RESEARCH ARTICLE

Winding up the molecular clock in the genus *Carabus* (Coleoptera: Carabidae) -- Effects of methodological decisions on rate and node age estimation

Carmelo Andújar[1]*, Jesús Gómez-Zurita[2], José Serrano[1]

Address: [1]Departamento de Zoología y Antropología Física. Facultad de Veterinaria, Universidad de Murcia. 30071 Murcia, Spain, [2]Institut de Biologia Evolutiva (CSIC-UPF), Pg. Marítim de la Barceloneta 37, 08003 Barcelona, Spain.

Email: Carmelo Andújar - candujar@um.es; Jesús Gómez-Zurita - j.gomez-zurita@ibe.upf-csic.es; José Serrano - jserrano@um.es

*Corresponding author:

Key words: Molecular clock, rates of molecular evolution, deep node ages, partitioning model, clock model, outgroup selection, Gblocks, mitochondrial genes, nuclear genes, Coleoptera, *Carabus*.

Abstract

Background: Rates of molecular evolution are known to vary across taxa and among genes, and this requires rate calibration for each specific dataset based on external information. Calibration is sensitive to evolutionary model parameters, partitioning schemes and clock model. However, the way in which these and other analytical aspects influence both the rates and the resulting clade ages from calibrated phylogenies are not yet well understood. To investigate these aspects we have conducted calibration analyses for the genus *Carabus* (Coleoptera, Carabidae) on five mitochondrial and four nuclear DNA fragments with 7888 nt total length, testing different clock models and partitioning schemes to select the most suitable using Bayes Factors comparisons.

Results: Results support an origin of the genus *Carabus* during the Oligocene in the Eurasian continent followed by a Miocene differentiation that originated all main extant lineages. We used these data to investigate the effect of ambiguous character and outgroup inclusion on both rates of molecular evolution and time to the most recent common ancestor of *Carabus*. We corroborate the existence of considerable variation in rates of molecular evolution depending on the fragment studied, but also on analytical conditions, including choice of clock model, partitioning scheme, treatment of ambiguous characters, and outgroup inclusion.

Conclusions: The combination of several genes is found the best strategy to minimise both the idiosyncratic behaviours of stand-alone markers and the effect of analytical aspects in rate and age estimations. Our results highlight the importance of estimating rates of

46 molecular evolution for each specific dataset, selecting for optimal clock and partitioning

47 model as well as other methodological issues potentially affecting rate estimation.

48

Background

Time calibration of phylogenetic trees is a key factor to reconstruct the evolutionary history of taxa [1]. This is usually accomplished by extrapolating the known age of a node, e.g. based on fossil data, to the remainder of the tree, and assuming a molecular clock. Some animal groups, such as mammals or birds, are especially suited for time estimation based on a rather complete fossil record, but other organisms, typically species-rich groups of invertebrates with poor or inexistent fossil record, may represent more of a challenge for a similar exercise. In the case of insects, for instance, reliable paleontological evidence is frequently lacking, what leads to apply a proposed standard rate of 2.3% divergence/My for the insect mitochondrial genome [2]. However, the development of independent calibrations analyses, mainly based on the age of geologic events that underlie the origin of particular cladogenetic events, have found rates either slower [e.g., 3-5] or faster [e.g., 6-9] than the above mentioned standard. These studies illustrate how often the routine application of a standard rate may lead to incorrect reconstructions of evolutionary histories.

The discrepancies in the rates of molecular evolution can be attributed to lineage specific effects or to the molecular marker employed [1, 10-12]. However, they could also reflect biases in the calibration procedure [13]. Decisions relative to the analytical procedure with potential effects on estimated rates include the suitability of selected lineage splits and the strategy used to enforce ages to nodes [14-16], methodological aspects such as the method for branch length estimation (i.e., maximum likelihood vs. Bayesian methods; [17, 18]), model of among-branch rate variation (i.e., strict vs. relaxed clock models; [19-22]),

70 selection of evolutionary model [e.g., 9], partitioning of data [e.g., 12, 23], taxon sampling
71 [e.g., 24] or inclusion of ambiguously aligned regions [25]. These studies show that the
72 effects of the methodology have not been fully explored, and highlight the importance of
73 investigating how these factors influence evolutionary rate and node age estimations in real
74 datasets.

75 Within the Coleoptera, the family Carabidae has been the focus of several calibration
76 attempts. For instance, Contreras-Diaz et al. [7] and Ruiz et al. [5] estimated *cox1-cox2*
77 rates of 3.04% and 0.92% divergence/My for Canarian species of *Trechus* and Sphodrini
78 ground beetles, respectively. Prüser and Mossakowski [3] used a strict global clock method
79 and the opening of the Gibraltar Strait at the end of the Miocene to calibrate hypothetic
80 vicariance events between Iberian and North African populations of *Carabus* species. They
81 found rates between 0.39 and 0.98% divergence/My for *nd1* data. Su et al. [26] and
82 Tominaga et al. [27] investigated *nd5* rates for two endemic subgenera of Japanese *Carabus*
83 using a similar approach and the isolation of Japan from the continent at 15 Mya. These
84 authors found a very low evolutionary rate for this gene, 0.28% divergence/My.

85 Indeed, the genus *Carabus* represents a good research subject to conduct comparative
86 calibration analyses, since there are a fair number of available DNA sequences, and the
87 evolutionary history of these apterous beetles can be linked to geologic events that provide
88 multiple potential calibration hypotheses about the origin of clades. In a recent study, we
89 identified eight reliable calibration hypotheses for the genus C*arabus* among a pool of 16
90 such hypotheses, based on their reciprocal consistency and with an *nd5* based phylogeny
91 (Andújar et al., unpublished data; alignment accession numbers XXXX). Here we
92 extrapolate ages of the most recent common ancestor (TMRCA) for three cladogenetic

events, as obtained from the *nd5* data and these eight calibration points, to perform calibration analyses on nine gene fragments that include protein coding and ribosomal genes belonging to both mitochondrial and nuclear genomes. Overall, we have conducted a total of 166 independent calibration analyses on individual and concatenated DNA matrices, using different outgroups, clock models, partition schemes and alternative treatments of ambiguous characters. We used these data to address several specific aims: (i) to obtain a reliable time scale for the origin and evolution of the genus *Carabus*, (ii) to discuss the obtained rates of molecular evolution with those reported in other studies, and (iii) to evaluate the effect of methodological decisions relative to calibration analyses in the resulting calibrated phylogenies.

103 # Methods

104 ***Taxon and gene sampling***

105 Thirty-four specimens belonging to the family Carabidae have been studied (Table 1).
106 Samples correspond to 18 western Palearctic species of the genus *Carabus* representing 13
107 of the 91 conservatively recognized subgenera [28], with at least one representative of each
108 of the eight main subdivisions of the genus [28]. Six taxa were incorporated as outgroups:
109 *Calosoma* as sister group to *Carabus*, *Ceroglossus* and *Cychrus* as related members of the
110 supertribe Carabitae, and *Leistus* and *Laemostenus* as more distantly related taxa. DNA was
111 extracted from a leg of each specimen using the Dneasy Blood and Tissue kit (Qiagen,
112 Hilden, Germany) or Invisorb Spin Tissue Mini Kit (Invitek, Berlin, Germany) following
113 manufacturers' instructions.

114 Each specimen was characterized for nine DNA fragments corresponding to seven different
115 ribosomal and protein coding genes from mitochondrial (*cox1-A, cox1-B, nd5, cytb, rrnL*)
116 and nuclear (*LSU-A, LSU-B, ITS2, HUWE1*) genomes (Table S1). The sequences of *HUWE1*
117 are homologous to the Anonymous gene described by Sota and Vogler [29] for the genus
118 *Carabus*, and to the predicted HUWE1 gene as identified by BLAST searches against the
119 *Tribolium castaneum* genome. PCR reactions were made using PuReTaq Ready-To-Go PCR
120 beads (GE Healthcare, UK) or Qiagen Taq Polymerase with 39 cycles at 50-54 ºC for primer
121 annealing. Purification of PCR products and sequencing in both directions with the same
122 primers used for PCR was performed by Macrogen Inc. (Seoul, Korea). Sequence accession
123 numbers are given in Table 1.

124 ***Sequence alignment***

125 Mitochondrial protein coding genes were unambiguously aligned and checked for their
126 correct translation to amino acids using Mega 4 [30]. The 5'-end of the *HUWE1* fragment
127 was also unambiguously aligned and correctly translated to amino acids, while the 3'-end
128 showed length variation. All ribosomal markers (*rrnL, LSU-A, LSU-B* and *ITS2*) also showed
129 length variation and required objective alignment prior to phylogenetic analysis.

130 Variable-length DNA fragments for the dataset including outgroups were aligned under each
131 combination of five iterative refinement methods (FFT-NS-i, E-INS-i, G-INS-I, L-INS-I and
132 Q-INS-I) and three scoring matrices (1-PAM, 20-PAM and 200-PAM) in Mafft 6.240 [32,
133 33]. Each individual alignment was assessed for congruence with respect to a combined
134 matrix including unambiguously aligned regions of every gene. To get this combined matrix,
135 every fragment was independently aligned in MAFFT with FFT-NS-i parameters, and local
136 ambiguities were removed with Gblocks [31] with the "No-gaps" option and other default
137 parameters; the resulting fragments were concatenated. Congruence was measured using
138 both the incongruence length difference index (ILD; [34]) and the rescaled ILD [35]. ILD
139 values were estimated from parsimony based tree lengths in every case using PAUP* 4.0
140 [36]. The alignment conditions maximizing character congruence for every length-variable
141 marker, i.e. producing the lowest rescaled ILD value, were objectively selected as those
142 generating the best homology hypothesis for these data and employed in subsequent
143 analyses.

144 Favoured alignments were used to produce three concatenated matrices, including all
145 mitochondrial fragments (MIT), all nuclear fragments (NUC) and both datasets (MIT-NUC).
146 In the case of ribosomal genes, selected alignments were previously processed with Gblocks

147 [31] using the *All-gaps* option and default parameters; only selected positions were included
148 in the concatenated matrices.

149 ***Phylogenetic analyses***

150 All phylogenetic and calibration analyses described below were run independently for both
151 the complete dataset (*outgroup* dataset) and a subset of data representing only the 28 ingroup
152 *Carabus* taxa (*ingroup* dataset).

153 Bayesian phylogenetic inference for each individual and concatenated dataset, without
154 specifying partitions and without clock assumptions, were run with MrBayes 3.1 [37,38] to
155 assess the reliability of nodes to be used in subsequent calibration tests. A complex
156 GTR+Γ+I model was used in every analysis as their parameters are co-estimated with the
157 tree and they can match eventually any simpler model. Analyses enforced two independent
158 runs, each with three hot and one cold chain, for 20,000,000 generations, whereby trees were
159 sampled every 1,000 generations. Convergence of independent runs was checked in Tracer
160 1.5 [39] and their half compatible consensus tree was calculated excluding 10% of initial
161 trees, after the plateau in tree likelihood values had been reached. Trees were visualized
162 using FigTree 1.1.2 [40] and node posterior probabilities were interpreted as support values.

163 ***Calibration analyses***

164 Tertiary fossils are scarce for *Carabus*, however there are some biogeographic scenarios with
165 relatively well-known ages that hint to potential vicariance events represented in the
166 phylogeny of this genus. Andújar *et al*. (unpublished data) identified 16 potential calibration
167 points on an *nd5* phylogeny of the subtribe Carabinae and applied a novel Bayes Factors
168 Cluster Analysis to select a subset of eight consistent calibration points as the most reliable

169 for subsequent calibration tests. Using the same *nd5* gene dataset and these mutually
170 congruent calibration points, we have conducted BEAST calibration analyses
171 (Supplementary materials Table S2) to obtain ages for three specific, well-supported
172 cladogenetic events in the phylogeny of *Carabus*, namely (Fig. 1): node A, the split between
173 *Carabus (Macrothorax) rugosus* and *C. (M.) morbillosus*; node B, the split of *Carabus*
174 *(Mesocarabus) riffensis* from European *Mesocarabus*; and node C, the split between the
175 subgenera *Eurycarabus* and *Nesaeocarabus*. These nodes were selected to be old enough to
176 avoid time dependence effects [41] but not so deep as to be excessively affected by
177 saturation of molecular change. We used TreeStat 1.6.1 [42] to retrieve these node ages from
178 the sample of the MCMC search in BEAST and used the "fitdistr" option of the R package
179 MASS to obtain a gamma function adjusting the distribution of sampled ages. These ages
180 were used as prior age information in subsequent calibration analyses for each DNA
181 fragment and for the concatenated datasets in BEAST 1.5.4 [43]. All calibration analyses
182 were based upon eight independent runs of 50 million generations each, sampling every
183 2,000th generation, and using a Yule tree prior and a GTR+Γ+I evolutionary model, with ten
184 categories for the Γ distribution. Samples from these independent runs were compared,
185 checked for convergence and combined after conservatively removing a 10% of initial trees
186 in Logcombiner 1.5.4 [43], drawing one sample every 16,000th generation. Mean, standard
187 error, highest posterior density intervals (95%HPD) and effective sample size of
188 evolutionary rates and other parameters of interest were inspected using Tracer 1.5.
189 Consensus trees were obtained in TreeAnotator 1.5.4 [43] using the mean age option.

190 *Calibration analyses on individual protein coding genes*. Protein coding genes, including the
191 5'-end of the *HUWE1* fragment, were analyzed under different clock assumptions including

192 strict (SC) and uncorrelated lognormal (ULN) clocks, as well as different codon partition
193 schemes (1P: no partitioning; 2P: two partitions, considering first and second codon
194 positions together; 3P: each codon position as a different partition), for both the "outgroup"
195 and "ingroup" datasets. Additionally, the complete *HUWE1* gene fragment (i.e. including the
196 non-coding 3'-end) was analyzed without codon partitioning. Results were analyzed to
197 simultaneously select the best clock and partition model using Bayes factors (BF). Marginal
198 likelihoods were calculated in Tracer 1.5 and used for the BF comparison, whereby BF were
199 interpreted as requiring at least a ten units increase in marginal likelihood per additional free
200 parameter before accepting a more complex model [44, 45]. We assumed one extra
201 parameter in ULN analyses compared to the SC assumption [20], and ten extra parameters
202 per additional partition under a GTR+Γ+I model.

203 *Calibration analyses on individual ribosomal genes.* Calibration analyses for trees based on
204 ribosomal genes were conducted considering (i) all positions (*complete* dataset), (ii) only
205 positions selected by Gblocks with the *No-gaps* option and other default parameters (*nogaps*
206 dataset), and (iii) only positions selected with the *All-gaps* option (*allgaps* dataset). Each
207 individual analysis was done under different clock assumptions (SC/ULN) and for the
208 *outgroup* and *ingroup* matrices. The best clock model was assessed in every case using BF
209 comparisons as above.

210 *Calibration analyses on concatenated datasets. Outgroup* and *ingroup* datasets for the MIT,
211 NUC and MIT-NUC concatenated matrices were analyzed in BEAST under different clock
212 assumptions (SC and ULN) following different partition schemes, which included no data
213 partitioning (NP), partitioning by gene but not by codon in the case of protein coding genes
214 (G-1P), partitioning by gene, and each protein coding gene with two codon partitions where

first and second positions are together (G-2P), and partitioning by gene and by codon position (G-3P). Among-branch rate variation was always treated as linked among partitions, and BF comparisons were used again to select for the best clock and partitioning models.

The effect of different treatments of data was explored using linear regression analyses on the standardized values of two relevant parameters, including the estimated rate of molecular evolution for the marker or markers investigated and the estimated age of the ingroup node (TMRCA of *Carabus*).

222 # Results

223 ***Phylogenetic framework for calibration tests***

224 The alignment of choice based on congruence optimization with unambiguously aligned data
225 was that obtained implementing the Q-INS-i algorithm for all nuclear genes, except for the
226 *LSU-A* fragment in the case of the *ingroup* dataset. The latter, together with the
227 mitochondrial ribosomal gene *rrnl* data, were optimal using the E-INS-i method (Table S3).

228 Bayesian phylogenetic analyses for the concatenated datasets resulted in the recovery of
229 focal nodes A, B and C with posterior probability of 1.0 in all instances. Node A—split
230 between *Carabus (Macrothorax) rugosus* and *C. (M.) morbillosus*—was found with a
231 posterior probability higher than 0.95 for all individual DNA fragments except for *LSU-B*
232 (pp=0.74) and *LSU-A* (pp<0.5). Node B—split of *Carabus (Mesocarabus) riffensis* from
233 European *Mesocarabus*—appeared highly supported (pp>0.95) for most individual DNA
234 fragments, except for *ITS2* (pp=0.9), *LSU-A* (pp=0.67), *cox1-A* and *cob* (pp<0.5). Finally,
235 node C—split between the subgenera *Eurycarabus* and *Nesaeocarabus*—was recovered for
236 all individual fragments although it showed a posterior probability <0.85 in *nd5, rrnl* and
237 *LSU-B*. Figure 1 shows support of nodes A, B and C as obtained with MrBayes and BEAST
238 analyses.

239 Analysis of individual genes frequently failed to recover the monophyly of the genus
240 *Carabus* (node T) and/or its sister relationship with the genus *Calosoma* (node K)
241 (supplementary figs. S1-S9). However, these nodes were recovered with high support in all
242 combined datasets (Figs. 1-2). Bayesian analyses conducted in BEAST, where the

243 calibration age prior was applied together with the favoured partition and clock scheme,
244 slightly improved node support within the *Carabus* clade.

245 ***Selection of optimal analytical conditions***

246 For each individual and combined dataset and alternative partition and clock model schemes,
247 the eight independent BEAST runs resulted in similar global rates, TMRCA of *Carabus* and
248 likelihood values, reaching stationary equilibrium and ESS values always higher than 500
249 for the likelihood parameter and higher than 200 for all the other parameters, with very few
250 exceptions. The results from these independent runs and for each dataset were thus pooled
251 together to generate samples of 360 million generations, with ESS values always higher than
252 200.

253 BF comparisons resulted in the selection of the 2P codon partition strategy for mitochondrial
254 coding genes, the combination of all mtDNA markers, and the combination of all markers,
255 while no data partitioning was selected for the nuclear *HUWE1* fragment alone and for all
256 nuclear markers combined (Table S4). The strict clock was favoured for all protein coding
257 genes (including the entire *HUWE1* gene fragment), the combined mtDNA including
258 outgroup taxa, and also for some ribosomal genes using the *ingroup* dataset and after gap
259 exclusion. Otherwise, the relaxed (ULN) clock was preferred for all other individual and
260 combined markers (Table S4). Relative to the effect of ambiguously aligned characters or
261 outgroup inclusion/exclusion no direct BF comparisons were possible, and a pragmatic
262 decision was taken in every case. *Nogaps* datasets were discarded due to their major effect
263 on rates and age estimations (see below). *Allgaps* and *complete* matrices showed similar
264 rates and ages, and the first option was selected. Finally, we discarded outgroups for

265 estimation of molecular rates and TMRCA of *Carabus* since this genus was not found
266 monophyletic in most individual marker analyses, producing a net overestimation of the age
267 for this node, taken as the one including all ingroup taxa but not only. When monophyly was
268 recovered, such as in concatenated datasets, the differences in both rates and ages between
269 *outgroup* and *ingroup* datasets was low.

270 ***Evolutionary rates and ingroup ages***

271 Table 2 shows the estimated rates of molecular evolution and the TMRCA of *Carabus* for
272 each gene and their combinations, for the *ingroup* datasets and treated under optimal
273 partitioning and clock assumptions. Rates of mitochondrial genes ranged from 0.0017 (95%
274 HPD 0.0011-0.0025) substitutions per site per My per lineage (subs/s/My/l) for the *rrnL*
275 fragment, to 0.0271 (0.0162-0.0396) subs/s/My/l for the *cob* fragment. Important differences
276 were also found for the estimated rates of nuclear genes, from 0.0013 (0.0008-0.002) for the
277 *LSU-A* gene fragment to 0.0069 (0.0037-0.0107) subs/s/My/l for *LSU-B*. Genes
278 characterized by two non-overlapping fragments (i.e., *cox1*, *LSU*) showed some differences
279 despite using identical approaches. Thus, the 3'-end fragment of *cox1* (*cox1-B*) showed a
280 faster rate of evolution —0.0161 (0.0109-0.0218)— than the *cox1-A* fragment generally used
281 as barcode —0.0126 (0.009-0.0165) subs/s/My/l—. Similarly, *LSU-B* was faster evolving—
282 0.0069 (0.0037-0.0107) subs/s/My/l— than *LSU-A* —0.0013 (0.0008-0.002)—. The rate
283 obtained for the concatenation of all mitochondrial fragments was 0.015 (0.0117-0.0185)
284 subs/s/My/l, roughly equivalent to a divergence rate of 3% per My. The combination of
285 nuclear fragments resulted in a rate of 0.0031 (0.0022-0.0042) subs/s/My/l (i.e., 0.62% per
286 My).

287 The mean estimated age of the ingroup oscillated between 14.8 (9.5-23.0) Mya as estimated
288 for *LSU-A* data, and 30.9 (16.3-48.7) Mya, in the case of *ITS2,* whereas for most of genes
289 this value was between 20 and 30 Ma with widely overlapping 95% HPD intervals (Fig. 1,
290 Table 2). The combined analyses of all genes resulted in a TMRCA of *Carabus* of 23.3
291 (16.5-30.4) Mya. The dispersion of values around the mean was higher for genes analyzed
292 under a relaxed clock. The split between *Carabus* and *Calosoma* (node K) was estimated to
293 have occurred at around 28.1 Mya (95% HPD 22.1-34.8) with the combined genes and the
294 dataset including outgroup taxa (Fig. 2).

295 ***Effect of partitioning scheme***

296 Data partitioning affected protein coding gene fragments (partitioning by codon positions)
297 and concatenated datasets (partitioning by gene and by codon positions). A general trend was
298 observed whereby the average values for the evolutionary rates and the estimated ingroup
299 age increased with the number of partitions considered both for individual markers (linear
300 regression, rate: R=0.420, P=0.001, n=60; TMRCA of *Carabus*: R=0.498, P=0.000, n=60)
301 and their concatenation (rate: R=0.934, P=0.000, n=40; TMRCA of *Carabus*: R=0.770,
302 P=0.000, n=40), and irrespective of the clock model enforced (Figs. 3 and 5). This trend was
303 particularly exacerbated in the case of the rate estimated for *nd5* and the entire dataset, which
304 tripled its value compared to non-partitioned data when three partitions where considered,
305 without remarkable effects on the estimation of the ingroup age. On the other hand, the
306 coding region of the *HUWE1* gene fragment showed invariable rates and estimated ingroup
307 age independently of the partitioning strategy employed. Values for mean rates and TMRCA
308 of *Carabus* and their associated 95% HPD intervals for each analysis are provided in
309 supplementary tables S5 and S6, respectively.

310 ***Effect of ambiguously aligned characters***

311 The effect of gapped characters on rate and node age estimation was assessed in all gene
312 fragments showing sequence length variation (Fig. 4). In the case of ribosomal genes, the
313 exclusion of gapped positions (*nogaps* option in Gblocks) had a noticeable effect lowering
314 the estimates of evolutionary rates and ingroup age, up to three-fold, for the most variable
315 gene fragments, hence with gappier alignments (*ITS2* and *LSU-B*). These gene fragments
316 diminished by 50 and 90% of aligned nucleotide positions, respectively, with a dramatic loss
317 of phylogenetic information and great oscillations in the estimated parameters. Expectedly,
318 more length-conserved fragments (*LSU-A* and *rrnL*) were only slightly affected by character
319 culling in Gblocks, with 27 and 1%, respectively, of character loss under the most
320 conservative *nogaps* treatment, and consequently showed lower variation, particularly in
321 estimation of evolutionary rates. Less restrictive character culling approaches (*allgaps* option
322 in Gblocks) preserved more characters to be used for branch length estimation and produced
323 intermediate results, still with significant effects on rate estimation for highly variable
324 markers, but not so much for that of the TMRCA of *Carabus* in the case of *ITS2* and *LSU-B*
325 data. The estimated mean node age in these cases was affected to a higher extent by outgroup
326 inclusion/exclusion and by the clock model.

327 The simultaneous analysis of non-coding sequence information with exon information for
328 the nuclear *HUWE1* gene had little effect on the estimation of evolutionary rates, although
329 the estimation of the ingroup age decreased or increased when assuming strict or relaxed
330 clocks, respectively (Fig. 4).

331 ***Effect of clock model***

332 The choice of strict *versus* relaxed clock had effects on the estimation of parameters of
333 interest, generally associated with the actual clock model best fitting the data. Individual and
334 combined genes in which the strict clock was preferred showed null to low effect of clock
335 model on the estimation of rates and ingroup ages, with both parameters showing at most a
336 trend to slightly higher values when a relaxed clock was enforced (rate: R=0.263, P=0.035,
337 n=64; TMRCA of *Carabus*: R=0.676, P=0.000, n=64). In all other cases, except for the total
338 evidence dataset, the use of the suboptimal strict clock resulted in lower rate estimates and
339 higher ingroup ages (Figs. 3-5).

340 ***Effect of outgroups***

341 The dataset including outgroups generally resulted in higher rates of molecular evolution
342 compared with the analyses using ingroup data only (R=0.318, P=0.000, n=148) (Figs. 3-5).
343 Exceptions to this pattern affected the fast evolving ribosomal markers *LSU-B* and *ITS2*, as
344 well as the combination of nuclear genes and the total evidence dataset when investigated
345 under a strict clock model. In turn, for the TMRCA of *Carabus* no general trend could be
346 identified (R=0.144, P=0.081, n=148), despite it was generally retrieved as older for most
347 treatments when including outgroups (e.g., individual mitochondrial genes: R=0.549,
348 P=0.000, n=54), a fact associated to most individual gene fragments failing to recover the
349 monophyly of *Carabus.* Exceptionally, this age was younger for all nuclear markers
350 independently (except *LSU-B*) and for their combination (NUC dataset) when the unfavoured
351 strict clock was enforced.

352

352 # Discussion

353 ***A time scale for the origin and evolution of* Carabus**

354 The analyses of MIT, NUC and MIT-NUC combined datasets produce highly congruent
355 topologies and high support for most nodes, including nodes T and K, representing the
356 monophyly of *Carabus* and its sister relationship with *Calosoma*, respectively (Fig. 2). In
357 addition, all combined datasets show nearly complete overlap of the 95% HPD intervals on
358 the estimated ages for these nodes; only the NUC dataset departs slightly from these values
359 producing an older mean age estimate (Fig. 1). The time scale obtained when all nuclear
360 and mitochondrial genes are analyzed together (MIT-NUC dataset) situates the initial split
361 between *Carabus* and *Calosoma* during the Oligocene, some 34 and 23 Ma, after the
362 opening of the Atlantic ocean and the split of the Nearctic and Palearctic regions [46].
363 Timing this split in this specific period and linked to this particular geological event is
364 congruent with the observation that *Carabus*, basically a flighless genus, is more diverse in
365 the Palaearctic region (ca. 800 species) than in the Nearctic (12 species), whereas
366 *Calosoma* is slightly more diverse in the Nearctic region (ca. 90 species) than in the
367 Palearctic (76 species). The evolutionary events that originated the main extant lineages of
368 *Carabus* took place according to our data during the early Miocene, between 23 and 16 Ma
369 (Fig. 2).

370 These findings disagree with the previous hypothesis assigning an older, Eocenic origin to
371 *Carabus* [47]. The occurrence of North American endemic subgenera *Tanaocarabus* and
372 *Lichnocarabus* was interpreted as evidence suggesting that the origin of the genus predated
373 the opening of the Atlantic Ocean, which was interpreted as the event responsible for the

374 isolation of Nearctic from Western Palearctic lineages of *Carabus*. However, molecular
375 data indicate that these Nearctic subgenera are instead related to Eastern Palearctic species
376 [48,49], which suggests an origin for the genus *Carabus* within the Paleartic region and a
377 more recent dispersal event across Bering Strait land bridges, in agree with the dates here
378 provided.

379 ***Rates of molecular evolution on* Carabus**

380 The evolutionary rates for the genus *Carabus* as estimated here are, in general terms, higher
381 than those reported in previous studies. The discrepancies are related to the choice of
382 molecular markers, but mostly to the use of inappropriate calibration points in these studies,
383 combined with simplistic corrections of genetic distances among species. Prüser and
384 Mossakowski [3] in a study based on the *nd1* gene in western Mediterranean species of the
385 genus *Carabus*, calibrated the separation of six pairs of taxa with the opening of the
386 Gibraltar Strait at the end of the Messinian (5.3 Ma). Five of these splits represented the
387 separation of North African and European subspecies in *C. (M.) morbillosus* and *C. (M.)*
388 *rugosus*, and the resulting rates ranged between 0.0020 to 0.0033 subs/s/My/l. However,
389 these splits seemingly occurred well after the opening of the Gibraltar Strait as demonstrated
390 by Andújar et al. (unpublished data). The sixth node represented the split of subspecies of
391 *Carabus (Rhabdotocarabus) melancholicus* from the Iberian Peninsula and North Africa, the
392 only one with compelling evidence to represent a vicariant event resulting from the opening
393 of Gibraltar Strait (Andújar et al., unpublished data). The rate estimated by Prüser and
394 Mossakowski [3] was 0.0049 subs/s/My/l, much lower than our estimated rate for the
395 slowest protein-coding gene —0.0126 (0.009-0.0165) subs/s/My/l—. Indeed, the divergence
396 estimated by Prüser and Mossakowski [3] for these taxa, based on an uncorrected p-distance,

397 was approximately half as low as distances corrected with appropriate evolutionary models
398 (e.g., GTR+Γ model of evolution; [9]).

399 The degree of relaxation of evolutionary constraints acting on different genes or their
400 portions ultimately reflects in differences on their inferred rate of molecular evolution [1,
401 50]. Our data showed an 18-fold difference between the slowest (*rrnL*) and the fastest (*cob*)
402 evolving markers. This important disparity cautions against the extrapolation of rates of
403 molecular evolution among different gene fragments or their combinations. On the other
404 hand, the extrapolation of evolutionary rates for the same marker across taxa, at least when
405 they are relatively closely related, appears as a safer assumption. Several studies using a
406 similar approach as described here produced roughly similar rates of evolution for the same
407 gene fragments, despite targeting different families (and suborders) of Coleoptera. Thus, for
408 instance, Papadopoulou et al. [9] and Ribera et al. [8] estimated rates of 3.54% and 4.08%
409 divergence/Ma in Tenebrionidae and Leiodidae, respectively, for a fragment homologous to
410 the *cox1-B* fragment investigated here, which in our case yielded a rate of 3.22%
411 divergence/Ma —0.0161 (0.0109-0.0218)— in *Carabus*.

412 ***Heterogeneity in rates of molecular evolution***

413 Invertebrate mitochondrial genomes have long been assumed to evolve at a standard
414 molecular clock rate of 2.3% divergence/My [2]. This rate was deduced from heterogeneous
415 mitochondrial data, including restriction fragment length polymorphism (RFLP), DNA-DNA
416 hybridization and sequence data for several genes. Its utilization in phylogenetic studies has
417 been frequent for datasets composed by individual and concatenated mitochondrial DNA
418 gene fragments [e.g., only in the case of beetles and for concatenated data: 7, 51-58]. It is

419 important to notice that in these studies similar rates for different taxa and different regions
420 of the mitochondrial genome were assumed. This assumption has been lately refuted [e.g., 9,
421 11, 12]. Moreover, in former studies where the standard rate is routinely applicated the
422 effects of methodological decisions on the calibration procedure has not been considered, a
423 question that has been shown to be important [18, 23], as we also stress here.

424 There is a fair variation in the rates of individual gene fragments in *Carabus* which are
425 affected by methodological aspects of the calibration procedure. Major differences on both
426 the rates and the TMRCA of *Carabus* are obtained in the case of the nuclear ribosomal
427 genes, reaching a fourfold variation between analyses involving different clock models, as
428 well as inclusion/exclusion of ambiguous characters and/or outgroups. Mitochondrial genes
429 show comparatively less variation due to methodological decisions, and only the *nd5*
430 fragment showed a twofold difference depending on treatment (but see below). The most
431 remarkable observation is that a calibration based on combined datasets results in the lowest
432 effect of prior analytical decisions on both evolutionary rate and node age estimation,
433 together with a reduction of 95% HPD intervals associated to these estimates. This effect
434 hints at the importance of using multiple gene data on calibration exercises, both as a way to
435 average across genes, diluting idiosyncratic behaviours of stand-alone markers, but also to
436 help the analyses to converge into more precise estimates.

437 ***Clock calibration with protein coding genes***

438 The response to changes in analytical conditions differs between protein coding
439 mitochondrial markers and the *HUWE1* gene fragment. Thus, for the latter, the selection of
440 a suboptimal clock model produces marked differences in the estimation of both the rates of

441 molecular evolution and the inferred TMRCA of *Carabus*, while this choice reveals no
442 remarkable effects in the case of the mitochondrial genes. This behaviour can be related to
443 the underlying rate heterogeneity for each marker class: clock-like in the case of mtDNA
444 data, but moderate in that of *HUWE1*. The selection of an unsuitable strict clock for the
445 nuclear gene distorts branch length estimation and all parameters derived from these. In
446 turn, the protein coding nuclear marker shows no apparent changes in the two parameters of
447 interest through different partition schemes, while the mitochondrial genes prove more
448 sensitive to complex evolutionary models, including the effect of adding an outgroup to the
449 estimations or not. As with clock model selection, the partition scheme employed affects
450 branch length estimation, and consequently the resulting inferred rate, a behaviour already
451 reported in other studies [23, 59]. Mitochondrial protein coding genes show a trend to
452 slightly higher inferred evolutionary rates and TMRCA of *Carabus* with increasing
453 partitioning; and the same is true for the MIT and MIT-NUC datasets. The influence of
454 codon site-specific models is similar to that found by Papadopoulou et al. [9], in their study
455 of darkling beetles from the Aegean islands based on two mitochondrial and two nuclear
456 fragments. These authors found up to 11% discrepancy in inferred rates between NP and 3P
457 partitioning, higher for the latter, both under ULN clock assumption and using relatively
458 recent calibration nodes (9-12 Ma), as used in our study. Conversely, other studies found
459 the opposite trend, with younger estimated ages when using complex codon partitioning
460 models for mitochondrial data (e.g., Nearctic and eastern Palaearctic skinks [23]). This
461 trend can be caused by the same underlying problem related to the specific way in which
462 saturation effects are corrected, but combined in this case with the use of deep calibration
463 points (96-148 Ma) to infer younger node ages.

464 There seems to be an effect of the relative position of calibrating nodes on the ages
465 extrapolated along the tree. Saturation and incorrect branch length estimation in deeper
466 parts of the tree despite complex model-based corrections could explain these differences.
467 The use of relative recent calibration nodes should produce reliable age estimates for other
468 nodes in areas of the tree not affected by saturation, with error accumulating progressively
469 for deeper nodes, possibly with underestimated branch lengths, and in this case providing
470 with minimum age estimates. Depth constraints restrain branch stretching in most of the
471 tree, and when deep nodes are affected by incorrect length estimation due to saturation
472 problems, errors in resulting node ages will be extrapolated to more recent parts of the tree,
473 resulting in older than true time estimates for such recent nodes [60].

474 Bayesian phylogenetic inference is known to produce incorrect long branch estimates at least
475 in the case mitochondrial codon-partitioned datasets [17, 61]. These flawed estimations have
476 been found to be stable across independent runs with fixed parameters, but also when
477 Bayesian prior parameters are modified, adding more difficulty to detect them. It appears to
478 be the case of the observed high rate increments when partitioning *nd5* data by codon
479 position (2P and 3P strategies) for both the *ingroup* and *outgroup* datasets. This type of
480 misbehaviour seems to be dependent on the characteristics of particular datasets, so that
481 slight changes in taxon sampling can provide correct estimates [61]. Taking advantage from
482 the availability of *nd5* sequences of *Carabus* in GenBank, we have conducted NP, 2P and 3P
483 partitioning calibration analyses for an additional 58 taxa dataset (alignment accession
484 number XXXX; Andújar et al., unpublished data). These analyses produced moderate
485 increments in the estimated evolutionary rate for *nd5* as more partitions were considered, in

486 agreement with the results for other mtDNA genes, supporting the reliability of these
487 estimates.

488 ***Clock calibration with non-coding genes***

489 The effect of methodological choices is markedly higher for non-coding gene fragments
490 than it is for protein coding genes, on both mean rate and node age estimations, but also on
491 their corresponding 95% HPD intervals. Indeed, calibration exercises based on non-coding
492 genes frequently face alignment ambiguity and departures from the molecular clock, which
493 can both jeopardize reliable branch length estimation. These problems increase with
494 evolutionary distance between taxa, thus raising concerns about deep node calibration using
495 slowly evolving ribosomal genes. Several alignment strategies have been proposed to deal
496 with length variation in homologous DNA sequences (see an evaluation of methods in
497 [62]), and discarding ambiguous DNA positions has been also proposed as a
498 complementary method [31]. We examined the effects of the latter procedure on global rate
499 and node age estimation. Expectedly, character removal, especially when applying
500 restrictive culling options (*nogaps* in Gblocks), has a marked effect which is negatively
501 correlated with marker conservation. Highly variable gene fragments, with gappy
502 alignments, can lose a significant amount of phylogenetic information when filtered out of
503 ambiguous alignment positions, and consequently contribute with unreliable calibrations.

504 As with protein coding genes, there is a strong effect of clock model selection for nuclear
505 ribosomal gene fragments as well. For these markers, SC was unfavoured in most cases and
506 its implementation resulted in lower mean values for inferred evolutionary rates and higher
507 ages (up to four-fold) on the estimated TMRCA of *Carabus*, compared to the use of a

508 favoured ULN clock. While this behaviour is not exclusive of non-coding genes, its
509 prevalence for non-coding data cautions about their use for node age estimation without
510 appropriate accounting for rate heterogeneity in calibration analyses.

511

511 ## Conclusions

512 Mitochondrial genes generally fit the strict clock model of evolution, providing an adequate
513 and simple frame to conduct calibration analyses. However, the combination of several
514 genes including nuclear and mitochondrial fragments results in the best strategy to minimise
515 the effect of both idiosyncratic behaviours of stand-alone markers and analytical aspects on
516 the estimation of evolutionary rates and node ages. But even for mitochondrial genes or
517 combined datasets, rate differences between markers, together with the effect of specific
518 analytical decisions on their estimation, advise against extrapolating rates between different
519 studies. As a summary of the findings reported here, it should be always desirable to check
520 for the appropriate data partition scheme, the same as the underlying evolutionary model for
521 each partition, but also to investigate whether the data really fit a molecular clock previous to
522 any calibration study, applying suitable corrections if needed. The analysis of time on trees
523 should avoid as much as possible regions with dubious homology due to alignment
524 ambiguity, or keep it to a minimum, since gap exclusion has a dramatic effect on branch
525 length estimation and concomitant rates and node ages. Finally, the analyses benefit from
526 enquiry circumscribed to ingroup taxa (and using relatively recent calibration nodes),
527 minimizing the biases introduced by highly divergent lineages and saturation.

528

528 ## Authors' contributions

529 CA and JS collected samples. CA, JG-Z and JS conceived the study, CA and JG-Z
530 designed the analyses, CA gathered the molecular data and performed the analyses. CA and
531 JG-Z and JS wrote the manuscript and all authors read and approved the final version.
532
533

533 ## Acknowledgments

534 This research was funded by projects CGL2006/06706, CGL2009-10906 (JS) and
535 CGL2008-00007 (JGZ) of the Spanish Ministry of Science and Innovation, as well as
536 project 08724PI08 of the Fundación Séneca (Murcia) (JS). CA received the support of FPU
537 predoctoral studentship of the Spanish Ministry of Education. Thanks are due to Obdulia
538 Sánchez, Ana Asensio (University of Murcia) and Gwenaelle Genson (CBGP Montpellier)
539 for their technical assistance. Carlos Ruiz (University of Murcia), Ignacio Ribera (IBE,
540 Barcelona) and specially Paula Arribas (University of Murcia) helped with their comments
541 and advise. Jean-Yves Rasplus (CBGP Montpellier) facilitated *C. (Tachypus) cancellatus*
542 and Achille Casale (University of Sassari) helped with the identification of some taxa.
543 MrBayes was run using the resources of the Computational Biology Service Unit from
544 Cornell University, which is partially funded by Microsoft Corporation; most final
545 computing was carried out using the facilities of the Ben Arabi supercomputer of the
546 Fundación Parque Científico de Murcia.

547

548

Abbreviations

BF - Bayes factor; ***LnBF*** - natural logarithm of the Bayes factor; ***HPD*** - highest posterior density; ***SD*** - standard deviation; ***TMRCA*** -Time to the most recent common ancestor.***NP*** - no partition; ***2P*** - two codon partitions, with 1^{est} and 2º position together; ***3P*** - three codon partitions; ***G-NP*** - partitioning by gene with no codon partition; ***G-2P*** - partitioning by gene, with two codon partition of coding genes with 1^{est} and 2º position together; ***G 3P*** - partitioning by gene, with three codon partition of coding genes; ***SC*** - strict clock; ***ULN*** - uncorrelated log normal clock; ***Nogaps*** - dataset excluding ambiguous character with the *Nogaps* option and default parameters in Gblocks; ***Allgaps*** - idem, applying *Allgaps* option; ***MIT*** - concatenation of mitochondrial genes; ***NUC*** - concatenation of nuclear genes; ***MIT-NUC*** - concatenation of all genes.

559 Reference List

560 1. Bromham L, Penny D: **The modern molecular clock.** *Nat Rev Genet* 2003, **4:**216-
561 224.

562 2. Brower AVZ: **Rapid morphological radiation and convergence among races of**
563 **the butterfly *Heliconius erato* inferred from patterns of mitochondrial-DNA**
564 **evolution.** *Proc Nat Acad Sci USA* 1994, **91:**6491-6495.

565 3. Pruser F, Mossakowski D: **Low substitution rates in mitochondrial DNA in**
566 **Mediterranean carabid beetles.** *Insect Mol Biol* 1998, **7:**121-128.

567 4. Gómez-Zurita J, Juan C, Petitpierre E: **The evolutionary history of the genus**
568 ***Timarcha* (Coleoptera, Chrysomelidae) inferred from mitochondrial COII Gene**
569 **and partial 16S rDNA sequences.** *Mol Phylogenet Evol* 2000, **14:**304-317.

570 5. Ruiz C, Jordal B, Serrano J: **Molecular phylogeny of the tribe Sphodrini**
571 **(Coleoptera: Carabidae) based on mitochondrial and nuclear markers.** *Mol*
572 *Phylogenet Evol* 2009, **50:**44-58.

573 6. Luchetti A, Marini M, Mantovani B: **Mitochondrial evolutionary rate and**
574 **speciation in termites: data on European Reticulitermes taxa (Isoptera,**
575 **Rhinotermitidae).** *Insectes Soc* 2005, **52:**218-221.

576 7. Contreras-Diaz HG, Moya O, Oromi P, Juan C: **Evolution and diversification of the**
577 **forest and hypogean ground-beetle genus *Trechus* in the Canary Islands.** *Mol*
578 *Phylogenet Evol* 2007, **42:**687-699.

579 8. Ribera I, Fresneda J, Bucur R, Izquierdo A, Vogler A, Salgado J, Cieslak A: **Ancient**
580 **origin of a Western Mediterranean radiation of subterranean beetles.** *BMC Evol*
581 *Biol* 2010, **10:**29.

582 9. Papadopoulou A, Anastasiou I, Vogler A: **Revisiting the insect mitochondrial**
583 **molecular clock: the mid-Aegean trench calibration.** *Mol Biol Evol* 2010, **27:**1659-
584 1672.

585 10. Kumar S: **Molecular clocks: four decades of evolution** *Nat Rev Genet* 2005, **6:**654-
586 662.

587 11. Mueller RL: **Evolutionary rates, divergence dates, and the performance of**
588 **mitochondrial genes in bayesian phylogenetic analysis.** *Syst Biol* 2006, **55:**289-
589 300.

590 12. Pons J, Ribera I, Bertranpetit J, Balke M: **Nucleotide substitution rates for the full**
591 **set of mitochondrial protein-coding genes in Coleoptera.** *Mol Phylogenet Evol*
592 2010, **56:**796-807.

593 13. Schwartz RS, Mueller RL: **Variation in DNA substitution rates among lineages**
594 **erroneously inferred from simulated clock-like data.** *Plos One* 2010, **5:**e9649.

595 14. Graur D, Martin W: **Reading the entrails of chickens: molecular timescales of**
596 **evolution and the illusion of precision.** *Trends in Genet* 2004, **20:**80-86.

597 15. Heads M: **Dating nodes on molecular phylogenies: a critique of molecular**
598 **biogeography.** *Cladistics* 2005, **21:**62-78.

599 16. Inoue J, Donoghue PCJ, Yang Z: **The impact of the representation of fossil**
600 **calibrations on bayesian estimation of species divergence times.** *Syst Biol* 2010,
601 **59:**74-89.

602 17. Brown JM, Hedtke SM, Lemmon AR, Lemmon EM: **When trees grow too long:**
603 **investigating the causes of highly inaccurate bayesian branch-length estimates.**
604 *Syst Biol* 2010, **59:**145-161.

605 18. Schwartz RS, Mueller RL: **Branch length estimation and divergence dating:**
606 **estimates of error in Bayesian and maximum likelihood frameworks.** *BMC Evol*
607 *Biol* 2010, **10**.

608 19. Aris-Brosou S, Yang Z: **Effects of models of rate evolution on estimation of**
609 **divergence dates with special reference to the metazoan 18S ribosomal RNA**
610 **phylogeny.** *Syst Biol* 2002, **51:**703-714.

611 20. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A: **Relaxed phylogenetics and**
612 **dating with confidence.** *Plos Biol* 2006, **4:**699-710.

613 21. Battistuzzi FU, Filipski A, Hedges SB, Kumar S: **Performance of relaxed-clock**
614 **methods in estimating evolutionary divergence times and their credibility**
615 **intervals.** *Mol Biol Evol* 2010, **27:**1289-1300.

616 22. Battistuzzi FU, Billing-Ross P, Paliwal A, Kumar S: **Fast and slow implementations**
617 **of relaxed clock methods show similar patterns of accuracy in estimating**
618 **divergence times.** *Mol Biol Evol* 2011.

619 23. Brandley MC, Wang Y, Guo X, Montes de Oca AN, Fería Ortiz M, Hikida T, Ota H:
620 **Accommodating heterogenous rates of evolution in molecular divergence dating**

621 **methods: an example using intercontinental dispersal of Plestiodon (Eumeces)**
622 **Lizards.** *Syst Biol* 2011, **6**: 3-15.

623 24. Linder HP, Hardy CR, Rutschmann F: **Taxon sampling effects in molecular clock**
624 **dating: An example from the African Restionaceae.** *Mol Phylogenet Evol* 2005,
625 **35:**569-582.

626 25. Lemmon AR, Brown JM, Stanger-Hall K, Lemmon EM: **The effect of ambiguous**
627 **data on phylogenetic estimates obtained by maximum likelihood and bayesian**
628 **inference.** *Syst Biol* 2009, **58:**130-145.

629 26. Su ZH, Tominaga O, Okamoto M, Osawa S: **Origin and diversification of**
630 **hindwingless *Damaster* ground beetles within the Japanese islands as deduced**
631 **from mitochondrial ND5 gene sequences (Coleoptera, Carabidae).** *Mol Biol Evol*
632 1998, **15:**1026-1039.

633 27. Tominaga O, Su ZH, Kim CG, Okamoto M, Imura Y, Osawa S: **Formation of the**
634 **Japanese carabina fauna inferred from a phylogenetic tree of mitochondrial**
635 **ND5 gene sequences (Coleoptera, Carabidae).** *J Mol Evol* 2000, **50:**541-549.

636 28. Deuve T: *Illustrated Catalogue of the Genus Carabus of the World (Coleoptera:*
637 *Carabidae).* Sofia-Moscow: PENSOFT; 2004.

638 29. Sota T, Vogler AP: **Incongruence of mitochondrial and nuclear gene trees in the**
639 **Carabid beetles *Ohomopterus*.** *Syst Biol* 2001, **50:**39-59.

640 30. Tamura K, Dudley J, Nei M, Kumar S. **MEGA4: Molecular evolutionary genetics**
641 **analysis (MEGA).** *Mol Biol Evol* 2007 **24:**1596-1599.

642 31. Castresana J: **Selection of conserved blocks from multiple alignments for their use**
643 **in phylogenetic analysis.** *Mol Biol Evol* 2000, **17:**540-552.

644 32. Katoh K, Misawa K, Kuma K, Miyata T: **MAFFT: a novel method for rapid**
645 **multiple sequence alignment based on fast Fourier transform.** *Nucleic Acids Res*
646 2002, **30:**3059-3066.

647 33. Katoh K, Asimenos G, Toh H: **Multiple alignment of DNA sequences with**
648 **MAFFT.** *Bioinf DNA Seq Anal* 2009, 39-64.

649 34. Farris JS, Kallersjo M, Kluge AG, Bult C: **Testing significance of incongruence.**
650 *Cladistics* 1994, **10:**315-319.

651 35. Wheeler WC, Hayashi CY: **The phylogeny of the extant chelicerate orders.**
652 *Cladistics* 1998, **14:**173-192.

653 36. Swofford DL. PAUP*. *Phylogenetic Analysis Using Parsimony (*and Other*
654 *Methods).* Sunderland, Massachusetts: Sinauer Associates; 2000.

655 37. Huelsenbeck JP, Ronquist F: **MrBayes: Bayesian inference of phylogenetic trees.**
656 *Bioinformatics* 2001, **17:**754-755.

657 38. Ronquist F, Huelsenbeck JP: **MrBayes 3: Bayesian phylogenetic inference under**
658 **mixed models.** *Bioinformatics* 2003, **19:**1572-1574.

659 39. Rambaut A, Drummond AJ. **Tracer v1.5** 2007 [http://beast.bio.ed.ac.uk/Tracer]

660 40. Rambaut A. **FigTree v.1.1.2** 2008. [http://tree.bio.ed.ac.uk/software/figtree/]

661 41. Ho SYW, Lanfear R, Bromham L, Phillips MJ, Soubrier J, Rodrigo AG, Cooper A:
662 **Time-dependent rates of molecular evolution.** *Mol Ecol* 2011, **20:**3087-3101.

663 42. Rambaut A, Drummond AJ. **TreeStat v1.6.1: tree statistic calculation tool.** 2010.

664 43. Drummond AJ, Rambaut A: **BEAST: Bayesian evolutionary analysis by sampling**
665 **trees.** *BMC Evol Biol* 2007, **7**.

666 44. Pagel M, Meade A: **A phylogenetic mixture model for detecting pattern-**
667 **heterogeneity in gene sequence or character-state data.** *Syst Biol* 2004, **53:**571-
668 581.

669 45. Miller KB, Bergsten J, Whiting MF: **Phylogeny and classification of the tribe**
670 **Hydaticini (Coleoptera: Dytiscidae): partition choice for Bayesian analysis with**
671 **multiple nuclear and mitochondrial protein-coding genes.** *Zool Scr* 2009, **38:**591-
672 615.

673 46. Ziegler PA: *Evolution of the Arctic-North Atlantic and the Western Tethys.* Tulsa,
674 OK: American Association of Petroleum Geologists Memoir; 1988.

675 47. Turin H, Penev L, Casale A: *The Genus* Carabus *in Europe. A synthesis.* Sofía-
676 Moscow-Leiden: PENSOFT & European Invertebrate Survey; 2003.

677 48. Su ZH, Imura Y, Okamoto M, Kim CG, Zhou HZ, Paik JC, Osawa S: **Phylogeny and**
678 **evolution of Digitulati ground beetles (Coleoptera, Carabidae) inferred from**
679 **mitochondrial ND5 gene sequences.** *Mol Phylogenet Evol* 2004, **30:**152-166.

680 49. Su ZH, Imura Y, Zhou HZ, Okamoto M, Osawa S: **Mode of morphological**
681 **differentiation in the Latitarsi-ground beetles (Coleoptera, Carabidae) of the**
682 **world inferred from a phylogenetic tree of mitochondrial ND5 gene sequences.**
683 Genes Genet Syst 2003, **78:**53-70.

684 50. Ohta T: **Slightly deleterious mutant substitutions in evolution.** *Nature* 1973,
685 **246:**96-98.

686 51. Ribera I, Hernando C, Aguilera P: ***Agabus alexandrae* sp n. from Morocco, with a**
687 **molecular phylogeny of the Western Mediterranean species of the *A. guttatus***
688 **group (Coleoptera : Dytiscidae).** *Insect Syst Evol* 2001, **32:**253-262.

689 52. Barraclough TG, Vogler AP: **Recent diversification rates in North American tiger**
690 **beetles estimated from a dated mtDNA phylogenetic tree.** *Mol Biol Evol* 2002,
691 **19:**1706-1716.

692 53. Ribera I, Vogler AP: **Speciation of Iberian diving beetles in Pleistocene refugia**
693 **(Coleoptera, Dytiscidae).** *Mol Ecol* 2004, **13:**179-193.

694 54. Gómez-Zurita J, Garneria I, Petitpierre E: **Molecular phylogenetics and**
695 **evolutionary analysis of body shape in the genus Cyrtonus (Coleoptera,**
696 **Chrysomelidae).** *J Zool Syst Evol Res* 2007, **45:**317-328.

697 55. Balke M, Gómez-Zurita J, Ribera I, Viloria A, Zillikens A, Steiner J, Garcia M,
698 Hendrich L, Vogler AP: **Ancient associations of aquatic beetles and tank**
699 **bromeliads in the Neotropical forest canopy.** *Proc Nat Acad Sci USA* 2008,
700 **105:**6356-6361.

701 56. Gómez-Zurita J, Funk DJ, Vogler AP: **The evolution of unisexuality in Calligrapha**
702 **leaf beetles: Molecular and ecological insights on multiple origins via**
703 **interspecific hybridization.** *Evolution* 2006, **60**: 328-347.

704 57. Hidalgo-Galiana A, Ribera I: **Late Miocene diversification of the genus *Hydrochus***
705 **(Coleoptera, Hydrochidae) in the west Mediterranean area.** *Mol Phylogenet Evol*
706 2011, **59:** 377-385

707 58. Faille A, Casale A, Ribera I: **Phylogenetic relationships of Western Mediterranean**
708 **subterranean Trechini groundbeetles (Coleoptera: Carabidae).** *Zool Scr* 2011,
709 **40:** 282-295

710 59. Brown JM, Lemmon AR: **The importance of data partitioning and the utility of**
711 **bayes factors in Bayesian phylogenetics.** *Syst Biol* 2007, **56:**643-655.

712 60. Zheng Y, Peng R, Kuro-o M, Zeng, X: **Exploring patterns and extent of bias in**
713 **estimating divergence time from mitochondrial DNA sequence data in a**
714 **particular lineage: A case study of salamanders (Order Caudata).** *Mol Biol Evol*
715 2011, **28:**2521-2535.

716 61. Marshall DC: **Cryptic failure of partitioned bayesian phylogenetic analyses: Lost**
717 **in the land of long trees.** *Syst Biol* 2010, **59:**108-117.

718 62. Golubchik T, Wise MJ, Easteal S, Jermiin LS: **Mind the gaps: Evidence of bias in**
719 **estimates of multiple sequence alignments.** *Mol Biol Evol* 2007, **24:**2433-2442.

720

720 # Figure captions

721 **Figure 1. Ultrametric time-calibrated trees obtained with BEAST for each individual (left) and**
722 **combined datasets (right) of *Carabus*.** Trees obtained with the *ingroup* dataset (including only
723 *Carabus* species) and optimal parameters. Black filled circles in nodes show posterior probabilities
724 (pp) higher than 0.9, bars represent 95% HPD intervals for node ages in Myr. The 95% HPD
725 intervals of the TMRCA of *Carabus* are shaded in grey. Pie charts at the bottom represent support
726 (black, pp≥0.95; dark grey, pp≥0.85; light grey, pp≥0.50; white, node not recovered) in Mrbayes
727 (*outgroup* dataset) and BEAST (*outgroup* and i*ngroup* datasets) for nodes A, B and C, employed as
728 calibration points for the different gene fragments: *cox1-A* (1), *cox1-B* (2), *cob* (3), *nd5* (4), *rrnL* (5),
729 *LSU-A* (6), *LSU-B* (7), *ITS2* (8), *HUWE1* (9), MIT (10), NUC (11) and MIT-NUC (12).

730 **Figure 2**. **Ultrametric time-calibrated tree for combined DNA markers (MIT-NUC dataset) of**
731 **Carabidae.** Analyses were conducted in BEAST including outgroups, partitioning by gene, with two
732 partitions for coding genes (1st and 2nd positions together) and applying a relaxed ULN clock.
733 Numbers besides nodes represent their posterior probabilities. Grey bars on nodes represent the 95%
734 confidence intervals for node ages in Myr. The vertical grey bar shows the 95% HPD interval for the
735 split between *Carabus* and *Calosoma.*

736 **Figure 3**. **Mean rates of molecular evolution and TMRCA of *Carabus* based on protein coding**
737 **genes.** Rates are given in substitutions per site per million years per lineage and TMRCA of *Carabus*
738 in millions of years before present. Different partitioning schemes, clock models and out-group
739 inclusion/exclusion are considered.

740 **Figure 4. Mean rates of molecular evolution and TMRCA of *Carabus* based on non-coding gene**
741 **fragments.** Rates are given in substitutions per site per million years per lineage and TMRCA of
742 *Carabus* in millions of years before present. Different clock models, ambiguous character and out-
743 group inclusion/exclusion are considered.

744 **Figure 5**. **Mean rates of molecular evolution and TMRCA of *Carabus* based upon the combined**
745 **datasets.** Rates are given in substitutions per site per million years per lineage and TMRCA of
746 *Carabus* in millions of years before present. Different partitioning schemes, clock models and out-
747 group inclusion/exclusion are considered.

748

Table 1: **Species, locality of collection, voucher reference and accession numbers for each specimen and sequence.** Division of genus *Carabus* follows Deuve [28].

Species	*Carabus* division	Locality	Voucher	*cox1*-A	*cox1*-B	*cob*	*rrnl*	*nd5*	*LSU*-A	*LSU*-B	*ITS2*	*HUWE*
Laemostenus terricola	OUTGROUP	Alicante, Spain	TEUL_43_09						n/a		n/a	n/a
Leistus spinibarbis	OUTGROUP	Albacete, Spain	CMOJ_39_09					n/a				n/a
Calosoma aeropunctatum	OUTGROUP	Susuz, Turkey	TURQ_520_01									
Calosoma sycophanta	OUTGROUP	Albacete, Spain	CALO_254_05									
Ceroglossus chilensis	OUTGROUP	Chiloe, Chile	CHIL_1-GA									n/a
Cychrus semigranosus	OUTGROUP	Pirin Mts., Bulgaria	CYCH_10_03						n/a			n/a
C. (Archicarabus) nemoralis	ARCHICARABOMORPHI	Navarra, Spain	RONC_1549									n/a
C. (Platycarabus) irregularis	ARCIFERA	Resita, Rumania	ROMA_940_06									
C. (Rhabdotocarabus) melancholicus	ARCIFERA	Cádiz, Spain	EALM_422_07									
C. (Rhabdotocarabus) melancholicus	ARCIFERA	Toledo, Spain	ROBU_37							n/a		
C. (Rhabdotocarabus) melancholicus	ARCIFERA	Tidiquin, Morocco	TIDI_391_04									
C. (Carabus) deyrollei	DIGITULATI	Lugo, Spain	GALI_1553								n/a	n/a
C. (Eurycarabus) famini	DIGITULATI	Ketama, Morocco	MORR_548_07									
C. (Nesaeocarabus) abbreviatus	DIGITULATI	Tenerife, Spain	TENE_150_07									
C. (Nesaeocarabus) abbreviatus	DIGITULATI	Tenerife, Spain	BABA_44					n/a	n/a			n/a
C. (Eurycarabus) famini	DIGITULATI	El Alia, Tunisia	EURY_1625									
C. (Morphocarabus) monilis	LIPASTRIMORPHI	Drome, France	SAOU_1538									
C. (Mesocarabus) dufourii	METACARABI	Cordoba, Spain	ZUHE_111									
C. (Mesocarabus) lusitanicus	METACARABI	Ciudad Real, Spain	VESC_5									
C. (Mesocarabus) macrocephalus	METACARABI	La Coruña, Spain	FORO_157									
C. (Mesocarabus) riffensis	METACARABI	Ketama, Morocco	KETA_569									
C. (Chrysocarabus) auronitens	NEOCARABI	Resita, Rumania	ROMA_939_06									
C. (Chrysocarabus) rutilans	NEOCARABI	Barcelona, Spain	SENY_1548									
C. (Lamprostus) coriaceus	NEOCARABI	Oysu, Turkey	TURQ_913_06									n/a
C. (Macrothorax) morbillosus	NEOCARABI	Murcia, Spain	MAZA_59_08									
C. (Macrothorax) morbillosus	NEOCARABI	Sejenane, Tunisia	TUNA_1622									
C. (Macrothorax) morbillosus	NEOCARABI	Bazia, Tunisia	TUNB_1623									
C. (Macrothorax) morbillosus	NEOCARABI	El Alia, Tunisia	TUNC_1624									
C. (Macrothorax) rugosus	NEOCARABI	Ksar-el-Kebir, Morocco	KSAR_658_07									
C. (Macrothorax) rugosus	NEOCARABI	Cádiz, Spain	SMAR_439_07									
C. (Megodontus) violaceus	NEOCARABI	Popovi Livadi, Bulgaria	POPO_830_00									n/a
C. (Limnocarabaus) clathratus	SPINULATI	Susuz, Turkey	TURQ_519_01									
C. (Autocarabus) cancellatus	TACHYPOGENICI	France	TACH_cc816									
C. (Autocarabus) cancellatus	TACHYPOGENICI	Resita, Rumania	ROMA_941_06									

Table 2: Rate of molecular evolution and TMRCA of *Carabus* for each individual fragment and combined datasets under optimal analytical conditions. Data correspond to the *ingroup* dataset.

Gene	Partition	Clock	Ambiguities treatment	Rate	TMRCA *Carabus*
nd5	NP*	SC	-	0.0175 (0.0113-0.0246)	19.68 (15-24.69)
cox1-A	2P	SC	-	0.0126 (0.009-0.0165)	18.3 (13.9-22.96)
cox1-B	2P	SC	-	0.0161 (0.0109-0.0218)	20.13 (15.2-25.53)
cob	2P	SC	-	0.0271 (0.0162-0.0396)	24.3 (18.24-30.63)
rrnL	*NP*	*SC*	*Allgaps*	0.0017 (0.0011-0.0025)	28.9 (18.03-40.98)
LSU-A	*NP*	ULN	*Allgaps*	0.0013 (0.0008-0.002)	14.75 (9.46-22.97)
LSU-B	*NP*	ULN	*Allgaps*	0.0069 (0.0037-0.0107)	21.01 (11.1-34.62)
ITS2	*NP*	ULN	*Allgaps*	0.0061 (0.0038-0.0085)	30.91 (16.3-48.7)
HUWE1	*NP*	SC	*Complete*	0.0023 (0.0017-0.0029)	28.39 (19.83-37.56)
MIT	G-2P	ULN	*Allgaps*	0.015 (0.0117-0.0185)	20.15 (15.89-24.69)
NUC	NP	ULN	*Allgaps*	0.0031 (0.0022-0.0042)	27.17 (15.71-40.97)
MIT-NUC	G-2P	ULN	*Allgaps*	0.0088 (0.0069-0.0107)	23.28 (16.5-30.43)

Rate: substitutions per site per million years per lineage (mean and 95%HPD interval)

TMRCA *Carabus*: Estimate for the origin of the genus *Carabus* in million years (mean and 95%HPD interval)

* Selection of NP dataset due to the anomalous branch length resulting from codon partitioned *nd5* analyses.

Additional files

Additional file. One PDF file including:

A) Supporting Figures and Legends

Figure S1 to S12. Phylogenetic trees obtained with MrBayes, BEAST (*outgroup dataset*) and BEAST (*ingroup dataset*) under the selected parameters. Bars represent 95% confident intervals for the node ages in Myr. Numbers inside nodes represent posterior probabilities. S1. *cox1-A*; S2. *cox1-B*; S3. *cob*; S4. nd5; S5. *rrnL*; S6. *LSU-A*; S7. *LSU-B*; S8. *ITS2*; S9. *HUWE1*. S10. *MIT*; S11. *NUC*; S12. *MIT-NUC.*

B) Supporting Tables

Table S1. Primers used in the study.
Table S2. Calibration hypotheses employed to time calibrate the phylogeny of the *nd5* gene
Table S3. Calculations for the objective selection of alignments.
Tabla S4. Marginal likelihood values for BEAST analyses of individual gene fragments.
Table S5. Mean rates of molecular evolution and 95% HPD interval of calibration analyses.
Table S6. Mean ages and 95% HPD interval of calibration analyses.

C) Supplementary Text

D) Supplementary References

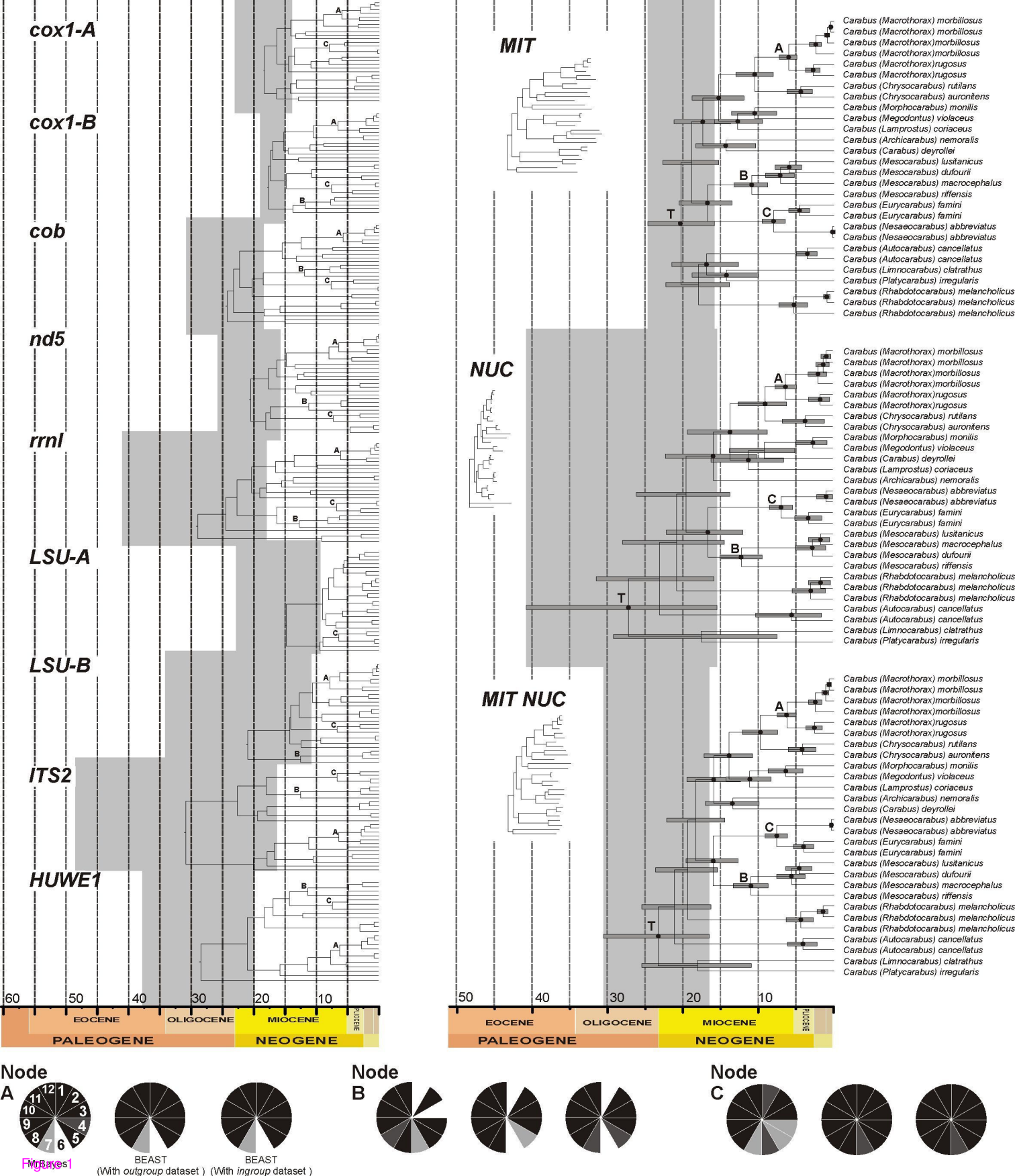
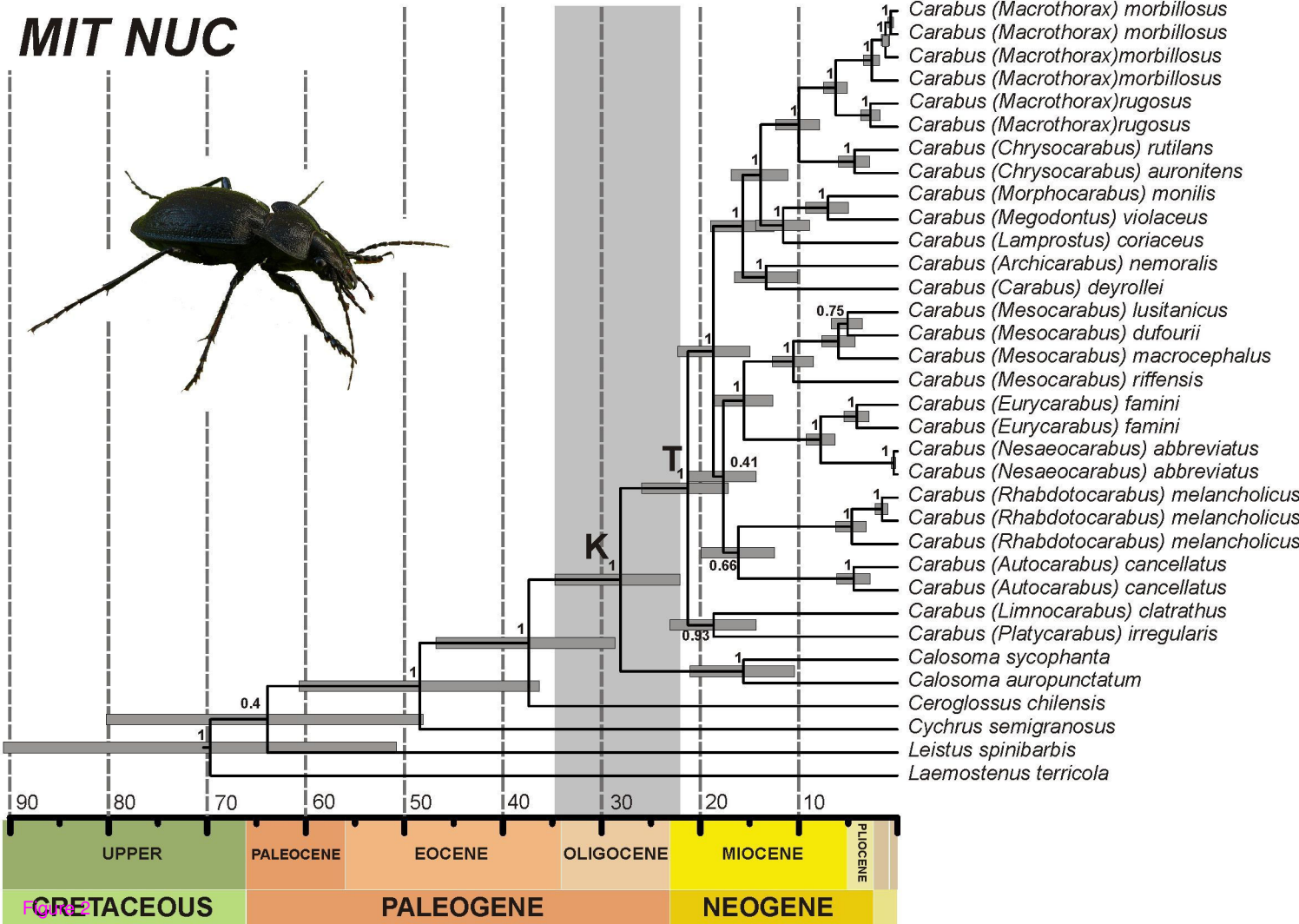

cox1-A
cox1-B
cob
nd5
rrnl
LSU-A
LSU-B
ITS2
HUWE1
MIT
NUC
MIT NUC
Carabus (Macrothorax) morbillosus
Carabus (Macrothorax) morbillosus
Carabus (Macrothorax)morbillosus
Carabus (Macrothorax)morbillosus
Carabus (Macrothorax)rugosus
Carabus (Macrothorax)rugosus
Carabus (Chrysocarabus) rutilans
Carabus (Chrysocarabus) auronitens
Carabus (Morphocarabus) monilis
Carabus (Megodontus) violaceus
Carabus (Lamprostus) coriaceus
Carabus (Archicarabus) nemoralis
Carabus (Carabus) deyrollei
Carabus (Mesocarabus) lusitanicus
Carabus (Mesocarabus) dufourii
Carabus (Mesocarabus) macrocephalus
Carabus (Mesocarabus) riffensis
Carabus (Eurycarabus) famini
Carabus (Eurycarabus) famini
Carabus (Nesaeocarabus) abbreviatus
Carabus (Nesaeocarabus) abbreviatus
Carabus (Autocarabus) cancellatus
Carabus (Autocarabus) cancellatus
Carabus (Limnocarabus) clatrathus
Carabus (Platycarabus) irregularis
Carabus (Rhabdotocarabus) melancholicus
Carabus (Rhabdotocarabus) melancholicus
Carabus (Rhabdotocarabus) melancholicus
60
50
40
30
20
10
EOCENE
OLIGOCENE
MIOCENE
PLIOCENE
PALEOGENE
NEOGENE
Node A
BEAST
(With outgroup dataset)
BEAST
(With ingroup dataset)
Node B
Node C

Figure 1


MIT NUC
Carabus (Macrothorax) morbillosus
Carabus (Macrothorax) morbillosus
Carabus (Macrothorax)morbillosus
Carabus (Macrothorax)morbillosus
Carabus (Macrothorax)rugosus
Carabus (Macrothorax)rugosus
Carabus (Chrysocarabus) rutilans
Carabus (Chrysocarabus) auronitens
Carabus (Morphocarabus) monilis
Carabus (Megodontus) violaceus
Carabus (Lamprostus) coriaceus
Carabus (Archicarabus) nemoralis
Carabus (Carabus) deyrollei
Carabus (Mesocarabus) lusitanicus
Carabus (Mesocarabus) dufourii
Carabus (Mesocarabus) macrocephalus
Carabus (Mesocarabus) riffensis
Carabus (Eurycarabus) famini
Carabus (Eurycarabus) famini
Carabus (Nesaeocarabus) abbreviatus
Carabus (Nesaeocarabus) abbreviatus
Carabus (Rhabdotocarabus) melancholicus
Carabus (Rhabdotocarabus) melancholicus
Carabus (Rhabdotocarabus) melancholicus
Carabus (Autocarabus) cancellatus
Carabus (Autocarabus) cancellatus
Carabus (Limnocarabus) clatrathus
Carabus (Platycarabus) irregularis
Calosoma sycophanta
Calosoma auropunctatum
Ceroglossus chilensis
Cychrus semigranosus
Leistus spinibarbis
Laemostenus terricola
T
K
0.75
0.41
0.66
0.93
0.4
90
80
70
60
50
40
30
20
10
UPPER
PALEOCENE
EOCENE
OLIGOCENE
MIOCENE
PLIOCENE
CRETACEOUS
PALEOGENE
NEOGENE

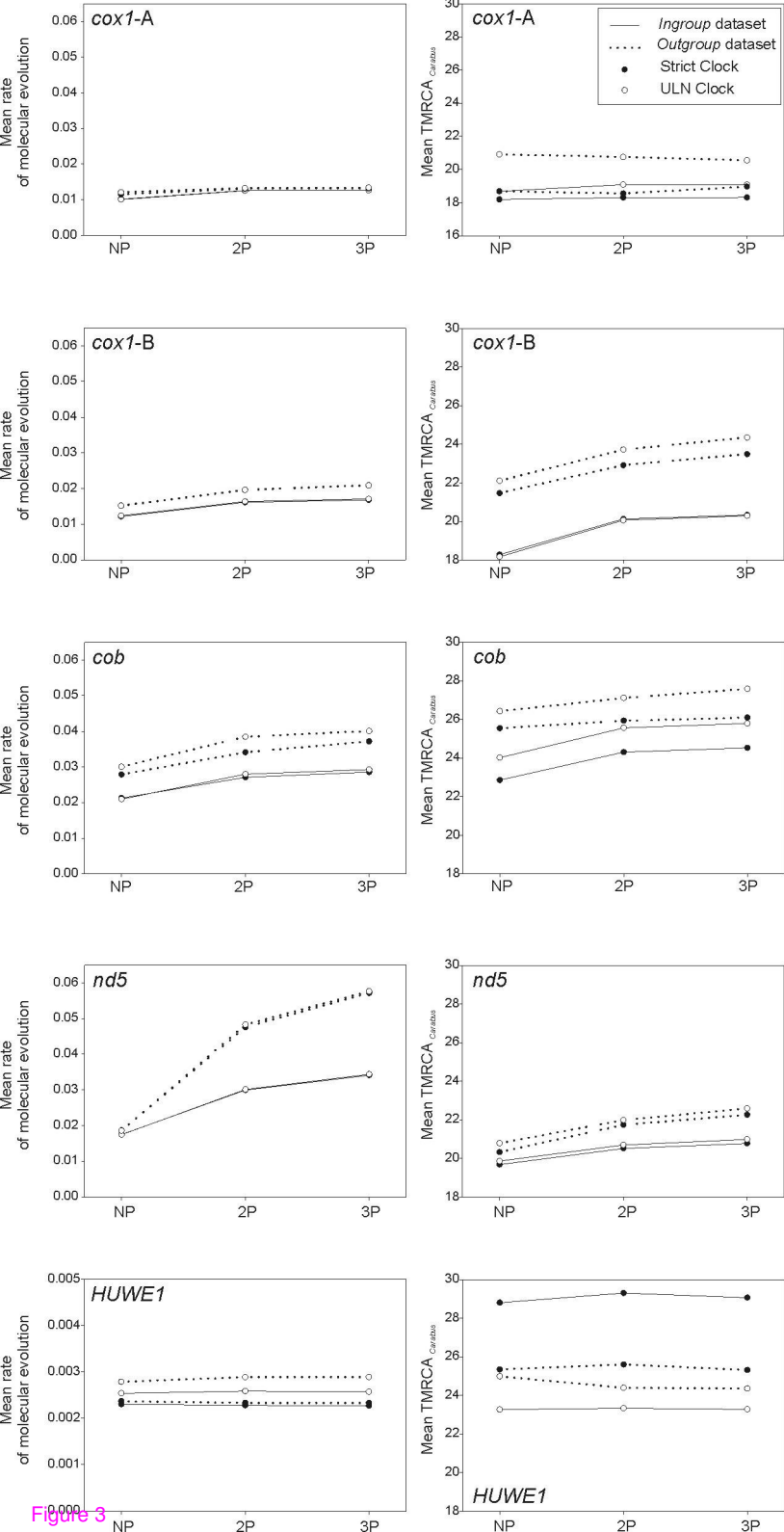

cox1-A
cox1-B
cob
nd5
HUWE1
Mean rate of molecular evolution
Mean TMRCA Carabus
Ingroup dataset
Outgroup dataset
Strict Clock
ULN Clock
NP
2P
3P

Figure 3

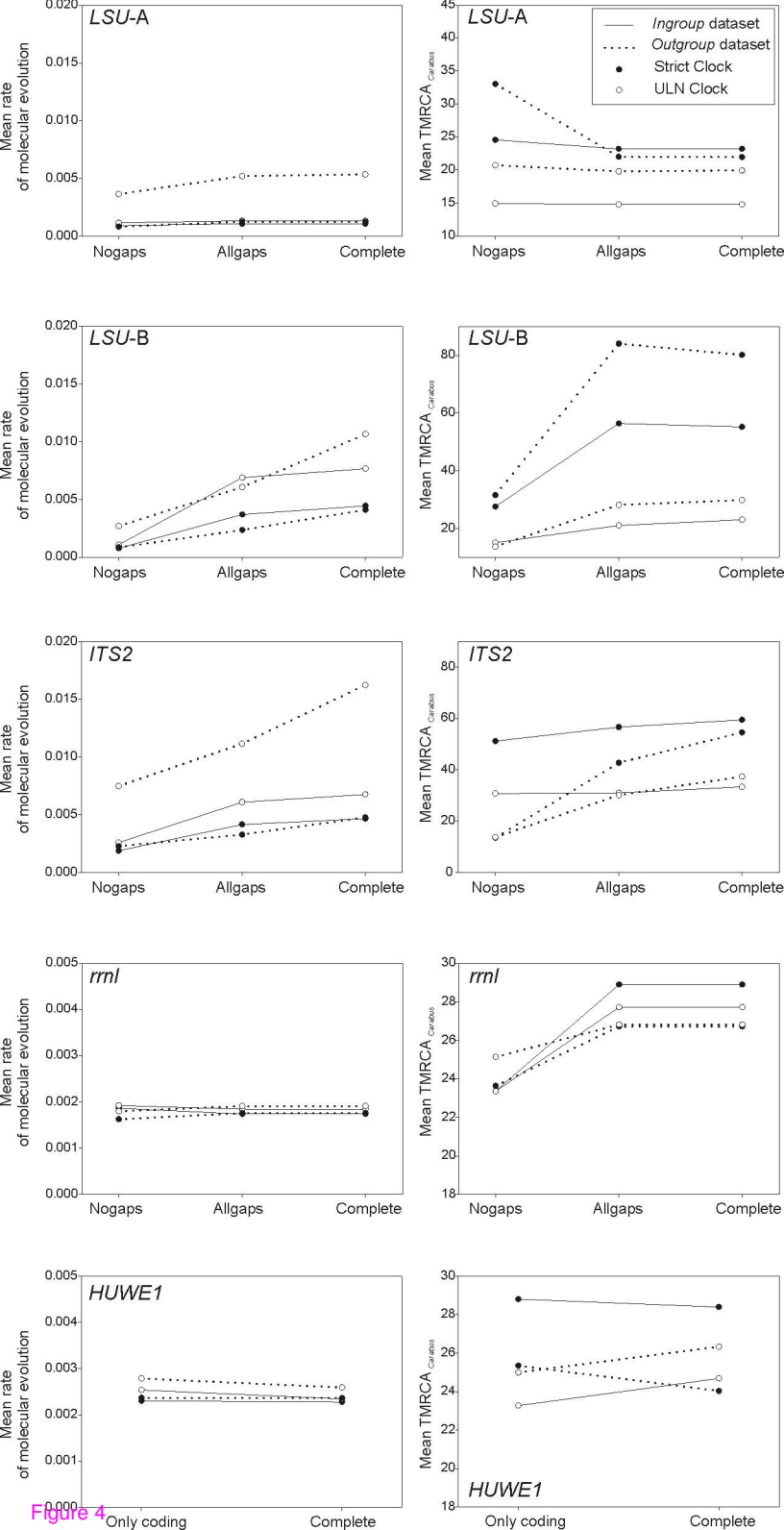

LSU-A
LSU-B
ITS2
rrnl
HUWE1
Mean rate of molecular evolution
Mean TMRCA Carabus
Ingroup dataset
Outgroup dataset
Strict Clock
ULN Clock
Nogaps
Allgaps
Complete
Only coding

Figure 4

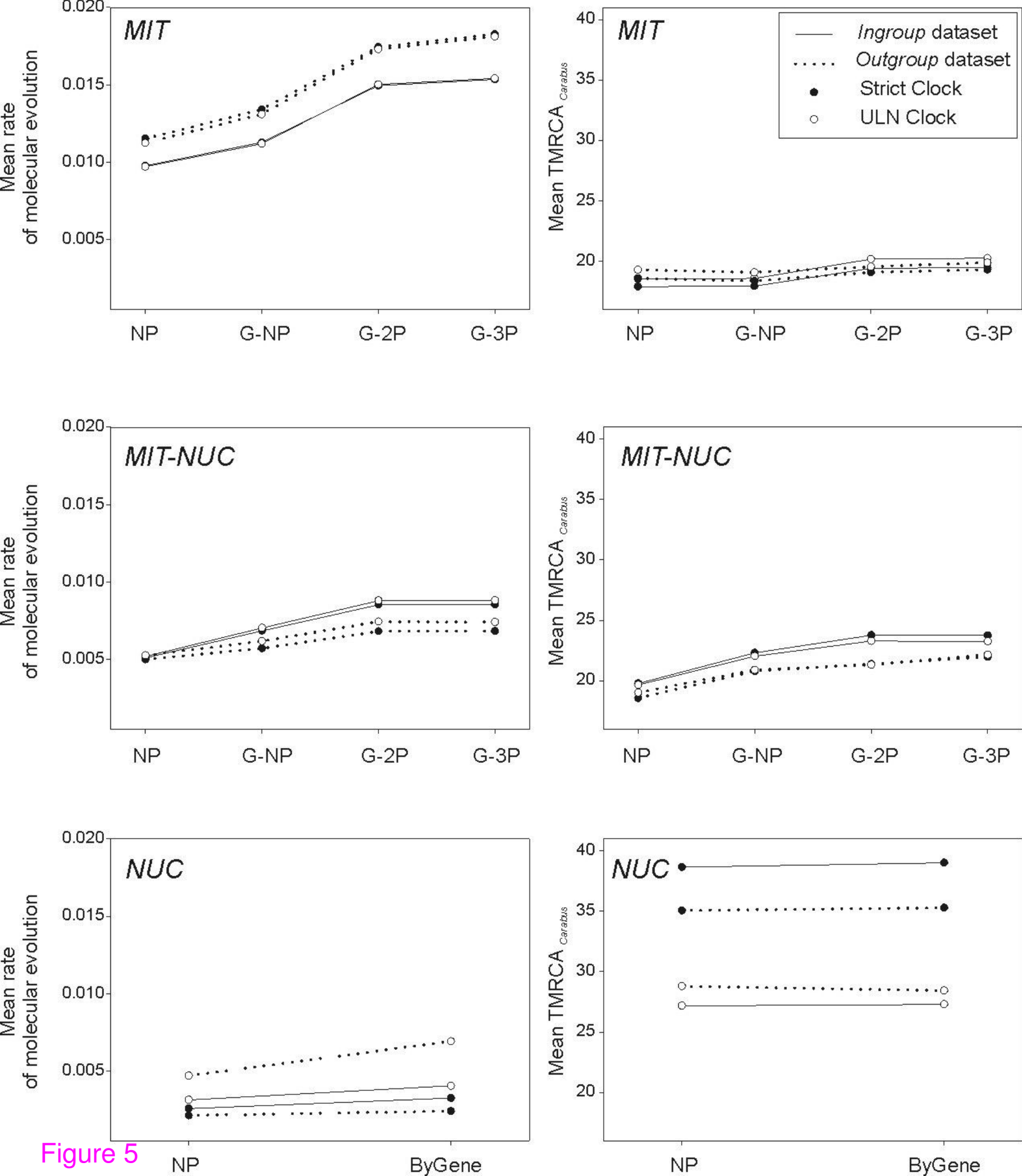

MIT
Mean rate of molecular evolution
0.020
0.015
0.010
0.005
NP
G-NP
G-2P
G-3P
MIT
Mean TMRCA Carabus
40
35
30
25
20
Ingroup dataset
Outgroup dataset
Strict Clock
ULN Clock
MIT-NUC
NUC
ByGene

Figure 5

Additional files provided with this submission:

Additional file 1: Additional file.pdf, 1866K
http://www.biomedcentral.com/imedia/1937603297595675/supp1.pdf