Phylogenomic Insights into the Cambrian Explosion, the Colonization of Land and the Evolution of Flight in Arthropoda

CHRISTOPHER W. WHEAT1,2,* AND NIKLAS WAHLBERG3

1Department of Biosciences, PL 65, Viikinkaari 1, 00014 University of Helsinki, Finland; 2Department of Zoology, Stockholm University, S-106 91 Stockholm, Sweden and 3Laboratory of Genetics, Department of Biology, University of Turku, FI-20014 Turku, Finland

*Correspondence to be sent to: Department of Biosciences, PL 65, Viikinkaari 1, 00014 University of Helsinki, Finland;
E-mail: christopherwheat.net.

Received 11 May 2012; reviews returned 30 July 2012; accepted 20 August 2012

Associate Editor: Brian Wiegmann

Abstract — The timing of the origin of arthropods in relation to the Cambrian explosion is still controversial, as are the timing of other arthropod macroevolutionary events such as the colonization of land and the evolution of flight. Here we assess the power of a phylogenomic approach to shed light on these major events in the evolutionary history of life on earth. Analyzing a large phylogenomic dataset (122 taxa, 62 genes) with a Bayesian-relaxed molecular clock, we simultaneously reconstructed the phylogenetic relationships and the absolute times of divergences among the arthropods. Simulations were used to test whether our analysis could distinguish between alternative Cambrian explosion scenarios with increasing levels of autocorrelated rate variation. Our analyses support previous phylogenomic hypotheses and simulations indicate a Cambrian origin of the arthropods. Our results provide insights into the 3 independent colonizations of land by arthropods and suggest that evolution of insect wings happened much earlier than the fossil record indicates, with flight evolving during a period of increasing oxygen levels and impressively large forests. These and other findings provide a foundation for macroevolutionary and comparative genomic study of Arthropoda. [Arthropod, Bayesian relaxed clock, Cambrian explosion, comparative genomics, macroevolution, phylogenomics, timing of divergence]

Arthropoda, as the largest animal phylum, may be considered the most successful group of living animal species. It includes the largest class of metazoans on earth, Insecta, which comprise more than half of all described species (Grimaldi and Engel 2005). Beyond sheer numbers, arthropods also exhibit an incredible range of morphological and ecological diversity and can be found in nearly all habitats on our planet. Understanding the evolutionary history of Arthropoda is therefore central to understanding the tempo and mode of evolution on earth. However, resolving the evolutionary history of the early arthropods is difficult as the fossil record is generally scant and episodic (Grimaldi and Engel 2005; Budd and Telford 2009). The timing of the origin of arthropods and its relation to the Cambrian explosion is still controversial, as are several macroevolutionary events, such as the colonization of land and the evolution of flight. Here we assess the power of a phylogenomic approach to shed light on these major events in the evolutionary history of life on earth.

Origins: A Short- or Long-Fuse Cambrian Explosion?

Prior to the Cambrian, very few metazoan body or trace fossils have been identified (Briggs and Fortey 2005). This paucity of metazoan fossils in the strata of Earth is broken by the sudden appearance of highly developed metazoan fossils in the Cambrian, a pattern colloquially referred to as the Cambrian evolutionary “explosion” (Conway Morris 2006). The cause of this “explosion” remains an unsolved question (Briggs and Fortey 2005), interpreted by some as evidence of a very dramatic evolutionary radiation, with short evolutionary divergence times among the ancestral lineages of the Cambrian fauna. Continuing with the explosion metaphor, this interpretation is referred to as a “short-fuse” scenario (Fig. 1a; Gould 1989; Conway Morris 2006). Others view this “explosion” as a historical artifact and argue for a deeper taxonomic age of arthropods, with longer divergence times among ancestral lineages extending back into the Precambrian. This alternative is referred to as a “long-fuse” scenario (Briggs and Fortey 2005). Although the short-fuse scenario necessarily requires a dramatic evolutionary radiation, the long-fuse scenario allows for either a gradual diversification or an older, yet still rapid radiation (Fig. 1b, c).

Numerous studies over the past 2 decades have worked to illuminate the origins of the Cambrian explosion using molecular data. Generally, molecular estimates of metazoan ages have been older than fossil based estimates, and among them there is considerable differences in the length of their Cambrian fuse, much of which can be attributed to the sampling and analytical biases (Bromham 2006). Significant advances in molecular dating with the advent of relaxed clock methods suggest the potential for analyses that are significantly less biased than previous studies (Drummond et al. 2006; Battistuzzi et al. 2010). In fact, a series of relaxed clock studies have reported molecular-based estimates claiming congruence with Cambrian fossil estimates (Aris-Brosou and Yang 2002; Aris-Brosou and Yang 2003; Peterson et al. 2004; Peterson et al. 2008). These results have been welcomed as long sought evidence of a possible congruence between molecular and fossil data (Briggs and Fortey 2005; Budd 2008). However, subsequent re-analysis of these studies identify these claims to derive from spurious results...
FIGURE 1. Representation of the competing hypotheses for the ancient evolutionary radiation of the Arthropoda. Squares are the first instance of a given taxonomic group in the fossil record. In gray are indicated the temporal positions of the primary Cambrian Konservat-Lagerstatte sites of the Burgess Shale and Chengjiang, along with the termination of the last global glaciation event (i.e., snowball earth). a) short-fuse; b) long-fuse, gradual; c) long-fuse, rapid. Phylogenetic relationships from Regier et al. (2010), fossils dates from Briggs and Fortey (2005).

(Blair and Hedges 2005; Ho et al. 2005), or result from excessive restrictions on priors that overly bias posterior estimates (Sanders and Lee 2010).

A very recent study took a different approach by exploiting EST databases, sacrificing even taxon coverage for large numbers of orthologous gene regions (Rehm et al. 2011). The times of divergence in that study were based on a fixed topology taken from a previous study (Meusemann et al. 2010b), and the evolution of amino acids was calibrated by giving 6 nodes minimum ages and one node both a minimum and a maximum age, but the distributions of these age priors are not clear from the paper. The results suggest that the arthropods originated in the Precambrian some 590 Ma. However, several aspects of this study complicate interpretation. First, an EST approach can result in a dataset with a high percentage of missing data that can be problematic for analyses (Sanderson et al. 2010). Second, the taxon coverage of the several important taxa is restricted to 1 or 2 individuals (e.g., Myriapoda, Pycnogonida and Pulmonata) or is missing crucial taxa (e.g., Xenocarida of Pancrustacea). Third, the use of hard constraints for age priors can lead to overparameterization and result in an inability of the actual molecular data to have meaningful effects on posterior estimates (Ho and Phillips 2009; Heled and Drummond 2012).

When Arthropods Colonized Land and Evolved Flight

Arthropods appear to be the first multicellular animals to colonize land and 3 groups are abundant in many ecosystems even today, i.e., hexapods, chelicerates, and myriapods. However, uncertainty remains regarding when (Labandeira 2005) and how many times ancient arthropods colonized land. The first identifiable terrestrial animal fossils, which were the ancestors of chelicerates and myriapods, appear in the late Silurian (ca. 419 Ma; Jeram et al Edwards 1990). The first fossilized tracks that are clearly terrestrial occur earlier, in the early Ordovician (ca. 490 Ma), and are mostly likely left by chelicerate ancestors (MacNaughton et al. 2002). The phylogenetic position of the hexapods, whose fossils appear in the Devonian (400 Ma), has been controversial, with one hypothesis suggesting that they are sister to the myriapods, which would infer one colonization of land by their common ancestor (Averof and Akam 1995). However, it is becoming increasingly clear that insects are a clade within crustaceans, suggesting independent colonization of land by myriapods and insects, as well as by chelicerates (Regier et al. 2010). Crustaceans, such as woodlice, land crabs, and coconut crabs, have presumably colonized land more recently and multiple times (Wildish 1988; Schubart et al. 1998; Labandeira 2005).

Few molecular studies have explicitly addressed the question of when and how many times land was colonized. Pisani et al. (2004) suggest that the ancestor of chelicerates and myriapods colonized land relatively late in their evolutionary history, i.e., very close to the time suggested by the fossil record. However, that study suffers from sparse taxonomic sampling (n = 8 species), which has been shown to affect estimates of times of
Divergences (Hug and Roger 2007). Other studies do not explicitly address the colonization of land (Sanders and Lee 2010; Rehm et al. 2011), but their estimates of times of divergence of the three major groups suggest that land was colonized much earlier, i.e., in the Cambrian.

Winged insects (Pterygota) are by far the most successful group of terrestrial arthropods (they comprise 98.5% of known hexapods; Mayhew 2003; Grimaldi and Engel 2005), evidently due to their ability to fly. Insects were the first group of organisms to take wing and their flight appears to have evolved only once. Although a diverse range of winged insect fossils are found in the late Carboniferous (ca. 325 Ma), and insect fossils are known before this period at ~420 and 390 Ma, the intervening time period is one of exceptional fossil paucity across nearly all life, known as Romer’s Gap (see Fig. 1 in Ward; Labandeira et al. 2006). As a result, very little is known about when flight evolved (Mayhew 2003). The recent study by Rehm et al. (2011) suggests that Pterygota diverged from its sister group ~450 Ma, which is in strong conflict with the fossil record.

Despite these promising advances, many molecular-based studies suffer from analytical biases, such as (i) assumptions of molecular-clock-like behavior (Aris-Brosou and Yang 2002; Welch and Bromham 2005), (ii) assumptions of no rate variation among positions (Welch and Bromham 2005), (iii) the use of priors that bias posterior estimates (Blair and Hedges 2005; Welch et al. 2005), (iv) systematic overestimation of interior, basal branches (Blair and Hedges 2005; Ho et al. 2005), and potentially (v) violations of model assumptions of either random- or autocorrelated rate variation (Battistuzzi et al. 2010). In general, the interpretation of many phylogenetic studies is complicated by some combination of these biases.

Here, we use an established Bayesian-relaxed molecular clock approach (Drummond et al. 2006) and explicitly avoid and quantify the previously mentioned biases by (i) using soft bounds on priors (Ho and Phillips 2009), (ii) assessing the effect of priors on posterior estimates, (iii) conducting simulations to assess how autocorrelated rate variation could affect our posterior estimates of the Cambrian explosion [as this violates the underlying assumptions of our analysis (Battistuzzi et al. 2010)], and (iv) assessing the effect of our fossil calibrations on temporal estimates (Hug and Roger 2007). Our data sampling takes advantage of phylogenomic advances in Arthropoda (Table 1) by using an extended phylogenomic dataset based on the study by Regier et al. (2010) for a computationally intensive, simultaneous estimation of phylogenetic relationships and times of divergence.

### Materials and Methods

#### Data

Sixty-two genes, identified as single copy within Arthropoda (Regier et al. 2010), were used in BLAST searches to find homologous sequences in 3 different types of databases [whole-genome
sequence (WGS), WGS predicted gene sets, and assembled transcriptomes from EST sequencing projects; see Table S1 for details, available online at: http://datadryad.org/review?wfID=3040&token=836474a-df6e-400b-8f86-9e0e5a90139]. We use data from all of the publicly available arthropod genomes (n = 25) and several EST databases (n = 17). Species scientific name, common name, length of DNA sequence obtained, data source (genomic vs. transcriptomic), database, and source are all provided (Supplementary Table S1).

Gene sequence from WGS was obtained by first identifying the beginning and end of the chromosomal region containing the gene of interest, which was then clipped and fed into the gene prediction program Genescan [http://genes.mit.edu/GENSCAN.html; (Burge and Karlin 1997)]. Using custom python scripts, identified genes collected from each species database were aligned with reference sequence using Blastalign software, which maintained codon positions, and concatenated. Final alignment was doublechecked by eye and third positions removed.

Time

Bayesian inference of phylogeny and times of divergence were performed using the BEAST v1.5.4 software package (Drummond and Rambaut 2007). Datasets were analyzed as one partition under the GTR+Γ model with a relaxed clock allowing branch lengths to vary according to an uncorrelated log-normal distribution (Drummond et al. 2006). The branch lengths were also allowed to vary according to an exponential distribution to assess the effects of changing priors on results. The tree prior was set to the birth–death process. Initial runs with BEAST showed that arthropods did not remain monophyletic with regard to the outgroup taxa Tardigrada and Onychophora. Because the monophyly of Arthropoda is not in question, we constrained it to be monophyletic in the analyses. All other priors, except calibration points described below, were left to the defaults in BEAST. Parameters were estimated using 26 independent runs of 10–20 million generations each, with parameters sampled every 2000 generations. Convergence and effective sample sizes of parameter estimates were checked in the Tracer v1.4.6 program; trimmed datasets were combined to yield output files with 147,465 sampled generations, from which summary trees were generated using TreeAnnotator v1.5.3.

Fossil Calibration Points

Analyses were based on 8 calibration points with their priors modeled as a normal distribution, with means set to the estimated age of the fossil and standard deviations about these means were set to give confidence intervals ±5% of the fossil age. The use of a mean distribution for priors, without upper or lower constraints, reflects the uncertainty in the fossil record and allows posterior estimates to vary in either direction based on their interactions with the other calibration points during analysis (Sanders and Lee 2007; Ho and Phillips 2009). The fossil calibrated nodes are as follows: (i) the first split in Pycnognonida set to a mean of 425 Ma (SD 11), based on the fossil Haliestes (Arango and Wheeler 2007), (ii) the first split in Pulmonata (spiders and scorpions) with a mean of 417 Ma based on the fossil Prosorhopus (Dunlop, Tettie and Lorenzo 2008), (iii) the node defining Communodonta (barnacles, crabs, lobsters, and woodlice) with a mean of 425 Ma (SD 11) based on the fossil Rhaphoverritor (Briggs et al. 2005), (iv) the first split in Chilopoda (centipedes) at 417 Ma (SD 11) based on the fossil Crassolum (Edgecombe and Giribet 2007), (v) the first split in Vericrustacea (fairly shrimp, copepods, barnacles, and crabs) with a mean of 516 Ma (SD 11) based on the fossils Yicaris and Rehbachiella (Olesen 2004; Zhang et al. 2007; Møller et al. 2008); (vi) the first split in Hexapoda was given a mean of 425 Ma (SD 7), based on the first known fossil of a hexapodan from the late Devonian/early Silurian (Grimaldi and Engel 2005), (vii) the first split in Holometabola with a mean of 300 Ma (SD 11) based on a fossil gall on a fern frond attributed to a holometabolous insect (Labandeira and Phillips 1996), and (viii) the first split in Diptera with a mean of 230 Ma (SD 11) based on the fossil Gruvogella arzwillariana (Krzeminski et al. 1994). We note that calibrations 1–5 have been used in Sanders and Lee (2010), calibrations 6 and 8 have been used by Rehm et al. (Rehm et al. 2011), and calibration 8 was used by Wiegmann et al. (2009).

We also tested the effects of using uniform priors instead of normal priors for the calibration points. In these cases, the above times were given as minimum bounds and a maximum bound was given as 100 myr older than the minimum bound.

Simulation Analyses

Previous assessment of the accuracy of relaxed clock methods for temporal inference has explored the effects of node constraints (i.e., fossil dates), tree topology, and taxonomic sampling (Hug and Roger 2007). Taxonomic subsets were found to give nearly identical temporal inferences to the full datasets when enough taxa for phylogenetic inference were used (i.e., tree topology as the same in the subset taxa dataset) and coupled with the same node constraints (Hug and Roger 2007). This is an important observation when assessing the temporal inferences from large phylogenetic datasets, as the BEAST runs of our full dataset (122 taxa for 27,984 bp) with a sufficient number of chains for convergence and sufficient effective sample sizes (147,465 sampled generations) took ∼3600 h on a high-performance computer cluster. In comparison, a taxonomic subset consisting of 21 taxa for 27,984 bp, representing the minimal taxa necessary to implement our 8 temporal constraints, generally finished in 40 h on the same cluster. This subset approach was necessary, because simulating and running the full dataset is currently
computationally prohibitive: running the subset data alone required roughly 400,000 CPU hours for the final simulation analyses used in this article (40 different parameter settings × 10 runs per parameter setting × 2 modeled scenarios × 12 CPUs per run × 40 h per run).

All simulations used the same topology as our full dataset. However, there were 2 types of trees generated, one with branch lengths based on our observed data (OD) and the other reflecting a Cambrian explosion scenario (CE). CE branch lengths were identical to the OD tree, except for those branch lengths >520 myr old, as they were shortened to conform to a radiation of the basal branches to have occurred within the 20 myr window between 540 and 520 Ma (Supplementary Fig. S1). BEAST is very robust to a range of violations of its lineage rate variation assumptions, except under conditions of autocorrelated rate variation analyzed with an uncorrelated log-normal model (Drummond et al. 2006), which we used in our analyses. Thus, by examining the potential effect of autocorrelated variation on our analyses, we assessed the conditions most likely to have the largest negative influence on our findings.

Autocorrelated rate change assumes the heritability of rate change, with closely related species having more similar rates. Autocorrelated rate change among branches was modeled by modifying the starting tree’s branch lengths. Beginning with the basal branch, new descendant lineage rates were drawn from an exponential distribution having a mean equal to the rate of their ancestral lineage as implemented in RateEvolver (Ho et al. 2005) (which was kindly modified by its creator Simon Ho to fit a topology matching our taxonomic subset). The variance of the distribution from which new rates can be drawn can be modified, as can the probability of a given branch experiencing a rate change and the mutation rate. All simulations had a probability of rate change per branch of 1, except for clock simulations. Our starting tree had branch lengths equal to millions of years, which were multiplied by the new autocorrelated rate change per branch of 1, except for clock simulations. The trimmed output was then combined to yield posterior estimates of mean rate, coefficient of variation in rates, ucld.mean, ucld.stdev, and the mean and 95% highest posterior density (HPD) for age of the arthropods. Please see schematic diagram provided for more details (Supplementary Fig. S1).

**RESULTS**

**Phylogenetic Relationships**

Available arthropod databases were searched for orthologs of the 62 nuclear genes used by Regier et al. (2008, 2010). We identified a total of 854,786 bp sequence data in 42 new species (average = 19,679 bp per species), 41 of which are from Insecta, represented by 25 species from genome sequence databases and 17 species from expressed sequence tag databases (Table S1). Sequences were aligned and concatenated with the Regier et al. dataset, for a total of 122 taxa in our study, and a total of 28% missing data (our new 42 species had an average of 45% missing data). Our sampling increased the representation of the most-species-rich group of macro-organisms, the insects, by 300% compared with the Regier et al. (2010) dataset (which contained 14 species of Insecta). All gene regions included are protein-coding, single copy, and orthologous (Regier et al. 2008), and thus alignment was relatively trivial. For analyses, third codon positions were removed, as they proved to be too variable to be phylogenetically informative (Regier et al. 2008). Analyses concentrated on estimating times of divergence, but topology was estimated concomitantly using the software BEAST (Drummond et al. 2006). The resulting tree file is provided, containing detailed information on node support and age estimates (Supplementary Text S1).

A stable topology is essential to having meaningful estimates of divergence times. Our estimated topology is essentially congruent with that reported by Regier et al. (2010), differing at a few poorly supported nodes (Fig. 2). Major lineages that were previously unrepresented, such as the social insects (Hymenoptera), beetles (Coleoptera), and the flies (Diptera), show expected relationships and reinforce the placement of Hymenoptera as the most basal branching lineage within the holometabolous insects (Fig. 2; Savard et al. 2006; Consortium 2007; Wiegmann et al. 2009; Mutanen et al. 2010). In addition,
Figure 2. Simultaneously estimated phylogenetic relationships and temporal divergences of the arthropods. Posterior probability support for each node is indicated above the branch to the left of the node. The 95% HPD of the node age estimate is given for each node. Nodes that were calibrated with fossils are indicated with orange 95% HPD bars. Taxa with Whole-Genome Sequences available are highlighted yellow; those with comprehensive Expressed Sequence Tag libraries available are highlighted orange. Named clades are discussed in the text.
we find Paraneoptera and Polynoeoptera to both be monophyletic. Thus, increased taxon sampling had no significant effects on topology, while providing greater resolution within Insecta, indicating a stable backbone for temporal inference.

Dates: An Assessment of Prior Bias on Posterior Estimates

Eight calibration points can be placed with relative certainty on the arthropod phylogeny (Fig. 2; Grimaldi and Engel 2005; Wiegmann et al. 2009; Sanders and Lee 2010). These were modeled as true soft constraints, using a normal distribution centered on the fossil age estimate, reflecting the bidirectionality of uncertainty inherent in such calibrations (Ho and Phillips 2009). Mean estimates and their 95% HPD for the age of clades of interest, and for the last common ancestor between pairwise comparisons among model genomic arthropods, were tabulated for easy access (Tables 2 and 3; a detailed pairwise species level list is also included in Supplementary Table S2).

We next assessed how our priors had influenced the posteriors by comparing our data driven posterior estimates to the results of a “null analysis,” which was solely based upon the fossil priors (i.e., without the DNA data). Of the 8 calibration points, the posterior distributions of age estimates were approximately the same as the prior distributions for 4 nodes (Vericrustacea, Commoonostraca, Dipitera, and Pycnogonida), yet had been “updated” by the data resulting in posteriors that were younger for 2 nodes (Pulmonata and Chilopoda; Fig. 3) and older for 2 nodes (Hexapoda and Holometabola; Fig. 3). Such effects were also reflected in analyses using uniform priors, with the first 4 nodes mentioned above having a posterior distribution whose 95% HPD tails were within the uniform prior bounds (not shown), the second 2 having posterior distributions that are highly skewed up against the lower bound, and the latter 2 similarly skewed to the upper bound (Fig. 3). In sum, our data did significantly influence our posterior estimates, suggesting our soft priors did not result in any over parameterization bias in our posterior estimates for other nodes in our tree (Heled and Drummond 2012).

Estimated Ages of Divergences in Arthropoda

The topological robustness and generally low level of rate variation among the branches of our resolved tree (Fig. 4) together provide an excellent foundation for temporal estimation of the age of Arthropoda. No maximal temporal constraints were set in order to minimize bias in our estimate, with the oldest of the 8 calibration points being from Middle Cambrian at 516 Ma for Vericrustacea (a clade containing fairy shrimp, copepods, barnacles and crabs). Based on these constraints, our estimated age for the crown group of Arthropoda was 706 Ma with a 95% HPD of 631–787 Ma (Table 2). The next divergence between Euchelicerata and Mandibulata is estimated to have happened 675 Ma (95% HPD: 608–746 Ma), followed by the divergence between Myriapoda and Pancrustacea at 639 Ma (95% HPD: 580–702 Ma).
<table>
<thead>
<tr>
<th>Species</th>
<th>Ixodes</th>
<th>Locusta</th>
<th>Pediculus</th>
<th>Rhodnius</th>
<th>Aphids</th>
<th>Wasps</th>
<th>Beetles</th>
<th>Choristoneura</th>
<th>Macromoths</th>
<th>Butterflies</th>
<th>Flies</th>
<th>Mosquitoes</th>
<th>Glossina</th>
<th>Drosophila</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ixodes</strong></td>
<td>631–787</td>
<td>—</td>
<td>706</td>
<td>706</td>
<td>706</td>
<td>706</td>
<td>706</td>
<td>706</td>
<td>706</td>
<td>706</td>
<td>706</td>
<td>706</td>
<td>706</td>
<td>706</td>
</tr>
</tbody>
</table>

Notes: Values in upper matrix are mean divergence estimates (Ma); values below are their 95% highest posterior densities. Lesser known species are described under the matrix.

Ixodes scapularis, deer tick, human pest (Arachnida; Ixodida)
Locusta migratoria, migratory locust, agricultural pest (Insecta; Orthoptera)
Pediculus humanus, head louse, human pest (Insecta; Phthiraptera)
Rhodnius prolixus, blood sucking bug, human pest (Insecta; Hemiptera)
Choristoneura fumiferana, spruce budworm, agricultural pest (Insecta; Lepidoptera)
Glossina morsitans, tsetse fly, human pest (Insecta; Diptera)
**Figure 4.** Variation in rates of molecular evolution across branches of the reconstructed arthropod phylogeny. Posterior estimates of branch rates vary between 0.0002 (dark blue) to 0.0013 (red) mutations per myr.

Euchelicerata is estimated to have begun diversifying 533 Ma (95% HPD: 476–592 Ma), while Myriapoda first diversified 538 Ma (95% HPD: 469–614 Ma) and Hexapoda at 432 Ma (95% HPD: 420–445 Ma).

**Test of the Cambrian Fuse Length: “Short” Versus “Long-Gradual”**

Our estimates for the age of Arthropoda and the branch lengths for the subsequent diversification significantly reject alternative hypotheses in favor of the “long-gradual fuse” scenario. The 95% HPD for Arthropoda crown group is entirely within the Cryogenian, with the 3 subsequent bifurcations leading to the major lineages of arthropods having all, or nearly all, of their 95% HPD entirely in the Ediacaran, with mean values for these nodes spanning a range of nearly 150 myr (Fig. 4; Table 2). Given these findings, we used simulations to explore how much confidence we can place in this rejection of the “short-fuse” hypothesis, given the possible violations of the underlying lineage rate variation model. In other words, we investigated how much confidence we could place in the rejection of the “short-fuse” scenario under “worst case” scenarios of our estimation method.

Analysis of our “short-fuse” and “long-gradual fuse” simulations focused on the estimated age of Arthropoda, assessing how close to the fossil Cambrian explosion event it was located. For both simulations, the 95% HPD always contained the simulated age of Arthropoda and the width between the lower and upper 95% HPD values increased significantly with higher levels of autocorrelated rate variation (Fig. 5, Table S3). Such positive effects of autocorrelated rate variation on the 95% HPD interval reflect increasing uncertainty in posterior estimations as autocorrelated rate variance increases, as previously observed (Drummond et al. 2006; Battistuzzi et al. 2010). However, this increase in the 95% HPD interval was not random in our simulations (Fig. 5). In the “short-fuse” scenario, only the upper 95% HPD increased dramatically with increases in autocorrelated rate variation, while the lower 95% HPD remained very stable, which likely derives from the calibration points in the younger areas of the dataset (Fig. 5a). In the “long-gradual fuse” scenario, a similar effect was observed although there was a greater decrease in the lower 95% HPD at high levels of autocorrelated rate variation (Fig. 5b). Importantly, while our observed 95% HPD does not include the Cambrian explosion event, the 95% HPD of the “short-fuse” simulations always contain the Cambrian explosion event, even at levels of autocorrelated rate variation much higher than our observed levels, and in this region the 95% HPD interval is extremely wide reflecting the large uncertainly in estimation.

In sum, our simulations have allowed us to assess the potential negative effects of autocorrelated rate variation on our analyses, as well as the potential limitations of our fossil calibration points. We find our observed branch length insights unaffected by autocorrelation at levels less than and well above the observed autocorrelated levels. Simulations indicate that a “short-fuse” Cambrian explosion could be detected at levels of autocorrelated rate variation nearly 4 times higher than observed levels. Thus, we conclude that we were able to significantly
FIGURE 5. Simulation results showing posterior age estimates and the 95% HPD for Arthropoda under alternative Cambrian evolution scenarios. Each simulation result (mean (circle), lower (square) and upper (triangle) 95% HPD) is derived from the combination of 10 independent topological replicates at a given level of autocorrelated rate variation (except when rate = 0). a) Modeling of the “short-fuse” scenario has all branch lengths <520 Ma and prior age constrains identical to our observed results. Branch lengths of basal lineages >520 Ma were modified to conform to a Cambrian explosion event between 540 and 530 Ma (shown in gray on both panels). b) Simulation results using observed branch lengths. Black filled symbols are results from the observed dataset. The x-axis shows increasing levels of autocorrelated rate variation among branches of the resulting simulated analyses. See Figure S1 for schematic.
distinguish between the alternative “short-fuse” and “long-gradual fuse” alternatives for the origin of the Arthropods (Fig. 1), and find strong evidence rejecting the “short-fuse” scenario (Fig. 5; Supplementary Table S3).

**Discussion**

Estimating times of divergence using molecular data is becoming a routine part of studies in molecular systematics and powerful algorithms designed to perform these analyses continue to be developed. However, the amount of data needed to estimate ancient times of divergence within and between phyla is not clear at the moment. Our empirical results complement simulation results (Battistuzzi et al. 2010) in suggesting that datasets comprising about 60 protein coding gene regions with a wide distribution of fossil calibration points are enough to generate meaningful estimates of times of divergence within a phylum.

**The Age of Arthropoda and Its Ancient Lineages**

Our estimates of the times of the deeper divergences in Arthropoda are younger than most previous attempts to date Precambrian divergences using molecular data and Bayesian relaxed clock methods (e.g., Pisani et al. 2004; Sanders and Lee 2010; Schaefer et al. 2010). Results of the recent phylogenomic analysis (Rehm et al. 2011) are the most comparable to our results based on taxon and character sampling, and we find that some nodes are estimated to be younger and others to be older in our study. We note that Rehm et al. (2011) did not sample some key taxa, based their times of divergence on rates of amino acid changes on a fixed topology taken from Meusemann et al. (2010), and were forced to place an age prior on the root of the topology due to the algorithm they used. In contrast, our study has a more inclusive taxon sampling scheme for basal divergences in Arthropoda (based on Regier et al. 2010), we estimate times of divergence on rates of nucleotide changes while simultaneously estimating topology, and we have no restrictions on the age of the root. Importantly, our phylogenetic hypothesis differs at some crucial nodes compared with that of Meusemann et al. (2010). For instance, Meusemann et al. (2010) sample only 2 Myriapoda and find them to be sister to Pycnogonida+Euchelicerata. They also find Xiphosura (1 sampled taxon) to be sister to Araneae (1 sampled taxon) and these 2 sister to Acari (9 sampled taxa), i.e., horseshoe crabs are found to be within the chelicerate clade. In contrast, Regier et al. (2010), based on maximum likelihood and our study based on Bayesian inference, find Myriapoda (11 sampled taxa) to be sister to Pancrustacea with strong support and Xiphosura (2 sampled taxa) to be sister to the chelicerate clade (which includes 14 sampled taxa of mites, scorpions, and spiders). We find Pycnogonida (5 sampled taxa) to be sister to the rest of Arthropoda, whereas Regier et al. (2010) have it as sister to Euchelicerata. The latter 2 relationships are weakly supported in all analyses.

Our results strongly suggest a Precambrian origin for the arthropods, which is consistent with new fossil finds indicating that spongiform animals had already diverged from eumetazoans at 630 Ma (Maloof et al. 2010). We estimate that the arthropods began diversifying in the Precambrian ∼706 Ma (95% HPD: 631–787), while Rehm et al. (2011) estimate a much younger age for the first split in Arthropoda at 562 Ma (95% confidence interval: 523–640), a result that may be influenced by their root priors.

The Paleozoic is an important era for the diversification of lineages leading to present day classes. Xiphosura (horseshoe crabs) diverged from their sister group Arachnida (mites, scorpions, and spiders) during the Middle Cambrian (Table 2, clade Euchelicerata), as did lineages leading to modern day classes of Myriapoda (centipedes and millipedes; Table 2). The times of divergence of these clades are not entirely comparable to Rehm et al. (2011), as their phylogenetic hypothesis is quite different at these nodes, but in general, their times are fairly similar to our estimates. For instance, their estimate for the divergence of centipedes from millipedes is almost identical to ours, placing it in the late Cambrian.

Early divergences in Pancrustacea (crabs, waterfleas, and insects) appear to have happened just prior to the Cambrian, although Cambrian divergences are not ruled out (Table 2). Rehm et al. (2011) do not sample any Oligostraca, thus they are likely missing the earliest divergence in Pancrustacea. Within Pancrustacea our results suggest that the clade Oligostraca needs attention, as Ostracoda is not monophyletic with regard to Mystacocarida, Pentastomida, and Branchiura. This is in contrast to Regier et al. (2010), who recovered Ostracoda as a monophyletic entity, although with little support. The divergence between Branchiura and Pentastomida is estimated to be in the Permian (Table 2, clade Ichnothyrostrocha), ∼200 myr younger than the recent estimate of Sanders and Lee (2010), although of note here is that the relationships are very different for this region of the topology compared with the latter study. Both Branchiura and Pentastomida are parasitic and have highly modified morphologies, thus the fossils attributed to Pentastomida from the Cambrian (Woloszek et al. 2006) may represent a stem group of the common ancestor of Branchiura and Pentastomida, which we estimate to have diverged from ostracodans in the early Ordovician or even the late Cambrian (445 Ma, 95% HPD: 362–526).

Our taxon sampling scheme covers a large number of insect lineages including all of the early divergences (Grimaolid and Engel 2005). The initial divergence between Entognatha (springtails) and Insecta (insects) appears to have happened in the Silurian (Table 2, clade Hexapoda), and the early divergences in Insecta during the Devonian (Fig. 2), leading to Archaeognatha, Zygentoma, and Pterygota (winged insects). Pterygota
diversified during the late Devonian and early Carboniferous (Fig. 2). This is in stark contrast to Rehm et al. (2011), who found that Pterygota diversified in the late Silurian or early Devonian, despite giving this node a minimum age in the late Carboniferous. Similarly, we find that the most diverse extant group, Holometabola, diversified into the extant orders during the late Carboniferous and early Permian (Fig. 2), whereas Rehm et al. (2011) suggest that they did so much earlier in the late Devonian. As in Regier et al. (2010), our analysis supports the Paleoptera hypothesis of Ephemeroptera (mayflies) being the sister group of Odonata (dragonflies), and the estimated ages of these groups as early Carboniferous (Table 2) is in accordance with the fossil record (Grimaldi and Engel 2005). Again, Rehm et al. (2011) suggest that this group is much older. Our estimate of the age of the divergence of Dermoptera (earwigs) from the common ancestor of Orthoptera (grasshoppers) and Blattodea (cockroaches) is somewhat younger (early Permian; Table 2) than the fossil record would suggest (early Carboniferous), although our credibility interval for the divergence does encompass the early Carboniferous. We estimate that Phthiraptera (lice) diverged from Hemiptera (aphids, bugs) in the early Permian (Table 2) and the Hemiptera diverged into Sternorrhyncha (aphids, white flies, scale insects) and Heteroptera (true bugs) in the early Triassic (Fig. 2). The latter is in line with fossil evidence (Grimaldi and Engel 2005).

The Colonization of Land and the Origins of Flight: Insights from Molecular Data

Our estimated times of divergence provide insights into the 3 independent colonizations of land by arthropods (Fig. 6). The first lineage likely to have colonized land was the common ancestor to myriapods in the early Cambrian (ca. 538 Ma), although the 95% HPD does stretch into the Ordovician (Table 2). Colonization by the chelicere clade Arachnida appears to have occurred during the Cambrian or early Ordovician (Fig. 6, Table 2). These estimates for Arachnida are very close to the proposed ages for the fossilized tracks of potential chelicerates at ~490 Ma (MacNaughton et al. 2002). A recent study on mite evolution and their colonization of land based on limited sampling of only one gene (Schafer et al. 2010) suggests that the common ancestor of mites existed some 570 Ma, which is older than our estimate of the colonization of land by the common ancestor of all arachnids (to which mites belong). The hexapods appear to have colonized land later, during the Ordovician or Silurian (Fig. 6, Table 2), although our estimate is earlier than the earliest fossil hexapods from the early Devonian (Grimaldi and Engel 2005). The crown clade of Hexapoda was one of our calibrated nodes, and the posterior estimates of the time of the first divergence are essentially the same, but somewhat older than the prior used to calibrate the node. Interestingly, a fossil insect believed to be a pterygote has been recorded from the early Devonian (Engel and Grimaldi 2004), suggesting that the age of insects in general may go back to the Silurian or indeed the Ordovician.

Given the overlap in the 95% HPDs among the myriapod and chelicerate lineages, we cannot exclude the possibility that the 2 colonizations happened at about the same time, although this appears unlikely. Rather, based on our mean estimates of the ages of the crown clades, myriapods appear to have left the aquatic life before plants, because fossil evidence for the first terrestrial plants appears roughly 60 myr later at 475 Ma (Gray 1985; Wellman et al. 2003). Arachnida also appears earlier than these plants at 501 Ma. Hexapods on the other hand appeared after the terrestrial plant fossils in the late Ordovician at around 433 Ma (Wellman et al. 2003; Gensel 2008).

Grimaldi and Engel (Grimaldi and Engel 2005) state that “[h]ow, when, and why insect wings originated is one of the most perplexing conundrums in evolution” (p. 158). Insect flight, which attains some of the highest mass-specific aerobic metabolic rates known to science (Sacktor 1976), has been hypothesized to have evolved during a period of hyperoxic conditions from the Late Devonian to the Late Carboniferous (ca. 375–250 Ma; (Dudley 1998). This period is also known for insect gigantism (Briggs 1985), even among flying insects (Carpenter 1992). The higher O2 conditions during this period, indicated in the geochemical record (Berner 1999), are argued to have facilitated the evolution of flight by simultaneously providing higher amounts of O2 for passive uptake by insects and increasing air density for greater lift (Dudley 1998). Although suggestive, molecular data are needed to provide independent confirmation of whether the origins of the Pterygota occurred during these exceptionally high ancient atmospheric oxygen levels or predated this event.

Based on our results, it is clear that the evolution of insect wings happened much earlier than the fossil record would suggest and led to a relatively rapid radiation of insects (Fig. 7). The ancestor to all Pterygota diverged from the common ancestor of Zygentoma (silverfish) in the late Devonian (ca. 384 Ma). The putative Devonian pterygote fossil (Engel and Grimaldi 2004) is very much inline with our estimate of the origin of Pterygota. The initial divergence was followed by a series of diversifications leading to clades with large numbers of species, i.e., the lineages leading to Paleoptera (367 Ma), Polyneoptera (346 Ma), Paraneoptera (330 Ma), Hymenoptera (308 Ma), Coleoptera (288 Ma), and finally Lepidoptera and Diptera (267 Ma). Within 120 myr, the basis for today’s incredible diversity of flying insects was established. Intriguingly, the first 100 myr of this period coincides with the period of increasing atmospheric oxygen levels (Berner 1999), which began rising in the late Devonian and peaked at about the same time as the major holometabolhan lineages diverged from each other in the early Permian. This rise in atmospheric oxygen is attributed to the appearance and spread of large
and woody vascular plants, resulting in an increased burial of organic carbon that in turn formed the most abundant coal deposits in Earth’s history, from which the time period’s name Carboniferous derives (Berner 1999). However, although flight clearly did evolve during a period of increasing oxygen levels and impressively large forests, the initial steps in the evolution of flight are primarily associated with the Devonian just prior to Romer’s Gap, when atmospheric O2 levels were relatively low although fossils of large forests were well established (Willis and McElwain 2002; Ward et al. 2006).

**Dating with Confidence**

Our findings suggest that in order to date with confidence (Drummond et al. 2006), size does matter. The necessity of multiple calibration points has been discussed by a number of authors (Hug and Roger 2007; Ho and Phillips 2009), but the size of a dataset, either number of taxa or number of characters sampled, has not received as much attention in studies estimating times of divergence (Wertheim and Sanderson 2011). As pointed out in the Introduction section, most previous studies have sampled either very few taxa and many gene regions or many taxa and few gene regions. Typical of early studies based on few taxa and few gene regions were much older estimates of divergence times than one would expect based on the fossil record. Increasing the number of gene regions has not appeared to alleviate this problem (Sanders and Lee 2010), unless very strong priors are used to calibrate the tree (Peterson et al. 2008), which may decrease informative input from the data itself (Sanders and Lee 2010). Increasing the number of taxa and relaxing prior constraints in order to get a better representation of the basal, ancient diversifications does appear to help in arriving at more realistic estimates of when they have happened (e.g., Wahlberg et al. 2009; Bell et al. 2010).

Several previous studies that have estimated the age of various clades within the arthropods, which, while using somewhat similar analysis methods to our own, had much smaller phylogenetic breadth, taxonomic depth, and gene sampling (see Introduction section). The exception to this is the recent study based on EST libraries (Rehm et al. 2011). The study by Rehm et al. (2011) differs in several respects to our study, as we have noted above. Our results conflict with theirs in several important respects, our estimate for the age of the basal split in Arthropoda is much older than theirs (ca. 700 vs. ca. 560 Ma) and our estimates for the ages of the basal splits in insects are much younger than theirs (between 50 and 100 myr younger). However, both studies do suggest that arthropods began diversifying in the Precambrian and our simulations suggest that this is a robust result. These age discrepancies for the
Figure 7. The origins and evolution of insect flight. The Insecta chronogram is taken from Fig. 2. The evolution of oxygen content of the atmosphere over time is shown below the chronogram and is based on (Berner 1999). The darker shaded area gives the interval during which major lineages of flying insects diverged.

Major insect lineages are probably explained by the way in which calibrations were implemented. Rehm et al. (2011) appear to have used hard minimum bounds with no maximum limits (either soft or hard), except at one node describing the age of the first split in Diptera, which has both a hard maximum and a hard minimum bound. Compared with the use of these hard bounds, we feel the use of soft bounds with mean ages provides for more interaction between our data and uncertainty in the calibration points during analysis. However, given the exceptional taxonomic diversity of the Arthropoda, estimations of the age of this clade will always be based on an extremely limited subset of taxa. Whether larger datasets of taxa and genes continue to disagree or
Substantial changes in substitution rates along branches could, in theory, complicate our ability to date with confidence, especially when rate changes are autocorrelated (Drummond et al. 2006; Battistuzzi et al. 2010). Using simulation analyses, we directly addressed the potential negative effects of such variation across a range of rate changes, from lower to much higher than observed levels (Fig. 5). We found no effect on the ability to detect a “short-fuse” Cambrian explosion with increasing levels of autocorrelated rate variation, with our observed results being significantly older than the Cambrian explosion. Exploration of similar levels of autocorrelated rate variation in our observed dataset, which is consistent with a “long-fuse, gradual” scenario of arthropod evolution, also finds no significant overlap with the Cambrian explosion unless autocorrelated rates are simulated at much higher levels.

Dating complications could also arise should rate changes be associated with basal nodes close to the Cambrian explosion. Indeed, previous study reported an increase in substitution rate among the basal branches of Arthropoda (Aris-Brosou and Yang 2002; Aris-Brosou and Yang 2003). However, these increases in rate appear to have been spurious results (Ho et al. 2005). Our findings suggest that genome-wide rate variation across the arthropod tree is low when averaged across a large phylogenomic dataset (Fig. 4), especially among the basal branches of the tree. However, previous study using a 5-taxon analysis of genomic data did identify an increased rate of molecular evolution for an internal region of the arthropod phylogeny, in the branches between Coleoptera and Diptera (Savard et al. 2006). Our results agree. The increased taxonomic sampling of our dataset localizes this rate increase to the basal branches of insects with a significant increase in the lineages leading to Diptera and Lepidoptera (Fig. 4). Further detailed sampling coupled with additional fossil data is needed to more accurately resolve the relationship between divergence timing and evolutionary rates among these ancestral branches.

A final concern is that of the age constraint structure and fossil placements when estimating times of divergence (Hug and Roger 2007; Ho and Phillips 2009), both of which can have large effects on results. Our use of multiple true soft constraints, relatively evenly spread throughout the topology and modeled as normal distributions, has allowed us to observe how these constraints affect the posterior results. Looking at the individual nodes with the prior constraints, we find no systematic bias in the effects, with some posterior estimates of ages being older than the priors, some being similar and some being younger than our imposed priors (Fig. 3). By allowing such shifts by using soft constraints, we feel, as others do, that it provides a more robust analysis of the data (Sanders and Lee 2007; Ho and Phillips 2009). Moreover, by including all of these constraints in our simulation analyses, we have simultaneously assessed the interaction of these constraints with various amounts of autocorrelated rate change, finding them sufficient for our temporal investigations.

Given these findings, should the arthropods have arisen quickly under a “short-fuse” Cambrian explosion scenario, our analyses would have detected this event even under some “worst case” scenarios of molecular evolution. Complimentary to this, our observed estimate of the crown age of Arthropoda was significantly older than the Cambrian explosion event (Fig. 5a). In sum, our observations and simulations are consistent with a “long-fuse” scenario of gradual evolution of the Arthropoda during the Precambrian, possibly beginning in the Cryogenian.

CONCLUSIONS
The results presented here provide estimates of times of divergence in the megadiverse phylum of Arthropoda using what we believe to be the most robust estimation methods available. In addition, we explicitly test the Cambrian explosion hypothesis using these methods on simulated data. Our molecular-based results provide independent temporal estimates for the study of macroevolutionary events that are complementary to fossil data. Knowing the age of the arthropods, as well as when subsequent major lineages appeared, provides a powerful tool for studying macroevolutionary events fundamental to our understanding of evolution. Recent fossil finds from the Ediacaran suggest that Metazoans are older than previously thought (Maloof et al. 2010; Yuan et al. 2011), and such discovery is an ongoing process that has continually pushed Metazoan origins deeper in time. This growing body of empirical data is concordant with our results in suggesting that conclusions based on the absence of data, such as the paucity of fossils in the Ediacaran, may be substantially revised over time with new fossil finds.

SUPPLEMENTARY MATERIAL
Supplementary material, including data files and/or online-only appendices, can be found in the Dryad data repository at http://datadryad.org, doi:10.5061/dryad.3r4j45d2.

FUNDING
Funding for this work was provided by the Academy of Finland (grant numbers 131155, 129811).

ACKNOWLEDGEMENTS
A special thank you to R. Robertson and the excellent guides at the Burgess Shale Geoscience Foundation for a tour of the Walcott Quary, which inspired much of this work. We thank Simon Ho for help with his program
RateEvolver, and Joona Lehtomäki and Jussi Nokso-Koivisto for help with Python. We are also grateful for access to the resources of CSC - IT Center for Science, Finland, which were used to perform the Bayesian analyses. We thank associate editor Brian Wiegmann, Karl Kjer and an anonymous reviewer for constructive comments on previous versions of this article.

REFERENCES


Wheat and Wahlberg—Divergence Time Estimates for Arthropoda 109