Whole-proteome Tree of Insects: Grouping and phylogeny without sequence alignment

JaeJin Choi[1,2)], Byung-Ju Kim[3)] & Sung-Hou Kim[1,2,3,#)]

[1]Department of Chemistry and Center for Computational Biology, University of California, Berkeley, CA, 94720, USA, [2] Biological Systems and Engineering Division, and Molecular Biophysics & Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA, and [3]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

419 Latimer Hall, University of California, Berkeley CA, 94720, USA

e-mail addresses of co-authors: JJC (jaejinchoi@berkeley.edu); BJK (bjk@inu.ac.kr)

Short running title: **Whole-Proteome Tree of Insects**

Keywords: Organism tree, Gene/protein tree, Feature Frequency Profile, Jensen-Shannon divergence, Evolutionary progression scale, Cumulative genomic divergence, Arthropodal burst

Abstract

An "organism tree" of insects, the largest and most species-diverse group of all living animals, can be considered as a conceptual tree to capture a simplified narrative of the complex evolutionary courses of the extant insects. Currently, the most common approach has been to construct a "protein tree", as a surrogate for the organism tree, by Multiple Sequence Alignment (MSA) of highly homologous regions of a set of select proteins to represent each organism. However, such selected regions account for a very small fraction of the whole-proteome of each organism.

Information Theory provides a method of comparing two sets of *all* proteins, two whole-proteomes, without MSA: By treating each whole-proteome sequence as a "book" of amino acid alphabets, the information contents of two whole-proteomes can be quantitatively compared using the text comparison method of the theory, without sequence alignment, providing an opportunity to construct a "whole-proteome tree" of insects as a surrogate for an organism tree of insects.

A whole-proteome tree of the insects in this study shows that: (a) all the founders of the major groups of the insects have emerged in an explosive "burst" near the root of the tree, (b) the most basal group of all the insects is a subgroup of Hemiptera consisting of aphids and psyllids, and (c) there are other notable differences in the phylogeny of the groups compared to those of the recent protein trees of insects.

Introduction

Sequence-alignment-based "protein trees": An "organism tree" of insects can be considered as a practically useful narrative to convey a simplified evolutionary relationship among the insects. However, it is a conceptual tree that cannot be experimentally validated. Thus, it is expected that the effort will continue to find one or more "surrogate trees" derived from various descriptors of the characteristics associated with each insect and to find improved methods to estimate evolutionary distances from the divergence of the descriptors under as few subjective assumptions as possible at the time of investigation.

At present, the best descriptor of an insect as an organism is its entire whole-genome sequence information, from which whole-proteome sequence can be derived. However, for several decades, due to the technical difficulties and high cost of whole genome sequencing, and to the difficult task of comparing unaligned sequences, the most practically feasible and common approach to construct a surrogate tree has been to construct a Multiple Sequence Alignment (MSA)-based "protein tree" under a few important, but debatable assumptions: (a) a set of regions with high homology selected among each of homologous proteins may have enough information to represent a whole organism, and (b) the divergence of certain characteristics, most commonly, point substitution rates within each MSA aligned-regions may be a reasonable measure to represent the evolutionary divergence/distances among the whole organisms, without considering possible evolutionary roles of all other proteins without high homologous

65 regions among them and of other mutational events including absence/presence of
66 proteins.

67 Such "alignment-based" protein trees represent the evolutionary phylogeny of the
68 selected regions of the selected proteins, but not full characteristics of all proteins, let
69 alone the whole organisms, because the aligned regions account, in general, for a very
70 small fraction of all proteins (Pace 2009).

71

72 ***Information-theory-based ("alignment-free") "whole-proteome trees"*:** This situation
73 has since changed significantly in two important aspects: (a) During last decades, a large
74 number (over 134 species, mostly insects, as of 2020) of whole-genome sequences of
75 extant insect species have been accumulating in public databases, and (b) Information
76 Theory, developed to analyze linear electronic signals, was found to be adaptable to
77 analyze other linear information, such as natural languages and genomic information,
78 without sequence alignment ("alignment-free") (Zielezinski et al. 2017; Blaisdell 1986).
79 In this approach, the whole content, not selected portions of the whole content, of a whole-
80 proteome sequence, can be described by "*n*-Gram" or "*k*-mers" (Zielezinski et al. 2017).
81 *N*-Gram of a whole-proteome is the collection of all overlapping short subsequences of
82 length *n*, and it contains all information necessary to reconstruct the original sequence.
83 Furthermore, the information divergence (difference) between two *n*-Grams can be
84 estimated by, for example, Jensen-Shannon divergence (JSD) without alignment of the
85 whole proteome sequences (Lin 1991). Such approach has been widely tested and
86 validated for comparing texts and books of natural languages for latent semantic analysis

since 1990s (Deerwester et al. 1990) and gene sequences consisting of coding and non-coding regions as well as amino acid sequences since 1986 (Blaisdell 1986).

Some of these validated methods have been adapted and optimized to handle whole-proteome sequences in Feature Frequency Profile (FFP) method (Sims et al. 2009). Since there is no "golden standard" for a phylogenetic evolutionary tree of a group of organisms that can be experimentally validated, the FFP method has been tested using 26 books in English alphabets from diverse authors and genres, after removing all spaces and delimiters as well as author names, book titles, headers, footers, etc. In general, the method performed well in grouping the "books" by the genre and authors (see Fig. 1 of Sims et al. 2009). In a recent bench-marking studies of 24 Alignment-free methods, FFP method was ranked among the top 5 best-performing tools for phylogeny prediction based on the input data of assembled whole genome sequences (Zielezinski et al. 2019).

In the FFP method, whole-genome/whole-proteome information is used under the assumptions very different from the protein trees: (a) whole-proteome sequence of an organism represents the organism better than the collection of short regions of highly homologous sequence from a set of selected genes/proteins used in the protein trees and (b) a combination of *all types of mutations*, such as point substitution, insertion/deletion of various length, recombination, duplication, transfer or swapping of genes etc., contribute to the evolutionary processes of the organisms, rather than only point substitution rates in the protein trees. Thus, whole-proteome tree may provide an independent view of the evolutionary relationship among the insect organisms.

109 ***Experiences from earlier whole-proteome trees**:* In the last decade, we have tested and
110 optimized the protocols for building whole-proteome trees using various different
111 populations such as Bacteria and Archaea Domains (Jun et al. 2010), Fungi Kingdom
112 (Choi & Kim 2017) and, most recently, all three Domains at a deep phylogenic level
113 (Choi & Kim 2020). From these studies we have learned that: (a) among three types of
114 the trees (whole-genome DNA tree, whole-transcriptome RNA tree, and whole-proteome
115 amino-acid tree) the whole-proteome trees produce the most topologically stable trees; (b)
116 for a give group of organisms, the optimal length of the sequence strings to be used in FFP
117 method can be empirically determined; and (c) cumulative genomic divergence (CGD) is
118 a useful and computable quantity for the point of the emergence for the founders of a
119 group in the "evolutionary progression scale".

120 In this study, we optimized various parameters and protocols specifically for the
121 population of insects and present a view of the whole-proteome tree based on whole-
122 proteome sequences of 134 diverse arthropod species (123 insects plus 11 non-insect
123 arthropods), available in the NCBI database (O'Leary et al. 2016), and discuss its
124 implications to phylogenic aspects of insect evolution.

125

126 **Results**

127

128 To compare the current protein trees with our whole-proteome Tree of Insects
129 (ToIn) we chose two recent and very comprehensively-analyzed protein trees: The first
130 one is the recent "alignment-based" tree of 144 insect taxa based on 1,478 single-copy

131 protein-coding nuclear genes (Fig. 1 of Misof et al. 2014). The second is the tree for 76
132 arthropod taxa (Fig. 2 of Thomas et al. 2020) based on up to 4,097 single copy protein-
133 coding genes. In both cases, the number of aligned genes used are a small fraction of
134 about 10,000 to 31,000 genes among their study insects. Both protein trees agree with
135 each other, in general, on the branching order of the Order groups of the respective
136 populations, and on similar time spread of the emergence of the founders of the groups in
137 chronological time scale estimated based on available fossils and calibration methods
138 under various assumptions.

139 In comparing our whole-proteome ToIn to these two protein trees, we focus on
140 two aspects separately: grouping patterns and phylogeny of the groups. For the former,
141 we use two methods: First, we cluster our study population by several unsupervised
142 clustering algorithms using only the distances estimated from the "divergence" among
143 whole-proteome FFPs with no explicit constraints of the presence of the common
144 ancestor(s) or specific evolutionary models (see Construction of whole-proteome Tree of
145 Insects in Materials and Methods). We then ask whether the "clustering pattern" is
146 similar to the "clading pattern" in the protein trees and in our whole-proteome ToIn,
147 recognizing that the both tree constructions assume the constrains of the common
148 ancestor(s) and specific evolutionary models. For the phylogeny of the groups, we
149 compare the order of branching of the groups and their emergence points on the
150 evolutionary progression scale in our tree and in chronological scale in the protein trees
151 (see "Cumulative Genomic Divergence (CGD)" as "Evolutionary Progression Scale" in
152 Materials and Methods).

153

154 ***A. Demographic grouping pattern by clustering and clading***

155 **Clustering:** We have tested the grouping pattern of the insects by several
156 unsupervised clustering algorithms, such as Principal Component Analysis (PCA), Multi-
157 Dimensional Scaling, and t-Distributed Stochastic Neighbor Embedding (t-SNE) (R Core
158 Team 2016; v.d. Maaten & Hinton 2008), all of which can be accomplished from the
159 same starting "distance" matrix constructed using the divergence of whole-proteome
160 sequences, as calculated by JSD (Lin 1991) of FFPs, among all pairs of the study
161 organisms. All three clustering methods showed the clustering pattern compatible with
162 the current grouping of arthropods with common and scientific names, mostly based on
163 morphological characteristics. Figure 1, a classical PCA clustering, which is very similar
164 to that of MDS method, shows that all our study population are distributed into 5
165 "spokes". Two long spokes (IV and V) corresponds to all the members of Diptera and
166 Hymenoptera of Insecta Class, respectively. The remaining three short spokes (I, II, and
167 III) correspond to: Members of Chelicerata and Crustacea of non-insect arthropods in
168 spoke I; those of Hemipters-A and Lepidoptera of Insecta Class in spoke II, and those of
169 Hemiptera-B, Coleoptera, and Blattodea of Insecta Class in spoke III. The most
170 noticeable difference with the grouping in the current protein trees (Misof et al. 2014;
171 Thomas et al. 2020) is that Hemiptera is split into two separate groups, labeled in this
172 study as Hemiptera-A and -B. Another clustering by t-SNE (see Supplemental
173 information, Fig. S1) also shows a similar split of Hemiptera into two, which is
174 unexpected, because the assumptions and algorithms in t-SNE and PCA are completely
175 different.

176 **Clading:** As a second method of the grouping, we use the clading pattern of the
177 organisms in our whole-proteome ToIn. Figure 2 shows the topology of the ToIn,
178 constructed using Neighbor-Joining method implemented in BIONJ (Saitou & Nei 1987;
179 Gascuel 1997). In this study, we use the divergence of whole-proteome sequences of two
180 organisms as the estimates for the evolutionary distances between them, as calculated by
181 JSD of pair-wise FFPs at an optimal Feature length (see Construction of whole-proteome
182 Tree of Insects in Materials and Methods and Supplementary information Fig. S2)**.** We
183 also assume an evolutionary model of Maximum Parsimony (minimum evolution) in a
184 way that the chosen neighbors to be joined are those that minimize the total sum of
185 branch-lengths at each stage of step-wise joining of neighbors starting from a star-like
186 tree (Saitou & Nei 1987). The tree shows that most of the clusters in Fig.1 can be
187 identified among the clades in the ToIn.

188 **Robustness of grouping**: The grouping pattern by clustering *and* clading in our
189 study agrees well with those of the protein trees (Misof et al. 2014; Thomas et al. 2020)
190 except for Hemiptera group (see Notable differences in grouping and phylogenic
191 positions in Discussions and Implications). Thus, it is surprising that the demographic
192 grouping pattern is robust, in general, regardless of not only the information type (select
193 protein-characteristics, or whole-proteome characteristics), but also of the methods
194 (clustering or clading) used in grouping. For an implication of this result, see Similarities
195 in grouping patterns in Discussion and Implications. However, not surprisingly, there
196 are significant differences from the protein trees (Misof et al. 2014; Thomas et al. 2020)
197 in branch-length and branching order of the groups (see Dissimilarities in branching
198 orders and branch-lengths in Discussions and Implications).

199

200

201 ***B. Emergence of the "Founders" of all major groups in a staged "burst"***

202 For the following results we define "Cumulative Genomic Divergence (CGD)"
203 for an internal node of the ToIn as the cumulative scaled-branch-length from the tree root
204 to the node (see Cumulative Genomic Divergence (CGD) as "Evolutionary Progression
205 Scale" in Materials and Methods) to represent the extent of the "evolutionary progression"
206 of the node. The progression is scaled such that the root node of ToIn is set at CGD = 0
207 (see Outgroup in Discussions and Implications) and the leaf nodes of the extant organisms
208 at CGD = 100, on average.

209 **"Arthropodal burst" near the root of ToIn:** Figures 3, 4 and 5 show the
210 whole-proteome tree with CGD values. They reveal that the "founders" (for definition,
211 see Supplemental information, Fig. S3) of all major groups of insects as well as non-
212 insect arthropods (at Subphyla and Order levels) emerged in a staged burst within a short
213 evolutionary progression span between CGD of 1.6 and 5.8 (marked by a small red arc in
214 Figs. 3A and 3B), near the root of the tree. This observation is dramatically different
215 from those of the protein trees (Misof et al. 2014; Thomas et al. 2020), where the
216 founders of the major groups of all arthropods emerged throughout a long time-span of
217 chronological scale.

218 **A subgroup of Hemiptera (Hemiptera-A) is the most basal group of all**
219 **Insecta:** The first founders of Class Insecta to emerge is the founders of Hemiptera-A
220 group (aphids and a psyllid) at CGD of about 3.7 (Figs. 3A, 3B). This is in stark contrast

221 to the protein trees, where all Hemiptera is the sister to Thysanoptera (thrips) (Misof et al.
222 2014) or a group of all Hemiptera, thrips and human louse is the sister to all other large
223 groups of insects except Blattodea group (Thomas et al. 2020) (see also Notable
224 differences in grouping and phylogenic positions in Discussions and Implications).

225 **Order of emergence of the "founders" of all major groups of Insecta:** Figure
226 4 shows a series of staged emergence of the founders of all major groups of Insecta.
227 After the most basal group of Hemiptera-A group (aphids and a psyllid) at CGD of
228 around 3.7, the founders of Diptera group emerged at CGD of 4.1, and those of the
229 remaining five Order-level groups (Lepidoptera, Hemiptera-B (bugs, a planthopper and a
230 whitefly), Coleoptera, Blattodea + a thrips, and Hymenoptera groups) at CGD of 4.4, 4.8,
231 5.2, 5.8, and 5.8, respectively. For possible implications see Notable differences in
232 grouping and phylogenic positions in Discussions and Implications below**)**

233

234 **Discussions and Implications**

235

236 ***Similarities in grouping patterns*:** As mentioned earlier, it is surprising that the grouping
237 patterns at Order level between the protein trees (Misof et al. 2014; Thomas et al. 2020)
238 and our whole-proteome ToIn are very similar (see below for one notable exception of
239 Hemiptera) despite the facts that the types of input data (multiple-aligned regions of
240 selected proteins vs. whole-proteome) and estimation methods for evolutionary distance
241 used (based on point mutational rates vs. whole genomic divergences) are very different.
242 A possible implication is that, after the "burst", the members of each group evolved

243 largely "isolated" within the group without significant genomic mixing between the
244 groups, thus, resulting in much smaller genomic variation within the group than between
245 the groups, as manifested by mostly isolated clusters.
246
247 ***Dissimilarities in branching orders and branch-lengths:*** It is not surprising that the
248 branching orders and branch-lengths are not similar between the protein trees (Misof et al.
249 2014; Thomas et al. 2020) and our whole-proteome tree, because the assumptions under
250 which the estimations for evolutionary distances among the organisms are calculated are
251 very different: in the protein trees, the distances are calculated only for the aligned
252 portions of the selected genes using, e.g., point-substitutional mutation rates, while in our
253 tree they are calculated by accounting presence/absence of all amino acid short strings,
254 Features, for all proteins due to all types of mutations.
255
256 ***Evolutionary progression scale vs. Chronological time scale:*** It is difficult to design a
257 scale that quantitatively measures the degree of evolutionary progression, because it is not
258 clear what characteristics of an organism can best reflect the progression and also are
259 quantitatively measurable. Since we are using whole-proteome sequence to represent each
260 organism, we use the divergence of the whole-proteome sequences as the evolutionary
261 progression scale (Choi & Kim 2020). In contrast to linear chronological time scale, the
262 evolutionary progression scale is most likely not strictly linear, because any significant
263 geological and ecological events may accelerate or decelerate the evolutionary progression
264 for a given organism. However, the direction of arrows in both scales are the same,
265 suggesting that the two scales may be calibrated when sufficient fossils, other independent

266 records, and improved calibration methods become available (see Cumulative Genomic
267 Divergence (CGD) as "Evolutionary Progression Scale" in Materials and Methods).
268 Meanwhile, we use the evolutionary progression scale to compare the order of emergence
269 of the founders of various major groups under the assumption that the whole-proteome
270 sequence divergence can be considered as informational entropy, which increases as
271 evolution progresses, similar to the physical entropy of universe increases as the universe
272 evolve.

273

274 ***"Burst" vs. Gradual emergence of the founders of major groups:*** While cognizant of
275 the difference and similarity of the two scales, the most dramatic difference is observed in
276 the span of the scales within which the founders of all major groups at Order level
277 emerged in the protein trees (Misof et al. 2014; Thomas et al. 2020) and in our whole-
278 proteome ToIn: In the protein trees, the founders of all the groups at Order level emerged
279 gradually during a long chronological time span of about 350 Million years (Myrs)
280 corresponding roughly 60% of about 570 Myrs between the tree root to the extant
281 arthropods (Fig. 1 of Misof et al. 2014), or about 210 Myrs corresponding to roughly 37%
282 of the same full chronological scale (Fig. 2 of Thomas et al. 2020). In drastic contrast, the
283 founders of all the major groups in our tree emerged within about 4% of the full
284 evolutionary progression scale in a sudden burst ("Arthropodal burst"; see Figs. 3A and
285 3B) near the root of our whole-proteome tree. This drastically contrasting observations
286 between the two types of trees may have an important qualitative evolutionary implication
287 in constructing the narrative for the birth of the insect diversity.

288

289 ***Notable differences in grouping and phylogenic positions:*** Despite the drastic difference
290 in the emergence pattern of the founders (burst vs. gradual) mentioned above, the order of
291 emergence of the major groups at Order and Subphylum levels agree between the two
292 protein trees (Misof et al. 2014; Thomas et al. 2020) and our whole-proteome tree with
293 some notable differences in Hemiptera and Blattodea as described in Results above.
294 These differences may get resolved once the whole-genome sequences of many more
295 relevant organisms become available. At present, we suggest some possible implications
296 as described below:

297 **Hemiptera:** As mentioned earlier, in our whole-proteome ToIn (Figs. 3A and
298 3B) as well as in PCA (Fig. 1) and t-SNE (Supplemental Fig. S1) clustering plots,
299 Hemiptera is divided into two separate clades/clusters, which we call Hemiptera-A
300 ("primitive" Hemiptera, such as aphids and a psyllid) and Hemiptera-B ("bugs" such a
301 planthopper, a whitefly, a stink bug and a bed bug), and their phylogenetic positions are
302 very far apart (see Figs. 2, 3 and 5): Hemiptera-A at the basal position of all Insecta and
303 Hemiptera-B as sister to the group consisting of Lepidoptera, Coleoptera and
304 Hymenoptera. But, in the protein trees (Misof et al. 2014; Thomas et al. 2020) the both
305 groups form a single clade, and is at basal or sister to all other large groups of insects
306 except Blattodea group. This difference in clustering and phylogenetic positioning
307 suggest that, when viewed at whole-proteome level, which includes *both* homologous and
308 non-homologous proteins, the members are more similar within each subgroup than
309 between the two subgroups in our tree. But, when viewed, as in the protein trees, only
310 for the select homologous proteins in the *absence* of the non-homologous proteins, which

311 are the overwhelming majority of the all proteins, they are similar among all of them to
312 form only one clade.

313 **Blattodea:** Two termites (Blattodea) and one thrips (Thysanoptera), both eusocial
314 and hemimetabolous, form a clade in our whole-proteome ToIn and the clade is sister (or
315 basal) to Hymenoptera group, which is also eusocial but holometabolous (see Fig. 5).
316 However, in one protein tree (Misof et al. 2014), Blattodea group (cockroaches and
317 termites, which are eusocial and hemimetabolous) is a member of a larger clade
318 Polyneoptera and placed at the basal position to all other Order groups of Insecta, which
319 are largely non-social and hemi- or holo-metabolous, while, in the other protein tree
320 (Thomas et al. 2020), Blattodea group forms a separate clade, and is placed near the basal
321 position of all other Order groups of Insecta. This is in contrast to what we observe in our
322 ToIn, where Hemiptera-A, is the basal group of Insecta.

323

324 ***Outgroup:*** Since our method does not require multiple sequence alignment, we
325 constructed, as was described in our earlier works on whole-proteome trees (Jun et al.
326 2010; Choi & Kim 2017; Choi & Kim 2020), the proteome sequence of an "artificial
327 (faux) arthropod" by "shuffling" (Knuth 1973; Fisher & Yates 1948) the alphabets of the
328 whole proteome sequence of an organism in the study group. We used two such artificial
329 arthropods (named R28612 and r12957) to form the outgroup for this study. Each has the
330 same size and amino acid composition of corresponding protein of an extant arthropod,
331 but does not have gene sequences information for the organism's survival.

332

333

334

335 **Materials and Methods**

336

337 ***Sources and selection of proteome sequences***

338 We downloaded the proteome sequences for 134 arthropods from NCBI RefSeq
339 DB using NCBI FTP depository (O'Leary et al. 2016). Protein sequences derived from
340 all organelles were excluded from this study. Also excluded from our study are those
341 derived from whole genome sequences assembled with "low" completeness based on two
342 criteria: (a) the genome assembly level indicated by NCBI as "contig" or lower (i.e. we
343 selected those with the assembly levels of 'scaffold', 'chromosome' or 'complete
344 genome'), and (b) the proteome size smaller than 80% of the smallest proteome size
345 among highly assembled arthropod genomes (*Anopheles gambiae str. PEST* with 14,089
346 proteins at "chromosome" assembly level; TaxID 180454).

347 All taxonomic names and their taxon identifiers (TaxIDs) of the organisms in this
348 study are from NCBI taxonomy database, and listed in Supplementary Information,
349 Dataset S1.

350

351 ***Construction of whole-proteome Tree of Insects***

352 Based on our earlier experiences of constructing whole-proteome trees of
353 prokaryotes (Jun 2010), fungi (Choi & Kim 2017) and all life forms (Choi & Kim 2020)
354 by Feature Frequency Profile (FFP) method (Sims et al. 2009), following choices have

355 been made to obtain a topologically stable whole proteome ToIn of maximum parsimony
356 (minimum evolution) by BIONJ (Saitou & Nei 1987): a) Among three types of genomic
357 information (DNA sequence of the whole genome, RNA sequence of whole
358 transcriptome and amino acid sequence of whole proteome) whole-proteome trees are
359 most "topologically stable" as estimated by Robinson-Foulds metric (Robinson &Foulds
360 1981) at respective "optimal Feature-length"; b) For FFP as the "descriptor" of the whole
361 proteome of each organism, the optimal Feature-length is about 10 amino-acid string (see
362 Supplementary Information**,** Fig. S2); and c) Jensen-Shannon Divergence (JSD) (Lin
363 1991) is an appropriate measure of "divergence of information content", as the measure
364 of dissimilarity between two whole-proteome descriptors, for constructing the distance
365 matrix of BIONJ (Saitou & Nei 1987; Gascuel 1997). It is important to note that such
366 FFP of a whole-proteome sequence of an organism has all the information necessary to
367 reconstruct the original whole proteome sequence.

368

369 ***"Cumulative Genomic Divergence (CGD)" as "Evolutionary progression scale"***

370 In Information Theory (Shannon 1948), the Jensen-Shannon Divergence (JSD)
371 (Lin 1991), bound between zero and one, is commonly used as a measure of the
372 dissimilarity between two probability distribution of informational features. The FFP as
373 the descriptor for a linear sequence information of the whole proteome of an organism is
374 such a probability distribution. Thus, a JSD value of two FFPs, used as a measure of the
375 information divergence between two proteome sequences, is also bound between 0 and 1,
376 corresponding to the JSD value between two FFPs of identical whole proteome sequences
377 and two completely different whole proteome sequences, respectively. Any whole

378 proteome-sequence "dissimilarity" between two extant organisms accumulated during the
379 evolution can be considered as caused by changes of, ultimately, genomic sequences of
380 all protein coding genes due to all types of mutational events, such as point substitutions,
381 insertion/deletion of various lengths, inversion, recombination, loss/gain of genes, etc. as
382 well as other unknown mechanisms, and they will bring JSD somewhere between 0.0 and
383 1.0 depending on the degree of the sequence divergence.

384 In this study the collection of the JSDs for all pairs of the study organisms plus 2
385 out-group members (see Outgroup in Discussions and Implications) constitutes the
386 "distance matrix" for BIONJ (Saitou & Nei1987; Gascuel 1997). Since all the branch-
387 lengths are derived from the JSD values, the cumulative branch-length of an internal
388 node, which we call "cumulative genomic divergence (CGD)" (to reflect the fact that the
389 proteomic divergence is ultimately derived by the genomic divergence during evolution)
390 of the node, can be considered as the point of evolutionary stage reached by the node on
391 an "evolutionary progression scale". For convenience of assigning the nodes on the
392 progression scale, CGDs are scaled, as mentioned earlier, such that the CGD value at the
393 root node of ToIn is set to zero and the leaf nodes of the extant organisms to 100, on
394 average, corresponding to the fully evolved genomic states of the organisms, which we
395 define as the beginning and ending point of the "evolutionary progression scale" for the
396 organisms (see Fig. 3A).

397

398 ***Clustering methods***

399 We use two unsupervised methods to observe the clustering patterns based solely
400 on whole-proteome sequences: Principal Component Analysis (PCA) and t-Distributed
401 Stochastic Neighbor Embedding (t-SNE) (R Core Team 2016; v.d. Maaten & Hinton
402 2008). Both are dimensional reduction methods, but with different strengths and
403 weaknesses for our purposes, which help to visualize any clustering pattern in the data
404 distribution. Both are based only on the evolutionary distances (CGD in this study),
405 estimated by the divergence of whole-proteome sequences among all pairs of the study
406 arthropods. In PCA, the distances within a cluster as well as between two clusters are
407 quantitative, thus, two close clusters nearby may not resolve well. In t-SNE, which
408 applies Machine Learning to emphasize the resolution of nearby clusters, but the inter-
409 cluster distances are de-emphasized, thus, not quantitative.

410

411

412 **Declarations**

413

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

420

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 and BJK; unsupervised clustering by various algorithms were performed by BJK; interpretation of computational results by SHK, JJC and BJK; manuscript preparation by SHK with extensive discussions with JJC and BJK; all figures are designed by SHK, JJC and BJK.

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

Blaisdell, B. E. (1986). A measure of the similarity of sets of sequences not requiring sequence alignment. *Proceedings of the National Academy of Sciences of the United States of America*, *83*(14), 5155–5159.

Choi, J.J., & Kim, S.-H. (2017). A genome Tree of Life for the Fungi kingdom. *Proceedings of the National Academy of Sciences*, *114*(35), 9391–9396

Choi, J.J., & Kim, S.-H. (2020). Whole-proteome tree of life suggests a deep burst of organism diversity. *Proceedings of the National Academy of Sciences*, *117*(7), 3678–3686

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

Fisher, R. A., & Yates, F. (1948). *Statistical tables for biological, agricultural and medical research*. London: Oliver and Boyd.

Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, *14*(7), 685–695.

Jun, S.-R., Sims, G. E., Wu, G. A., & Kim, S.-H. (2010). Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. *Proceedings of the National Academy of Sciences*, *107*(1), 133–138.

Knuth, D. E. (1973). *Seminumerical algorithms, in The art of computer programming* (3rd ed.). Boston: Addison-Wesley.

Letunic, I., & Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Research*. 47(W1), W256–W259

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, *37*(1), 145–151.

Misof, B., Liu, S., Meusemann, K., Peters, R. S. R., Donath, A., Mayer, C., … Zhou, X. (2014). Phylogenomics resolves the timing and pattern of insect evolution. *Science*, *346*(6210), 763–768.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733–D745.

Pace, N. R. (2009). Mapping the Tree of Life: Progress and Prospects. *Microbiology and Molecular Biology Reviews*, *73*(4), 565–576.

R Core Team. (2016). *R: A language and environment for statistical computing.* Retrieved from https://www.r-project.org/

Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, *53*(1), 131–147.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, *4*(4), 406–425.

Shannon, C. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, *Vol.27*(1948), 379–423.

Sims, G. E., Jun, S., Wu, G. A., & Kim, S. (2009). Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(8), 2677–2682.

Thomas, G. W. C., Dohmen, E., Hughes, D. S. T., Murali, S. C., Poelchau, M., Glastad, K., … Richards, S. (2020). Gene content evolution in the arthropods. *Genome Biology*, *21*(1), 15. 1-14

Van Der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Zielezinski, A., Girgis, H. Z., Bernard, G., Leimeister, C.-A., Tang, K., Dencker, T., … Karlowski, W. M. (2019). Benchmarking of alignment-free sequence comparison methods. *Genome Biology*, *20*(1), 144. 1–18

Zielezinski, A., Vinga, S., Almeida, J., & Karlowski, W. M. (2017, October 3). Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biology*, Vol. 18, pp. 1–17.

Figure legends

Fig. 1: Unsupervised clustering (grouping) of 134 extant arthropods (123 insects plus 11 non-insect arthropods) by classical PCA

Classical PCA plotted for the three principal axes reveals about 5 large clusters arranged in 5-spokes. Two long spokes (IV and V) corresponds to all the members of Diptera and Hymenoptera, respectively. The remaining three short spokes (I, II, and III) correspond to: Members of Chelicerata and Crustacea in spoke I; those of Hemipters-A and Lepidoptera in spoke II, and those of Hemiptera-B, Coleoptera, and Blattodea in spoke III.

Fig. 2: Topology of the linear representation of whole-proteome Tree of Insects (ToIn)

The colors of the first (inner) colored-band distinguish organisms in different Classes, and those of the second (outer) band among different Orders (the names of different color-bands are shown in Fig. 3A). 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 and Implications) constructed by shuffling (Knuth 1973; Fisher & Yates 1948) the whole-proteome sequences of the two arthropods. The visualization of the ToIn was made using iTOL (Letunic & Bork 2019).

Fig. 3A: "Pie" representation of whole-proteome ToIn with the cumulative branch-lengths scale.

This view of the whole-proteome ToIn shows all branch-lengths to emphasizes the progression of evolution of each member in the study population from the root of the tree at CGD = 0 to the extant forms of the members at CGD = 100, on average. 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 and gradual evolution of the founders and common ancestors *within* each major group toward their extant forms. The visualization of the ToIn was made using iTOL (Letunic & Bork 2019).

Fig. 3B: Expanded view of Fig. 3A near the root of the whole-proteome ToIn

Examples of the founders of all major groups are shown as blue dots, and the common ancestors of extant groups within two major groups, Diptera and Hymenoptera, as red dots. The visualization of the ToIn was made using iTOL (Letunic & Bork 2019).

Fig. 4: Simplified whole-proteome ToIn

The vertical axis shows cumulative genomic divergence (CGD) values, which ranges from zero to around 100, and they correspond to the extent of evolutionary progression from the root of the ToIn 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 node of the extant group shown as a blue sphere. Each internal node represents a "pool of founding ancestors" (see Supplementary Information Fig. S3). 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 organisms in this study have emerged in a "burst". For our interpretation of horizontal lines and vertical lines, see Supplementary Information Fig. S3.

Fig. 1

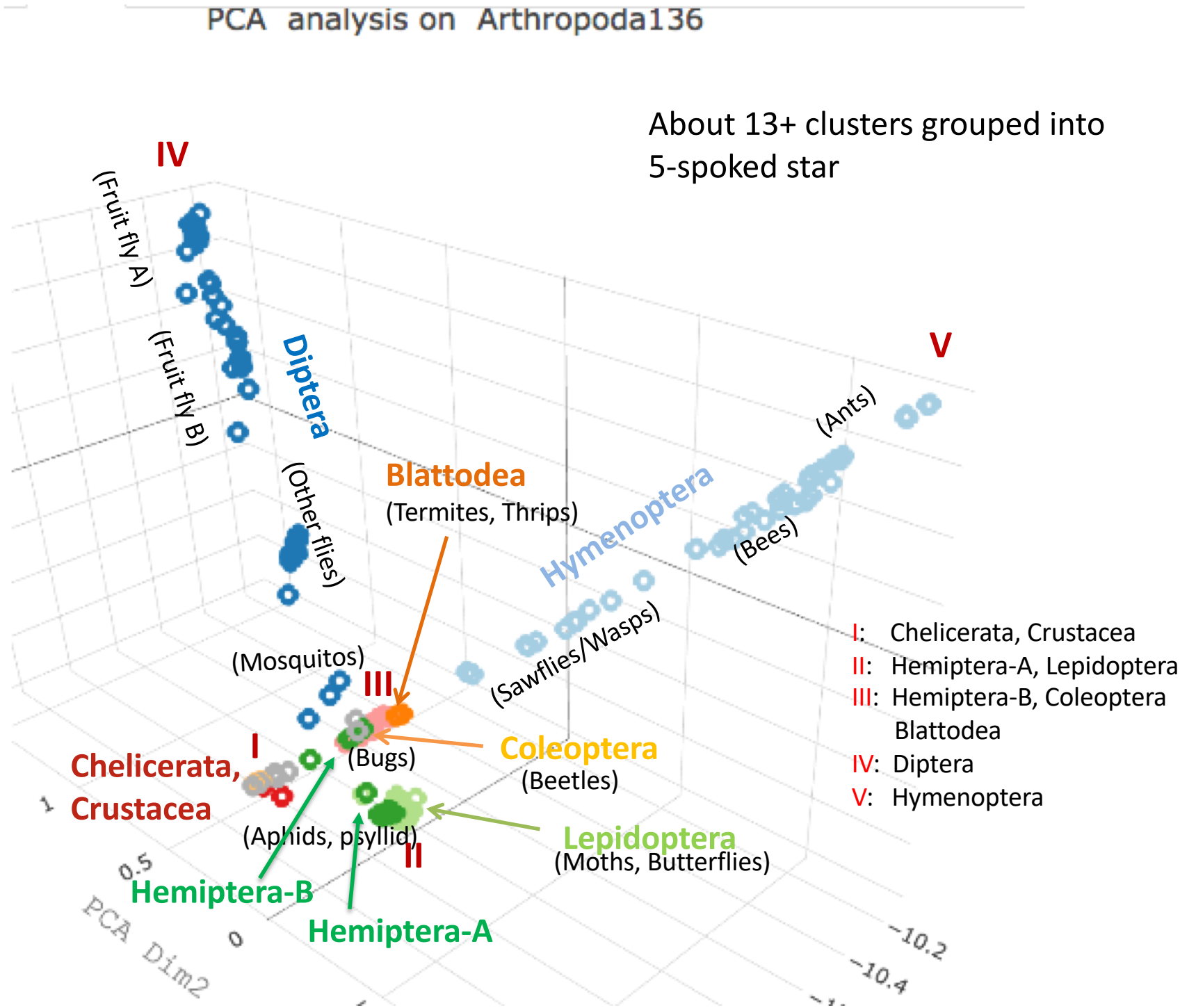

PCA analysis on Arthropoda136
About 13+ clusters grouped into 5-spoked star
IV
(Fruit fly A)
(Fruit fly B)
Diptera
(Other flies)
Blattodea
(Termites, Thrips)
Hymenoptera
V
(Ants)
(Bees)
(Sawflies/Wasps)
(Mosquitos)
III
I
Chelicerata, Crustacea
(Bugs)
Coleoptera
(Beetles)
(Aphids, psyllid)
II
Lepidoptera
(Moths, Butterflies)
Hemiptera-B
Hemiptera-A
PCA Dim2
1
0.5
0
-10.2
-10.4
I: Chelicerata, Crustacea
II: Hemiptera-A, Lepidoptera
III: Hemiptera-B, Coleoptera Blattodea
IV: Diptera
V: Hymenoptera

Fig. 2

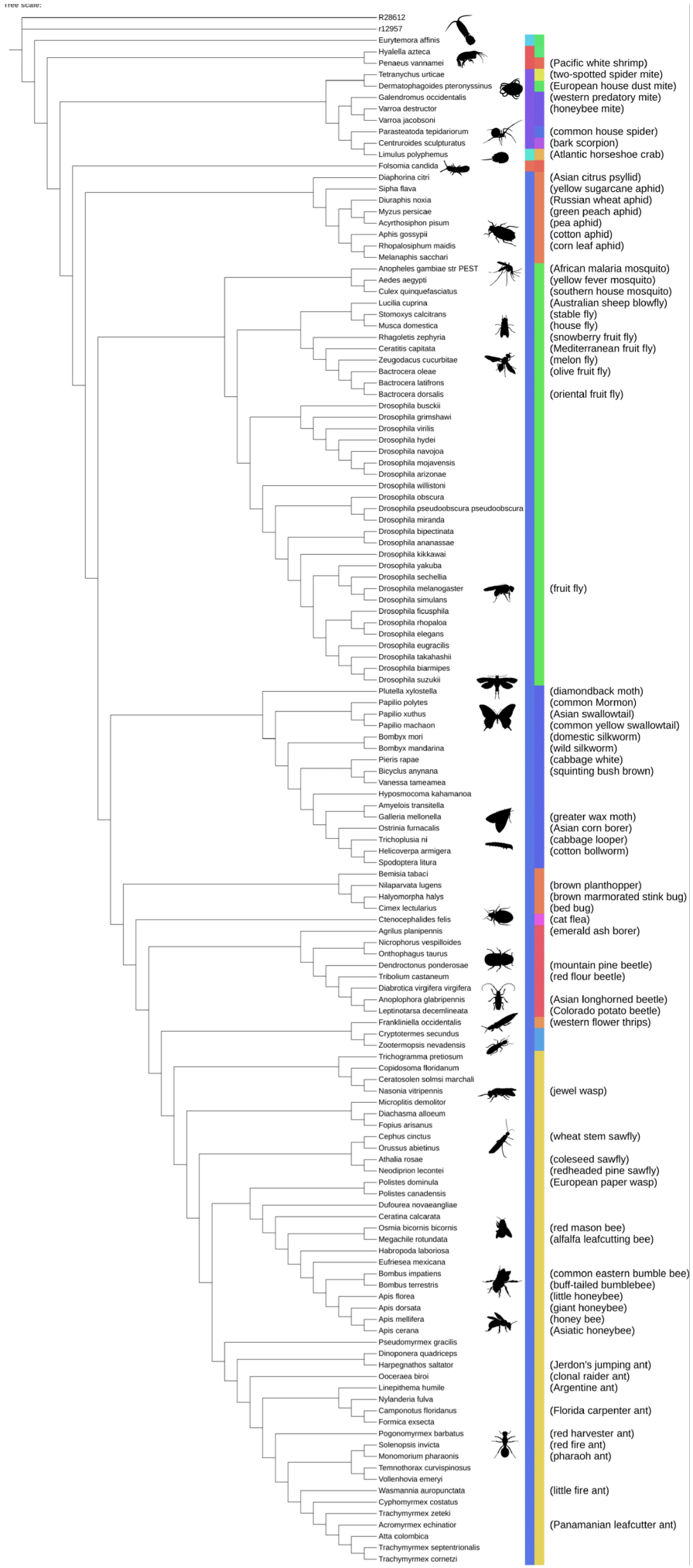

Tree scale:
R28612
r12957
Eurytemora affinis
Hyalella azteca
Penaeus vannamei
(Pacific white shrimp)
Tetranychus urticae
(two-spotted spider mite)
Dermatophagoides pteronyssinus
(European house dust mite)
Galendromus occidentalis
(western predatory mite)
Varroa destructor
Varroa jacobsoni
(honeybee mite)
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. 3A.

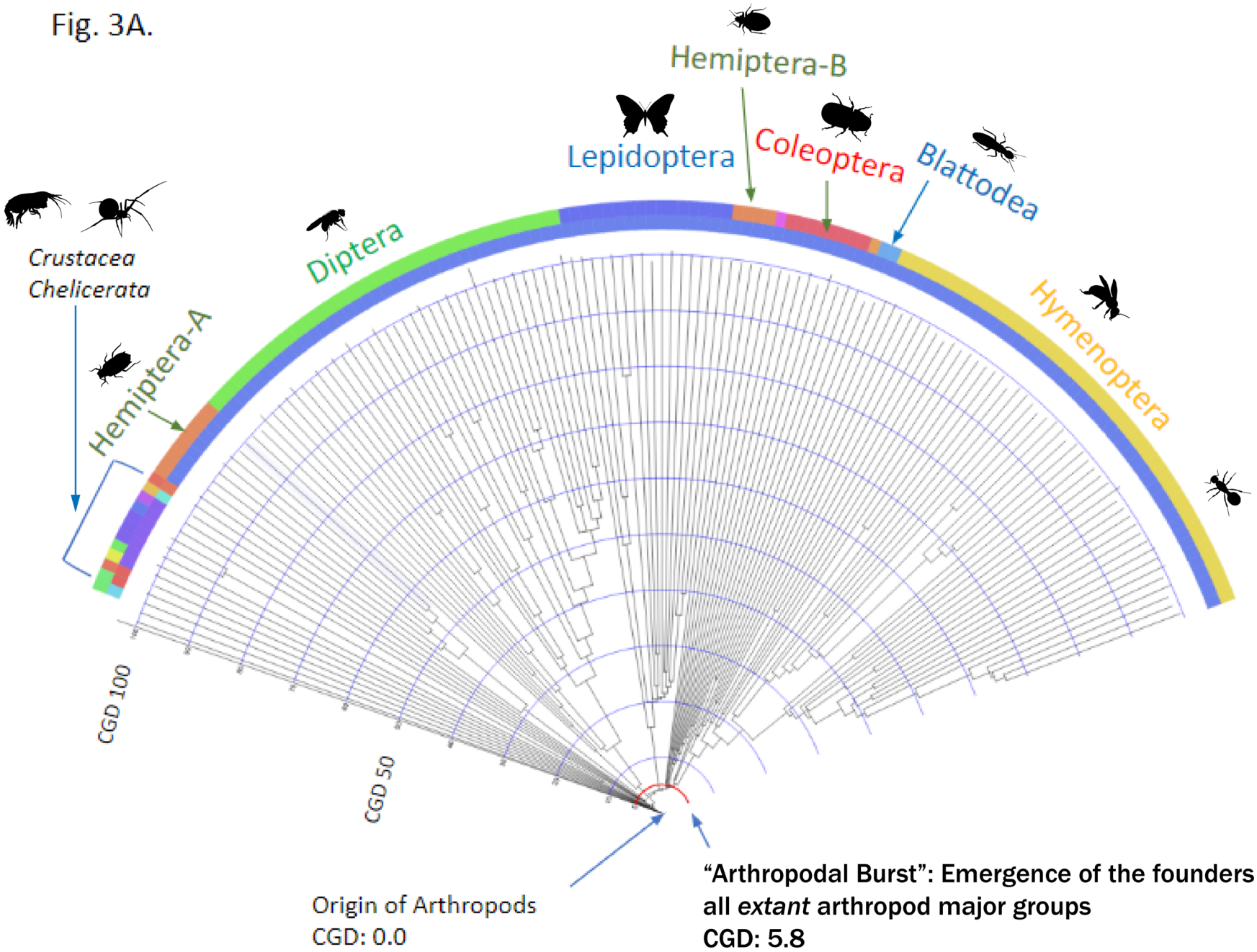
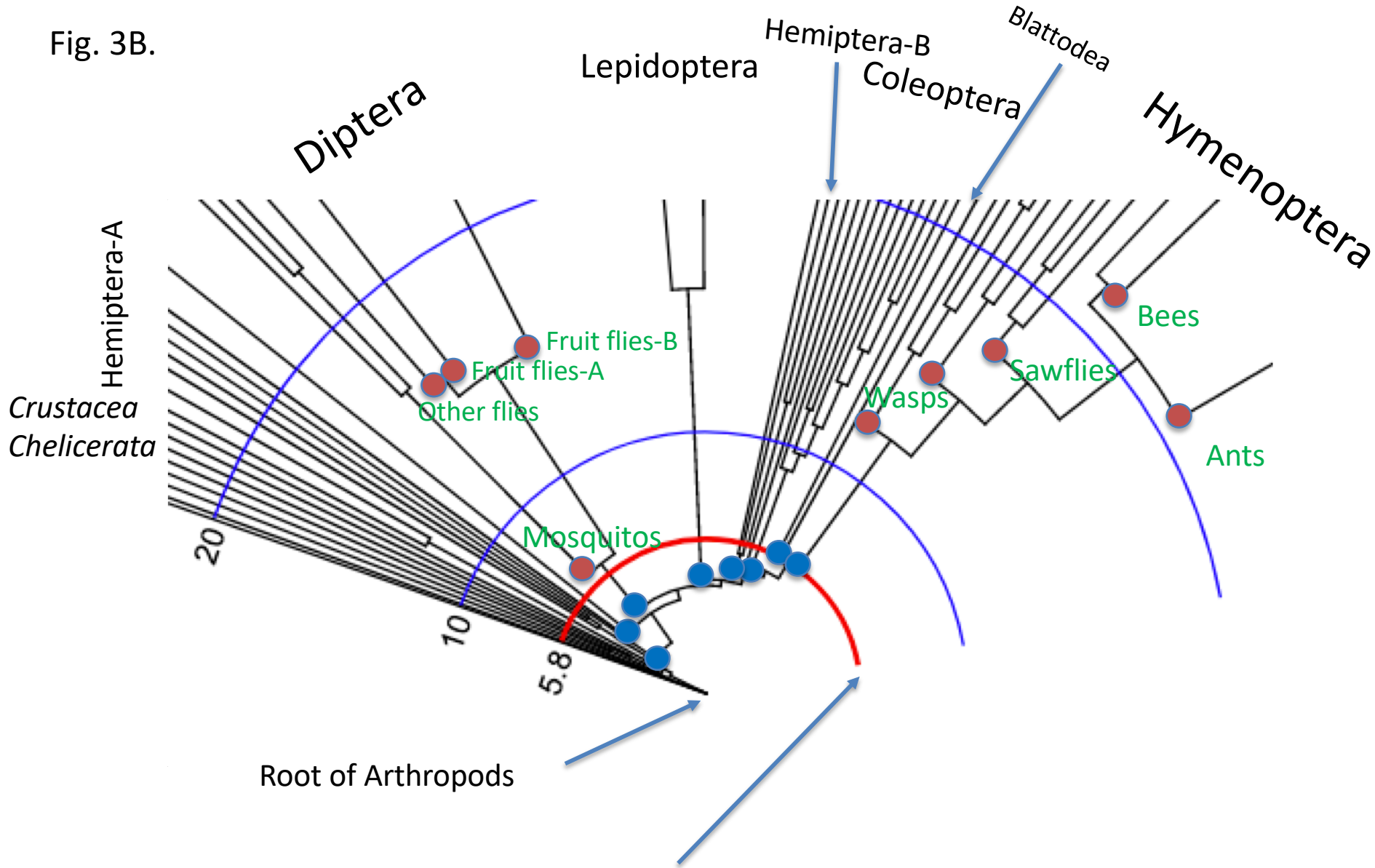
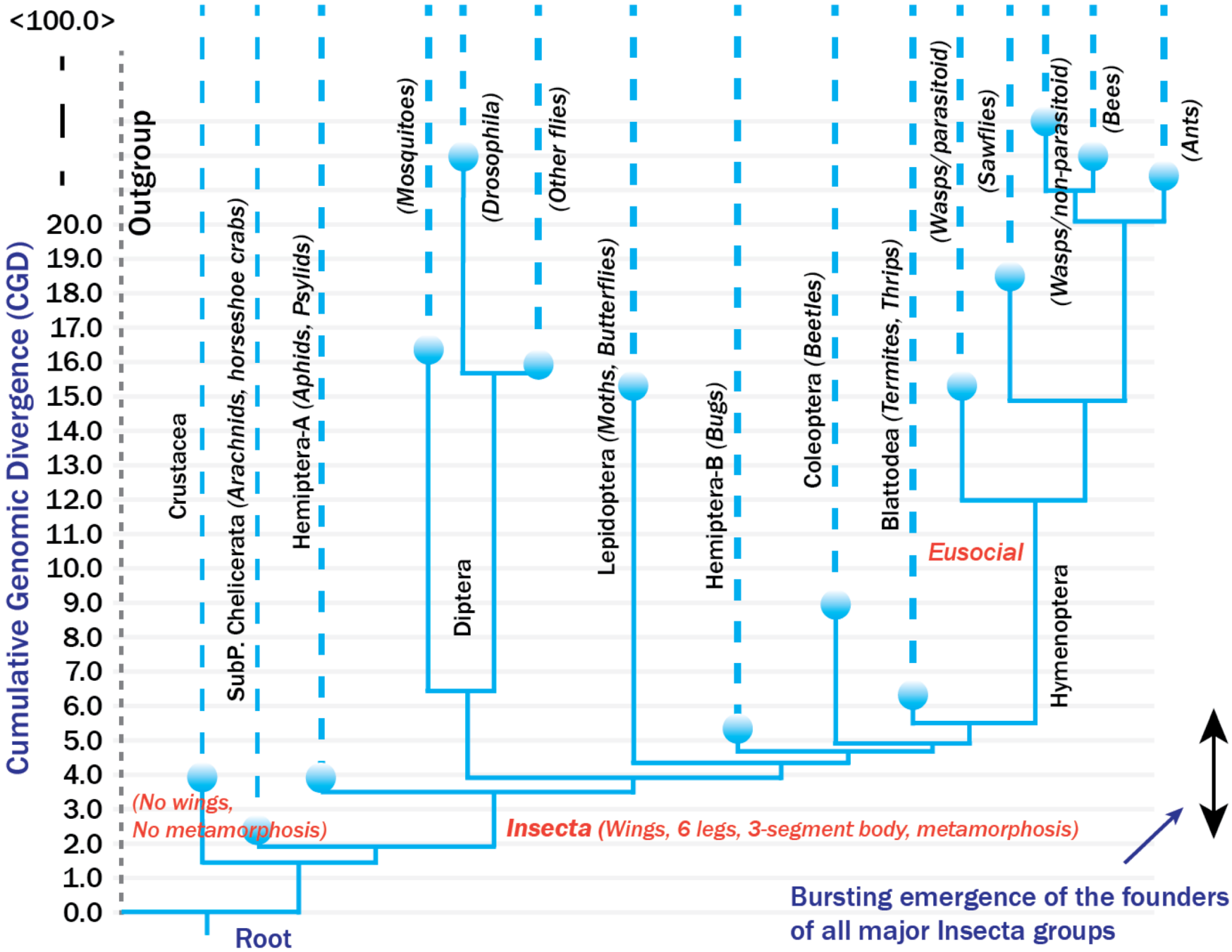

Hemiptera-B
Lepidoptera
Coleoptera
Blattodea
Diptera
Hymenoptera
Crustacea
Chelicerata
Hemiptera-A
CGD 100
CGD 50
Origin of Arthropods
CGD: 0.0
"Arthropodal Burst": Emergence of the founders of all *extant* arthropod major groups
CGD: 5.8

Fig. 3B.



"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

Fig. 4


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 crabs)
Hemiptera-A (Aphids, Psylids)
(Mosquitoes)
(Drosophila)
(Other flies)
Diptera
Lepidoptera (Moths, Butterflies)
Hemiptera-B (Bugs)
Coleoptera (Beetles)
Blattodea (Termites, Thrips)
(Wasps/parasitoid)
(Sawflies)
(Wasps/non-parasitoid)
(Bees)
(Ants)
Eusocial
Hymenoptera
(No wings,
No metamorphosis)
Insecta (Wings, 6 legs, 3-segment body, metamorphosis)
Root
Bursting emergence of the founders
of all major Insecta groups