


RESEARCH ARTICLE

Open Access



# Multi-faceted analysis provides little evidence for recurrent whole-genome duplications during hexapod evolution

Dick Roelofs<sup>1,2\*</sup> , Arthur Zwaenepoel<sup>3,4†</sup>, Tom Sisternans<sup>1</sup>, Joey Nap<sup>1</sup>, Andries A. Kampfraath<sup>1</sup>, Yves Van de Peer<sup>3,4,5</sup>, Jacintha Eilers<sup>1</sup> and Ken Kraaijeveld<sup>6,7</sup>

## Abstract

**Background:** Gene duplication events play an important role in the evolution and adaptation of organisms. Duplicated genes can arise through different mechanisms, including whole-genome duplications (WGDs). Recently, WGD was suggested to be an important driver of evolution, also in hexapod animals.

**Results:** Here, we analyzed 20 high-quality hexapod genomes using whole-paranome distributions of estimated synonymous distances ( $K_S$ ), patterns of within-genome co-linearity, and phylogenomic gene tree-species tree reconciliation methods. We observe an abundance of gene duplicates in the majority of these hexapod genomes, yet we find little evidence for WGD. The majority of gene duplicates seem to have originated through small-scale gene duplication processes. We did detect segmental duplications in six genomes, but these lacked the within-genome co-linearity signature typically associated with WGD, and the age of these duplications did not coincide with particular peaks in  $K_S$  distributions. Furthermore, statistical gene tree-species tree reconciliation failed to support all but one of the previously hypothesized WGDs.

**Conclusions:** Our analyses therefore provide very limited evidence for WGD having played a significant role in the evolution of hexapods and suggest that alternative mechanisms drive gene duplication events in this group of animals. For instance, we propose that, along with small-scale gene duplication events, episodes of increased transposable element activity could have been an important source for gene duplicates in hexapods.

**Keywords:** Polyploidy, Gene duplication and loss, Co-linearity, Insecta, Collembola, Gene tree reconciliation, Synonymous distance

## Background

Gene duplication is an important source of genetic variation that can propel adaptive evolution and speciation [1, 2]. Large-scale gene duplication events, such as large segmental or whole-genome duplications (WGDs), are

thought to have played a major role in evolution because they supply hundreds or even thousands of novel gene duplicates on which evolution can work. Such events enhance evolutionary innovation due to the creation of genetic redundancy [3], may increase mutational and environmental robustness [4], and reduce the risk of extinction [5, 6]. WGD events have also been linked to increased diversification [7, 8], either directly or after a lag-time period [7, 9], but see [10]. It has also been argued that WGD may facilitate adaptation and survival under specific conditions, for instance during periods of

\* Correspondence:

<sup>†</sup>Dick Roelofs and Arthur Zwaenepoel are shared first author, contributed equally to the work.

<sup>1</sup>Department of Ecological Science, Vrije Universiteit, De Boelelaan 1085, 1081HV Amsterdam, The Netherlands

<sup>2</sup>Keygene N.V, Agro Business Park 90, 6708 PW Wageningen, The Netherlands

Full list of author information is available at the end of the article

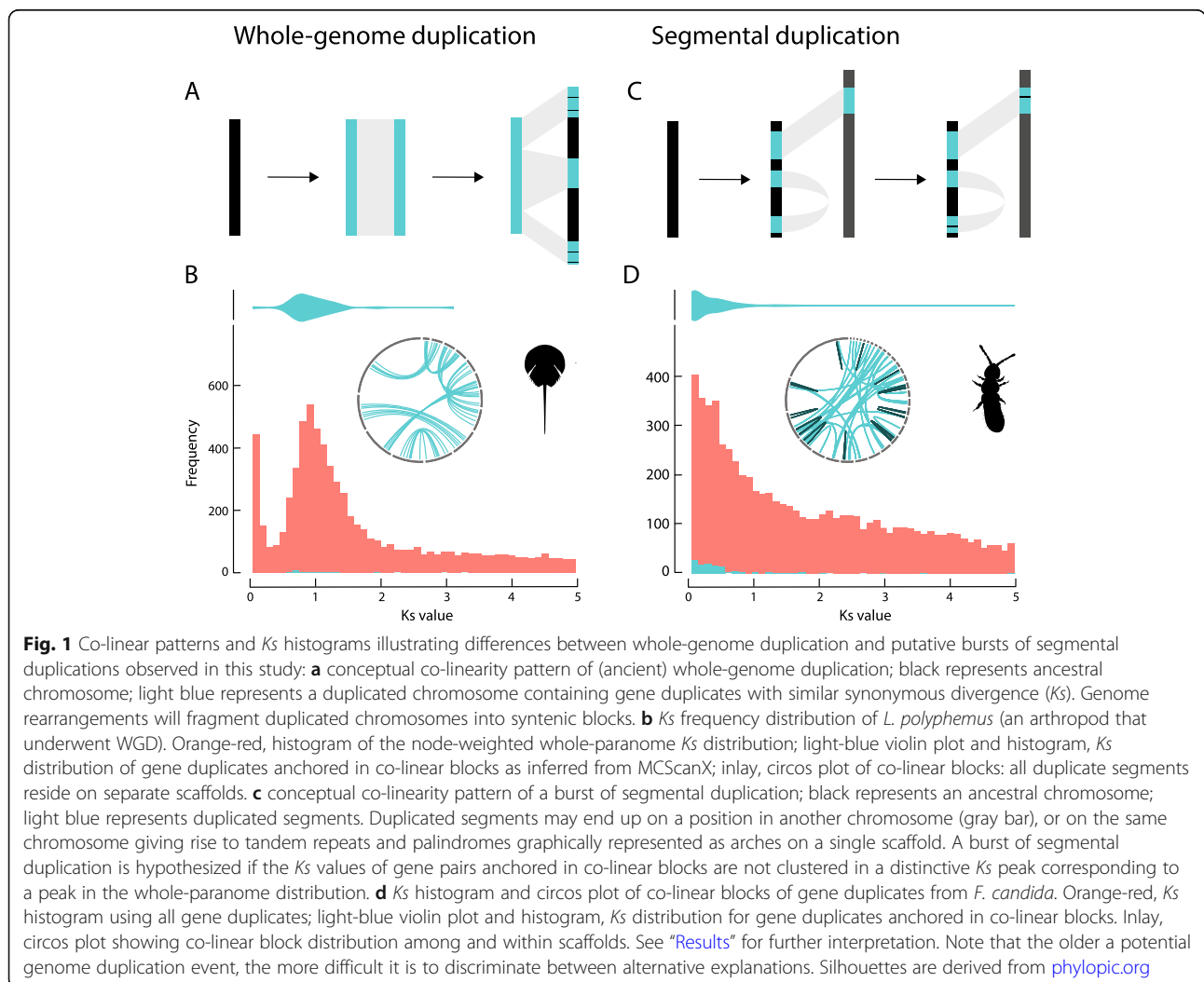


© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. 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 in a credit line to the data.

environmental turmoil or cataclysmic events [11]. In animals, WGD has been rarely detected, for which various explanations have been put forward, such as different reproductive modes [12], dosage-sensitive sex determination [12–14], and more intricate physiological and developmental constraints [12]. Nevertheless, ancient WGDs are hypothesized to have played a role in the evolution of teleost fish, mollusks [15], and particularly arthropods [16–18].

Arthropods are a highly speciose and diverse group of animals. Gene and genome duplications may have played an important role in generating this diversity, yet patterns of duplication in this group are still under discussion. With the recent availability of high-quality invertebrate genome sequences, several cases of large-scale gene duplication and potential WGDs have been identified. For example, genomes of arachnids were found to harbor many paralogous gene pairs and a duplicated Hox gene cluster, indicating a WGD event in their evolutionary history [19, 20]. The horseshoe crab

*Limulus polyphemus* even shows four copies of the Hox gene cluster, suggesting two rounds of WGD within this group [16, 20]. Recently, Li et al. [18] reported 18 ancient WGDs and six other large-scale bursts of gene duplication in 118 analyzed transcriptomes and 25 genomes of hexapods. The inferred pattern of scattered WGDs across the phylogenetic tree of hexapods would indicate that WGDs have been an important driver of evolutionary novelty and diversity in insects. However, inference of ancient WGDs remains challenging [21]. For instance, in a recent study, some of us showed that unaccounted variation in duplication and loss rates across lineages can strongly affect assessment of the presence or absence of WGDs [22]. Also, a recent re-analysis of the *Bombyx mori* data could not confirm a previously reported putative Lepidoptera-specific WGD [23]. Therefore, it remains unclear whether the observed patterns of duplications in hexapods are indeed indicative of frequent ancient WGDs. To reliably distinguish the presence of remnants of ancient WGD from



alternative scenarios, several independent lines of evidence are required. One line of evidence is peaks in  $K_S$  distributions (Fig. 1a, b), where the number of duplication events is plotted as a function of an estimate of the synonymous distance associated with these events ( $K_S$ , which serves as a proxy for age). Bursts of duplicates with similar synonymous divergence are indicative of a large-scale duplication event, although one has to be aware of the caveats in interpreting such distributions [21, 24, 25]. For one,  $K_S$  distributions cannot be used to infer very ancient WGDs, due to saturation of the synonymous distance and the stochastic nature of the molecular clock [24].

An important second line of evidence for uncovering remnants of large-scale or whole-genome duplications is within-genome co-linearity (Fig. 1a, b). In the absence of gene loss and rearrangements associated with the rediploidization process, we expect duplicated pairs (referred to as “anchor pairs” or “anchors”) to initially reside in syntenic and co-linear blocks (Fig. 1a). Genome rearrangements and gene loss will erode this signal over time, but even for ancient duplication events, substantial intragenomic co-linearity remains observable [26–29]. Such intragenomic co-linear blocks are usually assumed to result from ancient WGD, although other events, such as bursts of transposon activity, translocations, or aneuploidy, can also potentially generate similar signals (Fig. 1b) [30]. Furthermore, very ancient WGDs can often no longer be reliably identified from co-linear analyses, in particular when high-quality chromosome-level assemblies are lacking. In those cases, lastly, most evidence is based on the analysis of gene trees (e.g., [18, 31, 32]). However, assessing the support for a hypothetical WGD from gene trees remains very challenging, and results from such approaches should be treated with considerable caution [22, 33]. Generally, the combination of temporal (from  $K_S$  distributions and gene trees) and structural (co-linearity) evidence provides the most reliable means towards distinguishing WGD from other sources of gene duplication, yet requires high-quality genome data across multiple species.

Here, we present a multi-faceted analysis of highly contiguous, well-annotated genomes of 20 hexapod species and one outgroup, *Limulus polyphemus*, to study the occurrence of gene- and genome duplication events in this diverse species group. To this end, we (1) inferred whole-paranome  $K_S$  distributions, (2) performed co-linearity analysis, which classifies the genomic context of gene duplications as dispersed gene pairs or segmental duplications, and (3) employed a recently proposed probabilistic gene tree reconciliation approach designed to test hypotheses about ancient WGDs and to estimate lineage-specific duplication and loss rates. In contrast to the recently published study on this topic [18], we find

little support for an important role of WGDs during hexapod evolution. Alternatively, we propose that mainly small-scale gene duplication, together with instances of segmental duplication, possibly mediated through homologous recombination guided by surges in transposon activity, explains the observed duplication signal in hexapods.

## Results

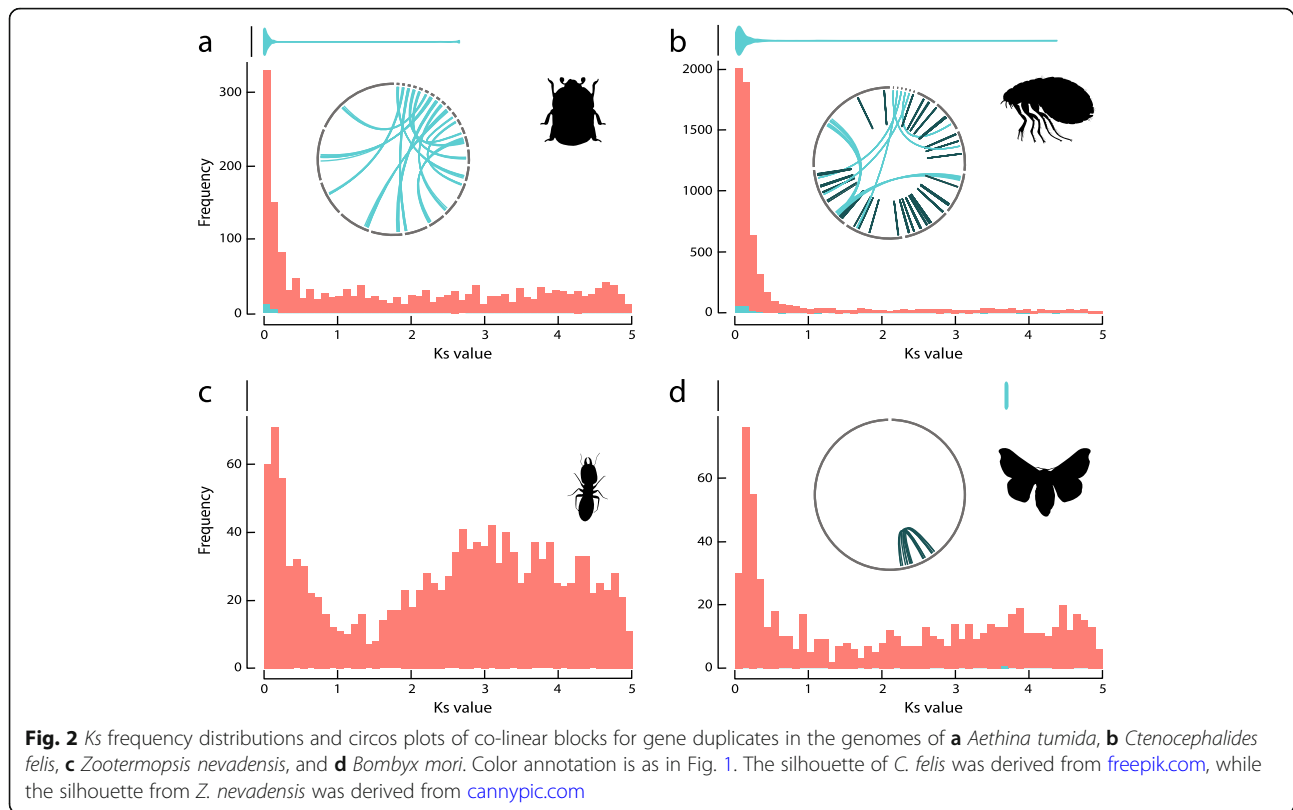
### Delineation of paranomes

We selected 20 high-quality, well-annotated genomes from the available hexapod genome sequences to represent the widest taxonomic diversity of this group. We included the genome of the chelicerid *Limulus polyphemus*, where there is compelling evidence for an ancient genome duplication [16], in our analysis as an outgroup. Table S1 lists the accession numbers of assembled genomes, as well as details on sequencing technology and assembly output. Within-genome sequence similarity searches, followed by MCL clustering, were used to detect paralogous gene pairs (see “Methods”), the number of which varied widely among the analyzed hexapod genomes, ranging from 1225 in the collembolan *Holacanthella duospinosa* to 21,073 in the dipteran *Aedes aegypti* (Table S1).

### Synonymous divergence and co-linearity among gene pairs

Gene duplicate age distributions were inferred by estimating the expected number of synonymous substitutions per synonymous site (synonymous distance or  $K_S$ ) across the paranome following the approach of Vanneste et al. [24] (see “Methods”). Ancient WGDs result in a characteristic pattern of a peak in distributions of gene duplicates of similar age (similar divergence at synonymous sites) that tend to be in co-linear regions within a genome. The genome of *L. polyphemus* showed such a distinct peak of  $K_S$  values around  $K_S \approx 0.8$  (Fig. 1b). The light blue violin plot above the  $K_S$  distributions in Fig. 1 represents the distribution of anchor pair  $K_S$  values found in co-linear blocks (drawn in the inlayed circos plots), which again shows an increase in frequency around  $K_S \approx 0.8$  in the case of *L. polyphemus* (Fig. 1b). This pattern is consistent with the ancient WGD reported for this species [16]. A second peak for the more ancient WGD event hypothesized for *L. polyphemus* is not observed in neither the whole-paranome nor anchor pair  $K_S$  distribution, likely due to the age of this event exceeding the window for which  $K_S$  distributions can be used for WGD detection [21, 25].

By contrast, none of the hexapod genomes in our study showed a similar pattern consistent with ancient WGD. Instead, we observed several different patterns of gene duplication among hexapod genomes (Figs. 1d and



2, and Figure S1). Most of the genomes in this study showed a high number of young duplicate pairs (low  $K_S$ , Figure S1), resulting in the L-shaped distribution that is characteristic for the continuous process of small-scale gene duplication (SSD) and loss [21, 34]. Representative patterns were, for instance, observed in the genomes of *Aethina tumida* and *Ctenocephalides felis* (Fig. 2). The collembolan hexapod *Folsomia candida* exhibited relatively high numbers of duplicates (over 20,000 duplicate pairs, Table S1) and has been assumed to have undergone a lineage-specific WGD in a previous study [18]. If true, we hypothesized that a distinct peak in the whole-paranome  $K_S$  distribution should be observed that coincides with an increase of anchor gene pairs in the same  $K_S$  range, as for instance in the case of *L. polyphemus* (Fig. 1c). Instead, in *F. candida* only a gradual decline in the number of duplication events for increasing  $K_S$  was observed, rather than a distinct peak. Although the *F. candida* genome did, indeed, contain a substantial number of co-linear blocks (55, Table S1), the distribution of anchor pair  $K_S$  values in these co-linear blocks exhibited a similar distribution with declining density for larger  $K_S$  values (blue violin plot Fig. 1d). This pattern suggests that the emergence and loss of such co-linear blocks is a continuous process, reminiscent of small-scale gene duplication (Fig. 1d,  $K_S$  histogram). The reason for this atypical shape of the collembolan anchor pair  $K_S$

distribution needs further investigation, but based on co-linearity analysis (as well as macrosynteny and phylogenomic analyses, see further), we see no reason to invoke a WGD event.

Similar observations were made for other hexapods: none of these genomes retained duplicates co-linearly organized in the way expected under WGD. As a matter of fact, in all hexapod species, most gene pairs were classified as dispersed duplicates not being physically linked in co-linear blocks (Table S1, Fig. 2, Figure S1). Co-linear segments were observed in six out of the 20 genomes (Fig. 2, Figure S1), but in most genomes, these segmental duplications were present in low numbers (Table S1). Larger numbers of duplicated co-linear segments, such as described above for *F. candida*, were also found in the genomes of *A. tumida* and *C. felis* (Table S1, Figure S1). However, similar to the pattern in *F. candida*, in these species the gene pairs organized in co-linear segments were recent and did not coincide with any peaks in  $K_S$  age distributions (Figs. 1d and 2a, b). In *B. mori*, a recent study [18] reported 728 syntenic chains of which 83 could potentially represent segmental duplications. In contrast, we only detected two co-linear segments (consisting of six and seven gene pairs, respectively, Fig. 2d), which is substantially less. To verify whether our more stringent parameter settings in MCScanX could explain the large difference in number

of segmental duplications retrieved, we first increased the *E*-value cutoff for all-against-all BLASTP searches with *B. mori* proteins from  $10e-10$  to  $10e-5$  (as applied by the Li et al. study [18]). This caused an increase of the number of co-linear genes from 13 to 94. We then used three genes to seed a co-linear block and applied a Manhattan distance of 40 in our MCScanX analysis. This resulted in the identification of 10 co-linear blocks, consisting of 94 co-linear gene pairs, which is still 8 times less than the number of co-linear blocks identified previously [18]. Further decreasing the stringency of these parameter settings eventually yielded 13 co-linear blocks, still a much lower number than previously reported [18]. Changing these parameter settings with regard to the analysis of the *L. polyphemus* genome resulted in an increase from 7 co-linear blocks with 44 gene pairs, to 14 co-linear blocks with 79 gene pairs. The  $K_S$  values of these gene pairs fall within the range of detected  $K_S$  peak of  $K_S \approx 0.8$  (Fig. 1b), which is in line with WGD in this species.

#### Genome structure of segmental duplications and macrosynteny patterns

Long co-linear blocks of paralogs covering large fractions of the genome are usually considered to support WGD events. In the case of fragmented genome assemblies, genuine WGD-derived co-linear blocks most probably reside on different scaffolds (Fig. 1a). Indeed, the previously inferred ancient WGD event in the chelicerid *L. polyphemus* shows exactly this pattern (Fig. 1b). Of the hexapods we studied, *A. tumida* is the only genome with substantial numbers of co-linear segments located on different scaffolds (Fig. 2a). However, a spurious pattern similar to this can also arise if excessive allelic variation in the genome assembly prevents the collapse of haplotigs into single contigs. Therefore, we tested this explanation by analyzing sequence read coverage of contigs, expecting a drop in coverage among two contigs if they are two haplotigs covering the same genomic region. Indeed, we found that the mean sequence read coverage for *A. tumida* is  $124.03\times$  with a 95% confidence interval (CI) between 123.8 and 124.5 assuming normal distribution (Figure S2A). By contrast, the coverage of the co-linear segments is on average  $88.82\times$  (95% CI 64.6–119.9, Figure S2B), which is below the lower bound of the 95% CI of coverage among 1000 random contigs. Thus, we conclude that at least some of these co-linear blocks may in fact correspond to haplotigs that did not collapse into one scaffold, and do not correspond to bona fide duplicated regions.

In the *C. felis* genome, we inferred 49 co-linear regions, of which 41 are located on the same scaffold (Fig. 2b, inlay circos plot). Thirty-one of these are organized as palindromes, while the remaining 10 are

organized as tandem repeats, which do not support WGD as a source of gene duplication in this species. Similarly, 14 out of 55 co-linear blocks in *F. candida* are located within the same scaffold (Fig. 1d). Importantly, no significant drop in sequence read coverage in these co-linear blocks was found, suggesting that these represent true segmental duplication events [35]. Given the low  $K_S$  values of anchor gene duplicates in *F. candida*'s co-linear blocks (blue violin plot in Fig. 1d, see section above), we infer that they evolved recently. However, if these co-linear blocks would have emerged as a result of a recent WGD event, most should still reside on different chromosomes. By contrast, we observe 14 co-linear blocks to reside within the same chromosome, which does not support WGD as a source of gene duplication in this species. In the case of *B. mori* (where a WGD event was proposed in its lepidopteran ancestor), the two co-linear blocks were also located on a single scaffold (Fig. 2d).

Extensive gene loss and genomic rearrangements may, however, cause substantial “erosion” of co-linear patterns through evolutionary time. Nevertheless, in the absence of conserved gene order, the chromosome-scale distribution of gene duplicates is still expected to be conserved to more or lesser extent, a pattern referred to as macrosynteny [28]. Following Nakatani and McLysaght [23], we further visualized the position of putative gene duplicates in a scatter plot representation for all species where at least one co-linear segment was found (Table S1; Figure S3). Although this visualization can be challenging due to the fragmented nature of some of the included assemblies, these plots clearly indicate that large-scale patterns of within-genome synteny are lacking in the examined genomes. A notable exception is *F. candida*, where peculiar non-random patterns can be observed (Figure S3G). However, the observed organization of gene duplicates for this species is not compatible with WGD either. Recently, we showed that at least some of *F. candida*'s segmental duplications are highly enriched in transposons [35], suggesting transposon-mediated gene duplication or TE proliferation. This remains however speculative and is subject of ongoing research. Taken together, the absence of both substantial co-linear segments and large-scale syntenic patterns suggests no role for WGD in the evolution of hexapods, at least in the evolutionary time frame that allows the inference of WGD events from genome structure.

#### Phylogenomic gene tree-species tree reconciliation

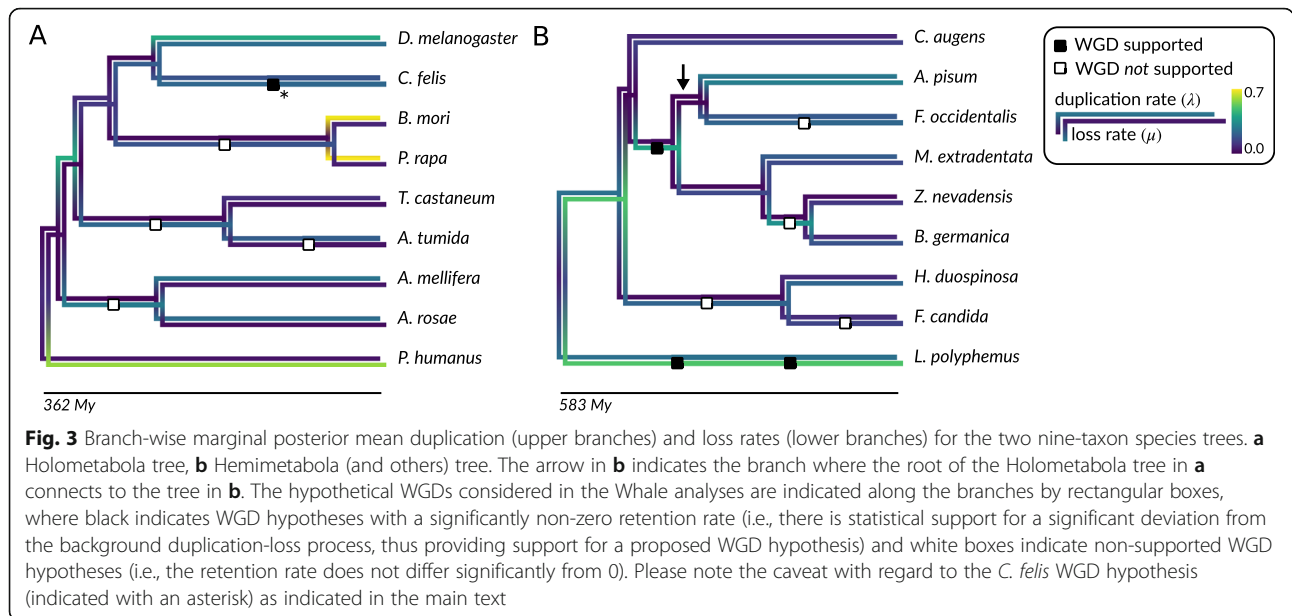
We conducted phylogenomic analyses to further investigate patterns of gene duplication and loss, and potential ancestral large-scale duplication events, in a phylogenetic context. To this end, we applied a recently developed Bayesian gene tree reconciliation approach to estimate

parameters of a stochastic model of gene family evolution that accounts for duplication, loss, and WGD events, while considering uncertainty in the gene trees by employing amalgamated likelihood estimation (ALE) [22]. In a gene tree reconciliation approach towards WGD inference, we seek to explain the evolution of a set of gene family trees in the context of a known species tree in terms of a set of evolutionary events, in our case, gene duplication and loss events. Most approaches do not use a model of gene family evolution and perform some flavor of parsimony-based reconciliation, effectively counting for each branch in the species tree the number of gene tree clades with a least common ancestor (LCA) that corresponds to that branch in the species tree, among some set of eligible gene tree clades (for instance focusing only on clades with high bootstrap support values in the gene tree or some other subset defined by a filtering criterion). There are several potential problems with such naive reconciliation approaches, among which the reliance on a single estimated gene tree topology is perhaps the most obvious one. Also the assumption that the LCA reconciliation is the true reconciliation can be troubling, as at least one study reported that the most parsimonious reconciliation differs from the true reconciliation in about 19% of the examined cases [36].

In the context of WGD inference, another issue is when to decide whether a given number of inferred duplication events on a particular branch is sufficiently high to infer a polyploidization event. The approach taken by Li et al. [18] is to estimate a background duplication and loss rate for the entire data set using WGDgc [37], and use this estimated rate to simulate a set of gene trees with and without WGD. Based on these simulations, they perform a test to decide whether a particular observed number of duplicates are significantly higher than their simulations without WGD, and not significantly less than their positive simulations with WGD. There are however again several potential issues with such an approach. In particular, the assumption of constant duplication and loss rates across lineages that underlies the simulation procedure has been shown to be inaccurate in the context of WGD inference [22]. In the Bayesian approach of Zwaenepoel and Van de Peer [22], a model of gene family evolution that includes the background duplication and loss process and WGD is used to perform gene tree reconciliation directly in a model-based framework, alleviating the need for potentially problematic simulation schemes. Furthermore, in the Bayesian approach, we can account—in a systematic way—for variation in gene duplication and loss rates across the species tree. We note that with the approach we take here, as with any method that does not take genome structure into account, we cannot distinguish

between a WGD and an episodic burst of small-scale or segmental gene duplication events.

For the sake of computational tractability, we considered two trees of 9 species, the first comprising the Holometabola with outgroup *Pediculus humanus* and the second representing the other hexapod groups included in this study with outgroup *L. polyphemus*. We marked 11 hypothetical WGDs on these trees, based on both the co-linearity analyses in this study and the results of Li et al. [18]. Using the simplest possible model, assuming constant rates across the species tree, we found the posterior mean duplication rate ( $\lambda$ ) across the Holometabola tree to be 0.00215 (events/gene lineage/My) with 95% of the posterior density in [0.00212, 0.00217]. Similarly, a loss rate ( $\mu$ ) of 0.00146 [0.00143, 0.00150] was obtained. Employing a branch-wise rates model with an independent rates (IR) prior (see “Methods”) indicated increased duplication rates in Lepidoptera, and an increased loss rate in the branch leading to *P. humanus* (although, in general, estimated rates on isolated branches near the root should be interpreted with caution [22] (Fig. 3a). For the branch leading to the model species *Drosophila melanogaster*, we estimated a duplication rate of 0.00427 [0.00416, 0.00438] and loss rate of 0.00248 [0.00236, 0.00261] events/gene lineage/My, estimates which are well in accord with previous estimates for *D. melanogaster* (e.g.,  $\lambda = 0.0023$  [34],  $\lambda = 0.0050$  [38]). We do not find support for WGD in the ancestor of the Diptera, as previously reported, and our analyses could not confirm recently suggested putative WGDs in the stem branches of Lepidoptera, Coleoptera, and Hymenoptera [18]. Performing an analysis with a strongly informative prior on the duplication and loss rates, assuming little rate heterogeneity across branches of the species tree—a model more akin to the simulations employed by Li et al. [18]—did not support any of these hypothetical WGDs. Of all WGD hypotheses indicated along this species tree, only a putative *C. felis* event received significant support with a decisively non-zero retention rate ( $q$ ) in all analyses (Figure S6). We note that an inspection of three MCMC chains revealed in one of the chains a second mode in the posterior distribution with  $q \approx 0$  and an increased duplication rate for this branch (Figure S7). However, all chains eventually converge on the same distribution (Figure S7), which likely represents a dominating mode in the posterior where the vast majority of posterior mass is located. Nevertheless, this suggests that the posterior distribution could be multimodal, with besides this dominating mode, the possibility of multiple small “peaks” separated by large “valleys” of low posterior probability, and that the MCMC algorithm may have trouble crossing these valleys. We further note for this putative event in *C. felis* that reconciled trees sampled from the posterior



revealed that a very large number of duplication events were reconciled to the hypothetical WGD node, but each of which with very low posterior probability (Figure S8). Furthermore, the  $K_S$  distribution for this species shows much more recent duplicates than any of the other examined genomes, but does not show a peak in the distribution nor co-linear blocks indicative of WGD. Together, this suggests that the non-zero retention rate for this branch is spurious and that the signal in the gene trees mistaken for a WGD event more likely corresponds to an increased gene duplication rate due to some other mechanism.

For the other species tree (Fig. 3b), analysis under the constant-rates model revealed a tree-wide duplication rate of 0.00160 [0.00158, 0.00161] and loss rate of 0.00195 [0.00192, 0.00198] events/gene lineage/My. Bayesian inference with Whale using models of branch-wise duplication and loss rate variation was more challenging for this tree, and convergence was not attained for the full branch-wise rate models, presumably due to the long outgroup branch that included hypothetical WGDs. To mitigate these issues, we constrained the two branches stemming from the root to have identical rates. Under the IR model, we obtained duplication and loss rates within a similar range as for the Holometabola tree (Fig. 3b). Posterior inferences under an autocorrelated prior were quite different from results for the IR prior, but did not result in any qualitative differences with respect to potential WGDs. In both cases, the data showed to be compatible with a previously suggested putative Insecta-shared WGD [18] ( $q = 31\%$  [29%, 33%] for the IR model), although the retention rate was highly sensitive to the prior used (Figure S11). The hypothesized WGDs

in *L. polyphemus* were also recovered with high retention rates, yet with a very high posterior variance (94% [77%, 100%] for the oldest WGD and 44% [32%, 58%] for the most recent putative WGD). This high variability is not surprising as this branch represents almost 600 My of evolution, and the only temporal information in the ALE-based approach used in Whale comes from tree topologies and not branch lengths per se. In line with the  $K_S$  distribution and co-linearity analyses in the present study, we do not find compelling evidence for large-scale duplication events in the stem of Dictyoptera (cockroaches, termites, and mantises), stem of Colembolla (springtails), the branch leading to *Frankliniella*, or the branch leading to *Folsomia* (Figure S11).

### Discussion

Gene duplication has been appreciated as an important factor in evolution for a long time [1, 39]. Our current study based on the analysis of whole-paranome  $K_S$  distributions, intragenomic co-linearity, and gene tree reconciliation for 20 high-quality hexapod genomes confirms gene duplicate abundance in this taxonomic group, with estimated rates of small-scale duplication and loss on the order of 0.002 events/gene/My. As expected, most duplicate gene pairs in hexapods are of recent origin, compatible with the continuous process of small-scale duplication and loss. However, some hexapod genomes, such as those of collembolans, show substantial retention of ancient gene duplicates, suggesting variation in the rates of the continuous duplication and loss process across the hexapod phylogeny.

Our current study does not provide evidence for multiple whole-genome duplication events to have occurred

during the evolution of hexapods. We unveiled a few segmental or large-scale duplications, but only in some of the genomes, and mostly of recent origin. The fact that these duplications were often located on the same scaffold is difficult to reconcile with WGDs, especially for relatively recent hypothesized WGDs. For example, *F. candida* and *C. felis* contained within-genome co-linear regions, which did not coincide with alternative peaks in  $K_S$  distributions, and of which a substantial fraction was intra-chromosomally distributed (Figs. 1d and 2b). Some intra-chromosomal co-linearity is expected to arise during rediploidization, a process associated with chromosomal rearrangements such as chromosomal fusions and translocations [40–42]. However, on average, the majority of co-linear blocks arising from WGD should reside on different chromosomes [23, 26–28], especially in case of recent WGD events. Therefore, the high frequency of recently evolved intra-chromosomal co-linearity observed in *C. felis* and *F. candida* is not in concordance with a WGD scenario. Large-scale macrosynteny pattern did not suggest a role for WGD in the evolution of hexapod genome structure either.

Identifying co-linearity is not straightforward, particularly in the case of ancient duplications where subsequent genome dynamics and restructuring may have erased the co-linearity signature to a large extent [22]. Also, the robustness of such analysis is dependent on assembly quality and applied analysis tool [43]. In this respect, the case of *B. mori* is interesting because this genome is of high quality and was analyzed using three independent co-linearity analysis tools [18, 23]. Li et al. [18] identified 728 potentially syntenic chains using the SynMap tool from the CoGe platform [44] that included 2210 genes, and suggested that 83 chains were associated with an ancient WGD event in Lepidoptera identified by their MultitAxon Paleopolyploidy Search (MAPS) algorithm [18]. More recently, Nakatani and McLysaght [23] visualized the position of these chained BlastP hits from the Li et al. analysis [18] on the silkworm chromosomes. They found that the majority of these duplicates were not chained, but were randomly distributed over the entire genome instead of organized in syntenic blocks. As mentioned above, extensive chromosomal rearrangements following WGD may have randomized paralog distribution throughout the *B. mori* genome over evolutionary time. This explanation seems unlikely, however, given the high levels of macrosynteny that was observed between *B. mori* and the coleopteran genome of *Tribolium castaneum* [23]. In the current study, we used MCSanX with conservative parameter settings and identified only two true segmental duplications, which were organized on one scaffold (Fig. 2d). Adjusting the parameter settings to the ones used by Li

et al. [18] did not retrieve comparable results. Taken together, our independent analysis confirmed the previous argument that WGD did not contribute to genome evolution in *B. mori* [23]. This conclusion is corroborated by our gene tree-species tree reconciliation analysis, which did not find support for a putative WGD in the lepidopteran stem branch.

Phylogenomic gene tree reconciliation analyses provided further insights into the phylogenetic patterns of gene duplication and loss across hexapods, as well as more ancient hypothetical large-scale duplication events. Duplication and loss rates varied across lineages, but remained within the same order of magnitude across the entire phylogeny. Using the statistical approach implemented in Whale to assess putative WGDs along those branches that were previously investigated [18], we failed to confirm the conclusions of this study in all but one case. Despite the fact that our taxon sampling was more limited, these results cast doubts on the methodology of the previous study and perhaps the suitability of transcriptomic data to infer gene family evolutionary processes. In particular, the assumption of constant rates across lineages, as applied in the previous hexapod WGD study [18], can seriously compromise inference of WGDs [22]. Conclusively refuting a hypothesized ancient WGD event is of course challenging, but model-based statistical inference can indicate under which assumptions what conclusions are acceptable. We showed that if we assumed the rate of gene duplication and loss to vary across lineages, i.e., duplication and loss, follow independent relaxed molecular clocks, gene trees of multi-copy gene families did not provide support for all but one of the entertained WGD hypotheses. We stipulate that to further substantiate these results, an increased taxon sampling remains desirable, breaking up long branches for which a WGD is hypothesized.

Our failure to confirm putative WGD events in hexapods seems also supported by Hox gene cluster organization. As mentioned previously, the *L. polyphemus* genome contains up to four copies of each Hox gene, supporting the hypothesis that this genome evolved through two rounds of WGD [16]. A literature survey among several published hexapod genomes (*D. melanogaster*, *T. castaneum*, *F. candida*, *Orchesella cincta*, *Acyrtosiphon pisum*, *Zootermopsis nevadensis*) showed that these genes are represented in single copy [20, 45–48], which is in line with our current findings. While Hox gene clusters are very tightly organized as one dense gene cluster in vertebrate genomes, hexapod Hox gene clusters seem to show a more differential pattern of gene dispersion, often interspersed by other open reading frames and long stretches of non-coding DNA [35, 45, 49].



An alternative scenario that would produce the observed signatures of large-scale gene duplication events in our and previous studies [18] are bursts of transposable element (TE) activity [30]. For example, duplication-dependent strand annealing was elucidated as the mechanism explaining their formation in the *D. melanogaster* genome [50]. Hotspots of TEs cause an increase in homologous regions, providing more opportunity for homologous recombination and unequal crossing over to drive gene amplification [51]. In such case, segmental duplications reside in genomic regions with high TE activity/abundance [52]. A recent systematic survey of TE activity across insect genomes provided evidence for ancient bursts of TE activity [53] in many insect species, which coincide with the  $K_S$  distribution of gene pairs detected in this study. For instance, *Z. nevadensis* showed an increased frequency of gene duplicates with  $K_S$  values between 2 and 4 (Fig. 2c). This coincides with a second broad peak of LINE transposon abundance in the Petersen study [53]. Similar concordant patterns between  $K_S$  distributions of gene duplicates and divergence distributions of the DNA transposon, LTR transposon, and rolling circle transposon families [53] were observed in *P. humanus* and *Apis mellifera*. Moreover, co-linear gene blocks in the genome of *F. candida* were found to be spatially associated with high densities of transposable element [35], suggesting a link between transposon activity and segmental duplication in the evolution of this genome. Finally, evidence was found for the involvement of high-density transposon regions facilitating gene family expansion of odorant receptors in ants [35]. Based on these observations, a tentative hypothesis emerges that bursts in transposon activity early during the evolution of some hexapod lineages may provide the basis for segmental duplication, by facilitating duplication-dependent strand annealing as main mechanism of gene duplication (Fig. 1c). In order to test this hypothesis, junctions of co-linear blocks could be examined for particular transposon sequence features. Such analysis was performed within the human genome, where enrichment of Alu short interspersed element (SINE) sequences near or within such junction was observed [51]. Moreover, assessing coincidence of our current age estimation based on  $K_S$  distributions with estimations of transposition rates in the evolutionary past could provide further support for our hypothesis. Historical transposition events could be dated in a phylogenetic context, as was shown for past transposition rate estimations of pogo-like transposable elements in different *Fusarium* species [54]. However, we are currently unaware of a method to reliably calibrate these two time indications against each other.

## Conclusions

The analysis of intragenomic co-linearity,  $K_S$  distributions, and gene tree-species tree reconciliation across a wide taxonomic range of hexapod genome sequences suggests that gene duplication is pervasive among hexapods and that species differ in the degree to which ancient gene duplicates have been retained. Interpreting our results in the light of recent studies, we speculate that TE activity might explain the observed patterns of bursts of gene duplication, while compelling evidence for an important role of WGD in hexapod evolution is missing.

## Methods

### Data sampling

Sample selection was guided by the availability of high-quality assembled genome sequences and taxonomic breadth. We aimed for a balanced distribution of genomes over hexapod diversity, and selected two genomes for each hexapod order, if available. In case multiple genomes were available for a given hexapod order, we selected the two most contiguous ones. In total, 20 high-quality hexapod genomes were compiled. The genome of the non-hexapod *Limulus polyphemus* was included as an exemplary lineage where WGD is well described [16]. Supplementary Table 1 lists sample names with associated accession numbers of assembled genomes, as well as details on sequencing technology and assembly output.

### $K_S$ distribution analysis

For paralogous gene families of two or more members, we estimated the expected number of synonymous substitutions per synonymous site ( $K_S$  value) for each node in the gene family tree using the approach of Vanneste et al. [24] as implemented in the “wgd” pipeline [55]. In brief, for each species, an all-against-all protein similarity search was done using BlastP with an  $e$ -value cutoff of  $1e-10$ . A sequence similarity graph was constructed and gene families were inferred using Markov Clustering with MCL (inflation factor = 2.0) [56]. The amino acid sequences of gene families were used to infer a multiple sequence alignment with MAFFT v7 (using options: “--amino --maxiterate 1000 --localpair”) [57]. This alignment was then used as a guide for obtaining a gap-stripped codon-level alignment. For each gene family, pairwise  $K_S$  values were estimated through maximum likelihood using CODEML (with runmode = -2) from the PAML package [58], using the Goldman & Yang model (GY94) and the F3x4 method for estimating equilibrium codon frequencies. For each family, an approximate phylogenetic tree was obtained using FastTree with default settings [59], which was then used to construct node-weighted empirical  $K_S$  distributions. The final set

of  $K_S$  values for each genome is represented as a weighted histogram, where the  $y$ -axis represents the number of duplication events (not duplicate pairs), to detect temporal patterns of gene duplication.

### Co-linearity analysis

To detect co-linear blocks, we used the Multiple Co-linearity Scan toolkit (MCScanX), using standard settings [60], except where indicated. The duplicate gene classifier within MCScanX uses the MCScan algorithm to classify the genomic context of gene duplications into three groups: segmental (including putative WGD-derived duplicates), tandem, or dispersed. Initially, all genes classify as singletons, while all gene pairs identified by BlastP are assigned dispersed duplicates. The segmental/WGD label is assigned to anchor pairs in intragenomic co-linear blocks [60]. Genome-wide co-linearity was visualized using Circos [61].

### Probabilistic gene tree-species tree reconciliation

We used the recently developed Whale approach for statistical assessment of WGD hypotheses [22]. For the analyses using Whale (v0.3), we considered two species trees of nine species, the first comprising the Holometabola with as outgroup *P. humanus*, and the second representing the other hexapod groups included in this study with as outgroup *L. polyphemus*. Species trees with branch lengths in calendar time were obtained from the TimeTree database [62]. We assumed 11 WGD hypotheses on these trees, informed by both the co-linearity analyses in this study and the results of Li et al. [18]. For both sets of species, we inferred gene families using OrthoFinder v2.3.3 [63]. To rule out de novo origin of gene families in arbitrary subtrees of the relevant species tree, we filtered out gene families that did not contain at least one gene in both species' clades stemming from the root in the associated species tree. We further filtered out large gene families according to a Poisson outlier criterion. Specifically, under the assumption that the total family size  $X$  across species is approximately Poisson distributed, we have that  $Y = 2\sqrt{X} \sim \text{Normal}(\text{median}(Y), 1)$ . Based on this assumption, we filtered out all gene families for which  $Y > \text{median}(Y) + 3$ . This resulted in 6937 gene families for the Holometabola set and 6712 gene families for the other species set. Pre-alignment masking of putatively non-homologous characters was performed using Prequal v1.01 [64], after which a protein multiple sequence alignment (MSA) was inferred for each gene family using MAFFT v7 [57]. For each alignment, a sample from the posterior distribution of phylogenetic trees was obtained using MrBayes v3.2.6 [65] using the LG+Γ4 substitution model and default exponential priors on the branch lengths. We ran the MCMC for 110,000 generations for each family,

recording a sample every 10 generations after discarding the first 10,000 generations as burn-in. The conditional clade distribution (CCD) was computed using the ALEobserve tool from the ALE suite v0.4 [66].

We performed Bayesian inference under the DL+WGD model in Whale using different priors and model structures for the duplication and loss rates across lineages of the species tree. For both species sets, we initially performed an analysis assuming constant rates of duplication and loss across the species tree and no WGD hypotheses. We assumed an exponential prior distribution with mean 0.005 events per gene lineage per million year (ev/lineage/My) on both the duplication and loss rate and a Beta (10, 2) hyper prior on the  $\eta$  parameter of the geometric prior distribution on the number of lineages in a gene family present at the root of the species tree. We used the marginal posterior mean  $\eta$  value from the constant-rates analysis for the relevant species set in all subsequent analyses for the same species set. We next performed an analysis using hierarchical independent branch-wise rates prior, where duplication and loss rate across branches are assumed to be independent and identically distributed following a log-normal distribution. We assumed an exponential distribution (with mean 0.5) on the mean duplication and loss rate and an InverseGamma (5, 1) prior on the variance of the log-normal distribution. We used uniform priors on the retention rates for all WGDs. Last, we performed Bayesian inference using the geometric Brownian motion (GBM) prior with strong assumed phylogenetic correlation ( $\nu = 0.1$ ), to investigate the influence of different assumptions on the branch-wise duplication and loss rates on the estimated WGD retention rates. For the analyses under the GBM prior, we used the same priors on the mean duplication and loss rates and retention rates. Throughout our analyses of the second (non-Holometabola) tree, convergence issues in the MCMC required us to constrain the duplication and loss rates to be identical for the two branches stemming from the root. Throughout our study, we use multiple pilot runs for subsets of 1000 gene families to investigate convergence and base our reported estimates on a chain ran for 11,000 generations with a burn-in of 1000 generations for the full data set. Analysis for the full data set is very computationally intensive, and therefore, we compare the obtained MCMC chains for the full data set with multiple chains for the random 1000 gene family subsets visually to assess convergence (e.g., Figures S1, S2).

### Sequence read coverage analysis

To test whether co-linear regions in *A. tumida* show a drop in sequence read coverage, we used bedtools genomecov package version 2.28 to calculate the coverage of

each basepair in the *A. tumida* genome. Subsequently, mean coverage over co-linear blocks was calculated and compared to the mean coverage of 1000 randomly sampled genome regions of the same length as the co-linear blocks including a 95% confidence interval.

## Supplementary information

**Supplementary information** accompanies this paper at <https://doi.org/10.1186/s12915-020-00789-1>.

**Additional file 1: Figure S1.** Ks frequency distribution graphs and circos plots of collinear syntenic blocks for gene duplicates in the genomes of *A. Aedes aegypti*, *B. Acyrthosiphon pisum*, *C. Apis mellifera*, *D. Athalia rosae*, *E. Bemisia tabaci*, *F. Blattella germanica*, *G. Campodea augens*, *H. Drosophila melanogaster*, *I. Frankliniella occidentalis*, *J. Holacanthella duospinosa*, *K. Medauroidea extrudentata*, *L. Orchesella cincta*, *M. Pediculus humanus*, *N. Pieris rapae*, *O. Tribolium castaneum*. Orange-red, frequency distribution of gene duplicate bins with identical Ks values; light blue, WGD/segmental duplication event predicted by MSscanX; inlay, circos plot co-linear blocks. **Figure S2.** Histograms of sequence read coverage distribution (bins of 20 counts) among scaffolds of *Aethina tumida*'s genome assembly: 1000 random contigs (A) and contigs with co-linear regions (B). **Figure S3.** Scatter plots of putative gene duplicates (BlastP hits with e-value < 10<sup>-10</sup>) for species that contain at least one segmental duplication. A) *L. polyphemus*, b) *A. aegypti* c) *A. tumida*, d) *B. germanica*, e) *B. mori*, f) *C. felis*, g) *F. candida*. Co-linear blocks identified by MCScanX are indicated as red dots. Scale on horizontal axis in bp. **Figure S4.** Trace plots for the MCMC samples for the Holometabola data set with the IR prior. In black results for the full data set are shown (10,000 generations after 1000 generations as burn-in, showing every iterate), whereas the other transparent colors show three replicate chains for a random subset of 1000 gene families (20,000 generations after 1000 as burn-in, showing every second iterate). Duplication ( $\lambda$ ) and loss ( $\mu$ ) rates are shown on a log<sub>10</sub> scale, and subscripts denote branches of the species tree. **Figure S5.** Marginal posterior distributions for the MCMC samples for the Holometabola data set with the IR prior. Interpretation is as in Figure S4, but here we show the rates on the original scale. **Figure S6.** Marginal posterior distributions for retention rates ( $q$ ) of the five hypothetical WGD events marked along the Holometabola tree. The upper row shows results under the IR prior, whereas the lower row corresponds to results under the GBM (autocorrelated rates) prior (see [methods](#)). Note that the distributions under the GBM prior for the Lepidoptera, Coleoptera and Hymenoptera events are vanishingly small but are shown on the same scale as the upper row for the sake of comparison. **Figure S7.** A distinct mode for the parameters associated with the *C. felis* branch was observed in one of the chains under the IR prior for the Holometabola tree, indicating the possible problem of inefficient sampling of multimodal distributions in Whale. Results from three independent chains are shown in blue, orange and green respectively. (a & d) Marginal posterior distributions for the duplication ( $\lambda$ ) and retention ( $q$ ) rate associated with the *C. felis* branch for two chains. (b,c & e) Trace plots for duplication, 2 loss ( $\mu$ ) and retention rates associated with the *C. felis* branch for the same two chains. (f) Trace plot of the log likelihood for these chains. **Figure S8.** Posterior reconciliation probabilities of gene duplicates reconciled to the hypothetical *C. felis* (A) or Insecta (B) WGDs. The posterior reconciliation probability is calculated as the fraction that a particular clade is reconciled to the WGD node of interest in 1000 reconciled trees sampled from the posterior. Boxplots show the same data but grouped by clade size, showing for the *C. felis* WGD hypothesis a slight trend towards lower reconciliation probabilities for larger clades, whereas this trend is not observed for the putative Insecta event. **Figure S9.** Trace plots for the MCMC samples for the non-Holometabola data set with the IR prior. In black results for the full data set are shown (10,000 generations after 1000 generations as burn-in, showing every iterate), whereas the other transparent colors show three replicate chains for a random subset of 1000 gene families (20,000 generations after 1000 as burn-in, showing every second iterate). Duplication ( $\lambda$ ) and loss ( $\mu$ ) rates are shown on a log<sub>10</sub> scale and subscripts denote branches of the species tree. **Figure S10.** Marginal posterior distributions for the MCMC samples for the non-Holometabola data set with the IR prior. Interpretation is as in Figure S8 but here we show the rates on the original scale. **Figure S11.** Marginal posterior

distributions for retention rates ( $q$ ) of the seven hypothetical WGD events marked along the non-Holometabola tree. The upper row shows results under the IR prior, whereas the lower row corresponds to results under the GBM (autocorrelated rates) prior (see [methods](#)). Note that the distributions under the GBM prior for the Colembolla and Polyneoptera events are vanishingly small but are shown on the same scale as the upper row for the sake of comparison. **Table S1.** General specifications of species included in this study. Gene pairs, the number of gene pairs per hexapod species used as input for Ks calculation, co-linearity analysis and gene tree-species tree reconciliation analysis. Gene pairs with Ks values of 0 and higher than 5 were filtered out.

## Acknowledgements

We acknowledge three anonymous reviewers for their comments and suggestions that improved the quality of our study and its presentation.

## Authors' contributions

DR, KK, JE, and DK initiated, set up the experiments, and quality controlled hexapod genome sequences. KK, TS, JN, and DR performed and interpreted K<sub>s</sub> analyses and co-linear analyses using MCScanX. KK and AZ performed node weighing to correct for redundant estimates of old gene duplications as implemented in the wgd pipeline. AZ and YVdP performed and interpreted phylogenomic reconciliation analyses. DR, KK, JE, and DK provided the first drafts of the manuscript. All authors contributed to the writing and revising the manuscript. All authors read and approved the final manuscript.

## Funding

This work was supported by a PhD Fellowship of the Research Foundation—Flanders (FWO) (to AZ), and the Netherlands Science Foundation TTW (NWO-TTW) Open Technology Program project DrCOMICS project no. 15494 (to DR). KK is funded through the StartImpuls of NWA.

## Availability of data and materials

The datasets analyzed during the current study are available in the National Center for Biotechnology Information (NCBI) repository, [ncbi.nlm.nih.gov/genome](https://ncbi.nlm.nih.gov/genome). Genome sequences of *blattella*, *holacanthella*, and *medauroidea* are available from the i5K initiative and can be downloaded from <https://i5k.nal.usda.gov>. The genome sequence of *Campodea augens* is published in BioRxiv [67] and was kindly provided by the authors. The following accession numbers were used: *A. pisum* GCF\_000142985.2 [48], *A. aegypti* GCF\_002204515.2 [68], *A. tumida* GCF\_001937115.1 [69], *A. mellifera* GCF\_000002195.4 [70], *Athalia rosae* GCF\_000344095.1 [71], *Bemisia tabaci* GCF\_001854935.1 [72], *Blattella germanica* GCA\_000762945.2 [46], *B. mori* GCF\_000151625.1 [73], *C. augens* *campodea\_augens\_genome\_v1.0* [67], *C. felis* GCF\_003426905.1 [74], *D. melanogaster* GCF\_000001215.4 [75], *F. candida* *fcand\_genome.fa* (Collemبولomics.nl) [35], *Frankliniella occidentalis* GCF\_000697945.2 [76], *H. duospinosa* GCA\_002738285.1 [77], *Medauroidea extrudentata* GCA\_003012365.1 [78], *O. cincta* *ocinc\_genome.fa* (Collemبولomics.nl) [79], *P. humanus* GCF\_000006295.1 [80], *Pieris rapae* GCF\_001856805.1 [81], *T. castaneum* GCA\_000002335.3 [82], *Z. nevadensis* GCA\_000696155.1 [47], *L. polyphemus* GCF\_000517525.1 [83].

## 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

<sup>1</sup>Department of Ecological Science, Vrije Universiteit, De Boelelaan 1085, 1081HV Amsterdam, The Netherlands. <sup>2</sup>Keygene N.V, Agro Business Park 90, 6708 PW Wageningen, The Netherlands. <sup>3</sup>Center for Plant Systems Biology, VIB, B-9052 Ghent, Belgium. <sup>4</sup>Department of Plant Biotechnology and Bioinformatics, Ghent University, B-9052 Ghent, Belgium. <sup>5</sup>Department of Biochemistry, Genetics and Microbiology, Center for Microbial Ecology and Genomics, University of Pretoria, Pretoria 0028, South Africa. <sup>6</sup>Origins Center,

Nijenborgh 7, 9747AG Groningen, The Netherlands. <sup>7</sup>Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Sciencepark 904, 1090 GE Amsterdam, The Netherlands.

Received: 7 November 2019 Accepted: 6 May 2020

Published online: 27 May 2020

## References

- Ohno S, Wolf U, Atkin NB. Evolution from fish to mammals by gene duplication. *Hereditas*. 1967;59:169–87. <https://doi.org/10.1111/j.1601-5223.1968.tb02169.x>.
- Zhang J. Evolution by gene duplication: an update. *Trends Ecol Evol*. 2003;18:292–8. [https://doi.org/10.1016/S0169-5347\(03\)00033-8](https://doi.org/10.1016/S0169-5347(03)00033-8).
- De Smet R, Sabaghian E, Li Z, Saeys Y, Van de Peer Y. Coordinated functional divergence of genes after genome duplication in *Arabidopsis thaliana*. *Plant Cell*. 2017;29:2786–800. <https://doi.org/10.1105/TPC.17.00531>.
- Osborn TC, Chris Pires J, Birchler JA, Auger DL, Jeffery Chen Z, Lee H-S, et al. Understanding mechanisms of novel gene expression in polyploids. *Trends Genet*. 2003;19:141–7. [https://doi.org/10.1016/S0168-9525\(03\)00015-5](https://doi.org/10.1016/S0168-9525(03)00015-5).
- Crow KD, Wagner GP. What is the role of genome duplication in the evolution of complexity and diversity? *Mol Biol Evol*. 2006;23:887–92. <https://doi.org/10.1093/molbev/msj083>.
- Yao Y, Carretero-Paulet L, Van de Peer Y. Using digital organisms to study the evolutionary consequences of whole genome duplication and polyploidy. *PLoS One*. 2019;14:e0220257. <https://doi.org/10.1371/journal.pone.0220257>.
- Landis JB, Soltis DE, Li Z, Marx HE, Barker MS, Tank DC, et al. Impact of whole-genome duplication events on diversification rates in angiosperms. *Am J Bot*. 2018;105:348–63. <https://doi.org/10.1002/ajb2.1060>.
- Tank DC, Eastman JM, Pennell MW, Soltis PS, Soltis DE, Hinchliff CE, et al. Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *Source New Phytol*. 2015;207:454–67. <https://doi.org/10.2307/newphytologist.207.2.454>.
- Schranz ME, Mohammadin S, Edger PP. Ancient whole genome duplications, novelty and diversification: the WGD radiation lag-time model. *Curr Opin Plant Biol*. 2012;15:147–53. <https://doi.org/10.1016/j.cupb.2012.03.011>.
- Mayrose I, Zhan SH, Rothfels CJ, Magnuson-Ford K, Barker MS, Rieseberg LH, et al. Recently formed polyploid plants diversify at lower rates. *Science*. 2011;333:1257. <https://doi.org/10.1126/science.1207205>.
- Van de Peer Y, Maere S, Meyer A. The evolutionary significance of ancient genome duplications. *Nat Rev Genet*. 2009;10:725–32. <https://doi.org/10.1038/nrg2600>.
- Mable BK. 'Why polyploidy is rarer in animals than in plants': myths and mechanisms. *Biol J Linn Soc*. 2004;82:453–66. <https://doi.org/10.1111/j.1095-8312.2004.00332.x>.
- Muller HJ. Why polyploidy is rarer in animals than in plants. *Am Nat*. 1925;59:346–53. <https://doi.org/10.1086/280047>.
- Orr HA. 'Why polyploidy is rarer in animals than in plants'; revisited. *Am Nat*. 1990;136:759–70. <https://doi.org/10.1086/285130>.
- Hallinan NM, Lindberg DR. Comparative analysis of chromosome counts infers three Paleopolyploidies in the Mollusca. *Genome Biol Evol*. 2011;3:1150–63. <https://doi.org/10.1093/gbe/evr087>.
- Kenny NJ, Chan KW, Nong W, Qu Z, Maeso I, Yip HY, et al. Ancestral whole-genome duplication in the marine chelicerate horseshoe crabs. *Heredity*. 2016;116:190–9. <https://doi.org/10.1038/hdy.2015.89>.
- Clarke TH, Garb JE, Hayashi CY, Arensburger P, Ayoub NA. Spider transcriptomes identify ancient large-scale gene duplication event potentially important in silk gland evolution. *Genome Biol Evol*. 2015;7:1856–70. <https://doi.org/10.1093/gbe/evw110>.
- Li Z, Tiley GP, Galuska SR, Reardon CR, Kidder TJ, Rundell RJ, et al. Multiple large-scale gene and genome duplications during the evolution of hexapods. *Proc Natl Acad Sci U S A*. 2018;115:4713–8. <https://doi.org/10.1073/pnas.1710791115>.
- Schwager EE, Sharma PP, Clarke T, Leite DJ, Wierschin T, Pechmann M, et al. The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. *BMC Biol*. 2017;15:62. <https://doi.org/10.1186/s12915-017-0399-x>.
- Leite DJ, McGregor AP. Arthropod evolution and development: recent insights from chelicerates and myriapods. *Curr Opin Genet Dev*. 2016;39:93–100. <https://doi.org/10.1016/j.cde.2016.06.002>.
- Zwaenepoel A, Li Z, Lohaus R, Van de Peer Y. Finding evidence for whole genome duplications: a reappraisal. *Mol Plant*. 2019;12:133–6. <https://doi.org/10.1016/j.molp.2018.12.019>.
- Zwaenepoel A, Van de Peer Y. Inference of ancient whole-genome duplications and the evolution of gene duplication and loss rates. *Mol Biol Evol*. 2019;36:1384–404. <https://doi.org/10.1093/molbev/msz088>.
- Nakatani Y, McLysaght A. Macrosynteny analysis shows the absence of ancient whole-genome duplication in lepidopteran insects. *Proc Natl Acad Sci U S A*. 2019;116:1816–8. <https://doi.org/10.1073/pnas.1817937116>.
- Vanneste K, Van de Peer Y, Maere S. Inference of genome duplications from age distributions revisited. *Mol Biol Evol*. 2013;30:177–90. <https://doi.org/10.1093/molbev/mss214>.
- Tiley GP, Barker MS, Burleigh JG. Assessing the performance of Ks plots for detecting ancient whole genome duplications. *Genome Biol Evol*. 2018;10:2882–98. <https://doi.org/10.1093/gbe/evy200>.
- Jaillon O, Aury J-M, Brunet F, Petit J-L, Stange-Thomann N, Mauceli E, et al. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature*. 2004;431:946–57. <https://doi.org/10.1038/nature03025>.
- Kellis M, Birren BW, Lander ES. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*. 2004;428:617–24. <https://doi.org/10.1038/nature02424>.
- Nakatani Y, McLysaght A. Genomes as documents of evolutionary history: a probabilistic macrosynteny model for the reconstruction of ancestral genomes. *Bioinformatics*. 2017;33:i369–78. <https://doi.org/10.1093/bioinformatics/btx259>.
- Van de Peer Y. Computational approaches to unveiling ancient genome duplications. *Nat Rev Genet*. 2004;5:752–63. <https://doi.org/10.1038/nrg1449>.
- Schrader L, Schmitz J. The impact of transposable elements in adaptive evolution. *Mol Ecol*. 2019;28:1537–49. <https://doi.org/10.1111/mec.14794>.
- Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, et al. Ancestral polyploidy in seed plants and angiosperms. *Nature*. 2011;473:97–100.
- Li Z, Baniaga AE, Sessa EB, Scacitelli M, Graham SW, Rieseberg LH, et al. Early genome duplications in conifers and other seed plants. *Sci Adv*. 2015;1:e1501084. <https://doi.org/10.1126/sciadv.1501084>.
- Ruprecht C, Lohaus R, Vanneste K, Mutwil M, Nikoloski Z, Van de Peer Y, et al. Revisiting ancestral polyploidy in plants. *Sci Adv*. 2017;3:e1603195. <https://doi.org/10.1126/sciadv.1603195>.
- Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. *Science*. 2000;290:1151–5. <https://doi.org/10.1126/science.290.5494.1151>.
- Faddeeva-Vakhrusheva A, Kraaijeveld K, Derks MFL, Anvar SY, Agamennone V, Suring W, et al. Coping with living in the soil: the genome of the parthenogenetic springtail *Folsomia candida*. *BMC Genomics*. 2017;18:493–506. <https://doi.org/10.1186/s12864-017-3852-x>.
- Mahmudi O, Sjöstrand J, Sennblad B, Lagergren J. Genome-wide probabilistic reconciliation analysis across vertebrates. *BMC Bioinformatics*. 2013;14(Suppl 15):S10. <https://doi.org/10.1186/1471-2105-14-S15-S10>.
- Rabier C-E, Ta T, Ané C. Detecting and locating whole genome duplications on a phylogeny: a probabilistic approach. *Mol Biol Evol*. 2014;31:750–62. <https://doi.org/10.1093/molbev/mst263>.
- Hahn MW, Han MV, Han S-G. Gene family evolution across 12 *Drosophila* genomes. *PLoS Genet*. 2007;3:e197. <https://doi.org/10.1371/journal.pgen.0030197>.
- Taylor JS, Raes J. Duplication and divergence: the evolution of new genes and old ideas. *Annu Rev Genet*. 2004;38:615–43.
- Simillion C, Vandepoel K, Van Montagu MCE, Zabeau M, Van de Peer Y. The hidden duplication past of *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A*. 2002;99:13627–32. <https://doi.org/10.1073/pnas.212522399>.
- Mandáková T, Lysak MA. Post-polyploid diploidization and diversification through dysploid changes. *Curr Opin Plant Biol*. 2018;42:55–65. <https://doi.org/10.1016/j.cupb.2018.03.001>.
- Lysak MA, Berr A, Pecinka A, Schmidt R, McBrean K, Schubert I. Mechanisms of chromosome number reduction in *Arabidopsis thaliana* and related Brassicaceae species. *Proc Natl Acad Sci U S A*. 2006;103:5224–9. <https://doi.org/10.1073/pnas.0510791103>.
- Liu D, Hunt M, Tsai IJ. Inferring synteny between genome assemblies: a systematic evaluation. *BMC Bioinformatics*. 2018;19:26. <https://doi.org/10.1186/s12859-018-2026-4>.
- Lyons E, Pedersen B, Kane J, Freeling M. The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates the rosids. <https://doi.org/10.1007/s12042-008-9017-y>.

45. Pace RM, Grbić M, Nagy LM. Composition and genomic organization of arthropod Hox clusters. *Evodevo*. 2016;7:1–11. <https://doi.org/10.1186/s13227-016-0048-4>.
46. Harrison MC, Jongepier E, Robertson HM, Arning N, Bitard-Feildel T, Chao H, et al. Hemimetabolous genomes reveal molecular basis of termite eusociality. *Nat Ecol Evol*. 2018;2:557–66. <https://doi.org/10.1038/s41559-017-0459-1>.
47. Terrapon N, Li C, Robertson HM, Ji L, Meng X, Booth W, et al. Molecular traces of alternative social organization in a termite genome. *Nat Commun*. 2014;5:3636. <https://doi.org/10.1038/ncomms4636>.
48. Richards S, Gibbs RA, Gerardo NM, Moran N, Nakabachi A, Stern D, et al. Genome sequence of the pea aphid *Acyrtosiphon pisum*. *PLoS Biol*. 2010;8:e1000313. <https://doi.org/10.1371/journal.pbio.1000313>.
49. Duboule D. The rise and fall of Hox gene clusters. *Development*. 2007;134:2549–60.
50. Fiston-Lavier A-S, Anxolabehere D, Quesneville H. A model of segmental duplication formation in *Drosophila melanogaster*. *Genome Res*. 2007;17:1458–70. <https://doi.org/10.1101/gr.6208307>.
51. Bailey JA, Liu G, Eichler EE. An Alu transposition model for the origin and expansion of human segmental duplications. *Am J Hum Genet*. 2003;73:823–34. <https://doi.org/10.1086/378594>.
52. Schrader L, Kim JW, Ence D, Zimin A, Klein A, Wyschetzki K, et al. ARTICLE transposable element islands facilitate adaptation to novel environments in an invasive species. *Nat Commun*. 2014;5 <https://doi.org/10.1038/ncomms6495>.
53. Petersen M, Armisén D, Gibbs RA, Hering L, Khila A, Mayer G, et al. Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. *BMC Evol Biol*. 2019;19:11. <https://doi.org/10.1186/s12862-018-1324-9>.
54. Le Rouzic A, Payen T, Hua-Van A. Reconstructing the evolutionary history of transposable elements. *Genome Biol Evol*. 2013;5:77–86. <https://doi.org/10.1093/gbe/evs130>.
55. Zwaenepoel A, Van de Peer Y. Wgd—simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics*. 2019;35:2153–5. <https://doi.org/10.1093/bioinformatics/bty915>.
56. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*. 2002;30:1575–84. <https://doi.org/10.1093/nar/30.7.1575>.
57. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013;30:772–80. <https://doi.org/10.1093/molbev/mst010>.
58. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 2007;24:1586–91. <https://doi.org/10.1093/molbev/msm088>.
59. Price MN, Dehal PS, Arkin AP. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One*. 2010;5:e9490. <https://doi.org/10.1371/journal.pone.0009490>.
60. Wang Y, Tang H, DeBarry JD, Tan X, Li J, Wang X, et al. MScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res*. 2012;40:2–14. <https://doi.org/10.1093/nar/gkr1293>.
61. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. *Genome Res*. 2009;19:1639–45. <https://doi.org/10.1101/gr.092759.109>.
62. Kumar S, Stecher G, Suleski M, Heddes SB. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol Biol Evol*. 2017;34:1812–9.
63. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol*. 2015;16:157. <https://doi.org/10.1186/s13059-015-0721-2>.
64. Whelan S, Irisarri I, Burki F. PREQUAL: detecting non-homologous characters in sets of unaligned homologous sequences. *Bioinformatics*. 2018; <https://doi.org/10.1093/bioinformatics/bty448>.
65. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol*. 2012;61:539–42. <https://doi.org/10.1093/sysbio/sys029>.
66. Szöllősi GJ, Rosikiewicz W, Bousseau B, Tannier E, Daubin V. Efficient exploration of the space of reconciled gene trees. *Syst Biol*. 2013;62:901–12. <https://doi.org/10.1093/sysbio/syt054>.
67. Manni M, Simao FA, Robertson HM, Gabaglio MA, Waterhouse RM, Misof B, et al. The genome of the blind soil-dwelling and ancestrally wingless dipluran *Campodea augens*, a key reference hexapod for studying the emergence of insect innovations. Preprint at <https://www.biorxiv.org/content/10.1101/585695v3>. <https://doi.org/10.1101/585695>.
68. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*. 2017;356:92–5.
69. Evans JD, McKenna D, Scully E, Cook SC, Dainat B, Egekwu N, et al. Genome of the small hive beetle (*Aethina tumida*, Coleoptera: Nitidulidae), a worldwide parasite of social bee colonies, provides insights into detoxification and herbivory. *Giga Sci*. 2018;7:1–16. <https://doi.org/10.1093/gigascience/giy138>.
70. Weinstock GM, Robinson GE, Gibbs RA, Worley KC, Evans JD, Maleszka R, et al. Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature*. 2006;443:931–49.
71. Mine S, Sumitani M, Aoki F, Hatakeyama M, Suzuki MG. Identification and functional characterization of the sex-determining gene doublesex in the sawfly, *Athalia rosae* (Hymenoptera: Tenthredinidae). *Appl Entomol Zool*. 2017;52:497–509.
72. Chen W, Hasegawa DK, Kaur N, Kliot A, Pinheiro PV, Luan J, et al. The draft genome of whitefly *Bemisia tabaci* MEAM1, a global crop pest, provides novel insights into virus transmission, host adaptation, and insecticide resistance. *BMC Biol*. 2016;14:110. <https://doi.org/10.1186/s12915-016-0321-y>.
73. Xia Q, Wang J, Zhou Z, Li R, Fan W, Cheng D, et al. The genome of a lepidopteran model insect, the silkworm *Bombyx mori*. *Insect Biochem Mol Biol*. 2008;38:1036–45.
74. Driscoll TP, Verhoeve VI, Gillespie JJ, Johnston JS, Guillotte ML, Rennoll-Bankert KE, et al. Cat fleas in flux: rampant gene duplication, genome size plasticity, and paradoxical *Wolbachia* infection. Preprint at <https://www.biorxiv.org/content/10.1101/2020.04.14.038018v1>. <https://doi.org/10.1101/2020.04.14.038018>.
75. Hoskins RA, Carlson JW, Wan KH, Park S, Mendez I, Galle SE, et al. The release 6 reference sequence of the *Drosophila melanogaster* genome. *Genome Res*. 2015;25:445–58.
76. Rotenberg D, Baumann AA, Ben-Mahmoud S, Christiaens O, Dermauw W, Ioannidis P, et al. Genome-enabled insights into the biology of thrips as crop pests. Preprint at <https://www.biorxiv.org/content/10.1101/2020.02.12.941716v1.full>. <https://doi.org/10.1101/2020.02.12.941716>.
77. Wu C, Jordan MD, Newcomb RD, Gemmill NJ, Bank S, Meusemann K, et al. Analysis of the genome of the New Zealand giant collembolan (*Holacanthella duosipinosa*) sheds light on hexapod evolution. *BMC Genomics*. 2017;18:1–19. <https://doi.org/10.1186/s12864-017-4197-1>.
78. Brand P, Lin W, Johnson BR. The draft genome of the invasive walking stick, *Medauroidea extradentata*, reveals extensive lineage-specific gene family expansions of cell wall degrading enzymes in Phasmatodea. *G3 genes, genomes*. *Genet*. 2018;8:1403–8.
79. Faddeeva-Vakhrusheva A, Derks MFL, Anvar SY, Agamenone V, Suring W, Smit S, et al. Gene family evolution reflects adaptation to soil environmental stressors in the genome of the collembolan *Orchesella cincta*. *Genome Biol Evol*. 2016;8:2106–17. <https://doi.org/10.1093/gbe/eww134>.
80. Johnson JS, Yoon KS, Strycharz JP, Pittendrigh BR, Clark JM. Body lice and head lice (Anoplura: Pediculidae) have the smallest genomes of any hemimetabolous insect reported to date. *J Med Entomol*. 2007;44:1009–12. <https://doi.org/10.1093/jmedent/44.6.1009>.
81. Grishin NV, Shen J, Cong Q, Kinch LN, Borek D, Otwinowski Z. Complete genome of *Pieris rapae*, a resilient alien, a cabbage pest, and a source of anti-cancer proteins. *F1000Research*. 2016;5:2631. <https://doi.org/10.12688/f1000research.9765.1>.
82. Kim HS, Murphy T, Xia J, Caragea D, Park Y, Beeman RW, et al. BeetleBase in 2010: revisions to provide comprehensive genomic information for *Tribolium castaneum*. *Nucleic Acids Res*. 2009;38:D437–42. <https://doi.org/10.1093/nar/gkp807>.
83. Battelle B-A, Ryan JF, Kempler KE, Saraf SR, Marten CE, Warren WC, et al. Opsin repertoire and expression patterns in horseshoe crabs: evidence from the genome of *Limulus polyphemus* (Arthropoda: Chelicerata). *Genome Biol Evol*. 2016;8:1571–89. <https://doi.org/10.1093/gbe/eww100>.

## Publisher's Note

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