The Genomic Basis of Arthropod Diversity

Gregg W.C. Thomas[1], Elias Dohmen[2], Daniel S.T. Hughes[3,a], Shwetha C. Murali[3,b], Monica Poelchau[4], Karl Glastad[5,c], Clare A. Anstead[6], Nadia A. Ayoub[7], Phillip Batterham[8], Michelle Bellair[3,d], Gretta J. Binford[9], Hsu Chao[3], Yolanda H. Chen[10], Christopher Childers[4], Huyen Dinh[3], HarshaVardhan Doddapaneni[3], Jian J. Duan[11], Shannon Dugan[3], Lauren A. Esposito[12], Markus Friedrich[13], Jessica Garb[14], Robin B. Gasser[6], Michael A.D. Goodisman[5], Dawn E. Gundersen-Rindal[15], Yi Han[3], Alfred M. Handler[16], Masatsugu Hatakeyama[17], Lars Hering[18], Wayne B. Hunter[19], Panagiotis Ioannidis[20, e], Joy C. Jayaseelan[3], Divya Kalra[3], Abderrahman Khila[21], Pasi K. Korhonen[6], Carol Eunmi Lee[22], Sandra L. Lee[3], Yiyuan Li[23], Amelia R.I. Lindsey[24,f], Georg Mayer[18], Alistair P. McGregor[25], Duane D. McKenna[26], Bernhard Misof[27], Mala Munidasa[3], Monica Munoz-Torres[28,g], Donna M. Muzny[3], Oliver Niehuis[29], Nkechinyere Osuji-Lacy[3], Subba R. Palli[30], Kristen A. Panfilio[31], Matthias Pechmann[32], Trent Perry[8], Ralph S. Peters[33], Helen C. Poynton[34], Nikola-Michael Prpic[35], Jiaxin Qu[3], Dorith Rotenberg[36], Coby Schal[37], Sean D. Schoville[38], Erin D. Scully[39], Evette Skinner[3], Daniel B. Sloan[40], Richard Stouthamer[24], Michael R. Strand[41], Nikolaus U. Szucsich[42], Asela Wijeratne[26,h], Neil D. Young[6], Eduardo E. Zattara[43], Joshua B. Benoit[44], Evgeny M. Zdobnov[20], Michael E. Pfrender[23], Kevin J. Hackett[45], John H. Werren[46], Kim C. Worley[3], Richard A. Gibbs[3], Ariel D. Chipman[47], Robert M. Waterhouse[48], Erich Bornberg-Bauer[49], Matthew W. Hahn[1], Stephen Richards[3]

[1]Department of Biology and Department of Computer Science, Indiana University, Bloomington, IN, USA, [2]Institute for Evolution and Biodiversity, University of Münster, Münster 48149, Germany and Institute for Bioinformatics and Chemoinformatics, Westphalian University of Applied Sciences, Recklinghausen 45665, Germany, [3]Human Genome Sequencing Center, Department of Human and Molecular Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030 USA, [4]National Agricultural Library, USDA, Beltsville, MD 20705, USA, [5]School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA 30332, [6]Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Parkville, Victoria 3010, Australia, [7]Department of Biology, Washington and Lee University, 204 West Washington Street, Lexington, VA 24450, USA, [8]School of BioSciences Science Faculty, The University of Melborne, Victoria 3010, Australia, [9]Department of Biology, Lewis & Clark College, Portland, OR 97219 USA, [10]Department of Plant and Soil Sciences, University of Vermont, Burlington, USA, [11]Beneficial Insects Introduction Research Unit, United States Department of Agriculture, Agricultural Research Service, Newark, DE, USA, [12]Institute for Biodiversity Science and Sustainability, California Academy of Sciences, 55 Music Concourse Drive, San Francisco, CA 94118, USA, [13]Department of Biological Sciences, Wayne State University , Detroit Michigan 48202 United States, [14]Department of Biological Sciences, University of Massachusetts Lowell, 198 Riverside Street, Lowell, MA 01854, [15]USDA-ARS Invasive Insect Biocontrol and Behavior Laboratory, Beltsville, Maryland, United States of America, [16]USDA-ARS, Center for Medical, Agricultural, and Veterinary Entomology, 1700 S.W. 23rd Drive, Gainesville, FL 32608, USA, [17]Division of Insect Sciences, National Institute of Agrobiological Sciences, Owashi, Tsukuba 305-8634, Japan, [18]Department of Zoology, Institute of Biology, University of Kassel, 34132 Kassel, Germany, [19]USDA ARS, U. S. Horticultural Research Laboratory, Ft. Pierce, FL 34945, USA, [20]Department of Genetic Medicine and Development and Swiss Institute of Bioinformatics, University of Geneva, Geneva 1211, Switzerland, [21]Université de Lyon, Institut de Génomique Fonctionnelle de Lyon, CNRS UMR 5242, Ecole Normale Supérieure de Lyon, Université Claude Bernard Lyon 1, 46 allée d'Italie, 69364 Lyon, France., [22]Department of Integrative Biology, University of Wisconsin, Madison, WI 53706, USA, [23]Department of Biological Sciences, University of Notre Dame, 109B Galvin Life Sciences, Notre Dame, IN 46556 USA, [24]Department of Entomology, University of California Riverside, Riverside, CA, USA, [25]Department of Biological and Medical Sciences, Oxford Brookes University, Gipsy Lane, Oxford OX3 0BP, UK, [26]Department of Biological Sciences, University of Memphis, 3700 Walker Ave., Memphis, TN 38152, USA, [27]Center for Molecular Biodiversity Research,

Zoological Research Museum Alexander Koenig, Bonn, Germany, [28]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, USA, [29]Evolutionary Biology and Ecology, Institute of Biology I (Zoology), Albert Ludwig University of Freiburg, 79104 Freiburg (Brsg.), Germany, [30]Department of Entomology, University of Kentucky, Lexington, Kentucky 40546, USA, [31]School of Life Sciences, University of Warwick, Gibbet Hill Campus, Coventry CV4 7AL, UK, [32]Cologne Biocenter, Zoological Institute, Department of Developmental Biology, University of Cologne, 50674 Cologne, Germany, [33]Centre of Taxonomy and Evolutionary Research, Arthropoda Department, Zoological Research Museum Alexander Koenig, Bonn, Germany, [34]School for the Environment, University of Massachusetts Boston, Boston, MA 02125 United States, [35]Johann-Friedrich-Blumenbach-Institut für Zoologie und Anthropologie, Abteilung für Entwicklungsbiologie, Georg-August-Universität Göttingen, Göttingen, Germany and Göttingen Center for Molecular Biosciences (GZMB), Georg-August-Universität Göttingen, Göttingen, Germany, [36]Department of Entomology and Plant Pathology, North Carolina State University, Raleigh, NC 27606 United States, [37]Department of Entomology and W.M. Keck Center for Behavioral Biology, North Carolina State University, Raleigh, North Carolina 27695, USA, [38]Department of Entomology, University of Wisconsin-Madison, Madison, USA, [39]Stored Product Insect and Engineering Research Unit, USDA-ARS Center for Grain and Animal Health Research, Manhattan, KS 66502, [40]Department of Biology, Colorado State University, Ft. Collins CO, USA, [41]Department of Entomology, University of Georgia, Athens, Georgia, United States of America, [42]Natural History Museum Vienna, Burgring 7, 1010 Vienna, Austria, [43]INIBIOMA, Univ. Nacional del Comahue – CONICET, Bariloche, Argentina, [44]Department of Biological Sciences, University of Cincinnati, Cincinnati, Ohio 45221, USA, [45]Crop Production and Protection, U.S. Department of Agriculture-Agricultural Research Service, Beltsville, MD 20705, USA, [46]Department of Biology, University of Rochester, Rochester, New York 14627, USA, [47]Department of Ecology, Evolution and Behavior, The Alexander Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, Edmond J. Safra Campus, Givat Ram 91904, Jerusalem, Israel, [48]Department of Ecology & Evolution and Swiss Institute of Bioinformatics, University of Lausanne, 1015 Lausanne, Switzerland, [49]Institute for Evolution and Biodiversity, University of Münster, Münster 48149, [a] Current address: Institute for Genomic Medicine, Columbia University, New York, NY 10032 United States, [b] Current address: Howard Hughes Medical Institute, Department of Genome Sciences, University of Washington, Seattle, WA 98195, [c] Current address: Penn Epigenetics Institute, Department of Cell and Developmental Biology, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania 19104, USA, [d] Current Address: CooperGenomics Houston TX, USA, [e]Current address: Foundation for Research and Technology Hellas, Institute of Molecular Biology and Biotechnology, 70013 Vassilika Vouton, Heraklion, Greece. [f] Current address: Department of Biology, Indiana University, Bloomington, IN, USA. [g] Current address: Phoenix Bioinformatics. 39221 Paseo Padre Parkway, Ste. J. Fremont, CA 94538, USA, [h]Current address: Arkansas Biosciences Institute, Arkansas State University, Jonesboro, AR, USA,

Summary

Arthropods comprise the largest and most diverse phylum on Earth and play vital roles in nearly every ecosystem. Their diversity stems in part from variations on a conserved body plan, resulting from and recorded in adaptive changes in the genome. Dissection of the genomic record of sequence change enables broad questions regarding genome evolution to be addressed, even across hyper-diverse taxa within arthropods. Using 76 whole genome sequences representing 21 orders spanning more than 500 million years of arthropod evolution, we document changes in gene and protein domain content and provide temporal and phylogenetic context for interpreting these innovations. We identify many novel gene families that arose early in the evolution of arthropods and during the diversification of insects into modern orders. We reveal unexpected variation in patterns of DNA methylation across arthropods and examples of gene family and protein domain evolution coincident with the appearance of notable phenotypic and physiological adaptations such as flight, metamorphosis, sociality and chemoperception. These analyses demonstrate how large-scale comparative genomics can provide broad new insights into the genotype to phenotype map and generate testable hypotheses about the evolution of animal diversity.

Introduction

Arthropods (hexapods, crustaceans, myriapods and chelicerates) constitute the most species-rich and diverse phylum on Earth, having adapted, innovated, and expanded into all major habitats within all major ecosystems. They are found as detritivores, herbivores, carnivores, and parasites. As major components of the world's biomass, their diversity and ubiquity lead naturally to significant interactions with humanity, as disease vectors, crop pests and pollinators, food sources, and synanthropes. Despite their diversity, arthropods share a deeply conserved and highly modular body plan. They are bilaterally symmetrical, with serially repeated segments along the anterior-posterior axis. Many segments bear paired appendages, which can take the form of jointed legs, gills, feeding appendages or antennae. Many arthropods have evolved specialized secretions such as venom or silk, extruded from dedicated structures that further capitalize on this segmental modularity. Arthropods also have a hard exoskeleton, composed mostly of chitin, which molts as the animal grows in size. One group of arthropods, the winged insects (Pterygota), took to the skies, bearing up to two pairs of wings as outgrowths of that exoskeleton.

The extraordinary diversity of arthropods is manifested in a series of genomic changes and innovations selected for throughout their evolutionary history. However, linking this phenotypic diversity to underlying genomic changes remains an elusive challenge. The major transitions in arthropod evolution include the differential grouping of body segments into morphological units with a common function (*e.g.*, head, thorax, and abdomen in the Hexapoda) in different taxa, the independent and parallel process of terrestrialization following freshwater colonizations from marine origins in several lineages[1,2], the emergence of active flight in insects[3,4], and the evolution of insect metamorphosis[5]. Multiple genomic mechanisms might be responsible for such innovations, but the underlying molecular transitions have not been explored on a broad

phylogenomic scale. Tracing these transitions at the genomic-level requires mapping whole genome data to a robust phylogenetic framework. Here we explore the evolution of arthropod genomes using a phylogeny-mapped genomic resource of 76 species representing the breath of arthropod diversity.

An Arthropod Evolution Resource

As a pilot project for the i5K initiative to sequence 5,000 arthropod genomes[6], we sequenced and annotated the genomes of 28 arthropod species. These include a combination of species of agricultural or ecological importance, emerging laboratory models, and species occupying key positions in the arthropod phylogeny. We combined these newly sequenced genomes with those of 48 previously sequenced arthropods creating a dataset comprising 76 species representing the four extant arthropod subphyla and spanning 21 taxonomic orders. Using the OrthoDB gene orthology database[7], we annotated 38,195 orthologous protein groups (orthogroups/gene families) among all 76 species (Fig. 1). Based on single-copy orthogroups within and between orders, we then built a phylogeny of all major arthropod lineages (Fig. 2). This phylogeny is mostly consistent with previous arthropod phylogenies[8,9], with the exception being that we recover a monophyletic Crustacea with currently available species, rather than the generally accepted paraphyletic nature of Crustacea in respect to Hexapoda (see Methods). We reconstructed the gene content and protein domain arrangements for all 38,195 orthogroups in each of the lineages for the 76 species in the arthropod phylogeny. This resource (available at https://i5k.gitlab.io/ArthroFam/ and Table S10) forms the basis for the analyses detailed below and is an unprecedented tool for identifying and tracking genomic changes over arthropod evolutionary history.

Genomic Change Throughout Arthropod History

Evolutionary innovation can result from diverse genomic changes. New genes can arise either by duplication or, less frequently, by *de novo* gene evolution[10]. Genes can also be lost over time, constituting an underappreciated mechanism of evolution[11,12]. Protein domains are the basis of reusable modules for protein innovation and the rearrangement of domains to form new combinations plays an important role in molecular innovation[13]. Together, gene family expansions and contractions and protein domain rearrangements, may coincide with phenotypic innovations in arthropods. We therefore searched for signatures of such events corresponding with pivotal phenotypic shifts in the arthropod phylogeny.

Using ancestral reconstructions of gene counts (see Methods), we tracked gene family expansions and losses across the arthropod phylogeny. In all, we inferred 181,157 gene family expansions and 87,505 gene family contractions. 68,430 gene families were inferred to have gone extinct in at least one lineage, and 9,115 families emerged in different groups. The most dynamically-changing gene families encode proteins involved in functions of xenobiotic defense (cytochrome P450s, sulfotransferases), digestion (peptidases), chitin exoskeleton structure and metabolism, multiple zinc finger transcription factor types, HSP20 domain stress response, fatty acid metabolism, chemosensation, and ecdysteroid (molting hormone) metabolism (Table S14). Using the estimates of where in the phylogeny these events occurred, we can infer characteristics of

ancestral arthropods. For example, we identified 9,601 genes in the last insect common ancestor (LICA) and estimate ~14,700 LICA genes after correcting for unobserved gene extinctions. We reconstructed similar numbers for the ancestor of each of the six well-represented arthropod taxa in our sample (Table 1). Of the 9,601 genes present in LICA, we identified 147 emergent gene families (*i.e.*, those with no close homologs in other clades) which appeared concurrently with the evolution of insects (Figs. 3A & 1 node 62, Table S15). GO-term analysis of these 147 gene families recovered multiple key functions, including cuticle and cuticle development (suggesting changes in exoskeleton development), visual learning and behavior, pheromone and odorant binding (suggesting the ability to sense in terrestrial/aerial environments rather than in water), ion transport, neuronal activity, larval behavior, imaginal disc development, and wing morphogenesis. These emergent gene families likely allowed insects to undergo substantial diversification by expanding chemical sensing, such as an expansion in odorant binding to locate novel food sources and fine-tune species self-recognition[14-16]. Others, such as cuticle proteins underlying differences in exoskeleton structure, may enable optimized cuticle properties for diverse environmental niches or life history stages[17]. In contrast, the data reveal only ten gene families that arose along the ancestral lineage of the Holometabola (Fig. 3B, Table S16), implying that genes and processes required for the transition to holometabolous development, such as imaginal disc development, were already present in the hemimetabolous ancestors. This is consistent with Truman and Riddiford's model that the holometabolous insect larva corresponds to a late embryonic state in hemimetabolous insects[18].

We identified numerous genes that emerged in specific orders of insects. Strikingly, we found 1,038 emergent gene families in the first ancestral Lepidoptera node (Fig. 3F). This node has by far the most emergent gene families, with the next highest being the node leading to the bumble bee genus *Bombus* with 860 emergent gene families. A recent analysis suggested whole genome duplication in the Lepidoptera lineage[19], consistent with the high number of emergent gene families. Emergent lepidopteran gene families show enrichment for functional categories such as peptidases and odorant binding. Among the other insect orders, we find 227 emergent families in the node leading to the Hymenoptera, 205 in Coleoptera, and 156 in Diptera.

Similarly, we reconstructed the protein domain arrangements for all nodes of the arthropod phylogeny, that is, the permutations in protein domain type per (multi-domain) gene. In total, we can explain the underlying events for more than 40,000 domain arrangement changes within the arthropods. The majority of domain arrangements (48% of all observable events) were formed by a fusion of two ancestral arrangements, while the fission of an existing arrangement into two new arrangements accounts for 14% of all changes. Interestingly, 37% of observed changes can be explained by losses (either as part of an arrangement (14%) or the complete loss of a domain in a proteome (23%)), while emergence of a novel protein domain is a very rare event, comprising only 1% of total events.

We observe high concordance between rates of gene family dynamics and protein domain rearrangement (Figs. 4 & S27). In some cases, we find specific examples of overlap between gene family and protein domain evolution. For example, spiders have

the characteristic ability to spin silk and are venomous. Correspondingly, we identify ten gene families associated with venom or silk production that are rapidly expanding within the Araneae (spiders) clade (Table S17). In parallel, we find a high rate of newly emerged protein domains in the subphylum Chelicerata, including a large number within the Araneae that are associated with venom and silk production. For example, 'spider silk protein 1' (Pfam ID: PF16763), 'Major ampullate spidroin 1 and 2' (PF11260), 'Tubuliform egg casing silk strands structural domain' (PF12042) and 'Toxin with inhibitor cystine knot ICK or Knottin scaffold' (PF10530) are all domains that emerged within the spider clade. Venom domains also emerged in other venomous chelicerates, such as the bark scorpion, *Centruroides sculpturatus*.

We identified other gene family changes that may underlie unique phenotypic transitions. The evolution of eusociality among three groups in our study, bees and ants (both Hymenoptera), and termites (Blattodea), requires these insects to be able to recognize other individuals of their colony (such as nest mates of the same or different caste), or invading individuals (predators, slave-makers and hosts) for effective coordination. We find 41 functional terms enriched for gene family changes in all three groups, with multiple gene family gains related to olfactory reception and odorant binding (Table S18) in agreement with previous chemoreceptor studies of these species[20,21].

Finally, we observe species-specific gene family expansions that suggest biological functions under selection. The German cockroach *Blattella germanica*, a pervasive tenant in human dwellings across the world, has experienced the highest number of rapidly evolving gene families among the arthropods studied here, in agreement with a previously reported major expansion of chemosensory genes[22]. We also find the largest number of domain rearrangement events at the terminal branch of *B. germanica*. The impressive capability of this cockroach to quickly adapt to changing environments could be linked to these numerous and rapid evolutionary changes at the genomic level and warrants more detailed investigation.

Evolutionary Rates Within Arthropod History

The rate of genomic change can reflect key events during evolution along a phylogenic lineage. Faster rates might imply small population sizes or strong selective pressure, possibly indicative of rapid adaptive radiations, and slower rates may indicate stasis. Studying rates of change requires a time-calibrated phylogeny. For this, we used 22 fossil calibration points[8,23] and obtained branch lengths for our phylogeny in millions of years (My) (Fig. 2) that are very similar to those obtained by Misof et al.[8] and Rota-Stabelli et al.[9].

We examined the rates of three types of mutational events: (i) amino acid substitutions, (ii) gene duplications and gene losses, and (iii) protein domain rearrangements, emergence, and loss. While clearly not changing in a clock-like manner, all types of events have a strikingly small amount of variation in rate among the investigated species (Fig. 4). We estimate an average amino acid substitution rate of 2.54 x 10^{-3} substitutions per site per My with a standard deviation of 1.11 x 10^{-3}. The slowest rate is found in the branch leading to the insect order Blattodea (cockroaches and

termites), while the fastest rates are found along the short branches during the early diversification of Holometabola, suggesting a period of rapid evolution, a pattern similar to that found for amino acid sequence evolution during the Cambrian explosion[24]. Other branches with elevated amino acid divergence rates include those leading to the Acarina (mites), and the overall path to the Diptera (flies).

Rates of gene duplication and loss show remarkably little variation, both across the tree and within the six multi-species orders (Table S12). Overall, we estimate an average rate of 43.0 gains/losses per My, but with a high standard deviation of 59.0 that is driven by a few lineages with greatly accelerated rates. Specifically, the terminal branches leading to the leafcutter ants *Atta cephalotes* and *Acromyrmex echinatior* have exceptionally high gene gain/loss rates of 266 and 277 per My, an order of magnitude higher than average, as previously reported[25]. Removing these and two other outliers, the average becomes 27.2 gains/losses per My (SD 19.7). Interestingly, the high gain/loss rates observed in these ants, in contrast to other arthropods, are not due to large gene content change in a small number of gene families. They are instead due mostly to single gene gains or losses in a large number of gene families.

Regarding protein domain rearrangements, which mainly arise from duplication, fusion and terminal losses of domains[26], we estimate an average rate of 5.27 events per My, approximately eight-fold lower than the rate of gene gain/loss. Interestingly, we discovered a strong correlation between rates of gene gain/loss and domain rearrangement (Figs. 4 & S27) even though these processes follow largely from different underlying genetic events. For example, terminal branches within the Hymenoptera have an accelerated rate of domain rearrangement, which coincides with the increased rate of gene gains and losses observed along those branches.

Although amino acid substitutions are mediated by distinctly different processes than gene gain, loss, and rearrangement, it is possible that similar life-history traits or similar selective regimes could have led to parallel acceleration or deceleration of these different types of genomic change. However, our examination of rates of variation in all types of change among lineages found no correlation between variation in amino acid substitution rates and rates of gene gain/loss (Figs. 4 & S27). Branches with accelerated rates of amino acid substitution, such as the lineage leading to the most recent common ancestor of the insect superorder Holometabola, do not show corresponding increases in gene gain/loss rates. Similarly, the hymenopteran lineages displaying the fastest rate of gene gain/loss in our analysis do not display higher rates of amino acid substitutions.

We observe accelerated rates of amino acid substitutions in several deep branches dating to 300-600 Mya. Descendants from those branches display accelerated rates of gene gain/loss and domain rearrangement. This suggests that, in some lineages, accelerated amino acid substitution rates may lay the foundation for subsequent genomic changes. Interestingly, we did not record a corresponding increase in gene gain/loss or domain rearrangement with the increased amino acid substitution rates, instead, we observed a generally steady rate. In Hymenoptera specifically, we again observe accelerated rates of amino acid substitutions in lineages preceding increased rates of gene gain/loss within the last 100 My.

Methylation Signals in Arthropod Genomes

In addition to records of gene content and protein domain change in the genome, we also searched for signals of gene regulation. DNA methylation is an important epigenetic modification that affects gene regulation in widely divergent organisms[27]. We find several independent losses of the DNA methylation machinery across the arthropods (Fig. S7)[28]. This indicates that DNA methylation is not universally necessary for development and that the DNA methyltransfereases in insects may function in ways not previously appreciated[29]. Additionally, putative levels of DNA methylation vary considerably across arthropod species (Figs. S7, S8). Notably, the hemimetabolous insects and non-insect arthropods show higher levels of DNA methylation than the holometabolous insects[28]. Araneae (spiders), in particular, show clear bimodal patterns of methylation (Fig. S8); where some genes are highly methylated and others are not. This pattern is also found in some holometabolous insects, suggesting that the division of genes into methylated and unmethylated categories is a relatively ancient trait in Arthropoda, although many species have since lost this clear distinction. Finally, some taxa, particularly in Hymenoptera, show higher levels of CpG di-nucleotides than expected by chance alone, which may be a signal of strong effects of gene conversion in the genome[30].

Concluding remarks and future directions

The i5K pilot initiative has assembled an unparalleled genomic dataset for research on arthropods and conducted a detailed phylogenetic analysis of evolutionary changes at the genomic level within this diverse and fascinating phylum. The combined research output of this species-level i5K work has been substantial and wide-ranging, addressing pests of agricultural crops[31,32] and animals[33], urban[20,34] and forest[35] pests, biocontrol species[36], along with developmental models[17,37,38], indicators of water quality and models for toxicology[14,39].

Here, in contrast, we take a broad overview generating a comparative genomics resource for a phylum with an evolutionary history of over 500 million years. Our analysis identifies multiple broad patterns such as the very small number of novel protein domains, and a surprising lack of variation in the rates of some types of genomic change. We pinpoint the origin of specific gene families and trace key transitions during which specific gene families or protein domains have undergone rapid radiations or contractions. Nonetheless, drawing functional biological conclusions from these data is not straightforward. In some cases, the link between specific gene families and their biological function is clear. This is true for genes related to specific physiological function (*e.g.,* olfaction), or to the production of specific compounds (*e.g.,* silk or venom). However, for many gene families, there is no known function, highlighting the need for functional genomic studies. This is true both for emergent gene families, such as those identified in the Lepidoptera which cannot be studied in the dipteran *Drosophila* model, and for rapidly evolving and diverging gene families.

A key result is that major morphological transitions cannot be assigned to particular genomic changes. Indeed, there is no genomic signature for morphological

evolution, and no novel genes or significant gene expansions or domain rearrangements can be linked to novel morphological structures or to major events in body plan evolution, such as the evolution of wings or holometabolous metamorphosis. This supports the emerging understanding that morphology is effected by complex gene networks, which are active mostly during ontogenetic processes[40], rather than by individual "morphology genes". Morphological innovations are often based on modulating the timing and location of expression, rewiring of existing gene networks, and assembling new networks using existing developmental toolkit genes[41].

The advent of affordable and widely transferable genomics opens up many avenues for evolutionary analyses. The genome is both the substrate and record of evolutionary change, and it encodes these changes, but the connection is far from simple. A better understanding of the genotype-phenotype map requires in-depth experimental studies to test hypotheses generated by genomic analyses, such as those presented here. The diversity of arthropods provides unparalleled taxonomic resolution for phenotypic change, which, combined with the experimental tractability of many arthropods, suggests a productive area of future research using and building upon the resource established herein.

Online Methods

Sequencing, assembly and annotation

Twenty-eight arthropod species were sequenced using Illumina short read technology. In total, 126 short read libraries were generated and sequenced to generate 4,883 Gbp of raw sequence (Table S2). For individual species, reads were assembled using AllpathsLG[42,43] followed by refinements employing Atlas-Link[44] and Gapfill[45]. Version 1.0 assemblies had minimum, mean, and maximum scaffold N50 lengths of 13.8 kbp, 1.0 Mbp and 7.1 Mbp (Table S3). Following re-assembly and collapsing of unassembled haplotypes using Redundans[46], version 2.0. assemblies had minimum, mean and maximum contig N50 lengths of 11.1, 166.2 and 857.0 kbp with a mean scaffold N50 lengths of 619 kbp (Table S3).

To support the annotation, RNAseq data were generated from 25 species for which no data were available (Table S4). A MAKER[47] based automated annotation pipeline was applied to the 1.0 assembly of each species with species-specific input RNAseq data and alignment data from a non-redundant metazoan protein sequence set containing all available arthropod protein sequences (see Supplementary methods). This pipeline was applied to 28 species with annotatable genome assemblies generating 533,636 gene models, with minimum, mean, and maximum gene model numbers of 10,901, 19,058, and 33,019 per species (Table S5, see Table S6 for completeness statistics). Many of these gene models were manually curated using the i5k Workspace@NAL[48]. Given the magnitude of this manual task, the greatest fraction of gene models manually confirmed for a species was 15 %. The analyses presented here were performed on the automatically generated gene models.

Orthology prediction

Orthology delineation is a cornerstone of comparative genomics, offering qualified hypotheses on gene function by identifying "equivalent" genes in different species. We used the OrthoDB[7] (www.orthodb.org) orthology delineation process that is based on the clustering of best reciprocal hits (BRHs) of genes between all pairs of species. Clustering proceeds first by triangulating all BRHs and then subsequently adding in-paralogous groups and singletons to build clusters of orthologous genes. Each of these orthologous groups represent all descendants of a single gene present in the genome of the last common ancestor of all the species considered for clustering[49].

The orthology datasets computed for the analyses of the 28 i5K pilot species, together with existing sequenced and annotated arthropod genomes were compiled from OrthoDB v8[50], which comprises 87 arthropods and an additional 86 other metazoans (including 61 vertebrates). Orthology clustering at OrthoDB included ten of the i5K pilot species (*Anoplophora glabripennis, Athalia rosae, Ceratitis capitata, Cimex lectularius, Ephemera danica, Frankliniella occidentalis, Ladona fulva, Leptinotarsa decemlineata, Orussus abietinus, Trichogramma pretiosum*). The remaining 18 i5K pilot species were subsequently mapped to OrthoDB v8 ortholog groups at several major nodes of the metazoan phylogeny. Orthology mapping proceeds by the same steps as for BRH clustering, but where existing ortholog groups are only permitted to accept new members, *i.e.*, the genes from species being mapped are allowed to join existing groups if the BRH criteria are met. The resulting ortholog groups of clustered and mapped genes were filtered to select all groups with orthologs from at least two species from the full set of 76 arthropods, as well as retaining all orthologs from any of 13 selected outgroup species for a total of 47,281 metazoan groups with orthologs from 89 species. Mapping was also performed for the relevant species at the following nodes of the phylogeny: Arthropoda (38,195 groups, 76 species); Insecta (37,079 groups, 63 species); Endopterygota (34,614 groups, 48 species); Arachnida (8,806 groups, 8 species); Hemiptera (8,692 groups, 7 species); Hymenoptera (21,148 groups, 24 species); Coleoptera (12,365 groups, 6 species); and Diptera (17,701, 14 species). All identified BRHs, protein sequence alignment results, and orthologous group classifications were made available for downstream analyses: http://ezmeta.unige.ch/i5k.

Arthropod phylogeny

We reconstructed the arthropod phylogeny (Fig. 2) using protein sequences from the 76 genomes. Six different phylogenetic reconstruction approaches generated a consistent relationship among the orders (see Supplemental Methods), corresponding with previously inferred arthropod phylogenies[8,9]. However, our analysis placed the Psocodea, represented here by the human body louse, *Pediculus humanus*, with low bootstrap support as a sister to Thysanoptera and Hemiptera, contradicting a recent report placing this order sister to Holometabola[8].

Of the six orders in our dataset represented by multiple species (Figs. S11-16), relationships within the Araneae, Hemiptera, Coleoptera, and Lepidoptera were identical, regardless of the tree building method used. Within the Hymenoptera, the only disagreement between methods concerned the position of the parasitoid wasps within the

Chalcidoidea, with three methods placing *Copidosoma floridanum* as sister to *Nasonia vitripennis* (in agreement with recent phylogenomic research[51]), and three sister to *Trichogramma pretiosum* (Fig. S13). Within the Diptera, we obtained a sister group relationship between the sand fly, *Lutzomyia longipalpis*, and the Culicidae, but this was not a stable topology across methods (Fig. S16).

The most contentious nodes in the phylogeny involve the relationship of crustaceans and hexapods. Here, we recover a monophyletic Crustacea that represent the sister clade to Hexapoda (Fig. 2), in contrast to recent analyses suggesting this group is paraphyletic, with Hexapoda as an in-group, and sister to the remipedes and branchiopods (Cladocera)[52]. However, an extensive phylogenetic investigation (Supplementary Results, Fig. S9) shows that regardless of the inference method used, the relationships among the crustacean and hexapod lineages remain uncertain. Aside from these few discrepancies, nodal support values across the tree were high for all tree building methods used. Even when bootstrap support was < 100 %, all methods still inferred the same topology among the species included. Most importantly, remipedes are missing from our taxon sampling, as are ostracods, pentatomids and mystacocarids.

Divergence time estimation

Phylogenetic branch lengths calibrated in terms of absolute time are required to study rates of evolution and to reconstruct ancestral gene counts. We used a non-parametric method of tree smoothing implemented in the software r8s[53] to estimate these divergence times. Fossil calibrations are required to scale the smoothed tree by absolute time. We relied on Wolfe et al.'s[23] aggregation of deep arthropod fossils with additional recent fossils used by Misof et al.[8] (Table S13). The results indicate that the first split within arthropods (the chelicerate-mandibulate split) occurred ~570 million years ago (mya). We estimate that within the chelicerates, arachnids radiated from a common ancestor ~500 mya. Within the mandibulates, myriapods split from other mandibulates ~570 mya. Crustaceans started radiating ~506 mya, and insects started radiating ~430 mya.

Substitution rate estimation

To estimate substitution rates per year on each lineage of the arthropod phylogeny, we divided the expected number of substitutions (the branch lengths in the unsmoothed tree) by the estimated divergence times (the branch lengths in the smoothed tree) (Fig. 4).

Gene family analysis

With the 38,195 orthogroups and the ultrametric phylogeny we were able to perform the largest gene family analysis of any group of taxa to date. In this analysis we were able to estimate gene turnover rates (λ) for the six multi-species taxonomic orders, to infer ancestral gene counts for each taxonomic family on each node of the tree, and to estimate gene gain/loss rates for each lineage of the arthropod phylogeny. The size of the dataset and the depth of the tree required several methods to be utilized.

Gene turnover rates (λ) for the six multi-species orders were estimated with CAFE 3.0, a likelihood method for gene-family analysis[54]. This version of CAFE is able to estimate the amount of assembly and annotation error (ε) present in the input data using a distribution across the observed gene family counts and a pseudo-likelihood search. CAFE is then able to correct for this error and obtain a more accurate estimate of λ (Table S12). However, with such deep divergence times of some orders, estimates of ε may not be accurate. CAFE has a built-in method to assess significance of changes along a lineage, given an estimated λ, and this method was used to identify rapidly evolving families within each order. We partitioned the full dataset of 38,195 orthogroups for each order such that taxa not in the order were excluded for each family and only families that had genes in a given order were included in the analysis. This led to the counts of gene families presented in Table S11.

For nodes with deeper divergence times across Arthropoda, likelihood methods to reconstruct ancestral gene counts such as CAFE, become inaccurate. Instead, a parsimony method was used to infer these gene counts across all 38,195 orthogroups[55]. Parsimony methods for gene family analysis do not include ways to assess significant changes in gene family size along a lineage. Hence, we performed a simple statistical test procedure for each branch to assess whether a given gene family was changing significantly: under a stochastic birth-death process of gene family evolution, and within a given family, the expected relationship between any node and its direct ancestor is that no change will have occurred. Therefore, we took all differences between nodes and their direct descendants in a family and compared them to a one-to-one linear regression. If any of the points differ from this one-to-one line by more than two standard deviations of the variance within the family, it was considered a significant change and that family is rapidly evolving along that lineage. Rates of gene gain and loss were estimated in a similar fashion to substitution rates. We counted the number of gene families inferred to be changing along each lineage and divided that by the estimated divergence time of that lineage (Fig. 4).

To estimate ancestral gene content (*i.e.*, the number of genes at any given node in the tree), we had to correct for gene losses that are impossible to infer given the present data. To do this, we first regressed the number of genes at each internal node with the split time of that node and noticed the expected negative correlation of gene count and time (Fig. S5) (r^2=0.37; P=4.1 x 10^{-9}). We then took the predicted value at time 0 (present day) as the number of expected genes if no unobserved gene loss occurs along any lineage, and shifted the gene count of each node so that the residuals from the regression matched the residuals of the 0 value.

Protein domain evolution analysis

We annotated the proteomes of all 76 arthropod species and 13 outgroup species with protein domains from the Pfam database (v30)[56]. Thereby, every protein was represented as a domain arrangement, defined by its order of domains in the amino acid sequence. To prevent evaluating different isoforms of proteins as additional rearrangement events, we removed all but the longest isoform. Repeats of a same domain were collapsed to one instance of the domain (A-B-B-B-C → A-B-C), since copy numbers of some repeated

domains can vary strongly even between closely related species[57,58]. To be able to infer all rearrangement events over evolutionary time, we reconstructed the ancestral domain content of all inner nodes in the phylogenetic tree via the DomRates tool (http://domainworld.uni-muenster.de/programs/domrates/) based on a combined parsimony approach (see Supplementary Methods). Six different event types were considered in this study (Fig. S6): fusion, fission, terminal loss/emergence and single domain loss/emergence. For the rate calculation just all arrangement changes were considered that could be explained by exactly one of these event types, while all arrangements were ignored that could not be explained by one of these events in a single step or if multiple events could explain a new arrangement.

References

1 Lozano-Fernandez, J. *et al.* A molecular palaeobiological exploration of arthropod terrestrialization. *Philosophical Transactions of the Royal Society B-Biological Sciences* **371**, 20160133, doi:ARTN 20150133 10.1098/rstb.2015.0133 (2016).

2 Glenner, H., Thomsen, P. F., Hebsgaard, M. B., Sorensen, M. V. & Willerslev, E. Evolution. The origin of insects. *Science* **314**, 1883-1884, doi:10.1126/science.1129844 (2006).

3 Haug, J. T., Haug, C. & Garwood, R. J. Evolution of insect wings and development - new details from Palaeozoic nymphs. *Biological Reviews of the Cambridge Philosophical Society* **91**, 53-69, doi:10.1111/brv.12159 (2015).

4 Medved, V. *et al.* Origin and diversification of wings: Insights from a neopteran insect. *Proc. Natl. Acad. Sci. USA* **112**, 15946-15951, doi:10.1073/pnas.1509517112 (2015).

5 Nel, A. *et al.* The earliest known holometabolous insects. *Nature* **503**, 257-261, doi:10.1038/nature12629 (2013).

6 i5K Consortium. The i5K Initiative: Advancing Arthropod Genomics for Knowledge, Human Health, Agriculture, and the Environment. *The Journal of heredity* **104**, 595-600, doi:10.1093/jhered/est050 (2013).

7 Zdobnov, E. M. *et al.* OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic acids research* **45**, D744-D749, doi:10.1093/nar/gkw1119 (2017).

8 Misof, B. *et al.* Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**, 763-767, doi:10.1126/science.1257570 (2014).

9 Rota-Stabelli, O., Daley, A. C. & Pisani, D. Molecular timetrees reveal a Cambrian colonization of land and a new scenario for ecdysozoan evolution. *Current biology : CB* **23**, 392-398, doi:10.1016/j.cub.2013.01.026 (2013).

10 Santos, M. E., Le Bouquin, A., Crumiere, A. J. J. & Khila, A. Taxon-restricted genes at the origin of a novel trait allowing access to a new environment. *Science* **358**, 386-390, doi:10.1126/science.aan2748 (2017).

11 Albalat, R. & Canestro, C. Evolution by gene loss. *Nature reviews. Genetics* **17**, 379-391, doi:10.1038/nrg.2016.39 (2016).

12 Porter, M. L. & Crandall, K. A. Lost along the way: the significance of evolution in reverse. *Trends in Ecology & Evolution* **18**, 541-547, doi:https://doi.org/10.1016/S0169-5347(03)00244-1 (2003).

13 Lees, J. G., Dawson, N. L., Sillitoe, I. & Orengo, C. A. Functional innovation from changes in protein domains and their combinations. *Curr Opin Struct Biol* **38**, 44-52, doi:10.1016/j.sbi.2016.05.016 (2016).

14 Eyun, S. I. *et al.* Evolutionary History of Chemosensory-Related Gene Families across the Arthropoda. *Molecular biology and evolution* **34**, 1838-1862, doi:10.1093/molbev/msx147 (2017).

15 Arguello, J. R. *et al.* Extensive local adaptation within the chemosensory system following Drosophila melanogaster's global expansion. *Nat Commun* **7**, ncomms11855, doi:10.1038/ncomms11855 (2016).

16 Leal, W. S. Odorant reception in insects: roles of receptors, binding proteins, and degrading enzymes. *Annu Rev Entomol* **58**, 373-391, doi:10.1146/annurev-ento-120811-153635 (2013).
17 Panfilio, e. A. Molecular evolutionary trends and feeding ecology diversification in the Hemiptera, anchored by the milkweed bug genome. *BioRxiv pre-print server* **201731**, doi:10.1101/201731. (2018).
18 Truman, J. W. & Riddiford, L. M. The origins of insect metamorphosis. *Nature* **401**, 447-452, doi:10.1038/46737 (1999).
19 Li, Z. *et al.* Multiple large-scale gene and genome duplications during the evolution of hexapods. *Proceedings of the National Academy of Sciences of the United States of America*, doi:10.1073/pnas.1710791115 (2018).
20 Harrison, M. C. *et al.* Hemimetabolous genomes reveal molecular basis of termite eusociality. *Nat Ecol Evol* **2**, 557-566, doi:10.1038/s41559-017-0459-1 (2018).
21 Zhou, X. *et al.* Chemoreceptor Evolution in Hymenoptera and Its Implications for the Evolution of Eusociality. *Genome biology and evolution* **7**, 2407-2416, doi:10.1093/gbe/evv149 (2015).
22 Robertson, H. M., Baits, R. L., Walden, K. K. O., Wada-Katsumata, A. & Schal, C. Enormous expansion of the chemosensory gene repertoire in the omnivorous German cockroach Blattella germanica. *J Exp Zool B Mol Dev Evol*, doi:10.1002/jez.b.22797 (2018).
23 Wolfe, J. M., Daley, A. C., Legg, D. A. & Edgecombe, G. D. Fossil calibrations for the arthropod Tree of Life. *Earth-Science Reviews* **160**, 43-110, doi:https://doi.org/10.1016/j.earscirev.2016.06.008 (2016).
24 Lee, M. S., Soubrier, J. & Edgecombe, G. D. Rates of phenotypic and genomic evolution during the Cambrian explosion. *Current biology : CB* **23**, 1889-1895, doi:10.1016/j.cub.2013.07.055 (2013).
25 Wissler, L., Gadau, J., Simola, D. F., Helmkampf, M. & Bornberg-Bauer, E. Mechanisms and dynamics of orphan gene emergence in insect genomes. *Genome biology and evolution* **5**, 439-455, doi:10.1093/gbe/evt009 (2013).
26 Moore, A. D., Grath, S., Schuler, A., Huylmans, A. K. & Bornberg-Bauer, E. Quantification and functional analysis of modular protein evolution in a dense phylogenetic tree. *Biochimica et biophysica acta* **1834**, 898-907, doi:10.1016/j.bbapap.2013.01.007 (2013).
27 Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature reviews. Genetics* **13**, 484-492, doi:10.1038/nrg3230 (2012).
28 Provataris, P., Meusemann, K., Niehuis, O., Grath, S. & Misof, B. Signatures of DNA Methylation across Insects Suggest Reduced DNA Methylation Levels in Holometabola. *Genome biology and evolution* **10**, 1185-1197, doi:10.1093/gbe/evy066 (2018).
29 Bewick, A. J., Vogel, K. J., Moore, A. J. & Schmitz, R. J. Evolution of DNA Methylation across Insects. *Molecular biology and evolution* **34**, 654-665, doi:10.1093/molbev/msw264 (2017).
30 Kent, C. F., Minaei, S., Harpur, B. A. & Zayed, A. Recombination is associated with the evolution of genome structure and worker behavior in honey bees. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 18012-18017, doi:10.1073/pnas.1208094109 (2012).
31 Schoville, S. D. *et al.* A model species for agricultural pest genomics: the genome of the Colorado potato beetle, Leptinotarsa decemlineata (Coleoptera: Chrysomelidae). *Sci Rep* **8**, 1931, doi:10.1038/s41598-018-20154-1 (2018).
32 Papanicolaou, A. *et al.* The whole genome sequence of the Mediterranean fruit fly, Ceratitis capitata (Wiedemann), reveals insights into the biology and adaptive evolution of a highly invasive pest species. *Genome Biol* **17**, 192, doi:10.1186/s13059-016-1049-2 (2016).
33 Anstead, C. A. *et al.* A blow to the fly - Lucilia cuprina draft genome and transcriptome to support advances in biology and biotechnology. *Biotechnol Adv* **34**, 605-620, doi:10.1016/j.biotechadv.2016.02.009 (2016).
34 Benoit, J. B. *et al.* Unique features of a global human ectoparasite identified through sequencing of the bed bug genome. *Nat Commun* **7**, 10165, doi:10.1038/ncomms10165 (2016).
35 McKenna, D. D. *et al.* Genome of the Asian longhorned beetle (Anoplophora glabripennis), a globally significant invasive species, reveals key functional and evolutionary innovations at the beetle-plant interface. *Genome Biol* **17**, 227, doi:10.1186/s13059-016-1088-8 (2016).
36 Lindsey, A. R. I. *et al.* Comparative genomics of the miniature wasp and pest control agent Trichogramma pretiosum. *BMC Biol* **16**, 54, doi:10.1186/s12915-018-0520-9 (2018).

37 Armisen, D. *et al.* The genome of the water strider Gerris buenoi reveals expansions of gene repertoires associated with adaptations to life on the water. *bioRxiv*, 242230 (2018).
38 Schwager, E. E. *et al.* The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. *BMC Biology* **15**, 62, doi:10.1186/s12915-017-0399-x (2017).
39 Poynton, H. C. *et al.* The Toxicogenome of Hyalella azteca: a model for sediment ecotoxicology and evolutionary toxicology. *Environmental Science and Technology - In Press* (2018).
40 Auman, T. & Chipman, A. D. The Evolution of Gene Regulatory Networks that Define Arthropod Body Plans. *Integr Comp Biol* **57**, 523-532, doi:10.1093/icb/icx035 (2017).
41 Davidson, E. H. *The Regulatory Genome: Gene Regulatory Networks In Development And Evolution*. (Elsevier Science, 2010).
42 Maccallum, I. *et al.* ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads. *Genome Biol* **10**, R103, doi:10.1186/gb-2009-10-10-r103 (2009).
43 Butler, J. *et al.* ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome research* **18**, 810-820, doi:10.1101/gr.7337908 (2008).
44 Deng, J. & Worley, K. C. *Atlas-Link*, <https://www.hgsc.bcm.edu/software/atlas-link> (2010).
45 Song, X., Liu, Y., Qu, J., Gibbs, R. A. & Worley, K. C. *ATLAS gapfill 2.2*, <https://www.hgsc.bcm.edu/software/atlas-gapfill> (2012).
46 Pryszcz, L. P. & Gabaldon, T. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic acids research* **44**, e113, doi:10.1093/nar/gkw294 (2016).
47 Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC bioinformatics* **12**, 491, doi:10.1186/1471-2105-12-491 (2011).
48 Poelchau, M. *et al.* The i5k Workspace@NAL--enabling genomic data access, visualization and curation of arthropod genomes. *Nucleic acids research* **43**, D714-719, doi:10.1093/nar/gku983 (2015).
49 Waterhouse, R. M., Zdobnov, E. M., Tegenfeldt, F., Li, J. & Kriventseva, E. V. OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011. *Nucleic acids research* **39**, D283-288, doi:10.1093/nar/gkq930 (2011).
50 Kriventseva, E. V. *et al.* OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic acids research* **43**, D250-256, doi:10.1093/nar/gku1220 (2015).
51 Peters, R. S. *et al.* Transcriptome sequence-based phylogeny of chalcidoid wasps (Hymenoptera: Chalcidoidea) reveals a history of rapid radiations, convergence, and evolutionary success. *Mol Phylogenet Evol* **120**, 286-296, doi:10.1016/j.ympev.2017.12.005 (2018).
52 Schwentner, M., Combosch, D. J., Pakes Nelson, J. & Giribet, G. A phylogenomic solution to the origin of insects by resolving crustacean-hexapod relationships. *Curr. Biol.* **27**, 1818-1824 e1815, doi:10.1016/j.cub.2017.05.040 (2017).
53 Sanderson, M. J. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**, 301-302 (2003).
54 Han, M. V., Thomas, G. W., Lugo-Martinez, J. & Hahn, M. W. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Molecular biology and evolution* **30**, 1987-1997, doi:10.1093/molbev/mst100 (2013).
55 Ames, R. M., Money, D., Ghatge, V. P., Whelan, S. & Lovell, S. C. Determining the evolutionary history of gene families. *Bioinformatics* **28**, 48-55, doi:10.1093/bioinformatics/btr592 (2012).
56 Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic acids research* **44**, D279-285, doi:10.1093/nar/gkv1344 (2016).
57 Schuler, A. & Bornberg-Bauer, E. Evolution of Protein Domain Repeats in Metazoa. *Molecular biology and evolution* **33**, 3170-3182, doi:10.1093/molbev/msw194 (2016).
58 Ekman, D., Bjorklund, A. K. & Elofsson, A. Quantification of the elevated rate of domain rearrangements in metazoa. *Journal of molecular biology* **372**, 1337-1348, doi:10.1016/j.jmb.2007.06.022 (2007).

Acknowledgments

We thank the staff at the Baylor College of Medicine Human Genome Sequencing Center for their contributions. Genome sequencing, assembly and annotation was funded by National Human Genome Research Institute grant U54 HG003273 to R.A.G. GWCT and MWH are funded by NSF DBI-1464611. ED was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 281125614 / GRK2220. RMW, PI and EMZ were funded by The Swiss National Science Foundation (PP00P3_170664 to RMW, 31003A_143936 to EMZ). Contributions by DDM and AW were supported in part by NSF-DEB grant 1355169 and USDA-APHIS Cooperative Agreement 15-8130-0547-CA to DDM. BM and ON acknowledge the German Research foundation (NI 1387/3-1, MI 649/12-1) and the Leibnitz Graduate School on Genomic Biodiversity Research. CS was supported by the Blanton J. Whitmire endowment, Housing and Urban Development NCHHU-0007-13, National Science Foundation 1557864 and Alfred P. Sloan Foundation 2013-5-35 MBE. Funding from Australian Wool Innovation (to P.B. and R.B.G.) and the Australian Research Council (to R.B.G.) is gratefully acknowledged. Support to R.B.G.'s laboratory by YourGene Bioscience and Melbourne Water Corporation is gratefully acknowledged. This project was also supported by a Victorian Life Sciences Computation Initiative (VLSCI; grant number VR0007) on its Peak Computing Facility at the University of Melbourne, an initiative of the Victorian Government (R.B.G.). C.A.A. holds an NSERC Postdoctoral Fellowship. N.D.Y. holds an NHMRC Early Career Research Fellowship. P.K.K. is the recipient of a scholarship (STRAPA) from the University of Melbourne.

Author Contributions

GWCT performed phylogenetic analyses/reconstructions and designed the website. GWCT, ED and RMW performed gene content and protein domain analysis and interpretation and contributed data to the website. KG and MADG performed methylation analysis. MB, HC, HD, HVD, SD, YH, JCJ, SLL, MM, NO-L, DMM, RAG and SR managed and performed sequence library and sequence generation. SCM, JQ, DSTH, KCW and SR performed genome assemblies, DSTH generated automated annotations, and DK, ES and SR submitted data to public databases. MP, CC, MM-T performed and supported manual annotation. NAA, JBB, DB, HC, JJD, LE, CEL, JG, RBG, CAA, PKK, NDY, PB, TP, DEG-R, AMH, MH, LH, WBH, AK, ARIL, GM, APM, DDM, BM, ON, SRP, KAP, MP, RSP, HCP, N-MP, DR, CS, SDS, EDS, DBS, RS, MRS, NUS, and EEZ provided species materials and expertise. RMW, PI and EMZ performed orthology analysis. YL and MEP performed GO analysis, AW, DDM and MF assessed coleopteran and dipteran gene content change. JG summarized chelicerate gene families. MEP, KJH, JHW, KCW, GWCT, ED, RMW, ADC, EBB, MWH and SR prepared the manuscript contributed to project management and provided leadership.

Figure Legends

Figure 1. OrthoDB orthology delineation for the i5K pilot species.

The bars show Metazoa-level orthologs for the 76 selected arthropods and three outgroup species (of 13 outgroup species used for analysis) partitioned according to their presence and copy-number, sorted from the largest total gene counts to the smallest. The 28 i5K species generated in this study with a total of 533,636 gene models are indicated in bold green font with species images positioned above each bar. A total of 38,195 orthologous protein groups were annotated among the total 76 genomes.

Figure 2: **Arthropod phylogeny inferred from 150 single copy genes among taxonomic orders and 1,627 to 4,097 single-copy genes within the six multi-species orders.** Divergence times estimated with non-parametric rate smoothing and fossil calibrations at 22 nodes. Species in bold are those sequenced within the framework of the i5K pilot project. All nodes, except those indicated with red shapes, have bootstrap support of 100. Nodes of particular interest are labeled in orange and referred to in the text. Larger fonts indicate multi-species orders enabling CAFE 3.0 likelihood analyses (see methods). See Fig. S17 and Table S10 for full node labels.

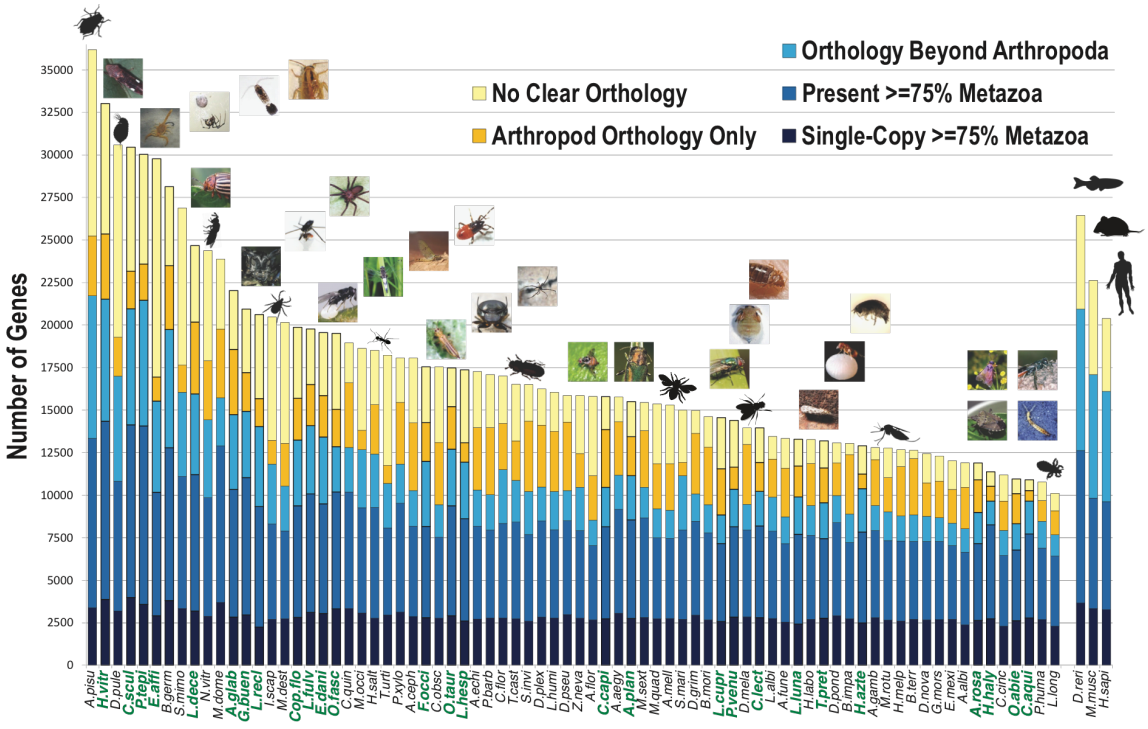
Figure 3: Summary of major results from gene family, protein domain, and methylation analyses. A. We identify 147 gene families emerging during the evolution of insects, including several which may play an important role in insect development and adaptation. **B.** Contrastingly, we find only 10 emergent gene families during the evolution of holometabolous insects, indicating many gene families were already present during this transition. **C.** Among all lineage nodes, we find that the node leading to Lepidoptera has the most emergent gene families. **D.** leafcutter ants have experienced high rates of both gene gain/loss and protein domain rearrangement. **E.** *Blattella germanica* has experienced the highest number of rapid gene family changes, possibly indicating its ability to rapidly adapt to new environments. **F.** We observe signals of methylation patterns in all Aranae (spiders) genomes investigated (species shown: The Brown Recluse spider, *Loxosceles reclusa*) and the genome of the bark scorpion *Centruroides exilicauda*. The two peaks show different CG counts in different gene features, with depletion of CG sequences in the left peak due to methylated C's mutating to T. This suggests epigenetic control of a significant number of spider genes. Additional plots for all species in this study are shown in Fig. S8.

Figure 4: Rate of genomic change along the arthropod phylogeny: **A**, frequency of amino acid substitutions per site, **B**, gene gains/losses and **C**, domain changes. All rates are averaged per My and color-indicated as branches of the phylogenetic tree. Species names are shown on the right; specific subclades are highlighted by colors according to the taxonomic groups noted in Fig. 2.

Table 1. Inferred and corrected gene counts at ancestral arthropod nodes.

Branch leading to	Rank	Node	Inferred gene count	Corrected gene count
Chelicerata	sub-phylum	74	8,360	14,602
Araneae	order	70	12,688	15,847
Crustacea	sub-phylum	65	8,876	15,230
Hexapoda	sub-phylum	63	8,776	14,882
Insecta	class	62	9,601	14,695
Hemiptera	order	57	10,968	15,250
Holometabola	super-order	51	9,076	13,982
Hymenoptera	order	25	9,609	12,460
Coleoptera	order	30	10,355	13,399
Lepidoptera	order	34	11,136	12,831
Diptera	order	48	9,020	12,920

Figure 1



Orthology Beyond Arthropoda
No Clear Orthology
Present >=75% Metazoa
Arthropod Orthology Only
Single-Copy >=75% Metazoa
Number of Genes
35000
32500
30000
27500
25000
22500
20000
17500
15000
12500
10000
7500
5000
2500
0
A.pisu
H.vitr
D.pule
C.scul
P.tepi
E.affi
B.germ
S.mimo
L.dece
N.vitr
M.dome
A.glab
G.buen
L.recl
I.scap
M.dest
Cop.flo
L.fulv
E.dani
O.fasc
C.quin
M.occi
H.salt
T.urti
P.xylo
A.ceph
F.occi
C.obsc
O.taur
L.hesp
A.echi
P.barb
C.flor
T.cast
S.invi
D.plex
L.humi
D.pseu
Z.neva
A.flor
C.capi
A.aegy
A.plan
M.sext
M.quad
A.mell
S.mari
D.grim
B.mori
L.cupr
P.venu
D.mela
C.lect
L.albi
A.fune
L.luna
H.labo
T.pret
D.pond
B.impa
H.azte
A.gamb
M.rotu
H.melp
B.terr
D.nova
G.mors
E.mexi
A.albi
A.rosa
H.haly
C.cinc
O.abie
C.aqui
P.huma
L.long
D.reri
M.musc
H.sapi

Figure 2

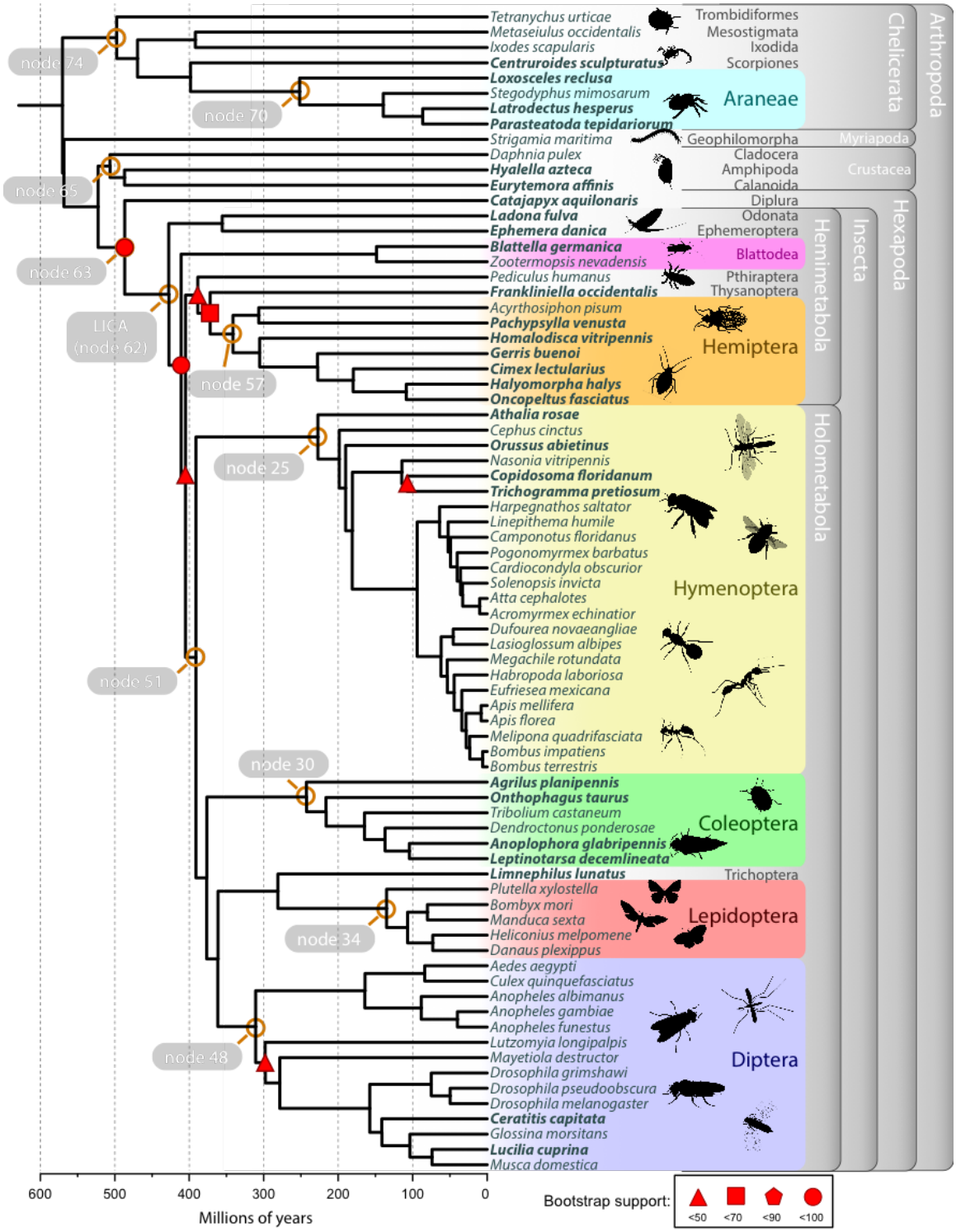

Tetranychus urticae
Metaseiulus occidentalis
Ixodes scapularis
Centruroides sculpturatus
Loxosceles reclusa
Stegodyphus mimosarum
Latrodectus hesperus
Parasteatoda tepidariorum
Strigamia maritima
Daphnia pulex
Hyalella azteca
Eurytemora affinis
Catajapyx aquilonaris
Ladona fulva
Ephemera danica
Blattella germanica
Zootermopsis nevadensis
Pediculus humanus
Frankliniella occidentalis
Acyrthosiphon pisum
Pachypsylla venusta
Homalodisca vitripennis
Gerris buenoi
Cimex lectularius
Halyomorpha halys
Oncopeltus fasciatus
Athalia rosae
Cephus cinctus
Orussus abietinus
Nasonia vitripennis
Copidosoma floridanum
Trichogramma pretiosum
Harpegnathos saltator
Linepithema humile
Camponotus floridanus
Pogonomyrmex barbatus
Cardiocondyla obscurior
Solenopsis invicta
Atta cephalotes
Acromyrmex echinatior
Dufourea novaeangliae
Lasioglossum albipes
Megachile rotundata
Habropoda laboriosa
Eufriesea mexicana
Apis mellifera
Apis florea
Melipona quadrifasciata
Bombus impatiens
Bombus terrestris
Agrilus planipennis
Onthophagus taurus
Tribolium castaneum
Dendroctonus ponderosae
Anoplophora glabripennis
Leptinotarsa decemlineata
Limnephilus lunatus
Plutella xylostella
Bombyx mori
Manduca sexta
Heliconius melpomene
Danaus plexippus
Aedes aegypti
Culex quinquefasciatus
Anopheles albimanus
Anopheles gambiae
Anopheles funestus
Lutzomyia longipalpis
Mayetiola destructor
Drosophila grimshawi
Drosophila pseudoobscura
Drosophila melanogaster
Ceratitis capitata
Glossina morsitans
Lucilia cuprina
Musca domestica
Trombidiformes
Mesostigmata
Ixodida
Scorpiones
Araneae
Geophilomorpha
Cladocera
Amphipoda
Calanoida
Diplura
Odonata
Ephemeroptera
Blattodea
Pthiraptera
Thysanoptera
Hemiptera
Hymenoptera
Coleoptera
Trichoptera
Lepidoptera
Diptera
Chelicerata
Myriapoda
Crustacea
Hexapoda
Insecta
Hemimetabola
Holometabola
Arthropoda
node 74
node 70
node 65
node 63
LICA (node 62)
node 57
node 25
node 51
node 30
node 34
node 48
600
500
400
300
200
100
0
Millions of years
Bootstrap support:
<50
<70
<90
<100

Figure 3

A

Last Insect Common Ancestor: 147 emergent gene families

Function	Emergent families
Wing morphogenesis	EOG86HJQQ EOG8TMTG9 EOG80ZTDS
Exoskeleton development and pigmentation	EOG8Q2GZG EOG8RZ1DS EOG8VDSCK EOG8WHC14 EOG8XPT03
Adaptation to terrestrial environment	EOG83XXJ1 EOG82VBZ4 EOG8PVRGC EOG8HTC7X
Larval behavior	EOG81K1SK

B

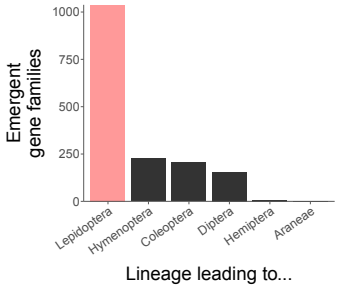
Last Holometabolous Common Ancestor: 10 emergent gene families

Function	Emergent families
Anterior head segmentation	EOG8HDW8X
Nucleosome assembly	EOG8G1PZD
Transporter activity	EOG847J8K
Transferase activity	EOG8ZPH98
Serine-type endopeptidase	EOG8QJV3F

+ 5 families with no known function

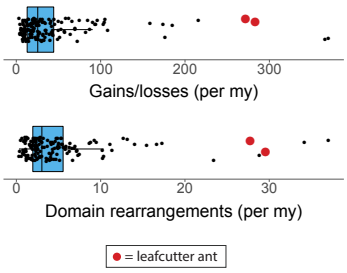
C

Last Lepidopteran common ancestor: 1,038 emergent gene families



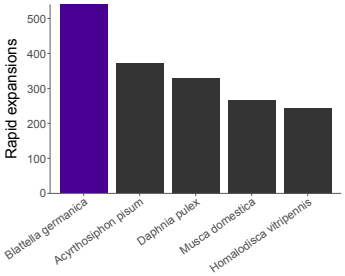
D

Leafcutter ants: High rates of gene gain/loss and domain rearrangements



E

Blattella germanica: Most rapid gene family expansions



F

Arachnids: High rates of DNA methylation

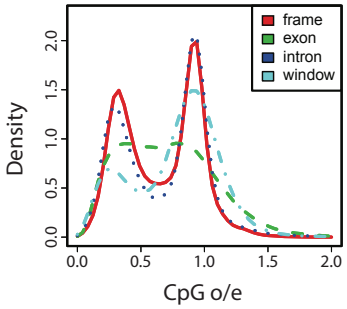
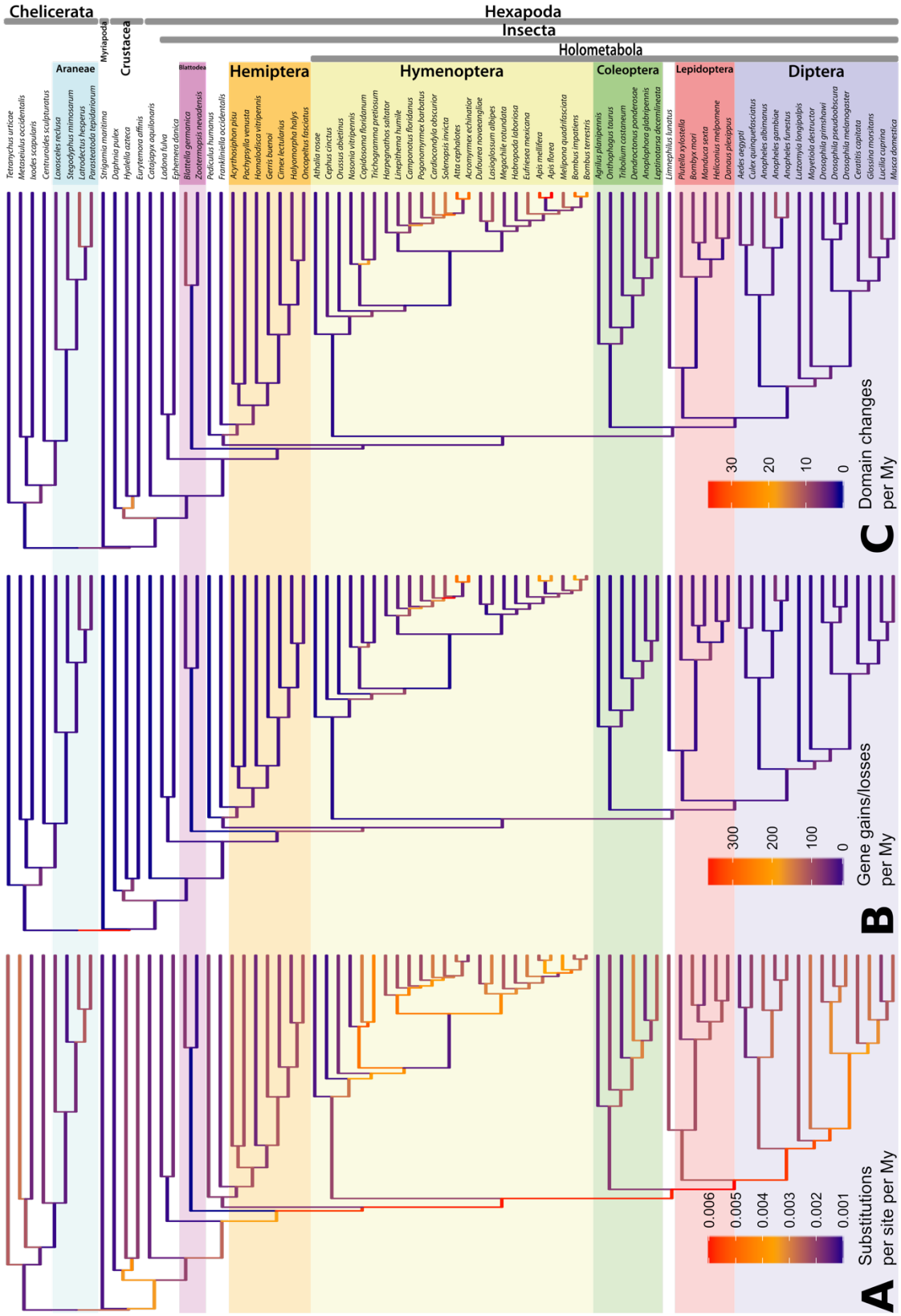


Figure 4


Chelicerata
Hexapoda
Insecta
Holometabola
Myriapoda
Crustacea
Araneae
Blattodea
Hemiptera
Hymenoptera
Coleoptera
Lepidoptera
Diptera
Tetranychus urticae
Metaseiulus occidentalis
Ixodes scapularis
Centruroides sculpturatus
Loxosceles reclusa
Stegodyphus mimosarum
Latrodectus hesperus
Parasteatoda tepidariorum
Strigamia maritima
Daphnia pulex
Hyalella azteca
Eurytemora affinis
Catajapyx aquilonaris
Ladona fulva
Ephemera danica
Blattella germanica
Zootermopsis nevadensis
Pediculus humanus
Frankliniella occidentalis
Acyrthosiphon pisu
Pachypsylla venusta
Homalodisca vitripennis
Gerris buenoi
Cimex lectularius
Halyomorpha halys
Oncopeltus fasciatus
Athalia rosae
Cephus cinctus
Orussus abietinus
Nasonia vitripennis
Copidosoma floridanum
Trichogramma pretiosum
Harpegnathos saltator
Linepithema humile
Camponotus floridanus
Pogonomyrmex barbatus
Cardiocondyla obscurior
Solenopsis invicta
Atta cephalotes
Acromyrmex echinatior
Dufourea novaeangliae
Lasioglossum albipes
Megachile rotundata
Habropoda laboriosa
Eufriesea mexicana
Apis mellifera
Apis florea
Melipona quadrifasciata
Bombus impatiens
Bombus terrestris
Agrilus planipennis
Onthophagus taurus
Tribolium castaneum
Dendroctonus ponderosae
Anoplophora glabripennis
Leptinotarsa decemlineata
Limnephilus lunatus
Plutella xylostella
Bombyx mori
Manduca sexta
Helicoverpa melpomene
Danaus plexippus
Aedes aegypti
Culex quinquefasciatus
Anopheles albimanus
Anopheles gambiae
Anopheles funestus
Lutzomyia longipalpis
Mayetiola destructor
Drosophila grimshawi
Drosophila pseudoobscura
Drosophila melanogaster
Ceratitis capitata
Glossina morsitans
Lucilia cuprina
Musca domestica
A
Substitutions per site per My
0.006
0.005
0.004
0.003
0.002
0.001
B
Gene gains/losses per My
300
200
100
0
C
Domain changes per My
30
20
10
0