Update TRENDS in Genetics Vol.xxx No.x

Full text provided by www.sciencedirect.com ScienceDirect

Genome Analysis

Quantification of insect genome divergence

Evgeny M. Zdobnov[1] and Peer Bork[2,3]

[1] University of Geneva Medical School, Department of Genetic Medicine and Development, Geneva 1211, Switzerland
[2] EMBL Heidelberg, Meyerhofstrasse 1, D-69117 Heidelberg, Germany
[3] MDC Berlin-Buch, Robert-Roessle-Strasse 10, D-13092 Berlin, Germany

The recent sequencing of twelve insect genomes has enabled us to quantify their divergence using synteny conservation and sequence identity of single-copy orthologs. Protein identity correlates well with synteny and is about three times more conserved, an observation consistent with comparisons among vertebrates. The observed distribution of the lengths of synteny blocks follows a power law and differs from the expectations of the currently accepted random breakage model. Our results show that there is only limited selection for conservation of gene order and reveal a few hundred genes, proximity among which seems to be vital.

Introduction

Insects are the largest and the most diverse group of invertebrate animals on earth, comprising over 800 000 species. They first appeared in the fossil record ~350 million years ago [1] as already specialized species representing >6 different orders. Insects greatly affect human agriculture and health, transmitting devastating parasitic diseases, such as malaria, and include well-studied model organisms, such as *Drosophila*. This has provided strong justification for several completed and on-going whole-genome sequencing projects. Although each of the organisms is unique and has its own story to be told, here we quantify the divergence among 12 insect lineages: the honeybee (*Apis mellifera*) representing the Hymenoptera order, the red flour beetle (*Tribolium castaneum*) from the Coleoptera, the silkworm moth (*Bombyx mori*) from the Lepidoptera, seven fly species (*Drosophila melanogaster, Drosophila erecta, Drosophila ananassae, Drosophila pseudoobscura, Drosophila mojavensis, Drosophila virilis and Drosophila grimshawi*) and two mosquitoes (*Anopheles gambiae* and *Aedes aegypti*) from the Diptera to investigate the evolutionary mechanisms shaping animal genomes. Many measures of genome divergence seem to correlate with each other [2], hence we used the multiple insect genomes to reliably quantify the correlation between genome shuffling (synteny; see Glossary) and protein sequence identity, and to investigate the extent to which gene order in animal genomes is functionally constrained.

Identification of the conserved core of genes

Despite tremendous efforts in gene prediction, our knowledge of the full gene repertoire is incomplete [3,4] and even the definition of a gene remains fuzzy [5]. However, genes that are well conserved among different species are easier to annotate, and thus the preliminary subset of the conserved cores of genes of the 12 recently sequenced insect genomes compiled for this study is likely to be of sufficient quality to reveal accurate relations among these genomes. To quantify the divergence of these genomes, we focused on the fraction of single-copy orthologs, that is, genes that have exactly one ortholog in each of the genomes. This is the ideal marker set to study rates of evolution as these genes are most likely to retain their ancestral function and thus to evolve under similar constraints (e.g. [6]). A quantification of divergence (Figure 1a) is also an essential reference for other evolutionary studies, as most comparative techniques are applicable only within a certain window of sequence divergence (e.g. [2]). Although the trends identified here based on 4632 universal single-copy orthologs (Box 1) are likely to be robust, the exact numbers given below should be considered as lower limits because they are derived using automatic methods for gene prediction and orthology identification, which have limitations (Box 1).

Quantification of protein substitutions

Insect genomes are much more diverse than those of comparable vertebrate lineages [2,7]. When using single-copy orthologs, the pairwise insect genome comparisons show an average protein identity conservation ranging from 53 to 95%, depending on the evolutionary distance (Figure 1a), which accords with a generally clock-like mode

Glossary

Genome shuffling: the process by which the linear gene order in genomes is disrupted by chromosomal rearrangements (e.g. inversions), which occur frequently during evolution. After speciation, independently evolving genomes accumulate many of such diversifying variations, so that they appear as mosaics of chromosomal regions with common evolutionary ancestry. The term **synteny** nowadays is often used to refer to orthologous genomic regions in these mosaics.

Power law: A function is a power law if the dependent variable (*x*) *is raised to some power (n), known therefore as "x to the power of n". A power law relationship between x and y can be written as* $y = ax^k$*, and thus it can be seen as a straight line on a log–log graph (log (y) = k log (x) + log (a)).*

The amino acid substitution model: a model of protein sequence evolution, a substitution matrix, in which amino acids mutate randomly and independently from one another but according to some predefined probabilities depending on the amino acids themselves. Such models are essential for homology searches and phylogenetic analysis.

The random breakage model: A model put forward in 1973 by Susumu Ohno, who proposed that chromosomal rearrangements occur randomly and that breaks of the ancestral genome are uniformly distributed along the chromosomal length. This model suggests an exponential probability to observe a synteny block of certain length.

Corresponding author: Zdobnov, E.M. (zdobnov@medecine.unige.ch).
Available online xxxxxx.

  doi:10.1016/j.tig.2006.10.004

Please cite this article in press as: Zdobnov, E.M., Bork, P., Quantification of insect genome divergence, Trends Genet. (2006), doi:10.1016/j.tig.2006.10.004

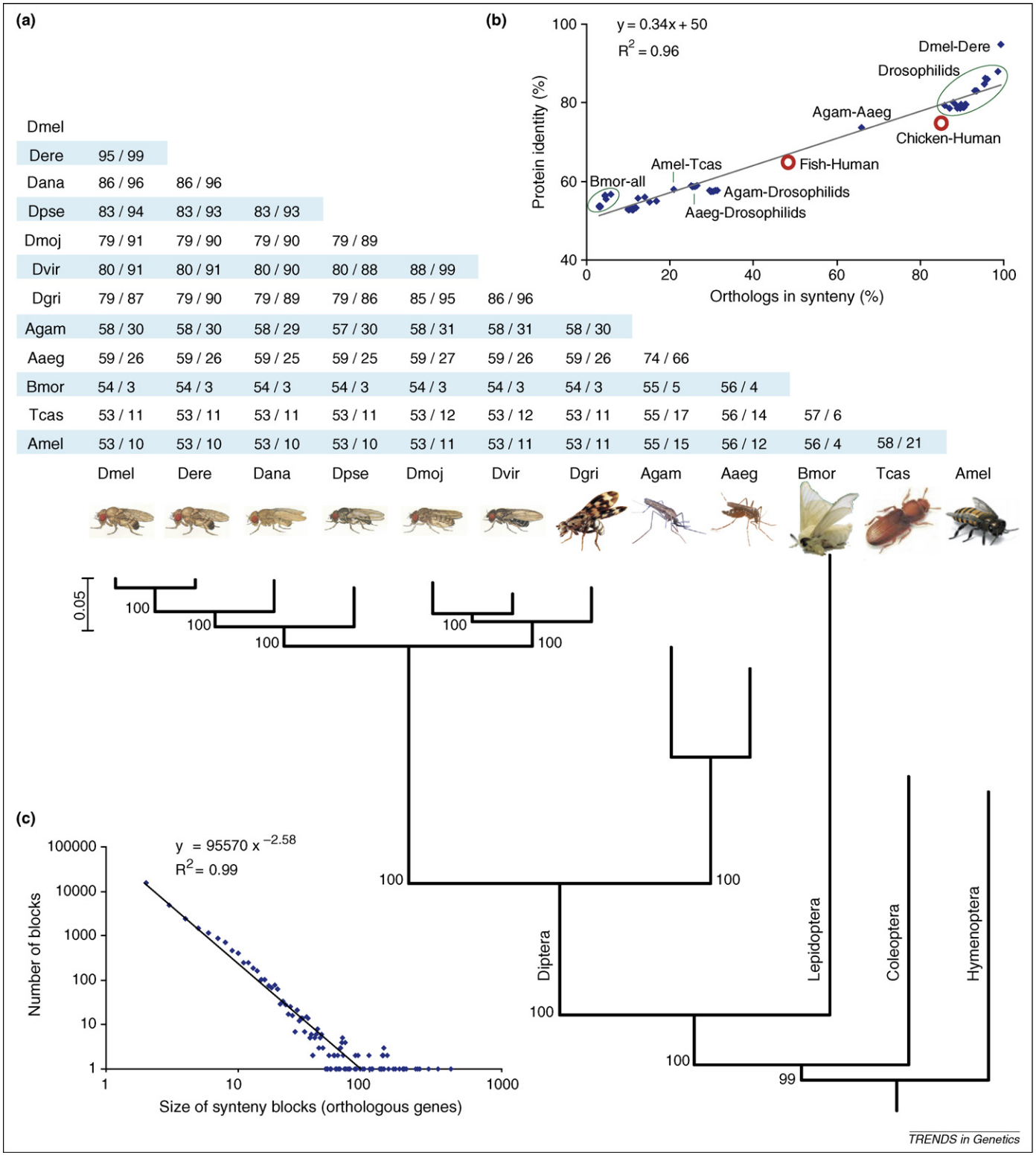

Figure 1. Pairwise divergence of sequenced insect genomes. **(a)** Table of the divergence of the insect genomes (top), measured in terms of average percent of protein identity of single-copy orthologs (the left figure in each pair) and the percentage of these orthologs remaining in synteny (the right figure), and a species divergence tree (bottom), robustly reconstructed using the maximum-likelihood method to model amino acid substitutions in genome-wide alignment of single-copy orthologs. The tree corresponds well to the known phylogeny of the species [8] and reliably quantifies the lineage-specific variation of evolutionary rates. **(b)** Correlation of the pairwise average protein sequence identity of single-copy orthologs and the fraction of these genes remaining in synteny. It is also consistent with vertebrate data comparing human to chicken and to fish genomes [12]. **(c)** Graph of the frequency of synteny block lengths. The distribution is governed by the power law rather than the exponential law implied by the random breakage model. Synteny blocks from all pairwise comparisons were mixed together for this analysis. Abbreviations: Dmel, *D. melanogaster*; Dere, *D. erecta*; Dana, *D. ananassae*; Dpse, *D. pseudoobscura*; Dmoj, *D. mojavensis*; Dvir, *D. virilis*; Dgri, *D. grimshawi*; Agam, *A. gambiae*; Aaeg, *A. aegypti*; Bmor, *B. mori*; Amel, *A. mellifera*; Tcas, *T. castaneum*.

Please cite this article in press as: Zdobnov, E.M., Bork, P., Quantification of insect genome divergence, Trends Genet. (2006), doi:10.1016/j.tig.2006.10.004

Box 1. Methods

Genomes
This study was based on the following genome assemblies:
T. castaneum: Baylor, Tcas_1.0 (NHGRI U54 HG003273);
A. mellifera: Baylor, Amel_2.0 (NHGRI and USDA);
B. mori: BGI, 2003-10-01 [16];
A. gambiae: *Anopheles* Genome Consortium, AgamP3 [17];
A. aegypti: *Aedes* Genome Consortium, aedes_aegyti_1 (NIAID U01-AI050936);
D. erecta: ARACHNE assembly from Agencourt 20050801 (NHGRI);
D. ananassae: ARACHNE assembly from Agencourt 20050801 (NHGRI);
D. pseudobscura: Flybase, release 1.04 [18];
D. melanogaster: Flybase, release 4 [18];
D. virilis: ARACHNE assembly from Agencourt 20050801 (NHGRI);
D. mojavensis: ARACHNE assembly from Agencourt 20050801 (NHGRI);
D. grimshawi: ARACHNE assembly from Agencourt 20050801 (NHGRI).

Gene prediction
For each of the mentioned genomes, the homology-based gene prediction pipeline (E. Zdobnov, unpublished) was applied, which relies on similarity to known proteins to identify putative genes using tBlastN [19] followed by gene model prediction using Fgenesh++ [20]. The tentative predictions were then used for orthology identification, as described below, to filter out the inter-species conserved core of genes. This effort aimed to provide an unbiased view on multiple genomes rather than to produce a complete catalog of genes for each species.

Orthology
Groups of orthologous genes were automatically identified using a variant of a strategy employed previously [7,12], based on all-against-all protein comparisons using the Smith-Waterman algorithm, followed by clustering of reciprocally best matching triangles between each set of three species that overlap by at least 30aa to avoid the domain walking effect. Furthermore, the orthologous groups were expanded by genes that are more similar to each other within a genome than to any gene in any of the other genomes.

We then focused only on the single-copy orthologs shared among all of the genomes. However, instead of the strict counting of single-copy orthologs, we applied more relaxed criteria to allow for a missing gene or a copy-number run-away in one of the twelve species, to compensate for the draft quality of genomes such as *Bombyx mori*.

Synteny
To identify conserved blocks of gene arrangements (synteny), we enumerated 4632 single-copy orthologs (as loosely defined above) along the chromosomes and then grouped them into synteny blocks using a variant of an earlier strategy [7,12] that requires at least two orthologs to be next to each other in both genomes and not allowing more than one intervening gene. Assuming random gene order in different genomes, the probability of finding such a minimal synteny block by chance can be estimated as $<4 \times 10^{-3}$ for 2-way synteny and $<4 \times 10^{-6}$ for three-way synteny.

Phylogeny
The 2302 orthologous genes found exactly in one copy in all of the 12 genomes were used to produce multiple protein alignments for each of the orthologous groups using Muscle [21]. The well aligned regions of these alignments were extracted using Gblocks [22] with default parameters and concatenated into one alignment comprising 705 502 amino acids for phylogenetic tree reconstruction using the maximum likelihood method as implemented in PHYML [23] and TREE-PUZZLE [24], using the JTT [25] model for amino acid substitutions with a γ correction with four discrete classes, an estimated α parameter and proportion of invariable sites. The values of statistical support shown in Figure 1 were obtained from 500 replicates of bootstrap analyses. The tree was rooted by applying the same procedure to 1150 single-copy orthologs between insects and vertebrates (see Figure S1 in the Supplementary Data online).

of genome divergence. Although only the most conservative parts of the proteomes have been used, the distributions of identity in each pair of organisms are relatively broad and vary between different functional gene classes. Having such a broad distribution of sequence identity at the level of encoded proteins limits the applicability of many methods that rely on comparative analysis of nucleotide substitutions in distantly related insects. We therefore used the amino acid substitution model to quantify more accurately the level of species divergence and the lineage-specific evolutionary rates, applying both maximum-likelihood and neighbor-joining methods to the concatenated alignments of the single-copy orthologs. This produced completely robust trees (Figure 1a) that are consistent with the previously estimated insect phylogeny [1,8]. Because the rates of evolution are variable, a defined mapping of the radiation time points in accordance with fossil records is complicated and beyond the aims of this study. Nevertheless, this genome-wide phylogenetic analysis is statistically robust and provides a reliable quantification of lineage-specific evolutionary rates (Figure 1a), confirming the previously suggested more ancestral state of the honeybee [9] and flour beetle [10] genomes. Using single-copy orthologs between insects and vertebrates enables confident rooting of the phylogenetic tree (see Figure S1 in the Supplementary Data), supporting the recently suggested [11] more basal position of Hymenoptera with respect to Coleoptera.

Quantification of genome shuffling

Another important genome comparison measure is the difference in gene arrangements along the chromosomes (synteny), which is affected by genome rearrangement events that are not obviously related to single-nucleotide mutations causing protein substitutions discussed above. Synteny analysis is more prone to errors than ortholog comparison, as additional artifacts in current datasets, such as fragmented and imperfect genome assemblies, can lead to underestimates of the length of synteny blocks. Nevertheless, synteny analysis enables confident assignment of larger orthologous genomic regions and hints at global trends in genome shuffling.

As we observed previously in a comparative study of the malaria mosquito and the fruit fly genomes, the chromosomal content (macro-synteny) is much better preserved than local gene arrangements (micro-synteny) [7]. A good quantification of micro-synteny, which is also rather robust to the currently fragmented chromosome assemblies, is the fraction of single-copy orthologs that are retained in synteny blocks of the total number of shared orthologs. Micro- synteny reveals an almost complete spectrum of levels of genome shuffling in insects (Figure 1a), something not previously observed among comparable vertebrate lineages; for example, even fish and human have ~50% of orthologs in synteny [12]. In insects this ranges from 99% of orthologs having the same neighbors in *D. melanogaster* and *D. erecta* to only ~10%

Please cite this article in press as: Zdobnov, E.M., Bork, P., Quantification of insect genome divergence, Trends Genet. (2006), doi:10.1016/j.tig.2006.10.004

remaining together when comparing drosophilids with honeybee (the figure drops to 3% when comparing drosophilids with silkworm, but this might be an underestimate because of the highly fragmented nature of the current silkworm genome sequence). The current data confirms our previous observation [2] that these seemingly unrelated measures of protein sequence identity and gene orders correlate well (Figure 1b), despite the substantial variation in the rates of evolution in some lineages as measured by the amino acid substitution model described above (Figure 1a). Although the actual relationship between these measures is likely to be more complex, it fits a linear approximation, whereby the rate of genome shuffling is ~3 times greater than the rate of accumulation of substitutions in protein sequences. Extrapolating this trend suggests that ancestral gene order is completely lost when the average protein identity is lower than 50% and implies limited constraints for preserving synteny.

The commonly accepted random breakage model of chromosome evolution [13] suggests that the probability of observing synteny blocks of a certain length varies exponentially with length. Yet, the current data from all pairwise comparisons do not support the model and can be fitted much better to a power function (Figure 1c). The previously available low-resolution synteny data did not have a sufficient variation range to distinguish between power law and exponential distributions. This finding therefore suggests a more complex pattern of genome rearrangements. An exponential (rather than uniform) distribution of breakpoints along chromosomes, such as around rearrangement hotspots, could explain the observed data, but more detailed studies are required.

Identification of functionally constrained blocks of genes

Despite the rapid decay of genomic gene order, the arrangement of a few genes is known to be constrained, such as the Hox gene cluster. To identify such conserved blocks of genes, we considered three-way synteny, requiring conservation of gene arrangements in three genomes of the most distantly related insects: (i) honeybee, fruit fly and mosquito, (ii) honeybee, beetle and fruit fly, and (iii) honeybee, beetle and mosquito. Multi-species synteny is particularly prone to underestimates, but it is the best filter for discriminating the gene proximities under functional selection from the remnants of ancestral gene arrangements. The probability of finding a random pair of orthologs next to each other (Box 1) in three genomes is $\sim 1 \times 10^{-5}$, and this figure indicates the significance of finding 266–380 genes organized into 126–178 synteny blocks, mostly overlapping between the considered genome triples (see Table S1 in the Supplementary Data). The most prominent example of gene order conservation is a cluster of glucose-methanol-choline oxidoreductases nested in the 80 kb intron of the *D. melanogaster Flo-2* gene on the X chromosome. Although this arrangement might have been frozen in evolution as a result of selection pressure to keep the parent gene intact, it seems that nested genes make up only a minority of the constrained genes identified, supporting a more direct functional constraint for the remaining cases. For example, although the disintegration of the Hox gene complex in *Drosophila* has been noted before [14] and the functional necessity of the gene order in this complex has been questioned, we found the *pb* and *Dfd* Hox genes to be persistent neighbors in these three-way comparisons. Another example is the conserved order of the *Wnt6*, *Wnt10* and *ninaC* genes, involved in developmental and signaling pathways, having the vertebrate orthologs of *Wnt6* and *Wnt10* also next to each other [15].

Identification of homologous chromosomal elements

Beyond local gene arrangement, macro-synteny can be established at the level of chromosomal elements for species as divergent as the fruit fly and the malaria mosquito [7], both of which have three major chromosomes (plus a minor one in *D. melanogaster*). However, the number of chromosomes varies in more distantly related insects: 28 in silkworm, 10 in the flour beetle and 16 in honeybee. When comparing the content of the 16 honeybee chromosomes with the fly and mosquito chromosomes, only a few confident correspondences could be established, with the most prominent ones linking *D. melanogaster* chromosomal arm 3R with *A. mellifera* Group 15 and 4 and *D. melanogaster* chromosomal arm 2L with *A. mellifera* Group 3; the rest of the relations are at the level of random expectation (Table S2 in the Supplementary Data).

Concluding remarks

The availability of a considerable number of the insect genomes with variable degrees of divergence now enables the study of genome rearrangements in much greater detail, and indeed we were able to quantify the rate of genome shuffling and to model the decay of ancestral synteny. This revealed a limited selection on gene order in insects and highlighted functionally constrained gene arrangements. The quantified divergence of the insect genomes provides an essential reference for further in-depth analyses aiming to identify, on the one hand, distinct molecular functions of genes in each of the organisms, and on the other hand, general evolutionary mechanisms.

Acknowledgements

We are grateful for the DNA sequence data from Agencourt, Baylor, BGI, Broad Institute, Celera Genomics, Flybase and TIGR and for the Fgenesh++ software kindly provided by Prof. V.V. Solovyev. We thank Dr E.V. Kriventseva and R.M. Waterhouse for discussions and help with the manuscript, and we also acknowledge support from Swiss National Science Foundation (SNF 3100A0–112588/1).

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tig.2006.10.004.

References

1 Grimaldi, D. and Engel, M. (2005) *Evolution of the Insects,* Cambridge University Press
2 Zdobnov, E.M. *et al.* (2005) Consistency of genome-based methods in measuring Metazoan evolution. *FEBS Lett.* 579, 3355–3361
3 Eyras, E. *et al.* (2005) Gene finding in the chicken genome. *BMC Bioinformatics* 6, 131
4 Brent, M.R. and Guigo, R. (2004) Recent advances in gene structure prediction. *Curr. Opin. Struct. Biol.* 14, 264–272

Please cite this article in press as: Zdobnov, E.M., Bork, P., Quantification of insect genome divergence, Trends Genet. (2006), doi:10.1016/j.tig.2006.10.004

5 Snyder, M. and Gerstein, M. (2003) Genomics. Defining genes in the genomics era. *Science* 300, 258–260
6 Ciccarelli, F.D. *et al.* (2005) Complex genomic rearrangements lead to novel primate gene function. *Genome Res.* 15, 343–351
7 Zdobnov, E.M. *et al.* (2002) Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science* 298, 149–159
8 Whiting, M.F. (2002) Phylogeny of the holometabolous insect orders based on 18S ribosomal DNA: when bad things happen to good data. *EXS* 92, 69–83
9 Raible, F. *et al.* (2005) Vertebrate-type intron-rich genes in the marine annelid *Platynereis dumerilii*. *Science* 310, 1325–1326
10 Savard, J. *et al.* (2006) Genome-wide acceleration of protein evolution in flies (Diptera). *BMC Evol. Biol.* 6, 7
11 Savard, J. et al. (2006) Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects. *Genome Res.* DOI:10.1101/gr.5204306 (www.genome.org)
12 Hillier, L.W. *et al.* (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432, 695–716
13 Nadeau, J.H. and Taylor, B.A. (1984) Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Natl. Acad. Sci. U. S. A.* 81, 814–818
14 Negre, B. *et al.* (2005) Conservation of regulatory sequences and gene expression patterns in the disintegrating *Drosophila* Hox gene complex. *Genome Res.* 15, 692–700
15 Miller, J.R. (2002) The Wnts. *Genome Biol* 3, REVIEWS3001
16 Xia, Q. *et al.* (2004) A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science* 306, 1937–1940
17 Holt, R.A. *et al.* (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* 298, 129–149
18 Drysdale, R.A. and Crosby, M.A. (2005) FlyBase: genes and gene models. *Nucleic Acids Res.* 33, D390–D395
19 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
20 Salamov, A.A. and Solovyev, V.V. (2000) Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.* 10, 516–522
21 Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797
22 Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17, 540–552
23 Guindon, S. and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52, 696–704
24 Schmidt, H.A. *et al.* (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18, 502–504
25 Jones, D.T. *et al.* (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* 8, 275–282

Please cite this article in press as: Zdobnov, E.M., Bork, P., Quantification of insect genome divergence, Trends Genet. (2006), doi:10.1016/j.tig.2006.10.004