Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Insect Science

Power, resolution and bias: recent advances in insect phylogeny driven by the genomic revolution

David K Yeates[1], Karen Meusemann[1], Michelle Trautwein[2], Brian Wiegmann[3] and Andreas Zwick[1]

Our understanding on the phylogenetic relationships of insects has been revolutionised in the last decade by the proliferation of next generation sequencing technologies (NGS). NGS has allowed insect systematists to assemble very large molecular datasets that include both model and non-model organisms. Such datasets often include a large proportion of the total number of protein coding sequences available for phylogenetic comparison. We review some early entomological phylogenomic studies that employ a range of different data sampling protocols and analyses strategies, illustrating a fundamental renaissance in our understanding of insect evolution all driven by the genomic revolution. The analysis of phylogenomic datasets is challenging because of their size and complexity, and it is obvious that the increasing size alone does not ensure that phylogenetic signal overcomes systematic biases in the data. Biases can be due to various factors such as the method of data generation and assembly, or intrinsic biological feature of the data *per se*, such as similarities due to saturation or compositional heterogeneity. Such biases often cause violations in the underlying assumptions of phylogenetic models. We review some of the bioinformatics tools available and being developed to detect and minimise systematic biases in phylogenomic datasets. Phylogenomic-scale data coupled with sophisticated analyses will revolutionise our understanding of insect functional genomics. This will illuminate the relationship between the vast range of insect phenotypic diversity and underlying genetic diversity. In combination with rapidly developing methods to estimate divergence times, these analyses will also provide a compelling view of the rates and patterns of lineagenesis (birth of lineages) over the half billion years of insect evolution.

Addresses
[1] Australian National Insect Collection, CSIRO National Research Collections Australia, Canberra, ACT 2601, Australia
[2] California Academy of Sciences, 55 Music Concourse Drive, San Francisco, CA 94118, USA
[3] Department of Entomology, North Carolina State University, Raleigh, NC 27695-7613, USA

Corresponding author: Yeates, David K (david.yeates@csiro.au)

Current Opinion in Insect Science 2016, **13**:16–23

This review comes from a themed issue on **Insect genomics**

Edited by **Owain Edwards**

For a complete overview see the Issue and the Editorial

Available online 14th November 2015

http://dx.doi.org/10.1016/j.cois.2015.10.007

Introduction

The current revolution in arthropod phylogenetics has been reviewed recently [1–3]. The field has evolved quickly over the last decade, from relying on comparatively small, hand-assembled datasets from single or a small set of genes using Sanger sequencing, to very large datasets that are assembled in semi-automated analysis pipelines and typically contain thousands of protein-coding genes obtained from NGS platforms [reviewed in 4]. This transition to phylogenomics has caused dramatic changes in the way datasets are assembled, assessed, manipulated, analysed and interpreted. The transformation is nascent, with analyses protocols being retooled frequently, and entirely new processes inserted into previous practice. The resulting output from these analytical approaches is a very large and rapidly growing volume of molecular data, most of which are publicly available for download from NCBI (Figure 1).

Here, we focus on the increasing resolution and confidence in insects ordinal-level relationships provided by EST (Expressed Sequence Tags) and RNA-Seq sequencing projects over the past few years. These projects have built on advances in RNA sequencing techniques [5]. We also review the findings of studies that have addressed the source of incongruence between genes in these datasets, and examine the influence of random and systematic error on the results [6••]. Many of the studies we refer to include both, insects *sensu stricto* and their closest relatives, the entognathous hexapods (Collembola, Diplura and Protura), together forming a clade called the Hexapoda (hexapods).

The very first hexapods and their Crustacean ancestors

Although morphological analyses had always placed the insects and their closest relatives among other terrestrial arthropods such as myriapods (centipedes, millipedes and their relatives), even early molecular studies showed support for hexapod origin within Pancrustacea (crustaceans plus hexapods) rendering crustaceans paraphyletic (see e.g. [7]).

Although Ertas *et al.* [8] already suggested a close relationship of the enigmatic Remipedia (crustaceans) and hexapods based on hemocyanin, Regier and colleagues [9] assembled the largest Sanger-sequenced dataset with 62 single-copy, nuclear, protein-coding genes of 75 arthropod species, which includes all classes of

pancrustaceans (crustaceans plus hexapods). The study supports xenocarid crustaceans (Remipedia and Cephalocarida) being closest relatives of Hexapoda. A 129 gene EST dataset with a rich arthropod taxon sampling, careful attention to ortholog assignment, and data quality, also found strong support for Pancrustacea, but lacked xenocarid terminals [10]. Rota-Stabelli and colleagues [11] also found the pancrustacean clade using a phylogenomic dataset of 198 protein coding genes, 393 morphological characters and 9 microRNAs. A much larger dataset including data from 1886 single-copy, protein coding genes and 46 terminals [12] and the study recently published by Misof *et al.* using 1478 single-copy, protein coding genes and 144 terminals [13••] respectively found Remipedia as closest relative to hexapods. These phylogenomic studies compiled from RNA-Seq data lack cephalocarid sequence data. Adding data from 9 new transcriptomes of crustaceans to the sequence data of Regier *et al.* [9] and other sources to sample a molecular dataset with more than 1000 genes, Oakley and colleagues [14] found remipedes as closest relatives to Hexapoda with cephalocarids as sister group to branchiopods. However, there was generally weak support for nodes that linked hexapods to their crustacean relatives.

Entognathous hexapods: anything is possible

The relationships of the three orders of entognathous hexapods Collembola (springtails), Protura (coneheads) and Diplura (two-pronged bristletails) have been particularly difficult to establish with any dataset, phylogenomic or otherwise. Most studies find the three orders to be a monophyletic sister lineage to the insects (Entognatha, see [10,12]). However, relationships among the three orders are not recovered consistently. Using a dataset of 253 single copy, protein-coding, orthologous genes and 62 terminals derived from RNA-Seq data, Dell'Ampio *et al.* [15••] were not able to establish robustly supported relationships among the entognathous hexapods, but showed that unevenly distributed missing data can inflate node support. Using methods to reduce the effects of unevenly distributed missing data, Misof *et al.* [13••] found that Diplura are the closest relatives to insects. Low bootstrap support for the nodes adjacent to the three entognathous orders in this phylogeny suggest that the relationships of these three orders have not yet been settled.

The first winged insects: old wings and new analyses

A study using RNA-Seq data with ~150 genes published by Simon *et al.* [16,17] addressed phylogenetic relationships in early winged insects (Pterygota). They focused on relationships of the first winged insects, that is, dragonflies, damselflies (Odonata), and mayflies (Ephemeroptera) and their relationship to all other insects, the Neoptera. Although analyses of morphological and single-gene molecular datasets were not conclusive, the authors found Ephemeroptera were closest to Neoptera. The EST dataset with 129 genes by Meusemann and colleagues [10] found similar results, except that the mayflies formed a clade with the Hemiptera with low support, a novel and unexpected placement, undermining confidence in the result and identifying the mayfly as a rogue taxon. Misof *et al.* [13••] inferred a monophyletic Palaeoptera (Odonata plus Ephemeroptera) as sister to the remaining pterygotes, with low support. This suggests again that relationships among early winged insects are not settled.

Polyneoptera: clarity emerging with phylogenomic datasets

Although relationships of Holometabola (insects with complete metamorphosis) have been well-established using morphological and Sanger sequence datasets (see review [1,2], and below), this does not hold for the remaining insect orders, the hemimetabolous (incomplete metamorphosis) insects. They are classified into two supra-ordinal groups, (1) the Polyneoptera (including grasshoppers, cockroaches, stick insects, earwigs and their relatives) and (2) the Paraneoptera (Hemiptera, thrips, lice and their relatives). Studies using transcriptome and genome sequence data have focused on the Holometabola, including model organisms such as the fruit fly *Drosophila* (Diptera), and honey bee *Apis* (Hymenoptera), while little attention has been paid to hemimetabolous insects. By adding new transcriptome data from a few representative polyneopteran orders, Simon *et al.* [17] addressed polyneopteran relationships with a dataset of 1579 genes and 78 terminals. Their analyses provided strong support for the monophyly of Polyneoptera, including earwigs (Dermaptera), stoneflies (Plecoptera) and ground lice (Zoraptera). Slightly different results were obtained by analysing matrices that minimised missing data in different ways, by (1) removing genes with patchy taxon coverage (73 terminals and 102 genes), and (2) removing terminals with patchy gene coverage (62 terminals, 285 genes), but all supported a monophyletic Polyneoptera. A monophyletic Polyneoptera was also found by Letsch *et al.* including 1579 genes and 40 terminals, Letsch and Simon including 1574 genes, 54 terminals [18,19] and Misof *et al.* [13••].

Paraneoptera: monophyletic sister to the Holometabola?

The sucking bugs and their relatives lack convincing synapomorphies, and have inconsistently been recovered as monophyletic and closest relatives of the Holometabola in phylogenomic studies [17–19]. However, Psocodea (sucking lice and bark lice) alone were found to be closest relatives of Holometabola by Misof *et al.* [13••]. This new placement did not receive support in additional statistical tests, and implies that further gene and taxonomic sampling and more cautious analyses are required.