Intron Presence–Absence Polymorphisms in *Daphnia*

Angela R. Omilian,[1] *Douglas G. Scofield,*[2] *and Michael Lynch,*
Department of Biology, Indiana University

Here, we report 2 novel intron gains segregating in populations of *Daphnia pulex* endemic to Oregon. These novel introns do not have an obvious source and are not present in any *D. pulex* populations outside Oregon, other species of *Daphnia* that we examined, or any other organism for which sequence data are available. Furthermore, the novel introns are both found in the same gene, a Rab GTPase (*rab4*), and they appear to differ in their insertion site by one base pair, providing some support to the proto-splice site hypothesis. The rarity of intron-gain polymorphisms is questioned as we discovered 2 events in an initial survey of only 6 nuclear loci in 36 *Daphnia* individuals. Neutrality tests failed to ascertain a clear selective effect for either intron insertion, and a significant difference in recombination rate was not observed in alleles that contain the novel intron insertion versus alleles lacking it. We conclude that one novel intron insertion segregating at high frequencies in *Daphnia* populations in Oregon is unlikely to be adaptive and may result from the reduced efficacy of selection in isolated populations of small effective size.

Introduction

Since the discovery of spliceosomal introns (Berget et al.1977; Chow et al. 1977; Evans et al. 1977; Goldberg et al. 1977), their origins and evolutionary roles have been debated (reviewed in Roy and Gilbert 2006; Lynch 2007). The introns early or exon theory of genes proposes that introns originated in early prokaryotes prior to the emergence of eukaryotes. Under the predominant version of this hypothesis, introns played a formative role in early protein evolution, with subsequent intron evolution being dominated by intron loss, and complete loss having occurred in prokaryotes (Doolittle 1978; Gilbert 1978, 1987; Roy and Gilbert 2006). In contrast, the introns-late hypothesis maintains that introns are largely adventitious embellishments of eukaryotic genes, with little initial role in adaptation (Orgel and Crick 1980; Cavalier-Smith 1985; Palmer and Logsdon 1991). There are now many variants on these extreme themes, and recent focus has been given to the full spectrum from ancient stable to recently gained introns (de Souza 2003; Fedorova and Fedorov 2003; Koonin 2006; Martin and Koonin 2006).

A puzzling aspect of intron evolution is the tremendous variation in intron numbers among eukaryotic species. Considerable disagreement exists over the source of such variation, but some phylogenetic groups appear to be characterized by extensive intron gains, whereas others have experienced substantial intron losses (e.g., Seo et al. 2001; Rogozin et al. 2003; Roy et al. 2003; Cho et al. 2004; Edvardsen et al. 2004; Qiu et al. 2004; Nguyen et al. 2005, 2007; Raible et al. 2005; Roy and Gilbert 2005; Roy and Hartl 2006; Stajich and Dietrich 2006). However, the mechanisms and evolutionary forces responsible for such gains and losses are largely unknown. The fixation of alternative intron presence/absence states among lineages must have been accompanied by transient phases of within-species intron presence/absence polymorphisms, but observations of such conditions are extremely rare. To our knowledge, there is only one published report—an intron loss via genomic deletion in the *jingwei* gene from *Drosophila teissieri* (Llopart et al. 2002).

The further study of intron presence/absence polymorphisms from natural populations may elucidate some longstanding questions regarding the origins and evolution of introns. First, novel introns that are segregating in natural populations are likely to reflect recent intron gains, the identification of which is an important step in determining the sources of introns and their potential insertion-site preferences. Second, modern intron distributions may be nonadaptive by-products of genetic drift in small populations, or the result of natural selection associated with the roles that introns can play in processes such as nonsense-mediated decay, alternative splicing, exon shuffling, or recombination. With intron presence/absence polymorphisms, population-genetic analyses can be applied to assess the importance of selection or genetic drift on the fate of novel intron insertions.

Here we report 2 intron-gain polymorphisms in the same gene of a cosmopolitan species of freshwater microcrustacean, *Daphnia pulex*. Both novel introns appear to be restricted to *Daphnia* collected in Oregon, USA, and their insertion sites differ from one another by one nucleotide. Common hypotheses for the origins and stability of introns are discussed in the context of our results.

Materials and Methods

Locus Information

A previous survey of 6 protein-coding loci in 27 populations of *Daphnia* revealed one locus that appeared to be polymorphic for an intron insertion in *D. pulex* populations located in Oregon (Omilian 2006). Blast searches against GenBank revealed this locus to be a member of the Rab family of small GTPases, which serve as central regulators of membrane traffic pathways (Pereira-Leal and Seabra 2001; Stenmark and Olkkonen 2001; Zerial and McBride 2001). We chose the name *rab4* as Blast results indicated closest similarity to this subfamily of Rab GTPases. Alleles at the *D. pulex rab4* locus shared 64.0% amino acid identity with the *Drosophila melanogaster* protein Rab4 isoform A (GenBank gi: 24654467) and 58.6% amino acid identity with the human Rab4b (GenBank gi: 82659107).

Gene duplication has increased the number of Rab GTPases in a number of organisms (Stenmark and

[1] Present address: Department of Biological Sciences, University at Buffalo, Buffalo, NY 14260.
[2] Present address: Department of Ecology and Evolutionary Biology, University of California, Los Angeles.

Key words: *Daphnia*, intron insertion, intron gain, intron polymorphism.

E-mail: alr2@buffalo.edu.

Mol. Biol. Evol. 25(10):2129–2139. 2008
doi:10.1093/molbev/msn164
Advance Access publication July 29, 2008

Table 1
Characteristic Features of Introns 2a and 2b in the *rab4* Locus in *Daphnia pulex* Individuals Collected in Oregon

Feature	Intron 2a	Intron 2b
Populations	AZ, CC, LOG, OP	GI, LOG
Length (bp)	75	62
Phase	2	0
In-phase stop codons	3	2
5′ and 3′ splice-site sequence	GT–AG	GT–AG
GC content (%)	9	21
Segregating in populations?	Fixed in AZ (4/4), CC (6/6), OP (8/8) and 83% (10/12) in LOG	50% (6/12) in GI and 8% (1/12) in LOG

NOTE.—The LOG population consisted of 10 alleles with intron 2a, 1 allele with intron 2b, and 1 allele of *Daphnia pulicaria* origin (in LOG52) that did not have intron 2.

Olkkonen 2001), and closely related Rab GTPase subfamilies differ in the positions and/or number of introns present (Pereira-Leal and Seabra 2001). However, 2 lines of evidence indicate that *D. pulex* has a single copy of *rab4*. First, all data obtained from exhaustive cloning (see below) never revealed evidence for more than 2 alleles per individual. Second, Blast searches of individual exons of *rab4* against the most recent assembly of the *D. pulex* genome (*Daphnia* Genomics Consortium, http://wfleabase.org) found only a single copy of *rab4*. This result is particularly noteworthy given that the individual sequenced for the *D. pulex* genome project was collected from the same pond (LOG collection site, table 1) as individuals from this study that contain the intron insertion and the sequenced genome includes the novel intron insertion reported here.

Species and Populations Examined

For the present study, we used previously reported *D. pulex* sequences (Omilian 2006) and collected additional samples from 3 sites in Oregon and 14 sites in the eastern United States and Canada for a total of 26 *D. pulex* populations (table 1, fig. 1, supplementary table 1, Supplementary Material online). Altogether, 5 populations with individuals that contained a novel intron (designated AZ, CC, GI, LOG, and OP) were examined. Nine *Daphnia pulicaria*, 2 *Daphnia melanica*, and 8 *Daphnia obtusa* populations were also screened for the novel introns. *Daphnia pulicaria* is commonly regarded as a sister species to *D. pulex* (Colbourne and Hebert 1996), although a study of 6 nuclear protein-coding loci revealed essentially no divergence between these species at silent sites (silent-site divergence [standard error] is 0.0026 [0.0037]; Omilian 2006). It has been proposed that certain populations of *D. pulex* endemic to Oregon are a separate species called *Daphnia arenata* (Hebert 1995), but because *D. pulex* is paraphyletic with respect to these populations (Colbourne et al. 1998; Lynch unpublished data) we will simply refer to members of this clade as "Oregon *D. pulex*."

Because taxonomic differentiation based on morphology is notoriously difficult within *Daphnia*, we verified the taxonomic identification of *D. pulex* and *D. obtusa* with sequence data from the 12S rDNA gene (Colbourne and Hebert 1996). Allozyme analysis for the lactate dehydrogenase locus was used to differentiate between *D. pulex* and *D. pulicaria*, following the conventional notion that *D. pulicaria* is homozygous for the F allele and *D. pulex* is homozygous for the S allele (Hebert et al. 1989, 1993). Allozyme analysis revealed that one individual (LOG52) was a *D. pulex–D. pulicaria* hybrid.

To obtain sequence for *rab4* homologues in taxa other than *Daphnia*, we used TBlastN (Altschul et al. 1997) against RefSeq entries containing intron–exon structure derived from sequenced genomes within GenBank. We aligned these sequences using ClustalW (Thompson et al. 1994) and inferred intron positions from GenBank annotations. Because our amplified sequences represented a fragment of the locus (see below), we obtained structural information for the complete gene from the most recent version of the *D. pulex* genome sequence (v1.1, Gene ID 304304, sequence position scaffold_28:735442..737181, +strand).

Polymerase Chain Reaction Amplification, Sequencing, and Cloning

Genomic DNA was extracted from field collections using the 2× cetyltrimethylammonium bromide extraction protocol (Doyle JJ and Doyle JL 1987). *Rab4* primers were designed from conserved regions based on complementary DNA (cDNA) sequences from *D. pulex* and *Daphnia magna* (cDNA libraries provided by Hajime Watanabe and John Colbourne) and are F6for 5′-CGTTTCGAATTGGCTTACTGA-3′ and F12rev 5′-CATGGTTATCTGTCTACGTCTTGAA-3′. Each polymerase chain reaction (PCR) consisted of 37.4 μl molecular grade water, 5 μl 10× PCR buffer, 12 nmoles deoxynucleoside triphosphates, 12 pmoles of each primer, 0.5 μl *Taq* polymerase (Clontech, Mountain View, CA), and 25–50 ng DNA template. PCR was conducted on an MJ Thermocycler with the following conditions: 40 cycles of 1 min at 94 °C, 1 min at 53 °C, 1.5 min at 72 °C; followed by 1 cycle of 10 min at 72 °C. PCR products were purified with solid phase reversible immobilization (Elkin et al. 2001), cycle sequenced, and analyzed on an ABI3730 DNA sequencer (Applied Biosystems, Foster City, CA). Sequence data have been deposited in GenBank under accession numbers EU918429–EU918560.

Our primers amplified part of the *rab4* locus in 132 *Daphnia* individuals for a total of 264 alleles. The amplified fragment consisted of 3 complete exons, 1 partial exon, 2 or 3 introns (depending on the individual sampled), and a small portion of the 3′ untranslated region (UTR). The aligned length of the fragment was 580 bp (includes intron insertion). Because DNA was extracted from *Daphnia* collected from natural populations (rather than inbred stock populations), several individuals were heterozygous for the *rab4* locus. An individual locus was considered heterozygous if 2 overlapping peaks were observed at any given site on the DNA sequence electropherogram for both forward and reverse sequencing primers. Putative heterozygous sites were detected with CodonCode Aligner v1.4.3 set to detect mutations at highest sensitivity and then verified by eye. PCR fragments with multiple heterozygous sites were cloned with the Invitrogen TOPO TA kit to determine gametic phase. The QIAprep Spin Miniprep Kit (QIAGEN, Valencia, CA) was used for plasmid purification, and a T7 primer was

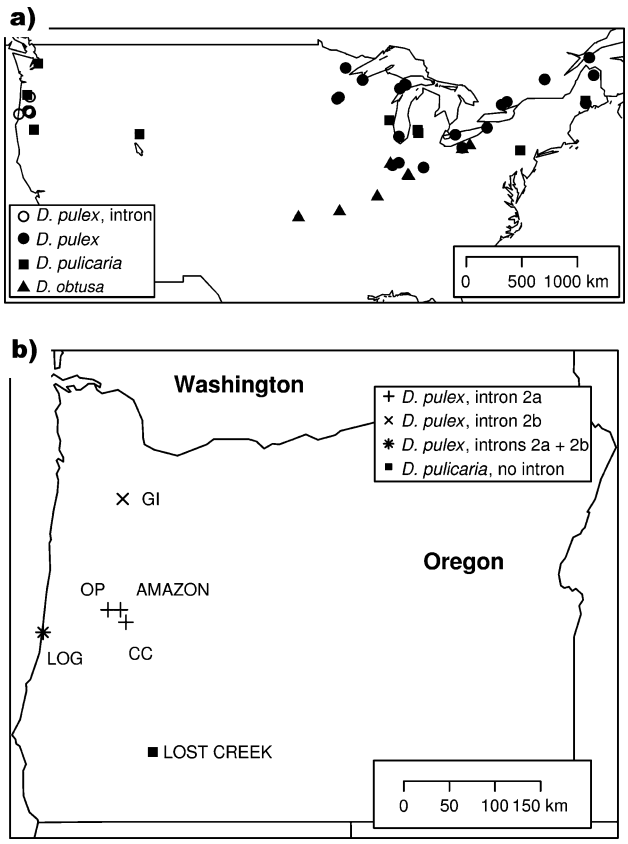

FIG. 1.—(*a*) Collection locations of *Daphnia* screened for novel introns in the *rab4* locus. (*b*) Collection locations of *Daphnia pulex* and *Daphnia pulicaria* populations located in Oregon; *D. pulex* populations from Oregon have individuals that contain a novel intron insertion.

used to sequence the cloned inserts. To guard against PCR and cloning errors (Cronn et al. 2002), 4–16 cloned fragments were sequenced per individual. Additionally, to ensure that polymorphisms were not the result of cloning-induced errors, sequences from cloned PCR products were compared with the directly sequenced PCR products.

To confirm that the observed insertion polymorphism was an intron spliced from the primary transcript, we extracted RNA from various *Daphnia* individuals using the RNeasy Mini Kit (QIAGEN). Reverse transcriptase–polymerase chain reaction (RT-PCR) was used to amplify RNA using the aforementioned F6for and F12rev primers and the QIAGEN OneStep RT-PCR Kit. RT-PCR products were purified and sequenced in both directions and then aligned with the DNA sequences in MEGA version 3.1 (Kumar et al. 2004). The full alignment is available as Supplementary Material online.

Evolutionary Relationships and Population-Genetic Analyses

MrBayes v3.1.2 was applied (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) to elucidate the genealogical relationships of the *Daphnia rab4* alleles with Bayesian inference. To avoid spurious clustering on the basis of the polymorphic intron, it was excluded from the phylogenetic analyses. The sequence data were partitioned into 1 noncoding (introns and UTR) and 3 codon positions (first, second, and third codon positions). Markov chain Monte Carlo analyses were run for 12 million generations, sampling from the chain every 100 generations. After determining chain convergence (average standard deviation of split frequencies <0.01), we discarded the initial 25% of trees as the "burn-in period." A 50% majority-rule consensus tree with posterior probability (PP) values for each node was constructed from the remaining Bayesian trees.

Population-genetic parameters and tests of neutrality were calculated with DnaSP (Rozas et al. 2003). Insertion–deletion mutations and the intron polymorphism were excluded from these analyses. Two measures of nucleotide diversity were estimated: π, the average of pairwise differences among DNA sequences (Tajima 1983), and θ, based on the total number of segregating mutations in the sample (Watterson 1975; Tajima 1996). Both π and θ were estimated for all nucleotide sites (π_T, θ_T), nonsynonymous sites (π_n, θ_n), and synonymous sites (π_s, θ_s). The following tests of neutrality were conducted: Tajima's *D* (Tajima 1989), Fu and Li's *D* and *F* (with outgroup, Fu and Li 1993), and Fay and Wu's *H* (Fay and Wu 2000).

We used a full-likelihood coalescence-based approach in LAMARC 2.0 (Kuhner 2006) to estimate the recombination rate in *rab4* alleles that contained or lacked the novel intron. LAMARC coestimates θ with the overall recombination rate, $r_{LAM} = c/\mu$, where c is the recombination rate per site per generation and μ is the neutral mutation rate per site per generation. Because LAMARC could not implement the best-fit model of nucleotide substitution determined by Modeltest v.3.7 (Posada and Crandall 1998), we used the Felsenstein 84 (F84) model with empirical base frequencies (Kishino and Hasegawa 1989; Felsenstein 1993). The transition/transversion ratio was set to 1.5, and 2 categories of relative mutation rate were assigned, accounting for mutation rate differences between nonsynonymous sites and all other sites. Our sampling strategy included 20 initial chains of 1,000 and 2 final chains of 50,000 genealogies with 1,000 genealogies discarded per chain. Adaptive heating was used to improve the search of parameter space; relative temperatures were initially set to 1, 1.1, and 2.7. The entire analysis was replicated 5 times and then the results were combined using the algorithm of Geyer (1991).

Results

Novel Introns Are Geographically Restricted

Our examination of the intron–exon structure of *rab4* in 26 populations of *D. pulex* sampled from the United States and Eastern Canada revealed 2 novel intron insertions in populations located in Oregon (figs. 1 and 2, supplementary table 1, Supplementary Material online). The insertion sites for these introns appear to differ by only one base pair within the *rab4* locus and occur outside of conserved Rab functional and structural domains, downstream of strand $\beta 6$ and immediately 5′ to helix $\alpha 5$ (fig. 3; for summaries of Rab domains, see Pereira-Leal and Seabra 2000; Stenmark and Olkkonen 2001). We designate these novel introns as 2a and 2b as they represent the second intron site in the fragment of the *rab4* locus used for

this study. Introns 2a and 2b differ in several ways (table 1) and do not share any apparent similarity aside from brief oligonucleotide stretches (fig. 2). Most Oregon populations were segregating or fixed for either intron 2a or intron 2b. The LOG population was the only exception and here both novel introns were segregating (table 1).

To verify that the polymorphic insertions are intronic in nature, RT-PCR amplified RNA was sequenced and aligned with corresponding genomic sequences (fig. 2). This confirmed that the insertions are novel introns with canonical splice-site sequences. Furthermore, the inclusion of either novel intron into the coding DNA would have created multiple in-phase stop codons resulting in truncated proteins.

In the *D. pulicaria* populations sampled, there is an intron-boundary sequence polymorphism for the first intron in our sequenced fragment, with 39% (7/18) of their intron sequences having the traditional GT–AG boundary and 61% (11/18) having a GC–AG boundary (supplementary table 2, Supplementary Material online). One *D. pulicaria* individual from the DUTCH collection site is heterozygous for this intron-boundary sequence polymorphism.

Intron–Exon Structures of *rab4* Homologs Are Highly Conserved in Other Metazoans

In a search of GenBank for homologs of *rab4* in other metazoans, we found associated intron–exon structures for the dipterans *Aedes aegypti* (GenBank gi: 108873366), *Anopheles gambiae* (gi: 118783599), *D. melanogaster* (gi: 24654467), and *Drosophila pseudoobscura* (gi: 125808107); the echinoderm *Strongylocentrotus purpuratus* (gi: 115653131); and the vertebrates *Bos taurus* (gi: 119910544), *Canis familiaris* (gi: 73946370), *Danio rerio* (gi: 57524538), *Gallus gallus* (gi: 50741401), *Rattus norvegicus* (gi: 8394136), human (gi: 82659107), and mouse (gi: 21313012) (fig. 3). Most species show a high degree of conservation of both intron positions and phases, with the exception of the dipterans, which have experienced some intron loss. However, no organisms for which a strong sequence homolog for *rab4* was identified have a homologous intron corresponding to introns 2a and 2b in *D. pulex* (fig. 3). Furthermore, *D. pulex* sampled from other locations in the United States and Canada do not have an intron in this region, nor do any of the sampled *D. pulicaria*, *D. melanica*, and *D. obtusa*.

No Sequence Matches for Novel Introns within the *D. pulex* Genome

Blast searches of the novel *Daphnia* intron sequences against the *Daphnia* genome assembly (v1.1), *Daphnia* genome trace files, and GenBank did not reveal a likely source. Blast searches of intron 2a against nucleotide records in GenBank yielded a poor best hit against a chloroplast photosystem II gene from *Euglena deses* (GenBank gi: 9049727; $E = 0.010$). The Blast result for intron 2b also yielded a poor best hit against a zebrafish DNA sequence (GenBank gi: 123844202; $E = 0.002$). Numerous mismatches and gaps for both Blast results suggest that these are not potential sources for either intron.

Evolutionary Relationships and Population-Genetic Analyses

The 50% majority-rule consensus Bayesian topology was reconstructed from 180,000 post burn-in trees that were obtained from 2 simultaneous and independent runs (fig. 4). The consensus tree revealed that *rab4* alleles containing a novel intron are found in 2 distinct clades on the tree. Alleles containing novel intron 2a are found in a strongly supported (PP = 1) monophyletic group that is a sister group to both *D. pulex* and *D. pulicaria* (fig. 4), thereby supporting the notion that alleles containing intron 2a may in fact belong to a distinct species—*D. arenata*. Most (83%) individuals containing intron 2b are heterozygous at the *rab4* locus, having one allele that contains the novel intron and one allele lacking it. Alleles containing novel intron 2b are reconstructed as a monophyletic group (PP = 0.99) that is a sister group to *D. pulicaria* (PP = 0.79), whereas the intron-lacking alleles group with *D. pulex* (fig. 4). *Daphnia pulex* is paraphyletic with respect to *D. pulicaria*, which is itself a monophyletic group (fig. 4).

Alleles containing intron 2a are found in 4 populations (AZ, CC, OP, and LOG) and appear to be fixed in the AZ, CC, and OP populations. Alleles containing intron 2b are segregating in the LOG (8%) and GI (50%) populations. Because new mutations (in this case, the intron insertions) occur at low frequency within populations upon their first appearance, their probability of going to fixation is also low. Our observation of 2 new introns segregating at intermediate to high frequencies in most populations suggests that directional selection has acted to increase the frequency of the intron-bearing alleles, either directly or indirectly via hitchhiking. However, genetic drift can leave a similar footprint depending upon population size, and it is important to differentiate between these disparate explanations. Following Llopart et al. (2002), we investigated 3 predictions associated with directional selection using the sequence data surrounding the intron presence–absence polymorphism: 1) the lineage of allelic variants containing the novel intron will have a frequency distribution that is skewed toward rare variants (Braverman et al. 1995; Simonsen et al. 1995), 2) the lineage of allelic variants that do not have a novel intron will have a frequency distribution skewed toward an excess of intermediate frequency variants, and 3) the complete set of allelic variants (i.e., both with and without the novel intron) will have an excess of derived variants segregating at high frequency (Fay and Wu 2000).

The frequency distribution of allelic variants can be assessed with the tests of Tajima (1989), Fu and Li (1993), and Fay and Wu (2000). Tajima's D measures the normalized difference between the average number of pairwise nucleotide differences (π) and the scaled number of segregating sites (θ_w, Watterson 1975; Tajima 1989). Fu and Li's test calculates the difference between the number of polymorphic sites in external versus internal phylogenetic branches using an outgroup (Fu and Li 1993). These tests were conducted on intron-containing and intron-lacking alleles for total sites, nonsynonymous sites, and synonymous sites for the 5 populations in Oregon that contained individuals with either intron 2a or intron 2b. Under neutrality, panmixia, and equilibrium, values for Tajima's

```
AST          GAAAATGTCG AAGA------ ---------- ---------- ---------- ---------- ---------- ---------- ---------G GCATTCCTGC AATGT
EB4          GAAAATGTCG AAGA------ ---------- ---------- ---------- ---------- ---------- ---------- ---------G GCATTCCTCC AATGT
FAT          GAAAATGTCG AAGA------ ---------- ---------- ---------- ---------- ---------- ---------- ---------G GCATTCCTCC AATGT
MAR6         GAAAATGTCG AAGA------ ---------- ---------- ---------- ---------- ---------- ---------- ---------G GCATTCCTCC AATGT
MAR6_RNA     GAAAATGTCG AAGA------ ---------- ---------- ---------- ---------- ---------- ---------- ---------G GCATTCCTCC AATGT
PA35         GAAAATGTCG AAGA------ ---------- ---------- ---------- ---------- ---------- ---------- ---------G GCATTCCTCC AATGT
RON          GAAAATGTCG AAGA------ ---------- ---------- ---------- ---------- ---------- ---------- ---------G GCATTCCTCC AATGT
SHK          GAAAATGTCG AAGA------ ---------- ---------- ---------- ---------- ---------- ---------- ---------G GCATTCCTCC AATGT
TRO          GAAAATGTCG AAGA------ ---------- ---------- ---------- ---------- ---------- ---------- ---------G GCATTCCTCC AATGT
WCH          GAAAATGTCG AAGA------ ---------- ---------- ---------- ---------- ---------- ---------- ---------G GCATTCCTCC AATGT
AMAZ1_1      GAAAATGTCG AAGAgttaag atagcaatta aattattata gttattaatt atatttatta aatatatctt attttaatta aattattagA GCATTCCTCC AATGT
AMAZ1_2      GAAAATGTCG AAGAgttaag atagcaatta aattattata gttattaatt atatttatta aatatatctt attttaatta aattattagA GCATTCCTCC AATGT
AMAZ2_1      GAAAATGTCG AAGAgttaag atagcaatta aattattata gttattaatt atatttatta aatatatctt attttaatta aattattagA GCATTCCTCC AATGT
AMAZ2_2      GAAAATGTCG AAGAgttaag atagcaatta aattattata gttattaatt atatttatta aatatatctt attttaatta aattattagA GCATTCCTCC AATGT
CC3_1        GAAAATGTCG AAGAgttaag atagcaatta aattattata gttattaatt atatttatta aatatatctt attttaatta aattattagA GCATTCCTCC AATGT
CC3_2        GAAAATGTCG AAGAgttaag atagcaatta aattattata gttattaatt atatttatta aatatatctt attttaatta aattattagA GCATTCCTCC AATGT
CC3_RNA      GAAAATGTCG AAGA------ ---------- ---------- ---------- ---------- ---------- ---------- ---------A GCATTCCTCC AATGT
CC6_1        GAAAATGTCG AAGAgttaag atagcaatta aattattata gttattaatt atatttatta aatatatctt attttaatta aattattagA GCATTCCTCC AATGT
CC6_2        GAAAATGTCG AAGAgttaag atagcaatta aattattata gttattaatt atatttatta aatatatctt attttaatta aattattagA GCATTCCTCC AATGT
CC7_1        GAAAATGTCG AAGAgttaag atagcaatta aattattata gttattaatt atatttatta aatatatctt attttaatta aattattagA GCATTCCTCC AATGT
CC7_2        GAAAATGTCG AAGAgttaag atagcaatta aattattata gttattaatt atatttatta aatatatctt attttaatta aattattagA GCATTCCTCC AATGT
OP1_1        GAAAATGTCG AAGAgttaag atagcaatta aattattata gttattaatt atatttatta aatatatctt attttaatta aattattagA GCATTCCTCC AATGT
OP1_2        GAAAATGTCG AAGAgttaag atagcaatta aattattata gttattaatt atatttatta aatatatctt attttaatta aattattagA GCATTCCTCC AATGT
OP1_RNA      GAAAATGTCG AAGA------ ---------- ---------- ---------- ---------- ---------- ---------- ---------A GCATTCCTCC AATGT
OP10_1       GAAAATGTCG AAGAgttaag atagcaatta aattattata gttattaatt atatttatta aatatatctt attttaatta aattattagA GCATTCCTCC AATGT
OP10_2       GAAAATGTCG AAGAgttaag atagcaatta aattattata gttattaatt atatttatta aatatatctt attttaatta aattattagA GCATTCCTCC AATGT
OP11_1       GAAAATGTCG AAGAgttaag atagcaatta aattattata gttattaatt atatttatta aatatatctt attttaatta aattattagA GCATTCCTCC AATGT
OP11_2       GAAAATGTCG AAGAgttaag atagcaatta aattattata gttattaatt atatttatta aatatatctt attttaatta aattattagA GCATTCCTCC AATGT
OP103_1      GAAAATGTCG AAGAgttaag atagcaatta aattattata gttattaatt atatttatta aatatatctt attttaatta aattattagA GCATTCCTCC AATGT
OP103_2      GAAAATGTCG AAGAgttaag atagcaatta aattattata gttattaatt atatttatta aatatatctt attttaatta aattattagA GCATTCCTCC AATGT
LOG13_1      GAAAATGTCG AAGAgttaag atagcaatta aattattata gttattaatt atatttatta aatatatctt attttaatta aattattagA GCATTCCTCC AATGT
LOG13_2      GAAAATGTCG AAGAgttaag atagcaatta aattattata gttattaatt atatttatta aatatatctt attttaatta aattattagA GCATTCCTCC AATGT
LOG18_1      GAAAATGTCG AAGAgttaag atagcaatta aattattata gttattaatt atatttatta aatatatctt attttaatta aattattagA GCATTCCTCC AATGT
LOG18_2      GAAAATGTCG AAGAgttaag atagcaatta aattattata gttattaatt atatttatta aatatatctt attttaatta aattattagA GCATTCCTCC AATGT
LOG29_1      GAAAATGTCG AAGAgttaag atagcaatta aattattata gttattaatt atatttatta aatatatctt attttaatta aattattagA GCATTCCTCC AATGT
LOG29_2      GAAAATGTCG AAGAgttaag atagcaatta aattattata gttattaatt atatttatta aatatatctt attttaatta aattattagA GCATTCCTCC AATGT
LOG29_RNA    GAAAATGTCG AAGA------ ---------- ---------- ---------- ---------- ---------- ---------- ---------A GCATTCCTCC AATGT
LOG50_1      GAAAATGTCG AAGAgttaag atagcaatta aattattata gttattaatt atatttatta aatatatctt attttaatta aattattagA GCATTCCTCC AATGT
LOG50_2      GAAAATGTCG AAGAgttaag atagcaatta aattattata gttattaatt atatttatta aatatatctt attttaatta aattattagA GCATTCCTCC AATGT
LOG69_1      GAAAATGTCG AAGAgttaag atagcaatta aattattata gttattaatt atatttatta aatatatctt attttaatta aattattagA GCATTCCTCC AATGT
LOG69_2      GAAAATGTCG AAGAgttaag atagcaatta aattattata gttattaatt atatttatta aatatatctt attttaatta aattattagA GCATTCCTCC AATGT
LOG52_1      GAAAATGTCG AAGA------ ---------- ---------- ---------- ---------- ---------- ---------- ---------A GCATTCCTAC AATGT
LOG52_2      GAAAATGTCG AAGAGgtaag ttatttttaag ttgttgttat cttaactaat aacttattgt tgttaacct- attttaag-- ---------- GCATTCCTGC AATGT
LOG52_RNA    GAAAATGTCG AAGAR----- ---------- ---------- ---------- ---------- ---------- ---------- ---------- GCATTCCTRC AATGT
GI2_1        GAAAATGTCG AAGAG----- ---------- ---------- ---------- ---------- ---------- ---------- ---------- GCATTCCTGC AATGT
GI2_2        GAAAATGTCG AAGAG----- ---------- ---------- ---------- ---------- ---------- ---------- ---------- GCATTCCTGC AATGT
GI3_1        GAAAATGTCG AAGAGgtaag ttatttttaag ttgttgttaa cttaactaat aacttattgt tgttaacct- attttaag-- ---------- GCATTCTTGC AATGC
GI3_2        GAAAATGTCG AAGAGgtaag ttatttttaag ttgttgttaa cttaactaat aacttattgt tgttaacct- attttaag-- ---------- GCATTCTTGC AATGC
GI8_1        GAAAATGTCG AAGAG----- ---------- ---------- ---------- ---------- ---------- ---------- ---------- GCATTCCTGC AATGT
GI8_2        GAAAATGTCG AAGAGgtaag ttatttttaag ttgttgttaa cttaactaat aacttattgt tgttaacct- attttaag-- ---------- GCATTCTTGC AATGC
GI8_RNA      GAAAATGTCG AAGAG----- ---------- ---------- ---------- ---------- ---------- ---------- ---------- GCATTCYTGC AATGY
GI9_         GAAAATGTCG AAGAG----- ---------- ---------- ---------- ---------- ---------- ---------- ---------- GCATTCCTGC AATGT
GI9_2        GAAAATGTCG AAGAGgtaag ttatttttaag ttgttgttaa cttaactaat aacttattgt tgttaacct- attttaag-- ---------- GCATTCTTGC AATGC
GI11_1       GAAAATGTCG AAGAG----- ---------- ---------- ---------- ---------- ---------- ---------- ---------- GCATTCCTGC AATGT
GI11_2       GAAAATGTCG AAGAGgtaag ttatttttaag ttgttgttaa cttaactaat aacttattgt tgttaacct- attttaag-- ---------- GCATTCTTGC AATGC
GI13_1       GAAAATGTCG AAGAG----- ---------- ---------- ---------- ---------- ---------- ---------- ---------- GCATTCCTGC AATGT
GI13_2       GAAAATGTCG AAGAGgtaag ttatttttaag ttgttgttaa cttaactaat aacttattgt tgttaacct- attttaag-- ---------- GCATTCTTGC AATGC
BAK          GAAAATGTCG AAGA------ ---------- ---------- ---------- ---------- ---------- ---------- ---------A GCATTCCTGC AATGT
DAN          GAAAATGTCG AAGA------ ---------- ---------- ---------- ---------- ---------- ---------- ---------A GCATTCCTGC AATGT
DUT          GAAAATGTCG AAGA------ ---------- ---------- ---------- ---------- ---------- ---------- ---------A GCATTCCTAC AATGT
LL           GAAAATGTCG AAGA------ ---------- ---------- ---------- ---------- ---------- ---------- ---------A GCATTCCTGC AATGT
LCR          GAAAATGTCG AAGA------ ---------- ---------- ---------- ---------- ---------- ---------- ---------A GCATTCCTGC AATGT
MIN          GAAAATGTCG AAGA------ ---------- ---------- ---------- ---------- ---------- ---------- ---------A GCATTCCTGC AATGT
MOO          GAAAATGTCG AAGA------ ---------- ---------- ---------- ---------- ---------- ---------- ---------A GCATTCCTAC AATGT
WAS          GAAAATGTCG AAGA------ ---------- ---------- ---------- ---------- ---------- ---------- ---------A GCATTCCTGC AATGT
WIN          GAAAATGTCG AAGA------ ---------- ---------- ---------- ---------- ---------- ---------- ---------A GCATTCCTAC AATGT
SQL          GAAAATGNCG AAGA------ ---------- ---------- ---------- ---------- ---------- ---------- ---------A GCATTCCTAC AATGT
WATT         GAAAATGTCG AAGA------ ---------- ---------- ---------- ---------- ---------- ---------- ---------A GCATTCCTAC AATGT
```

FIG. 2.—Alignment of intron presence/absence polymorphism and 30 bp of the surrounding exonic region in the *rab4* locus in *Daphnia pulex*, *Daphnia pulicaria*, and *Daphnia melanica*. Exonic sequence is denoted with uppercase letters, and intronic sequence is shown with lowercase letters. The "Oregon *D. pulex*" are denoted with bold typeface. Sequence labels include the name of the population followed by a numeric identifier for each individual collected per population (no number is given when only one individual was collected per population). Both alleles are shown for individuals from populations that contain the novel introns; alleles are differentiated by the label extensions _1 and _2. Nine representative alleles are shown for *D. pulex* and *D. pulicaria* populations that do not contain the intron insertion. Names of individuals that are heterozygous for the intron insertion are italicized. Individual 2 from the GI population is the only *D. pulex* individual sampled in Oregon that did not have the intron insertion.

D, Fu and Li's *D* and *F*, and Fay and Wu's H should be close to zero, whereas a significant deviation from zero suggests a departure from neutral model expectations. No values obtained from our neutrality tests were significantly different from zero, indicating that our data are in agreement with neutral expectations (table 2).

Calculations of nucleotide diversity revealed a range of values for synonymous site diversity for the Oregon populations, being especially low for the CC and OP populations ($\pi_s < 0.0050$) and slightly higher in LOG ($\pi_s = 0.0188$), presumably due to the inclusion of LOG52, a *D. pulex–D. pulicaria* hybrid. When LOG52 is excluded, synonymous diversity is zero for the LOG population. Synonymous site diversity is extraordinarily high for the GI collection site ($\pi_s = 0.0444$), rivaling the species-wide diversity estimates for *D. pulex* and exceeding

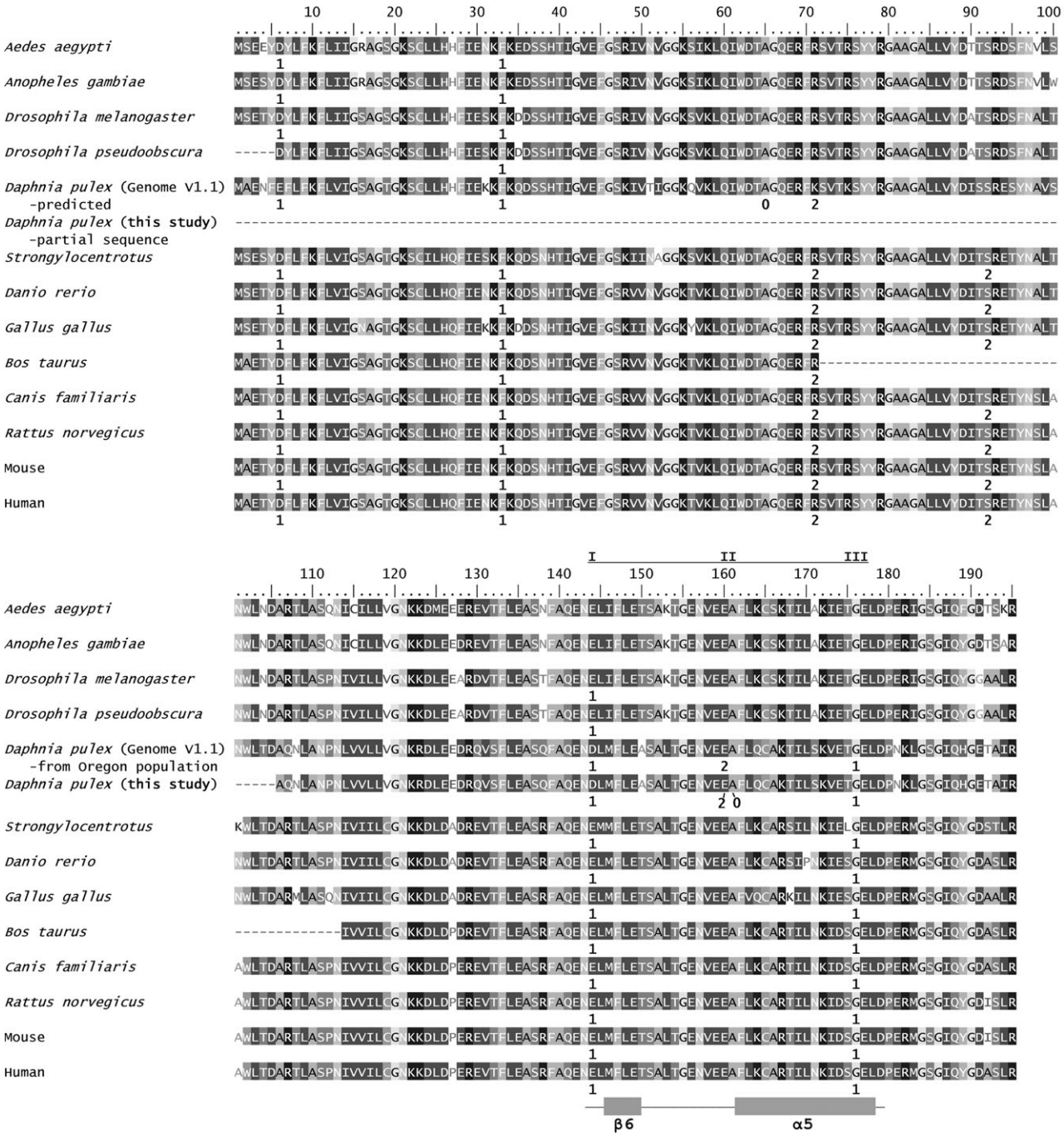

FIG. 3.—Alignment of amino acid sequences for *rab4* from *Daphnia pulex* (both from this study and from the preliminary genome assembly v1.1 available at http://wfleabase.org) and sequence homologs of *rab4* from 12 other metazoans. Numbers below the amino acid sequence indicate the position of introns in sequence and the phase of the intron within the amino acid position. Roman numerals I, II, and III indicate the intron positions in the *D. pulex rab4* amplicon discussed in the text, with the locations and phases of novel introns 2a and 2b in *D. pulex* indicated at II. Protein structural features in the region of the novel intron insertions are also indicated.

the species-wide diversity estimates of both *D. pulicaria* and *D. obtusa* (table 3).

Genome-wide analyses show a weak significant trend of more abundant and longer introns in regions of low recombination in *D. melanogaster* and some vertebrates (Duret et al. 1995; Carvalho and Clark 1999; Comeron and Kreitman 2000). This association has been explained by attributing a beneficial role for introns—they increase intragenic recombination (e.g., Gilbert 1978; Comeron and Kreitman 2000). Alternatively, introns may persist in regions of low recombination due to the reduced efficiency of selection against their presence in such regions (Carvalho and Clark 1999). We removed the novel intron insertion from the *rab4* sequences and calculated the population

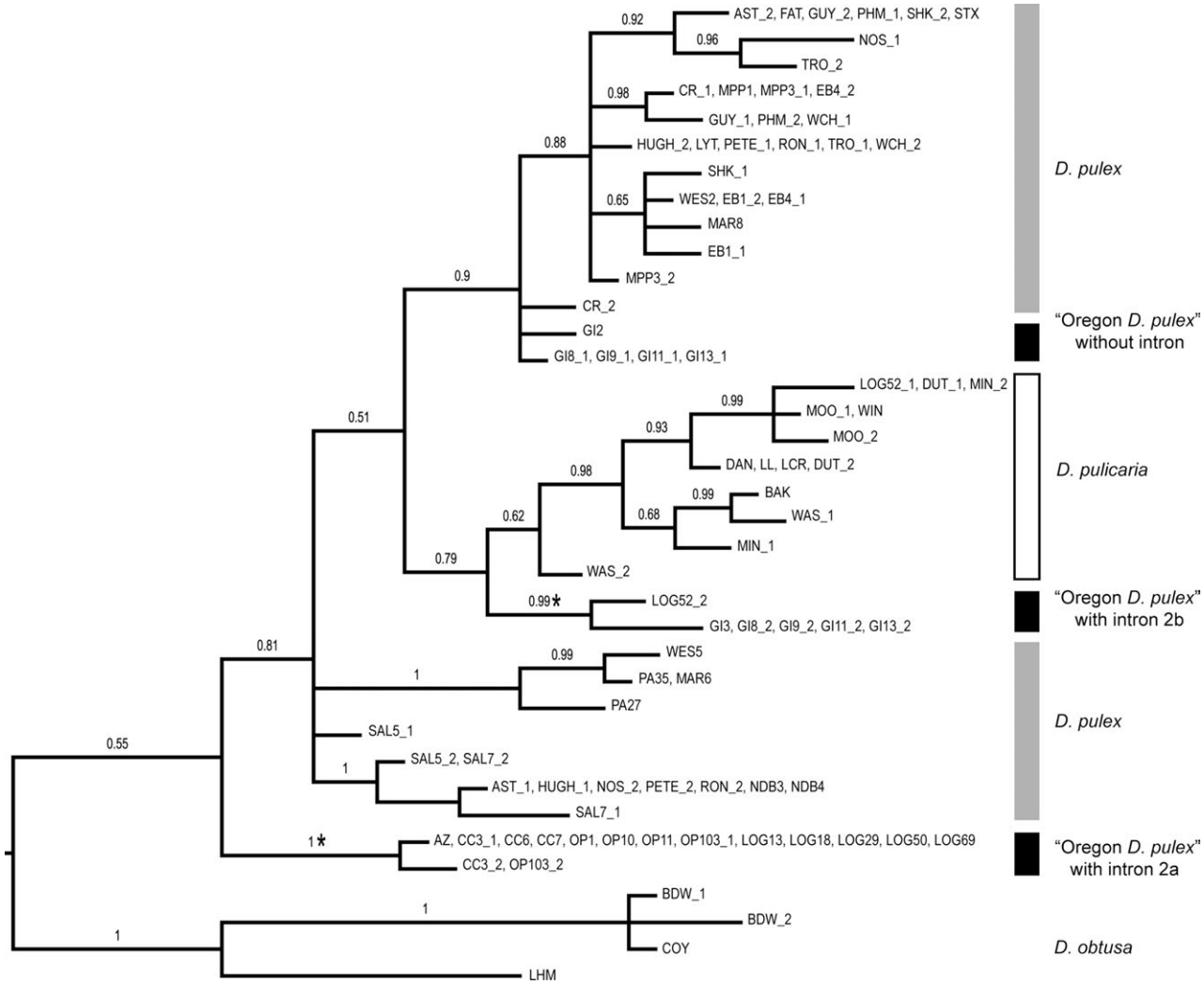

FIG. 4.—The 50% majority-rule consensus Bayesian genealogy of *rab4* alleles in *Daphnia pulex* and *Daphnia pulicaria*. Numbers at nodes are PPs. The tree is rooted with the outgroup species, *Daphnia obtusa*. Labels consist of the population name, followed by the individual number (if more than one individual was sampled) and the allele number (if alleles differed within an individual). Lineages from Oregon are noted by the black rectangular bars to the right of the tree, and lineages with a novel intron are also marked with an asterisk. All individuals containing intron 2a belong to a monophyletic group with a PP of 1.0. Most GI individuals (and LOG52 individual) that contain intron 2b are heterozygotes. Alleles that contain intron 2b are a sister group to *D. pulicaria* (PP = 0.79), and alleles lacking 2b are most closely allied with *D. pulex* (PP = 0.90).

recombination rate, $R = 4N_e r$ (where N_e is the effective population size and r is the rate of recombination per nucleotide per generation), for 3 different categories of *D. pulex* alleles: 1) alleles that contained novel intron 2a, 2) alleles that contained novel intron 2b, and 3) alleles that did not have either intron insertion. The population recombination rate for intron-containing alleles was not significantly different from alleles that lack the intron insertion. However, the confidence intervals for both estimates of recombination rate for the intron-containing alleles were enormous, which is likely due to the low number of variable sites in these alleles (table 4).

Discussion

We observe 2 novel introns that inserted at adjacent nucleotide sites in the *rab4* locus in Oregon populations of *D. pulex*. These introns are not found in any populations of *D. pulex* outside of Oregon or in other *Daphnia* species that we examined, with the exception of one *D. pulex*–*D. pulicaria* hybrid (LOG52). Nor is an intron found at this position in sequence homologs in any other surveyed metazoan species, ranging from *D. melanogaster* to humans. Thus, these novel introns are likely to be evolutionarily recent innovations that originated in the *D. pulex* species complex. The relative isolation of *D. pulex* populations located in Oregon may have restricted the spread of these novel introns. One important mechanism of long-distance dispersal in *Daphnia* is hitchhiking on waterfowl, and because few major migratory flyways cross the Rocky Mountains (Crease et al. 1997), the novel intron-containing alleles are endemic to populations in Oregon.

At present, considerable uncertainty exists over the relative rates of intron loss versus intron gain and the role

Table 2
Tests of Neutrality Were Performed on All Nucleotide Sites of the *rab4* Locus and Included Tajima's *D*, Fu and Li's *D* and *F* (FL—*D* and FL—*F*), and Fay and Wu's *H*

	Tajima's *D*	FL—*D*	FL—*F*	Fay and Wu's *H*
Collection site				
AZ	NA	NA	NA	—
CC	−0.933	−1.133	−1.257	—
GI (alleles containing intron)	NA	NA	NA	—
GI (alleles without intron)	0.851	0.883	1.005	—
GI (all alleles)	—	—	—	0.242
LOG	NA	NA	NA	—
OP	−1.055	−1.262	−1.406	—
All Oregon *D. pulex* alleles	—	—	—	−0.395

NOTE.—Tests performed on synonymous and nonsynonymous sites separately did not show a significant departure from neutrality (results not shown). Individuals from the AZ, CC, and OP populations were fixed for intron 2a. Neutrality tests in the LOG collection site were conducted on alleles containing intron 2a because only one allele-containing intron 2b was present. NA indicates that tests could not be conducted because no polymorphisms existed in the data set. Significance was determined by generating random samples under the hypothesis of selective neutrality and population equilibrium using coalescent simulations; no values were significant.

of introns in genome evolution (e.g., Seo et al. 2001; Rogozin et al. 2003; Roy et al. 2003; Cho et al. 2004; Edvardsen et al. 2004; Qiu et al. 2004; Nguyen et al. 2005; Raible et al. 2005; Roy and Gilbert 2005; Roy and Hartl 2006; Stajich and Dietrich 2006). However, very few observations of population-level polymorphisms have been available for addressing hypotheses of the origins and maintenance of new introns. Most data used to study intron evolution are based upon alignments of conserved loci from a relatively small number of widely divergent model organisms. Although this approach has generated a number of useful insights, it may be biased in that highly conserved genes that can be aligned unambiguously across much of eukaryotic life may also be less tolerant of sequence and/or structural disruptions caused by intron birth and death events (but see Carmel et al. 2007). Furthermore, the nature of sequence data associated with large indel mutations selects against their detection. Individuals that are heterozygous for intron presence/absence states usually cannot be detected without cloning PCR products; these data may often be sidelined due to the monetary and labor expenses associated with cloning. Thus, it may be incorrect to assume that intron presence–absence polymorphisms are rare, at least in *Daphnia*, where we have found 2 parallel events in an initial screen of only 6 loci in 36 individuals (Omilian 2006). It is noteworthy that *Drosophila*, the only other genus for which this type of polymorphism has been reported (Llopart et al. 2002), is a genus for which population-level genetic data are abundant. Comprehensive population-genetic studies of nuclear protein-coding loci in other organisms may yield further examples.

Our findings are relevant to another idea at the heart of the introns-early versus introns-late debate—the preferential insertion of new introns at proto-splice sites. Identical intron positions among divergent taxa are frequently interpreted as representing conservation of intron presence during evolution (Rogozin et al. 2003; Roy and Gilbert 2005). However, it is also possible that intron-position correspondence results from independent insertions into proto-splice sites, that is, "hot spots" for intron insertions (e.g., Dibb and Newman 1989; Coghlan and Wolfe 2004; Qiu et al. 2004; Sadusky et al. 2004; Tordai and Patthy 2004). We have shown that 2 distinct and seemingly unrelated introns have inserted independently into nearly the same site/region (fig. 2). Although the possibility remains that the 2 introns have a single origin and 1 has descended from the other, this is unlikely for 2 major reasons. First, these introns are considerably different both in sequence and GC composition (table 1, fig. 2), yet the silent-site divergence of the surrounding exons is low enough that the gene genealogy across the entire *D. pulex* complex is readily discernible. The novel introns would have to be accumulating mutations at several times the silent-site rate to show no homology. Because sequence diversity within both introns is virtually nonexistent, this is unlikely. Second, the insertion sites of each intron differ by one base pair (fig. 2). The possibility remains, however, that the current introns are 2 different remnants of a larger piece of DNA that was inserted as a single event.

Table 3
Polymorphism Statistics for the *rab4* Locus in Populations of *Daphnia pulex*

	π_T	θ_T	π_n	θ_n	π_s	θ_s
Collection site in Oregon						
AZ	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CC	0.0007	0.0009	0.0000	0.0000	0.0043	0.0056
GI	0.0132	0.0082	0.0012	0.0013	0.0444	0.0256
LOG	0.0073	0.0110	0.0000	0.0000	0.0188	0.0295
LOG (excluding LOG52)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
OP	0.0005	0.0008	0.0000	0.0000	0.0032	0.0049
All *D. pulex*	0.0183	0.0190	0.0010	—[a]	0.0611	—[a]
Daphnia pulicaria	0.0061	0.0066	0.0000	0.0000	0.0112	0.0111
Daphnia obtusa	0.0024	0.0041	0.0000	0.0000	0.0067	0.0130

NOTE.—*Daphnia pulicaria* and *D. obtusa* information is also included. π and θ were estimated for all nucleotide sites (π_T, θ_T), nonsynonymous sites (π_n, θ_n), and synonymous sites (π_s, θ_s).

[a] θ values are not reported where codons differ by multiple changes.

Table 4
Results for Recombination Analyses and Coalescent-Based Estimates of θ (θ_{LAM})

	θ_{LAM}	r_{LAM}	95% support intervals	$4N_e r$
Daphnia pulex alleles with intron 2a	0.0010	5.5×10^{-1}	$<1.0 \times 10^{-9}$, 3.2×10^{-0}	5.5×10^{-4}
Daphnia pulex alleles with intron 2b	0.0029	1.8×10^{-5}	$<1.0 \times 10^{-10}$, 9.0×10^{-1}	5.2×10^{-8}
Daphnia pulex without intron 2	0.0183	3.8×10^{-1}	1.1×10^{-1}, 8.7×10^{-1}	6.9×10^{-3}

NOTE.—The overall recombination rate $r_{LAM} = c/\mu$, where c is the recombination rate per site per generation and μ is the neutral mutation rate per site per generation. Ninety-five percent support intervals are shown for r_{LAM}.

What is the source of these new intron sequences? The novel introns do not have significant homology either to surrounding sequence in the locus or any other sequenced part of the *Daphnia* genome, as evidenced by a Blast search to both the genome assembly and the trace files. Thus, it is unlikely that some of the commonly proposed mechanisms of intron gain are responsible (reviewed in Roy and Gilbert 2006; Lynch 2007); these include transposable elements as a source of new introns (Purugganan and Wessler 1992; Kidwell and Lisch 2000), the duplication of a released intron from an mRNA and reintegration at an ectopic site (Sharp 1985), or the tandem duplication of an internal fragment of coding DNA that contains an AGGT tetramer (Rogers 1990). Like all genome sequencing projects, a small portion of the *Daphnia* genome is likely to remain unsequenced, so the remote possibility exists that the novel introns are derived from some unsequenced portion of the *Daphnia* genome.

It has been suggested that the lack of intron presence–absence polymorphisms observed in natural populations indicates that this type of polymorphism is rarely neutral (Llopart et al. 2002). A previous report of an intron presence–absence polymorphism in the *jingwei* gene in *D. teissieri* determined that positive Darwinian selection was acting on the intron-absent variant (Llopart et al. 2002). Our tests of neutrality applied to the *rab4* locus do not reveal a significant departure from a population in mutation-drift equilibrium (table 2). Furthermore, we were unable to determine whether the polymorphic intron affects the recombination rate in *Daphnia*; estimates of recombination rate in *rab4* for intron-containing versus intron-lacking alleles are not significantly different. However, the lack of variation observed in the intron-containing alleles is likely the cause of the enormous support intervals that accompany our estimates of recombination rate—a situation that cannot be improved except perhaps by sequencing more DNA upstream and downstream of the fragment included in the present analysis.

In the absence of an obvious selective advantage to either intron, the demographic processes that influence the success of a new intron-containing allele may explain our observations. Oregon populations that are fixed (or nearly fixed) for intron insertion 2a are associated with a pronounced reduction in diversity at the *rab4* locus (table 3), thereby suggesting small effective population sizes. The reduction in the efficiency of natural selection in species with smaller effective population size might magnify the probability of retention of a new intron-containing allele (Lynch 2002). If demographic forces are responsible for the patterns we observe, then other unlinked loci should also have low estimates of silent-site diversity because demographic forces are expected to affect all loci in the genome equally, whereas selection has localized effects. An analysis of diversity in 5 additional loci was conducted in a subset of the Oregon populations included in the present study using previously published data (Omilian 2006). Here, it is shown that the LOG population has substantially lower diversity than all other populations of *D. pulex* (see supplementary fig. 1, Supplementary Material online). Thus, novel intron 2a may have gone to near fixation simply due to genetic drift in a population with small effective size.

Because the novel intron-containing (2b) allele that is segregating in the GI population is associated with extremely high genetic diversity, invoking small effective population size as a catalyst for the moderate spread of this allele is unwarranted. Rather, inspection of the alignment and phylogenetic tree indicates that most individuals with intron 2b are heterozygotes at the *rab4* locus. Most individuals containing intron 2b have one intron-lacking allele that groups with *D. pulex rab4* alleles and one intron-containing allele that clusters with *D. pulicaria* alleles. Thus, the possibility exists that novel intron 2b originated in *D. pulicaria*, which is capable of hybridizing with *D. pulex*. However, intron 2b was not observed in any of the other 9 sampled *D. pulicaria* populations. *Daphnia melanica*, a closely related congener that is endemic to Oregon, was also examined; but sequencing of the *rab4* locus in 2 populations of *D. melanica* failed to reveal intron 2b. So, it is unclear if intron 2b originated within a diverging population of *D. pulex* or a closely related species capable of hybridizing with *D. pulex*. Regardless of the species of origin, the ultimate source of the novel intron-containing alleles remains unknown.

Supplementary Material

Supplementary figure 1 and tables 1 and 2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Teri Crease, Derek Taylor, Amanda Seyfert, Brian Eads, Brian Molter, Niles Lehman, and John Colbourne for their generous assistance with various aspects of this project. Jeff Dudycha and Emily Williams assisted with allozyme analyses. Desiree Allen, Carla Caceres, Sandy Connelly, Jeff Dudycha, John Havel, David Innes, Rebecca Klaper, Karen Looper, Susanne Paland, Mike Pfrender, Sarah Schaack, and Emily Williams provided

Daphnia specimens. This work was supported by a National Science Foundation Integrative Graduate Education and Research Traineeship fellowship to A.R.O., National Science Foundation grant DBI-0434671 to D.G.S., and National Science Foundation grants DEB-0196450 and EF-0328516 and National Institutes of Health grant R01-GM36827 to M.L.

Literature Cited

Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Berget SM, Moore C, Sharp PA. 1977. Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. Proc Natl Acad Sci USA. 74:3171–3175.

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics. 140:783–796.

Carmel L, Rogozin IB, Wolf YI, Koonin EV. 2007. Evolutionarily conserved genes preferentially accumulate introns. Genome Res. 17:1045–1050.

Carvalho AB, Clark AG. 1999. Intron size and natural selection. Nature. 401:344.

Cavalier-Smith T. 1985. Selfish DNA and the origin of introns. Nature. 315:283–284.

Cho S, Jin SW, Cohen A, Ellis RE. 2004. A phylogeny of *Caenorhabditis* reveals frequent loss of introns during nematode evolution. Genome Res. 14:1207–1220.

Chow LT, Gelinas RE, Broker TR, Roberts RJ. 1977. An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. Cell. 12:1–8.

Coghlan A, Wolfe KH. 2004. Origins of recently gained introns in *Caenorhabditis*. Proc Natl Acad Sci USA. 101: 11362–11367.

Colbourne JK, Crease TJ, Weider LJ, Hebert PDN, Dufresne F, Hobaek . 1998. Phylogenetics and evolution of a circumarctic species complex (Cladocera: *Daphnia pulex*). Biol J Linn Soc. 65:347–365.

Colbourne JK, Hebert PDN. 1996. The systematics of North American *Daphnia* (Crustacea: Anomopoda): a molecular phylogenetic approach. Proc R Soc Lond B Biol Sci. 351:349–360.

Comeron JM, Kreitman M. 2000. The correlation between intron length and recombination in *Drosophila*. Dynamic equilibrium between mutational and selective forces. Genetics. 156: 1175–1190.

Crease TJ, Lee SK, Yu SL, Spitze K, Lehman N, Lynch M. 1997. Allozyme and mtDNA variation in populations of the *Daphnia pulex* complex from both sides of the Rocky Mountains. Heredity. 79:242–251.

Cronn R, Cedroni M, Haselkorn T, Grover C, Wendel JF. 2002. PCR-mediated recombination in amplification products derived from polyploid cotton. Theor Appl Genet. 104:482–489.

de Souza SJ. 2003. The emergence of a synthetic theory of intron evolution. Genetica. 118:117–121.

Dibb NJ, Newman AJ. 1989. Evidence that introns aroses at proto-splice sites. EMBO J. 8:2015–2021.

Doolittle WF. 1978. Genes in pieces: were they ever together? Nature. 272:581–582.

Doyle JJ, Doyle JL. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull. 19:11–15.

Duret L, Mouchiroud D, Gautier C. 1995. Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. J Mol Evol. 40:308–317.

Edvardsen RB, Lerat E, Maeland AD, Flat M, Tewari R, Jensen MF, Lehrach H, Reinhardt R, Seo HC, Chourrout D. 2004. Hypervariable and highly divergent intron-exon organizations in the chordate *Oikopleura dioica*. J Mol Evol. 59:448–457.

Elkin CJ, Richardson PM, Fourcade HM, Hammon NM, Pollard MJ, Predki PF, Glavina T, Hawkins TL. 2001. High-throughput plasmid purification for capillary sequencing. Genome Res. 11:1269–1274.

Evans RM, Fraser N, Ziff E, Weber J, Wilson M, Darnell JE. 1977. The initiation sites for RNA transcription in Ad2 DNA. Cell. 12:733–739.

Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. Genetics. 155:1405–1413.

Fedorova L, Fedorov A. 2003. Introns in gene evolution. Genetica. 118:123–131.

Felsenstein J. 1993. PHYLIP (phylogeny inference package), version 3.5c. Seattle (WA): Department of Genetics, University of Washington.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. Genetics. 133:693–709.

Geyer CJ. 1991. Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. Technical Report No. 568. School of Statistics, University of Minnesota, Minneapolis.

Gilbert W. 1978. Why genes in pieces? Nature. 271:501.

Gilbert W. 1987. The exon theory of genes. Cold Spring Harb Symp Quant Biol. 52:901–905.

Goldberg S, Schwartz H, Darnell JE. 1977. Evidence from UV transcription mapping in HeLa cells that heterogeneous nuclear RNA is the messenger RNA precursor. Proc Natl Acad Sci USA. 74:4520–4523.

Hebert PDN. 1995. The Daphnia of North America: an illustrated Fauna. CD-ROM. Ontario (Canada): Department of Zoology, University of Guelph.

Hebert PDN, Beaton MJ, Schwartz SS, Stanton DJ. 1989. Polyphyletic origins of asexuality in *Daphnia pulex*. I. Breeding system variation and levels of clonal diversity. Evolution. 43:1004–1015.

Hebert PDN, Schwartz SS, Ward RD, Finston TL. 1993. Macrogeographic patterns of breeding system diversity in the *Daphnia pulex* group. I. Breeding systems of Canadian populations. Heredity. 70:148–161.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 17:754–755.

Kidwell MG, Lisch DR. 2000. Transposable elements and host genome evolution. Trends Ecol Evol. 15:95–99.

Kishino H, Hasegawa M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data and the branching order in Hominoidea. J Mol Evol. 29:170–179.

Koonin EV. 2006. The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? Biol Direct. 1:22.

Kuhner MK. 2006. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. Bioinformatics. 22:768–770.

Kumar S, Tamura K, Nei M. 2004. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform. 5:150–163.

Llopart A, Comeron JM, Brunet FG, Lachaise D, Long M. 2002. Intron presence-absence polymorphism in *Drosophila* driven by positive Darwinian selection. Proc Natl Acad Sci USA. 99:8121–8126.

Lynch M. 2002. Intron evolution as a population-genetic process. Proc Natl Acad Sci USA. 99:6118–6123.

Lynch M. 2007. The origins of genome architecture. Sunderland (MA): Sinauer Associates, Inc.

Martin W, Koonin EV. 2006. Introns and the origin of nucleus-cytosol compartmentalization. Nature. 440:41–45.
Nguyen HD, Yoshihama M, Kenmochi N. 2005. New maximum likelihood estimators for eukaryotic intron evolution. PLoS Comput Biol. 1:631–638.
Nguyen HD, Yoshihama M, Kenmochi N. 2007. The evolution of spliceosomal introns in Alveolates. Mol Biol Evol. 24: 1093–1096.
Omilian AR. 2006. Features of *Daphnia* genome evolution [dissertation]. [Bloomington (IN)]: Indiana University. p. 49–91.
Orgel LE, Crick FH. 1980. Selfish DNA: the ultimate parasite. Nature. 284:604–607.
Palmer JD, Logsdon JM. 1991. The recent origins of introns. Curr Opin Genet Dev. 1:470–477.
Pereira-Leal JB, Seabra MC. 2000. The mammalian Rab family of small GTPases: definition of family and subfamily sequence motifs suggests a mechanism for functional specificity in the Ras superfamily. J Mol Biol. 301:1077–1087.
Pereira-Leal JB, Seabra MC. 2001. Evolution of the Rab family of small GTP-binding proteins. J Mol Biol. 313:889–901.
Posada D, Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics. 14:817–818.
Purugganan M, Wessler S. 1992. The splicing of transposable elements and its role in intron evolution. Genetica. 86:295–303.
Qiu WG, Schisler N, Stoltzfus A. 2004. The evolutionary gain of spliceosomal introns: sequence and phase preferences. Mol Biol Evol. 21:1252–1263.
Raible F, Tessmar-Raible K, Osoegawa K, et al. (12 co-authors). 2005. Vertebrate-type intron-rich genes in the marine annelid *Platynereis dumerilii*. Science. 310:1325–1326.
Rogers JH. 1990. The role of introns in evolution. FEBS Lett. 268:339–343.
Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV. 2003. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. Curr Biol. 13:1512–1517.
Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19:1572–1574.
Roy SW, Fedorov A, Gilbert W. 2003. Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. Proc Natl Acad Sci USA. 100:7158–7162.
Roy SW, Gilbert W. 2005. Rates of intron loss and gain: implications for early intron evolution. Proc Natl Acad Sci USA. 102:5773–5778.
Roy SW, Gilbert W. 2006. The evolution of spliceosomal introns: patterns, puzzles, and progress. Nat Rev Genet. 7:211–221.
Roy SW, Hartl DL. 2006. Very little intron loss/gain in *Plasmodium*: intron loss/gain mutation rates and intron number. Genome Res. 16:750–756.
Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics. 19:2496–2497.
Sadusky TA, Newman J, Dibb NJ. 2004. Exon junction sequences as cryptic splice sites: implications for intron origin. Curr Biol. 14:505–509.
Seo HC, Kube M, Edvardsen RB, et al. (11 co-authors). 2001. Miniature genome in the marine chordate *Oikopleura dioica*. Science. 294:2506.
Sharp PA. 1985. On the origin of RNA splicing and introns. Cell. 42:397–400.
Simonsen KL, Churchill GA, Aquadro CF. 1995. Properties of statistical tests of neutrality for DNA polymorphism data. Genetics. 141:413–429.
Stajich JE, Dietrich FS. 2006. Evidence of mRNA-mediated intron loss in the human-pathogenic fungus *Cryptococcus neoformans*. Eukaryot Cell. 5:789–793.
Stenmark H, Olkkonen VM. 2001. The Rab GTPase family. Genome Biol. 2(5):1–7 reviews3007.
Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics. 105:437–460.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 123: 585–595.
Tajima F. 1996. The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. Genetics. 143:1457–1465.
Thompson JD, Higgins DG, Gibson TJ. 1994. ClustalW—improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties, and weight matrix choice. Nucleic Acids Res. 22:4673–4680.
Tordai H, Patthy L. 2004. Insertion of spliceosomal introns in proto-splice sites: the case of secretory signal peptides. FEBS Lett. 575:109–111.
Watterson GA. 1975. Number of segregating sites in genetic models without recombination. Theor Popul Biol. 7:256–276.
Zerial M, McBride H. 2001. Rab proteins as membrane organizers. Nat Rev Mol Cell Biol. 2:107–117.

Kenneth Wolfe, Associate Editor

Accepted July 16, 2008