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

David K. Yeates, Australian National Insect Collection, CSIRO National Research Collections
Australia, Canberra ACT 2601 AUSTRALIA, david.yeates@csiro.au

Karen Meusemann, Australian National Insect Collection, CSIRO National Research Collections
Australia, Canberra ACT 2601 AUSTRALIA, karen.meusemann@csiro.au

Michelle Trautwein, California Academy of Sciences
55 Music Concourse Drive, San Francisco, CA 94118, USA mtrautwein@calacademy.org

Brian Wiegmann, Department of Entomology, North Carolina State University, Raleigh NC 27695-
7613, bwiegman@ncsu.edu

and

Andreas Zwick, Australian National Insect Collection, CSIRO National Research Collections
Australia, Canberra ACT 2601 AUSTRALIA, andreas.zwick@csiro.au

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

Keywords: next generation sequencing, NGS, non-model organisms, transcriptomes, RNA-Seq, target enrichment, hybrid probes, genomes, insects

Abstract

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.

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 (Fig. 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

While 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 for example [7]).

While 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 1,886 single-copy, protein coding genes and 46 terminals [12] and a the study recently published by Misof *et al.* using 1,478 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 1,000 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, i.e. dragonflies, damselflies (Odonata), and mayflies (Ephemeroptera) and their relationship to all other insects, the Neoptera. While 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

While 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 1,579 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 1,579 genes and 40 terminals, Letsch and Simon including 1,574 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.

Holometabola: Order-level relationships established

In contrast to hemimetabolous orders, a consensus has emerged over the past few years on the relationships of the Holometabola, with rare congruence between morphology (20), Sanger sequenced molecular datasets (e.g. [21], reviewed in [1-2]), phylogenomic datasets (Peters *et al.* [22] comprising a dataset with 1,343 genes and 88 terminals), and a genomic dataset including 4,485 protein-coding genes plus 8,983 near intron pairs for 13 terminals [23]). Wasps, bees and ants (Hymenoptera) are sister to all other holometabolous insects (the Aparaglossata, see [22]). Aparaglossata is divided into two major monophyletic groups: firstly flies and relatives (Diptera), butterflies and moths (Lepidoptera) and smaller orders (fleas, scorpion flies, caddisflies), together called Mecopterida, and secondly beetles (Coleoptera), twisted winged parasites (Strepsiptera), and lacewings plus relatives (Neuropteridea). These studies have always placed the enigmatic twisted-winged insects (Strepsiptera) as closest relatives to Coleoptera, rather than to Diptera as suggested in an early study using ribosomal DNA [24]. Recent studies have examined the relationships of ants, bees and wasps using transcriptome data including 5,214 genes and 19 terminals [25] and the ditrysian Lepidoptera (e.g., Bazinet *et al.* [26] with 741 genes and 46 terminals and Kawahara and Breinholt [27] including 2,696 genes and 46 terminals).

Systematic Error in Phylogenomic Datasets

Analysis of phylogenomic datasets is challenging because of their size and complexity, and massive amounts of data alone do not ensure that phylogenetic signal overcomes systematic biases in the data [13, 28, 29]. Some of these challenges are new, and some are common to all phylogenetic analyses. For example, taxon sampling remains an important issue in phylogenomics, with inadequate sampling capable of leading to long-branch attraction artefacts [see 18; 30-32].

The amount of missing data in phylogenomic datasets is generally higher than in datasets derived from Sanger sequencing (e.g., up to 80%, see [6]). The proportion of data present can for instance be measured as matrix fill, the percentage of cells in a matrix with unambiguous amino acids or base pairs [6]. Data assembled from EST or RNA-Seq studies are particularly prone to missing data, for example some genes may not have been transcribed at the time the organism was preserved, or sequence coverage was too low to detect rare transcripts. A software tool available to analyse the amount of missing data at individual amino acid or nucleotide sites within multiple alignments is AliStat (see Supplementary Materials [13]B, available upon request from L. Jermin, lars.jermin@csiro.au), which also produces heat maps illustrating the level of data overlap between pairs of taxa, and calculates matrix fill (Supplementary Materials [13]). While the debate over the impact and effect of missing data has a long history, see [33] for a recent update, the general effects are likely to be dataset dependent, and not recoverable from simulation or a sample of empirical studies. Those researchers trying to minimise the amount of missing data in a matrix do so in a number of ways, for example by removing taxa with abundant missing data, removing genes absent from many taxa, or finding a trade off between these strategies that optimise matrix fill. The software MARE [34] offers one way to measure

the potential phylogenetic information content of partitions (e.g., genes) using geometry mapping. This information is used in combination with information on missing data to select an optimal data subset. This last step is based on a simple optimisation algorithm that penalises matrix reduction and favours the total information content of the matrix [6].

Combining data from different sources into a phylogenomic dataset can also cause artefacts because it can bias the distribution of missing data. Letsch *et al.* [18] combined terminals with data from genomes with those from transcriptome studies. Terminals derived from full genomes had complete coverage, but those from transcriptomes had many missing genes. Non-randomly distributed data can distort the phylogenetic position of taxa having the full gene set relative to the transcriptome-derived terminals with less gene overlap. The same problems, and additional node support inflation, occurred even in data derived only from transcriptomes if large blocks of data are absent [15].

The effect of compositional heterogeneity on phylogenetic analysis was first identified as a potential problem decades ago [35, 36]. The software SymTest can be used to evaluate whether sequence data compiled into datasets have evolved under globally stationary, time-reversible and homogeneous conditions (SRH) for nucleotide and amino acid data (see [13], available upon request from L. Jermini, lars.jermini@csiro.au). The software implements three matched-pairs statistical tests of symmetry and homogeneity across the taxa [37-39]. Misof and colleagues [13] found that a large proportion of their dataset did not evolve under SRH conditions, but excluding the most heterogeneous blocks had only minor local effects on tree topology (compare the results of supermatrix C and D, Supplementary Materials). This empirical example shows that compositional heterogeneity may not have much effect on phylogenetic results, but there is no doubt that this is dataset dependent. Phylogenetic studies on arthropods [9, 40, 41] and Lepidoptera [42, 43] provide clear examples of compositional heterogeneity influencing phylogenetic estimates, which can result in artificial groupings with 100% bootstrap support (e.g., Regier *et al.* Fig. 3 [43]). Comparisons between the topologies of trees derived from phylogenetic analysis and those derived from analysis of compositional distance alone, can be used to assess and pin-point potential influence of compositional heterogeneity on phylogenetic estimates [44]. Within arthropods, compositional heterogeneity is largely restricted to synonymous nucleotide changes, i.e., those differences in codons that don't translate to different amino acids [9]. An efficient approach to eliminating synonymous changes only, and therefore to minimise compositional heterogeneity at the nucleotide level, is degeneracy coding (refer to [9, 44], see www.phylotools.com).

A number of recent studies have shown that there can be strong signal for alternative resolutions in a phylogeny, even for nodes with high bootstrap support [13, 16]. Four-cluster likelihood quartet mapping (FCLM, see Strimmer and von Haeseler [45]) has been used to detect alternative signals in these datasets. One example is that FCLM analyses in Misof *et al.* [13] found some support for a sister group relationship between mayflies and Neoptera supporting the Chiasmomyaria hypothesis, see [17], though being incompatible with their ML tree reconstructions supporting Palaeoptera.

aTRAM – easing the computational burden of genome assembly

The *de-novo* assembly of entire genomes for non-model organisms is difficult and time consuming, in particular if no reference genome from closely related species is available for guidance. Most phylogenomic studies utilise only protein-coding single-copy genes, which make up only a tiny fraction of the entire genome. To avoid the difficult assembly of large genome regions that are not used in phylogenomic analyses, Johnson and colleagues [46] published a BLAST-based computational approach that restricts short reads generated from whole genomes to a subset needed to assemble a set of targeted genes (TRAM: Target Restricted Assembly Method). Software for an improved and parallelised version (aTRAM: automated TRAM) is available for download [47].

Ultra Conserved Elements, Anchors and Exons

Entire genomes and transcriptomes provide the most promising data source for deep-level phylogenetics and higher classification, but are too costly to sequence for adequate sampling of the enormous biodiversity of insects at lower taxonomic levels. Methods of genome reduction are required to cut costs for broader taxon sampling, and target enrichment with hybrid probes has emerged as the most cost effective approach (anchored hybrid enrichment, see [48, 49]). Compared to transcriptomes, about ten times more samples per Illumina HiSeq lane can be sequenced for many hundreds of targeted genes. However, the greatest benefit of this approach is the less stringent requirement for sample quality. Transcriptome studies start with intact mRNA that has to be obtained from live, flash-frozen or specifically preserved specimens (e.g., ground in RNAlater or RNA-Shield), typically requiring the dedicated collection of live specimens. In contrast, different hybrid probe approaches can utilise samples that are stored in ethanol (e.g., molecular collections, trap catches, soft-bodied insects) or even old, dried specimens from traditional pinned insect collections. These approaches have the potential to expand museum collections worldwide into molecular treasure troves that hold many millions of identified specimens amenable to target enrichment with hybrid probes and covering the entire known global diversity of insects.

Current applications of hybrid probes for phylogenomics fall into two categories. Firstly, direct targeting is the traditional approach where probe design is typically based on the exon sequences of protein-coding single-copy genes (exome capture) that are derived from orthologous genes in whole genome or transcriptome data, e.g., Ilves and López-Fernández [50]. These genomes and transcriptomes should ideally include the most divergent lineages of the taxonomic group of interest, because the probes based on these sequences will only efficiently capture matching sample DNA that differs by no more than 10-15% [51, 52] (for probe design software BaitFisher, Oliver Niehuis and Christoph Mayer, pers. comm.). Consequently, probes should be designed for a targeted taxon divergence level, but multiple designs for different levels can be included in a capture. Because probes cover targeted genetic regions in their entirety, this method is suitable for capturing even small fragments from degraded museum specimens, e.g., [51, 53-55]. Such probes can be generated by, e.g., producing clean PCR products for relatively few targeted regions [52, 56-58-57], by Restriction site Associated DNA (RAD) probes [59], or much more costly by synthesising many smaller probes that tile the entire sequence of many hundreds of targeted regions [60, 61] .

The second category of hybrid probes targets either "ultra conserved elements" [62-65] or "anchored elements" [48, 49]. These similar methods require genome or transcriptome data from ideally ingroup and outgroup taxa to identify highly conserved regions for probe design that are flanked by more variable, phylogenetically informative regions. The methods can be applied to relatively large DNA fragments (e.g., from fresh, ethanol preserved or recently dried specimens) and are not suitable for strongly degraded DNA (e.g., old museum specimens). Because of the conserved nature of the probe target, these probes can be universally suitable for highly divergent taxa. The desired variation of the by-caught flanking sequences often spans a range of divergences [47], potentially providing informative characters across multiple taxonomic levels, but with fewer data for any particular level as a trade-off.

Conclusions

The first entomological phylogenetic analyses using genome and transcriptome data [13, 23] have been published, establishing a backbone-phylogeny for Hexapoda and Holometabola respectively. Several entomological studies including transcriptome data (www.1kite.org) and anchored hybrid enrichment mainly covering holometabolous insects are currently in preparation. We expect these to mark a new era in which molecular phylogenies of the mega-diverse insect lineages will have much broader and more representative taxon sampling across all levels of classification. Analyses of these datasets are in their infancy, and this review demonstrates that their complexity commands caution and careful inspection before and after analysis. We have outlined a few of the recently developed techniques to interrogate, mask, analyse and interpret these datasets, and the resulting phylogenetic trees. We predict that insect phylogenomic analyses will become much more sophisticated, and produce more reliable results, in the near future.

Acknowledgments

DKY, MT and KM are funded by Schlinger Foundation Trusts to The Australian National Insect Collection (CSIRO) and The Californian Academy of Sciences, USA. BM, MT and DY acknowledge support from grant # DEB-1257960 "Beyond *Drosophila*: Comparative transcriptomics of schizophoran flies to resolve a rapid radiation" from the National Science Foundation, USA.

References

- [1] Trautwein MD, Wiegmann BM, Beutel RG, Kjer KM, Yeates DK: **Advances in insect phylogeny at the dawn of the postgenomic era.** Annu Rev Entomol 2012, **57**:449-468.
- [2] Yeates DK, Cameron SL, Trautwein, M: **A view from the edge of the forest: recent progress in understanding the relationships of the insect orders.** Austral Entomology 2012, **51**: 79-87
- [3] Behura SK: **Insect phylogenomics.** Insect Mol Biol 2015, **24**:403-411.

[4] Oppenheim SJ, Baker RH, Simon S, DeSalle R: **We can't all be supermodels: the value of comparative transcriptomics to the study of non-model insects. *Insect Mol Biol* 2015, **24**:139-154.
The authors review of insect comparative transcriptomics focusing on functional transcriptomics relevant in insect evolution, and identifying the importance of rapid changes in the overall pattern of gene expression.

[5] Ozsolak F, Milos PM: **RNA sequencing: advances, challenges and opportunities**. *Nat Rev Genet* 2011, **12**:87-98.

[6] Misof B, Meusemann K, von Reumont BM, Kück P, Prohaska SJ, Stadler PF: **A priori assessment of data quality in molecular phylogenetics. *Algorithms Mol Biol* 2013, **9**:22.
The authors summarise challenges in large-scale datasets and identify possible methods to optimize data matrices by filtering out data that cannot be distinguished from random, and unevenly distributed missing data.

[7] Friedrich M, Tautz D: **Ribosomal DNA phylogeny of the major extant arthropod classes and the evolution of myriapods**. *Nature* 1995, **376**:165-167.

[8] Ertas B, von Reumont BM, Wägele JW, Misof B, Burmester T: **Hemocyanin suggests a close relationship of Remipedia and Hexapoda**. *Mol Biol Evol* 2009, **26**:2711-2718.

[9] Regier JC, Shultz JW, Zwick A, Hussey A, Ball B, Wetzer R, Martin JW, Cunningham CW: **Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences**. *Nature* 2010, **463**:1079-1083.

[10] Meusemann K, von Reumont BM, Simon S, Roeding F, Strauss S, Kück P, Ebersberger I, Walz M, Pass G, Breuers S *et al.*: **A phylogenomic approach to resolve the arthropod tree of life**. *Mol Biol Evol* 2010, **27**:2451-2464.

[11] Rota-Stabelli O, Campbell L, Brinkmann H, Edgecombe GD, Longhorn SJ, Peterson KJ, Pisani D, Philippe H, Telford MJ: **A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata**. *Proc Biol Sci* 2011, **278**:298-306.

[12] von Reumont BM, Jenner RA, Wills MA, Dell'Ampio E, Pass G, Ebersberger I, Meyer B, Koenemann S, Iliffe TM, Stamatakis A *et al.*: **Pancrustacean phylogeny in the light of new phylogenomic data: support for Remipedia as the possible sister group of Hexapoda**. *Mol Biol Evol* 2012, **29**:1031-1045.

[13] Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG *et al.*: **Phylogenomics resolves the timing and pattern of insect evolution. *Science* 2014, **346**: 763-767.

The authors provide the first “phylogenomic” insect backbone tree based on 1,478 orthologous single-copy protein-coding genes covering all hexapod orders with an exhaustive and well designed analytical approach.

[14] Oakley TH, Wolfe JM, Lindgren AR, Zaharoff AK: **Phylotranscriptomics to bring the**

understudied into the fold: monophyletic Ostracoda, fossil placement, and Pancrustacean phylogeny. Mol Biol Evol 2011, **30**:215-233.

[15] Dell'Ampio E, Meusemann K, Szucsich NU, Peters RS, Meyer B, Borner J, Petersen M, Aberer AJ, Stamatakis A, Walz MG *et al.*: **Decisive data sets in phylogenomics: lessons from studies on the phylogenetic relationships of primarily wingless insects. Mol Biol Evol 2014, **31**:239-249.

The authors provide important insights into the design of meaningful phylogenomic datasets and demonstrate that artifacts can appear if missing data is unevenly distributed or there is insufficient sequence overlap among taxa of interest.

[16] Simon S, Strauss S, von Haeseler A, Hadrys H: **A phylogenomic approach to resolve the basal pterygote divergence.** Mol Biol Evol 2009, **26**: 2719-2730.

[17] Simon S, Narechania A, Desalle R, Hadrys H: **Insect phylogenomics: exploring the source of incongruence using new transcriptomic data.** Genome Biol Evol 2012, **4**:1295-1309.

[18] Letsch HO, Meusemann K, Wipfler B, Schütte K, Beutel R, Misof B: **Insect phylogenomics: results, problems and the impact of matrix composition.** Proc R Soc London Ser B 2012, **279**:3282-3290.

[19] Letsch H, Simon S: **Insect phylogenomics: new insights on the relationships of lower neopteran orders (Polyneoptera).** Syst Entomol 2013, **38**:783-793.

[20] Beutel RG, Friedrich F, Hörnschemeyer T, Pohl H, Hünefeld F, Beckmann F, Meier R, Misof B, Whiting MF, Vilhemsen L: **Morphological and molecular evidence converging upon a robust phylogeny of the megadiverse Holometabola.** Cladistics 2011, **27**:341-355.

[21] Wiegmann BM, Trautwein MD, Kim JW, Cassek BK, Bertone MA, Winterton SL, Yeates DK: **Single-copy nuclear genes resolve the phylogeny of the holometabolous insects.** BMC Biol. 2009, **7**:34. doi: 10.1186/1741-7007-7-34.

[22] Peters RS, Meusemann K, Petersen M, Mayer C, Wilbrandt J, Ziesmann T, Donath A, Kjer KM, Aspöck U, Aspöck H, *et al.*: **The evolutionary history of holometabolous insects inferred from transcriptome-based phylogeny and comprehensive morphological data. BMC Evol Biol 2014, **14**:52. doi: 10.1186/1471-2148-14-52.

The first phylogeny including all orders of holometabolous based on transcriptome and morphological data.

[23] Niehuis O, Hartig G, Grath S, Pohl H, Lehmann J, Tafer H, Donath A, Krauss V, Eisenhardt C, Hertel J *et al.*: **Genomic and morphological evidence converge to resolve the enigma of Strepsiptera. Curr Biol 2012, **22**:1309-1313.

This is the first study based on complete genomes including the Strepsiptera (twisted winged parasites) based on more than 4000 orthologous genes and including other characters (e.g., near intron pairs) placing Strepsiptera as closest relatives of Coleoptera with high support.

- [24] Whiting MF, Carpenter JC, Wheeler QD, Wheeler WC: **The Strepsiptera problem: phylogeny of the holometabolous insect orders inferred from 18S and 28S ribosomal DNA sequences and morphology.** Syst Biol 1997, **46**:1-68.
- [25] Johnson BR, Borowiec ML, Chiu JC, Lee EK, Atallah J, Ward PS: **Phylogenomics resolves evolutionary relationships among ants, bees, and wasps.** Curr Biol 2013, **23**:2058-2062.
- [26] Bazinet AL, Cummings MP, Mitter KT, Mitter CW: **Can RNA-Seq resolve the rapid radiation of advanced moths and butterflies (Hexapoda Lepidoptera: Apoditrysia)? An exploratory study.** PLoS ONE 2013, **8**:e82615. doi: 10.1371/journal.pone.0082615
- [27] Kawahara AY, Breinholt JW: **Phylogenomics provides strong evidence for relationships of butterflies and moths.** Proc Biol Sci 2014, **281**:20140970.
- [28] Delsuc F, Brinkmann H, Philippe H: **Phylogenomics and the reconstruction of the tree of life.** Nat Rev Genet. 2005, **6**:361-375.
- [29] Philippe H, Brinkmann H, Lavrov DV, Littlewood DT, Manuel M, Wörheide G, Baurain D: **Resolving difficult phylogenetic questions: why more sequences are not enough.** PLoS Biol 2011, **9**:e1000602. doi: 10.1371/journal.pbio.1000602.
- [30] Hillis DM, Pollock DD, McGuire JA, Zwickl DJ: **Is sparse taxon sampling a problem for phylogenetic inference?** Syst Biol 2003, **52**:124-126.
- [31] Brinkmann H, van der Giezen M, Zhou Y, Poncelin de Raucourt G, Philippe H: **An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics.** Syst Biol 2005, **54**:743-757.
- [32] Thomas JA, Trueman JW, Rambaut A, Welch JJ: **Relaxed phylogenetics and the palaeoptera problem: resolving deep ancestral splits in the insect phylogeny.** Syst Biol 2013, **62**:285-297.
- [33] Jiang W, Chen SY, Wang H, Lia DZ, Wiens JJ: **Should genes with missing data be excluded from phylogenetic analyses?** Mol Phylogen Evol 2014, **80**:308-318.
- [34] Misof B, Meyer B, von Reumont BM, Kück P, Misof K, Meusemann K: **Selecting informative subsets of sparse supermatrices increases the chance to find correct trees.** BMC Bioinformatics 2013, **14**:348. doi:10.1186/1471-2105-14-348.
- [35] Jermini LS, Ho SYW, Ababneh F, Robinson J, Larkum AWD: **The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated.** Syst Biol 2004, **53**:638-643.
- [36] Roure B, Philippe H: **Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference.** BMC Evol Biol 2011, **14**:11. doi: 10.1186/1471-2148-11-17.

- [37] Bowker AH: **A test for symmetry in contingency tables.** J Am Stat Assoc 1948, **43**: 572-574.
- [38] Stuart A: **A test for homogeneity of the marginal distributions in a two-way classification.** Biometrika 1955, **42**:412-416.
- [39] Ababneh F, Jermini LS, Ma C, Robinson J: **Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences.** Bioinformatics 2006, **22**:1225-1231.
- [40] Regier JC, Shultz JW, Ganley ARD, Hussey A, Shi D, Ball B, Zwick A, Stajich JE, Cummings MP, Martin JW, Cunningham CW: **Resolving arthropod phylogeny: Exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence.** Syst Biol 2008, **57**:920-938.
- [41] Regier, JC, Zwick, A: **Sources of signal in 62 protein-coding nuclear genes for higher-level phylogenetics of arthropods.** PLoS ONE 2011, **6**:e23408. doi: 10.1371/journal.pone.0023408.
- [42] Regier JC, Mitter C, Zwick A, Bazinet AL, Cummings MP, Kawahara AY, Sohn JC, Zwickl DJ, Cho S, Davis DR *et al.*: **A large-scale, higher-level, molecular phylogenetic study of the insect order Lepidoptera (moths and butterflies).** PLoS ONE 2013, **8**:e58568. doi: 10.1371/journal.pone.0058568.
- [43] Regier JC, Mitter C, Davis DR, Harrison TL, Sohn JC, Cummings MP, Zwick A, Mitter KT: **A molecular phylogeny and revised classification for the oldest ditrysian moth lineages (Lepidoptera: Tineoidea), with implications for ancestral feeding habits of the mega-diverse Ditrysia.** Systematic Entomology 2015, **40**:409-432.
- [44] Zwick A, Regier JC, Zwickl DJ: **Resolving discrepancy between nucleotides and amino acids in deep-level arthropod phylogenomics: differentiating Serine codons in 21-amino-acid models.** PLoS ONE 2012, **7**:e47450. doi: 10.1371/journal.pone.0047450.
- [45] Strimmer K, von Haeseler A: **Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment.** Proc Natl Acad Sci USA 1997, **94**:6815-6819.
- [46] Johnson KP, Walden KK, Robertson HM: **Next-generation phylogenomics using a target restricted assembly method.** Mol Phylogenet Evol 2013, **66**:417-422.
- **[47] Allen JM, Huang DI, Cronk QC, Johnson KP: **aTRAM - automated target restricted assembly method: a fast method for assembling loci across divergent taxa from next-generation sequencing data.** BMC Bioinformatics 2015, **16**:98.
- The software tool aTRAM uses a targeted approach to assembly loci of interest across distantly-related taxa without needing a reference or draft genome.
- [48] Lemmon AR, Emme SA, Lemmon EM: **Anchored hybrid enrichment for massively high-throughput phylogenomics.** Syst Biol 2012, **61**:727-744.
- **[49] Lemmon EM, Lemmon AR: **High-Throughput Genomic Data in Systematics and**

Phylogenetics. Annu Rev Ecol Evol Syst 2013, **44**:99-121.

This review summarizes protocols and data-analysis workflows for phylogenomics including different wet-lab approaches and addressing locus selection and model choice.

[50] Ilves KL, López-Fernández H: **A targeted next-generation sequencing toolkit for exon-based cichlid phylogenomics.** Mol Ecol Resour 2014, **14**:802-811.

[51] Mason VC, Li G, Helgen KM, Murphy WJ: **Efficient cross-species capture hybridization and next-generation sequencing of mitochondrial genomes from noninvasively sampled museum specimens.** Genome Res 2011, **21**:1695-1704.

[52] Peñalba JV, Smith LL, Tonione MA, Sass C, Hykin SM, Skipwith PL, McGuire JA, Bowie RCK, Moritz C: **Sequence capture using PCR-generated probes: a cost-effective method of targeted high-throughput sequencing for nonmodel organisms. Mol Ecol Resour 2014, **14**:1000-1010.

Presents a modified approach to sequence capture using PCR-generated probes that are flexible and economical for up to 100 loci, and suitable for a wide range of taxa.

[53] Guschanski K, Krause J, Sawyer S, Valente LM, Bailey S, Finstermeier K, Sabin R, Gilissen E, Sonet G, Nagy ZT *et al.*: **Next-generation museomics disentangles one of the largest primate radiations.** Syst Biol 2013, **62**:539-554.

[54] Bi K, Linderöth T, Vanderpool D, Good JM, Nielsen R, Moritz C: **Unlocking the vault: next-generation museum population genomics. Mol Ecol 2013, **22**:6018-6032.

The authors address the use, advantages and disadvantages of target enrichment techniques on degraded DNA typical of historical samples and show that such techniques can enhance the value of museum collections.

[55] Enk J, Devault A, Kuch M, Murgha Y, Rouillard JM, Poinar H: **Ancient whole genome enrichment using baits built from modern DNA.** Mol Biol Evol 2014, **31**: 1292-1294.

[56] Noonan JP, Coop G, Kudaravalli S, Smith D, Krause J, Alessi J, Chen F, Platt D, Pääbo S, Pritchard JK, Rubin EM: **Sequencing and analysis of Neanderthal genomic DNA.** Science 2006, **314**:1113-1118.

[57] Maricic T, Whitten M, Pääbo S: **Multiplexed DNA sequence capture of mitochondrial genomes using PCR products.** PLoS ONE 2010, **5**:e14004. doi: 10.1371/journal.pone.0014004.

[58] Horn S: **Target enrichment via DNA hybridization capture.** Methods Mol Biol 2012, **840**: 177-188.

[59] Suchan T, Pitteloud C, Gerasimova N, Kostikova A, Arrigo N, Pajkovic M, Ronikier M, Alvarez N: **Hybridization capture using RAD probes (hyRAD), a new tool for performing genomic analyses on museum collection specimens. 2015. doi: <http://dx.doi.org/10.1101/025551>.

The authors introduce a new approach and efficient for generating hybrid probes for target enrichment using Restriction Site associated DNA (RAD) sequence data.

[60] McCormack JE, Harvey MG, Faircloth BC, Crawford NG, Glenn TC, Brumfield RT: **A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing.** PloS ONE 2013, **8**:e54848. doi: 10.1371/journal.pone.0054848.

[61] Smith BT, Harvey MG, Faircloth BC, Glenn TC, Brumfield RT: **Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales.** Syst Biol 2014, **63**: 83-95.

[62] Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC: **Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales.** Syst Biol 2012, **61**: 717-726.

[63] Faircloth BC, Branstetter MG, White ND, Brady SG: **Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. Mol Ecol Resour 2015, **15**:489-501.

The authors identify and provide a large set of ultraconserved elements (UCEs) shared among Hymenoptera which they used for hybrid enrichment to reconstruct phylogenetic relationships spanning more than 200 Mya. They also summarize the methods used in previous studies using UCE for hybrid enrichment.

[64] McCormack JE, Faircloth BC, Crawford NG, Gowaty PA, Brumfield RT, Glenn TC. **Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species tree analysis.** Genome Res 2012, **22**:746-754.

[65] McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT: **Applications of next-generation sequencing to phylogeography and phylogenetics.** Mol Phylogen Evol 2013, **66**:526-538.

Fig. 1 Number of released and registered transcriptome and genome Bioprojects in Genbank, NCBI.

A) Insect phylogenetic tree, modified from Fig. 1, Misof *et al.* 2014. Number of released and registered transcriptome- and genome projects in Genbank, Bioprojects, NCBI. Source: www.ncbi.nlm.nih.gov/bioproject/; assessed 28 August 2015. Note that bioprojects can cover more than one species, so the numbers are an underestimate of the species covered. Different Bioprojects in addition can cover same species. For transcriptome data, the 1KITE project contributes ca. 1260 species, for which sequencing has been finished (see http://www.1kite.org/1kite_species.php). Since not all have been registered or released yet, the contribution of 1KITE is not fully reflected in the table. Several genome sequencing projects are contributed by the i5K project (<http://arthropodgenomes.org/wiki/i5K>). Only ~35% of all registered transcriptome projects are available for download. For released data, also data in the Sequence Reads Archive (SRA) are available; registered projects include projects for which data are released.

¹ Hymenoptera: numbers excluding *Apis*, all Hymenoptera: 117 released, 465 registered transcriptome projects; 35 released, 60 genome projects;

² Coleoptera: numbers excluding *Tribolium*; all Coleoptera: 78 released, 236 registered transcriptome projects; 12 released, 24 registered genome projects;

³ Lepidoptera: numbers excluding *Bombyx*; all Lepidoptera: 194 released, 338 registered transcriptome projects; 35 released, 48 registered genome projects;

⁴ Diptera excluding *Drosophila*; all Diptera 574 released, 1523 transcriptome projects; 191 released, 246 genome projects. ⁵ Psocodea includes Psocoptera and Phthiraptera.