Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Insect Science

A maturing understanding of the composition of the insect gene repertoire

Robert M Waterhouse[1,2,3,4]

Recent insect genome sequencing initiatives have dramatically accelerated the accumulation of genomics data resources sampling species from different lineages to explore the incredible diversity of insect biology. These efforts have built a comprehensive catalogue of the insect gene repertoire, which is expanded with each newly-sequenced genome and continually refined using knowledge from cross-species comparisons and new sources of evidence. Since the sequencing of the very first insect genomes, comparative analyses have identified shared (homologous) and equivalent (orthologous) genes, as well as subsets of genes that appear to be unique. With the number of available insect genomes fast approaching one hundred, a maturing understanding of the composition of the insect gene repertoire broadly partitions it into an expected core of universally-present orthologues and a diverse array of lineage-specific and species-specific genes. While homology and orthology help to build evolutionarily-informed functional hypotheses for many genes from these newly-sequenced genomes, experimental interrogations are required to test such hypotheses and to probe the functions of genes for which homology offers no clues. Such taxonomically-restricted genes may represent the current contents of an evolutionary melting pot, out of which novel adaptations have emerged to make insects the most successful group of animals on Earth.

Addresses
[1] Department of Genetic Medicine and Development, University of Geneva Medical School, rue Michel-Servet 1, 1211 Geneva, Switzerland
[2] Swiss Institute of Bioinformatics, rue Michel-Servet 1, 1211 Geneva, Switzerland
[3] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar Street, Cambridge, MA 02139, USA
[4] The Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, MA 02142, USA

Corresponding author: Waterhouse, Robert M (Robert.Waterhouse@unige.ch, robert.waterhouse@gmail.com)

Current Opinion in Insect Science 2015, **7**:15–23

This review comes from a themed issue on **Insect genomics**

Edited by **Susan Brown** and **Denis Tagu**

For a complete overview see the Issue and the Editorial

Available online 13th January 2015

http://dx.doi.org/10.1016/j.cois.2015.01.004

From flies and moths to bees and beetles: discovering shared, equivalent, and unique insect genes

Advances in DNA and RNA sequencing technologies have dramatically increased the opportunities to explore the molecular composition of life and begin to relate it to the fascinating plethora of biological diversity. Insects, as the most successful group of terrestrial animals, demonstrate countless evolutionary adaptations that exploit almost every ecological niche on Earth. This, together with their generally compact genomes (although genome sizes do vary considerably e.g. [1]), makes them ideal for investigating how conservation or divergence and gains or losses of functional genomic elements give rise to the observed splendour of insect biology.

The discovery and functional interrogation of such genomic elements has focused on protein-coding genes, undoubtedly influenced by molecular biology's central dogma from transcription through translation to build the proteins necessary to sustain cellular life. With an initial set of 13,601 protein-coding genes, the fruit fly, *Drosophila melanogaster*, as a model organism and pioneer in animal genetics, was the first insect to have its genome sequenced and annotated [2]. One of the first questions to be asked of these genes was how they compared to those of other organisms, that is which genes were shared (homology), which were equivalent (orthology), and which appeared to be unique. Comparisons with the only two other sequenced eukaryotes at the time identified worm orthologues for about 30% of fly genes and a conserved core of about 20% with orthologues in both worm and yeast [3]. Subsequent sequencing of the genome of the *Anopheles gambiae* malaria mosquito [4] created the first opportunity for the comparative analysis of two insect genomes. Over this relatively much shorter evolutionary span, pairwise orthology for 55% of fly (and 61% of mosquito) genes could be identified (figure 1A from [5]), with 6089 single-copy orthologues [5]. These observations provided some of the first quantifications of the comparative genomics axiom that there should be more genes recognisably in common amongst closely-related organisms than amongst more distantly-related species.

Over the following few years, insect genome sequencing projects were driven by agricultural, economic, environmental, or human health rationale and sampled the major insect orders, for example, the lepidopteran silk moth, *Bombyx mori* [6,7], a second fruit fly, *Drosophila* *pseudoobscura* [8], the hymenopteran honey bee, *Apis mellifera* [9,10], a second mosquito, *Aedes aegypti* [11,12], and the coleopteran red flour beetle, *Tribolium castaneum* [13] (Table 1). While the initial honey bee and malaria mosquito gene sets were both smaller than that of *D. melanogaster*, the other new insect genomes were predicted to encode some 2000–5000 more genes, although none had as many as the $\sim$24,000 genes from human or mouse. Comparisons with vertebrates provided the first detailed views of gene repertoire evolution across animals, from universally maintained orthologues to lineage-specific and species-specific genes (e.g. figure 5 from [9] and figure 2 from [13]). Taking advantage of these and five additional *Drosophila* genomes, initial large-scale quantitative analyses identified a conserved core of 4632 insect-universal single-copy orthologues that revealed strikingly faster rates of molecular evolution and genome shuffling in insects compared with vertebrates [14]. Such evolutionary dynamism was even observed at within-genus level, with the comparative analysis of 12 *Drosophila* genomes [15]: a larger set of 6698 fly-universal single-copy orthologues, but still many lineage-specific and species-specific genes (figure 2 from [15]), and spanning a greater molecular evolutionary distance than humans to reptiles (figure 1C from [16]).

Beyond holometabola: more comprehensive species sampling and improved gene sets

Both beneficial and harmful impacts on humans continued to influence the sampling of insect species for genome sequencing, for example, the parasitoid *Nasonia* wasps [19], the pea aphid pest [20], the human body louse [21], and the *Culex* mosquito [24] (Table 1). Gene discovery in the first hemimetabolous insect genomes offered strikingly contrasting views: the hemipteran aphid with extensive gene duplications and a total of 34,604 genes,

Table 1

Published insect genomes with their total gene counts, genome browser websites, and principal references.

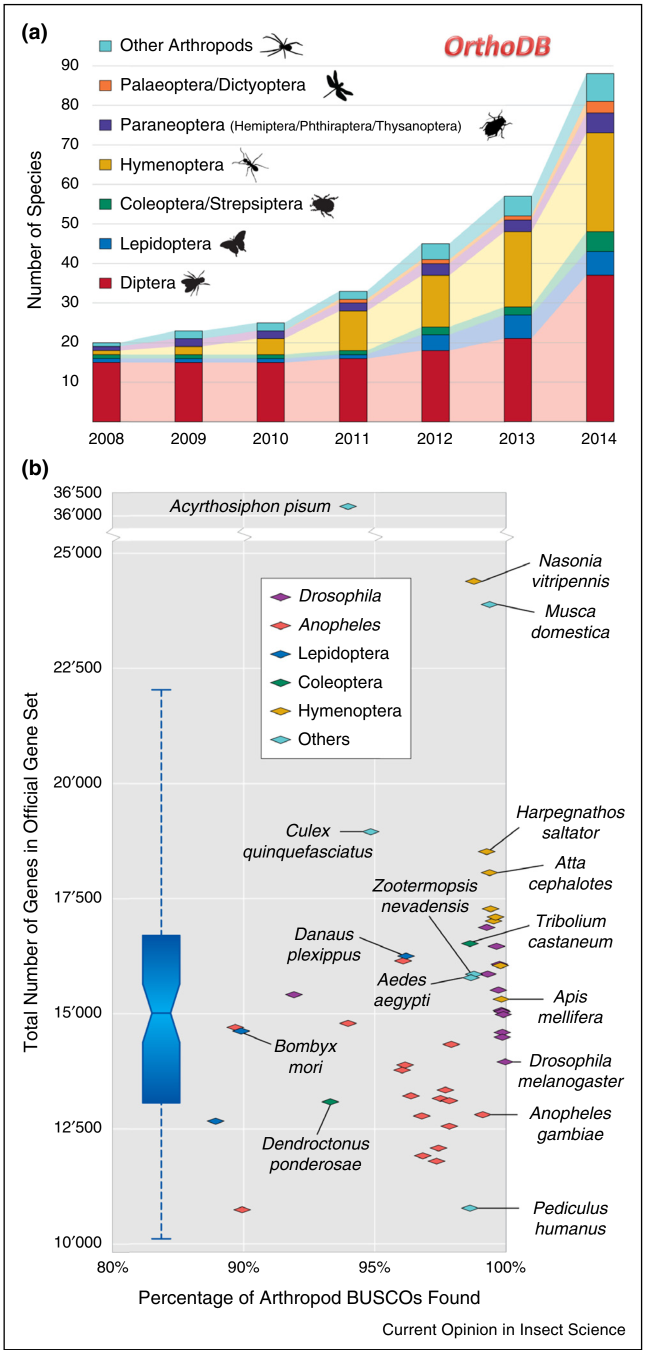
Publication	Species	Common name	Number of genes[†]	Genome browser	Reference(s)
2000.03	*Drosophila melanogaster*	Fruit fly	13,954	http://flybase.org	[2,3]
2002.10	*Anopheles gambiae*	African malaria mosquito	12,810	http://www.vectorbase.org	[4,5]
2004.12	*Bombyx mori*	Domestic silkworm moth	14,623	http://www.silkdb.org	[6,7]
2005.01	*Drosophila pseudoobscura*	Fruit fly	15,864	http://flybase.org	[8]
2006.10	*Apis mellifera*	Honey bee	15,314	http://hymenopteragenome.org	[9,10,17••]
2007.06	*Aedes aegypti*	Yellowfever mosquito	15,784	http://www.vectorbase.org	[11,12]
2007.11	10 *Drosophila*	Fruit flies	15,452[‡]	http://flybase.org	[15,16,18]
2008.04	*Tribolium castaneum*	Red flour beetle	16,524	http://agripestbase.org	[13]
2010.01	*Nasonia vitripennis*	Parasitic jewel wasp	24,389	http://hymenopteragenome.org	[19]
2010.02	*Acyrthosiphon pisum*	Pea aphid	36,195	http://www.aphidbase.com	[20]
2010.07	*Pediculus humanus*	Body louse	10,773	http://www.vectorbase.org	[21]
2010.08	*Harpegnathos saltator*	Jerdon's jumping ant	18,518	http://hymenopteragenome.org	[22,23••]
2010.08	*Camponotus floridanus*	Florida carpenter ant	17,015	http://hymenopteragenome.org	[22,23••]
2010.10	*Culex quinquefasciatus*	Southern house mosquito	18,955	http://www.vectorbase.org	[24]
2011.02	*Atta cephalotes*	Leafcutter ant	18,062	http://hymenopteragenome.org	[25,23••]
2011.04	*Linepithema humile*	Argentine ant	16,048	http://hymenopteragenome.org	[26,23••]
2011.04	*Pogonomyrmex barbatus*	Red harvester ant	17,100	http://hymenopteragenome.org	[27,23••]
2011.04	*Solenopsis invicta*	Red fire ant	16,513	http://hymenopteragenome.org	[28,23••]
2011.08	*Acromyrmex echinatior*	Panamanian leafcutter ant	17,277	http://hymenopteragenome.org	[29,23••]
2011.11	*Danaus plexippus*	Monarch butterfly	16,254	http://monarchbase.umassmed.edu	[30]
2012.07	*Heliconius melpomene*	Postman butterfly	12,669	http://www.butterflygenome.org	[31]
2013.02	*Plutella xylostella*	Diamondback moth	18,073	http://iae.fafu.edu.cn/DBM	[32]
2013.03	*Dendroctonus ponderosae*	Mountain pine beetle	13,088	http://metazoa.ensembl.org	[33]
2013.08	*Anopheles darlingi*	South American malaria mosquito	10,457	http://www.vectorbase.org	[34]
2014.04	*Glossina morsitans*	Tsetse fly	12,308	http://www.vectorbase.org	[35]
2014.05	*Zootermopsis nevadensis*	Dampwood termite	15,860	http://termitegenome.org	[36]
2014.07	*Chilo suppressalis*	Striped riceborer moth	10,221[§]	http://ento.njau.edu.cn/ChiloDB	[37]
2014.09	*Anopheles stephensi*	Indo-Pakistan malaria mosquito	11,789	http://www.vectorbase.org	[38]
2014.09	*Melitaea cinxia*	Glanville fritillary butterfly	16,667	http://metazoa.ensembl.org	[39]
2014.09	*Polypedilum nubifer*	Chironomid midge	16,553[§]	http://bertone.nises-f.affrc.go.jp/midgebase	[40]
2014.09	*Polypedilum vanderplanki*	Sleeping chironomid midge	17,137[§]	http://bertone.nises-f.affrc.go.jp/midgebase	[40]
2014.10	*Musca domestica*	House fly	23,884	http://www.vectorbase.org	[41]
2014.11	16 *Anopheles*	Mosquitoes	13,377[‡]	http://www.vectorbase.org	[42,43••]

[†] Counts from OrthoDB.

[‡] Mean number of genes.

[§] Counts from publication.

Figure 1

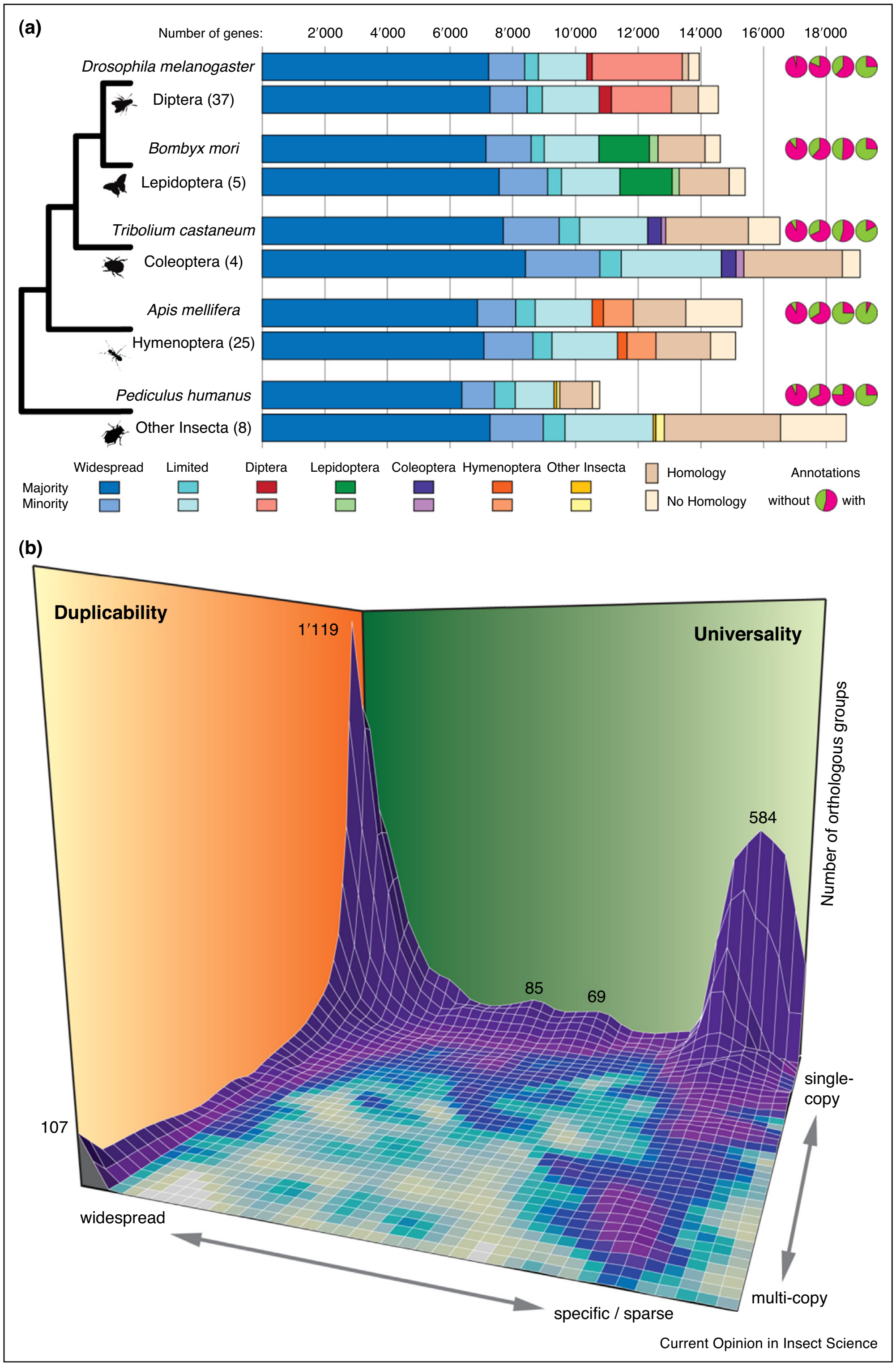


The growing numbers of available insect gene sets show variable total gene counts and completeness. **(a)** The number of insect species with sequenced genomes and annotated gene sets included in OrthoDB, the hierarchical catalogue of orthologues, has more than quadrupled over the last decade. Diptera remains the most well-represented insect order, but current and future genome projects, such as those selected for the *i*5K initiative, aim to improve sampling across arthropods and will lead to the generation of many more insect genomics resources. **(b)** Most insect genomes sequenced to date encode between about 13,000 and 16,500 protein-coding genes, with notable exceptions including the body louse, *Pediculus humanus*, with 10,773 genes and the pea aphid, *Acyrthosiphon pisum*, with 36,195 genes. The boxplot shows the median (15,013) of 80 insect gene sets with the first (13,078) and third quartiles (16,612) of the distribution, and dashed lines extending to 1.5 times the inter-quartile range. Assessing annotations using arthropod Benchmarking Universal Single-Copy Orthologues (BUSCOs) reveals variations in gene set completeness, influenced by the quality of the assembled genome (e.g. contiguity) and the applied annotation approaches (e.g. *ab initio* and/or homology-based methods, combiners, and supporting data like transcriptomes). Comparing gene set size to gene set completeness reveals that small gene sets are not necessarily incomplete and larger gene sets are not necessarily more complete.

many of which appeared to be species-specific; and the phthirapteran louse, with a compact but remarkably complete set of 10,773 genes (figure 3 from [20] and figure 1 from [21]). More recent and ongoing genome projects have focused on harnessing the power of multi-species comparisons, for example, the 16 *Anopheles* mosquito genomes [42,43••], as well as exploring different questions, such as gene flow amongst closely-related members of a species complex [44], or the evolution of sociality in ants [22,23••,25–29], and termites [36]; or of migration [30], mimicry [31], and holocentric chromosomes [39] in butterflies. Species sampling has also been expanded through large transcriptome projects such as 1KITE (1000 Insect Transcriptome Evolution) to detail the timings and patterns of insect evolution through phylogenomics [46]. As part of the *i*5K [47,48•] and other insect genome sequencing initiatives (see '*Best practices in insect genome sequencing from 30 insect genomes*' in this issue), the sampling of species from different lineages continues to improve, exploring the incredible diversity of insect biology and building a more comprehensive catalogue of the insect gene repertoire.

For each newly-sequenced insect genome, an official set of predicted protein-coding genes and associated functional annotations is generated. This process of automated genome annotation must balance different sources of evidence such as *ab initio* gene predictions and alignments of homologues from other insects or RNAseq transcripts from experimental samples (for further details see '*Towards automated gene annotation*' in this issue). *D. melanogaster* and FlyBase (http://flybase.org) have pioneered the model by which the genome sequence itself provides a logical framework for building a comprehensive biological knowledgebase. Importantly, this includes the collation of additional gene functional annotations, as well as editing of the gene models themselves to improve the completeness and accuracy of the gene set as new evidence becomes available. Comparisons with multiple, closely-related species provide a rich source of such evidence, as demonstrated by sequence signature analyses of 12 *Drosophila* genomes to accurately define encoded functional elements, thereby improving protein-coding gene predictions, as well as discovering novel genes [16,18]. The re-annotation of the honey bee genome [17••] also used evidence from closely-related

Figure 2



Insect gene homology landscapes. **(a)** Partitioning insect gene sets by their traceable orthology and homology reveals a spectrum of conservation from widespread orthologues found across the major insect clades to species-specific genes with no recognisable homologues. Insect orthology data from OrthoDB, for representative species and averaged across each clade, delineates genes with widespread (present in all five clades), limited (present in two, three, or four clades), and lineage-specific (present in only one clade) orthologues, found in the majority (>50%) or the minority (<50%) of species in each clade, and species-specific genes with (*e*-value < 1e − 3) and without homologues. The subsets of genes with widespread orthologues have the highest proportions of genes with Gene Ontology and/or InterPro domain annotations species, and combined this with multiple gene-prediction strategies, as well as RNAseq transcriptome and mass spectrometry peptide support, to add about 5000 more protein-coding genes to the initial gene set of only 10,157 genes [10]. The catalogued insect gene repertoire is thus not only expanded with each newly-sequenced genome, but is also continually refined using knowledge from cross-species comparisons and new sources of evidence.

Access to insect gene sets and their associated functional annotations and supporting evidence is provided through online genome browsers at research-community-focused resources such as AgripestBase, FlyBase, Hymenoptera Genome Database, and VectorBase (Table 1). For additional clues about gene function through homology inferences, or evidence supporting improvements to predicted gene models, researchers often turn to interspecies comparisons provided by dedicated orthology resources (e.g. members of the 'Quest for Orthologs' consortium [49]). Having quadrupled over the last decade to a current total of 80 insect species (Figure 1a), the OrthoDB hierarchical catalogue of orthologues (http://www.orthodb.org) provides the most comprehensive orthology catalogue of the insect gene repertoire [50•,51–53]. The database of orthologues facilitates flexible user queries with available gene identifiers and protein descriptors, Gene Ontology and InterPro attributes, as well as gene copy-number profiles or sequence homology searches. With a total of 1,248,883 genes from 80 insects, counts per species range from 10,110 to 36,195 with a median of 15,013 genes (Figure 1b, boxplot). Some variation stems from variable amounts of supporting data, or whether any closely-related species were sequenced, as well as from variable annotation strategies. For example, the unusually large initial gene set of the pea aphid comprised a core of only 10,249 'high-quality' gene models with transcript and/or protein homology support, and a further 24,355 predictions that included unsupported *ab initio* models and likely partial gene models [20].

While different approaches to gene discovery will influence total predicted gene counts, the unique biology of each organism and the evolutionary pressures shaping genomic architectures offer more interesting insights into gene repertoire variations. For example, the relatively small gene set of the body louse might have been explained by reductive evolution common in obligate parasites. However, representative hymenopteran and coleopteran species share more orthologues with the body louse than they do with *D. melanogaster*, suggesting a remarkably complete body louse gene repertoire that is small due to fewer gene duplications rather than many losses (figures 1 and S4.E from [21]). With respect to genomic architectures, higher total gene counts in culicine compared to anopheline mosquitoes may be linked to gene duplications driven by the much more abundant transposable elements in culicine genomes [11,24]. Even without quite as many duplications as the culicines, gene turnover analysis across multiple anopheline genomes revealed much faster rates of gene gain and loss compared to fruit flies [43••].

Total gene counts can therefore be influenced by both annotation strategies and genome biology, making it difficult to judge the completeness of any new insect gene set given only the total number of genes. Quantification of completeness is especially important given how 'next-generation' sequencing technologies usually generate short read-lengths that are difficult to assemble into long, contiguous sequences, which in turn hinders accurate gene annotation. One indication of completeness would be to quantify the proportions of multiple tissue and life-stage transcriptomes that map back to the assembled genome, but this would require further sample collections and sequencing, and is not easily comparable amongst different species. An alternative approach would be to let evolution guide the selection of sets of 'expected' genes by identifying orthologues conserved in genomes of the majority of already-sequenced species. OrthoDB defines such sets of 'Benchmarking Universal Single-Copy Orthologues' (BUSCOs) [51••,54] by selecting representative genes from orthologous groups with single-copy orthologues in at least 90% of species. As these genes are near-universally present as single-copy orthologues across each lineage, they are expected to be found (as single-copy genes) in any newly-sequenced genome of a species from within the lineage. Assessing the completeness of a subset of published insect gene sets using arthropod BUSCOs reveals that most are of high quality, each showing more than 95% recovery of 2753 BUSCOs, and shows that larger gene sets are not necessarily more complete, and that smaller gene sets are not necessarily incomplete (Figure 1b). Such comparative analyses will help to improve the annotations of the conserved core of universal insect orthologues, while further transcriptome sequencing will be required to support and refine lineage-specific or species-specific gene annotations, as well to discover new genes.

The evolutionary histories of 1.25 million insect genes

Speciations, followed by genomic evolutionary events including gene duplications and losses and gene sequence

(Figure 2 Legend Continued) (pie charts from left to right: widespread, limited, lineage-specific, species-specific). **(b)** Dissecting the *Drosophila melanogaster* gene set by orthologous group universality and duplicability highlights how the largest fractions of genes are preserved as single-copy orthologues across all 80 insects or specific to the 12 drosophilids. Orthologous groups with 80 insect species from OrthoDB: universality, from widespread to specific or sparse species representation; duplicability, from mostly single-copy to mostly multi-copy orthologue counts.

mutations, have led to the diversity of genes found in extant species. The common ancestries — or homologies — of such genes are recognized by assessing the statistical significance of the similarities of their aligned sequences. For any given set of species, these homologies define groups of orthologues comprising all surviving descendants of a gene from their last common ancestor. With each newly-sequenced insect genome, orthology delineation contextualizes its gene set by defining evolutionary relationships with genes from other species; with relatively closely-related insects (e.g. figure 1C from [24] and figure 2B from [38]), or more widely across and beyond insects (e.g. figure 3 from [19] and figure 2 from [28]). This effectively partitions insect gene sets according to their evolutionary histories, from taxonomically-widespread orthologues, to those with traceable orthologues from only closely-related species, or species-specific genes with no recognizable homologues.

With almost 1.25 million genes across four major insect orders and several outgroup insect species, OrthoDB orthology delineation identifies a conserved core of 6000–8000 genes per species that are found in the majority of organisms sampled from each clade (Figure 2a, 'Widespread, majority'). Although still widespread, a smaller fraction of the insect gene repertoire comprises 1000–2000 genes per species where orthologues have not been maintained in the majority of organisms from all clades (Figure 2a, 'Widespread, minority'), highlighting how gene losses can play a substantial role in shaping insect gene repertoires (e.g. see [55]). A further 2000–3000 genes per species have orthologues in organisms from at least one other major clade, but are not found in all clades (Figure 2a, 'Limited'), suggesting ancient gene gains or losses during the radiations of the major insect orders.

Amongst the remaining, taxonomically-restricted genes (TRGs), some have identifiable orthologues only with species from the same order or even only with the most closely-related species, while others show no clear orthology and appear as species-specific genes, sometimes called 'orphans' (Figure 2a). These fractions may vary widely as they are influenced both by the evolutionary breadth of the examined lineage, as well as by the density of species sampled within the lineage (see box 2 from [56]). The origins of such genes are often difficult to ascertain; some, with homology to conserved, widespread genes, may have resulted from gene duplications followed by rapid divergence, while those without homology may have diverged beyond recognition, or have emerged de novo. Studies in *Drosophila* indicate that many orphan genes are apparently very short-lived [57], nevertheless, maintained TRGs likely survive due to acquired functions and may be important for the evolution of lineage-specific novelties [56]. This is exemplified by a large family of TRGs from the hessian fly, *Mayetiola destructor*, where functional genomics data provided evidence for the involvement of these genes in modulating insect–plant interactions in an evolutionary arms race that determines successful feeding by *M. destructor* larvae or successful resistance of their wheat plant hosts [58••]. Comparative analysis of ant genomes revealed an abundance of TRGs that may be linked to eusocial adaptations, and comprehensive genome re-annotations identified almost 30,000 of these genes across the seven examined ant species, about 40% of which appeared to be species-specific orphans [23••]. The rate of TRG emergence in Hymenoptera was about twice as fast as in Diptera and may be due to differences in the rate of gene loss rather than gene gain [23••]. Thus, although the biological roles of such TRGs often remain completely unknown, it seems likely that they represent an evolutionary melting pot out of which novel lineage-specific and species-specific adaptations may emerge.

The paucity of TRG functional annotations compared with those of widespread orthologues is highlighted by splitting each subset of evolutionarily-partitioned genes into those with and without Gene Ontology terms or InterPro domains (Figure 2a, pie charts). About 90% or more of widespread orthologues exhibit some clues about gene function, but the proportion of genes with such annotations declines to less than a quarter when examining species-specific genes. Orthology is thus a useful starting point for inferring functions of genes from newly- sequenced organisms, although there remain subsets of genes for which no confident orthology-based functional inferences can be made. This highlights another important axiom of comparative genomics: although orthology is not defined by gene function, inferences of common functions of orthologues remain the most plausible evolutionary scenario, and they thereby help generate functional hypotheses for many newly-discovered genes.

This concept may be further refined to qualify the confidence with which such inferences can be made, for example, more confident hypotheses may be made for universal single-copy orthologues than for orthologues with multiple duplications and/or losses. Partitioning the *D. melanogaster* gene set according to the universality and duplicability of orthologues across 80 insect species highlights a large fraction of mostly-universal, mostly-single-copy orthologues and a second sizeable fraction made up of *Drosophila*-specific, mostly-single-copy orthologues (Figure 2b). These distributions of orthologues demonstrate the dichotomy of gene evolution either under 'single-copy control' or with a 'multi-copy licence' [59]. This suggests that gene dosage constraints likely preserve most universal orthologues as single-copy genes ('single-copy control'), with only a few cases where relaxed copy-number restrictions seem to allow multiple duplications in the majority of descendant lineages ('multi-copy licence') (see figure 1B from [59]). Thus, at least for the sizeable fraction of widespread orthologues evolving under 'single-copy control', orthology offers useful clues about gene function, while such clues are generally less informative, or indeed completely lacking, for lineage-specific or species-specific genes.

Conclusions

The catalogue of insect protein-coding genes has grown substantially since the initial sequencing and annotation of the *D. melanogaster* genome revealed the very first insect gene set. Subsequent sampling of species from across Insecta has allowed comparative genomics to explore the insect gene repertoire and broadly partition it into an expected core of universally-present orthologues and a diverse array of lineage-specific and species-specific genes. Nevertheless, this maturing understanding of the composition of the insect gene repertoire continues to be refined, with new experimental evidence of transcription, as well as from leveraging the power of comparisons across multiple closely-related species. Importantly, such approaches have also begun to catalogue the repertoire of non-protein-coding genes, which seems to exhibit an even more dynamic evolutionary history than that of protein-coding genes. While ongoing efforts that extend species sampling may make only relatively minor revisions to the expected conserved core, they will undoubtedly detail new levels of taxonomically-restricted genes and continue to uncover many novelties. With careful considerations of gene evolutionary histories, orthology helps to drive hypotheses about the functions of genes from newly-sequenced genomes, but the most confident inferences are generally limited to genes of the conserved core. Beyond this core, experimental interrogation holds the key to elucidating the roles of the genes most likely to hold the secrets behind the countless evolutionary adaptations that have allowed insects to exploit almost all ecological niches and become the most successful group of animals on Earth.

Acknowledgements

RMW thanks Stuart A. Mansfield for useful discussions and comments on the draft manuscript. RMW was supported by Marie Curie International Outgoing Fellowship PIOF-GA-2011-303312.

References

1. Hanrahan SJ, Johnston JS: **New genome size estimates of 134 species of arthropods**. *Chromosome Res* 2011, **19**:809-823.

2. Adams M, Celniker S, Holt R, Evans C, Gocayne J, Amanatides P, Scherer S, Li P, Hoskins R, Galle R *et al.*: **The genome sequence of *Drosophila melanogaster***. *Science* 2000, **287**:2185-2195.

3. Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, Fortini ME, Li PW, Apweiler R, Fleischmann W *et al.*: **Comparative genomics of the eukaryotes**. *Science* 2000, **287**:2204-2215.

4. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JMC, Wides R *et al.*: **The genome sequence of the malaria mosquito *Anopheles gambiae***. *Science* 2002, **298** 129-+.

5. Zdobnov EM, von Mering C, Letunic I, Torrents D, Suyama M, Copley RR, Christophides GK, Thomasova D, Holt RA, Subramanian GM *et al.*: **Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster***. *Science* 2002, **298**:149-159.

6. Xia Q, Zhou Z, Lu C, Cheng D, Dai F, Li B, Zhao P, Zha X, Cheng T, Chai C *et al.*: **A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*)**. *Science* 2004, **306**:1937-1940.

7. International Silkworm Genome Consortium: **The genome of a lepidopteran model insect, the silkworm *Bombyx mori***. *Insect Biochem Mol Biol* 2008, **38**:1036-1045.

8. Richards S, Liu Y, Bettencourt BR, Hradecky P, Letovsky S, Nielsen R, Thornton K, Hubisz MJ, Chen R, Meisel RP *et al.*: **Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution**. *Genome Res* 2005, **15**:1-18.

9. Weinstock GM, Robinson GE, Gibbs RA, Worley KC, Evans JD, Maleszka R, Robertson HM, Weaver DB, Beye M, Bork P *et al.*: **Insights into social insects from the genome of the honeybee *Apis mellifera***. *Nature* 2006, **443**:931-949.

10. Elsik C, Mackey A, Reese J, Milshina N, Roos D, Weinstock G: **Creating a honey bee consensus gene set**. *Genome Biol* 2007, **8**:R13.

11. Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu ZJ, Loftus B, Xi ZY, Megy K, Grabherr M *et al.*: **Genome sequence of *Aedes aegypti*, a major arbovirus vector**. *Science* 2007, **316**:1718-1723.

12. Waterhouse RM, Wyder S, Zdobnov EM: **The *Aedes aegypti* genome: a comparative perspective**. *Insect Mol Biol* 2008, **17**:1-8.

13. Richards S, Gibbs RA, Weinstock GM, Brown SJ, Denell R, Beeman RW, Gibbs R, Bucher G, Friedrich M, Grimmelikhuijzen CJP *et al.*: **The genome of the model beetle and pest *Tribolium castaneum***. *Nature* 2008, **452**:949-955.

14. Zdobnov EM, Bork P: **Quantification of insect genome divergence**. *Trends Genet* 2007, **23**:16-20.

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

16. Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN *et al.*: **Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures**. *Nature* 2007, **450**:219-232.

17. •• Elsik CG, Worley KC, Bennett AK, Beye M, Camara F, Childers CP, de Graaf DC, Debyser G, Deng JX, Devreese B *et al.*: **Finding the missing honey bee genes: lessons learned from a genome upgrade**. *BMC Genomics* 2014:15.

The initial sequencing and annotation of the honey bee genome identified a conservative gene set of only 10,157 genes. In this study, a comprehensive re-annotation strategy that applied multiple new methods and incorporated several new lines of supporting evidence raised the total protein-coding gene count to just over 15,000, with support for 92–97% of genes. This highlights the importance of revisiting initial genome annotations to improve gene set quality and completeness, as well as to discover new genes.

18. Lin MF, Carlson JW, Crosby MA, Matthews BB, Yu C, Park S, Wan KH, Schroeder AJ, Gramates LS, St Pierre SE *et al.*: **Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using 12 fly genomes**. *Genome Res* 2007, **17**:1823-1836.

19. Werren JH, Richards S, Desjardins CA, Niehuis O, Gadau J, Colbourne JK, Beukeboom LW, Desplan C, Elsik CG, Grimmelikhuijzen CJP *et al.*: **Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species**. *Science* 2010, **327**:343-348.

20. International Aphid Genomics Consortium: **Genome sequence of the pea aphid *Acyrthosiphon pisum***. *PLoS Biol* 2010, **8**:e1000313.

21. Kirkness EF, Haas BJ, Sun W, Braig HR, Perotti MA, Clark JM, Lee SH, Robertson HM, Kennedy RC, Elhaik E *et al.*: **Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle**. *Proc Natl Acad Sci USA* 2010, **107**:12168-12173.

22. Bonasio R, Zhang G, Ye C, Mutti N, Fang X, Qin N, Donahue G, Yang P, Li Q, Li C *et al.*: **Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator***. *Science* 2010, **329**:1068-1071.

23. •• Simola DF, Wissler L, Donahue G, Waterhouse RM, Helmkampf M, Roux J, Nygaard S, Glastad KM, Hagen DE, Viljakainen L *et al.*: **Social insect genomes exhibit dramatic evolution in gene composition and regulation while preserving regulatory features linked to sociality**. *Genome Res* 2013, **23**:1235-1247.
This study identified a remarkable abundance of taxonomically-restricted genes from seven ant genomes, as well as a higher rate in the emergence of such genes in ants compared to flies, suggesting links between the birth of lineage-specific genes and adaptations to eusociality.

24. Arensburger P, Megy K, Waterhouse RM, Abrudan J, Amedeo P, Antelo B, Bartholomay L, Bidwell S, Caler E, Camara F *et al.*: **Sequencing of *Culex quinquefasciatus* establishes a platform for mosquito comparative genomics**. *Science* 2010, **330**:86-88.

25. Suen G, Teiling C, Li L, Holt C, Abouheif E, Bornberg-Bauer E, Bouffard P, Caldera EJ, Cash E, Cavanaugh A *et al.*: **The genome sequence of the leaf-cutter ant *Atta cephalotes* reveals insights into its obligate symbiotic lifestyle**. *PLoS Genet* 2011, **7**:e1002007.

26. Smith CD, Zimin A, Holt C, Abouheif E, Benton R, Cash E, Croset V, Currie CR, Elhaik E, Elsik CG *et al.*: **Draft genome of the globally widespread and invasive Argentine ant (*Linepithema humile*)**. *Proc Natl Acad Sci USA* 2011, **108**:5673-5678.

27. Smith CR, Smith CD, Robertson HM, Helmkampf M, Zimin A, Yandell M, Holt C, Hu H, Abouheif E, Benton R *et al.*: **Draft genome of the red harvester ant *Pogonomyrmex barbatus***. *Proc Natl Acad Sci USA* 2011, **108**:5667-5672.

28. Wurm Y, Wang J, Riba-Grognuz O, Corona M, Nygaard S, Hunt BG, Ingram KK, Falquet L, Nipitwattanaphon M, Gotzek D *et al.*: **The genome of the fire ant *Solenopsis invicta***. *Proc Natl Acad Sci USA* 2011, **108**:5679-5684.

29. Nygaard S, Zhang G, Schiøtt M, Li C, Wurm Y, Hu H, Zhou J, Ji L, Qiu F, Rasmussen M *et al.*: **The genome of the leaf-cutting ant *Acromyrmex echinatior* suggests key adaptations to advanced social life and fungus farming**. *Genome Res* 2011, **21**:1339-1348.

30. Zhan S, Merlin C, Boore JL, Reppert SM: **The monarch butterfly genome yields insights into long-distance migration**. *Cell* 2011, **147**:1171-1185.

31. Heliconius Genome Consortium: **Butterfly genome reveals promiscuous exchange of mimicry adaptations among species**. *Nature* 2012, **487**:94-98.

32. You M, Yue Z, He W, Yang X, Yang G, Xie M, Zhan D, Baxter SW, Vasseur L, Gurr GM *et al.*: **A heterozygous moth genome provides insights into herbivory and detoxification**. *Nat Genet* 2013, **45**:220-225.

33. Keeling CI, Yuen MM, Liao NY, Docking TR, Chan SK, Taylor GA, Palmquist DL, Jackman SD, Nguyen A, Li M *et al.*: **Draft genome of the mountain pine beetle, *Dendroctonus ponderosae* Hopkins, a major forest pest**. *Genome Biol* 2013, **14**:R27.

34. Marinotti O, Cerqueira GC, de Almeida LG, Ferro MI, Loreto EL, Zaha A, Teixeira SM, Wespiser AR, Almeida e Silva A, Schlindwein AD *et al.*: **The genome of *Anopheles darlingi*, the main neotropical malaria vector**. *Nucleic Acids Res* 2013, **41**:7387-7400.

35. International Glossina Genome Initiative: **Genome sequence of the tsetse fly (*Glossina morsitans*): vector of African trypanosomiasis**. *Science* 2014, **344**:380-386.

36. Terrapon N, Li C, Robertson HM, Ji L, Meng X, Booth W, Chen Z, Childers CP, Glastad KM, Gokhale K *et al.*: **Molecular traces of alternative social organization in a termite genome**. *Nat Commun* 2014, **5**:3636.

37. Yin C, Liu Y, Liu J, Xiao H, Huang S, Lin Y, Han Z, Li F: **ChiloDB: a genomic and transcriptome database for an important rice insect pest *Chilo suppressalis***. *Database (Oxford)* 2014:2014.

38. Jiang X, Peery A, Hall A, Sharma A, Chen XG, Waterhouse RM, Komissarov A, Riehl MM, Shouche Y, Sharakhova MV *et al.*: **Genome analysis of a major urban malaria vector mosquito, *Anopheles stephensi***. *Genome Biol* 2014, **15**:459.

39. Ahola V, Lehtonen R, Somervuo P, Salmela L, Koskinen P, Rastas P, Välimäki N, Paulin L, Kvist J, Wahlberg N *et al.*: **The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera**. *Nat Commun* 2014, **5**:4737.

40. Gusev O, Suetsugu Y, Cornette R, Kawashima T, Logacheva MD, Kondrashov AS, Penin AA, Hatanaka R, Kikuta S, Shimura S *et al.*: **Comparative genome sequencing reveals genomic signature of extreme desiccation tolerance in the anhydrobiotic midge**. *Nat Commun* 2014, **5**:4784.

41. Scott JG, Warren WC, Beukeboom LW, Bopp D, Clark AG, Giers SD, Hediger M, Jones AK, Kasai S, Leichter CA *et al.*: **Genome of the house fly, *Musca domestica* L., a global vector of diseases with adaptations to a septic environment**. *Genome Biol* 2014, **15**:466.

42. Neafsey DE, Christophides GK, Collins FH, Emrich SJ, Fontaine MC, Gelbart W, Hahn MW, Howell PI, Kafatos FC, Lawson D *et al.*: **The evolution of the *Anopheles* 16 genomes project**. *G3-Genes Genom Genet* 2013, **3**:1191-1194.

43. •• Neafsey DE, Waterhouse RM, Abai MR, Aganezov SS, Alekseyev MA, Allen JE, Amon J, Arcà B, Arensburger P, Artemov G *et al.*: **Highly evolvable malaria vectors: the genomes of 16 *Anopheles* mosquitoes**. *Science* 2015, **347**:1258522.
This study revealed a much faster rate of gene turnover in anopheline mosquitoes compared with drosophilid fruit flies, highlighting the importance of gene gains and losses in shaping the gene repertoires of extant insect species, even within a single genus.

44. Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, Sharakhov IV, Jiang X, Hall AB, Catteruccia F, Kakani E *et al.*: **Extensive introgression in a malaria vector species complex revealed by phylogenomics**. *Science* 2015, **347**:1258524.

46. Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG *et al.*: **Phylogenomics resolves the timing and pattern of insect evolution**. *Science* 2014, **346**:763-767.

47. Robinson GE, Hackett KJ, Purcell-Miramontes M, Brown SJ, Evans JD, Goldsmith MR, Lawson D, Okamuro J, Robertson HM, Schneider DJ: **Creating a buzz about insect genomes**. *Science* 2011, **331**:1386.

48. • i5K Consortium: **The i5K initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment**. *J Hered* 2013, **104**:595-600.
The 5000 arthropod genomes initiative (*i*5K) is a coordinated effort by scientists to prioritize and promote arthropod genome sequencing projects that will extend species sampling and contribute greatly to the growing catalogue of the insect protein-coding gene repertoire.

49. Sonnhammer EL, Gabaldón T, Sousa da Silva AW, Martin M, Robinson-Rechavi M, Boeckmann B, Thomas PD, Dessimoz C, consortium QfO: **Big data and other challenges in the quest for orthologs**. *Bioinformatics* 2014, **30**:2993-2998.

50. • Kriventseva EV, Tegenfeldt F, Petty TJ, Waterhouse RM, Simão FA, Pozdnyakov IA, Ioannidis P, Zdobnov EM: **OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software**. *Nucleic Acids Res* 2015, **43(Database issue)**:D250-D256.
With a total of 80 insect species, the latest release of the OrthoDB hierarchical catalogue of orthologues represents the most comprehensive resource for insect orthology data. OrthoDB presents both functional and evolutionary traits of orthologues, facilitating evolutionarily-informed hypotheses about the functions of genes from newly-sequenced genomes.

51. •• Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva EV: **OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs**. *Nucleic Acids Res* 2013, **41**:D358-D365.

Sets of Benchmarking Universal Single-Copy Orthologues (BUSCOs) represent a conserved core of the repertoire of genes universally-maintained across insects. Quantifying the presence of BUSCOs in newly-sequenced genome assemblies and their annotated gene sets provides an intuitive measure of completeness in terms of expected gene content.

52. Waterhouse RM, Zdobnov EM, Tegenfeldt F, Li J, Kriventseva EV: **OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011**. *Nucleic Acids Res* 2011, **39**:D283-D288.

53. Kriventseva EV, Rahman N, Espinosa O, Zdobnov EM: **OrthoDB: the hierarchical catalog of eukaryotic orthologs**. *Nucleic Acids Res* 2008, **36**:D271-D275.

54. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM: **BUSCOs: assessing genome assembly and annotation completeness with single-copy orthologs**. Submitted for publication.

55. Wyder S, Kriventseva EV, Schroder R, Kadowaki T, Zdobnov EM: **Quantification of ortholog losses in insects and vertebrates**. *Genome Biol* 2007:8.

56. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TC: **More than just orphans: are taxonomically-restricted genes important in evolution?** *Trends Genet* 2009, **25**:404-413.

57. Palmieri N, Kosiol C, Schlötterer C: **The life cycle of *Drosophila* orphan genes**. *Elife* 2014, **3**:e01311.

58. •• Zhao C, Escalante LN, Chen H, Benatti TR, Qu J, Chellapilla S, Waterhouse RM, Wheeler D, Andersson MN, Bao R *et al.*: **A massive expansion of effector genes underlies gall-formation in the wheat pest Mayetiola destructor**. *Curr Biol* 2015, **25**:1-8.

Newly-sequenced insect genomes reveal subsets of unique genes with no homology to any other insect genes. This study of the genome of the hessian fly, Mayetiola destructor, not only uncovered a large number of taxonomically-restricted genes, but also identified some of them as key components of the larval salivary gland excretions that modulate host plant gene expression.

59. Waterhouse RM, Zdobnov EM, Kriventseva EV: **Correlating traits of gene retention, sequence divergence, duplicability and essentiality in vertebrates, arthropods, and fungi**. *Genome Biol Evol* 2011, **3**:75-86.