Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Insect Science

Advances using molecular data in insect systematics

Karl Kjer[1], Marek L Borowiec[2], Paul B Frandsen[3], Jessica Ware[1] and Brian M Wiegmann[4]

The size of molecular datasets has been growing exponentially since the mid 1980s, and new technologies have now dramatically increased the slope of this increase. New datasets include genomes, transcriptomes, and hybrid capture data, producing hundreds or thousands of loci. With these datasets, we are approaching a consensus on the higher level insect phylogeny. Huge datasets can produce new challenges in interpreting branch support, and new opportunities in developing better models and more sophisticated partitioning schemes. Dating analyses are improving as we recognize the importance of careful fossil calibration selection. With thousands of genes now available, coalescent methods have come of age. Barcode libraries continue to expand, and new methods are being developed for incorporating them into phylogenies with tens of thousands of individuals.

Addresses
[1] Rutgers University, Department of Biological Sciences, 415 Boyden Hall, Newark, NJ 07012, USA
[2] University of Rochester, 226 Hutchison Hall, Rochester, NY 14627, USA
[3] Smithsonian Institution, Office of Research Information Services, Office of the Chief Information Officer, Washington, D.C. 20024, USA
[4] Department of Entomology & Plant Pathology, North Carolina State University, Raleigh, NC 27695, USA

Corresponding author: Wiegmann, Brian M (bwiegman@ncsu.edu)

Current Opinion in Insect Science 2016, **18**:40–47

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

Edited by **Gregory W Courtney** and **Brian M Wiegmann**

For a complete overview see the Issue and the Editorial

Available online 14th October 2016

http://dx.doi.org/10.1016/j.cois.2016.09.006

Introduction

While molecular data has revolutionized the higher level phylogeny of many taxa, Börner's [1] phylogeny of insects was remarkably close to our current understanding. Subsequent morphological treatments refined insect phylogeny, and corroborated many nodes [2] but differences among hypotheses had been difficult to resolve. The period from 1995 to 2010 was dominated by molecular studies from Sanger sequencing. Datasets typically ranged from 1000 to 10 000 nucleotides and usually included nuclear rRNA, one or two mitochondrial genes, and sometimes one or two nuclear single copy genes. Seemingly revolutionary discoveries from this period, such as Nonocculata, Halteria, and mecopteran paraphyly, have not been confirmed by much larger phylogenomic datasets [3]. Conflict was amplified by philosophical and analytical differences, recently reviewed [4]. At the dawn of the phylogenomic age, much controversy over higher-level insect phylogeny remained, and we awaited the age of 'big data' to mediate our differences.

Datasets of extraordinary size are now common in phylogenetics, involving hundreds or thousands of genes. The size of datasets has been growing exponentially since the earliest studies in the 1980s (Figure 1) and the recent works [3–9] are converging on consensus in higher level insect phylogeny (Figure 2).

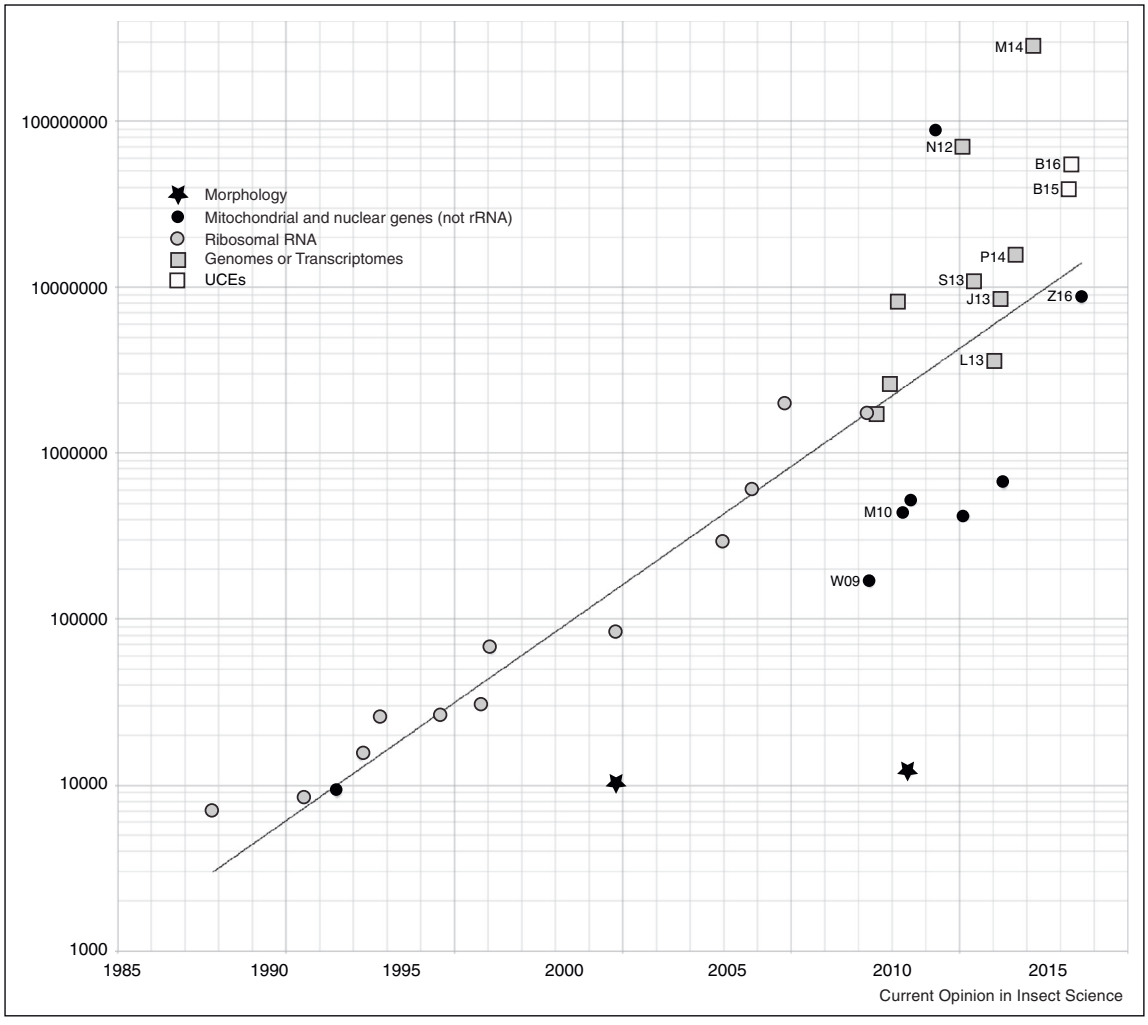
Branch support and confidence

Branch support measures such as bootstrap values and posterior probabilities provide confidence when a range of values are needed to distinguish signal from stochasticity. However, with very large datasets, stochasticity is effectively eliminated, and support values are often 100%. Of course, this is not a bad thing. However, even small biases amplified by millions of nucleotides can result in strong support for erroneous results. Recognizing this, Misof *et al.* [3] examined model mis-specification [10] and used quartet mapping [11], which led them to question certain strongly supported relationships, most notably the monophyly of Palaeoptera and the placement of Psocodea as sister to Holometabola. Congruence among data sources can also be used for building confidence for a particular hypothesis. Although it is often stated that morphological data should simply be mapped onto molecular trees, we find this opinion to be too limiting, especially for insects, whose morphological characters are abundant, and still accumulating with advanced techniques such as micro computed tomography (μCT) [12]. Morphological data can provide corroboration and focus attention on problematic nodes. Corroboration can also come from embryological data [13] or from comparative studies of spermatozoa [14]. Despite its utility, congruence does not give us a quantitative or statistically meaningful value, and conflict from quartet mapping or the discovery of model misspecification simply provides reason for skepticism. We are still looking for a more meaningful method of branch support for very large datasets.

Data partitioning and model selection

As phylogenetic datasets increase in size, the variation in the pattern of evolution among those data also increases.

Figure 1



Size of phylogenetic datasets through time. *X*-axis: dates of publication of selected phylogenetic works. *Y*-axis: number of sites multiplied by the number of taxa. For transcriptome datasets this number was then multiplied by (1 minus the percent missing data). 'Ribosomal RNA' points frequently included other data. W09 = [5]; M10 = [6]; N12 = [77]; J13 = [42]; S13 = [7]; L13 = [8]; P14 = [9]; M14 = [3]; B15 = [66]; B16 = [68]; Z16 = [50]. Others are as in Kjer *et al*. [4], Fig. 5B.

Two methods are often used for accounting for this increased variation: mixture models and partitioning. For large datasets, partitioning is arguably the more popular method, though some studies have used mixture model methods on large datasets [15,16]. Partitioned models account for evolutionary variation by estimating independent model parameters for different subsets of sites within a concatenated alignment. In early model based molecular studies, the decision of how to partition the dataset was generally made *a priori* by the researcher based on some known biological feature of the data, for example, gene boundaries, codon positions within a gene, or stems and loops of rRNA. The process of determining these boundaries is sometimes referred to as 'more of an art than a science' [17]. More recently, algorithms have been proposed for the selection of partitioning schemes from pre-defined data blocks [18,19]. These heuristic algorithms join the pre-defined data blocks and accept joins based on whether they improve the score of an information theoretic metric such as AICc or BIC. For algorithms like these, the more data blocks pre-defined, *a priori*, the better (since subsequent improvements are made by joining data blocks). For example, Misof *et al.* [3] identified protein domains as their initial data blocks with the argument that the domain, not the gene, is the unit of selection, and showed that partitioning by domain outperformed partitioning by gene. Other methods have explored estimating partitioning schemes without *a priori* knowledge of partitioning boundaries and, instead, partition by clustering sites with similar site patterns or rates [20,21].