Gene content evolution in the arthropods

Gregg W. C. Thomas1, Elias Dohmen2,3,4, Daniel S. T. Hughes5,6, Shwetha C. Murali5,7, Monica Poelchau8, Karl Glästraid9,10, Clare A. Anstead11, Nadia A. Ayoub12, Phillip Batterham13, Michelle Bellais5,14, Greta J. Binford15, Hsu Chao5, Yolanda H. Chen16, Christopher Childers9, Huyen Dinh5, Harsha Vardhan Doddapaneni9, Jian J. Duan17, Shannon Dugan5, Lauren A. Esposito18, Markus Friedrich19, Jessica Garb20, Robin B. Gasser11, Michael A. D. Goodisman9, Dawn E. Gundersen-Rindal11, Yi Han5, Alfred M. Handler22, Masatsugu Hatakeyama23, Lars Herwig24, Wayne B. Hunter25, Panagiota Ioannidou26,27, Joy C. Jayaseelan5, Divya Kalra5, Abderrahman Khila28, Pasi K. Korhonen11, Carol Eunmi Lee29, Sandra L. Lee5, Yiyuan Li30, Amelia R. Lindsey31,32, Georg Mayer24, Alistair P. McGregor33, Duane D. McKenna34, Bernhard Misof35, Mala Munidas5, Monica Munoz-Torres36,37, Donna M. Muzny5, Oliver Niehuis38, Nkechinyere Osuji-Lacy5, Subba R. Palli39, Kristen A. Panfilio40, Matthias Pechmann41, Trent Perry13, Ralph S. Peters42, Helen C. Poynton43, Nikola-Michael Prpic44,45, Jiaxin Qu5, Dorith Rotenberg46, Coby Schal47, Sean D. Schoville48, Erin D. Scully49, Evette Skinner5, Daniel B. Sloan50, Richard Stouthamer51, Michael R. Strand52, Nikolaus U. Szczech53, Asela Wijeratne34,35, Neil D. Young11, Eduardo E. Zattara34, Joshua B. Benoit55, Evgeny M. Zdobnov26, Michael E. Pfrender30, Kevin J. Hackett56, John H. Werren57, Kim C. Worley5, Richard A. Gibbs5, Ariel D. Chipman58, Robert M. Waterhouse59, Erich Bornberg-Bauer1,3,60, Matthew W. Hahn1 and Stephen Richards5,61*

Abstract

Background: 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.

Results: 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.

Conclusions: 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.

Keywords: Arthropods, Genome assembly, Genomics, Protein domains, Gene content, Evolution, DNA methylation

© The Author(s). 2020 Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.
Background

Arthropods (chelicerates, myriapods, crustaceans, and hexapods) 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 carnivores, detritivores, herbivores, and parasites. As major components of the world’s biomass, their diversity and ubiquity lead naturally to significant interactions with humanity, as crop pests, disease vectors, food sources, pollinators, 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 antennae, feeding appendages, gills, and jointed legs. 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 melts 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 colonizations of terrestrial and freshwater habitats by ancestrally marine 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 breadth of arthropod diversity.

Results

An arthropod evolution resource

As a pilot project for the i5K initiative to sequence 5000 arthropod genomes [6], we sequenced and annotated the genomes of 28 arthropod species (Additional file 1: Table S1). 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 protein ortholog 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 [8–10], with the exception being that we recover a monophyletic Crustacea, rather than the generally accepted paraphyletic nature of Crustacea with respect to Hexapoda; the difference is likely due to our restricted taxon sampling (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://arthrofam.org and Additional file 1: Table S1) forms the basis for the analyses detailed below and is an unprecedented tool for identifying and tracking genomic changes over arthropod evolutionary history.

Genomic changes 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 [11]. Genes can also be lost over time, constituting an underappreciated mechanism of evolution [12, 13]. 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 [14]. 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. Overall, we inferred 181,157 gene family expansions and 87,505 gene family contractions. A total of 68,430 gene families were inferred to have gone extinct in at least one lineage, and 9115 families emerged in different groups. We find that, of the 268,662 total gene family changes, 5843 changes are statistically rapid (see “Methods”), with the German cockroach, Blattella germanica, having the most rapid gene family changes (Fig. 3e). 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 (Additional file 1: Table S15). Using the estimates of where in the phylogeny these events occurred, we can infer characteristics of ancestral arthropods. For example, we identified 9601 genes in the last insect common ancestor (LICA) and estimate ~14,700 LICA genes after correcting for unobserved gene extinctions (Fig. 2, Additional file 2: Figure S1 and Additional file 1: Table S16). We reconstructed similar numbers for ancestors of the six well-represented arthropod taxa in our sample (Fig. 2 and Additional file 1: Table S16). Of the 9601 genes present in LICA, we identified 147 emergent gene families (i.e., lineage-restricted families with no traceable orthologs in other clades) which appeared concurrently with the evolution of insects (Fig. 3a, Fig. 2 node 62, Additional file 1: Table S18). Gene Ontology 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 aquatic), 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 [15–17]. Others, such as cuticle proteins underlying differences in exoskeleton structure, may enable cuticle properties optimized for diverse environmental habitats or life history stages [18]. In contrast, the data reveal only ten gene families that arose along the ancestral lineage of the Holometabola (Fig. 3b, Additional file 1: Table S19), 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 of hemimetabolous insects [19].

We identified numerous genes that emerged in specific orders of insects. Strikingly, we found 1038 emergent gene families in the first ancestral Lepidoptera node (Fig. 3c). 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 (Additional file 2: Figure S2). 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 that leading to Coleoptera, and 156 in that leading to Diptera. Though our sampling is extensive, it is possible that gene families we have classified as emergent may be present in unsampled lineages.

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 (Fig. 4 and Additional file 2: Figure S3). 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 Araneae (spiders, Additional file 1: Table S20). In parallel, we find a high rate of new protein domains in the subphylum Chelicerata, including a large number within Araneae 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 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 (Additional file 1: Table S21) 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, 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 in B. germanica. The impressive capability of this cockroach to survive many environments and its social behavior 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 genomic change: (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 genomic change have a strikingly small amount of variation in rate among the investigated species (Fig. 4). We estimate an average amino acid substitution rate of $2.54 \times 10^{-3}$ substitutions per site per My with a standard deviation of $1.11 \times 10^{-3}$. 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 Acarina (mites), and to the Diptera (flies).

Though we observe thousands of genomic changes across the arthropod phylogeny, they are mostly evenly distributed (Fig. 3d). Rates of gene duplication and loss show remarkably little variation, both across the tree and within the six multi-species orders (Additional file 1: Table S13). 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* along with the internal node leading to the leafcutter ants and the red fire ant (node HY29) have exceptionally high gene gain/loss rates of 266, 277, and 370 per My, respectively (Fig. 3d). This is an order of magnitude higher than average, as previously reported among leafcutter ants [25]. Removing these nodes, 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 eightfold lower than the rate of gene gain/loss. Interestingly, we discovered a strong correlation between rates of gene gain/loss and domain rearrangement (Figs. 3d and 4 and Additional file 2: Figure S3). 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. This novel finding is surprising, given that these processes follow largely from different underlying genetic events (see [27] for discussion of these processes).

Our examination found no correlation between variation in amino acid substitution rates and rates of gene gain/loss or domain rearrangement rates (Fig. 4 and Additional file 2: Figure S3). 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.

**Control of novel genes: methylation signals in arthropod genomes**

Our description of gene family expansions in arthropods by gene duplication naturally suggests the need for differential control of duplicated genes. Insect epigenetic control by CpG methylation is important for caste development in honey bees [28] and polyphenism in aphids [29]. However, signals of methylation are not seen in every insect, and the entire Dipteran order appears to have lost the capacity for DNA methylation. Given this diversity in the use of, and capacity for epigenetic control by DNA methylation, we searched for signals of CpG methylation in our broader sampling of arthropod genomes. We find several independent losses of the DNA methylation machinery across the arthropods (Additional file 2: Figure S4) [30]. This indicates that DNA methylation is not universally necessary for development and that the DNA methyltransferases in insects may function in ways not previously appreciated [31]. Additionally, putative levels of DNA methylation vary
A key consequence of the relatively stable rate of gene family and protein domain change across the arthropod tree is that major morphological transitions (e.g., full metamorphosis, wing emergence, Additional file 1: Table S17) could not easily be identified by surges in gene content or protein domain change. There are two possible exceptions in our data. We see an increased rate of gene family extinction along the ancestral nodes from the ancestor of the cockroach and termites and hemimetabolous insects to the ancestor of Lepidoptera and Diptera (Additional file 2: Figure S6), suggesting the possibility of evolution by gene loss [12, 43]. This rate increase is not seen in wing evolution. The second possible exception is that of whole genome duplications (as proposed in spiders [40]), when there is a temporary opening of the “evolutionary search space” of gene and protein domain content. This overall finding is in line with the emerging understanding that morphology is effected by complex gene networks, which are active mostly during ontogenetic processes [44], 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 [45]. The current study was unable to address the evolution of non-coding sequences such as enhancers, promoters, and small and other non-coding RNAs underlying these networks due to the lack of sequence conservation over large evolutionary distances; however, our results underscore their evolutionary importance.

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.

Conclusions
We have generated annotated draft genome assemblies for 28 species sampled from across the phylum Arthropoda. Combined with previously sequenced genomes, we documented changes in gene and protein domain content across 76 species sampled from 21 orders, spanning more than 500 million years. The resulting Arthropod
resource comprises reconstructed gene content and protein domain arrangements for 38,195 orthogroups at each node of the Arthropod phylogeny. It enables inference and identification of gene content in terms of both families and domains at ancestral phylogenetic nodes. Rates of gene content change and protein domain change appear to be correlated, but neither gene content change nor protein domain change rates are correlated with amino-acid change. This work is a first look at the history of arthropod gene evolution, and an example of the power of comparative genomic analysis in a phylogenetic context to illuminate the evolution of life on earth.

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 4.9 Tb of raw nucleotide sequence (Additional file 1: Table S2). For individual species, reads were assembled using AllpathsLG [46, 47] followed by refinements employing Atlas-Link [48] and Gapfill [49]. Version 1.0 assemblies had minimum, mean, and maximum scaffold N50 lengths of 13.8 kb, 1.0 Mb, and 7.1 Mb (Additional file 1: Table S3). Following re-assembly and collapsing of unassembled haplotypes using Redundans [50], version 2.0, assemblies had minimum, mean, and maximum contig N50 lengths of 11.1 kb, 166.2 kb, and 857.0 kb with a mean scaffold N50 lengths of 619 kb (Additional file 1: Table S3). The redundans software and new assemblies became available late in the project timeline, and thus automated gene annotations, orthologous gene family identification in OrthoDB, and analysis were performed on the Version 1 ALLPATHS-LG-based assemblies.

To support the annotation, RNAseq data were generated from 25 species for which no data were available (Additional file 1: Table S4). A MAKER [51] 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 Additional file 2: 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 (Additional file 1: Table S5, see Additional file 1: Table S7 for completeness statistics). Many of these gene models were manually curated using the i5k Workspace@NAL [52]. 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 ortholog groups represent all descendants of a single gene present in the genome of the last common ancestor of all the species considered for clustering [53].

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 [54], which comprises 87 arthropods and an additional 86 other metazoans (including 61 vertebrates). Although the majority of these gene sets were built using MAKER (Additional file 1: Table S6), variation in annotation pipelines and supporting data, introduce a potential source of technical gene content error in our analysis.

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 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 (8806 groups, 8 species); Hemiptera (8692 groups, 7 species); Hymenoptera (21,148 groups, 24 species); Coleoptera (12,365 groups, 6 species); and Diptera (17,701, 14 species). All identified BRHs, amino acid 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 [8–10].

Of the six orders in our dataset represented by multiple species (Additional file 2: Figures S7–S12), 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 [55]), and the three other methods placing *C. floridanum* as sister to *Trichogramma pretiosum* (Additional file 2: Figure S9). Within the Diptera, we obtained a sister group relationship between the sand fly, *Luizomyia longipalpis*, and the Culicidae, but this was not a stable topology across methods (Additional file 2: Figure S12).

The most contentious nodes in the phylogeny involve the relationship of crustaceans and hexapods. We recover a monophyletic Crustacea that represent the sister clade to Hexapoda (Fig. 2), in contrast to recent analyses suggesting this group is paraphyletic in respect to Hexapoda [56]. However, an extensive phylogenetic investigation (Additional file 2: Supplementary Results, Additional file 2: Figure S13) shows that regardless of the inference method used, the relationships among the crustacean and hexapod lineages remain uncertain. Aside from these few discrepancies, branch 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. The most likely reason for the difference from the current consensus is poor taxon sampling. Importantly, remipedes (the possible sister group of the hexapods) are missing from our taxon sampling, as are mystacocarids, ostracods, and pentatomids, and may change this result to the current consensus when added as was seen in [56].

**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* [57] 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] (Additional file 1: Table S14). 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 [58]. *CAFE* 3.0 is able to estimate the amount of assembly and annotation error (ε) present in the input gene count data. This is done by treating the observed gene family counts as distributions rather than certain observations. *CAFE* can then be run repeatedly on the input data while varying these error distributions to calculate a pseudo-likelihood score for each one. The error model that is obtained as the minimum score after such a search is then used by *CAFE* to obtain a more accurate estimate of λ and reconstruct ancestral gene counts throughout the tree (Additional file 1: 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 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 seen in Additional file 1: 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 [59]. 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 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 quantify the effect of any single species on the parsimony gene family reconstructions, we performed 100 jackknife replicates while randomly removing 5 species from each replicate. We find that ancestral gene counts are not greatly impacted by the presence or absence of any single genome (Additional file 2: Figure S14).

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 (Additional file 2: Figure S1) ($r^2 = 0.37; P = 4.1 \times 10^{-9}$). 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) [60]. 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 [61, 62]. 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 (Additional file 2: Figure S15): 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.

Supplementary information

The online version of this article (https://doi.org/10.1186/s13059-019-1925-7) contains supplementary material, which is available to authorized users.

Acknowledgments

We thank the staff at the Baylor College of Medicine Human Genome Sequencing Center for their contributions.

Review history

The review history is available as Additional file 3.

Authors’ 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, JCI, SLL, MW, NO-L, DMM, RAG, and SR managed and performed sequence library and sequence generation. SCM, IQ, DSTH, KOW, and SR performed genome assemblies. DSTH generated automated annotations, and DK, ES, and SR submitted data to public databases. MP, CC, and MM-T performed and supported manual annotation. NAA, JBB, DB, HC, JJD, LE, CEL, JG, RBG, CAA, PKN, NDY, PB, TP, DEG-R, AMH, MH, LH, WIB, AK, ARIL, GM, APM, DDM, BM, ON, SRP, KAP, MP, RSP, HCP, N-MP, DR, CS, SDS, EDS, DS, BS, RFS, NUS, and EEE provided species materials and expertise. RMW, PI, and EMZ performed orthology analysis. YL and MEP performed GO analysis, and AW, DDM, and MF assessed coleopteran and dipteran gene content change. JG summarized chelicerate gene families. MEP, KIH, JHW, KCW, GWCT, ED, RMW, ADC, EBB, MWH, and SR prepared the manuscript contributed to project management and provided leadership. All authors read and approved the final manuscript.

Funding

Genome sequencing, assembly, and annotation were funded by National Human Genome Research Institute grant U54 HG003273 to R.A.G. GWCT and MWH are funded by NSF DBI-1564611. ED was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 281125614 / GK2220. RMW, PI, and EMZ were funded by The Swiss National Science Foundation (P300P3_170664 to RMW, 31003A_143936 to EMZ). Contributions by DDM and AW were supported in part by NSF-DEB grant 1335169 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
Availability of data and materials
All datasets generated and/or analyzed during the current study are publicly available. All reconstituted gene content for the lineages of the 76 species in this arthropod phylogeny are freely available at https://arthrofam.org and in Additional file 1: Table S1. All DNA, RNA, genome assembly, and transcriptome assembly sequences can be found in the NCBI under the i5k Umbrella BioProject PRJNA163973 [63].

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
1Department of Biology and Department of Computer Science, Indiana University, Bloomington, IN, USA. 2Institute for Evolution and Biodiversity, University of Münster, 48149 Münster, Germany. 3Institute for Bioinformatics and Chemoinformatics, University of Hamburg, Hamburg, Germany. 4Westphalian University of Applied Sciences, 45665 Recklinghausen, Germany. 5Human Genome Sequencing Center, Department of Human and Molecular Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA. 6Present Address: Institute for Genomic Medicine, Columbia University, New York, NY 10032, USA. 7Present Address: Howard Hughes Medical Institute, Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA. 8National Agricultural Library, USDA, Beltsville, MD 20705, USA. 9School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA 30332, USA. 10Present Address: Penn Epigenetics Research Unit, United States Department of Agriculture, Agricultural Research Service, Newark, DE, USA. 11Department of Biology and Environmental Sciences, Wayne State University, Detroit, MI 48202, USA. 12Department of Biological Sciences, University of Massachusetts Lowell, 198 Riverside Street, Lowell, MA 01854, USA. 13USDA-ARS Invasive Insect Biocontrol and Behavior Laboratory, Beltsville, MD, USA. 14USDA-ARS, Center for Medical, Agricultural, and Veterinary Entomology, 1700 S.W. 23rd Drive, Gainesville, FL 32608, USA. 15Department of Insect Sciences, National Institute of Agrobiological Sciences, Tsukuba, 305-8634, Japan. 16Department of Zoology, Institute of Biology, University of Kassel, 34132 Kassel, Germany. 17USDA ARS, U. S. Horticultural Research Laboratory, Ft. Pierce, FL 34945, USA. 18Department of Genetic Medicine and Development and Swiss Institute of Bioinformatics, University of Geneva, 1211 Geneva, Switzerland. 19Present Address: Foundation for Research and Technology Hellas, Institute of Molecular Biology and Biotechnology, Vassilikia Voutou, 70013 Heraklion, Greece. 20Université 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. 21Department of Integrative Biology, University of Wisconsin, Madison, WI 53706, USA. 22Department of Biological Sciences, University of Notre Dame, 109B Galvin Life Sciences, Notre Dame, IN 46556, USA. 23Department of Entomology, University of California Riverside, Riverside, CA, USA. 24Present Address: Department of Biology, Indiana University, Bloomington, IN, USA. 25Department of Biological and Medical Sciences, Oxford Brookes University, Gipsy Lane, Oxford OX3 0BP, UK. 26Department of Biological Sciences, University of Memphis, 3700 Walker Ave, Memphis, TN 38152, USA. 27Center for Molecular Biodiversity Research, Zoological Research Museum Alexander Koenig, Bonn, Germany. 28Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, USA. 29Present Address: Phoenix Bioinformatics, 39221 Paseo Padre Parkway, Ste. J, Fremont, CA 94538, USA. 30Evolutionary Biology and Ecology, Institute of Biology I (Zoology), Albert Ludwig University of Freiburg, 79104 Freiburg (Brsg), Germany. 31Department of Entomology, University of Kentucky, Lexington, KY 40546, USA. 32School of Life Sciences, University of Warwick, Gibbet Hill Campus, Coventry CV4 7AL, UK. 33Collinge Bioceenter, Zoological Institute, Department of Developmental Biology, University of Cologne, 50674 Cologne, Germany. 34Centre of Taxonomy and Evolutionary Research, Arthropoda Department, Zoological Research Museum Alexander Koenig, Bonn, Germany. 35School for the Environment, University of Massachusetts Boston, Boston, MA 02125, USA. 36Johann-Friedrich-Blumenbach-Institut für Zoologie und Anthropologie, Abteilung für Entwicklungsbiologie, Georg-August-Universität Göttingen, Göttingen, Germany. 37Göttingen Center for Molecular Biosciences (GZMB), Georg-August-Universität Göttingen, Göttingen, Germany. 38Department of Entomology and Plant Pathology, North Carolina State University, Raleigh, NC 27606, USA. 39Department of Entomology and W.M. Keck Center for Behavioral Biology, North Carolina State University, Raleigh, NC 27695, USA. 40Department of Entomology, University of Wisconsin-Madison, Madison, USA. 41Stored Product Insect and Engineering Research Unit, USDA-ARS Center for Grain and Animal Health Research, Manhattan, KS 66502, USA. 42Department of Biology, Colorado State University, Ft. Collins, CO, USA. 43Department of Entomology, University of Georgia, Athens, GA, USA. 44Present Address: Arkansas Biosciences Institute, Arkansas State University, Jonesboro, AR, USA. 45Natural History Museum Vienna, Burggraben 7, 1010 Vienna, Austria. 46INIBIOMA, Univ. Nacional del Comahue – CONICET, Bariloche, Argentina. 47Department of Biological Sciences, University of Cincinnati, Cincinnati, OH 45221, USA. 48Crop Production and Protection, U.S. Department of Agriculture-Agricultural Research Service, Beltsville, MD 20705, USA. 49Department of Biology, University of Rochester, Rochester, NY 14627, USA. 50Department 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. 51Department of Ecology & Evolution and Swiss Institute of Bioinformatics, University of Lausanne, 1015 Lausanne, Switzerland. 52Department Protein Evolution, Max Planck Institute for Developmental Biology, Tübingen, Germany. 53Present Address: UC Davis Genome Center, University of California, Davis, CA 95616, USA.

Received: 6 November 2019 Accepted: 26 December 2019

References
1. Lozano-Fernandez J, Carton R, Tanner AR, Puttick MN, Blaxter M, Vinther J, Olesen J, Giribet G, Edgecombe GD, Pisani D. A molecular palaeobiological exploration of arthropod terrestrialization. Philos Trans R Soc B. 2016;371: 20160133.
48. Atlas-Link [https://www.hgsc.bcm.edu/software/atlas-link].
49. ATLAS gapfill 2.2 [https://www.hgsc.bcm.edu/software/atlas-gapfill].
50. Pryszcz LP, Gabaldon T. Redundans: an assembly pipeline for highly heterozygous genomes. Nucleic Acids Res. 2016;44:e113.
51. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011;12:491.
52. Poelchau M, Childers C, Moore G, Tsavatapalli V, Evans J, Lee CY, Lin H, Lin JW, Hackett K. The i5k workspace@NAL—enabling genomic data access, visualization and curation of arthropod genomes. Nucleic Acids Res. 2015;43:D714–9.
53. Waterhouse RM, Zdobnov EM, Tegenfeldt F, Li J, Kriventseva EV. OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011. Nucleic Acids Res. 2011;39:D283–8.
54. Kriventseva EV, Tegenfeldt F, Petty TJ, Waterhouse RM, Simao 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:D250–6.
55. Peters RS, Niehuis O, Gunkel S, Blaser M, Mayer C, Podsiadloowski L, Koslov A, Donath A, van Noort S, Liu S, et al. Transcriptome sequence-based phylogeny of chalcidid wasps (Hymenoptera: Chalcidoidea) reveals a history of rapid radiations, convergence, and evolutionary success. Mol Phylogenetics. 2018;120:286–96.
56. Schwentner M, Combosch DJ, Pakes Nelson J, Gribet G. A phylogenomic solution to the origin of insects by resolving crustacean-hexapod relationships. Curr Biol. 2017;27:1818–24. e1815
57. Sanderson MJ. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics. 2003;19:301–2.
58. Han MW, Thomas GW, Lugo-Martinez J, Hahn MW. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Mol Biol Evol. 2013;30:1987–97.
59. Ames RM, Money D, Ghatge VP, Whelan S, Lovell SC. Determining the evolutionary history of gene families. Bioinformatics. 2012;28:48–55.
60. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Quevreux M, Sangrador-Vegas A, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44:D279–85.
61. Schuler A, Bornberg-Bauer E. Evolution of protein domain repeats in Metazoa. Mol Biol Evol. 2016;33:3170–82.
62. Ekman D, Bjorklund AK, Eloffsson A. Quantification of the elevated rate of domain rearrangements in metazoa. J Mol Biol. 2007;372:1337–48.
63. Thomas GWC, Dohmen E, Hughes DST, Murali SC, Poelchau M, Glastad K, et al. i5k arthropod genome pilot project. NCBI Seq Read Arch. https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA163973

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.