Whole-Proteome Tree of Arthropods: An "alignment-free" phylogeny of proteome "books"

JaeJin Choi*, Byung-Ju Kim^ and Sung-Hou Kim*^#

*Department of Chemistry and Center for Computational Biology University of California, Berkeley, CA, 94720, USA, and Molecular Biophysics & Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

^Human Genome Research Center, Incheon National University, Incheon, Republic of Korea (South Korea)

ORCID: JJ Choi (0000-0002-2860-9606), BJ Kim (0000-0002-8906-5622), SH Kim (0000-0002-7293-5994)

#Corresponding author: Sung-Hou Kim; sunghou@berkeley.edu; 510-708-2564

Abstract

Background: An "organism tree" of a group of extant-organisms can be considered as a conceptual tree to capture a simplified narrative of the evolutionary course among the organisms. Due to the difficulties of whole-genome sequencing for many organisms, the most common approach has been to construct a "gene tree" by selecting a group of genes common among the organisms, "align" each gene family and estimate evolutionary distances. Despite broad acceptance of the gene trees as the surrogates for the organism trees, there are important limitations and confounding issues with the approach. During last decades, whole-genome sequences of many extant-arthropods became available, providing an opportunity to construct a "whole-proteome tree" of the arthropods, the largest and most species-diverse group of all living animals.

Results: An "alignment-free" whole-proteome tree of the arthropods shows that (a) the demographic grouping-pattern is similar to those in the gene trees, but there are notable differences in the branching orders of the groups and the sisterhood relationships between pairs of the groups; and (b) almost all the "founders" of the groups have *emerged* in an "explosive burst" near the root of the tree.

Conclusion: Since the *whole*-proteome sequence of an organism can be considered as a "book" of amino-acid alphabets, a tree of the books can be constructed, without alignment of sequences, using a text analysis method of Information Theory, which allows comparing the information content of whole-proteomes. Such tree provides another view-point to consider in telling the narrative of kinship among the arthropods.

Keywords: Whole genome tree, Organism tree, Alignment-free method, Feature Frequency Profile, Jensen-Shannon divergence, Evolutionary progression scale, Arthropod burst, Divergence of genomic information content

Background

For several decades, due to the technical difficulties of whole genome sequencing, the most practically feasible and common approach to construct an evolutionary tree of a group of extant organisms has been to construct a "gene tree", where a group of genes common among the organisms are selected, "align" highly homologous sequence-regions for each gene family, and quantify the divergence of certain characteristics, e.g. point substitution rates, to estimate evolutionary distances among the genes. Such "alignment-based" gene trees may represent the phylogeny of the selected genes, and may reveal, depending on the selected genes, various aspects, but not full aspects of the "organism tree", because the aligned regions of the selected genes account, in general, for a very small fraction of the genome coding for *all* genes. Furthermore, this approach makes a set of *assumptions* about evolutionary process such as: (a) the collection of such highly homologous sequence-regions contains enough information to represent each organism and (b) one type of mutation, e.g., point substitutions, within the aligned regions account for the most important mutational events of evolution. In addition, there are several other limitations and confounding issues with the approach [1, 2]. Strictly speaking, such gene trees may reveal the evolution of the selected genes, but not of the organisms [1]. "Alignment-free" approach [2, 3] applied to whole-proteome sequence can reduce some of the limitations and the confounding issues of the alignment-based approach.

Recent availability of a large number of whole-genome sequences prompted us to consider whether an alternative approach, such as alignment-free methods applied to whole-proteome, may provide information about the similarities and differences between the two types of trees, since the basic assumptions used are very different. In alignment-free method, whole-genome information can be used under the assumptions that: (a) all or most of whole genome sequence of an organism represents the organism and (b) a combination of *all types of mutations*, such as point substitution, insertion/deletion of various length, recombination, duplication, transfer or swapping of gene, domain or motif, etc., contribute to the evolution of the organism. At present, complete or near complete sequences of many arthropods (over 134 species as of 2020) of broad diversity are available in a public database [4], which allows one to construct a whole-proteome tree of arthropods (ToAr).

We have constructed a whole-proteome ToAr, which turned out to be more stable *topologically* than those of whole genome or whole transcriptome ToAr (see

Supplementary Information Fig. S1), constructed using Information-Theory based Feature Frequency Profile method [5], an "alignment-free" method, on the extant arthropods. Here we describe the main features of the ToAr on grouping and kinship among the groups.

Results

To compare the gene trees with the whole-proteome tree of arthropods we chose two recent and very comprehensively-analyzed gene trees: The first one is the recent "alignment-based" gene tree of 144 insect taxa based on 1,478 single-copy protein-coding nuclear genes [6], where the tree was "inferred through a maximum-likelihood analysis of 413,459 amino acid sites divided into multiple metapartitions, branch lengths were optimized and node ages estimated from 1,050,000 trees sampled from trees separately generated for 105 partitions that included all taxa". The authors also suggest the relationship between the geological origin of insect clades and major geological and biological events in chronological time scale. The second is the gene tree for 76 arthropod taxa [7], where the tree was "inferred from 569 to 4,097 single copy protein-coding genes among the six multi-species orders, crustaceans, and non-spider chelicerates and 150 single copy genes for the orders represented by a single species and deeper nodes". Both gene trees agree on the branching order of the Order groups, and on similar time of the emergence of the founders of the groups in chronological time scale.

In comparing our whole-proteome tree to these two gene trees, we focus on two aspects: grouping patterns and phylogeny of the groups. For the former, we "cluster" our study population by several *unsupervised* clustering algorithms using *only* the evolutionary distances estimated from the "divergence" distances among whole-proteome characters (see **Construction of whole-proteome Tree of Arthropods in Materials and Methods**) with no assumption of specific evolutionary models; we then ask whether the "clustering pattern" is similar to the "clading pattern" in the gene trees and in our whole-proteome tree. For the phylogeny of the groups, we compare the order of branching of the groups and their emergence on the evolutionary progression scale (see "**Cumulative Genomic Divergence(CGD)" in "Evolutionary Progression Scale" in Materials and Methods)**, which are derived from the evolutionary distance estimation *and* specific evolutionary models assumed.

A. Demographic grouping by clustering and clading

Clustering: We have tested several unsupervised clustering algorithms, such as Principal Component Analysis (PCA), Multi-Dimensional Scaling (MDS), and t-Distributed Stochastic Neighbor Embedding (t-SNE) [8, 9], all of which are based only on the evolutionary distances, estimated by the divergence of whole-proteome sequences among all pairs of the arthropods (see **Construction of whole-proteome Tree of Arthropods in Materials and Methods)**. Of these, t-SNE and classical PCA showed the clustering pattern most compatible to the current grouping of arthropods based on common or scientific names. Figure 1, a t-SNE clustering, shows about 17 clusters of the extant arthropods in this study that are most stable under a wide range of "perplexity", the adjustable parameter of t-SNE, which is roughly related to the "looseness" of each cluster. Another clustering by PCA (see Supplemental information, Fig. S1) also shows about similar number of clusters, although some of them are not as well resolved as in t-SNE. This is not expected, because the assumptions and algorithms in t-SNE and PCA are completely different.

Clading: Figure 2 shows the topology of whole-proteome ToAr, constructed using neighbor-joining method implemented in BIONJ [10, 11], where we use the divergence distances of whole-proteome sequences as the estimates for the evolutionary distances (see **Construction of whole-proteome Tree of Arthropods** in **Materials and Methods)**, combined with an evolutionary model of maximum parsimony (minimum evolution) in a way that the "correct neighbors" to be joined are those that minimize the total sum of branch-lengths at each stage of step-wise joining of neighbors starting from a star-like tree [10]. The tree shows that all 17 clusters in Fig.1 can be identified among the clades in the ToAr.

Robustness of grouping: Furthermore, the grouping pattern by clustering *and* clading based on whole-proteome information also agrees well with those of the gene trees except for Hemiptera group (see **Comparison with recent gene trees** in **Discussions**).

Thus, it is surprising that the demographic grouping pattern is robust regardless not only of the information type (morphology, select gene-characters or whole proteome characters), but also of the methods (clustering or clading) used in grouping the extant organisms. However, not surprisingly, there are significant differences from the gene trees in branch-length and branching order of the groups, which are used to predict the position, in the evolutionary progression scale, at which the "founders" of each group emerged (see below).

B. Emergence of the "Founders" of all major groups

For the following results we define "Cumulative Genomic Divergence (CGD)" for an internal node of the ToAr as the cumulative scaled-branch-length from the tree root to the node (see "**Cumulative Genomic Divergence(CGD)" in "Evolutionary Progression Scale" in Materials and Methods)** to represent the extent of "Evolutionary Progression", where the progression is scaled such that the root of ToAr is set at CGD=0 (see **Outgroup** in **Discussions**) and the leaf positions of the extant organisms at CGD near 100.

"Arthropodal burst" near the root of ToAr: Figures 3 shows the whole-proteome ToAr with CGD values. It reveals that the "founders" (for definition, see Supplemental information, Fig. S3) of all major groups (at Subphyla and Order levels) emerged in a burst within a short evolutionary progression span between CGD of 1.6 and 5.8 (marked by a small red arc), near the root of the tree in Figs. 3, 4 and 5. This burst is reminiscent of the "deep burst" of the founders of all kingdoms of Life in the whole-proteome tree of Life [12].

Order of emergence of the "founders" for all major groups: Figures 4 and 5 show a series of staged emergence of the founders of all major groups within the burst. The first groups to emerge (branch out) in a narrow range of CGD values of 1.6 and 1.9 are the founders of Subphylum Crustacea and Chelicerata (arachnids and Atlantic horseshoe crab). The founders of the next groups to emerge are Hemiptera-A group (aphids and a psyllid) at CGD of 3.7, followed by the founders of Diptera group at CGD of 4.1. The founders of the remaining five Order-level groups (Lepidoptera, Hemiptera-B (bugs, a planthopper and a whitefly), Coleoptera, Blattodea + a thrips, and Hymenoptera groups) emerged at CGD of 4.4, 4.8, 5.2, 5.8, and 5.8, respectively.

Discussions

Similarities in grouping patterns: It is surprising that the grouping patterns at Order level between the gene trees [6, 7] and our whole proteome tree are very similar (see below for one notable exception of Hemiptera) despite the facts that the types of input data (multiply aligned regions of selected genes vs. whole proteome) and grouping methods (clustering vs. clading) are different. A possible implication

is that each group members evolved, during the most of the evolutionary progression, largely "isolated" within the group without genomic intermixing between the groups, thus, resulting in each group having much smaller genomic variation within the group than between the groups. This interpretation is consistent with our whole-proteome tree, where the founders of all Order groups emerged in a burst near the root of the tree, followed by a long gradual evolution spanning the most of the evolutionary progression scale.

Dissimilarities in branching orders and branch-lengths: It is not surprising that the branching orders and branch-lengths are not similar, because the assumptions under which the estimations for evolutionary distances among the organisms are very different: in the gene trees, the distances are calculated for the aligned portions of the selected genes for, e.g., point-substitutional mutations, while in our tree they are done for all proteins in the whole proteome for all types of mutations.

Evolutionary progression scale vs. chronological time scale: "Chronological time scale" is linear, but the "evolutionary progression scale" is most likely non-linear, because any catastrophic events on earth may accelerate, decelerate or diversify the evolutionary progression. However, the direction of arrows in both scales are the same, suggesting that the two scales need to be calibrated when sufficient fossils and other independent record become available (see "**Cumulative Genomic Divergence (CGD)" in "Evolutionary Progression Scale" in Materials and Methods)**. Meanwhile, we use the evolutionary progression scale to compare the order of emergence of the founders of various major groups under the assumption that the scale can be calculated by the whole proteomic information divergence.

Emergence of the founders of major groups: While cognizant of the difference and similarity of the two scales, the most dramatic difference is observed in the span of the scales when the founders of all major groups at Order level emerged. In the gene trees, the founders of all the groups at Order level emerged gradually during a long chronological time of about 350 Myrs corresponding roughly 60% of about 570 Myrs between the tree root to the extant arthropods (Fig. 1 of reference 6) or about 210 Myrs corresponding to roughly 37% of the same full chronological scale (Fig. 2 of reference 7). In drastic contrast, the founders of all the major groups in our tree emerged within about 4% of the full evolutionary progression scale in a sudden burst ("Arthropodal burst": see Figs. 3 and 4) near the root of the ToAr. This observation has an important evolutionary implication in constructing the narrative for the birth of the arthropods.

On notable differences in grouping and phylogenic positions: As mentioned earlier all grouping and most of the order of emergence of the groups at Order and Subphylum levels agree between the two types of trees except some notable differences in Hemiptera and Blattodea as discussed below. These differences may get resolved once the whole genome sequences of more diverse members of the two groups become available.

Hemiptera: In the gene trees [6, 7], the bugs, cicadas, and aphids form Hemiptera clade. But, in our ToAr as well as in t-SNE (Fig. 1) and PCA clustering plots (Supplemental Information, Fig. S1), Hemiptera is divided into two separate clades, which we call Hemiptera-A (aphids and a psyllid) and Hemiptera-B (planthopper, whitefly, stink bug, bed bug, cat flea, ash borer). The former is consisting of two of the most "primitive" Hemiptera feeding passively on mostly one plant species or its closely related plants, is at the basal position of *all* Hexapoda and the latter is at basal or sister position to a larger clade consisting of Coleoptera, Blattodea and Hymenoptera (see Figs. 3 and 5 for their phylogeny and Fig. 1 for clustering). Our clustering results (which are based solely on the whole proteome sequences, and not constrained by (a) the assumptions and related confounding issues associated with algorithms of tree building under given evolutionary models or (b) any or combination of particular phenotypes, such as metamorphosis, eusociality, or passive/active defensive survival modes etc.), show that the difference in the *whole-proteome information content*s between the two clades is greater than those within each clade.

Blattodea: Two termites (Blattodea) and one thrips (Thysanoptera), both eusocial and hemimetabolous, form a clade in our ToAr and the clade is sister (or basal) to Hymenoptera group, which is also eusocial but holometabolous (see Fig. 5). However, in the gene trees, Blattodea (cockroaches and termites, which are eusocial and hemimetabolous), as a member of a larger clade Polyneoptera [6] or as a separate clade [7], is basal or sister to all other Order groups of Insecta, which are mostly non-social and hemi- or holo-metabolous [7]. This is in contrast to what we observe in our ToAr, where non-social Hemiptera-A, is the basal group of all Hexapoda, consistent with the notion that eusociality of Blattodea is likely to be a trait acquired late in arthropod evolution.

Outgroup: As was described in our earlier work on whole-proteome trees [12, 13, 14] we constructed the proteome sequence of an "artificial (faux) arthropod" to use as an outgroup member of a ToAr by "shuffling" [15, 16] the alphabets of the whole proteome sequence of an organism in the study group. We used two such artificial arthropods (named R28612 and r12957) to form the outgroup for the study. Each has the same size and amino acid composition of an extant arthropod, but does not have information to life.

Conclusions

A whole-proteome tree constructed using "alignment-free" approach is based on a set of assumptions and methods that are fundamentally different from those used in the most commonly accepted "alignment-based" gene trees. Our tree provides another view-point to consider in telling the narrative of kinship among the arthropods. Some of the significant differences between the two types of trees of extant arthropods may have important implications in constructing a narrative of the evolutionary kinship among arthropods. Our tree raises the following debatable issues:

1. The founders of *all* order-level groups of arthropods emerged in a "burst" rather than through a long stretch of chronological time as suggested in gene trees;
2. Hemiptera-A, consisting of some of the most "primitive" Hemiptera, may be the basal group of all members of Insecta, rather than eusocial Blattodea as suggested in gene trees;
3. Blattodea may be sister to Hymenoptera, a eusocial group.

Methods

Aim: Construct a "whole-proteome tree" to provide another view-point of building a narrative for the evolution of the extant arthropods.

Design: 1) To treat each whole proteome-sequence of an arthropod as a "book" of the amino acid alphabets; 2) To compare the information content among all books of the extant arthropods for which whole genome sequences are available by

calculating the divergence of the information content using a text comparison method of Information Theory [17, 18, 19]; 3) To construct the whole proteome tree of the arthropods to estimate a cumulative genomic divergence as the evolutionary distance for each internal node of the tree; 4) To compare the whole proteome tree with recent gene trees [6, 7] in grouping and the kinship among the groups.

Sources and selection of proteome sequences

We downloaded the proteome sequences for 134 arthropods from NCBI RefSeq DB using NCBI FTP depository [4; as of December 2019]. Protein sequences derived from all organelles were excluded from this study. Also excluded from our study are those derived from whole genome sequences assembled with "low" completeness based on two criteria: (a) the genome assembly level indicated by NCBI as "contig" or lower (i.e. we selected those with the assembly levels of 'scaffold', 'chromosome' or 'complete genome'), and (b) the proteome size smaller than the smallest proteome size among highly assembled arthropod genomes (*Anopheles gambiae str. PEST* with 14,089 proteins at "chromosome" assembly level; TaxID 180454).

All taxonomic names and their taxon identifiers (TaxIDs) of the organisms in this study are from NCBI taxonomy database, and listed in Supplementary Information, Dataset S1, where "N/A" indicates an unassigned taxonomic order.

Construction of whole-proteome Tree of Arthropods

Based on our earlier experiments of constructing whole-proteome trees of prokaryotes [13], fungi [14] and all life forms [12] by Feature Frequency Profile (FFP) method, following choices have been made to obtain a topologically stable whole proteome ToAr of maximum parsimony (minimum evolution) by BIONJ [10, 11]: a) Among three types of genomic information (DNA sequence of the whole genome, RNA sequence of whole transcriptome and amino acid sequence of whole proteome) whole-proteome trees are most "topologically stable" as estimated by Robinson-Foulds metric [20] at respective "optimal Feature-length" (see Supplementary Information, Fig. S2 ; b) For FFP as the "descriptor" of the whole proteome of each organism, the optimal Feature-length is about 10 amino-acid string (see also Supplementary Information, Fig. S2); and 3) Jensen-Shannon Divergence (JSD) [17] is an appropriate measure of "divergence of information content" or distance of dissimilarity between two whole-proteome descriptors for constructing the distance matrix of BIONJ. It is important to note that such FFP of a whole-proteome sequence of an organism has all the information necessary to reconstruct the original whole proteome sequence.

"Cumulative Genomic Divergence (CGD)" and "Evolutionary progression scale"

In Information Theory, the Jensen-Shannon Divergence (JSD), bound between zero and one, is commonly used as a measure of the dissimilarity between two probability distribution of informational features. The FFP as the descriptor for a linear sequence information of the whole proteome of an organism is such a probability distribution. Thus, a JSD value of two FFPs, used as a measure of the divergent distance between two proteome sequences, is also bound between 0 and 1, corresponding to the JSD value between two FFPs of identical whole proteome sequences and two completely different whole proteome sequences, respectively. (For easy visibility, we multiply the branch-lengths in iTOL [21] by 200). Any whole proteome-sequence "dissimilarity" between two extant organisms accumulated during the evolution can be considered as caused by changes of, ultimately, genomic sequences due to point substitutions, indels, inversion, recombination, loss/gain of genes, etc. as well as other unknown mechanisms, and they will bring JSD somewhere between 0.0 and 1.0 depending on the degree of the sequence divergence.

In this study the collection of the JSDs for all pairs of the study organisms plus 2 out-group members (see **Outgroup** in **Discussions**) constitutes the "distance matrix" for BIONJ [10, 11]. Since all the branch-lengths are derived from the JSD values, the cumulative branch-length of an internal node, which we call "cumulative genomic divergence (CGD)" of the node along the presumed evolutionary lineage, can be considered as the point of evolutionary stage reached by the node on an "evolutionary progression scale". For convenience of assigning the nodes on the progression scale, CGDs are scaled such that the CGD value at the root node of ToAr is set to zero and the leaf nodes of the extant organisms to come out around 100, corresponding to the fully evolved genomic states of the organisms, which we define as the end point of the "evolutionary progression scale" for the organisms at present (see Fig. 3).

Clustering methods

We use two unsupervised methods to observe the clustering patterns based solely on whole proteome sequences: Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) [8, 9]. Both are dimensional reduction methods, but with strength and weakness for our purposes. which help to visualize any clustering pattern in the data distribution. Both are based only on the evolutionary distances, estimated by the divergence of whole-proteome sequences

among all pairs of the study arthropods. In PCA, the distances within a cluster as well as between two clusters are quantitative, thus, two close clusters nearby may not resolve well. In t-SNE, which applies Machine Learning to emphasize the resolution of nearby clusters, but the inter-cluster distances are ignored, thus, not quantitative.

Declarations

Acknowledgement: We gratefully acknowledge the comments and advices from the arthropod experts at University of California, Berkeley, CA., Profs. Kipling Will and Peter Oboyski, and the constructive suggestions on our application of Information Theory and Unsupervised Clustering methods from Prof. Se-Ran Jun of College of Medicine, University of Arkansas for Medical Sciences, Little Rock, AR, and Dr. Chao Zhang, Chief Scientific Officer, of Plexxikon, Inc. All the silhouettes are generated by Rigel Sison.

Funding: This research was partly supported by a grant (to SHK) from World Class University Project, Ministry of Education, Science and Technology, Republic of Korea and a gift grant to University of California, Berkeley, CA. (to support JJC). SHK acknowledges having appointments as Visiting Professorships at Yonsei University, Korea Advanced Institute of Science and Technology, and Incheon National University in South Korea during the manuscript preparation.

Competing financial interests: Authors declare that there are no competing financial interests in connection with this paper.

Author contribution: Conceptual design of the study and speculative interpretations and implications of the results by SHK; filtering and curation of genomic and proteomic sequence data from NCBI database, computational-algorithm design, programming and execution by JJC; interpretation of computational results by SHK and JJC; unsupervised clustering by various algorithms were performed by BJK; manuscript preparation by SHK with extensive discussions with JJC and BJK; all figures are designed by SHK and JJC.

Computer code availability: The FFP programs for this study (2v.3.0) written in GCC(g++) is available in Github: https://github.com/jaejinchoi/FFP.

References

1. Pace NR. Mapping the Tree of Life: Progress and Prospects. Microbiol Mol Biol Rev. 2009;73:565–76. doi:10.1128/mmbr.00033-09.

2. Zielezinski A, Vinga S, Almeida J, Karlowski WM. Alignment-free sequence comparison: benefits, applications, and tools. Genome Biology. 2017;18:1–17.

3. Blaisdell BE. A measure of the similarity of sets of sequences not requiring sequence alignment. Proc Natl Acad Sci U S A. 1986;83:5155–9. doi:10.1073/pnas.83.14.5155.

4. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44:D733–45. doi:10.1093/nar/gkv1189.

5. Sims GE, Jun S, Wu GA, Kim S. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proc Natl Acad Sci U S A. 2009;106:2677–82.

6. Misof B, Liu S, Meusemann K, Peters RSR, Donath A, Mayer C, et al. Phylogenomics resolves the timing and pattern of insect evolution. Science. 2014;346:763–8. doi:10.1126/science.1257570.

7. Thomas GWC, Dohmen E, Hughes DST, Murali SC, Poelchau M, Glastad K, et al. Gene content evolution in the arthropods. Genome Biol. 2020;21:15. doi:10.1186/s13059-019-1925-7.

8. R Core Team. R: A language and environment for statistical computing. 2016. https://www.r-project.org/.

9. Van Der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9:2579–605.

10. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4:406–25. http://www.ncbi.nlm.nih.gov/pubmed/18343690.

11. Gascuel O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol. 1997;14:685–95.

12. Choi JJ, Kim S-H. Whole-proteome tree of life suggests a deep burst of organism diversity. Proc Natl Acad Sci. 2019;117:756155. doi:10.1101/756155.

13. Jun S-R, Sims GE, Wu GA, Kim S-H. Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. Proc Natl Acad Sci. 2010;107:133–8. doi:10.1073/pnas.0913033107.

14. Choi JJ, Kim S-H. A genome Tree of Life for the Fungi kingdom. Proc Natl Acad Sci.

2017;114:201711939. doi:10.1073/pnas.1711939114.

15. Knuth DE. Seminumerical algorithms, in The art of computer programming. 3rd edition. Boston: Addison-Wesley; 1973.

16. Fisher RA, Yates F. Statistical tables for biological, agricultural and medical research. London: Oliver and Boyd; 1948.

17. Lin J. Divergence measures based on the Shannon entropy. IEEE Trans Inf Theory. 1991;37:145–51. doi:10.1109/18.61115.

18. Claude S. A Mathematical Theory of Communication. Bell Syst Tech J. 1948;Vol.27:379–423.

19. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. J Am Soc Inf Sci. 1990;41:391–407.

20. Robinson DF, Foulds LR. Comparison of phylogenetic trees. Math Biosci. 1981;53:131–47.

21. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019. doi:10.1093/nar/gkz239.

Figure legends

Fig. 1: Unsupervised clustering (grouping) of 134 extant arthropods by t-SNE.
t-SNE [9] is a machine learning algorithm of clustering a high-dimensional data into a two or three dimensional space by emphasizing resolution of clusters, but de-emphasizing the distances between the clusters. There are about 17 clusters with their common names in parentheses. These clusters can be assorted into six Order groups and two Subphylum groups (in colored-bold letters), correspond to Hemiptera, Lepidoptera, Diptera, Coleoptera, Blattodea, Hymenoptera, Chelicerata and Crustacea.

Fig. 2: Topology of the linear representation of whole-proteome ToAr
The colors of the first colored-band distinguish organisms in different Classes, and those of the second among different Orders (the names of different color-bands are shown in Fig. 3). Scientific names and common names, when available, of each organism are also listed. The silhouettes of sampled organisms are shown next to their names. To emphasize the clading pattern, all branch-lengths are ignored. The first two items refer to two members of the outgroup (see **Outgroup** in **Discussions**)

constructed by shuffling [15, 16] the whole-proteome sequences of two arthropods. The visualization of the ToAr was made using iTOL [21].

Fig. 3: Cumulative branch-lengths of the "pie" representation of whole-proteome ToAr.
This view of the tree shows all branch-length to emphasizes the progression of evolution of each member in the study from the origin at CGD=0 to the extant forms of the arthropods at CGD= near100. The small red arc near the root is at CGD=5.8, by which point of the evolutionary progression, the founders of "all major groups" (consisting of 7 Order groups and 2 Subphylum groups shown in Fig. 1) have emerged, suggesting that the remaining 94.2 on CGD scale corresponds to further diversification of the founders *within* each major group toward their extant forms. The visualization of the ToAr was made using iTOL [21].

Fig. 4: Expanded view of Fig. 3 near the root of the whole-proteome ToAr
Examples of the founders of all major groups are shown as blue dots, and the groups within two major groups, Diptera and Hymenoptera, as red dots. The visualization of the ToAr was made using iTOL [21].

Fig. 5: Simplified whole-proteome ToAr
On the left is cumulative genomic divergence values, which ranges from zero to around 100, and they correspond to the extent of evolutionary progression from the root of the ToAr to the extant leaves. For simplicity, "singletons" (that do not belong to any named groups) are not shown and all the leaf nodes and their branches of a common-named group (in parenthesis) are combined into a single dotted line coming out from their common ancestor (internal) node shown as a sphere. Each internal node represents a "pool of founding ancestors" (see Supplementary Information Fig. S3), and each circle represents an internal node containing all members of a clade with known name (in parenthesis). Dotted vertical lines are to indicate that they are arbitrarily shortened to accommodate large jumps of CGD values within a limited space of the figure. The double-headed arrow at bottom right indicates the short range of the CGD values, within which the founders of all the major groups of the extant arthropods in the study have emerged. For our interpretation of horizontal lines and vertical lines, see Supplementary Information Fig. S3.

Fig. 1

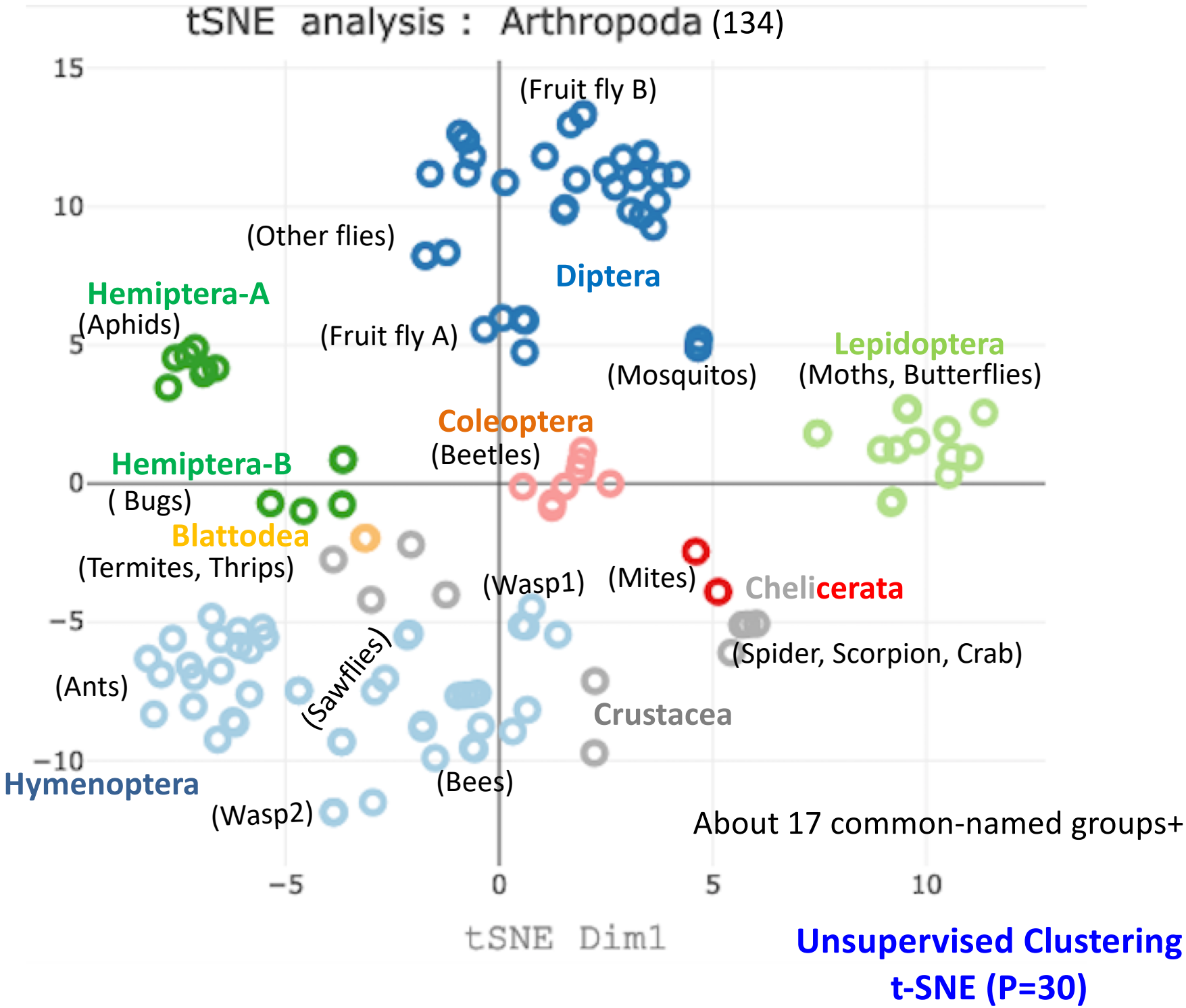

tSNE analysis : Arthropoda (134)
15
10
5
0
−5
−10
−5
0
5
10
tSNE Dim1
(Fruit fly B)
(Other flies)
Diptera
Hemiptera-A
(Aphids)
(Fruit fly A)
(Mosquitos)
Lepidoptera
(Moths, Butterflies)
Coleoptera
(Beetles)
Hemiptera-B
(Bugs)
Blattodea
(Termites, Thrips)
(Wasp1)
(Mites)
Chelicerata
(Spider, Scorpion, Crab)
(Sawflies)
(Ants)
Crustacea
Hymenoptera
(Bees)
(Wasp2)
About 17 common-named groups+
Unsupervised Clustering
t-SNE (P=30)

Fig. 2

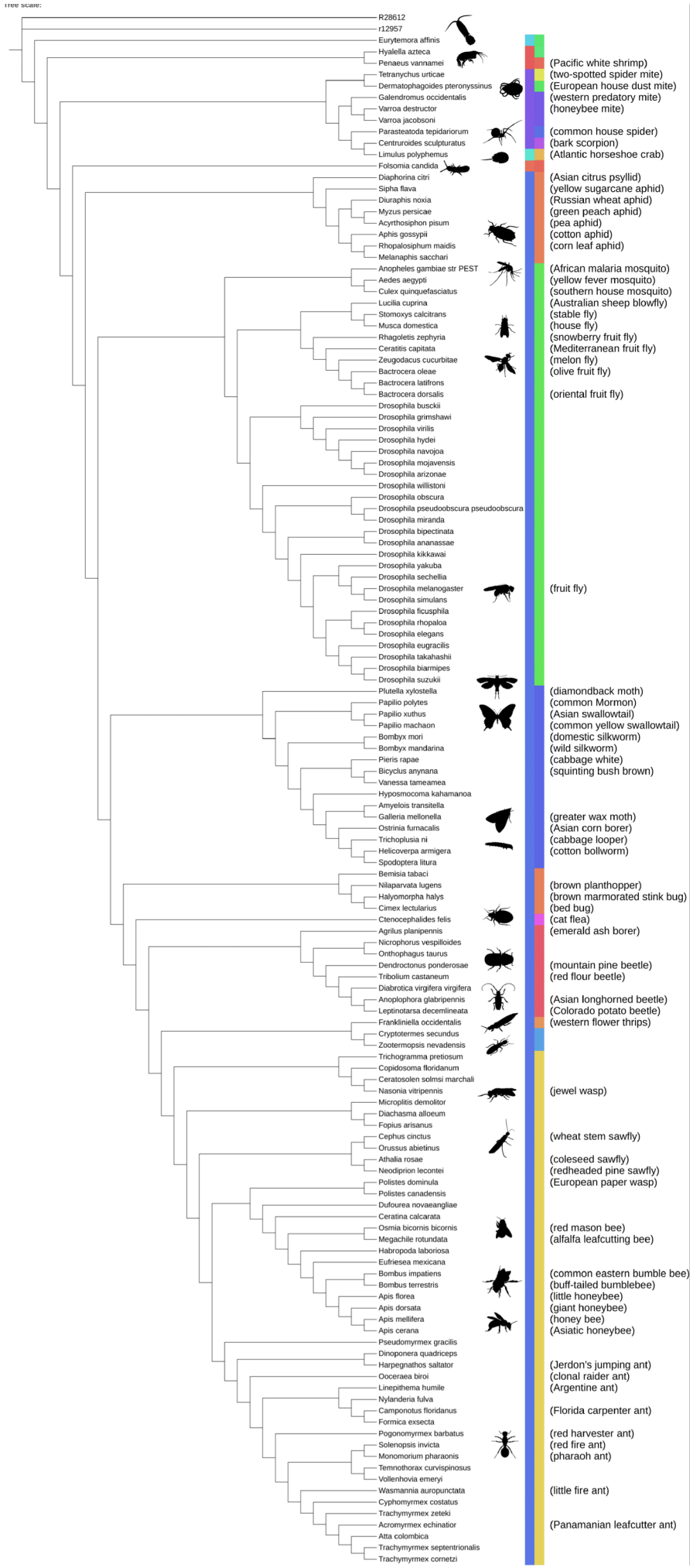

Tree scale:
R28612
r12957
Eurytemora affinis
Hyalella azteca
Penaeus vannamei
(Pacific white shrimp)
Tetranychus urticae
(two-spotted spider mite)
Dermatophagoides pteronyssinus
(European house dust mite)
Galendromus occidentalis
(western predatory mite)
Varroa destructor
Varroa jacobsoni
(honeybee mite)
Parasteatoda tepidariorum
(common house spider)
Centruroides sculpturatus
(bark scorpion)
Limulus polyphemus
(Atlantic horseshoe crab)
Folsomia candida
Diaphorina citri
(Asian citrus psyllid)
Sipha flava
(yellow sugarcane aphid)
Diuraphis noxia
(Russian wheat aphid)
Myzus persicae
(green peach aphid)
Acyrthosiphon pisum
(pea aphid)
Aphis gossypii
(cotton aphid)
Rhopalosiphum maidis
(corn leaf aphid)
Melanaphis sacchari
Anopheles gambiae str PEST
(African malaria mosquito)
Aedes aegypti
(yellow fever mosquito)
Culex quinquefasciatus
(southern house mosquito)
Lucilia cuprina
(Australian sheep blowfly)
Stomoxys calcitrans
(stable fly)
Musca domestica
(house fly)
Rhagoletis zephyria
(snowberry fruit fly)
Ceratitis capitata
(Mediterranean fruit fly)
Zeugodacus cucurbitae
(melon fly)
Bactrocera oleae
(olive fruit fly)
Bactrocera latifrons
Bactrocera dorsalis
(oriental fruit fly)
Drosophila busckii
Drosophila grimshawi
Drosophila virilis
Drosophila hydei
Drosophila navojoa
Drosophila mojavensis
Drosophila arizonae
Drosophila willistoni
Drosophila obscura
Drosophila pseudoobscura pseudoobscura
Drosophila miranda
Drosophila bipectinata
Drosophila ananassae
Drosophila kikkawai
Drosophila yakuba
Drosophila sechellia
Drosophila melanogaster
(fruit fly)
Drosophila simulans
Drosophila ficusphila
Drosophila rhopaloa
Drosophila elegans
Drosophila eugracilis
Drosophila takahashii
Drosophila biarmipes
Drosophila suzukii
Plutella xylostella
(diamondback moth)
Papilio polytes
(common Mormon)
Papilio xuthus
(Asian swallowtail)
Papilio machaon
(common yellow swallowtail)
Bombyx mori
(domestic silkworm)
Bombyx mandarina
(wild silkworm)
Pieris rapae
(cabbage white)
Bicyclus anynana
(squinting bush brown)
Vanessa tameamea
Hyposmocoma kahamanoa
Amyelois transitella
Galleria mellonella
(greater wax moth)
Ostrinia furnacalis
(Asian corn borer)
Trichoplusia ni
(cabbage looper)
Helicoverpa armigera
(cotton bollworm)
Spodoptera litura
Bemisia tabaci
Nilaparvata lugens
(brown planthopper)
Halyomorpha halys
(brown marmorated stink bug)
Cimex lectularius
(bed bug)
Ctenocephalides felis
(cat flea)
Agrilus planipennis
(emerald ash borer)
Nicrophorus vespilloides
Onthophagus taurus
Dendroctonus ponderosae
(mountain pine beetle)
Tribolium castaneum
(red flour beetle)
Diabrotica virgifera virgifera
Anoplophora glabripennis
(Asian longhorned beetle)
Leptinotarsa decemlineata
(Colorado potato beetle)
Frankliniella occidentalis
(western flower thrips)
Cryptotermes secundus
Zootermopsis nevadensis
Trichogramma pretiosum
Copidosoma floridanum
Ceratosolen solmsi marchali
Nasonia vitripennis
(jewel wasp)
Microplitis demolitor
Diachasma alloeum
Fopius arisanus
Cephus cinctus
(wheat stem sawfly)
Orussus abietinus
Athalia rosae
(coleseed sawfly)
Neodiprion lecontei
(redheaded pine sawfly)
Polistes dominula
(European paper wasp)
Polistes canadensis
Dufourea novaeangliae
Ceratina calcarata
Osmia bicornis bicornis
(red mason bee)
Megachile rotundata
(alfalfa leafcutting bee)
Habropoda laboriosa
Eufriesea mexicana
Bombus impatiens
(common eastern bumble bee)
Bombus terrestris
(buff-tailed bumblebee)
Apis florea
(little honeybee)
Apis dorsata
(giant honeybee)
Apis mellifera
(honey bee)
Apis cerana
(Asiatic honeybee)
Pseudomyrmex gracilis
Dinoponera quadriceps
Harpegnathos saltator
(Jerdon's jumping ant)
Ooceraea biroi
(clonal raider ant)
Linepithema humile
(Argentine ant)
Nylanderia fulva
Camponotus floridanus
(Florida carpenter ant)
Formica exsecta
Pogonomyrmex barbatus
(red harvester ant)
Solenopsis invicta
(red fire ant)
Monomorium pharaonis
(pharaoh ant)
Temnothorax curvispinosus
Vollenhovia emeryi
Wasmannia auropunctata
(little fire ant)
Cyphomyrmex costatus
Trachymyrmex zeteki
Acromyrmex echinatior
(Panamanian leafcutter ant)
Atta colombica
Trachymyrmex septentrionalis
Trachymyrmex cornetzi

Fig.3

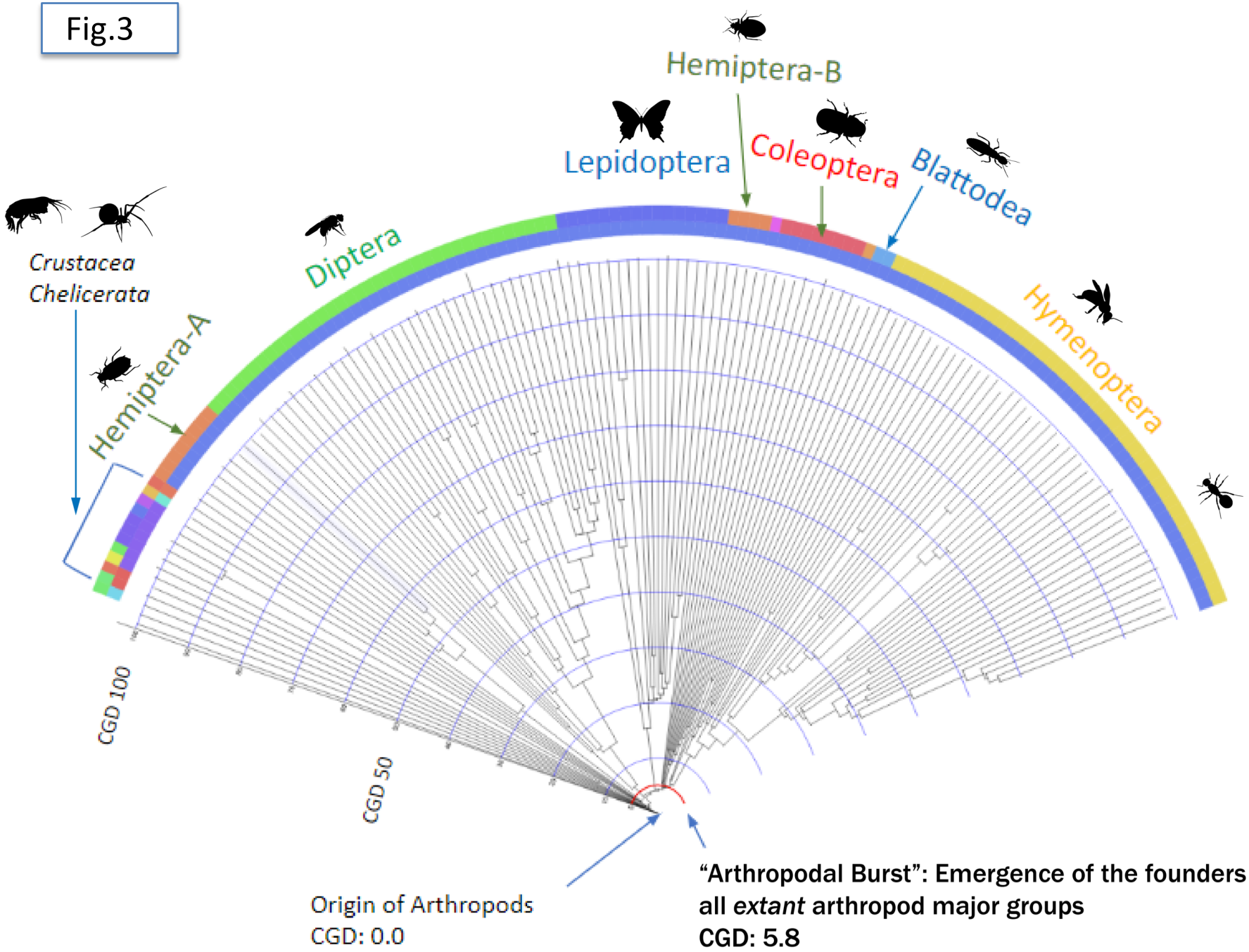
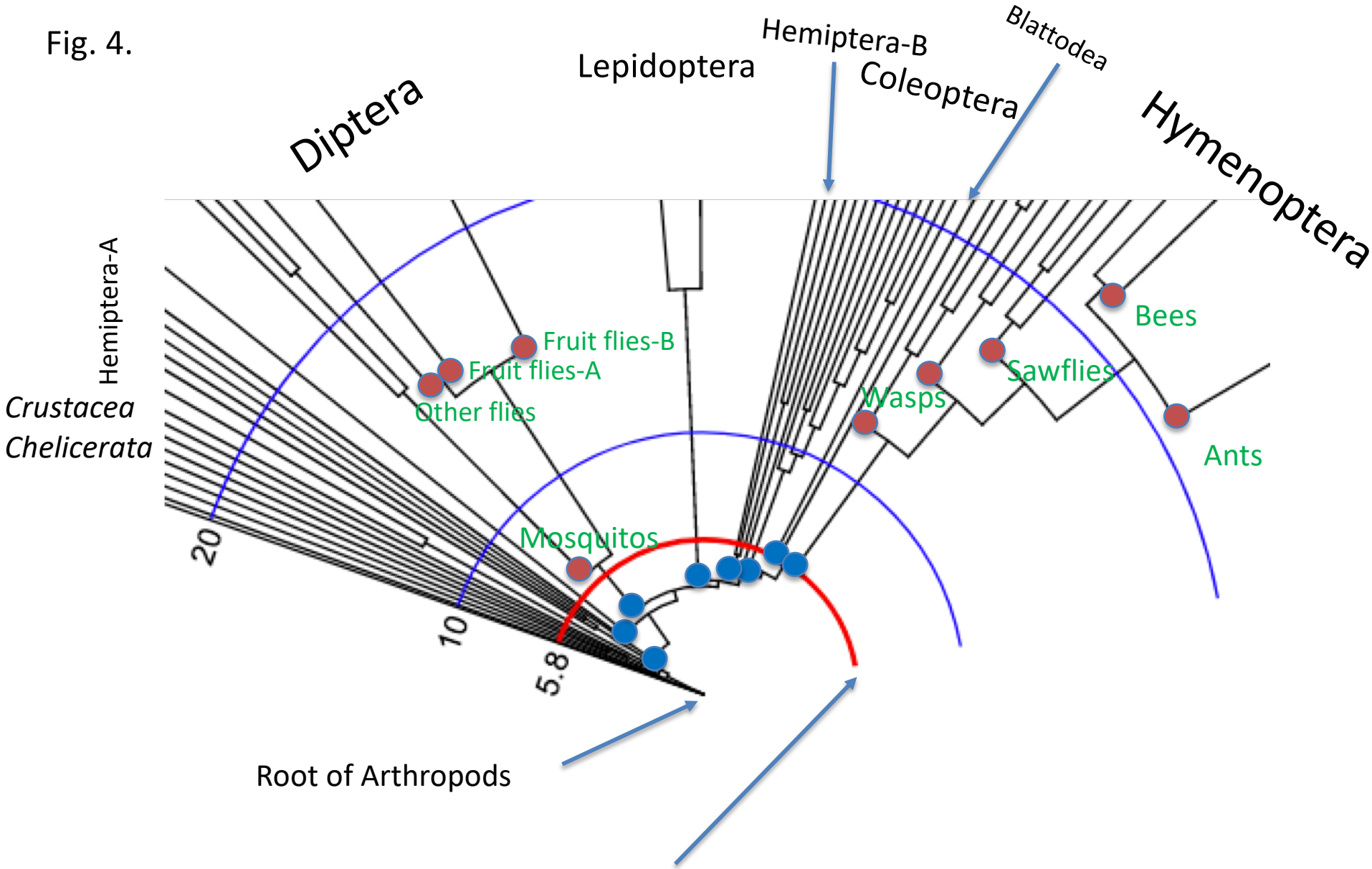
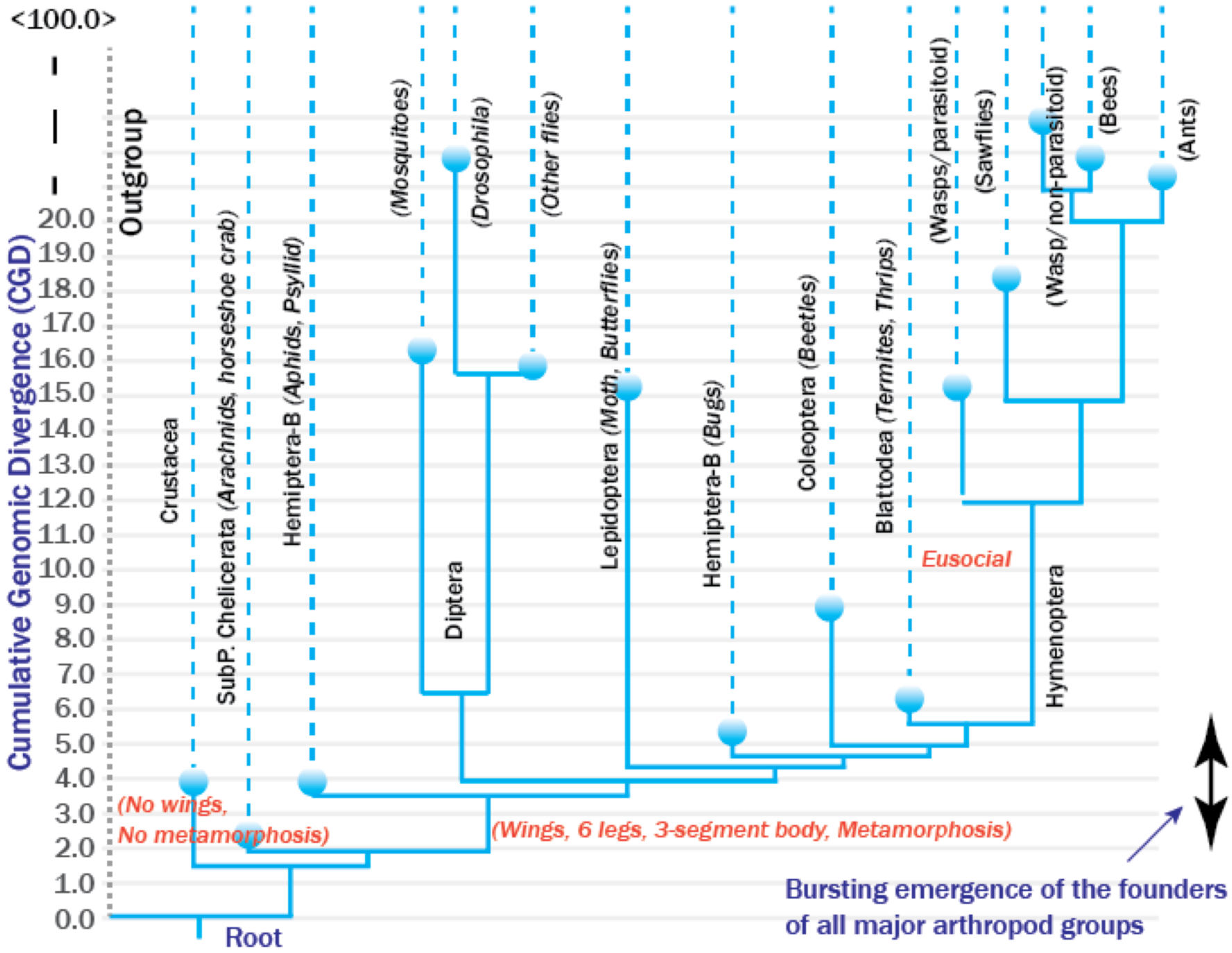

Hemiptera-B
Lepidoptera
Coleoptera
Blattodea
Diptera
Hymenoptera
Crustacea
Chelicerata
Hemiptera-A
CGD 100
CGD 50
Origin of Arthropods
CGD: 0.0
"Arthropodal Burst": Emergence of the founders of all *extant* arthropod major groups
CGD: 5.8

Fig. 4.



"Arthropodal Burst": ● Emergence of the founders of all extant major groups (Subphyla Crustacea and Chelicerata, and all Order groups of Hexapoda); ● Emergence of the founders of common-named groups in Diptera and Hymenoptera, two largest groups


Cumulative Genomic Divergence (CGD)
<100.0>
20.0
19.0
18.0
17.0
16.0
15.0
14.0
13.0
12.0
11.0
10.0
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
Outgroup
Crustacea
SubP. Chelicerata (Arachnids, horseshoe crab)
Hemiptera-B (Aphids, Psyllid)
(Mosquitoes)
(Drosophila)
(Other flies)
Diptera
Lepidoptera (Moth, Butterflies)
Hemiptera-B (Bugs)
Coleoptera (Beetles)
Blattodea (Termites, Thrips)
(Wasps/parasitoid)
(Sawflies)
(Wasp/non-parasitoid)
(Bees)
(Ants)
Eusocial
Hymenoptera
(No wings,
No metamorphosis)
(Wings, 6 legs, 3-segment body, Metamorphosis)
Root
Bursting emergence of the founders
of all major arthropod groups