 727712.8 | 4.410380606 | | | | | |
| AMDK | 1520 | 1420 | 1244 | 578.8 | 720027.2 | 4.363801212 | | | | | |
| AMDG | 1520 | 1438 | 1215 | 572.8 | 695952 | 4.217890909 | | | | | |
| AMBS | 1520 | 1290 | 1192 | 580 | 691360 | 4.190060606 | | | | | |
| AMEW | 1330 | 1252 | 1143 | 604 | 690372 | 4.184072727 | | | | | |
| AMCK | 1330 | 1224 | 1107 | 618.9 | 685122.3 | 4.152256364 | | | | | |
| AMHE | 1140 | 1092 | 1061 | 642.4 | 681586.4 | 4.130826667 | | | | | |
| AMET | 1234 | 1090 | 1010 | 673.4 | 680134 | 4.122024242 | | | | | |
| AMCW | 1330 | 1239 | 1125 | 602.9 | 678262.5 | 4.110681818 | | | | | |
| AMEA | 1804 | 1285 | 1232 | 550.3 | 677969.6 | 4.108906667 | | | | | |
| AMEX | 1330 | 1120 | 1027 | 643.3 | 660669.1 | 4.004055152 | | | | | |
| AMCY | 1140 | 1078 | 980 | 657.9 | 644742 | 3.907527273 | | | | | |
| AMGJ | 1330 | 1186 | 1008 | 631.9 | 636955.2 | 3.860334545 | | | | | |
| AMCS | 1140 | 1076 | 995 | 639.6 | 636402 | 3.856981818 | | | | | |
| AMAE | 1520 | 1120 | 1065 | 594 | 632610 | 3.834 | | | | | |
| AMFO | 1520 | 1246 | 1125 | 561.8 | 632025 | 3.830454545 | | | | | |
| AMDB | 1140 | 1073 | 1033 | 609 | 629097 | 3.812709091 | | | | | |
| AMFT | 1140 | 1057 | 970 | 647.9 | 628463 | 3.808866667 | | | | | |
| AMCA | 1424 | 1129 | 972 | 644.6 | 626551.2 | 3.79728 | | | | | |
| AMAG | 1140 | 1048 | 1006 | 619 | 622714 | 3.774024242 | | | | | |
| AMDU | 1140 | 1020 | 985 | 625.8 | 616413 | 3.735836364 | | | | | |
| AMBK | 1520 | 1225 | 1127 | 545.3 | 614553.1 | 3.724564242 | | | | | |
| AMEV | 1330 | 1155 | 1079 | 567.4 | 612224.6 | 3.710452121 | | | | | |
| AMGE | 1330 | 1155 | 1063 | 572.6 | 608673.8 | 3.688932121 | | | | | |
| AMAY | 1330 | 1094 | 960 | 602.9 | 578784 | 3.507781818 | | | | | |
| AMEN | 1140 | 1033 | 986 | 578.7 | 570598.2 | 3.458170909 | | | | | |
| AMGF | 1140 | 1036 | 997 | 557.5 | 555827.5 | 3.368651515 | | | | | |
| AMFY | 1140 | 1051 | 937 | 592.4 | 555078.8 | 3.364113939 | | | | | |
| AMHU | 950 | 889 | 846 | 647.4 | 547700.4 | 3.319396364 | | | | | |
| AMEZ | 1140 | 1001 | 932 | 585.2 | 545406.4 | 3.305493333 | | | | | |
| AMBO | 1140 | 1008 | 931 | 583.5 | 543238.5 | 3.292354545 | | | | | |
| AMCT | 950 | 847 | 816 | 646.9 | 527870.4 | 3.199214545 | | | | | |
| AMFH | 1330 | 1139 | 985 | 531.4 | 523429 | 3.17229697 | | | | | |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AMCR | 1520 | 1084 | 1059 | 470.2 | 497941.8 | 3.017829091 | | | | | |
| AMFB | 760 | 743 | 706 | 675.9 | 477185.4 | 2.892032727 | | | | | |
| AMAM | 760 | 735 | 698 | 678.9 | 473872.2 | 2.871952727 | | | | | |
| AMAQ | 950 | 847 | 779 | 592.9 | 461869.1 | 2.799206667 | | | | | |
| AMHG | 1046 | 850 | 757 | 603.8 | 457076.6 | 2.770161212 | | | | | |
| AMHO | 760 | 700 | 669 | 672.3 | 449768.7 | 2.725870909 | | | | | |
| AMFF | 2090 | 1086 | 1003 | 446.8 | 448140.4 | 2.716002424 | | | | | |
| AMHB | 760 | 739 | 707 | 616.9 | 436148.3 | 2.64332303 | | | | | |
| AMHQ | 760 | 737 | 710 | 613.7 | 435727 | 2.640769697 | | | | | |
| AMGU | 760 | 710 | 688 | 628.7 | 432545.6 | 2.621488485 | | | | | |
| AMAT | 760 | 688 | 634 | 677 | 429218 | 2.601321212 | | | | | |
| AMCN | 760 | 678 | 643 | 667.5 | 429202.5 | 2.601227273 | | | | | |
| AMGC | 950 | 808 | 776 | 552.6 | 428817.6 | 2.598894545 | | | | | |
| AMGP | 760 | 714 | 684 | 625.5 | 427842 | 2.592981818 | | | | | |
| AMES | 760 | 725 | 658 | 648.3 | 426581.4 | 2.585341818 | | | | | |
| AMAF | 760 | 699 | 664 | 637.7 | 423432.8 | 2.566259394 | | | | | |
| AMGW | 760 | 716 | 690 | 601.6 | 415104 | 2.515781818 | | | | | |
| AMHW | 760 | 703 | 672 | 616.5 | 414288 | 2.510836364 | | | | | |
| AMBW | 1520 | 839 | 760 | 544.4 | 413744 | 2.507539394 | | | | | |
| AMCQ | 760 | 701 | 647 | 633.2 | 409680.4 | 2.482911515 | | | | | |
| AMAR | 760 | 706 | 673 | 608.1 | 409251.3 | 2.480310909 | | | | | |
| AMHL | 760 | 733 | 625 | 653.3 | 408312.5 | 2.474621212 | | | | | |
| AMGR | 760 | 699 | 646 | 630.9 | 407561.4 | 2.470069091 | | | | | |
| AMFS | 760 | 702 | 657 | 619.9 | 407274.3 | 2.468329091 | | | | | |
| AMAN | 760 | 726 | 671 | 606.3 | 406827.3 | 2.46562 | | | | | |
| AMGY | 760 | 709 | 670 | 607 | 406690 | 2.464787879 | | | | | |
| AMDS | 760 | 705 | 591 | 687.7 | 406430.7 | 2.463216364 | | | | | |
| AMFI | 760 | 711 | 679 | 597.9 | 405974.1 | 2.460449091 | | | | | |
| AMIB | 760 | 721 | 687 | 587.2 | 403406.4 | 2.444887273 | | | | | |
| AMCM | 760 | 694 | 640 | 627.7 | 401728 | 2.434715152 | | | | | |
| AMBA | 1140 | 743 | 698 | 572.3 | 399465.4 | 2.421002424 | | | | | |
| AMGM | 760 | 669 | 630 | 631.9 | 398097 | 2.412709091 | | | | | |
| AMGZ | 760 | 719 | 644 | 616.7 | 397154.8 | 2.406998788 | | | | | |
| AMEI | 760 | 708 | 654 | 605.8 | 396193.2 | 2.401170909 | | | | | |
| AMIA | 760 | 702 | 645 | 613.7 | 395836.5 | 2.399009091 | | | | | |
| AMHR | 760 | 712 | 644 | 614.2 | 395544.8 | 2.397241212 | | | | | |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AMHA | 760 | 704 | 652 | 600.7 | 391656.4 | 2.373675152 | | | | | |
| AMGB | 760 | 728 | 666 | 585.8 | 390142.8 | 2.364501818 | | | | | |
| AMGT | 760 | 704 | 625 | 622.7 | 389187.5 | 2.358712121 | | | | | |
| AMCZ | 760 | 670 | 652 | 592.4 | 386244.8 | 2.340877576 | | | | | |
| AMGX | 760 | 683 | 623 | 605.9 | 377475.7 | 2.287731515 | | | | | |
| AMAZ | 760 | 723 | 680 | 553.9 | 376652 | 2.282739394 | | | | | |
| AMBT | 760 | 670 | 610 | 613.8 | 374418 | 2.2692 | | | | | |
| AMDN | 760 | 707 | 592 | 627.5 | 371480 | 2.251393939 | | | | | |
| AMGD | 570 | 558 | 531 | 699.3 | 371328.3 | 2.250474545 | | | | | |
| AMHI | 760 | 671 | 633 | 583.1 | 369102.3 | 2.236983636 | | | | | |
| AMCU | 1140 | 642 | 595 | 614.8 | 365806 | 2.217006061 | | | | | |
| AMAH | 760 | 661 | 614 | 580.7 | 356549.8 | 2.160907879 | | | | | |
| AMEQ | 760 | 673 | 612 | 579 | 354348 | 2.147563636 | | | | | |
| AMAS | 760 | 687 | 616 | 572.6 | 352721.6 | 2.137706667 | | | | | |
| AMDE | 760 | 704 | 656 | 535.2 | 351091.2 | 2.127825455 | | | | | |
| AMHH | 760 | 667 | 624 | 561 | 350064 | 2.1216 | | | | | |
| AMFG | 760 | 675 | 554 | 621.2 | 344144.8 | 2.085726061 | | | | | |
| AMCP | 570 | 534 | 506 | 678.3 | 343219.8 | 2.08012 | | | | | |
| AMFK | 760 | 642 | 604 | 564.3 | 340837.2 | 2.06568 | | | | | |
| AMCF | 760 | 664 | 613 | 555.7 | 340644.1 | 2.064509697 | | | | | |
| AMAV | 760 | 642 | 551 | 617.5 | 340242.5 | 2.062075758 | | | | | |
| AMHP | 570 | 549 | 505 | 662.5 | 334562.5 | 2.027651515 | | | | | |
| AMGO | 570 | 548 | 507 | 654.6 | 331882.2 | 2.011407273 | | | | | |
| AMFQ | 570 | 536 | 499 | 657.8 | 328242.2 | 1.989346667 | | | | | |
| AMFA | 760 | 638 | 571 | 572.9 | 327125.9 | 1.982581212 | | | | | |
| AMEM | 570 | 526 | 512 | 612.3 | 313497.6 | 1.899985455 | | | | | |
| AMCV | 1520 | 704 | 656 | 475 | 311600 | 1.888484848 | | | | | |
| AMHD | 570 | 522 | 472 | 653.8 | 308593.6 | 1.870264242 | | | | | |
| AMCE | 570 | 554 | 477 | 643.2 | 306806.4 | 1.859432727 | | | | | |
| AMBE | 570 | 527 | 498 | 614.3 | 305921.4 | 1.854069091 | | | | | |
| AMGH | 760 | 644 | 584 | 523.3 | 305607.2 | 1.852164848 | | | | | |
| AMAB | 760 | 650 | 589 | 518.6 | 305455.4 | 1.851244848 | | | | | |
| AMBF | 760 | 628 | 572 | 529.2 | 302702.4 | 1.83456 | | | | | |
| AMGS | 570 | 516 | 483 | 605.3 | 292359.9 | 1.771878182 | | | | | |
| AMBN | 570 | 536 | 514 | 562.7 | 289227.8 | 1.752895758 | | | | | |
| AMBX | 760 | 552 | 534 | 541.3 | 289054.2 | 1.751843636 | | | | | |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AMBG | 570 | 513 | 453 | 632.5 | 286522.5 | 1.7365 | | | | | |
| AMHJ | 760 | 657 | 619 | 457.7 | 283316.3 | 1.717068485 | | | | | |
| AMAW | 760 | 613 | 492 | 568.7 | 279800.4 | 1.69576 | | | | | |
| AMFZ | 760 | 619 | 585 | 466 | 272610 | 1.652181818 | | | | | |
| AMDR | 760 | 595 | 539 | 501.7 | 270416.3 | 1.638886667 | | | | | |
| AMAU | 760 | 505 | 453 | 541.1 | 245118.3 | 1.485565455 | | | | | |
| AMDA | 380 | 368 | 348 | 693.5 | 241338 | 1.462654545 | | | | | |
| AMFN | 380 | 369 | 343 | 682.3 | 234028.9 | 1.41835697 | | | | | |
| AMAJ | 380 | 359 | 340 | 646.2 | 219708 | 1.331563636 | | | | | |
| AMHC | 380 | 357 | 318 | 686.1 | 218179.8 | 1.322301818 | | | | | |
| AMBI | 380 | 356 | 326 | 662 | 215812 | 1.307951515 | | | | | |
| AMHM | 380 | 349 | 329 | 643.8 | 211810.2 | 1.283698182 | | | | | |
| AMDF | 380 | 361 | 336 | 626.6 | 210537.6 | 1.275985455 | | | | | |
| AMFC | 380 | 361 | 330 | 636.5 | 210045 | 1.273 | | | | | |
| AMDI | 760 | 466 | 392 | 486.1 | 190551.2 | 1.154855758 | | | | | |
| AMCD | 760 | 270 | 239 | 563.2 | 134604.8 | 0.815786667 | | | | | |
| AMBV | 760 | 308 | 298 | 417.7 | 124474.6 | 0.754391515 | | | | | |
| AMAO | 190 | 184 | 171 | 635.4 | 108653.4 | 0.658505455 | | | | | |
| AMER | 190 | 179 | 165 | 610.1 | 100666.5 | 0.6101 | | | | | |
| AMGG | 380 | 265 | 253 | 396.7 | 100365.1 | 0.608273333 | | | | | |

## Table S4. A. Honey bee chromosome structure.

| Chromosome | Length (µm) | Arm Ratio | Paracentromeric AT-Rich Band Ratio (%) [a] | Type [b] |
|---|---|---|---|---|
| 1 | 3.48 | 1.16 | 18.85 | Metacentric |
| 2 | 2.49 | 3.18 | 29.36 | Subtelocentric |
| 3 | 2.28 | 4.61 | 27.46 | Subtelocentric |
| 4 | 2.06 | 2.77 | 34.30 | Subtelocentric |
| 5 | 2.16 | 4.12 | 30.22 | Subtelocentric |
| 6 | 1.86 | 2.40 | 38.79 | Submetacentric |
| 7 | 1.91 | 3.78 | 33.06 | Subtelocentric |
| 8 | 1.74 | 2.87 | 32.19 | Subtelocentric |
| 9 | 1.75 | 3.57 | 44.20 | Subtelocentric |
| 10 | 1.60 | 1.99 | 47.39 | Submetacentric |
| 11 | 1.67 | 2.64 | 36.32 | Submetacentric |
| 12 | 1.58 | 2.33 | 44.33 | Submetacentric |
| 13 | 1.59 | 2.32 | 44.85 | Submetacentric |
| 14 | 1.44 | 2.22 | 43.09 | Submetacentric |
| 15 | 1.33 | 1.88 | 52.20 | Submetacentric |
| 16 | 1.16 | 1.92 | 60.00 | Submetacentric |
| Total | 30.09 | 2.41 | 35.84 | |

[a] Proportion of total chromosome length in DAPI bright (heterochromatic, dense DNA and/or higher AT content) bands on both sides of the centromere.

[b] Classification based on Levan indexes[2].

The stages examined to produce the different designations are described in a companion paper. Only the indexes values for the chromosomes 3, 5, 7 and 9 are close to the telocentric range values.

**B. Mapped BAC clones.**

| Locus # | Accession # | Clones | NCBI linkage group | Map position in the current NCBI http://www.ncbi.nlm.nih.gov/mapview/ | |
|---|---|---|---|---|---|
| | | | | Position using Locus # | Position using Acces # |
| Ac005 | AJ509634 | 1F2 | 1 | 0.8676 | 0.757 and 0.86765 |
| Ac012 | AJ509638 | 1C6 | 2 | 0.5714 | 0.4167 |
| Ac032 | AJ509641 | 2E1 | 6 | 0.5428 | 0.5428 and 0.8285 |
| Ac045 | AJ509644 | 2D6 | 6 | 0.6857 | 0.657 |
| Ac140 | AJ509682 | 5E2 | 15 | 0.5 | 0.5 |
| Ac179 | AJ509698 | 6B9 | 11 | 0.8387 | 0.80645 |
| Ac184 | AJ509700 | 6G8 | 3 | 0.5758 | 0.6061 |
| Ac191 | AJ509701 | 6D11 | 1 | 0.86765 | 0.8971 |
| Ac194 | AJ509703 | 6H12 | 1 | 0.2058 | 0.2058 |
| Ac216 | AJ509711 | 8H7 | 10 | 0.3333 | 0.4074 |
| Ac303 | AJ509719 | 49H2 | (1) | No Match | No Match |
| Ag005a | AJ509722 | 56F6 | 1 | 0.3971 | 0.4117 |
| Av006 | AJ509738 | 6H3 | 5 | 0.625 | 0.59375 |
| ANTP | AJ276511 | 22F1 | 16 | 0.0.4187 | 0.4187 |

The designations of BACs that were located to chromosomal position, genetic map position and physical sequence. The BACs located to a chromosomal position by FISH are from M. Solignac (personal communication). Each BAC contains a locus from the Solignac genetic map as given. The accession number identifies the physical sequence of the BAC and its mapped locus.

**Table S5: Nucleotide and dinucleotide composition of honey bee non-coding regions, and comparison with other genomes.**

| Species | | *Apis mellifera* | *Anopheles gambiae* | *Drosophila melanogaster* | *Caenorhabditis elegans* | *Arabidopsis thaliana* | *Tetraodon nigroviridis* | *Gallus gallus* | *Mus musculus* |
|---|---|---|---|---|---|---|---|---|---|
| Total amount of DNA analyzed (Mb): | | | | | | | | | |
| Intergenic | | 11.0 | 143.8 | 99.7 | 50.3 | 95.9 | 97.3 | 370.4 | 1,341.2 |
| Introns | | 32.3 | 47.5 | 52.3 | 28.4 | 23.2 | 65.3 | 574.7 | 945.7 |
| G+C content: | | | | | | | | | |
| | (a) | 31.60% | 43.70% | 40.30% | 35.00% | 32.50% | 44.60% | 41.40% | 42.10% |
| | (b) | 31.50% | 42.50% | 39.70% | 33.60% | 32.50% | 44.50% | 40.80% | 42.50% |
| Dinucleotide occurrence (observed/expected): | | | | | | | | | |
| ApA | (a) | 1.16 | 1.23 | 1.23 | 1.31 | 1.14 | 1.18 | 1.13 | 1.08 |
| | (b) | 1.17 | 1.22 | 1.23 | 1.35 | 1.17 | 1.20 | 1.15 | 1.08 |
| ApC | (a) | 0.80 | 0.96 | 0.86 | 0.83 | 0.93 | 0.93 | 0.85 | 0.87 |
| | (b) | 0.78 | 0.97 | 0.86 | 0.81 | 0.96 | 0.93 | 0.85 | 0.88 |
| ApG | (a) | 0.80 | 0.85 | 0.87 | 0.93 | 0.98 | 1.06 | 1.19 | 1.21 |
| | (b) | 0.81 | 0.85 | 0.88 | 0.90 | 0.97 | 1.08 | 1.20 | 1.24 |
| ApT | (a) | 1.03 | 0.92 | 0.95 | 0.83 | 0.90 | 0.83 | 0.84 | 0.86 |
| | (b) | 1.02 | 0.92 | 0.95 | 0.81 | 0.90 | 0.82 | 0.83 | 0.85 |
| CpA | (a) | 0.86 | 1.12 | 1.12 | 1.05 | 1.08 | 1.23 | 1.27 | 1.23 |
| | (b) | 0.85 | 1.12 | 1.12 | 1.00 | 1.10 | 1.23 | 1.27 | 1.24 |
| CpC | (a) | 1.07 | 0.98 | 1.07 | 1.10 | 1.08 | 1.04 | 1.10 | 1.21 |
| | (b) | 1.07 | 0.99 | 1.08 | 1.14 | 0.94 | 1.05 | 1.11 | 1.21 |
| CpG | (a) | 1.67 | 1.06 | 0.94 | 0.96 | 0.78 | 0.60 | 0.25 | 0.19 |
| | (b) | 1.68 | 1.04 | 0.92 | 0.99 | 0.56 | 0.58 | 0.23 | 0.19 |
| CpT | (a) | 0.79 | 0.85 | 0.87 | 0.92 | 0.99 | 1.06 | 1.20 | 1.21 |
| | (b) | 0.80 | 0.85 | 0.88 | 0.89 | 1.08 | 1.06 | 1.22 | 1.21 |
| GpA | (a) | 1.15 | 0.94 | 0.88 | 1.13 | 1.07 | 0.98 | 0.98 | 1.03 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | (b) 1.18 | 0.94 | 0.89 | 1.10 | 1.12 | 0.99 | 0.99 | 1.02 |
| GpC | (a) 1.07 | 1.15 | 1.32 | 0.97 | 0.90 | 1.08 | 1.13 | 0.93 |
| | (b) 1.05 | 1.13 | 1.30 | 1.05 | 0.95 | 1.07 | 1.15 | 0.95 |
| GpG | (a) 1.06 | 0.97 | 1.06 | 1.09 | 1.08 | 1.03 | 1.09 | 1.20 |
| | (b) 1.07 | 0.97 | 1.05 | 1.12 | 0.96 | 1.00 | 1.04 | 1.17 |
| GpT | (a) 0.79 | 0.97 | 0.86 | 0.84 | 0.94 | 0.93 | 0.86 | 0.88 |
| | (b) 0.78 | 0.98 | 0.88 | 0.82 | 0.96 | 0.96 | 0.89 | 0.90 |
| TpA | (a) 0.84 | 0.72 | 0.76 | 0.61 | 0.78 | 0.66 | 0.70 | 0.74 |
| | (b) 0.82 | 0.74 | 0.77 | 0.60 | 0.76 | 0.65 | 0.71 | 0.75 |
| TpC | (a) 1.14 | 0.94 | 0.88 | 1.12 | 1.08 | 0.97 | 0.98 | 1.02 |
| | (b) 1.17 | 0.94 | 0.89 | 1.09 | 1.07 | 0.97 | 0.97 | 1.00 |
| TpG | (a) 0.86 | 1.13 | 1.13 | 1.04 | 1.09 | 1.23 | 1.27 | 1.23 |
| | (b) 0.85 | 1.13 | 1.13 | 1.02 | 1.21 | 1.24 | 1.27 | 1.23 |
| TpT | (a) 1.16 | 1.23 | 1.23 | 1.29 | 1.13 | 1.18 | 1.11 | 1.07 |
| | (b) 1.17 | 1.20 | 1.21 | 1.33 | 1.05 | 1.15 | 1.09 | 1.06 |

(a) Intergenic DNA; (b) Introns

## Table S6: Comparison of Gene Sets.

| Predicted Gene Set | No. Genes | No. Perfect Alignments / weighted by no. gene models | No. Present / weighted by no. gene models |
|---|---|---|---|
| GLEAN | 10,157 | 111 / .011 | 356 / .035 |
| Fgenesh | 32,664 | 100 / .003 | 385 / .012 |
| Fgenesh++ | 19,201 | 97 / .005 | 350 / .018 |
| NCBI | 9,759 | 88 / .009 | 340 / .035 |
| Evolutionary Conserved Core | 10,966 | 39 / .004 | 284 / .026 |
| Ensembl | 27,755* | 32 / .0012 | 217 / .008 |
| *Drosophila* Orthologs | 8,878* | 4 / .0005 | 116 / .013 |

*includes splice variants

## Table S7: MicroRNAs

| Identifier, contig location | Sequence |
|---|---|
| >ame-miR-9b MIMAT0001492 | GCTTTGGTAATCTAGCTTTATGA |
| >ame-miR-12 MIMAT0001472 | TGAGTATTACATCAGGTACTGGT |
| >ame-miR-124 MIMAT0001473 | TAAGGCACGCGGTGAATGCCAAG |
| >ame-miR-125 MIMAT0001474 | CCCCTGAGACCCTAACTTGTGA |
| >ame-miR-133 MIMAT0001475 | TTGGTCCCCTTCAACCAGCTGT |
| >ame-miR-184 MIMAT0001476 | TGGACGGAGAACTGATAAGGGC |
| >ame-miR-210 MIMAT0001477 | TTGTGCGTGTGACAGCGGCTA |
| >ame-miR-219 MIMAT0001478 | TGATTGTCCAAACGCAATTCTTG |
| >ame-miR-263 MIMAT0001479 | GTAAATGGCACTGGAAGAATTCAC |
| >ame-miR-276 MIMAT0001480 | TAGGAACTTCATACCGTGCTCT |
| >ame-miR-277 MIMAT0001481 | TAAATGCACTATCTGGTACGACA |
| >ame-miR-278 MIMAT0001482 | TCGGTGGGACTTTCGTCCGTTT |
| >ame-miR-281 MIMAT0001483 | TGTCATGGAGTTGCTCTCTTTGT |
| >ame-miR-282 MIMAT0001484 | GATTTAGCCTCTCCTAGGCTTTGTCTGT |
| >ame-miR-305 MIMAT0001486 | ATTGTACTTCATCAGGTGCTCTG |
| >ame-miR-315 MIMAT0001487 | TTTTGATTGTTGCTCAGAAAGC |
| >ame-miR-317 MIMAT0001488 | TGAACACAGCTGGTGGTATCTCAGT |
| >1:3-27:Contig5303:28849:28875:+ | AACTACGTGTATTCTCAAGCAATAACA |
| >2:3-26:Contig5152:16524:16549:- | AACAACCAAGAATATCAAACATATCT |

| | |
|---|---|
| >3:275-22:Contig5152:16564:16585:+ | CCAGGAATCAAACATATTATTA |
| >4:16-23:Contig5560:7389:7411:- | AAATTGACTCTAGTAGGGAGTCC |
| >5:15-24:Contig3345:18299:18322:- | GGTAAAGCGTAGGAATTCTAAAAC |
| >6:230-22:Contig689:10381:10402:+ | CTGCAATGCACTACGGAATTGA |
| >7:5-28:Contig5581:22777:22804:+ | AGTTTTCAACTAGCAATAATCGCACCTC |
| >8:3-23:Contig4904:1877:1899:+ | ACCACGCACAAGAGCCTGCAGCA |
| >9:13-23:Contig2989:21112:21134:+ | GCGGCCAGGTTGGCGGTGTACGA |
| >10:49-23:Contig2370:12765:12787:+ | TGGGGTTGCTTCGACGAGTTCAA |
| >11:83-23:Contig5267:19004:19026:+ | AAGCACAAGGAGTCGAAGCACCT |
| >12:4-29:Contig6617:8346:8374:+ | GCCGTCACCCAGTCCTGCAGCACCGGCGA |
| >13:2-27:Contig4870:17222:17248:- | GACAACGTTGGCTTCAACGTGAAGAAC |
| >14:2-25:Contig1504:192:216:- | AGGGATTCGGTTTTGTAACATTCGC |
| >15:4-25:Contig2364:7116:7140:- | AGCCCAAGATCCAAGTCGCCTCCAA |
| >16:5-24:Contig7280:17004:17027:- | GCTCTACCACTGAGCTATATCCCC |
| >17:24-24:Contig5599:11767:11790:- | CAGGTGAAGATCTGGTTCCAGAAC |
| >18:84-23:Contig2187:462:484:+ | ATCTCGTTGGCGCACTCGATGCA |
| >19:111-22:Contig2327:9616:9637:+ | GTGATGATCATCTCGGTGCCGA |
| >20:125-22:Contig4109:28566:28587:+ | ACATGTACTCCTGCACGATGTA |
| >21:212-22:Contig5564:13893:13914:+ | AGTGCGACTGCGGCTGGGAGGA |
| >22:267-22:Contig461:87433:87454:- | AGGTTGAAGATGGTGTAGATGA |
| >23:37-21:Contig4131:17187:17207:+ | CCTTGCAGCCCTCGCAGGTGA |

| | |
|---|---|
| >24:174-21:Contig4222:29953:29973:+ | CTGGCTGTGGAAGCTGGCGAA |
| >25:605-21:Contig2856:13986:14006:- | ACGATCAGGATCTCCTGCAGG |
| >HC_mir-283_mature Amel2.0 | AAATATCAGCTGGTAATTCT |

## Table S8. Ortholog coverage of different gene prediction approaches.

A) The following sets were compared:

| Acronym | Proteins | Source |
|---|---|---|
| FGAB | 32,664 | Fgenesh softberry (ab initio) |
| ENSM | 27,755 | EnsEMBL EBI |
| HSAP | 22,218 | *Homo_sapiens.NCBI35.LONGESTpep* |
| FGPL | 19,201 | Fgenesh++ softberry |
| AGAM | 14,364 | *Anopheles_gambiae.MOZ2a.LONGESTpep* |
| DMEL | 13,450 | *D.melanogaster.LONGESTpep Flybase r4.1* |
| EZs1 | 10,966 | Homology core (E. Zdobnov) |
| GLN2 | 10,044 | GLEAN predictions version 2 |
| NCBI | 9,759 | NCBI |
| MEIS | 8,878 | M. Eisen |

B) Counts of the number of best reciprocal hits in Smith-Waterman protein comparisons between each of the sets and proteomes of fruit fly, mosquito and human.

| *D.melanogaster* | | *A. gambiae* | | *H. sapiens* | |
|---|---|---|---|---|---|
| AGAM | 7,390 | - | - | AGAM | 5,585 |
| - | - | DMEL | 7,390 | DMEL | 5,681 |
| ENSM | 5,904 | ENSM | 5,954 | ENSM | 5,270 |
| EZs1 | 6,566 | EZs1 | 6,608 | EZs1 | 5,630 |
| FGAB | 6,762 | FGAB | 6,855 | FGAB | 5,963 |
| FGPL | 5,940 | FGPL | 5,924 | FGPL | 5,254 |
| GLN2 | 6,698 | GLN2 | 6,580 | GLN2 | 5,848 |
| HSAP | 5,681 | HSAP | 5,585 | - | - |
| MEIS | 5,551 | MEIS | 5,074 | MEIS | 4,383 |

| | | | | | |
|---|---|---|---|---|---|
| NCBI | 6,275 | NCBI | 6,220 | NCBI | 5,539 |

**Table S9. 3-way overlap of best reciprocal hits between proteomes of fruit fly and mosquito with each pair of bee gene prediction sets shows discrepancies between sets.**

The diagonal cells show the number of best reciprocal hits as in Table S6 and, for example, even though the number of best reciprocal hits between FGAB and GLN2 sets and *D. melanogaster* proteome are similar (6762 and 6698 respectively) they agree on only 6308 genes etc.

A) *D. melanogaster*

| | AGAM | FGAB | GLN2 | EZs1 | NCBI | FGPL | ENSM | HSAP | MEIS |
|---|---|---|---|---|---|---|---|---|---|
| **AGAM** | 7,390 | 6,029 | 5,999 | 5,901 | 5,661 | 5,338 | 5,311 | 5,140 | 4,942 |
| **FGAB** | 6,029 | 6,762 | 6,308 | 6,142 | 5,930 | 5,658 | 5,479 | 5,098 | 5,089 |
| **GLN2** | 5,999 | 6,308 | 6,698 | 6,213 | 6,143 | 5,692 | 5,542 | 5,080 | 5,136 |
| **EZs1** | 5,901 | 6,142 | 6,213 | 6,566 | 5,836 | 5,606 | 5,442 | 5,016 | 5,039 |
| **NCBI** | 5,661 | 5,930 | 6,143 | 5,836 | 6,275 | 5,366 | 5,257 | 4,804 | 4,760 |
| **FGPL** | 5,338 | 5,658 | 5,692 | 5,606 | 5,366 | 5,940 | 4,924 | 4,522 | 4,478 |
| **ENSM** | 5,311 | 5,479 | 5,542 | 5,442 | 5,257 | 4,924 | 5,904 | 4,531 | 4,425 |
| **HSAP** | 5,140 | 5,098 | 5,080 | 5,016 | 4,804 | 4,522 | 4,531 | 5,681 | 4,170 |
| **MEIS** | 4,942 | 5,089 | 5,136 | 5,039 | 4,760 | 4,478 | 4,425 | 4,170 | 5,551 |

B) *A. gambiae*

| | DMEL | FGAB | EZs1 | GLN2 | NCBI | ENSM | FGPL | HSAP | MEIS |
|---|---|---|---|---|---|---|---|---|---|
| **DMEL** | 7,390 | 6,019 | 5,864 | 5,974 | 5,644 | 5,277 | 5,304 | 5,118 | 4,880 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **FGAB** | 6,019 | 6,855 | 6,125 | 6,226 | 5,895 | 5,455 | 5,616 | 5,011 | 4,738 |
| **EZs1** | 5,864 | 6,125 | 6,608 | 6,128 | 5,778 | 5,390 | 5,524 | 4,918 | 4,688 |
| **GLN2** | 5,974 | 6,226 | 6,128 | 6,580 | 6,063 | 5,457 | 5,591 | 4,971 | 4,776 |
| **NCBI** | 5,644 | 5,895 | 5,778 | 6,063 | 6,220 | 5,198 | 5,294 | 4,708 | 4,449 |
| **ENSM** | 5,277 | 5,455 | 5,390 | 5,457 | 5,198 | 5,954 | 4,872 | 4,402 | 4,165 |
| **FGPL** | 5,304 | 5,616 | 5,524 | 5,591 | 5,294 | 4,872 | 5,924 | 4,419 | 4,180 |
| **HSAP** | 5,118 | 5,011 | 4,918 | 4,971 | 4,708 | 4,402 | 4,419 | 5,585 | 3,922 |
| **MEIS** | 4,880 | 4,738 | 4,688 | 4,776 | 4,449 | 4,165 | 4,180 | 3,922 | 5,074 |

## Table S10: Comparative Ortholog Patterns.

Patterns of orthologous genes distribution among the three insect and three vertebrate species considered. For each pattern the number of such ortholous groups (OGs) and the total number of genes in each of the organisms classified into these groups are shown. 'N' denotes more than one gene member in each of the orthologous groups. The table is sorted in the descending number of bee genes.

| | | | | | | **OGs** | ***A.gam*** | ***D.mel*** | ***A.mel*** | ***H.sap*** | ***G.gal*** | ***T.nig*** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 1,428 | 1,428 | 1,428 | 1,428 | 1,428 | 1,428 | 1,428 |
| 1 | 1 | 1 | n | n | n | 704 | 704 | 704 | 704 | 2,059 | 1,686 | 2,373 |
| n | n | n | n | n | n | 120 | 729 | 572 | 463 | 934 | 441 | 634 |
| 1 | 1 | 1 | 1 | 1 | n | 352 | 352 | 352 | 352 | 352 | 352 | 819 |
| 1 | 1 | 1 | n | 1 | n | 326 | 326 | 326 | 326 | 770 | 326 | 900 |
| 1 | 1 | 1 | 1 | 0 | 1 | 267 | 267 | 267 | 267 | 267 | 0 | 267 |
| 1 | 1 | 1 | n | 1 | 1 | 220 | 220 | 220 | 220 | 608 | 220 | 220 |
| 0 | 0 | 1 | 1 | 1 | 1 | 116 | 0 | 0 | 116 | 116 | 116 | 116 |
| 1 | 1 | n | n | n | n | 52 | 52 | 52 | 107 | 168 | 158 | 217 |
| 1 | 1 | 1 | 1 | n | 1 | 106 | 106 | 106 | 106 | 106 | 230 | 106 |
| 1 | 1 | 1 | n | n | 1 | 105 | 105 | 105 | 105 | 243 | 220 | 105 |
| 0 | 1 | 1 | 1 | 1 | 1 | 101 | 0 | 101 | 101 | 101 | 101 | 101 |
| n | 1 | 1 | 1 | 1 | 1 | 100 | 246 | 100 | 100 | 100 | 100 | 100 |
| 1 | 1 | 1 | 1 | 1 | 0 | 94 | 94 | 94 | 94 | 94 | 94 | 0 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | n | 1 | 1 | 1 | 1 | 92 | 92 | 209 | 92 | 92 | 92 | 92 |
| 1 | 0 | 1 | 1 | 1 | 1 | 82 | 82 | 0 | 82 | 82 | 82 | 82 |
| n | n | n | 1 | 1 | 1 | 23 | 119 | 139 | 81 | 23 | 23 | 23 |
| n | n | n | n | 1 | n | 29 | 107 | 106 | 77 | 87 | 29 | 93 |
| 1 | 1 | n | 1 | 1 | 1 | 30 | 30 | 30 | 72 | 30 | 30 | 30 |
| n | 1 | 1 | n | n | n | 71 | 162 | 71 | 71 | 239 | 190 | 278 |
| 1 | n | 1 | n | n | n | 70 | 70 | 193 | 70 | 251 | 195 | 234 |
| 1 | 1 | 1 | 1 | n | n | 68 | 68 | 68 | 68 | 68 | 141 | 170 |
| 1 | 1 | 1 | 1 | 0 | n | 63 | 63 | 63 | 63 | 63 | 0 | 143 |
| 0 | 0 | 1 | 0 | 0 | 0 | 58 | 0 | 0 | 58 | 0 | 0 | 0 |
| n | n | 1 | n | n | n | 53 | 201 | 178 | 53 | 258 | 187 | 270 |
| 1 | n | n | n | n | n | 21 | 21 | 55 | 49 | 104 | 71 | 113 |
| 0 | 1 | 1 | n | n | n | 43 | 0 | 43 | 43 | 136 | 104 | 163 |
| n | 1 | n | n | n | n | 15 | 53 | 15 | 40 | 87 | 62 | 75 |
| 0 | 1 | 1 | 1 | 1 | n | 39 | 0 | 39 | 39 | 39 | 39 | 86 |
| n | 1 | 1 | 1 | 1 | n | 38 | 77 | 38 | 38 | 38 | 38 | 91 |
| n | n | n | n | n | 1 | 11 | 46 | 56 | 38 | 31 | 24 | 11 |
| 1 | 1 | n | n | 1 | n | 17 | 17 | 17 | 36 | 45 | 17 | 47 |
| 1 | n | 1 | n | 1 | n | 35 | 35 | 82 | 35 | 106 | 35 | 89 |
| 1 | 1 | 1 | 0 | 1 | 1 | 34 | 34 | 34 | 34 | 0 | 34 | 34 |
| 1 | n | n | n | 1 | n | 15 | 15 | 36 | 34 | 50 | 15 | 47 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n | 1 | 1 | n | 1 | n | 32 | 66 | 32 | 32 | 82 | 32 | 84 |
| 0 | 0 | 1 | 1 | 1 | n | 30 | 0 | 0 | 30 | 30 | 30 | 72 |
| 1 | 0 | 1 | n | n | n | 30 | 30 | 0 | 30 | 75 | 67 | 111 |
| 1 | n | 1 | 1 | 1 | n | 30 | 30 | 66 | 30 | 30 | 30 | 66 |
| n | n | 1 | 1 | 1 | 1 | 27 | 68 | 82 | 27 | 27 | 27 | 27 |
| 1 | 1 | n | 1 | n | n | 11 | 11 | 11 | 26 | 11 | 25 | 26 |
| 1 | 1 | 1 | 1 | 0 | 0 | 25 | 25 | 25 | 25 | 25 | 0 | 0 |
| 1 | 1 | n | 1 | 1 | n | 10 | 10 | 10 | 25 | 10 | 10 | 24 |
| 1 | 1 | 1 | n | 0 | n | 24 | 24 | 24 | 24 | 58 | 0 | 62 |
| 1 | 1 | n | n | 1 | 1 | 9 | 9 | 9 | 24 | 21 | 9 | 9 |
| n | n | n | n | 1 | 1 | 9 | 24 | 29 | 24 | 34 | 9 | 9 |
| 0 | 0 | 1 | 1 | 0 | 1 | 21 | 0 | 0 | 21 | 21 | 0 | 21 |
| 0 | 0 | 1 | n | n | n | 21 | 0 | 0 | 21 | 58 | 51 | 69 |
| 0 | 1 | 1 | n | 1 | n | 21 | 0 | 21 | 21 | 51 | 21 | 64 |
| 1 | 0 | 1 | 1 | 1 | n | 21 | 21 | 0 | 21 | 21 | 21 | 59 |
| 1 | 1 | 1 | 0 | 0 | 0 | 21 | 21 | 21 | 21 | 0 | 0 | 0 |
| n | 1 | 1 | 1 | 0 | 1 | 21 | 42 | 21 | 21 | 21 | 0 | 21 |
| n | n | n | 1 | 1 | n | 8 | 28 | 31 | 21 | 8 | 8 | 19 |
| 0 | 0 | 1 | 1 | 0 | 0 | 20 | 0 | 0 | 20 | 20 | 0 | 0 |
| 0 | 0 | 1 | n | 1 | n | 20 | 0 | 0 | 20 | 45 | 20 | 52 |
| n | 1 | 1 | n | 1 | 1 | 20 | 40 | 20 | 20 | 43 | 20 | 20 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n | 1 | n | n | 1 | n | 10 | 22 | 10 | 20 | 25 | 10 | 24 |
| 0 | 1 | 1 | 1 | 0 | 1 | 19 | 0 | 19 | 19 | 19 | 0 | 19 |
| n | 1 | n | 1 | 1 | 1 | 9 | 21 | 9 | 19 | 9 | 9 | 9 |
| 1 | 1 | 1 | 0 | 0 | 1 | 18 | 18 | 18 | 18 | 0 | 0 | 18 |
| 1 | 1 | 1 | n | 1 | 0 | 18 | 18 | 18 | 18 | 46 | 18 | 0 |
| 1 | n | 1 | n | 1 | 1 | 18 | 18 | 39 | 18 | 52 | 18 | 18 |
| n | n | n | 1 | n | n | 6 | 22 | 18 | 18 | 6 | 12 | 18 |
| 0 | 0 | 1 | 0 | 1 | 0 | 17 | 0 | 0 | 17 | 0 | 17 | 0 |
| 1 | 0 | 1 | 0 | 0 | 0 | 17 | 17 | 0 | 17 | 0 | 0 | 0 |
| 1 | n | n | 1 | 1 | 1 | 8 | 8 | 20 | 17 | 8 | 8 | 8 |
| 0 | 0 | 1 | 0 | 0 | 1 | 16 | 0 | 0 | 16 | 0 | 0 | 16 |
| 0 | 0 | 1 | n | 1 | 1 | 16 | 0 | 0 | 16 | 40 | 16 | 16 |
| 1 | 0 | 1 | n | 1 | n | 16 | 16 | 0 | 16 | 44 | 16 | 47 |
| 1 | 1 | 1 | n | 0 | 1 | 16 | 16 | 16 | 16 | 37 | 0 | 16 |
| 1 | 1 | n | 1 | 1 | 0 | 8 | 8 | 8 | 16 | 8 | 8 | 0 |
| 0 | 1 | 1 | 1 | 1 | 0 | 15 | 0 | 15 | 15 | 15 | 15 | 0 |
| n | n | 1 | n | 1 | n | 15 | 34 | 46 | 15 | 48 | 15 | 47 |
| n | n | n | 1 | n | 1 | 5 | 18 | 17 | 15 | 5 | 11 | 5 |
| 0 | 1 | n | 1 | 1 | 1 | 4 | 0 | 4 | 14 | 4 | 4 | 4 |
| 1 | 1 | n | 1 | 0 | 1 | 7 | 7 | 7 | 14 | 7 | 0 | 7 |
| 1 | n | n | 1 | 1 | n | 5 | 5 | 14 | 14 | 5 | 5 | 11 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 1 | 1 | 0 | 13 | 0 | 0 | 13 | 13 | 13 | 0 |
| 0 | 0 | n | 1 | 1 | 1 | 6 | 0 | 0 | 13 | 6 | 6 | 6 |
| 1 | 0 | 1 | 1 | 0 | 1 | 13 | 13 | 0 | 13 | 13 | 0 | 13 |
| 1 | 0 | 1 | n | 1 | 1 | 13 | 13 | 0 | 13 | 31 | 13 | 13 |
| 0 | 1 | 1 | 1 | n | n | 12 | 0 | 12 | 12 | 12 | 26 | 37 |
| 0 | 1 | 1 | n | 1 | 1 | 12 | 0 | 12 | 12 | 90 | 12 | 12 |
| 1 | 0 | n | 1 | 1 | 1 | 6 | 6 | 0 | 12 | 6 | 6 | 6 |
| 1 | 1 | n | 1 | n | 1 | 6 | 6 | 6 | 12 | 6 | 12 | 6 |
| 1 | n | 1 | 1 | 0 | 1 | 12 | 12 | 26 | 12 | 12 | 0 | 12 |
| 1 | n | 1 | n | n | 1 | 12 | 12 | 28 | 12 | 31 | 32 | 12 |
| n | n | n | 1 | 0 | 1 | 5 | 15 | 12 | 12 | 5 | 0 | 5 |
| n | 1 | 1 | 1 | n | n | 11 | 22 | 11 | 11 | 11 | 23 | 29 |
| 1 | 1 | 1 | 0 | 0 | n | 10 | 10 | 10 | 10 | 0 | 0 | 22 |
| n | n | 1 | 1 | 1 | n | 10 | 28 | 30 | 10 | 10 | 10 | 21 |
| 0 | 1 | 1 | 1 | 0 | 0 | 9 | 0 | 9 | 9 | 9 | 0 | 0 |
| 0 | n | 1 | 1 | 1 | 1 | 9 | 0 | 19 | 9 | 9 | 9 | 9 |
| 1 | 1 | 1 | 0 | 1 | 0 | 9 | 9 | 9 | 9 | 0 | 9 | 0 |
| 1 | n | 1 | 1 | 1 | 0 | 9 | 9 | 23 | 9 | 9 | 9 | 0 |
| n | n | 1 | n | 1 | 1 | 9 | 21 | 20 | 9 | 20 | 9 | 9 |
| 0 | 0 | 1 | n | n | 1 | 8 | 0 | 0 | 8 | 17 | 16 | 8 |
| 0 | 1 | 1 | 1 | 0 | n | 8 | 0 | 8 | 8 | 8 | 0 | 21 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | n | n | n | 1 | 4 | 4 | 4 | 8 | 9 | 8 | 4 |
| 1 | n | n | 1 | 0 | 1 | 3 | 3 | 7 | 8 | 3 | 0 | 3 |
| 1 | n | n | n | 1 | 1 | 4 | 4 | 9 | 8 | 13 | 4 | 4 |
| n | 0 | 1 | 1 | 1 | 1 | 8 | 18 | 0 | 8 | 8 | 8 | 8 |
| 0 | 0 | 1 | 1 | n | 1 | 7 | 0 | 0 | 7 | 7 | 14 | 7 |
| 1 | 0 | 1 | 0 | 0 | 1 | 7 | 7 | 0 | 7 | 0 | 0 | 7 |
| 1 | 0 | 1 | 1 | 1 | 0 | 7 | 7 | 0 | 7 | 7 | 7 | 0 |
| 1 | 1 | 1 | 0 | 1 | n | 7 | 7 | 7 | 7 | 0 | 7 | 15 |
| 1 | n | n | 1 | 0 | n | 3 | 3 | 10 | 7 | 3 | 0 | 13 |
| 1 | n | n | 1 | n | n | 3 | 3 | 10 | 7 | 3 | 6 | 8 |
| n | 1 | 1 | n | n | 1 | 7 | 15 | 7 | 7 | 18 | 15 | 7 |
| n | 1 | n | 1 | 1 | n | 3 | 8 | 3 | 7 | 3 | 3 | 7 |
| n | n | n | 1 | 1 | 0 | 2 | 5 | 5 | 7 | 2 | 2 | 0 |
| 0 | 0 | 1 | 1 | n | n | 6 | 0 | 0 | 6 | 6 | 12 | 17 |
| 0 | 1 | 1 | n | n | 1 | 6 | 0 | 6 | 6 | 17 | 15 | 6 |
| 0 | n | 1 | n | n | n | 6 | 0 | 13 | 6 | 28 | 15 | 17 |
| 1 | 0 | 1 | 1 | n | n | 6 | 6 | 0 | 6 | 6 | 12 | 14 |
| 1 | 0 | n | n | n | n | 3 | 3 | 0 | 6 | 12 | 8 | 19 |
| 1 | 1 | n | 0 | 1 | 1 | 3 | 3 | 3 | 6 | 0 | 3 | 3 |
| n | n | 1 | 0 | 0 | 0 | 6 | 40 | 55 | 6 | 0 | 0 | 0 |
| n | n | 1 | 1 | n | n | 6 | 15 | 16 | 6 | 6 | 14 | 16 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n | n | n | n | n | 0 | 2 | 5 | 4 | 6 | 6 | 5 | 0 |
| 0 | n | n | 1 | 1 | 1 | 1 | 0 | 9 | 5 | 1 | 1 | 1 |
| 0 | n | n | n | n | n | 1 | 0 | 2 | 5 | 4 | 2 | 5 |
| 1 | 0 | 1 | 1 | n | 1 | 5 | 5 | 0 | 5 | 5 | 13 | 5 |
| 1 | n | 1 | 1 | n | 1 | 5 | 5 | 10 | 5 | 5 | 11 | 5 |
| n | 1 | 1 | n | 0 | 1 | 5 | 11 | 5 | 5 | 23 | 0 | 5 |
| n | n | 1 | 1 | n | 1 | 5 | 14 | 33 | 5 | 5 | 96 | 5 |
| n | n | n | 1 | 0 | n | 2 | 7 | 13 | 5 | 2 | 0 | 4 |
| n | n | n | n | 0 | n | 2 | 10 | 10 | 5 | 16 | 0 | 5 |
| 0 | 0 | n | 1 | 1 | 0 | 2 | 0 | 0 | 4 | 2 | 2 | 0 |
| 0 | 0 | n | n | 1 | n | 2 | 0 | 0 | 4 | 5 | 2 | 11 |
| 0 | 1 | 1 | 1 | n | 1 | 4 | 0 | 4 | 4 | 4 | 11 | 4 |
| 0 | 1 | 1 | n | 0 | 1 | 4 | 0 | 4 | 4 | 8 | 0 | 4 |
| 0 | 1 | n | n | n | n | 2 | 0 | 2 | 4 | 6 | 4 | 8 |
| 0 | n | n | 1 | 1 | 0 | 1 | 0 | 2 | 4 | 1 | 1 | 0 |
| 1 | 1 | n | 0 | 0 | 0 | 2 | 2 | 2 | 4 | 0 | 0 | 0 |
| 1 | n | 1 | 1 | 0 | n | 4 | 4 | 9 | 4 | 4 | 0 | 8 |
| 1 | n | 1 | n | 0 | 1 | 4 | 4 | 8 | 4 | 8 | 0 | 4 |
| n | 0 | 1 | 1 | 1 | n | 4 | 8 | 0 | 4 | 4 | 4 | 12 |
| n | 1 | 1 | 0 | 0 | 0 | 4 | 9 | 4 | 4 | 0 | 0 | 0 |
| n | 1 | 1 | 0 | 0 | 1 | 4 | 8 | 4 | 4 | 0 | 0 | 4 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n | 1 | 1 | 1 | 0 | n | 4 | 8 | 4 | 4 | 4 | 0 | 8 |
| n | 1 | 1 | 1 | n | 1 | 4 | 9 | 4 | 4 | 4 | 8 | 4 |
| n | 1 | n | n | n | 0 | 2 | 4 | 2 | 4 | 5 | 4 | 0 |
| n | 1 | n | n | n | 1 | 2 | 5 | 2 | 4 | 4 | 5 | 2 |
| n | n | 1 | 1 | 0 | n | 4 | 8 | 11 | 4 | 4 | 0 | 10 |
| n | n | 1 | n | n | 1 | 4 | 11 | 9 | 4 | 10 | 8 | 4 |
| n | n | n | 0 | 0 | n | 1 | 7 | 4 | 4 | 0 | 0 | 2 |
| n | n | n | n | 0 | 1 | 2 | 5 | 7 | 4 | 5 | 0 | 2 |
| 0 | 0 | 1 | 1 | 0 | n | 3 | 0 | 0 | 3 | 3 | 0 | 7 |
| 0 | 0 | 1 | n | 0 | 0 | 3 | 0 | 0 | 3 | 8 | 0 | 0 |
| 0 | 0 | n | 1 | 1 | n | 1 | 0 | 0 | 3 | 1 | 1 | 3 |
| 0 | 1 | 1 | 0 | 0 | 0 | 3 | 0 | 3 | 3 | 0 | 0 | 0 |
| 0 | 1 | 1 | 0 | 0 | 1 | 3 | 0 | 3 | 3 | 0 | 0 | 3 |
| 0 | n | 1 | 1 | 1 | 0 | 3 | 0 | 7 | 3 | 3 | 3 | 0 |
| 0 | n | n | 1 | n | n | 1 | 0 | 3 | 3 | 1 | 2 | 2 |
| 1 | 0 | 1 | 0 | 1 | 1 | 3 | 3 | 0 | 3 | 0 | 3 | 3 |
| 1 | 0 | 1 | 1 | 0 | n | 3 | 3 | 0 | 3 | 3 | 0 | 7 |
| 1 | 0 | n | 1 | 1 | n | 1 | 1 | 0 | 3 | 1 | 1 | 4 |
| 1 | 1 | 1 | n | n | 0 | 3 | 3 | 3 | 3 | 6 | 6 | 0 |
| 1 | n | 1 | 1 | 0 | 0 | 3 | 3 | 12 | 3 | 3 | 0 | 0 |
| n | 0 | 1 | n | 1 | 1 | 3 | 7 | 0 | 3 | 6 | 3 | 3 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n | 0 | 1 | n | n | n | 3 | 7 | 0 | 3 | 14 | 7 | 11 |
| n | 0 | n | 1 | 1 | 1 | 1 | 3 | 0 | 3 | 1 | 1 | 1 |
| n | 1 | 1 | 1 | 0 | 0 | 3 | 7 | 3 | 3 | 3 | 0 | 0 |
| n | 1 | 1 | n | 0 | n | 3 | 8 | 3 | 3 | 6 | 0 | 10 |
| n | 1 | n | n | 0 | n | 1 | 3 | 1 | 3 | 3 | 0 | 2 |
| n | n | 1 | 0 | 0 | 1 | 3 | 8 | 7 | 3 | 0 | 0 | 3 |
| n | n | 1 | 1 | 1 | 0 | 3 | 6 | 6 | 3 | 3 | 3 | 0 |
| n | n | n | 0 | 0 | 0 | 1 | 3 | 2 | 3 | 0 | 0 | 0 |
| n | n | n | n | 0 | 0 | 1 | 2 | 4 | 3 | 3 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 | n | 2 | 0 | 0 | 2 | 0 | 0 | 4 |
| 0 | 0 | 1 | 0 | 1 | 1 | 2 | 0 | 0 | 2 | 0 | 2 | 2 |
| 0 | 0 | 1 | n | 1 | 0 | 2 | 0 | 0 | 2 | 4 | 2 | 0 |
| 0 | 0 | n | n | 1 | 1 | 1 | 0 | 0 | 2 | 3 | 1 | 1 |
| 0 | 0 | n | n | n | 0 | 1 | 0 | 0 | 2 | 2 | 3 | 0 |
| 0 | 0 | n | n | n | 1 | 1 | 0 | 0 | 2 | 3 | 2 | 1 |
| 0 | 0 | n | n | n | n | 1 | 0 | 0 | 2 | 3 | 3 | 4 |
| 0 | 1 | n | 1 | 0 | 0 | 1 | 0 | 1 | 2 | 1 | 0 | 0 |
| 0 | 1 | n | 1 | 0 | 1 | 1 | 0 | 1 | 2 | 1 | 0 | 1 |
| 0 | 1 | n | 1 | n | 1 | 1 | 0 | 1 | 2 | 1 | 2 | 1 |
| 0 | 1 | n | 1 | n | n | 1 | 0 | 1 | 2 | 1 | 2 | 4 |
| 0 | 1 | n | n | 0 | n | 1 | 0 | 1 | 2 | 2 | 0 | 2 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | n | n | 1 | 1 | 1 | 0 | 1 | 2 | 2 | 1 | 1 |
| 0 | 1 | n | n | 1 | n | 1 | 0 | 1 | 2 | 3 | 1 | 6 |
| 0 | n | 1 | 0 | 1 | 1 | 2 | 0 | 4 | 2 | 0 | 2 | 2 |
| 0 | n | 1 | n | 1 | n | 2 | 0 | 4 | 2 | 5 | 2 | 5 |
| 1 | 0 | 1 | 0 | 0 | n | 2 | 2 | 0 | 2 | 0 | 0 | 4 |
| 1 | 0 | 1 | 0 | 1 | 0 | 2 | 2 | 0 | 2 | 0 | 2 | 0 |
| 1 | 0 | 1 | n | 0 | 1 | 2 | 2 | 0 | 2 | 4 | 0 | 2 |
| 1 | 0 | n | 0 | 0 | 0 | 1 | 1 | 0 | 2 | 0 | 0 | 0 |
| 1 | 0 | n | 1 | 0 | 1 | 1 | 1 | 0 | 2 | 1 | 0 | 1 |
| 1 | 0 | n | n | 1 | 1 | 1 | 1 | 0 | 2 | 2 | 1 | 1 |
| 1 | 1 | 1 | 0 | n | 1 | 2 | 2 | 2 | 2 | 0 | 6 | 2 |
| 1 | 1 | 1 | n | 0 | 0 | 2 | 2 | 2 | 2 | 5 | 0 | 0 |
| 1 | 1 | n | 0 | 0 | 1 | 1 | 1 | 1 | 2 | 0 | 0 | 1 |
| 1 | 1 | n | 0 | 1 | n | 1 | 1 | 1 | 2 | 0 | 1 | 2 |
| 1 | 1 | n | 1 | 0 | 0 | 1 | 1 | 1 | 2 | 1 | 0 | 0 |
| 1 | 1 | n | n | 0 | 1 | 1 | 1 | 1 | 2 | 6 | 0 | 1 |
| 1 | 1 | n | n | 0 | n | 1 | 1 | 1 | 2 | 3 | 0 | 4 |
| 1 | 1 | n | n | n | 0 | 1 | 1 | 1 | 2 | 4 | 2 | 0 |
| 1 | n | 1 | 0 | 0 | 0 | 2 | 2 | 5 | 2 | 0 | 0 | 0 |
| 1 | n | 1 | 0 | 0 | 1 | 2 | 2 | 6 | 2 | 0 | 0 | 2 |
| 1 | n | 1 | 0 | 0 | n | 2 | 2 | 4 | 2 | 0 | 0 | 4 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | n | 1 | 0 | 1 | 1 | 2 | 2 | 4 | 2 | 0 | 2 | 2 |
| 1 | n | 1 | 1 | n | n | 2 | 2 | 5 | 2 | 2 | 4 | 5 |
| 1 | n | 1 | n | 0 | 0 | 2 | 2 | 5 | 2 | 16 | 0 | 0 |
| 1 | n | 1 | n | 0 | n | 2 | 2 | 5 | 2 | 4 | 0 | 6 |
| 1 | n | n | 1 | n | 1 | 1 | 1 | 3 | 2 | 1 | 2 | 1 |
| 1 | n | n | n | 0 | 1 | 1 | 1 | 2 | 2 | 2 | 0 | 1 |
| 1 | n | n | n | 1 | 0 | 1 | 1 | 2 | 2 | 2 | 1 | 0 |
| n | 0 | 1 | 1 | n | n | 2 | 12 | 0 | 2 | 2 | 12 | 18 |
| n | 0 | 1 | n | 1 | n | 2 | 4 | 0 | 2 | 7 | 2 | 17 |
| n | 1 | 1 | 1 | 1 | 0 | 2 | 4 | 2 | 2 | 2 | 2 | 0 |
| n | 1 | 1 | n | n | 0 | 2 | 4 | 2 | 2 | 4 | 4 | 0 |
| n | 1 | n | 0 | 0 | 0 | 1 | 11 | 1 | 2 | 0 | 0 | 0 |
| n | 1 | n | 1 | 1 | 0 | 1 | 2 | 1 | 2 | 1 | 1 | 0 |
| n | 1 | n | 1 | n | n | 1 | 2 | 1 | 2 | 1 | 3 | 6 |
| n | 1 | n | n | 1 | 1 | 1 | 2 | 1 | 2 | 2 | 1 | 1 |
| n | n | 1 | 1 | 0 | 0 | 2 | 4 | 4 | 2 | 2 | 0 | 0 |
| n | n | 1 | 1 | 0 | 1 | 2 | 4 | 6 | 2 | 2 | 0 | 2 |
| n | n | n | 0 | 0 | 1 | 1 | 3 | 2 | 2 | 0 | 0 | 1 |
| n | n | n | 0 | 1 | 1 | 1 | 4 | 3 | 2 | 0 | 1 | 1 |
| n | n | n | 0 | n | n | 1 | 8 | 2 | 2 | 0 | 2 | 2 |
| n | n | n | n | 1 | 0 | 1 | 4 | 7 | 2 | 2 | 1 | 0 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 1 | n | 1 | 0 | 0 | 1 | 0 | 1 | 2 |
| 0 | 0 | 1 | 1 | n | 0 | 1 | 0 | 0 | 1 | 1 | 2 | 0 |
| 0 | 0 | 1 | n | 0 | 1 | 1 | 0 | 0 | 1 | 3 | 0 | 1 |
| 0 | 1 | 1 | 0 | 0 | n | 1 | 0 | 1 | 1 | 0 | 0 | 3 |
| 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |
| 0 | 1 | 1 | n | 0 | n | 1 | 0 | 1 | 1 | 2 | 0 | 2 |
| 0 | 1 | 1 | n | 1 | 0 | 1 | 0 | 1 | 1 | 8 | 1 | 0 |
| 0 | 1 | 1 | n | n | 0 | 1 | 0 | 1 | 1 | 2 | 2 | 0 |
| 0 | n | 1 | 1 | 0 | 0 | 1 | 0 | 3 | 1 | 1 | 0 | 0 |
| 0 | n | 1 | 1 | 0 | 1 | 1 | 0 | 2 | 1 | 1 | 0 | 1 |
| 0 | n | 1 | 1 | 0 | n | 1 | 0 | 2 | 1 | 1 | 0 | 2 |
| 0 | n | 1 | 1 | 1 | n | 1 | 0 | 2 | 1 | 1 | 1 | 2 |
| 0 | n | 1 | 1 | n | 1 | 1 | 0 | 2 | 1 | 1 | 2 | 1 |
| 0 | n | 1 | 1 | n | n | 1 | 0 | 2 | 1 | 1 | 2 | 3 |
| 0 | n | 1 | n | 0 | 1 | 1 | 0 | 2 | 1 | 2 | 0 | 1 |
| 0 | n | 1 | n | 1 | 1 | 1 | 0 | 2 | 1 | 2 | 1 | 1 |
| 0 | n | 1 | n | n | 1 | 1 | 0 | 5 | 1 | 2 | 2 | 1 |
| 1 | 0 | 1 | 0 | 1 | n | 1 | 1 | 0 | 1 | 0 | 1 | 2 |
| 1 | 0 | 1 | 0 | n | 0 | 1 | 1 | 0 | 1 | 0 | 2 | 0 |
| 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 |
| 1 | 0 | 1 | n | 0 | 0 | 1 | 1 | 0 | 1 | 2 | 0 | 0 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | n | n | 1 | 1 | 1 | 0 | 1 | 2 | 3 | 1 |
| 1 | 1 | 1 | 0 | n | 0 | 1 | 1 | 1 | 1 | 0 | 2 | 0 |
| 1 | 1 | 1 | 0 | n | n | 1 | 1 | 1 | 1 | 0 | 2 | 2 |
| 1 | 1 | 1 | 1 | n | 0 | 1 | 1 | 1 | 1 | 1 | 2 | 0 |
| 1 | n | 1 | 0 | n | n | 1 | 1 | 2 | 1 | 0 | 2 | 3 |
| 1 | n | 1 | n | 1 | 0 | 1 | 1 | 2 | 1 | 2 | 1 | 0 |
| n | 0 | 1 | 0 | 0 | 0 | 1 | 2 | 0 | 1 | 0 | 0 | 0 |
| n | 0 | 1 | 0 | 0 | n | 1 | 2 | 0 | 1 | 0 | 0 | 2 |
| n | 0 | 1 | 1 | 0 | 0 | 1 | 2 | 0 | 1 | 1 | 0 | 0 |
| n | 0 | 1 | 1 | 0 | n | 1 | 2 | 0 | 1 | 1 | 0 | 4 |
| n | 0 | 1 | n | n | 0 | 1 | 2 | 0 | 1 | 2 | 2 | 0 |
| n | 0 | 1 | n | n | 1 | 1 | 2 | 0 | 1 | 2 | 2 | 1 |
| n | 1 | 1 | 0 | 1 | 0 | 1 | 2 | 1 | 1 | 0 | 1 | 0 |
| n | 1 | 1 | 1 | n | 0 | 1 | 2 | 1 | 1 | 1 | 2 | 0 |
| n | 1 | 1 | n | 0 | 0 | 1 | 4 | 1 | 1 | 3 | 0 | 0 |
| n | n | 1 | 0 | 0 | n | 1 | 3 | 3 | 1 | 0 | 0 | 2 |
| n | n | 1 | 0 | n | n | 1 | 4 | 3 | 1 | 0 | 2 | 2 |
| n | n | 1 | 1 | n | 0 | 1 | 2 | 2 | 1 | 1 | 2 | 0 |
| n | n | 1 | n | 0 | 1 | 1 | 6 | 3 | 1 | 2 | 0 | 1 |
| n | n | 1 | n | 0 | n | 1 | 2 | 2 | 1 | 2 | 0 | 3 |
| n | n | 1 | n | n | 0 | 1 | 2 | 2 | 1 | 5 | 2 | 0 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 0 | 0 | 0 | 0 | 7 | 0 | 7 | 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 | 1 | 9 | 0 | 9 | 0 | 0 | 0 | 9 |
| 0 | 1 | 0 | 0 | 0 | n | 1 | 0 | 1 | 0 | 0 | 0 | 3 |
| 0 | 1 | 0 | 0 | 1 | 0 | 5 | 0 | 5 | 0 | 0 | 5 | 0 |
| 0 | 1 | 0 | 0 | 1 | 1 | 3 | 0 | 3 | 0 | 0 | 3 | 3 |
| 0 | 1 | 0 | 0 | n | n | 1 | 0 | 1 | 0 | 0 | 2 | 2 |
| 0 | 1 | 0 | 1 | 0 | 0 | 11 | 0 | 11 | 0 | 11 | 0 | 0 |
| 0 | 1 | 0 | 1 | 0 | 1 | 7 | 0 | 7 | 0 | 7 | 0 | 7 |
| 0 | 1 | 0 | 1 | 1 | 0 | 10 | 0 | 10 | 0 | 10 | 10 | 0 |
| 0 | 1 | 0 | 1 | 1 | 1 | 19 | 0 | 19 | 0 | 19 | 19 | 19 |
| 0 | 1 | 0 | 1 | 1 | n | 9 | 0 | 9 | 0 | 9 | 9 | 22 |
| 0 | 1 | 0 | 1 | n | 1 | 1 | 0 | 1 | 0 | 1 | 2 | 1 |
| 0 | 1 | 0 | n | 0 | 1 | 1 | 0 | 1 | 0 | 2 | 0 | 1 |
| 0 | 1 | 0 | n | 1 | 1 | 2 | 0 | 2 | 0 | 4 | 2 | 2 |
| 0 | 1 | 0 | n | 1 | n | 3 | 0 | 3 | 0 | 9 | 3 | 8 |
| 0 | 1 | 0 | n | n | 1 | 2 | 0 | 2 | 0 | 4 | 4 | 2 |
| 0 | 1 | 0 | n | n | n | 3 | 0 | 3 | 0 | 7 | 8 | 6 |
| 0 | n | 0 | 1 | 0 | 0 | 1 | 0 | 2 | 0 | 1 | 0 | 0 |
| 0 | n | 0 | 1 | 1 | 0 | 4 | 0 | 43 | 0 | 4 | 4 | 0 |
| 0 | n | 0 | 1 | 1 | 1 | 1 | 0 | 3 | 0 | 1 | 1 | 1 |
| 0 | n | 0 | 1 | 1 | n | 1 | 0 | 2 | 0 | 1 | 1 | 2 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | n | 0 | 1 | n | 1 | 1 | 0 | 2 | 0 | 1 | 2 | 1 |
| 0 | n | 0 | n | 0 | 1 | 1 | 0 | 20 | 0 | 2 | 0 | 1 |
| 0 | n | 0 | n | 1 | 1 | 1 | 0 | 2 | 0 | 2 | 1 | 1 |
| 0 | n | 0 | n | 1 | n | 2 | 0 | 4 | 0 | 5 | 2 | 12 |
| 0 | n | 0 | n | n | n | 1 | 0 | 4 | 0 | 12 | 2 | 3 |
| 1 | 0 | 0 | 0 | 0 | 0 | 123 | 123 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 1 | 20 | 20 | 0 | 0 | 0 | 0 | 20 |
| 1 | 0 | 0 | 0 | 0 | n | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 1 | 0 | 0 | 0 | 1 | 0 | 14 | 14 | 0 | 0 | 0 | 14 | 0 |
| 1 | 0 | 0 | 0 | 1 | 1 | 3 | 3 | 0 | 0 | 0 | 3 | 3 |
| 1 | 0 | 0 | 0 | n | 1 | 1 | 1 | 0 | 0 | 0 | 2 | 1 |
| 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 |
| 1 | 0 | 0 | 1 | 0 | 1 | 10 | 10 | 0 | 0 | 10 | 0 | 10 |
| 1 | 0 | 0 | 1 | 0 | n | 1 | 1 | 0 | 0 | 1 | 0 | 2 |
| 1 | 0 | 0 | 1 | 1 | 0 | 9 | 9 | 0 | 0 | 9 | 9 | 0 |
| 1 | 0 | 0 | 1 | 1 | 1 | 24 | 24 | 0 | 0 | 24 | 24 | 24 |
| 1 | 0 | 0 | 1 | 1 | n | 4 | 4 | 0 | 0 | 4 | 4 | 8 |
| 1 | 0 | 0 | 1 | n | 0 | 1 | 1 | 0 | 0 | 1 | 3 | 0 |
| 1 | 0 | 0 | 1 | n | 1 | 3 | 3 | 0 | 0 | 3 | 6 | 3 |
| 1 | 0 | 0 | n | 0 | n | 2 | 2 | 0 | 0 | 7 | 0 | 4 |
| 1 | 0 | 0 | n | 1 | 0 | 1 | 1 | 0 | 0 | 4 | 1 | 0 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | n | 1 | 1 | 3 | 3 | 0 | 0 | 7 | 3 | 3 |
| 1 | 0 | 0 | n | 1 | n | 3 | 3 | 0 | 0 | 7 | 3 | 8 |
| 1 | 0 | 0 | n | n | 1 | 1 | 1 | 0 | 0 | 2 | 2 | 1 |
| 1 | 0 | 0 | n | n | n | 2 | 2 | 0 | 0 | 5 | 7 | 5 |
| 1 | 1 | 0 | 0 | 0 | 0 | 5 | 5 | 5 | 0 | 0 | 0 | 0 |
| 1 | 1 | 0 | 0 | 0 | 1 | 8 | 8 | 8 | 0 | 0 | 0 | 8 |
| 1 | 1 | 0 | 0 | 0 | n | 2 | 2 | 2 | 0 | 0 | 0 | 5 |
| 1 | 1 | 0 | 0 | 1 | 0 | 6 | 6 | 6 | 0 | 0 | 6 | 0 |
| 1 | 1 | 0 | 0 | 1 | 1 | 5 | 5 | 5 | 0 | 0 | 5 | 5 |
| 1 | 1 | 0 | 1 | 0 | 0 | 5 | 5 | 5 | 0 | 5 | 0 | 0 |
| 1 | 1 | 0 | 1 | 0 | 1 | 24 | 24 | 24 | 0 | 24 | 0 | 24 |
| 1 | 1 | 0 | 1 | 0 | n | 4 | 4 | 4 | 0 | 4 | 0 | 11 |
| 1 | 1 | 0 | 1 | 1 | 0 | 8 | 8 | 8 | 0 | 8 | 8 | 0 |
| 1 | 1 | 0 | 1 | 1 | 1 | 96 | 96 | 96 | 0 | 96 | 96 | 96 |
| 1 | 1 | 0 | 1 | 1 | n | 36 | 36 | 36 | 0 | 36 | 36 | 77 |
| 1 | 1 | 0 | 1 | n | 1 | 10 | 10 | 10 | 0 | 10 | 23 | 10 |
| 1 | 1 | 0 | 1 | n | n | 2 | 2 | 2 | 0 | 2 | 4 | 6 |
| 1 | 1 | 0 | n | 0 | 1 | 1 | 1 | 1 | 0 | 2 | 0 | 1 |
| 1 | 1 | 0 | n | 0 | n | 6 | 6 | 6 | 0 | 14 | 0 | 16 |
| 1 | 1 | 0 | n | 1 | 0 | 1 | 1 | 1 | 0 | 2 | 1 | 0 |
| 1 | 1 | 0 | n | 1 | 1 | 15 | 15 | 15 | 0 | 64 | 15 | 15 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | n | 1 | n | 22 | 22 | 22 | 0 | 54 | 22 | 69 |
| 1 | 1 | 0 | n | n | 0 | 1 | 1 | 1 | 0 | 2 | 2 | 0 |
| 1 | 1 | 0 | n | n | 1 | 6 | 6 | 6 | 0 | 16 | 45 | 6 |
| 1 | 1 | 0 | n | n | n | 31 | 31 | 31 | 0 | 86 | 77 | 112 |
| 1 | n | 0 | 0 | 0 | 1 | 1 | 1 | 17 | 0 | 0 | 0 | 1 |
| 1 | n | 0 | 0 | 0 | n | 1 | 1 | 2 | 0 | 0 | 0 | 2 |
| 1 | n | 0 | 1 | 0 | 0 | 3 | 3 | 7 | 0 | 3 | 0 | 0 |
| 1 | n | 0 | 1 | 0 | 1 | 3 | 3 | 7 | 0 | 3 | 0 | 3 |
| 1 | n | 0 | 1 | 1 | 1 | 6 | 6 | 16 | 0 | 6 | 6 | 6 |
| 1 | n | 0 | 1 | 1 | n | 3 | 3 | 8 | 0 | 3 | 3 | 7 |
| 1 | n | 0 | 1 | n | 0 | 1 | 1 | 2 | 0 | 1 | 2 | 0 |
| 1 | n | 0 | 1 | n | 1 | 1 | 1 | 2 | 0 | 1 | 2 | 1 |
| 1 | n | 0 | n | 1 | 1 | 1 | 1 | 2 | 0 | 2 | 1 | 1 |
| 1 | n | 0 | n | 1 | n | 3 | 3 | 7 | 0 | 7 | 3 | 6 |
| 1 | n | 0 | n | n | n | 3 | 3 | 6 | 0 | 7 | 6 | 10 |
| n | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 |
| n | 0 | 0 | 0 | 0 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | 1 |
| n | 0 | 0 | 0 | 1 | 0 | 1 | 2 | 0 | 0 | 0 | 1 | 0 |
| n | 0 | 0 | 1 | 0 | 1 | 1 | 2 | 0 | 0 | 1 | 0 | 1 |
| n | 0 | 0 | 1 | 1 | 1 | 2 | 5 | 0 | 0 | 2 | 2 | 2 |
| n | 0 | 0 | 1 | 1 | n | 1 | 3 | 0 | 0 | 1 | 1 | 2 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n | 0 | 0 | n | 0 | n | 1 | 3 | 0 | 0 | 2 | 0 | 2 |
| n | 0 | 0 | n | 1 | 0 | 1 | 29 | 0 | 0 | 2 | 1 | 0 |
| n | 0 | 0 | n | 1 | n | 3 | 8 | 0 | 0 | 11 | 3 | 22 |
| n | 0 | 0 | n | n | 1 | 1 | 10 | 0 | 0 | 2 | 2 | 1 |
| n | 0 | 0 | n | n | n | 2 | 29 | 0 | 0 | 24 | 7 | 35 |
| n | 1 | 0 | 0 | 0 | 0 | 1 | 2 | 1 | 0 | 0 | 0 | 0 |
| n | 1 | 0 | 0 | 1 | 0 | 2 | 4 | 2 | 0 | 0 | 2 | 0 |
| n | 1 | 0 | 0 | 1 | n | 1 | 2 | 1 | 0 | 0 | 1 | 4 |
| n | 1 | 0 | 0 | n | 1 | 1 | 2 | 1 | 0 | 0 | 2 | 1 |
| n | 1 | 0 | 1 | 0 | 0 | 1 | 2 | 1 | 0 | 1 | 0 | 0 |
| n | 1 | 0 | 1 | 0 | 1 | 1 | 2 | 1 | 0 | 1 | 0 | 1 |
| n | 1 | 0 | 1 | 1 | 0 | 1 | 2 | 1 | 0 | 1 | 1 | 0 |
| n | 1 | 0 | 1 | 1 | 1 | 12 | 24 | 12 | 0 | 12 | 12 | 12 |
| n | 1 | 0 | 1 | 1 | n | 3 | 18 | 3 | 0 | 3 | 3 | 6 |
| n | 1 | 0 | 1 | n | n | 1 | 3 | 1 | 0 | 1 | 2 | 3 |
| n | 1 | 0 | n | 1 | 1 | 2 | 4 | 2 | 0 | 5 | 2 | 2 |
| n | 1 | 0 | n | 1 | n | 1 | 2 | 1 | 0 | 3 | 1 | 5 |
| n | 1 | 0 | n | n | 0 | 1 | 3 | 1 | 0 | 2 | 2 | 0 |
| n | 1 | 0 | n | n | n | 2 | 6 | 2 | 0 | 10 | 9 | 12 |
| n | n | 0 | 0 | 0 | 0 | 1 | 3 | 4 | 0 | 0 | 0 | 0 |
| n | n | 0 | 1 | 0 | 1 | 1 | 2 | 2 | 0 | 1 | 0 | 1 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n | n | 0 | 1 | 1 | 1 | 5 | 13 | 11 | 0 | 5 | 5 | 5 |
| n | n | 0 | 1 | 1 | n | 1 | 2 | 4 | 0 | 1 | 1 | 5 |
| n | n | 0 | 1 | n | 0 | 1 | 2 | 2 | 0 | 1 | 2 | 0 |
| n | n | 0 | 1 | n | 1 | 1 | 3 | 8 | 0 | 1 | 2 | 1 |
| n | n | 0 | 1 | n | n | 1 | 2 | 2 | 0 | 1 | 2 | 2 |
| n | n | 0 | n | 0 | 1 | 1 | 5 | 2 | 0 | 3 | 0 | 1 |
| n | n | 0 | n | 1 | 1 | 1 | 3 | 2 | 0 | 2 | 1 | 1 |
| n | n | 0 | n | 1 | n | 1 | 2 | 23 | 0 | 4 | 1 | 8 |
| n | n | 0 | n | n | n | 1 | 3 | 2 | 0 | 2 | 2 | 2 |

## Table S11. Comparative Intron Patterns.

Patterns of shared intron distribution among the three insect and three vertebrate considered based on 4,441 orthologous groups defined above (Tables S8, S9, S10).

| D.mel | A.gam | A.mel | H.sap | G.gal | T.nig | Count |
|---|---|---|---|---|---|---|
| O | O | O | X | X | X | 16,694 |
| O | O | X | O | O | O | 7,209 |
| O | O | X | X | X | X | 4,617 |
| O | O | O | O | O | X | 4,360 |
| O | O | O | X | X | O | 4,120 |
| O | X | O | O | O | O | 3,236 |
| X | O | O | O | O | O | 2,452 |
| X | X | X | X | X | X | 2,199 |
| O | O | O | O | X | O | 2,110 |
| O | O | O | X | O | O | 1,710 |
| O | O | O | X | O | X | 1,612 |
| X | X | O | O | O | O | 1,546 |
| X | O | X | X | X | X | 998 |
| O | X | X | X | X | X | 993 |

| | | | | | | |
|---|---|---|---|---|---|---|
| X | X | X | O | O | O | 870 |
| O | O | X | X | X | O | 833 |
| X | X | O | X | X | X | 810 |
| X | O | O | X | X | X | 764 |
| O | X | O | X | X | X | 714 |
| O | O | O | O | X | X | 713 |
| X | O | X | O | O | O | 678 |
| O | X | X | O | O | O | 540 |
| X | X | X | X | X | O | 389 |
| O | O | X | X | O | X | 328 |
| O | O | X | O | O | X | 226 |
| X | O | X | X | X | O | 172 |
| O | X | X | X | X | O | 171 |
| X | O | O | X | X | O | 170 |
| X | X | X | X | O | X | 167 |
| O | O | X | X | O | O | 162 |
| O | O | X | O | X | X | 157 |
| O | X | O | X | X | O | 151 |
| X | X | O | X | X | O | 142 |
| O | O | X | O | X | O | 122 |
| O | X | O | O | O | X | 95 |

| | | | | | | |
|---|---|---|---|---|---|---|
| O | X | O | X | O | X | 74 |
| O | X | X | X | O | X | 74 |
| X | O | O | X | O | O | 64 |
| X | X | O | X | O | X | 62 |
| X | O | X | X | O | X | 61 |
| O | X | O | X | O | O | 61 |
| X | O | O | O | O | X | 58 |
| X | X | X | X | O | O | 55 |
| X | O | O | X | O | X | 46 |
| X | X | X | O | X | X | 45 |
| X | O | X | X | O | O | 45 |
| X | O | O | O | X | O | 44 |
| X | X | X | O | O | X | 42 |
| O | X | O | O | X | O | 41 |
| X | X | O | O | O | X | 39 |
| O | X | X | O | X | X | 38 |
| X | O | X | O | O | X | 37 |
| X | X | X | O | X | O | 35 |
| X | X | O | X | O | O | 31 |
| O | X | X | O | O | X | 31 |
| X | O | X | O | X | X | 30 |

| | | | | | | |
|---|---|---|---|---|---|---|
| X | X | O | O | X | O | 29 |
| X | X | O | O | X | X | 27 |
| X | O | O | O | X | X | 26 |
| O | X | X | X | O | O | 22 |
| O | X | O | O | X | X | 22 |
| O | X | X | O | X | O | 21 |
| X | O | X | O | X | O | 20 |

## Table S12: One to One to One Orthologs with Duplications in Honey Bee.

Duplications of bee genes that have single-copy orthologs in the six Metazoan analyzed. They are indicative of recently evolved, lineage-specific functions.

| A.mel | D.mel | A.gam | H.sap | G.gal | T.nig | Fly (human) gene | Annotation |
|---|---|---|---|---|---|---|---|
| 2 | 1 | 0 | 1 | 0 | 1 | CG8184 | protein modification; GO:0006464; proteolysis and peptidolysis; GO:0006508; ubiquitin cycle; GO:0006512 |
| 2 | 1 | 1 | 1 | 1 | 0 | norpA | calcium-mediated signaling; GO:0019722; cytosolic calcium ion concentration elevation; GO:0007204; diacylglycerol biosynthesis; GO:0006651; light-induced release of internally sequestered calcium ion (Ca2+); GO:0008377; perception of smell; GO:000 |
| 2 | 1 | 1 | 1 | 1 | 1 | mind-bomb | DNA metabolism; GO:0006259; DNA replication; GO:0006260; cell cycle; GO:0007049; cell proliferation; GO:0008283; eye morphogenesis (sensu Endopterygota); GO:0007456; nucleobase, nucleoside, nucleotide and nucleic acid metabolism; GO:0006139; p |
| 2 | 1 | 1 | 1 | 1 | 1 | Spt6 | RNA elongation; GO:0006354; RNA elongation from Pol II |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | promoter; GO:0006368; chromatin assembly or disassembly; GO:0006333; intracellular signaling cascade; GO:0007242; transcription initiation from Pol II promoter; GO:0006367 |
| 2 | 1 | 1 | 1 | 0 | 1 | CG9311 | G-protein coupled receptor protein signaling pathway; GO:0007186; protein amino acid dephosphorylation; GO:0006470; signal transduction; GO:0007165 |
| 2 | 1 | 1 | 1 | 1 | 1 | CG31004 | cell-matrix adhesion; GO:0007160 |
| 2 | 1 | 1 | 1 | 1 | 1 | CG31935 | |
| 2 | 1 | 1 | 1 | 1 | 1 | Top3beta | DNA catabolism, endonucleolytic; GO:0000737; DNA modification; GO:0006304; DNA topological change; GO:0006265; DNA unwinding; GO:0006268; tRNA aminoacylation for protein translation; GO:0006418 |
| 2 | 1 | 1 | 1 | 1 | 1 | CG17492 | protein ubiquitination; GO:0016567 |
| 2 | 1 | 1 | 1 | 1 | 1 | mei-9 | DNA recombination; GO:0006310; DNA repair; GO:0006281; chromosome segregation; GO:0007059; double-strand break repair; GO:0006302; meiosis; GO:0007126; meiotic chromosome segregation; |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | GO:0045132; meiotic recombination; GO:0007131; mismatch r |
| 2 | 1 | 1 | 1 | 1 | 1 | Klp61F | cell cycle; GO:0007049; microtubule-based movement; GO:0007018; mitotic centrosome separation; GO:0007100; mitotic spindle organization and biogenesis; GO:0007052; protein targeting; GO:0006605 |
| 2 | 1 | 1 | 1 | 0 | 1 | CG33138 | glycogen metabolism; GO:0005977 |
| 2 | 1 | 1 | 1 | 1 | 1 | car | endosome to lysosome transport; GO:0008333; eye pigment granule morphogenesis (sensu Endopterygota); GO:0008057; eye pigmentation (sensu Endopterygota); GO:0048072; intracellular protein transport; GO:0006886; intracellular transport; GO:0046907; |
| 2 | 1 | 1 | 1 | 1 | 1 | CG31716 | negative regulation of transcription; GO:0016481; protein ubiquitination; GO:0016567; regulation of transcription from Pol II promoter; GO:0006357 |
| 2 | 1 | 1 | 1 | 1 | 1 | bbx | regulation of transcription, DNA-dependent; GO:0006355; transport; GO:0006810 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 2 | 0 | 1 | 1 | 1 | 1 | ENSP00000293303 | Kelch-like protein 10 |
| 2 | 1 | 1 | 1 | 1 | 0 | CG14085 | |
| 2 | 1 | 1 | 1 | 1 | 1 | CG7430 | electron transport; GO:0006118; glycine catabolism; GO:0006546; glycolysis; GO:0006096; lipoamide metabolism; GO:0006748; tricarboxylic acid cycle; GO:0006099 |
| 2 | 1 | 1 | 1 | 1 | 1 | CG7433 | amino acid biosynthesis; GO:0008652; amino acid metabolism; GO:0006520; gamma-aminobutyric acid metabolism; GO:0009448 |
| 2 | 1 | 1 | 1 | 0 | 1 | CG32647 | DNA methylation; GO:0006306 |
| 2 | 1 | 1 | 1 | 1 | 0 | CG11963 | tricarboxylic acid cycle; GO:0006099 |
| 2 | 1 | 1 | 1 | 1 | 0 | Oatp26F | organic anion transport; GO:0015711 |
| 2 | 1 | 1 | 1 | 1 | 1 | CG13855 | |
| 2 | 1 | 0 | 1 | 1 | 1 | CG10555 | |
| 2 | 1 | 1 | 1 | 1 | 1 | CG1461 | amino acid catabolism; GO:0009063; aromatic amino acid family metabolism; GO:0009072; biosynthesis; GO:0009058; metal ion transport; GO:0030001 |
| 2 | 1 | 1 | 1 | 1 | 1 | b | amino acid metabolism; GO:0006520; beta-alanine biosynthesis; GO:0019483; |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | pigmentation; GO:0048066; sulfur metabolism; GO:0006790; synaptic transmission; GO:0007268; uracil catabolism; GO:0006212 |
| 2 | 1 | 1 | 1 | 1 | 1 | CG6903 | |
| 2 | 0 | 1 | 1 | 1 | 1 | ENSP00000347583 | Major facilitator superfamily MFS_1 |
| 2 | 1 | 1 | 1 | 1 | 1 | CG6321 | amino acid metabolism; GO:0006520; biosynthesis; GO:0009058 |
| 2 | 1 | 0 | 1 | 1 | 1 | Thd1 | mismatch repair; GO:0006298; nucleotide-excision repair; GO:0006289; regulation of transcription, DNA-dependent; GO:0006355 |
| 2 | 1 | 1 | 1 | 1 | 1 | CG11876 | pyruvate metabolism; GO:0006090; regulation of transcription, DNA-dependent; GO:0006355; tricarboxylic acid cycle; GO:0006099 |
| 2 | 0 | 1 | 1 | 0 | 1 | ENSP00000319429 | Muscleblind-like protein (Triplet-expansion RNA-binding protein). |
| 2 | 1 | 1 | 0 | 1 | 1 | CG13502 | |
| 2 | 0 | 1 | 1 | 1 | 1 | ENSP00000244490 | Werner helicase interacting protein |
| 2 | 1 | 1 | 1 | 1 | 0 | CG10793 | response to oxidative stress; GO:0006979 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 2 | 1 | 1 | 1 | 0 | 1 | CG12018 | DNA repair; GO:0006281; DNA replication; GO:0006260 |
| 2 | 1 | 1 | 1 | 1 | 1 | O-fut1 | O-linked glycosylation; GO:0006493; neurogenesis; GO:0007399; regulation of Notch signaling pathway; GO:0008593 |
| 2 | 1 | 1 | 1 | 1 | 1 | CG12030 | galactose metabolism; GO:0006012; nucleotide-sugar metabolism; GO:0009225 |
| 2 | 1 | 1 | 1 | 1 | 1 | CG10908 | ER-associated protein catabolism; GO:0030433; misfolded or incompletely synthesized protein catabolism; GO:0006515 |
| 2 | 1 | 1 | 1 | 1 | 1 | CG5281 | |
| 2 | 1 | 1 | 1 | 1 | 1 | CG9886 | carbohydrate metabolism; GO:0005975 |
| 2 | 0 | 1 | 1 | 1 | 1 | ENSP00000317998 | Ribonuclease P protein subunit p40 (EC 3.1.26.5) |
| 2 | 1 | 1 | 1 | 1 | 1 | Rpb5 | mRNA transcription; GO:0009299; transcription from Pol II promoter; GO:0006366 |
| 2 | 1 | 0 | 1 | 1 | 1 | mars | cell-cell signaling; GO:0007267 |
| 2 | 0 | 1 | 1 | 1 | 1 | ENSP00000298420 | Homeobox / paired-like homeodomain protein |
| 2 | 1 | 1 | 1 | 1 | 0 | CG4743 | transport; GO:0006810 |
| 2 | 1 | 1 | 1 | 0 | 1 | ENSP00000306982 | Pancreas specific transcription |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | factor, 1a. |
| 2 | 0 | 1 | 1 | 1 | 1 | Fer1 | regulation of transcription; GO:0045449 |
| 2 | 1 | 1 | 1 | 0 | 0 | CG12935 | |
| 2 | 1 | 1 | 0 | 0 | 1 | CG12910 | |
| 2 | 1 | 1 | 1 | 0 | 1 | CG7215 | |
| 2 | 1 | 1 | 1 | 1 | 0 | CG6674 | |
| 2 | 1 | 1 | 1 | 1 | 1 | CG8407 | microtubule-based movement; GO:0007018 |
| 2 | 1 | 1 | 1 | 1 | 0 | mRpL50 | |
| 2 | 1 | 1 | 0 | 1 | 1 | CG18428 | |
| 2 | 1 | 0 | 1 | 0 | 0 | CG4757 | |
| 2 | 1 | 1 | 1 | 1 | 1 | CG8891 | nucleobase, nucleoside, nucleotide and nucleic acid metabolism; GO:0006139; regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism; GO:0019219 |
| 2 | 1 | 1 | 1 | 0 | 1 | CG9804 | lipoate biosynthesis; GO:0009107; protein modification; GO:0006464 |
| 2 | 1 | 1 | 0 | 1 | 1 | ox | mitochondrial electron transport, ubiquinol to cytochrome c; GO:0006122; spermatid development; GO:0007286 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 2 | 1 | 1 | 1 | 1 | 1 | CG4774 | phospholipid biosynthesis; GO:0008654 |

**Table S13. Comparative Domains.**

From file: URL http://azra.embl.de/~zdobnov/Bee2/top500_domains.html

Statistically significant differences are marked in bold (p-value of chi-square test < 0.01 without correction for multiple tests).

| **Apis mellifera (honey bee)** 10157 total genes ( 71% w domains) | **Drosophila melanogaster (fruit fly)** 13450 total genes ( 68% w domains) | **Homo sapiens (man)** 22218 total genes ( 72% w domains) | **Family** | **GO terms** |
|---|---|---|---|---|
| 2275 | 3171 | 5083 | Transmembrane | |
| 1081 | **2895** | **4083** | Signal peptide | |
| 247 (1) | 314 (1) | **778 (1)** | IPR007087: Zinc finger, C2H2-type | Molecular Function: nucleic acid binding (GO:0003676), Cellular Component: nucleus (GO:0005634), Molecular Function: zinc ion binding (GO:0008270) |
| 201 (2) | 207 (3) | 464 (3) | IPR000719: Protein kinase | Molecular Function: protein kinase activity (GO:0004672), Molecular Function: ATP binding (GO:0005524), Biological Process: protein amino acid phosphorylation (GO:0006468) |
| 184 (3) | **184 (4)** | **261 (9)** | IPR001680: WD-40 repeat | |

| | | | | |
|---|---|---|---|---|
| 120 (4) | 121 (8) | **462 (4)** | IPR003599: Immunoglobulin subtype | |
| 117 (5) | 119 (9) | 285 (7) | IPR001841: Zinc finger, RING-type | Cellular Component: ubiquitin ligase complex (GO:0000151), Molecular Function: ubiquitin-protein ligase activity (GO:0004842), Molecular Function: zinc ion binding (GO:0008270), Biological Process: protein ubiquitination (GO:0016567) |
| 116 (6) | 122 (7) | **133 (28)** | IPR003593: AAA ATPase | Molecular Function: nucleotide binding (GO:0000166), Molecular Function: nucleoside-triphosphatase activity (GO:0017111) |
| 114 (7) | 128 (6) | 236 (12) | IPR000504: RNA-binding region RNP-1 (RNA recognition motif) | Molecular Function: nucleic acid binding (GO:0003676) |
| 107 (8) | 107 (10) | 221 (15) | IPR001611: Leucine-rich repeat | |
| 100 (9) | 103 (11) | 244 (11) | IPR002290: Serine/threonine protein kinase | Molecular Function: protein serine/threonine kinase activity (GO:0004674), Molecular Function: ATP binding (GO:0005524), Biological Process: protein amino acid phosphorylation (GO:0006468) |
| 98 (10) | 130 (5) | **99 (38)** | IPR011701: Major facilitator superfamily MFS_1 | |
| 97 (11) | **82 (20)** | 247 (10) | IPR002110: Ankyrin | |
| 97 (12) | 102 (12) | **369 (5)** | IPR007110: Immunoglobulin-like | |
| 96 (13) | 100 (14) | 220 (16) | IPR003598: Immunoglobulin C2 type | |

| | | | | |
|---|---|---|---|---|
| 89 (14) | 99 (15) | 226 (14) | IPR001356: Homeobox | Molecular Function: DNA binding (GO:0003677), Molecular Function: transcription factor activity (GO:0003700), Cellular Component: nucleus (GO:0005634), Biological Process: regulation of transcription, DNA-dependent (GO:0006355) |
| 83 (15) | **61 (33)** | **()** | IPR004117: Olfactory receptor, Drosophila | Molecular Function: olfactory receptor activity (GO:0004984), Molecular Function: odorant binding (GO:0005549), Biological Process: perception of smell (GO:0007608), Cellular Component: membrane (GO:0016020) |
| 78 (16) | **67 (30)** | **265 (8)** | IPR001849: Pleckstrin-like | |
| 78 (17) | 74 (25) | **109 (34)** | IPR011545: DEAD/DEAH box helicase, N-terminal | Molecular Function: nucleic acid binding (GO:0003676), Molecular Function: helicase activity (GO:0004386), Molecular Function: ATP binding (GO:0005524) |
| 75 (18) | 76 (22) | 140 (26) | IPR001440: TPR repeat | |
| 74 (19) | 76 (23) | **107 (35)** | IPR001650: Helicase, C-terminal | Molecular Function: nucleic acid binding (GO:0003676), Molecular Function: helicase activity (GO:0004386), Molecular Function: ATP binding (GO:0005524) |
| 73 (20) | 71 (28) | 208 (17) | IPR001452: SH3 | |
| 72 (21) | 76 (24) | 160 (23) | IPR003591: Leucine-rich repeat, typical subtype | |
| 71 (22) | 68 (29) | 166 (22) | IPR000210: BTB/POZ | Molecular Function: protein binding (GO:0005515) |
| 69 (23) | 84 (19) | 183 (19) | IPR002048: Calcium-binding EF-hand | Molecular Function: calcium ion binding (GO:0005509) |

| | | | | |
|---|---|---|---|---|
| 69 (24) | 89 (16) | **232 (13)** | IPR003656: Zinc finger, BED-type predicted | Molecular Function: DNA binding (GO:0003677) |
| 67 (25) | 56 (36) | 102 (37) | IPR000357: HEAT | |
| 66 (26) | 63 (31) | 143 (25) | IPR001478: PDZ/DHR/GLGF | Molecular Function: protein binding (GO:0005515) |
| 63 (27) | 56 (39) | 180 (20) | IPR011511: Variant SH3 | |
| 58 (28) | **243 (2)** | 120 (30) | IPR001254: Peptidase S1 and S6, chymotrypsin/Hap | Molecular Function: trypsin activity (GO:0004295), Biological Process: proteolysis and peptidolysis (GO:0006508) |
| 58 (29) | 61 (32) | 179 (21) | IPR003961: Fibronectin, type III | |
| 57 (30) | 61 (34) | **191 (18)** | IPR006210: Type I EGF | |
| 55 (31) | 71 (27) | **704 (2)** | IPR000276: Rhodopsin-like GPCR superfamily | Molecular Function: rhodopsin-like receptor activity (GO:0001584), Biological Process: G-protein coupled receptor protein signaling pathway (GO:0007186), Cellular Component: integral to membrane (GO:0016021) |
| 53 (32) | 58 (35) | 112 (31) | IPR001092: Basic helix-loop-helix dimerisation region bHLH | Cellular Component: nucleus (GO:0005634), Molecular Function: transcription regulator activity (GO:0030528), Biological Process: regulation of transcription (GO:0045449) |
| 53 (33) | 72 (26) | **38 (116)** | IPR005828: General substrate transporter | Molecular Function: transporter activity (GO:0005215), Biological Process: transport (GO:0006810), Cellular Component: integral to membrane (GO:0016021) |

| | | | | |
|---|---|---|---|---|
| 51 (34) | 55 (40) | 130 (29) | IPR001806: Ras GTPase | Molecular Function: GTP binding (GO:0005525), Biological Process: small GTPase mediated signal transduction (GO:0007264) |
| 48 (35) | 40 (47) | 134 (27) | IPR000008: C2 | |
| 47 (36) | 86 (18) | **61 (64)** | IPR001128: Cytochrome P450 | Molecular Function: monooxygenase activity (GO:0004497), Molecular Function: iron ion binding (GO:0005506), Biological Process: electron transport (GO:0006118), Molecular Function: heme binding (GO:0020037) |
| 44 (37) | 40 (50) | **150 (24)** | IPR006209: EGF-like | |
| 43 (38) | 56 (37) | 57 (68) | IPR002198: Short-chain dehydrogenase/reductase SDR | Biological Process: metabolism (GO:0008152), Molecular Function: oxidoreductase activity (GO:0016491) |
| 40 (39) | 42 (45) | 81 (48) | IPR001965: Zinc finger, PHD-type | Molecular Function: protein binding (GO:0005515), Biological Process: regulation of transcription, DNA-dependent (GO:0006355), Molecular Function: zinc ion binding (GO:0008270) |
| 39 (40) | 56 (38) | **49 (83)** | IPR003439: ABC transporter related | Molecular Function: ATP binding (GO:0005524), Molecular Function: ATPase activity (GO:0016887) |
| 39 (41) | 41 (46) | **44 (102)** | IPR003959: AAA ATPase, central region | Molecular Function: ATP binding (GO:0005524) |
| 38 (42) | 81 (21) | **2 (1621)** | IPR002557: Chitin binding Peritrophin-A | Cellular Component: extracellular region (GO:0005576), Biological Process: chitin metabolism (GO:0006030), Molecular Function: chitin binding (GO:0008061) |

| | | | | |
|---|---|---|---|---|
| 33 (43) | 43 (44) | 44 (101) | IPR002172: Low density lipoprotein-receptor, class A | |
| 33 (44) | 38 (52) | 111 (32) | IPR005821: Ion transport protein | Molecular Function: ion channel activity (GO:0005216), Biological Process: ion transport (GO:0006811), Cellular Component: membrane (GO:0016020) |
| 32 (45) | 46 (43) | 54 (71) | IPR001993: Mitochondrial substrate carrier | Molecular Function: binding (GO:0005488), Biological Process: transport (GO:0006810), Cellular Component: membrane (GO:0016020) |
| 31 (46) | **100 (13)** | **()** | IPR000618: Insect cuticle protein | Molecular Function: structural constituent of cuticle (GO:0042302) |
| 31 (47) | 32 (62) | 109 (33) | IPR000980: SH2 motif | Biological Process: intracellular signaling cascade (GO:0007242) |
| 31 (48) | 34 (59) | 69 (54) | IPR001781: LIM, zinc-binding | Molecular Function: zinc ion binding (GO:0008270) |
| 30 (49) | 30 (67) | 92 (39) | IPR001245: Tyrosine protein kinase | Molecular Function: protein-tyrosine kinase activity (GO:0004713), Molecular Function: ATP binding (GO:0005524), Biological Process: protein amino acid phosphorylation (GO:0006468) |
| 29 (50) | 32 (63) | 103 (36) | IPR001881: EGF-like calcium-binding | Molecular Function: calcium ion binding (GO:0005509) |
| 28 (51) | 31 (66) | 86 (42) | IPR000048: IQ calmodulin-binding region | |
| 28 (52) | 27 (78) | 44 (98) | IPR001214: Nuclear protein SET | |
| 28 (53) | 36 (53) | 60 (66) | IPR001810: Cyclin-like F-box | |

| | | | | |
|---|---|---|---|---|
| 27 (54) | 32 (60) | 46 (90) | IPR000608: Ubiquitin-conjugating enzyme, E2 | Biological Process: protein modification (GO:0006464), Biological Process: ubiquitin cycle (GO:0006512), Molecular Function: ubiquitin-like activating enzyme activity (GO:0008642) |
| 27 (55) | 40 (49) | 49 (81) | IPR001623: Heat shock protein DnaJ, N-terminal | Biological Process: protein folding (GO:0006457), Molecular Function: heat shock protein binding (GO:0031072), Molecular Function: unfolded protein binding (GO:0051082) |
| 27 (56) | 35 (56) | **15 (317)** | IPR002018: Carboxylesterase, type B | |
| 26 (57) | 24 (93) | 50 (77) | IPR000571: Zinc finger, CCCH-type | Molecular Function: nucleic acid binding (GO:0003676) |
| 26 (58) | 29 (71) | 78 (51) | IPR001660: Sterile alpha motif SAM | |
| 26 (59) | 26 (84) | 45 (93) | IPR001752: Kinesin, motor region | Molecular Function: microtubule motor activity (GO:0003777), Molecular Function: ATP binding (GO:0005524), Cellular Component: microtubule associated complex (GO:0005875), Biological Process: microtubule-based movement (GO:0007018) |
| 26 (60) | 26 (85) | 62 (62) | IPR002219: Protein kinase C, phorbol ester/diacylglycerol binding | Biological Process: intracellular signaling cascade (GO:0007242) |
| 26 (61) | **89 (17)** | **1 (3049)** | IPR012934: Zinc finger, AD-type | |
| 25 (62) | 20 (123) | 57 (67) | IPR000219: DH | |
| 25 (63) | 25 (88) | 83 (46) | IPR000483: Cysteine-rich flanking region, C-terminal | |

| | | | | |
|---|---|---|---|---|
| 25 (64) | 27 (79) | 65 (59) | IPR001715: Calponin-like actin-binding | |
| 24 (65) | 24 (97) | 48 (87) | IPR001005: Myb, DNA-binding | Molecular Function: DNA binding (GO:0003677), Cellular Component: nucleus (GO:0005634) |
| 24 (66) | 24 (98) | 44 (100) | IPR001878: Zinc finger, CCHC-type | Molecular Function: nucleic acid binding (GO:0003676) |
| 24 (67) | **13 (218)** | **()** | IPR003534: Major royal jelly protein | |
| 23 (68) | 20 (122) | 68 (55) | IPR000198: RhoGAP | |
| 23 (69) | 24 (94) | 53 (73) | IPR000626: Ubiquitin | Biological Process: protein modification (GO:0006464) |
| 23 (70) | 22 (103) | 62 (61) | IPR000910: HMG1/2 (high mobility group) box | Molecular Function: DNA binding (GO:0003677), Biological Process: regulation of transcription, DNA-dependent (GO:0006355) |
| 23 (71) | 25 (89) | 53 (74) | IPR000961: Protein kinase, C-terminal | Molecular Function: protein serine/threonine kinase activity (GO:0004674), Molecular Function: ATP binding (GO:0005524), Biological Process: protein amino acid phosphorylation (GO:0006468) |
| 23 (72) | 18 (138) | 45 (94) | IPR001791: Laminin G | |
| 23 (73) | 29 (72) | 36 (126) | IPR004087: KH | Molecular Function: nucleic acid binding (GO:0003676) |
| 22 (74) | 21 (115) | 48 (85) | IPR000299: Band 4.1 | Cellular Component: cytoskeleton (GO:0005856) |
| 22 (75) | 23 (100) | 44 (97) | IPR001202: WW/Rsp5/WWP | |

| | | | | |
|---|---|---|---|---|
| 22 (76) | 22 (106) | 54 (70) | IPR001394: Peptidase C19, ubiquitin carboxyl-terminal hydrolase 2 | Molecular Function: cysteine-type endopeptidase activity (GO:0004197), Molecular Function: ubiquitin thiolesterase activity (GO:0004221), Biological Process: ubiquitin-dependent protein catabolism (GO:0006511) |
| 22 (77) | 30 (69) | 61 (65) | IPR003579: Ras small GTPase, Rab type | Molecular Function: GTP binding (GO:0005525), Biological Process: small GTPase mediated signal transduction (GO:0007264), Biological Process: protein transport (GO:0015031) |
| 22 (78) | 22 (111) | 72 (53) | IPR011510: Sterile alpha motif homology 2 | |
| 21 (79) | 21 (114) | 47 (88) | IPR000195: RabGAP/TBC | |
| 21 (80) | 17 (144) | 48 (84) | IPR000225: Armadillo | |
| 21 (81) | 16 (163) | 65 (58) | IPR000884: Thrombospondin, type I | |
| 21 (82) | 21 (118) | 47 (89) | IPR001628: Nuclear hormone receptor, DNA-binding | Molecular Function: transcription factor activity (GO:0003700), Cellular Component: nucleus (GO:0005634), Biological Process: regulation of transcription, DNA-dependent (GO:0006355) |
| 21 (83) | 26 (86) | 49 (82) | IPR002867: Zinc finger, C6HC-type | |
| 21 (84) | 28 (76) | 35 (129) | IPR004088: KH, type 1 | Molecular Function: nucleic acid binding (GO:0003676) |

| | | | | |
|---|---|---|---|---|
| 21 (85) | 21 (120) | 44 (103) | IPR006029: Neurotransmitter-gated ion-channel transmembrane region | Biological Process: ion transport (GO:0006811), Cellular Component: membrane (GO:0016020), Molecular Function: neurotransmitter receptor activity (GO:0030594) |
| 21 (86) | 21 (121) | 46 (91) | IPR006202: Neurotransmitter-gated ion-channel ligand-binding | Molecular Function: extracellular ligand-gated ion channel activity (GO:0005230), Biological Process: transport (GO:0006810), Cellular Component: membrane (GO:0016020) |
| 20 (87) | 15 (178) | **1 (2023)** | IPR000172: Glucose-methanol-choline oxidoreductase | Biological Process: electron transport (GO:0006118), Molecular Function: oxidoreductase activity (GO:0016491) |
| 20 (88) | 30 (68) | 22 (210) | IPR001251: Cellular retinaldehyde-binding/triple function, C-terminal | |
| 20 (89) | 20 (127) | 50 (78) | IPR004827: Basic-leucine zipper (bZIP) transcription factor | Molecular Function: DNA binding (GO:0003677), Cellular Component: nucleus (GO:0005634), Biological Process: regulation of transcription, DNA-dependent (GO:0006355) |
| 20 (90) | 12 (241) | **89 (40)** | IPR007125: Histone core | Molecular Function: DNA binding (GO:0003677) |
| 20 (91) | 15 (196) | **1 (2664)** | IPR007867: GMC oxidoreductase | Molecular Function: oxidoreductase activity, acting on CH-OH group of donors (GO:0016614), Molecular Function: FAD binding (GO:0050660) |
| 20 (92) | 17 (156) | 37 (122) | IPR012680: Laminin G, subdomain 2 | |
| 19 (93) | 26 (81) | 34 (131) | IPR000449: Ubiquitin-associated | |

| | | | | |
|---|---|---|---|---|
| 19 (94) | 18 (133) | 34 (133) | IPR000595: Cyclic nucleotide-binding | |
| 19 (95) | 28 (75) | **17 (284)** | IPR001509: NAD-dependent epimerase/dehydratase | Molecular Function: catalytic activity (GO:0003824), Biological Process: nucleotide-sugar metabolism (GO:0009225), Molecular Function: NAD binding (GO:0051287) |
| 19 (96) | 12 (229) | 38 (115) | IPR001609: Myosin head, motor region | Molecular Function: motor activity (GO:0003774), Molecular Function: ATP binding (GO:0005524), Cellular Component: myosin (GO:0016459) |
| 19 (97) | 15 (186) | 44 (99) | IPR001683: Phox-like | Biological Process: intracellular signaling cascade (GO:0007242) |
| 19 (98) | 23 (101) | 37 (118) | IPR001828: Extracellular ligand-binding receptor | |
| 19 (99) | 32 (65) | 27 (171) | IPR004843: Metallophosphoesterase | Molecular Function: hydrolase activity (GO:0016787) |
| 19 (100) | 17 (152) | 68 (56) | IPR006652: Kelch repeat | |
| 19 (101) | 17 (153) | 40 (112) | IPR006670: Cyclin | |
| 18 (102) | 34 (58) | 32 (141) | IPR000301: CD9/CD37/CD63 antigen | Cellular Component: integral to membrane (GO:0016021) |
| 18 (103) | 16 (161) | 49 (80) | IPR000536: Nuclear hormone receptor, ligand-binding | Molecular Function: transcription factor activity (GO:0003700), Molecular Function: steroid hormone receptor activity (GO:0003707), Cellular Component: nucleus (GO:0005634), Biological Process: regulation of transcription, DNA-dependent (GO:0006355) |
| 18 (104) | 24 (96) | 48 (86) | IPR000859: CUB | |

| | | | | |
|---|---|---|---|---|
| 18 (105) | 25 (90) | 20 (236) | IPR001054: Adenylyl cyclase class-3/4/guanylyl cyclase | Biological Process: intracellular signaling cascade (GO:0007242), Biological Process: cyclic nucleotide biosynthesis (GO:0009190), Molecular Function: phosphorus-oxygen lyase activity (GO:0016849) |
| 18 (106) | 18 (135) | 21 (226) | IPR001163: Like-Sm ribonucleoprotein, core | Cellular Component: nucleus (GO:0005634), Cellular Component: small nucleolar ribonucleoprotein complex (GO:0005732), Biological Process: mRNA processing (GO:0006397) |
| 18 (107) | 18 (136) | 40 (109) | IPR001487: Bromodomain | |
| 18 (108) | 26 (87) | 38 (117) | IPR005834: Haloacid dehalogenase-like hydrolase | Molecular Function: catalytic activity (GO:0003824), Biological Process: metabolism (GO:0008152) |
| 18 (109) | 16 (175) | 68 (57) | IPR011498: Kelch | |
| 18 (110) | 18 (143) | 40 (114) | IPR011700: Basic leucine zipper | Molecular Function: DNA binding (GO:0003677), Cellular Component: nucleus (GO:0005634), Biological Process: regulation of transcription, DNA-dependent (GO:0006355) |
| 17 (111) | 21 (113) | 19 (247) | IPR000182: GCN5-related N-acetyltransferase | Molecular Function: N-acetyltransferase activity (GO:0008080) |
| 17 (112) | 17 (148) | **80 (50)** | IPR002126: Cadherin | Molecular Function: calcium ion binding (GO:0005509), Biological Process: homophilic cell adhesion (GO:0007156), Cellular Component: membrane (GO:0016020) |
| 17 (113) | 16 (170) | 20 (241) | IPR002893: Zinc finger, MYND-type | |

| | | | | |
|---|---|---|---|---|
| 17 (114) | 17 (149) | 21 (228) | IPR002917: GTP-binding protein, HSR1-related | |
| 17 (115) | 19 (131) | 22 (218) | IPR006662: Thioredoxin-related | Molecular Function: electron transporter activity (GO:0005489), Biological Process: electron transport (GO:0006118) |
| 17 (116) | 16 (177) | 31 (151) | IPR012679: Laminin G, subdomain 1 | |
| 16 (117) | 15 (179) | 25 (177) | IPR000253: Forkhead-associated | |
| 16 (118) | 12 (225) | 27 (169) | IPR000306: Zinc finger, FYVE-type | Molecular Function: zinc ion binding (GO:0008270) |
| 16 (119) | 17 (145) | 31 (146) | IPR000330: SNF2-related | Molecular Function: DNA binding (GO:0003677), Molecular Function: ATP binding (GO:0005524) |
| 16 (120) | 15 (180) | **86 (43)** | IPR000372: Cysteine-rich flanking region, N-terminal | |
| 16 (121) | 16 (160) | 20 (233) | IPR000467: D111/G-patch | Molecular Function: nucleic acid binding (GO:0003676), Cellular Component: intracellular (GO:0005622) |
| 16 (122) | 14 (201) | 16 (295) | IPR000717: Proteasome component region PCI | |
| 16 (123) | 18 (134) | 29 (162) | IPR000795: Protein synthesis factor, GTP-binding | Molecular Function: GTP binding (GO:0005525), Biological Process: protein biosynthesis (GO:0006412) |
| 16 (124) | 14 (206) | 33 (140) | IPR004148: BAR | Molecular Function: protein binding (GO:0005515), Cellular Component: cytoplasm (GO:0005737) |
| 16 (125) | 18 (141) | 27 (172) | IPR006594: Lissencephaly type-1-like homology motif | |
| 15 (126) | 40 (48) | 24 (183) | IPR000073: Alpha/beta hydrolase fold | |

| | | | | |
|---|---|---|---|---|
| 15 (127) | 16 (158) | 41 (106) | IPR000242: Tyrosine specific protein phosphatase | Molecular Function: protein tyrosine phosphatase activity (GO:0004725), Biological Process: protein amino acid dephosphorylation (GO:0006470) |
| 15 (128) | 9 (293) | **9 (504)** | IPR000560: Histidine acid phosphatase | Molecular Function: acid phosphatase activity (GO:0003993) |
| 15 (129) | 28 (74) | 27 (170) | IPR000873: AMP-dependent synthetase and ligase | Molecular Function: catalytic activity (GO:0003824), Biological Process: metabolism (GO:0008152) |
| 15 (130) | 22 (105) | 18 (263) | IPR001320: Ionotropic glutamate receptor | Molecular Function: ionotropic glutamate receptor activity (GO:0004970), Molecular Function: glutamate-gated ion channel activity (GO:0005234), Cellular Component: membrane (GO:0016020) |
| 15 (131) | 18 (137) | 42 (104) | IPR001766: Fork head transcription factor | Molecular Function: transcription factor activity (GO:0003700), Cellular Component: nucleus (GO:0005634), Biological Process: regulation of transcription, DNA-dependent (GO:0006355) |
| 15 (132) | 9 (300) | 26 (173) | IPR002017: Spectrin repeat | |
| 15 (133) | 16 (168) | 37 (119) | IPR002130: Peptidyl-prolyl cis-trans isomerase, cyclophilin type | Biological Process: protein folding (GO:0006457) |
| 15 (134) | 12 (235) | 35 (128) | IPR003008: Tubulin/FtsZ, GTPase | |
| 15 (135) | 15 (189) | 34 (134) | IPR004000: Actin/actin-like | Molecular Function: protein binding (GO:0005515) |

| | | | | |
|---|---|---|---|---|
| 15 (136) | 10 (279) | 13 (377) | IPR004273: Dynein heavy chain | Molecular Function: microtubule motor activity (GO:0003777), Biological Process: microtubule-based movement (GO:0007018), Cellular Component: dynein complex (GO:0030286) |
| 15 (137) | 15 (192) | 22 (217) | IPR004841: Amino acid permease-associated region | Biological Process: transport (GO:0006810), Cellular Component: membrane (GO:0016020) |
| 15 (138) | 13 (222) | 30 (159) | IPR006671: Cyclin, N-terminal | Biological Process: regulation of progression through cell cycle (GO:0000074) |
| 15 (139) | 10 (289) | 64 (60) | IPR008160: Collagen triple helix repeat | Cellular Component: cytoplasm (GO:0005737), Biological Process: phosphate transport (GO:0006817) |
| 14 (140) | 11 (244) | 36 (123) | IPR000159: RA | Biological Process: signal transduction (GO:0007165) |
| 14 (141) | 21 (112) | 21 (221) | IPR000175: Sodium:neurotransmitter symporter | Molecular Function: neurotransmitter:sodium symporter activity (GO:0005328), Cellular Component: integral to plasma membrane (GO:0005887), Biological Process: neurotransmitter transport (GO:0006836), Cellular Component: membrane (GO:0016020) |
| 14 (142) | 17 (146) | **5 (811)** | IPR000412: ABC-2 | Molecular Function: ATP binding (GO:0005524), Biological Process: transport (GO:0006810), Cellular Component: membrane (GO:0016020), Molecular Function: ATPase activity, coupled to transmembrane movement of substances (GO:0042626) |

| | | | | |
|---|---|---|---|---|
| 14 (143) | 14 (200) | 28 (167) | IPR000569: HECT | Molecular Function: ubiquitin-protein ligase activity (GO:0004842), Cellular Component: intracellular (GO:0005622), Biological Process: protein modification (GO:0006464), Biological Process: ubiquitin cycle (GO:0006512) |
| 14 (144) | 32 (61) | **9 (505)** | IPR000734: Lipase | Molecular Function: catalytic activity (GO:0003824), Biological Process: lipid metabolism (GO:0006629) |
| 14 (145) | 12 (232) | **81 (49)** | IPR002035: von Willebrand factor, type A | |
| 14 (146) | 20 (125) | **7 (646)** | IPR002076: GNS1/SUR4 membrane protein | Cellular Component: integral to membrane (GO:0016021) |
| 14 (147) | 22 (108) | 40 (111) | IPR002350: Proteinase inhibitor I1, Kazal | |
| 14 (148) | 16 (171) | 25 (182) | IPR004161: Elongation factor Tu, domain 2 | Molecular Function: GTP binding (GO:0005525) |
| 14 (149) | 10 (285) | 15 (327) | IPR006935: Type III restriction enzyme, res subunit | Molecular Function: ATP binding (GO:0005524) |
| 14 (150) | 14 (208) | 18 (273) | IPR007502: Helicase-associated region | Molecular Function: helicase activity (GO:0004386) |
| 14 (151) | 12 (242) | 36 (127) | IPR008280: Tubulin/FtsZ, C-terminal | Molecular Function: GTPase activity (GO:0003924), Molecular Function: GTP binding (GO:0005525), Cellular Component: protein complex (GO:0043234), Biological Process: protein polymerization (GO:0051258) |
| 14 (152) | 14 (209) | 37 (121) | IPR011616: bZIP transcription factor, bZIP_1 | Molecular Function: DNA binding (GO:0003677), Cellular Component: nucleus (GO:0005634), Biological Process: regulation of transcription, |

| | | | | |
|---|---|---|---|---|
| | | | | DNA-dependent (GO:0006355) |
| 13 (153) | 16 (157) | 30 (152) | IPR000014: PAS | Molecular Function: signal transducer activity (GO:0004871), Biological Process: signal transduction (GO:0007165) |
| 13 (154) | 22 (104) | 23 (194) | IPR001140: ABC transporter, transmembrane region | Molecular Function: ATP binding (GO:0005524), Biological Process: transport (GO:0006810), Cellular Component: integral to membrane (GO:0016021), Molecular Function: ATPase activity, coupled to transmembrane movement of substances (GO:0042626) |
| 13 (155) | 26 (83) | 19 (253) | IPR001353: 20S proteasome, A and B subunits | Molecular Function: threonine endopeptidase activity (GO:0004298), Cellular Component: proteasome core complex (sensu Eukaryota) (GO:0005839), Biological Process: ubiquitin-dependent protein catabolism (GO:0006511) |
| 13 (156) | 15 (185) | 14 (345) | IPR001507: Endoglin/CD105 antigen | |
| 13 (157) | 12 (230) | 25 (179) | IPR001610: PAC motif | Biological Process: regulation of transcription, DNA-dependent (GO:0006355), Biological Process: signal transduction (GO:0007165) |
| 13 (158) | 11 (249) | 30 (156) | IPR002049: Laminin-type EGF-like | Molecular Function: structural molecule activity (GO:0005198) |

| | | | | |
|---|---|---|---|---|
| 13 (159) | 15 (188) | 37 (120) | IPR003595: Protein tyrosine phosphatase, catalytic region | Molecular Function: protein tyrosine phosphatase activity (GO:0004725) |
| 13 (160) | **54 (41)** | **2 (1694)** | IPR004119: Protein of unknown function DUF227 | Molecular Function: molecular function unknown (GO:0005554) |
| 13 (161) | 28 (77) | **()** | IPR004272: Odorant binding protein | Molecular Function: molecular function unknown (GO:0005554) |
| 13 (162) | 36 (54) | **()** | IPR006170: Pheromone/general odorant binding protein, PBP/GOBP | Molecular Function: odorant binding (GO:0005549), Biological Process: transport (GO:0006810) |
| 13 (163) | 13 (223) | 29 (166) | IPR006689: ARF/SAR superfamily | Molecular Function: GTP binding (GO:0005525) |
| 13 (164) | 22 (110) | 40 (113) | IPR011497: Protease inhibitor, Kazal-type | |
| 13 (165) | 19 (132) | 50 (79) | IPR011705: BTB/Kelch-associated | |
| 13 (166) | 20 (130) | **()** | IPR012464: Protein of unknown function DUF1676 | |
| 12 (167) | 12 (226) | 41 (107) | IPR000340: Dual specificity protein phosphatase | Biological Process: protein amino acid dephosphorylation (GO:0006470), Molecular Function: protein tyrosine/serine/threonine phosphatase activity (GO:0008138) |
| 12 (168) | 11 (245) | 19 (249) | IPR000403: Phosphatidylinositol 3- and 4-kinase, catalytic | Molecular Function: phosphotransferase activity, alcohol group as acceptor (GO:0016773) |
| 12 (169) | 14 (199) | 17 (276) | IPR000433: Zinc finger, ZZ-type | Molecular Function: zinc ion binding (GO:0008270) |
| 12 (170) | 13 (210) | 20 (237) | IPR001159: Double-stranded RNA binding | Molecular Function: double-stranded RNA binding (GO:0003725), Cellular Component: intracellular (GO:0005622) |
| 12 (171) | 16 (165) | 12 (389) | IPR001199: Cytochrome b5 | |

| | | | | |
|---|---|---|---|---|
| 12 (172) | 12 (228) | 22 (211) | IPR001357: BRCT | Cellular Component: intracellular (GO:0005622) |
| 12 (173) | 14 (204) | 22 (214) | IPR001876: Zinc finger, RanBP2-type | |
| 12 (174) | 10 (271) | 18 (266) | IPR002123: Phospholipid/glycerol acyltransferase | Biological Process: metabolism (GO:0008152), Molecular Function: acyltransferase activity (GO:0008415) |
| 12 (175) | 12 (234) | 14 (348) | IPR002225: 3-beta hydroxysteroid dehydrogenase/isomerase | Molecular Function: 3-beta-hydroxy-delta5-steroid dehydrogenase activity (GO:0003854), Biological Process: steroid biosynthesis (GO:0006694) |
| 12 (176) | 16 (169) | 18 (267) | IPR002422: Amino acid/polyamine transporter II | Molecular Function: amino acid-polyamine transporter activity (GO:0005279), Biological Process: amino acid transport (GO:0006865), Cellular Component: membrane (GO:0016020) |
| 12 (177) | 12 (239) | **87 (41)** | IPR003877: SPla/RYanodine receptor SPRY | |
| 12 (178) | 9 (313) | 9 (552) | IPR011704: ATPase associated with various cellular activities, AAA_5 | Molecular Function: ATP binding (GO:0005524), Molecular Function: ATPase activity (GO:0016887) |
| 11 (179) | 14 (198) | 23 (191) | IPR000086: NUDIX hydrolase | |
| 11 (180) | 6 (410) | **74 (52)** | IPR000315: Zinc finger, B-box | Cellular Component: intracellular (GO:0005622), Molecular Function: zinc ion binding (GO:0008270) |
| 11 (181) | 9 (292) | 17 (275) | IPR000408: Regulator of chromosome condensation, RCC1 | |
| 11 (182) | 16 (159) | 54 (69) | IPR000436: Sushi/SCR/CCP | |

| | | | | |
|---|---|---|---|---|
| 11 (183) | 20 (124) | 30 (153) | IPR000953: Chromo | Cellular Component: chromatin (GO:0000785), Molecular Function: chromatin binding (GO:0003682), Cellular Component: nucleus (GO:0005634), Biological Process: chromatin assembly or disassembly (GO:0006333) |
| 11 (184) | 15 (181) | 36 (124) | IPR001007: von Willebrand factor, type C | |
| 11 (185) | 21 (117) | 22 (213) | IPR001594: Zinc finger, DHHC-type | Molecular Function: metal ion binding (GO:0046872) |
| 11 (186) | 15 (187) | 18 (265) | IPR001932: Protein phosphatase 2C-like | Molecular Function: catalytic activity (GO:0003824) |
| 11 (187) | **35 (57)** | 14 (347) | IPR002213: UDP-glucuronosyl/UDP-glucosyltransferase | Biological Process: metabolism (GO:0008152), Molecular Function: transferase activity, transferring hexosyl groups (GO:0016758) |
| 11 (188) | 12 (236) | 51 (76) | IPR003131: K+ channel tetramerisation | Molecular Function: voltage-gated potassium channel activity (GO:0005249), Biological Process: potassium ion transport (GO:0006813), Cellular Component: voltage-gated potassium channel complex (GO:0008076), Cellular Component: membrane (GO:0016020) |
| 11 (189) | 13 (217) | 30 (158) | IPR003347: Transcription factor jumonji, jmjC | |
| 11 (190) | 10 (276) | 24 (186) | IPR003577: Ras small GTPase, Ras type | Molecular Function: GTP binding (GO:0005525), Biological Process: small GTPase mediated signal transduction (GO:0007264) |

| | | | | |
|---|---|---|---|---|
| 11 (191) | 12 (237) | 22 (215) | IPR003604: Zinc finger, U1-type | Molecular Function: nucleic acid binding (GO:0003676), Cellular Component: nucleus (GO:0005634), Molecular Function: zinc ion binding (GO:0008270) |
| 11 (192) | 11 (252) | 45 (96) | IPR003659: Plexin/semaphorin/integrin | Biological Process: development (GO:0007275) |
| 11 (193) | 7 (385) | 10 (480) | IPR003958: Transcription factor CBF/NF-Y/archaeal histone | Molecular Function: DNA binding (GO:0003677) |
| 11 (194) | 11 (255) | 10 (485) | IPR004344: Tubulin-tyrosine ligase | Molecular Function: tubulin-tyrosine ligase activity (GO:0004835), Biological Process: protein modification (GO:0006464) |
| 11 (195) | 8 (349) | 23 (205) | IPR008250: E1-E2 ATPase-associated region | Molecular Function: ATP binding (GO:0005524), Cellular Component: membrane (GO:0016020), Molecular Function: hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances (GO:0016820) |
| 11 (196) | 13 (224) | 17 (293) | IPR011709: Protein of unknown function DUF1605 | Molecular Function: ATP binding (GO:0005524), Molecular Function: ATP-dependent helicase activity (GO:0008026) |
| 10 (197) | 8 (315) | 19 (248) | IPR000313: PWWP | |
| 10 (198) | 10 (263) | 25 (178) | IPR000337: GPCR, family 3, metabotropic glutamate receptor-like | Molecular Function: metabotropic glutamate, GABA-B-like receptor activity (GO:0008067), Cellular Component: membrane (GO:0016020) |
| 10 (199) | 8 (316) | 31 (147) | IPR000342: Regulator of G protein signalling | Molecular Function: signal transducer activity (GO:0004871) |

| | | | | |
|---|---|---|---|---|
| 10 (200) | 22 (102) | 45 (92) | IPR000832: GPCR, family 2, secretin-like | Molecular Function: G-protein coupled receptor activity (GO:0004930), Cellular Component: membrane (GO:0016020) |
| 10 (201) | 24 (95) | 22 (208) | IPR000834: Peptidase M14, carboxypeptidase A | Molecular Function: carboxypeptidase A activity (GO:0004182), Biological Process: proteolysis and peptidolysis (GO:0006508) |
| 10 (202) | 15 (183) | 22 (209) | IPR001148: Carbonic anhydrase, eukaryotic | Molecular Function: carbonate dehydratase activity (GO:0004089), Biological Process: one-carbon compound metabolism (GO:0006730), Molecular Function: zinc ion binding (GO:0008270) |
| 10 (203) | 16 (164) | 23 (195) | IPR001173: Glycosyl transferase, family 2 | |
| 10 (204) | 16 (166) | 7 (634) | IPR001223: Glycoside hydrolase, family 18 | Molecular Function: hydrolase activity, hydrolyzing O-glycosyl compounds (GO:0004553), Biological Process: carbohydrate metabolism (GO:0005975) |
| 10 (205) | **35 (55)** | **81 (47)** | IPR001304: C-type lectin | Molecular Function: sugar binding (GO:0005529) |
| 10 (206) | 6 (431) | 40 (110) | IPR001590: Peptidase M12B, ADAM/reprolysin | Molecular Function: metalloendopeptidase activity (GO:0004222), Biological Process: proteolysis and peptidolysis (GO:0006508) |
| 10 (207) | 14 (203) | 17 (286) | IPR001753: Enoyl-CoA hydratase/isomerase | Molecular Function: catalytic activity (GO:0003824), Biological Process: metabolism (GO:0008152) |

| | | | | |
|---|---|---|---|---|
| 10 (208) | 22 (107) | 13 (370) | IPR001930: Peptidase M1, membrane alanine aminopeptidase | Molecular Function: membrane alanyl aminopeptidase activity (GO:0004179), Biological Process: proteolysis and peptidolysis (GO:0006508) |
| 10 (209) | 7 (374) | 20 (240) | IPR002085: Alcohol dehydrogenase superfamily, zinc-containing | Molecular Function: zinc ion binding (GO:0008270), Molecular Function: oxidoreductase activity (GO:0016491) |
| 10 (210) | 10 (272) | 10 (470) | IPR002129: Pyridoxal-dependent decarboxylase | Biological Process: amino acid metabolism (GO:0006520), Molecular Function: carboxy-lyase activity (GO:0016831) |
| 10 (211) | 13 (215) | 17 (289) | IPR002423: Chaperonin Cpn60/TCP-1 | Molecular Function: protein binding (GO:0005515), Molecular Function: ATP binding (GO:0005524), Biological Process: cellular protein metabolism (GO:0044267) |
| 10 (212) | 13 (216) | 26 (174) | IPR002999: Tudor | Molecular Function: nucleic acid binding (GO:0003676) |
| 10 (213) | 10 (274) | 12 (395) | IPR003029: RNA binding S1 | Molecular Function: RNA binding (GO:0003723) |
| 10 (214) | 10 (277) | 23 (202) | IPR003607: Metal-dependent phosphohydrolase, HD region | Molecular Function: catalytic activity (GO:0003824) |
| 10 (215) | 10 (280) | 15 (323) | IPR004839: Aminotransferase, class I and II | Biological Process: biosynthesis (GO:0009058), Molecular Function: transferase activity, transferring nitrogenous groups (GO:0016769) |
| 10 (216) | 12 (240) | 42 (105) | IPR005135: Endonuclease/exonuclease/phosphatase | |
| 10 (217) | 11 (256) | 31 (149) | IPR006020: Phosphotyrosine interaction region | |

| | | | | |
|---|---|---|---|---|
| 10 (218) | 29 (73) | **()** | IPR006625: Insect pheromone/odorant binding protein PhBP | |
| 10 (219) | 10 (284) | 22 (219) | IPR006688: ADP-ribosylation factor | Molecular Function: GTP binding (GO:0005525) |
| 9 (220) | 10 (265) | 13 (364) | IPR000555: Mov34/MPN/PAD-1 | |
| 9 (221) | 8 (320) | 10 (447) | IPR000594: UBA/THIF-type NAD/FAD binding fold | Molecular Function: catalytic activity (GO:0003824) |
| 9 (222) | 12 (227) | 21 (223) | IPR000727: Target SNARE coiled-coil region | |
| 9 (223) | 7 (363) | 32 (142) | IPR001164: Arf GTPase activating protein | Biological Process: regulation of GTPase activity (GO:0043087) |
| 9 (224) | 6 (427) | 11 (414) | IPR001258: NHL repeat | |
| 9 (225) | 11 (248) | 22 (212) | IPR001395: Aldo/keto reductase | Molecular Function: oxidoreductase activity (GO:0016491) |
| 9 (226) | 8 (327) | 32 (143) | IPR001496: SOCS protein, C-terminal | Biological Process: intracellular signaling cascade (GO:0007242) |
| 9 (227) | 16 (167) | 11 (417) | IPR001734: Na+/solute symporter | Molecular Function: transporter activity (GO:0005215), Biological Process: transport (GO:0006810), Cellular Component: membrane (GO:0016020) |
| 9 (228) | 10 (269) | 10 (467) | IPR002007: Animal haem peroxidase | Molecular Function: peroxidase activity (GO:0004601) |
| 9 (229) | 10 (270) | 8 (577) | IPR002068: Heat shock protein Hsp20 | |
| 9 (230) | 11 (250) | 19 (255) | IPR002086: Aldehyde dehydrogenase | Biological Process: metabolism (GO:0008152), Molecular Function: oxidoreductase activity (GO:0016491) |
| 9 (231) | 12 (233) | **3 (1246)** | IPR002159: CD36 antigen | Biological Process: cell adhesion (GO:0007155), Cellular Component: membrane (GO:0016020) |

| | | | | |
|---|---|---|---|---|
| 9 (232) | 8 (331) | 8 (578) | IPR002300: Aminoacyl-tRNA synthetase, class Ia | Molecular Function: tRNA ligase activity (GO:0004812), Molecular Function: ATP binding (GO:0005524), Biological Process: tRNA aminoacylation for protein translation (GO:0006418) |
| 9 (233) | 9 (301) | 11 (420) | IPR002314: tRNA synthetase, class II (G, H, P and S) | Molecular Function: tRNA ligase activity (GO:0004812), Molecular Function: ATP binding (GO:0005524), Biological Process: tRNA aminoacylation for protein translation (GO:0006418) |
| 9 (234) | 4 (654) | 13 (375) | IPR003616: SET-related region | |
| 9 (235) | 8 (338) | 20 (244) | IPR004012: RUN | |
| 9 (236) | 11 (254) | 10 (484) | IPR004274: NLI interacting factor | |
| 9 (237) | 10 (282) | 8 (602) | IPR005804: Fatty acid desaturase | Molecular Function: oxidoreductase activity (GO:0016491) |
| 9 (238) | 17 (150) | 21 (230) | IPR006076: FAD dependent oxidoreductase | Biological Process: electron transport (GO:0006118), Molecular Function: oxidoreductase activity (GO:0016491) |
| 9 (239) | 6 (472) | 14 (357) | IPR006560: AWS | |
| 9 (240) | **51 (42)** | **5 (915)** | IPR006578: MADF | |
| 9 (241) | 9 (310) | 26 (176) | IPR008145: Guanylate kinase/L-type calcium channel region | |
| 9 (242) | 15 (197) | 12 (401) | IPR008191: Maternal tudor protein | |

| | | | | |
|---|---|---|---|---|
| 9 (243) | 9 (312) | 10 (494) | IPR011547: Sulphate transporter | Molecular Function: transporter activity (GO:0005215), Biological Process: transport (GO:0006810), Cellular Component: integral to membrane (GO:0016021) |
| 9 (244) | 16 (176) | 7 (693) | IPR011583: Chitinase II | Molecular Function: chitinase activity (GO:0004568), Biological Process: chitin catabolism (GO:0006032) |
| 8 (245) | 3 (726) | 6 (699) | IPR000092: Polyprenyl synthetase | Biological Process: isoprenoid biosynthesis (GO:0008299) |
| 8 (246) | 6 (413) | 24 (184) | IPR000413: Integrins alpha chain | Biological Process: cell adhesion (GO:0007155), Biological Process: cell-matrix adhesion (GO:0007160), Cellular Component: integrin complex (GO:0008305) |
| 8 (247) | 8 (317) | 29 (161) | IPR000418: Ets | Molecular Function: transcription factor activity (GO:0003700), Cellular Component: nucleus (GO:0005634), Biological Process: regulation of transcription, DNA-dependent (GO:0006355) |
| 8 (248) | 17 (147) | 15 (308) | IPR000637: HMG-I and HMG-Y, DNA-binding | |
| 8 (249) | 21 (116) | 7 (627) | IPR000718: Peptidase M13, neprilysin | Molecular Function: neprilysin activity (GO:0004245), Biological Process: proteolysis and peptidolysis (GO:0006508), Cellular Component: membrane (GO:0016020) |
| 8 (250) | 10 (268) | 15 (311) | IPR000917: Sulfatase | Biological Process: metabolism (GO:0008152), Molecular Function: sulfuric ester hydrolase activity (GO:0008484) |

| | | | | |
|---|---|---|---|---|
| 8 (251) | 9 (296) | 10 (454) | IPR000994: Peptidase M24 | Biological Process: proteolysis and peptidolysis (GO:0006508), Molecular Function: metalloexopeptidase activity (GO:0008235) |
| 8 (252) | 11 (247) | 11 (413) | IPR001107: Band 7 protein | |
| 8 (253) | 8 (323) | 14 (341) | IPR001206: Diacylglycerol kinase, catalytic region | Molecular Function: diacylglycerol kinase activity (GO:0004143), Biological Process: protein kinase C activation (GO:0007205) |
| 8 (254) | 7 (364) | 8 (567) | IPR001208: MCM | Molecular Function: DNA binding (GO:0003677), Molecular Function: ATP binding (GO:0005524), Biological Process: DNA replication initiation (GO:0006270), Molecular Function: DNA-dependent ATPase activity (GO:0008094) |
| 8 (255) | 6 (425) | 8 (568) | IPR001236: Lactate/malate dehydrogenase | Biological Process: tricarboxylic acid cycle intermediate metabolism (GO:0006100), Molecular Function: oxidoreductase activity (GO:0016491) |
| 8 (256) | 15 (184) | 12 (390) | IPR001494: Importin-beta, N-terminal | Biological Process: protein-nucleus import, docking (GO:0000059), Cellular Component: nucleus (GO:0005634), Cellular Component: nuclear pore (GO:0005643), Cellular Component: cytoplasm (GO:0005737), Molecular Function: protein transporter activity (GO:0008565) |

| | | | | |
|---|---|---|---|---|
| 8 (257) | 7 (370) | 12 (391) | IPR001529: DNA-directed RNA polymerase, M/15 kDa subunit | Molecular Function: DNA binding (GO:0003677), Molecular Function: DNA-directed RNA polymerase activity (GO:0003899), Biological Process: transcription (GO:0006350) |
| 8 (258) | 6 (436) | 20 (238) | IPR001763: Rhodanese-like | |
| 8 (259) | 6 (441) | 20 (239) | IPR002073: 3'5'-cyclic nucleotide phosphodiesterase | Molecular Function: 3',5'-cyclic-nucleotide phosphodiesterase activity (GO:0004114), Biological Process: signal transduction (GO:0007165) |
| 8 (260) | 8 (330) | 29 (164) | IPR002165: Plexin | Molecular Function: receptor activity (GO:0004872), Cellular Component: membrane (GO:0016020) |
| 8 (261) | 4 (643) | 10 (474) | IPR002919: Protease inhibitor I8, cysteine-rich trypsin inhibitor-like | |
| 8 (262) | 5 (534) | 10 (477) | IPR003392: Patched | Molecular Function: hedgehog receptor activity (GO:0008158), Cellular Component: membrane (GO:0016020) |
| 8 (263) | 6 (456) | 6 (753) | IPR003395: SMC protein, N-terminal | Molecular Function: ATP binding (GO:0005524), Cellular Component: chromosome (GO:0005694), Biological Process: chromosome organization and biogenesis (GO:0051276) |
| 8 (264) | 7 (383) | 20 (243) | IPR003594: ATP-binding region, ATPase-like | Molecular Function: ATP binding (GO:0005524) |
| 8 (265) | 9 (305) | 14 (349) | IPR003689: Zinc/iron permease | Cellular Component: membrane (GO:0016020), Biological Process: metal ion transport (GO:0030001), Molecular Function: metal ion transporter activity (GO:0046873) |

| | | | | |
|---|---|---|---|---|
| 8 (266) | 8 (337) | 18 (269) | IPR003903: Ubiquitin interacting motif | |
| 8 (267) | **39 (51)** | 23 (203) | IPR004045: Glutathione S-transferase, N-terminal | |
| 8 (268) | 8 (344) | 18 (271) | IPR005824: KOW | |
| 8 (269) | 7 (393) | 13 (380) | IPR006055: Exonuclease | Molecular Function: exonuclease activity (GO:0004527), Cellular Component: intracellular (GO:0005622) |
| 8 (270) | 10 (283) | 11 (434) | IPR006090: Acyl-CoA dehydrogenase, C-terminal | Biological Process: electron transport (GO:0006118), Molecular Function: oxidoreductase activity (GO:0016491) |
| 8 (271) | 13 (219) | 15 (324) | IPR006091: Acyl-CoA dehydrogenase, central region | Molecular Function: acyl-CoA dehydrogenase activity (GO:0003995), Biological Process: electron transport (GO:0006118) |
| 8 (272) | 6 (471) | 18 (272) | IPR006212: Furin-like repeat | |
| 8 (273) | 9 (307) | 15 (325) | IPR006553: Leucine-rich repeat, cysteine-containing subtype | |
| 8 (274) | 9 (308) | 11 (437) | IPR006575: RWD | |
| 8 (275) | 20 (129) | **1 (2460)** | IPR006631: Protein of unknown function DM4/12 | |
| 8 (276) | 7 (397) | 7 (687) | IPR007863: Peptidase M16, C-terminal | |
| 8 (277) | 18 (142) | 7 (689) | IPR008753: Peptidase M13 | Biological Process: proteolysis and peptidolysis (GO:0006508), Molecular Function: metallopeptidase activity (GO:0008237) |
| 8 (278) | 17 (155) | 9 (551) | IPR011022: Arrestin, C-terminal | |

| | | | | |
|---|---|---|---|---|
| 7 (279) | 10 (261) | 15 (303) | IPR000033: Low-density lipoprotein receptor, YWTD repeat | Cellular Component: membrane (GO:0016020) |
| 7 (280) | 5 (485) | 7 (619) | IPR000061: SWAP/Surp | Molecular Function: RNA binding (GO:0003723), Biological Process: RNA processing (GO:0006396) |
| 7 (281) | 6 (406) | 9 (497) | IPR000089: Biotin/lipoyl attachment | |
| 7 (282) | **29 (70)** | 34 (130) | IPR000215: Proteinase inhibitor I4, serpin | Molecular Function: serine-type endopeptidase inhibitor activity (GO:0004867) |
| 7 (283) | 5 (491) | 21 (222) | IPR000421: Coagulation factor 5/8 type, C-terminal | Biological Process: cell adhesion (GO:0007155) |
| 7 (284) | 8 (318) | 33 (136) | IPR000488: Death | Molecular Function: protein binding (GO:0005515), Biological Process: signal transduction (GO:0007165) |
| 7 (285) | 6 (416) | 5 (817) | IPR000640: Elongation factor G, C-terminal | Molecular Function: GTP binding (GO:0005525) |
| 7 (286) | 5 (494) | 17 (277) | IPR000644: CBS | |
| 7 (287) | 6 (417) | 12 (387) | IPR000679: Zinc finger, GATA-type | Molecular Function: transcription factor activity (GO:0003700), Cellular Component: nucleus (GO:0005634), Biological Process: regulation of transcription, DNA-dependent (GO:0006355) |
| 7 (288) | 9 (294) | 5 (821) | IPR000760: Inositol monophosphatase | Molecular Function: inositol or phosphatidylinositol phosphatase activity (GO:0004437) |
| 7 (289) | 10 (267) | 31 (148) | IPR000863: Sulfotransferase | Molecular Function: sulfotransferase activity (GO:0008146) |
| 7 (290) | 9 (295) | 6 (711) | IPR000866: Alkyl hydroperoxide reductase/ Thiol specific antioxidant/ | |

| | | | | |
|---|---|---|---|---|
| | | | Mal allergen | |
| 7 (291) | 7 (359) | 9 (509) | IPR000938: CAP-Gly | |
| 7 (292) | 6 (421) | 10 (456) | IPR001025: Bromo adjacent region | Molecular Function: DNA binding (GO:0003677) |
| 7 (293) | 5 (504) | 18 (261) | IPR001060: Cdc15/Fes/CIP4 | |
| 7 (294) | 6 (423) | 18 (262) | IPR001093: IMP dehydrogenase/GMP reductase | Molecular Function: catalytic activity (GO:0003824) |
| 7 (295) | 8 (322) | 19 (252) | IPR001179: Peptidylprolyl isomerase, FKBP-type | Biological Process: protein folding (GO:0006457) |
| 7 (296) | 6 (426) | 7 (635) | IPR001247: 3' exoribonuclease | Molecular Function: 3'-5'-exoribonuclease activity (GO:0000175), Molecular Function: RNA binding (GO:0003723), Biological Process: RNA processing (GO:0006396) |
| 7 (297) | 7 (368) | 9 (517) | IPR001375: Peptidase S9, prolyl oligopeptidase active site region | Biological Process: proteolysis and peptidolysis (GO:0006508), Molecular Function: serine-type peptidase activity (GO:0008236) |
| 7 (298) | 9 (299) | 9 (520) | IPR001523: Paired box protein, N-terminal | Cellular Component: nucleus (GO:0005634), Biological Process: development (GO:0007275) |
| 7 (299) | 6 (433) | 13 (369) | IPR001606: AT-rich interaction region | Molecular Function: DNA binding (GO:0003677), Cellular Component: intracellular (GO:0005622) |
| 7 (300) | 7 (372) | 36 (125) | IPR001839: Transforming growth factor beta | Molecular Function: growth factor activity (GO:0008083) |

| | | | | |
|---|---|---|---|---|
| 7 (301) | 25 (91) | 8 (575) | IPR001873: Na+ channel, amiloride-sensitive | Molecular Function: sodium channel activity (GO:0005272), Biological Process: sodium ion transport (GO:0006814), Cellular Component: membrane (GO:0016020) |
| 7 (302) | 8 (329) | 29 (163) | IPR001895: Guanine-nucleotide dissociation stimulator CDC25 | Molecular Function: guanyl-nucleotide exchange factor activity (GO:0005085), Biological Process: intracellular signaling cascade (GO:0007242) |
| 7 (303) | 2 (1023) | 7 (644) | IPR001991: Sodium:dicarboxylate symporter | Biological Process: dicarboxylic acid transport (GO:0006835), Cellular Component: membrane (GO:0016020), Molecular Function: sodium:dicarboxylate symporter activity (GO:0017153) |
| 7 (304) | 7 (376) | 13 (371) | IPR002553: Adaptin, N-terminal | |
| 7 (305) | 10 (273) | 30 (157) | IPR002909: Cell surface receptor IPT/TIG | |
| 7 (306) | 7 (379) | 9 (531) | IPR003014: N/apple PAN | |
| 7 (307) | 5 (531) | 19 (256) | IPR003034: DNA-binding SAP | Molecular Function: DNA binding (GO:0003677) |
| 7 (308) | 7 (380) | 8 (588) | IPR003107: RNA-processing protein, HAT helix | Cellular Component: intracellular (GO:0005622), Biological Process: RNA processing (GO:0006396) |
| 7 (309) | 8 (335) | 15 (321) | IPR003409: MORN motif | |
| 7 (310) | 6 (457) | 23 (201) | IPR003578: Ras small GTPase, Rho type | Molecular Function: GTP binding (GO:0005525), Biological Process: small GTPase mediated signal transduction (GO:0007264) |
| 7 (311) | 6 (458) | 13 (374) | IPR003603: Leucine-rich-associated | |

| | | | | |
|---|---|---|---|---|
| 7 (312) | 8 (339) | 10 (482) | IPR004156: Organic anion transporter polypeptide OATP | Molecular Function: transporter activity (GO:0005215), Biological Process: transport (GO:0006810), Cellular Component: membrane (GO:0016020) |
| 7 (313) | 7 (388) | 18 (270) | IPR004160: Elongation factor Tu, C-terminal | Molecular Function: GTP binding (GO:0005525) |
| 7 (314) | 15 (191) | **1 (2329)** | IPR004262: Male sterility protein | |
| 7 (315) | 7 (390) | 11 (430) | IPR004365: nucleic acid binding, OB-fold, tRNA/helicase-type | Molecular Function: nucleic acid binding (GO:0003676) |
| 7 (316) | 8 (340) | 9 (539) | IPR004837: Sodium/calcium exchanger membrane region | Cellular Component: integral to membrane (GO:0016021) |
| 7 (317) | 8 (341) | **3 (1318)** | IPR005018: DOMON | Molecular Function: dopamine beta-monooxygenase activity (GO:0004500), Biological Process: catecholamine metabolism (GO:0006584) |
| 7 (318) | 21 (119) | 17 (291) | IPR005123: 2OG-Fe(II) oxygenase | |
| 7 (319) | 6 (464) | 8 (600) | IPR005475: Transketolase, central region | |
| 7 (320) | 7 (392) | 19 (258) | IPR005817: Wnt superfamily | Molecular Function: signal transducer activity (GO:0004871), Cellular Component: extracellular region (GO:0005576), Biological Process: frizzled-2 signaling pathway (GO:0007223), Biological Process: development (GO:0007275) |
| 7 (321) | 14 (207) | 7 (679) | IPR006047: Alpha amylase, catalytic region | Molecular Function: alpha-amylase activity (GO:0004556), Biological Process: carbohydrate metabolism |

| | | | | |
|---|---|---|---|---|
| | | | | (GO:0005975) |
| 7 (322) | 17 (151) | 14 (355) | IPR006186: Serine/threonine-specific protein phosphatase and bis(5-nucleosyl)-tetraphosphatase | Molecular Function: hydrolase activity (GO:0016787) |
| 7 (323) | 11 (257) | 10 (487) | IPR006569: Regulation of nuclear pre-mRNA protein | |
| 7 (324) | 8 (345) | 8 (603) | IPR006595: CTLH, C-terminal to LisH motif | |
| 7 (325) | 6 (473) | 11 (439) | IPR006612: Zinc finger, C2CH-type | Molecular Function: nucleic acid binding (GO:0003676) |
| 7 (326) | 9 (309) | **1 (2670)** | IPR007889: Helix-turn-helix, Psq | Molecular Function: DNA binding (GO:0003677), Cellular Component: nucleus (GO:0005634) |
| 7 (327) | 7 (398) | 20 (246) | IPR008144: Guanylate kinase | |
| 7 (328) | 5 (566) | 15 (328) | IPR008197: Whey acidic protein, core region | |
| 7 (329) | 17 (154) | 9 (550) | IPR011021: Arrestin, N-terminal | |
| 6 (330) | 5 (486) | 10 (445) | IPR000209: Peptidase S8 and S53, subtilisin, kexin, sedolisin | Molecular Function: subtilase activity (GO:0004289), Biological Process: proteolysis and peptidolysis (GO:0006508) |
| 6 (331) | 10 (264) | 14 (331) | IPR000348: emp24/gp25L/p24 | Biological Process: intracellular protein transport (GO:0006886), Molecular Function: protein carrier activity (GO:0008320), Cellular Component: membrane (GO:0016020) |

| | | | | |
|---|---|---|---|---|
| 6 (332) | 7 (354) | 15 (306) | IPR000425: Major intrinsic protein | Molecular Function: transporter activity (GO:0005215), Biological Process: transport (GO:0006810), Cellular Component: membrane (GO:0016020) |
| 6 (333) | 7 (356) | 7 (626) | IPR000577: Carbohydrate kinase, FGGY | Biological Process: carbohydrate metabolism (GO:0005975) |
| 6 (334) | 5 (495) | 19 (250) | IPR000651: Guanine nucleotide exchange factor for Ras-like GTPases, N-terminal | Molecular Function: guanyl-nucleotide exchange factor activity (GO:0005085), Biological Process: small GTPase mediated signal transduction (GO:0007264) |
| 6 (335) | 11 (246) | 15 (309) | IPR000668: Peptidase C1A, papain | Biological Process: proteolysis and peptidolysis (GO:0006508), Molecular Function: cysteine-type peptidase activity (GO:0008234) |
| 6 (336) | 14 (202) | 24 (185) | IPR000772: Ricin B lectin | |
| 6 (337) | 4 (596) | 8 (561) | IPR000850: Adenylate kinase | Molecular Function: ATP binding (GO:0005524), Biological Process: nucleobase, nucleoside, nucleotide and nucleic acid metabolism (GO:0006139), Molecular Function: nucleotide kinase activity (GO:0019201) |
| 6 (338) | 6 (419) | 16 (296) | IPR000904: SEC7-like | |
| 6 (339) | 4 (598) | 15 (310) | IPR000909: Phosphatidylinositol-specific phospholipase C, X region | Molecular Function: phospholipase C activity (GO:0004629), Biological Process: signal transduction (GO:0007165), Biological Process: intracellular signaling cascade (GO:0007242) |
| 6 (340) | 9 (297) | 14 (337) | IPR001012: UBX | |

| | | | | |
|---|---|---|---|---|
| 6 (341) | 6 (420) | 6 (713) | IPR001017: Dehydrogenase, E1 component | Biological Process: metabolism (GO:0008152), Molecular Function: oxidoreductase activity, acting on the aldehyde or oxo group of donors, disulfide as acceptor (GO:0016624) |
| 6 (342) | 7 (361) | 16 (297) | IPR001019: Guanine nucleotide binding protein (G-protein), alpha subunit | Molecular Function: signal transducer activity (GO:0004871), Biological Process: G-protein coupled receptor protein signaling pathway (GO:0007186), Molecular Function: guanyl nucleotide binding (GO:0019001) |
| 6 (343) | 15 (182) | 11 (412) | IPR001023: Heat shock protein Hsp70 | Molecular Function: ATP binding (GO:0005524) |
| 6 (344) | 2 (983) | 6 (715) | IPR001087: Lipolytic enzyme, G-D-S-L | Biological Process: lipid metabolism (GO:0006629), Molecular Function: hydrolase activity, acting on ester bonds (GO:0016788) |
| 6 (345) | 8 (324) | 10 (459) | IPR001279: Beta-lactamase-like | |
| 6 (346) | 8 (325) | 8 (571) | IPR001327: FAD-dependent pyridine nucleotide-disulphide oxidoreductase | Biological Process: electron transport (GO:0006118), Molecular Function: disulfide oxidoreductase activity (GO:0015036) |
| 6 (347) | 7 (366) | 14 (344) | IPR001345: Phosphoglycerate/bisphospho glycerate mutase | Molecular Function: catalytic activity (GO:0003824), Biological Process: metabolism (GO:0008152) |
| 6 (348) | 7 (367) | 9 (515) | IPR001373: Cullin | Biological Process: cell cycle (GO:0007049) |
| 6 (349) | 5 (507) | 9 (516) | IPR001374: Single-stranded nucleic acid binding R3H | Molecular Function: nucleic acid binding (GO:0003676) |

| | | | | |
|---|---|---|---|---|
| 6 (350) | 6 (429) | 13 (368) | IPR001433: Oxidoreductase FAD/NAD(P)-binding | Biological Process: electron transport (GO:0006118), Molecular Function: oxidoreductase activity (GO:0016491) |
| 6 (351) | 5 (511) | 5 (829) | IPR001451: Bacterial transferase hexapeptide repeat | |
| 6 (352) | 4 (619) | 3 (1222) | IPR001540: Glycoside hydrolase, family 20 | Molecular Function: beta-N-acetylhexosaminidase activity (GO:0004563), Biological Process: carbohydrate metabolism (GO:0005975) |
| 6 (353) | 8 (328) | 17 (285) | IPR001699: Transcription factor, T-box | Molecular Function: transcription factor activity (GO:0003700), Cellular Component: nucleus (GO:0005634), Biological Process: regulation of transcription, DNA-dependent (GO:0006355) |
| 6 (354) | 3 (769) | 14 (346) | IPR001711: Phosphatidylinositol-specific phospholipase C, Y domain | Molecular Function: phosphoinositide phospholipase C activity (GO:0004435), Biological Process: lipid metabolism (GO:0006629), Biological Process: signal transduction (GO:0007165), Biological Process: intracellular signaling cascade (GO:0007242) |
| 6 (355) | 7 (371) | 7 (641) | IPR001737: Ribosomal RNA adenine methylase transferase | Biological Process: rRNA modification (GO:0000154), Molecular Function: rRNA (adenine-N6,N6-)-dimethyltransferase activity (GO:0000179), Molecular Function: rRNA methyltransferase activity (GO:0008649) |
| 6 (356) | 5 (515) | 23 (198) | IPR001762: Disintegrin | |
| 6 (357) | 5 (517) | 16 (298) | IPR001846: von Willebrand | |

| | | | | |
|---|---|---|---|---|
| | | | factor, type D | |
| 6 (358) | 5 (518) | 15 (315) | IPR001936: Ras GTPase-activating protein | |
| 6 (359) | 2 (1026) | 3 (1240) | IPR002015: Proteasome/cyclosome, regulatory subunit | Biological Process: regulation of progression through cell cycle (GO:0000074) |
| 6 (360) | 7 (375) | 9 (526) | IPR002524: Cation efflux protein | Biological Process: cation transport (GO:0006812), Molecular Function: cation transporter activity (GO:0008324), Cellular Component: membrane (GO:0016020) |
| 6 (361) | 9 (302) | 13 (372) | IPR002659: Glycosyl transferase, family 31 | Biological Process: protein amino acid glycosylation (GO:0006486), Molecular Function: galactosyltransferase activity (GO:0008378), Cellular Component: membrane (GO:0016020) |
| 6 (362) | 6 (448) | 5 (849) | IPR002877: Ribosomal RNA methyltransferase RrmJ/FtsJ | |
| 6 (363) | 18 (140) | 7 (654) | IPR002933: Peptidase M20 | Biological Process: proteolysis and peptidolysis (GO:0006508), Molecular Function: metallopeptidase activity (GO:0008237) |
| 6 (364) | 8 (333) | 7 (655) | IPR002937: Amine oxidase | Biological Process: electron transport (GO:0006118), Molecular Function: oxidoreductase activity (GO:0016491) |

| | | | | |
|---|---|---|---|---|
| 6 (365) | 5 (530) | 8 (585) | IPR003000: Silent information regulator protein Sir2 | Molecular Function: DNA binding (GO:0003677), Cellular Component: chromatin silencing complex (GO:0005677), Biological Process: chromatin silencing (GO:0006342), Biological Process: regulation of transcription, DNA-dependent (GO:0006355) |
| 6 (366) | 8 (334) | 8 (586) | IPR003010: Nitrilase/cyanide hydratase and apolipoprotein N-acyltransferase | Biological Process: nitrogen compound metabolism (GO:0006807), Molecular Function: hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds (GO:0016810) |
| 6 (367) | 6 (450) | 13 (373) | IPR003104: Actin-binding FH2 | Molecular Function: actin binding (GO:0003779), Biological Process: cell organization and biogenesis (GO:0016043) |
| 6 (368) | 6 (453) | 7 (658) | IPR003126: Zinc finger, N-recognin | Molecular Function: ubiquitin-protein ligase activity (GO:0004842), Biological Process: ubiquitin cycle (GO:0006512) |
| 6 (369) | 6 (455) | 12 (397) | IPR003323: Ovarian tumour, otubain | |
| 6 (370) | 4 (651) | 7 (661) | IPR003405: Structural maintenance of chromosome protein SMC, C-terminal | Molecular Function: ATP binding (GO:0005524), Cellular Component: chromosome (GO:0005694), Biological Process: chromosome organization and biogenesis (GO:0051276) |
| 6 (371) | 6 (459) | 7 (665) | IPR003609: Apple-like | |

| | | | | |
|---|---|---|---|---|
| 6 (372) | 5 (537) | 7 (666) | IPR003613: U box | Cellular Component: ubiquitin ligase complex (GO:0000151), Molecular Function: ubiquitin-protein ligase activity (GO:0004842), Biological Process: protein ubiquitination (GO:0016567) |
| 6 (373) | 12 (238) | 11 (428) | IPR003650: Orange | Molecular Function: DNA binding (GO:0003677), Biological Process: regulation of transcription, DNA-dependent (GO:0006355) |
| 6 (374) | 10 (278) | 10 (479) | IPR003890: Initiation factor eIF-4 gamma, middle | Molecular Function: RNA binding (GO:0003723) |
| 6 (375) | 4 (656) | 14 (350) | IPR003954: RNA recognition, region 1 | Molecular Function: nucleic acid binding (GO:0003676) |
| 6 (376) | 5 (541) | 16 (300) | IPR004014: Cation transporting ATPase, N-terminal | Biological Process: cation transport (GO:0006812), Molecular Function: ATPase activity, coupled to transmembrane movement of ions, phosphorylative mechanism (GO:0015662), Cellular Component: membrane (GO:0016020) |
| 6 (377) | 6 (461) | 22 (216) | IPR004038: Ribosomal protein L7Ae/L30e/S12e/Gadd45 | |
| 6 (378) | **32 (64)** | 21 (229) | IPR004046: Glutathione S-transferase, C-terminal | |
| 6 (379) | 7 (387) | 10 (481) | IPR004154: Anticodon-binding | Molecular Function: tRNA ligase activity (GO:0004812), Molecular Function: ATP binding (GO:0005524), Biological Process: protein biosynthesis (GO:0006412) |

| | | | | |
|---|---|---|---|---|
| 6 (380) | 6 (462) | 14 (351) | IPR004367: Cyclin, C-terminal | Biological Process: regulation of progression through cell cycle (GO:0000074), Cellular Component: nucleus (GO:0005634) |
| 6 (381) | 8 (342) | 9 (540) | IPR005024: Snf7 | Molecular Function: molecular function unknown (GO:0005554) |
| 6 (382) | 5 (553) | 5 (899) | IPR005814: Aminotransferase class-III | Molecular Function: transaminase activity (GO:0008483), Molecular Function: pyridoxal phosphate binding (GO:0030170) |
| 6 (383) | 6 (467) | 14 (354) | IPR006068: Cation transporting ATPase, C-terminal | Biological Process: cation transport (GO:0006812), Molecular Function: ATPase activity, coupled to transmembrane movement of ions, phosphorylative mechanism (GO:0015662), Cellular Component: membrane (GO:0016020) |
| 6 (384) | 6 (468) | 9 (542) | IPR006092: Acyl-CoA dehydrogenase, N-terminal | Molecular Function: acyl-CoA dehydrogenase activity (GO:0003995), Biological Process: electron transport (GO:0006118) |
| 6 (385) | 3 (857) | 5 (905) | IPR006204: GHMP kinase | Molecular Function: ATP binding (GO:0005524), Molecular Function: kinase activity (GO:0016301), Biological Process: phosphorylation (GO:0016310) |

| | | | | |
|---|---|---|---|---|
| 6 (386) | 24 (99) | 15 (326) | IPR006620: Prolyl 4-hydroxylase, alpha subunit | Molecular Function: oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, 2-oxoglutarate as one donor, and incorporation of one atom each of oxygen into both donors (GO:0016706), Biological Process: protein metabolism (GO:0019538) |
| 6 (387) | 15 (194) | () | IPR006621: Nose resistant to fluoxetine-4, N-terminal | |
| 6 (388) | 16 (174) | 16 (301) | IPR006629: LPS-induced tumor necrosis factor alpha factor | |
| 6 (389) | 7 (396) | 13 (383) | IPR007123: Gelsolin region | |
| 6 (390) | 10 (286) | 3 (1399) | IPR007248: Mpv17/PMP22 | Cellular Component: integral to membrane (GO:0016021) |
| 6 (391) | 4 (701) | 3 (1430) | IPR008257: Peptidase M19, renal dipeptidase | Molecular Function: membrane dipeptidase activity (GO:0004237), Biological Process: proteolysis and peptidolysis (GO:0006508), Molecular Function: dipeptidyl-peptidase activity (GO:0008239) |
| 6 (392) | 4 (706) | 22 (220) | IPR010294: ADAM-TS Spacer 1 | Molecular Function: metalloendopeptidase activity (GO:0004222), Cellular Component: extracellular matrix (GO:0031012) |
| 6 (393) | 5 (572) | 13 (384) | IPR010569: Myotubularin-related | Molecular Function: inositol or phosphatidylinositol phosphatase activity (GO:0004437), Biological Process: phospholipid dephosphorylation (GO:0046839) |
| 6 (394) | 4 (709) | () | IPR010629: Insect allergen related | |

| | | | | |
|---|---|---|---|---|
| 5 (395) | 8 (314) | 23 (189) | IPR000024: Frizzled CRD region | |
| 5 (396) | 4 (576) | 7 (618) | IPR000034: Laminin B | |
| 5 (397) | 6 (407) | 14 (329) | IPR000095: PAK-box/P21-Rho-binding | |
| 5 (398) | 3 (731) | 13 (363) | IPR000156: RanBP1 | |
| 5 (399) | 4 (578) | 22 (206) | IPR000164: Histone H3 | Cellular Component: nucleosome (GO:0000786), Molecular Function: DNA binding (GO:0003677), Cellular Component: nucleus (GO:0005634), Biological Process: nucleosome assembly (GO:0006334), Biological Process: chromosome organization and biogenesis (sensu Eukaryota) (GO:0007001) |
| 5 (400) | 6 (409) | 9 (500) | IPR000300: Inositol polyphosphate related phosphatase | Molecular Function: inositol or phosphatidylinositol phosphatase activity (GO:0004437) |
| 5 (401) | 10 (262) | 15 (305) | IPR000326: Phosphoesterase, PA-phosphatase related | |
| 5 (402) | 4 (583) | 9 (501) | IPR000331: Rap/ran-GAP | |
| 5 (403) | 5 (490) | 6 (705) | IPR000409: Beige/BEACH | |
| 5 (404) | 6 (414) | 5 (813) | IPR000435: Tektin | Biological Process: microtubule cytoskeleton organization and biogenesis (GO:0000226), Cellular Component: microtubule (GO:0005874) |
| 5 (405) | 5 (492) | 5 (814) | IPR000462: CDP-alcohol phosphatidyltransferase | Biological Process: phospholipid biosynthesis (GO:0008654) |
| 5 (406) | 2 (948) | 8 (558) | IPR000533: Tropomyosin | |
| 5 (407) | 2 (949) | 33 (137) | IPR000566: Lipocalin-related protein and Bos/Can/Equ allergen | Molecular Function: binding (GO:0005488) |

| | | | | |
|---|---|---|---|---|
| 5 (408) | 6 (415) | 20 (235) | IPR000591: Pleckstrin/ G-protein, interacting region | Biological Process: intracellular signaling cascade (GO:0007242) |
| 5 (409) | 5 (498) | 11 (411) | IPR000697: EVH1 | |
| 5 (410) | 5 (499) | 9 (506) | IPR000756: Diacylglycerol kinase accessory region | Molecular Function: diacylglycerol kinase activity (GO:0004143), Biological Process: protein kinase C activation (GO:0007205) |
| 5 (411) | 6 (418) | 10 (450) | IPR000836: Phosphoribosyltransferase | Biological Process: nucleoside metabolism (GO:0009116) |
| 5 (412) | 4 (597) | 9 (507) | IPR000857: Unconventional myosin/plant kinesin-like protein/non-motor protein conserved region MyTH4 | Cellular Component: cytoskeleton (GO:0005856) |
| 5 (413) | 7 (358) | () | IPR000896: Arthropod hemocyanin/insect LSP | Molecular Function: oxygen transporter activity (GO:0005344), Biological Process: transport (GO:0006810) |
| 5 (414) | 4 (602) | 2 (1557) | IPR001117: Multicopper oxidase, type 1 | Molecular Function: copper ion binding (GO:0005507) |
| 5 (415) | 6 (424) | 8 (565) | IPR001132: Dwarfin protein | Cellular Component: intracellular (GO:0005622), Biological Process: regulation of transcription, DNA-dependent (GO:0006355) |
| 5 (416) | 4 (603) | 14 (340) | IPR001180: Citron-like | Molecular Function: small GTPase regulator activity (GO:0005083) |
| 5 (417) | 4 (611) | 9 (514) | IPR001298: Filamin/ABP280 repeat | |
| 5 (418) | 4 (617) | 10 (460) | IPR001370: Proteinase inhibitor I32, inhibitor of apoptosis | Cellular Component: intracellular (GO:0005622), Biological Process: anti-apoptosis (GO:0006916) |
| 5 (419) | 5 (509) | 9 (518) | IPR001388: Synaptobrevin | Cellular Component: integral to membrane (GO:0016021), Biological Process: vesicle-mediated transport |

| | | | | |
|---|---|---|---|---|
| | | | | (GO:0016192) |
| 5 (420) | 4 (618) | 10 (461) | IPR001392: Clathrin adaptor complex, medium chain | Biological Process: intracellular protein transport (GO:0006886), Cellular Component: clathrin vesicle coat (GO:0030125) |
| 5 (421) | 6 (428) | 13 (367) | IPR001401: Dynamin | Molecular Function: GTPase activity (GO:0003924), Molecular Function: GTP binding (GO:0005525) |
| 5 (422) | 1 (1517) | 4 (1021) | IPR001525: C-5 cytosine-specific DNA methylase | Molecular Function: DNA binding (GO:0003677), Biological Process: DNA methylation (GO:0006306) |
| 5 (423) | 5 (513) | 7 (639) | IPR001619: Sec1-like protein | Biological Process: vesicle docking during exocytosis (GO:0006904), Biological Process: vesicle-mediated transport (GO:0016192) |
| 5 (424) | 6 (434) | 30 (155) | IPR001627: Semaphorin/CD100 antigen | |
| 5 (425) | 4 (627) | 2 (1590) | IPR001747: Lipid transport protein, N-terminal | Molecular Function: lipid transporter activity (GO:0005319), Biological Process: lipid transport (GO:0006869) |
| 5 (426) | 5 (519) | 12 (393) | IPR002083: MATH | |
| 5 (427) | 2 (1033) | 25 (181) | IPR002119: Histone H2A | Cellular Component: nucleosome (GO:0000786), Molecular Function: DNA binding (GO:0003677), Cellular Component: nucleus (GO:0005634), Biological Process: nucleosome assembly (GO:0006334), Biological Process: chromosome organization and biogenesis (sensu Eukaryota) (GO:0007001) |

| | | | | |
|---|---|---|---|---|
| 5 (428) | 6 (443) | 5 (838) | IPR002125: CMP/dCMP deaminase, zinc-binding | Molecular Function: zinc ion binding (GO:0008270), Molecular Function: hydrolase activity (GO:0016787) |
| 5 (429) | 6 (444) | 7 (648) | IPR002155: Thiolase | |
| 5 (430) | **25 (92)** | 17 (288) | IPR002223: Proteinase inhibitor I2, Kunitz metazoa | Molecular Function: serine-type endopeptidase inhibitor activity (GO:0004867) |
| 5 (431) | 5 (522) | 6 (739) | IPR002469: Peptidase S9B, dipeptidylpeptidase IV N-terminal | Molecular Function: dipeptidyl-peptidase IV activity (GO:0004274), Biological Process: proteolysis and peptidolysis (GO:0006508), Cellular Component: membrane (GO:0016020) |
| 5 (432) | 11 (251) | 4 (1048) | IPR002502: N-acetylmuramoyl-L-alanine amidase, family 2 | Molecular Function: N-acetylmuramoyl-L-alanine amidase activity (GO:0008745), Biological Process: peptidoglycan catabolism (GO:0009253) |
| 5 (433) | 5 (525) | 5 (844) | IPR002562: 3'-5' exonuclease | Molecular Function: nucleic acid binding (GO:0003676), Cellular Component: intracellular (GO:0005622), Molecular Function: 3'-5' exonuclease activity (GO:0008408) |
| 5 (434) | 7 (377) | 9 (527) | IPR002610: Rhomboid-like protein | |
| 5 (435) | 3 (792) | 9 (529) | IPR002857: Zinc finger, CXXC-type | Molecular Function: DNA binding (GO:0003677), Molecular Function: zinc ion binding (GO:0008270) |
| 5 (436) | 9 (304) | 9 (530) | IPR002939: Chaperone DnaJ, C-terminal | Biological Process: protein folding (GO:0006457), Molecular Function: unfolded protein binding (GO:0051082) |

| | | | | |
|---|---|---|---|---|
| 5 (437) | 6 (451) | 15 (320) | IPR003118: Sterile alpha motif/pointed | Molecular Function: DNA binding (GO:0003677), Cellular Component: nucleus (GO:0005634) |
| 5 (438) | 6 (452) | 19 (257) | IPR003124: Actin-binding WH2 | |
| 5 (439) | 5 (532) | 5 (857) | IPR003152: PIK-related kinase, FATC | |
| 5 (440) | 5 (533) | 8 (592) | IPR003307: eIF4-gamma/eIF5/eIF2-epsilon | Molecular Function: translation initiation factor activity (GO:0003743), Biological Process: regulation of translational initiation (GO:0006446) |
| 5 (441) | 7 (382) | 5 (863) | IPR003397: Mitochondrial import inner membrane translocase, subunit Tim17/22 | Cellular Component: mitochondrial inner membrane (GO:0005743), Molecular Function: protein transporter activity (GO:0008565), Biological Process: protein transport (GO:0015031) |
| 5 (442) | 5 (538) | 12 (398) | IPR003619: Dwarfin protein, A | Cellular Component: intracellular (GO:0005622), Biological Process: regulation of transcription, DNA-dependent (GO:0006355) |
| 5 (443) | 4 (655) | 8 (596) | IPR003892: Ubiquitin system component Cue | |
| 5 (444) | 5 (545) | 17 (290) | IPR004182: GRAM | |
| 5 (445) | 20 (126) | () | IPR004210: BESS motif | Molecular Function: DNA binding (GO:0003677) |
| 5 (446) | 7 (389) | 12 (400) | IPR004299: Membrane bound O-acyl transferase, MBOAT | |

| | | | | |
|---|---|---|---|---|
| 5 (447) | 5 (548) | 5 (885) | IPR004364: tRNA synthetase, class II (D, K and N) | Molecular Function: tRNA ligase activity (GO:0004812), Molecular Function: ATP binding (GO:0005524), Cellular Component: cytoplasm (GO:0005737), Biological Process: tRNA aminoacylation for protein translation (GO:0006418) |
| 5 (448) | 4 (669) | 13 (379) | IPR005112: dDENN | |
| 5 (449) | 4 (670) | 11 (432) | IPR005113: uDENN | |
| 5 (450) | () | **1 (2384)** | IPR005119: LysR, substrate-binding | |
| 5 (451) | 10 (281) | **()** | IPR005203: Hemocyanin, C-terminal | |
| 5 (452) | 8 (343) | **()** | IPR005204: Hemocyanin, N-terminal | |
| 5 (453) | 2 (1181) | 5 (895) | IPR005476: Transketolase, C-terminal | |
| 5 (454) | 3 (847) | 3 (1345) | IPR005835: Nucleotidyl transferase | Biological Process: biosynthesis (GO:0009058), Molecular Function: nucleotidyltransferase activity (GO:0016779) |
| 5 (455) | 6 (466) | 11 (433) | IPR006011: Syntaxin, N-terminal | Cellular Component: membrane (GO:0016020) |
| 5 (456) | 15 (193) | 31 (150) | IPR006026: Peptidase, metallopeptidases | Biological Process: proteolysis and peptidolysis (GO:0006508), Molecular Function: metallopeptidase activity (GO:0008237) |
| 5 (457) | 6 (469) | 4 (1104) | IPR006139: D-isomer specific 2-hydroxyacid dehydrogenase, catalytic region | Biological Process: L-serine biosynthesis (GO:0006564), Molecular Function: oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor (GO:0016616) |

| | | | | |
|---|---|---|---|---|
| 5 (458) | 6 (470) | 4 (1105) | IPR006140: D-isomer specific 2-hydroxyacid dehydrogenase, NAD-binding | Biological Process: L-serine biosynthesis (GO:0006564), Molecular Function: oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor (GO:0016616) |
| 5 (459) | 5 (554) | 13 (381) | IPR006153: Sodium/hydrogen exchanger | Biological Process: regulation of pH (GO:0006885), Molecular Function: solute:hydrogen antiporter activity (GO:0015299), Cellular Component: integral to membrane (GO:0016021) |
| 5 (460) | 5 (557) | 5 (912) | IPR006571: TLDc | |
| 5 (461) | 13 (220) | 5 (917) | IPR006589: Alpha amylase, catalytic subdomain | Molecular Function: alpha-amylase activity (GO:0004556), Biological Process: carbohydrate metabolism (GO:0005975) |
| 5 (462) | 7 (394) | 7 (682) | IPR006596: Nucleotide binding protein, PINc | |
| 5 (463) | 5 (559) | 6 (780) | IPR006630: RNA-binding protein Lupus La | |
| 5 (464) | 4 (691) | 10 (488) | IPR006680: Amidohydrolase | Molecular Function: hydrolase activity (GO:0016787) |
| 5 (465) | 5 (562) | 6 (786) | IPR007109: Brix | |
| 5 (466) | 6 (474) | 5 (922) | IPR007259: Spc97/Spc98 | Biological Process: microtubule cytoskeleton organization and biogenesis (GO:0000226), Cellular Component: spindle pole (GO:0000922), Cellular Component: microtubule organizing center (GO:0005815) |
| 5 (467) | 15 (195) | 9 (546) | IPR007484: Peptidase M28 | Biological Process: proteolysis and peptidolysis (GO:0006508), Molecular Function: peptidase activity (GO:0008233) |

| | | | | |
|---|---|---|---|---|
| 5 (468) | 6 (476) | 14 (359) | IPR008211: Laminin, N-terminal | Cellular Component: extracellular matrix (sensu Metazoa) (GO:0005578) |
| 5 (469) | 3 (902) | 10 (490) | IPR008273: Cellular retinaldehyde-binding/triple function, N-terminal | |
| 5 (470) | 5 (569) | **1 (2770)** | IPR008774: A2 Phospholipase | Molecular Function: phospholipase A2 activity (GO:0004623), Molecular Function: calcium ion binding (GO:0005509), Cellular Component: extracellular region (GO:0005576), Biological Process: phospholipid metabolism (GO:0006644) |
| 5 (471) | 2 (1331) | 9 (549) | IPR009886: HCaRG | |
| 5 (472) | 6 (479) | 12 (403) | IPR010625: CHCH | |
| 5 (473) | 10 (290) | 3 (1479) | IPR011611: PfkB | |
| 5 (474) | 5 (575) | 4 (1164) | IPR011706: Multicopper oxidase, type 2 | Molecular Function: copper ion binding (GO:0005507), Molecular Function: oxidoreductase activity (GO:0016491) |
| 5 (475) | 4 (718) | **()** | IPR011707: Multicopper oxidase, type 3 | Molecular Function: copper ion binding (GO:0005507), Molecular Function: oxidoreductase activity (GO:0016491) |
| 5 (476) | 6 (482) | 6 (805) | IPR011765: Peptidase M16, N-terminal | |
| 4 (477) | 2 (930) | 18 (260) | IPR000001: Kringle | |
| 4 (478) | 5 (484) | 15 (304) | IPR000038: Cell division/GTP binding protein | Molecular Function: GTP binding (GO:0005525), Biological Process: cell cycle (GO:0007049) |

| | | | | |
|---|---|---|---|---|
| 4 (479) | 4 (577) | 3 (1172) | IPR000040: Acute myeloid leukemia 1 protein (AML 1)/Runt | Molecular Function: DNA binding (GO:0003677), Molecular Function: ATP binding (GO:0005524), Cellular Component: nucleus (GO:0005634), Biological Process: regulation of transcription, DNA-dependent (GO:0006355) |
| 4 (480) | 6 (408) | 4 (974) | IPR000120: Amidase | Molecular Function: amidase activity (GO:0004040) |
| 4 (481) | 9 (291) | 20 (231) | IPR000157: TIR | Molecular Function: transmembrane receptor activity (GO:0004888), Cellular Component: membrane (GO:0016020) |
| 4 (482) | 7 (353) | 6 (701) | IPR000194: H+-transporting two-sector ATPase, alpha/beta subunit, central region | Molecular Function: ATP binding (GO:0005524), Biological Process: ATP synthesis coupled proton transport (GO:0015986), Cellular Component: proton-transporting two-sector ATPase complex (GO:0016469), Molecular Function: hydrogen-transporting ATP synthase activity, rotational mechanism (GO:0046933), Molecular Function: hydrogen-transporting ATPase activity, rotational mechanism (GO:0046961) |
| 4 (483) | 5 (487) | 3 (1177) | IPR000246: Peptidase T2, asparaginase 2 | Molecular Function: asparaginase activity (GO:0004067), Biological Process: glycoprotein catabolism (GO:0006516) |
| 4 (484) | 4 (581) | 9 (499) | IPR000261: EPS15 homology (EH) | |
| 4 (485) | 5 (488) | 11 (408) | IPR000286: Histone deacetylase superfamily | |

| | | | | |
|---|---|---|---|---|
| 4 (486) | 4 (582) | 3 (1182) | IPR000323: Copper type II, ascorbate-dependent monooxygenase | Molecular Function: monooxygenase activity (GO:0004497), Molecular Function: copper ion binding (GO:0005507) |
| 4 (487) | 5 (489) | 16 (294) | IPR000327: POU | Molecular Function: transcription factor activity (GO:0003700), Cellular Component: nucleus (GO:0005634), Biological Process: regulation of transcription, DNA-dependent (GO:0006355) |
| 4 (488) | 3 (739) | 4 (978) | IPR000352: Class I peptide chain release factor | Molecular Function: translation release factor activity (GO:0003747), Biological Process: translational termination (GO:0006415) |
| 4 (489) | 6 (412) | 6 (704) | IPR000402: Na+/K+ ATPase, beta subunit | Molecular Function: sodium:potassium-exchanging ATPase activity (GO:0005391), Biological Process: potassium ion transport (GO:0006813), Biological Process: sodium ion transport (GO:0006814), Cellular Component: membrane (GO:0016020) |
| 4 (490) | 2 (947) | 5 (812) | IPR000432: DNA mismatch repair protein MutS, C-terminal | Molecular Function: damaged DNA binding (GO:0003684), Molecular Function: ATP binding (GO:0005524), Biological Process: mismatch repair (GO:0006298) |

| | | | | |
|---|---|---|---|---|
| 4 (491) | 1 (1423) | 22 (207) | IPR000558: Histone H2B | Cellular Component: nucleosome (GO:0000786), Molecular Function: DNA binding (GO:0003677), Cellular Component: nucleus (GO:0005634), Biological Process: nucleosome assembly (GO:0006334), Biological Process: chromosome organization and biogenesis (sensu Eukaryota) (GO:0007001) |
| 4 (492) | 4 (588) | 4 (981) | IPR000583: Glutamine amidotransferase, class-II | Biological Process: metabolism (GO:0008152) |
| 4 (493) | 4 (590) | 12 (386) | IPR000648: Oxysterol-binding protein | Biological Process: steroid metabolism (GO:0008202) |
| 4 (494) | 4 (591) | 4 (989) | IPR000649: Initiation factor 2B related | Biological Process: cellular biosynthesis (GO:0044249) |
| 4 (495) | 7 (357) | 6 (709) | IPR000793: H+-transporting two-sector ATPase, alpha/beta subunit, C-terminal | Biological Process: ATP biosynthesis (GO:0006754), Biological Process: ATP synthesis coupled proton transport (GO:0015986), Cellular Component: proton-transporting two-sector ATPase complex (GO:0016469), Molecular Function: hydrogen-transporting ATP synthase activity, rotational mechanism (GO:0046933), Molecular Function: hydrogen-transporting ATPase activity, rotational mechanism (GO:0046961) |

| | | | | |
|---|---|---|---|---|
| 4 (496) | 5 (500) | 3 (1195) | IPR000814: TATA-box binding | Molecular Function: DNA binding (GO:0003677), Molecular Function: RNA polymerase II transcription factor activity (GO:0003702), Cellular Component: nucleus (GO:0005634), Cellular Component: transcription factor TFIID complex (GO:0005669), Biological Process: regulation of transcription, DNA-dependent (GO:0006355), Biological Process: transcription initiation from RNA polymerase II promoter (GO:0006367) |
| 4 (497) | 2 (969) | 1 (2077) | IPR000897: GTP-binding signal recognition particle SRP54, G-domain | Molecular Function: RNA binding (GO:0003723), Molecular Function: GTP binding (GO:0005525), Cellular Component: signal recognition particle (sensu Eukaryota) (GO:0005786), Biological Process: SRP-dependent cotranslational protein-membrane targeting (GO:0006614) |
| 4 (498) | 2 (971) | 9 (508) | IPR000906: ZU5 | |
| 4 (499) | 5 (502) | 13 (365) | IPR000949: ELM2 | |
| 4 (500) | 7 (360) | () | IPR000990: Innexin | Cellular Component: gap junction (GO:0005921) |

## Table S14. Homeobox Genes.

| ID Name | Chromosome | Exon Number | Domain Name | Top Blast hit (not to be used as guide to orthology) |
|---|---|---|---|---|
| GB13163-RA | Group1.38 | 2 | HOX | Homeobox protein GBX-2 |
| GB13163-RB | Group1.38 | 2 | HOX | Gastrulation brain homeobox 1: GBX-1 |
| GB13163-RC | Group1.38 | 2 | HOX | Homeobox protein GBX-2 |
| GB13163-RD | Group1.38 | 3 | HOX | Gastrulation brain homeobox 1: GBX-1 |
| GB11694-RA | Group1.43 | 3 | Dicty HOX | Homeobox protein invected inv |
| GB15566-RA | Group1.43 | 4 | HOX | Segmentation polarity homeobox protein engrailed; 72% identity to homeobox protein En-1 |
| GB15566-RB | Group1.43 | 4 | HOX | Segmentation polarity homeobox protein engrailed; 72% identity to homeobox protein En-1; isoform2 |
| GB10613-RA | Group1.37 | 2 | HOX | Transcription factor LBX1 |
| GB13498-RA | Group1.37 | 3 | HOX | Homeobox protein Nkx-3.2 |
| GB15586-RA | Group1.37 | 4 | HOX | Homeobox protein Nkx-2.5 isoform1 |
| GB15586-RB | Group1.37 | 3 | HOX | Homeobox protein Nkx-2.5 isoform2 |
| GB13830-RA | Group1.37 | 3 | HOX | Muscle segmentation |

| | | | | |
|---|---|---|---|---|
| | | | | homeobox (Protein Drop) |
| GB13830-RB | Group1.37 | 2 | HOX | Muscle segmentation homeobox (Protein Drop) |
| GB18552-RA | Group1.37 | 3 | SelP HOX | Muscle segmentation homeobox (Protein Drop) |
| GB18552-RB | Group1.37 | 2 | SelP HOX | Muscle segmentation homeobox (Protein Drop) |
| GB11341-RA | Group1.48 | 3 | HOX PRP8 | Prd-like homeobox protein or aristaless protein |
| GB11566-RA | Group1.49 | 4 | HOX | Orthodenticle-2 protein |
| GB30099-RA | Group1.52 | 6 | HOX Spor | Orthopedia |
| GB10752-RA | Group1.34 | 4 | HOX | Homeobox protein six4.3 |
| GB12355-RA | Group1.3 | 8 | PAX HOX Prox1 Andr | Paired box isoform 1 gene 6 |
| GB12355-RB | Group1.3 | 8 | PAX HOX Prox1 Andr | Paired-box isoform 2 gene 6 |
| GB10821-RA | Group1.19 | 3 | Dicty HOX | Homeotic caudal protein; Parahox-related |
| GB12408-RA | Group2.8 | 7 | 2LIM HOX Herpes | Homeobox protein Lim-1 |
| GB14802-RA | Group2.7 | 2 | HOX | Gsx |
| GB16761-RA | Group2.10 | 1 | HOX | Homeobox protein Six3 (one exon gene) |
| GB19877-RA | Group3.20 | 5 | 2LIM 2HOX | LIM/homeobox protein LMX1B |

| | | | | |
|---|---|---|---|---|
| GB15837-RA | Group3.27 | 8 | PBX HOX | Homeobox protein extradenticle Exd |
| GB10569-RA | Group3.23 | 3 | HOX | "Dual bar protein , Homeobox protein B-H2" |
| GB16262-RA | Group4.18 | 7 | POU HOX | "POU domain, class 2, transcription factor 1 isoform A " |
| GB15295-RA | Group4.11 | 4 | HOX | Ptx1 homeodomain protein |
| GB18397-RA | Group4.19 | 11 | PAX GRP | "Paired box protein, sparkling protein" |
| GB18397-RB | Group4.19 | 2 | - | "Paired box protein, sparkling protein, the shortest isoform" |
| GB14533-RA | Group5.6 | 5 | 2Prox1 HemX | Tf hox Prospero |
| GB15894-RA | Group5.21 | 4 | HOX | Medaka OG-12; homeodomain protein;Short stature homeobox protein 2 |
| GB15213-RA | Group5.14 | 5 | HOX | Six2 protein; contains knotted-1-like domain from TALE family |
| GB18348-RA | Group5.15 | 13 | HOX | Homothorax homeoprotein hth |
| GB17945-RA | Group6.38 | 9 | 3CUT HOX BBC Herpes | Homeobox protein cut |
| GB11714-RA | Group6.23 | 9 | PAX HOX | Paired box transcription factor Pax6 |
| GB11714-RB | Group6.23 | 7 | PAX HOX | Paired box transcription factor Pax6 |
| GB30239-RA | Group8.5 | 12 | LIM-bind | LIM homeobox protein cofactor CLIM-1a |

| | | | | |
|---|---|---|---|---|
| GB30239-RB | Group8.5 | 9 | LIM-bind | LIM homeobox protein cofactor CLIM-1a |
| GB13745-RA | Group9.4 | 4 | HOX | Homeobox protein HLX1 |
| GB15027-RA | Group9.11 | 2 | HOX | Part of NK2 transcription factor related; NK2 transcription factor-like protein B; homeobox protein |
| GB16085-RA | Group10.10 | 1 | POU HOX | "Pou DOMAIN PROTEIN, drifter protein; CF1" |
| GB13229-RA | Group10.12 | 14 | PDZ 4LIM | Part of LIM domain binding 3 |
| GB10772-RA | Group10.15 | 3 | PAX 2Selp | Paired box protein Pax-1 |
| GB18585-RA | Group11.11 | 4 | 2LIM SelP HOX | LIM2 related protein |
| GB18585-RB | Group11.11 | 5 | 2LIM SelP HOX | LIM2 related protein |
| GB30111-RA | Group11.11 | 5 | LIM HOX | LIM homeobox 9 |
| GB13918-RB | Group11.9 | 3 | CUT HOX | Hepatocyte nuclear factor 6 (HNF-6) (One cut domain family member 1) |
| GB13918-RA | Group11.9 | 7 | CUT HOX | Hepatocyte nuclear factor 6 (HNF-6) (One cut domain family member 1) |
| GB16706-RA | Group11.17 | 6 | 4HOX Prox1 TolA | zinc finger homeodomain 4 isoform 3 |
| GB18111-RA | Group11.23 | 4 | HOX Atro | Caupolican homeoprotein |
| GB14516-RA | Group13.7 | 5 | HOX Selp | Homeotic distal-less protein (Protein |

| | | | | |
|---|---|---|---|---|
| | | | | brista); DLL; DLX-2 |
| GB30148-RA | Group13.14 | 7 | HOX Herpes | Transcription factor DRG11 homeodomain protein |
| GB15643-RA | Group13.10 | 14 | POU HOX | Retina-derived POU-domain factor-1 isoform 1 |
| GB15643-RB | Group13.10 | 13 | POU HOX | Retina-derived POU-domain factor-1 isoform 2 |
| GB15643-RC | Group13.10 | 12 | POU HOX | Retina-derived POU-domain factor-1 isoform 3 |
| GB15469-RA | Group14.4 | 5 | PAX HOX Atro | "Paired-box transcription factor, PRD-like homeobox" |
| GB15469-RB | Group14.4 | 5 | PAX HOX Atro | Segmentation protein paired ;PRD class homeobox protein; DMBX1 28-42% |
| GB15632-RA | Group14.12 | 6 | PAX HOX | segmentation polarity homeobox protein engrailed; 72% identity to homeobox protein En-1 |
| GB14483-RA | Group14.12 | 6 | PAX HOX | Paired box transcription factor BSH4 |
| GB16661-RA | Group14.3 | 3 | HOX | Aristaless 3; PRD-like homeobox protein |
| GB16259-RB | Group15.12 | 1 | HOX TFIIA | "Homeobox protein slou (S59/2), 45% similar to NK-1 homeobox protein; isoform 1" |

| | | | | |
|---|---|---|---|---|
| GB16259-RA | Group15.12 | 2 | HOX TFIIA | "Homeobox protein slou (S59/2), 45% similar to NK-1 homeobox protein; isoform 2" |
| GB11254-RA | Group15.15 | 3 | SelP HOX | Empty spiracles homeotic protein Emx |
| GB30063-RA | Group16.1 | 14 | Prox1 | Homeobox prospero-like protein (PROX1) protein |
| GB15651-RA | Group16.2 | 9 | LIM LIM HOX | LIM HOMEOBOX PROTEIN isoform 1 |
| GB15651-RB | Group16.2 | 8 | LIM LIM HOX | LIM HOMEOBOX PROTEIN isoform 2 |
| GB15651-RC | Group16.2 | 4 | LIM LIM HOX | LIM HOMEOBOX PROTEIN isoform 3 |
| GB18833-RA | Group16.2 | 3 | POU HOX | "POU DOMAIN PROTEIN, CLASS 4-RELATED " |
| GB13409-RB | Group16.6 | 3 | HOX Atro Andro | "HOMEOBOX PROTEIN: dfd, isoform 1" |
| GB13409-RA | Group16.6 | 3 | HOX | "HOMEOBOX PROTEIN: dfd, isoform 2" |
| GB18792-RA | Group16.6 | 3 | HOX | HOMEOBOX PROTEIN: zen-related |
| GB11988-RA | Group16.6 | 4 | HOX ROM1 | HOMEOTIC PROTEIN: pb |
| GB14027-RA | Group16.6 | 2 | HOX ZhuA | HOMEOBOX PROTEIN:lab |
| GB18940-RA | Group16.7 | 2 | HOX | Homeotic scr-related protein |
| GB13491-RA | Group16.7 | 2 | HOX Trbl | HOMEOBOX PROTEIN: ftz |

| | | | | |
|---|---|---|---|---|
| GB10341-RA | Group16.8 | 2 | HOX Herpes | HOMEOBOX-RELATED PROTEIN: AbdB |
| GB19738-RA | Group16.8 | 3 | HOX Herpes | HOMEOBOX PROTEIN: AbdA |
| GB30077-RA | Group16.8 | 2 | HOX | Ultrabithorax: Ubx |
| GB18813-RA | Group16.8 | 2 | HOX | Homeotic antennapedia protein Antp |
| GB18918-RA | GroupUn.1196 | 9 | HOX | Zinc-finger homeodomain protein 1 |
| GB10623-RA | GroupUn.41 | 3 | HOX | "Even skipped, HOX related eve" |
| GB12465-RA | GroupUn.531 | 6 | LIM LIM HOX | LIM homeobox protein |
| GB14318-RA | GroupUn.109 | 4 | LIM LIM HOX | Arrowhead ;45% identity to Lhx6 |
| GB11571-RA | GroupUn.118 | 4 | HOX | Homeodomain protein dbx |
| GB30353-RA | GroupUn.49 | 3 | HOX | "TCL: T-cell leukemia, homeobox 1 " |
| GB11491-RA | GroupUn.1 | 2 | HOX | Gnot1 homeodomain protein |
| GB30330-RA | GroupUn.3697 | 2 | HOX | Transcription factor DRG11 |
| GB14165-RA | GroupUn.30 | 4 | HOX | Reversed polarity ; goosecoid |
| GB20009-RA | GroupUn.1701 | 4 | HOX | Homeobox protein rough |
| GB11098-RA | GroupUn.209 | 3 | HOX | Homeo box HB9; Homeobox protein rough; |
| GB18266-RA | GroupUn108 | 4 | HOX | Mesenchyme homeo box 2 |

| | | | | |
|---|---|---|---|---|
| | | | | (growth arrest-specific homeo box) |
| GB30291-RA | GroupUn.1841 | 2 | HOX | Homeobox protein aristaless; HOX domain |
| GB11536-RA | GroupUn.548 | 6 | HOX | Aristaless-like homeobox protein |
| GB30426-RA | GroupUn.4568 | 2 | HOX | Homeobox protein B-H1 (Homeobox BarH1 protein) |
| GB10709-RA | GroupUn.1038 | 6 | 5LIM | LIM and senescent cell antigen-like domains 1 |

**Table S15. Candidate new bee venom components.**

**Homologues of known insect allergens**

| Allergen | Genbank Acc N° | Species | Function | Glean3 Acc N° | E-value |
|---|---|---|---|---|---|
| Aed a 1 | GI:556272 | Aedes aegypti | apyrase | GLEAN3_08103 | 1e-85 |
| Bla g 2 | GI:1703445 | *Blatella germanica* | aspartic protease | GLEAN3_04551 | 4e-38 |
| Bla g 5 | GI:2326190 | Blatella germanica | glutathione transferase | GLEAN3_08030 | 6e-49 |
| Per a 1 | GI:2580504 | *Periplaneta americana* | Cr-PII | GLEAN3_06235 | 5e-49 |
| Per a 3 | GI:1580797 | *Periplaneta americana* | Cr-PI | GLEAN3_05583 | 3e-61 |
| Per a 7 | GI:4468638 | *Periplaneta americana* | tropomyosin | GB17608-RA | 4e-27 |
| Chi k 10 | GI:42559556 | *Chironomus kiiensis* | tropomyosin | GLEAN3_04894 | 3e-80 |
| Lep s 1 | GI:20387026 | *Lepisma saccharina* | tropomyosin | GB10939-RA | 4e-27 |
| Dol m 1 | GI:548449 | *Dolichovespula maculata* | phospholipase A1 | GLEAN3_03364 | 5e-33 |
| Dol m 5 | GI:137395 | *Dolichovespula maculata* | antigen 5 | GLEAN3_01188 | 7e-18 |

| | | | | | |
|---|---|---|---|---|---|
| Dol a 5 | GI:465052 | *Dolichovespula arenaria* | antigen 5 | GLEAN3_01188 | 2e-16 |
| Pol a 1 | GI:14423833 | *Polistes annularies* | phospholipase A1 | GLEAN3_03364 | 3e-33 |
| Pol a 5 | GI:465053 | *Polistes annularies* | antigen 5 | GLEAN3_01188 | 1e-19 |
| Pol d 5 | GI:6136164 | *Polistes dominulus* | | GLEAN3_01188 | 1e-19 |
| Pol e 5 | GI:549187 | *Polistes exclamans* | | GLEAN3_01188 | 2e-19 |
| Pol f 5 | GI:549188 | *Polistes fuscatus* | | GLEAN3_01188 | 1e-20 |
| Pol g 5 | GI:25091511 | *Polistes gallicus* | | GLEAN3_01188 | 7e-19 |
| Ves f 5 | GI:549189 | *Vespula flavopilosa* | antigen 5 | GLEAN3_01188 | 9e-19 |
| Ves g 5 | GI:74035841 | *Vespula germanica* | antigen 5 | GLEAN3_01188 | 1e-18 |
| Ves m 5 | GI:85681830 | *Vespula maculifrons* | antigen 5 | GLEAN3_01188 | 4e-18 |
| Ves p 5 | GI:549192 | *Vespula pennsylvanica* | antigen 5 | GLEAN3_01188 | 1e-18 |
| Ves s 5 | GI:549193 | *Vespula squamosa* | antigen 5 | GLEAN3_01188 | 1e-17 |
| Ves vi 5 | GI:549194 | *Vespula vidua* | antigen 5 | GLEAN3_01188 | 3e-17 |
| Ves v 1 | GI:1352699 | *Vespula vulgaris* | phospholipase A1 | GLEAN3_03364 | 7e-34 |
| Ves v 5 | GI:11514279 | *Vespula vulgaris* | antigen 5 | GLEAN3_01188 | 5e-19 |

| | | | | | |
|---|---|---|---|---|---|
| Sol i 3 | GI:14424466 | *Solenopsis invicta* | | GLEAN3_01188 | 7e-21 |
| Tria p 1 | GI:15426413 | *Triatoma protracta* | procalin | GLEAN3_06096 | 4e-04 |

## Homologues of known snake and scorpion venom components

| Protein family | Genbank Acc N° | Species | Function | Glean3 Acc N° | E-value |
|---|---|---|---|---|---|
| desintegrins | GI:461932 | Calloselasma rhodostoma | hemorrhagic protein-rhodostomin precursor | GLEAN3_06023 | 2e-50 |
| | | | | GLEAN3_05408 | 6e-37 |
| | | | | GLEAN3_04160 | 4e-24 |
| desintegrins | GI:67462322 | Cryptelytrops albolabris | disintegrin albolabrin | GLEAN3_05408 | 5e-14 |
| | | | | GLEAN3_06023 | 3e-12 |
| | | | | GLEAN3_04160 | 6e-07 |
| desintegrins | GI:50400453 | *Agkistrodon piscivorus piscivorus* | zinc metalloproteinase | GLEAN3_06023 | 3e-59 |
| | | | | GLEAN3_05408 | 1e-47 |

| | | | | | |
|---|---|---|---|---|---|
| | | | | GLEAN3_04160 | 1e-18 |
| neurotoxin | GI:33187130 | *Vipera aspis* | ammodytin I1 isoform 1 | GLEAN3_00224 | 4e-04 |
| neurotoxin | GI:33187116 | *Vipera aspis* | vaspin B | GLEAN3_00224 | 1e-09 |
| neurotoxin | GI:68705 | Bungarus multicinctus | beta-1 bungarotoxin chain B | GLEAN3_01308 | 9e-08 |
| anticoagulant peptide | GI:39932463 | *Mesobuthus martensii* | venom peptide BmKAPi precursor | GLEAN3_08736 | 5e-06 |
| anticoagulant peptide | GI:27903821 | *Mesobuthus martensii* | venom peptide BmKAPi precursor | GLEAN3_08736 | 5e-06 |

## References


1. Beye, M. & Raeder, U. Rapid DNA preparation from bees and %GC fractionation. *Biotechniques* , 372-4 (1993).
2. Levan, A., Fredga, K. & Sandberg, A. Nomenclature for centromeric position on chromosomes. *Hereditas* , 201-220 (1964).

**Table S16. Mean population differentiation ($F_{ST}$) for evolutionary lineages of *Apis mellifera*, based on 1136 SNPs.**

| | Mean $F_{ST}$ |
|---|---|
| Among all subspecies (N-10) | 0.501 |
| Major major lineages (M, A, C and O) | 0.471 |
| M vs. A | 0.242 |
| M vs. C | 0.565 |
| M vs. O | 0.458 |
| A vs. C | 0.354 |
| A vs. O | 0.256 |
| C vs. O | 0.332 |

$F_{ST}$ values were calculated using Weir and Cockerham's unbiased estimator (Weir, B. S. & Cockerham, C. C. Estimating f-statistics for the analysis of population structure. Evolution 38, 1358-1370 (1984). Geographical subspecies (each represented by 9-21 individuals) are described in Supplementary Methods; these subspecies are divided into 4 major lineages (M, A, C and O) as represented in Figure 10.

**Table S17 Access to the Genome Assemblies.**

| Assembly | Date | BCM FTP Subdirectory | Browser and Identifier |
|---|---|---|---|
| 4.0 | March 10, 2006 | Amel20060310-freeze | BeeBase 4.0<br>NCBI AADG05000000 |
| 3.0 | May 1, 2005 | Amel20050501-freeze | BeeBase 3.0<br>NCBI AADG05000000 |
| 2.0 | January 20, 2005 | Amel20050120-freeze | BeeBase 2.0<br>NCBI AADG04000000<br>UCSC apiMel2 |
| 1.2 | July 20, 2004 | Amel20040720-freeze | NCBI AADG03000000<br>UCSC apiMel1 |
| 1.1 | January 20, 2004 | Amel20040120-freeze | NCBI AADG02000000 |
| 1.0 | December 10, 2003 | Amel20031211-freeze | NCBI AADG01000000 |