Trends in
Entomology
Vol. 15, 2019

Original Article

Suitability of COI, COII, 18S, and ITS2 for phylogeny and systematics of black fly (Diptera: Simuliidae)

Christine Gaudreau[1,2,#], Bernard LaRue[1,&,$] and Guy Charpentier[1,*,&]

[1]Département de Chimie-Biologie, Université du Québec à Trois-Rivières, 3351 Boulevard des Forges, Box 500, Trois-Rivières, G9A 5H7, Canada; [2]CHUM, 1100 rue Sanguinet, 7^{e} étage Pavillon F, Montréal, H2X OC1, Canada.

ABSTRACT

A sample of 15 blackfly (Diptera: Simuliidae) species collected in Quebec as larval specimens is used to evaluate COI, COII and 18S rDNA for their abilities to resolve distinct levels of evolutionary depth. Following characterization of the mutation pattern, the COI and COII computer simulations show that too frequent base interchanges at the degenerate third codon position quickly lead to early mutation saturation and to a loss of phylogenetic resolution. However, both mtDNA markers appear well suited for distinguishing closely related species. Molecular evidence also sheds doubts on the value of morphology alone for classifying Prosimuliini larvae, up to the point of calling for a taxonomic reappraisal of the whole tribe. At the opposite end of the phylogenetic spectrum, 18S elucidates remote divergence events while losing resolving power at low levels due to sequence conservatism as embodied by a limited number of informative sites being concentrated within two hot spot regions. In an attempt to fill the gap between ancient and recent divergence events, a composite tree was built from 18S, ITS2 and the COI/COII codons undergoing amino acid changes. Results show that a well weighed combination of markers evolving on different time scales can cover multiple taxonomic levels.

KEYWORDS: COI, COII, 18S rDNA, Simuliidae, evolutionary simulations.

INTRODUCTION

Classical systematic makes an intensive use of morphological and ecological criteria. However, the need to classify often conflicts with the reality of the evolutionary continuum. This is particularly true for complexes of cryptic species as frequently found in many groups of small invertebrates. To resolve them, cytogenetics and molecular phylogenies can be added to the previous criteria. Even so, a lack of congruence between and within these various approaches has been frequently noticed, in particular with respect to mitochondrial versus nuclear DNA markers.

The nuclear ribosomal DNA cluster (rDNA) and the mitochondrial cytochrome oxidase I (COI) and II (COII) genes are common genomic targets given the availability of universal polymerase chain reaction (PCR) primers [1, 2]. Due to differences in evolutionary conservatism, each probe looks at distinct levels of phylogenetic depth. The rDNA transcription unit, a mosaic of conserved and variable regions, exhibits little intragenomic divergence between its hundreds to thousands of individual copies. Although the 18S rDNA finds some applications in barcoding [3, 4], the highly conserved 18S, 5.8S and 28S genes help mostly to construct high level phylogenies [5-7]. The fast-evolving intergenic spacers ITS1 and ITS2 discarded during rRNA maturation are

*Corresponding author: guy.charpentier@uqtr.ca
[#]christine.gaudreau.chum@ssss.gouv.qc.ca
[&]Retired; [$]Deceased.

prone to insertions or deletions, complicating sequence alignment [8]. ITS2 is better suited to low or intermediate levels of phylogenetic reconstructions [9-11] and ITS1 to population studies, including Simuliidae [12, 13], and species authentication [14- 16].

Their coding function constrains variation within the COI and COII genes. The high conservatism of the two polypeptides provides for an unambiguous DNA sequence alignment, but the genes themselves evolve at a quick pace mainly through silent mutations of the third codon position [17]. COI finds a wide use at low phylogenetic levels [18-22] and its proposal as a standard for the molecular barcoding of species [23] has generated a large body of data for many animal and plant groups.

Black flies, a nematoceran sub-family with 2177 known living species [24], appear as the taxonomist's nightmare due to morphological and ecological similarities. An environmentally induced plasticity and a prevalence of biotypes and complexes of species also bring additional complications (see for instance [25, 26]). This classification problem has relevance to basic research and also to health issues involving the selective control of onchocerciasis which depends on a reliable identification of vector species from the *Simulium damnosum* complex Theobald, 1903, and on an understanding of their respective roles in pathogen transmission [27]. Based on fixed-inversion differences in polytene salivary gland chromosomes of larvae, many nominal black fly species now appear as complexes of siblings. However, cytological techniques do not allow the identification of biting adult females [15], but usually require samples from late larval instars [28] and have other severe limitations such as the required amount of work and a high failure rate [29]. At the molecular level, applications of COI and COII include species resolution [30-34], the study of phylogeographic variation [13, 17, 35, 36] and the deciphering of subgenera evolutionary history [37, 38].

On our data from a multigenetic sample of 15 black fly species from Quebec, we performed computer simulations to define more generally the operational time frame of COI and COII with respect to taxonomic depth. At shallow levels, their ability to resolve pairs of black fly morphospecies with an ambiguous status was also investigated and the validity of some larval traits currently used to classify members from the *Prosimulium* genus Roubaud, 1906, questioned by pitting morphology against molecular data. We also built a composite tree encompassing all 15 species from 18S, ITS2 and the more conserved segments of the COI and COII genes to investigate the potential of this composite approach in resolving simultaneously various levels of phylogenetic depth.

MATERIALS AND METHODS

General collecting

Actual specimens of black fly larvae from Quebec were part of a larger sample from a field study [39], with collection and storage prior to DNA extraction. Larvae were identified under a low-magnification microscope according to the morphological keys used in Wood *et al.* [40] and Adler *et al.* [41].

DNA amplification and sequencing

DNA extraction and purification were performed as described by St-Onge *et al.* [3]. PCRs for COII [13] and COI [17] were performed as described earlier. In rare instances of initial PCR failure, the COI region was recovered through other pairs of primers [42]. Our COI primers, selected for their efficiency with black fly DNA, amplify the 3' end of the gene as opposed to the 5' end-specific barcoding primers LCO1490 and HCO2198 [23]. The two kinds of PCR products do not overlap at all. The COII primers also target the 3' half of the gene. The 18S gene was amplified in two separate sections overlapping on a 130 nucleotide (nt) stretch. CTGGTTGATCCTGCCAGTAG (forward) and GTGGTGCCGTTCCGTCAATTCC (reverse) were 5' half primers, and GGATCGAAGGCGA TTAGATAC (forward) and CTTCCGCAGGTTC ACCTACG (reverse) targeted the 3' half. Either half was amplified in 50 µL of a 10 mM Tris-HCl, pH 8.3, 50 mM KCl, 1.5 mM $MgCl_2$ buffer containing 1 unit of Taq DNA polymerase (Roche), 200 µM of each dNTP and 4 µM of each primer. Thermocycling consisted of an initial 60 s denaturation step at 94 °C followed by 38 amplification cycles (30 s, 94 °C + 50 s, 51 °C +

120 s, 72 °C) and a final 180 s extension step at 72 °C. Sequencing in both directions ensured on the average an 85% redundancy for the entire 18S gene sequence. The split 5.8S gene (including 5.8S rDNA proper, spacer 2a and 2S rDNA) was recovered from the PCR products obtained through the combined use of the leftward ITS1 [13] and rightward ITS2 primers [10]. Prior to automated dideoxy sequencing, all PCR products underwent agarose gel electrophoresis followed by extraction and purification [13].

Sequence data analysis

Alignment of DNA sequences from rDNA (18S and 5.8S), COI and COII was performed with the Clustal W program implanted in Geneious software version 4.7.6. [43] and further refined by manual adjustments if required. For each gene, we performed maximum-likelihood (ML) analysis using PHYML, version 2.4.4 [44]. Modeltest 3.4 [45] indicated general time-reversible (GTR) to be the appropriate substitution model for ML based on a hierarchical likelihood ratio test or Akaike's information criterion. The credibility of nodes in the tree was assessed by 500 bootstrap replicates.

Simulation of COI and COII evolution

The evolution of the COI/COII genes, treated as a single unit, was simulated through an algorithm built in Excel software from Microsoft Office 2007. The simulation concerned exclusively the third codon position and assumed all mutations to be synonymous with respect to amino acid identity. It started by creating a random progenitor sequence having a 300 nt length – same as the number of codons in our concatenated COI/COII sequences – and a base composition replicating the one found at the third position in the average black fly COI/COII sequence. While doing so, the creation of the progenitor and the ensuing simulation considered separately codon families specifying a given amino acid, namely XXN, XXY or XXR (X, a given base; N, any base; Y, T/C; R, A/G). The simulation next assumed that each nucleotide interchange, more or less envisioned as a chemical equilibrium, occurs at equal rates in both directions. The rate of a given mutation was set on an arbitrary scale by multiplying the frequencies of the two exchanged bases and next readjusted to obtain an overall transition to transversion (Ts/Tv) ratio such as chosen for a specific simulation. Afterwards, the previous rates were normalized as standard probabilities on a 0 to 1 scale. To implement a mutation, the specific type of base substitution was first picked at random following the above probabilities and a target in the sequence was next determined by lottery among all the bases able to initiate this peculiar type of change. To mimic evolution more realistically, the number of mutations corresponding to a given branch length was considered as a mean and further modulated through the Poisson distribution before performing mutations. To create trees, offspring sequences from previous cycles were reintroduced into the procedure, branch lengths modified as required and so on.

RESULTS

Validating the sample

As described in Table 1 which also provides Genbank accession numbers for COI, COII and rDNA sequences, we analyzed 15 black fly species from Quebec spanning 5 genera. Morphological identifications were validated by two molecular criteria: (i) externally, through close matches with independent GenBank entries whenever available and (ii) internally, as all specimens from any given morphological unit should, and indeed do, cluster together as an exclusive group according to ITS1 (highly sensitive to species identity) [10], COI and COII sequences. Specimens of *S. pictipes* (Chutes Mont-à-Peine; 46°12`N, 73°34`W) and *S. longistylatum* (Chutes Dunbar; 46°57`N, 73°07`W) were collected about 90 km apart in distinct streams and identified according to the ancient key [40]; see discussion with respect to the use of these names. As described later, the sample also contained at least two *Prosimulium* molecular species (shortened as *P*.sp.1 and *P*.sp.2) that did not consistently relate to known morphospecies. Globally, COI or COII sequences varied in numbers from 1 to 28 depending on the species under study. Excluding a presumed 41 nt length missing at the 5` end of the gene due to the forward primer location, the entire 18S rDNA sequence was obtained for each species from a single specimen already authenticated through COI and COII sequences.

Table 1. List of species.

Species (no specimens)*, author	**COI/COII/rDNA (Genbank no)**	**Molecular validation of morphospecies identification****
Cnephia dacotensis (2), Dyar and Shannon 1927	GQ280358/GQ265806/FJ436346	Int (ITS1), Ext (COI / AF425841)
Prosimulium 1 (19)	GQ355808/GQ355810/FJ595476	Int (COI, COII, ITS1)
Prosimulium 2 (4)	GQ355809/GQ355811/FJ595477	Int (COI, COII, ITS1)
Simulium annulus (4), Lundström 1911	GQ280360/GQ265808/FJ437566	Int (COI)
Simulium aureum (3), Fries 1824	GQ280361/GQ280348/FJ437565	Int (ITS1), Ext (COII / AF083862)
Simulium decorum (2), Walker 1848	GQ280362/GQ280349/FJ437564	Int (COI, COII, ITS1), Ext (COII / DQ284522)
Simulium jenningsi (1), Malloch 1914	GQ280364/GQ280351/FJ436353	Ext (COII / DQ284509)
Simulium longistylatum (25), Shewell 1959	GQ280366/GQ280352/FJ436352	Int (COI, COII, ITS1)
Simulium pictipes (4), Hagen 1880	GQ280367/GQ280353/FJ436351	Int (COI, COII, ITS1)
Simulium quebecense (2), Twinn 1936	GQ280368/GQ280354/FJ436350	Int (ITS1)
Simulium tuberosum (2), Lundström 1911	GQ280369/GQ280355/FJ436349	Int (ITS1, ITS2), Ext (COII / DQ284515)
Simulium venustum (3), Say 1823	GQ280370/GQ280356/FJ436348	Int (COI, COII, ITS1)
Simulium vittatum (28), Zetterstedt 1838	FJ231202/FJ231200/FJ231203	Int (COI, COII, ITS1), Ext (rDNA / U48383)
Stegopterna mutata (2), Malloch 1914***	GQ280359/GQ265807/FJ436345	Int (ITS1)
Twinnia tibblesi (1) Stone and Jamnback 1955	GQ280371/GQ280357/FJ436347	N/A

*Specimen with known sequences of any type (COI, COII and/or ITS1).

**Int: identification supported by high sequence homology between all specimens. Ext: independant Genbank match (sequence/accession no).

***Previously known as *Cnephia mutata*; see Adler *et al.* [41].

COI and COII: general considerations

Edited sequences from the COI and COII genes were trimmed to 663 (221 codons) and 252 nt (84 codons) lengths, respectively, without any need for gaps during interspecific alignment. They all possessed a single open reading frame and did not show dual-base positions or unusual numbers of amino acid changes, which excludes the presence of mtDNA nuclear copies in the dataset. Since both gene segments evolve at comparable rates [13], they were usually treated as a single unit, designated as COI/COII. Only slight variations of base content, likely ascribed to random fluctuations, occurred between species. Composition was quite well balanced for the first codon position, while an excess of T at the second position (Table 2) reflected the high content (58%) of hydrophobic amino acids in the corresponding polypeptides. As usual for arthropod mtDNA genes, the third position showed a very heavy A + T bias. To get a closer look at the mutation pattern, a sample of one sequence per species was entered into a constrained tree identical to the one previously determined for ITS2 [10] and rooted at the basal divergence between Simuliini and Prosimuliini. The latter tree was judged reliable enough for the present purpose due to a generally good bootstrap support. The resulting count (Table 2), performed manually at each position so as to minimize the number of substitutions, indicated a huge majority of third position mutations, with only 0.8% of these involving amino acid substitutions. Given that equivalent counts often resulted from several possible pathways, the number of silent changes could be severely underestimated due to many undetected reversions, parallel mutations and multiple hits. Also, about two thirds of the first position mutations involved neutral substitutions between TTR and CTN leucine codons. When considering all three positions simultaneously, 94% of all mutations appeared of the silent type. If not the case (34/305 codons involved, mostly first position hits), amino acid changes produced a majority of presumably neutral switches of hydrophobic amino acids or within a trio made of serine, threonine and alanine (Table 3).

Mimicking COI and COII evolution

Since the first and second codon positions can be practically neglected due to their low impact on the overall mutation picture, composition imbalances and high substitution rates of the third one could easily lead to early mutation saturation and severely impair phylogenetic reconstruction [46]. Indeed, a COI/COII tree built under ML analysis showed extremely poor bootstrap support and erratic branching relative to ITS2 even within the *Simulium* genus Latreille, 1802, (data not shown). The effect of total saturation was evaluated from the proportions of the three codon families (XXN = 0.472, XXY = 0.330, XXR = 0.198) and their individual base contents (Table 2). At a randomly chosen position from two random sequences

Table 2. Base content and mutation pattern of COI/COII (sample of 15 species). X, a given base; N, any base; R, A or G; Y, C or T.

							Mutations		
		% A	% T	% G	% C	% variable	Total	Per position [1]	% silent
Codon position	1st	27.0	27.1	29.3	16.6	19.0	167	2.88	67.7
	2nd	19.5	41.3	15.5	23.7	2.3	11	1.57	0
	3nd	42.3	41.0	4.2	12.5	87.5	983	3.7	99.2
Codon family [2]	XXN	52.3	33.5	5.1	9.1				
	XXY	—	76.4	—	23.6				
	XXR	92.2	—	7.8	—				

[1] Considering only variable positions.

[2] Box of codons differing only at the third position and encoding the same amino acid.

Table 3. Amino acid substitutions in the COI and COII polypeptides. Dot: identity with the consensus sequence. Codon numbering relative to our Genbank sequence (see Table 1).

	COI																				COII													
No codon					1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2														
	6	6	7	7	2	3	4	4	4	4	5	7	8	8	0	0	0	1	1	1			1	2	2	3	3	3	3	4	4	5	5	6
	1	5	0	2	2	0	1	4	5	8	2	8	1	2	4	6	9	0	2	4	6	8	3	6	9	3	5	7	8	3	8	3	7	8
Consensus	T	Y	M	A	F	T	A	M	V	I	L	A	V	V	M	S	S	V	Y	I	V	T	T	V	M	I	I	V	T	L	I	V	G	L
T. tibblesi	•	F	•	S	•	S	T	•	I	•	•	T	I	I	I	T	•	A	F	•	I	•	I	•	L	•	V	I	•	•	•	I	A	•
Prosimulium 1	•	•	•	•	•	•	•	•	I	•	I	T	I	•	•	A	T	•	•	•	I	•	•	•	L	•	•	•	S	•	•	•	A	M
Prosimulium 2	•	•	•	•	•	•	•	•	I	•	I	T	I	•	•	A	T	•	•	•	I	•	•	•	L	•	•	•	S	•	•	•	A	M
S. mutata	A	•	•	•	L	•	•	•	I	•	•	T	•	•	•	•	T	•	•	•	I	•	•	I	L	•	•	•	•	•	•	•	A	•
C. dacotensis	•	•	•	•	•	S	•	L	I	•	•	T	I	•	•	T	T	I	•	•	•	•	•	•	L	•	•	•	•	•	•	•	A	•
S. vittatum	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	A	•	•	L	•	•	•	•	•	•	•	•	•	I	V	•	•	•
S. longistylatum	•	•	•	•	•	•	•	•	•	•	•	•	•	I	•	•	•	•	•	L	I	•	M	•	•	•	•	•	•	I	V	•	•	•
S. pictipes	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	L	•	•	•	•	•	•	•	•	•	I	V	•	•	•
S. tuberosum	•	•	I	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	L	•	S	•	•	•	•	•	•	•	•	V	•	•	•
S. venustum	•	•	I	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	V	•	•	•
S. decorum	•	•	I	•	•	•	•	•	•	•	•	•	I	•	•	A	•	•	•	•	•	•	M	•	•	•	•	•	•	•	V	•	•	•
S. jenningsi	•	•	I	•	•	•	•	•	•	•	•	•	•	•	•	A	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
S. aureum	•	•	I	•	•	•	•	•	•	T	•	T	•	•	•	•	•	•	F	•	•	•	•	•	•	•	•	•	•	•	•	•	A	•
S. annulus	•	•	I	•	•	•	•	•	I	•	•	•	•	•	•	A	T	•	•	•	•	•	•	•	•	V	•	•	•	•	V	•	•	•
S. quebecense	•	•	I	•	•	•	•	•	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	A	•

obeying these proportions and composition rules and by setting 'm' as N, Y and R alternately, the probability (p[match]) of base identity may be expressed as:

p[match within 'm'] = ε (individual base frequencies within 'm')^2,

hence: p[match] = ε p[family 'm') x p[match within 'm'],

The relative divergence of random sequences, which amounts to total saturation and is given by p[no match] = 1 - p[match], stood at 43.2% after computation. Incidentally, the same value is obtained by comparing large sets of computer-generated sequences. Since they diverged on average by 34%, real sequences had reached a point where they appeared as being 79% randomized. In addition, they provided a distribution of divergences with an asymmetry coefficient of only -0.02 which indicated that most sequences appear as equidistant and far away from each other. On the contrary and with respect to phylogenetic resolution, a good DNA marker would exhibit a bias towards lower values due to a significant proportion of comparisons involving closely or moderately related species.

To evaluate how fast sequences get randomized, we started from a unique randomly generated progenitor, built under the rules specified in 'Material and Methods', to create through our program a database of independently generated daughter sequences which could be sampled at will as they evolved. As shown in Figure 1, randomization increased asymptotically and reached the level seen in our black fly sample after 0.40-0.55 mutation hits per position, with the highest figure corresponding to saturation process slowing down at high Ts/Tv ratios. Even in this worst-case scenario, about 1.2 hits per position, was sufficient to reach near total saturation. With respect to absolute time, we used Brower's [47] estimate of the COI gene divergence rate, averaged over all three codon positions, of 2.3% My^{-1}. The corresponding figure was rounded to 6% My^{-1} for the third position being considered in isolation by removing the quite negligible contribution from the two others (Table 2). This meant a rate of change of 3% My^{-1} per lineage, a value to be used further on for the sake of consistency (see discussion). Based on Figure 1 and on this molecular clock rate, the average randomization level of 79% within our black fly sample would be reached after 13-18 My.

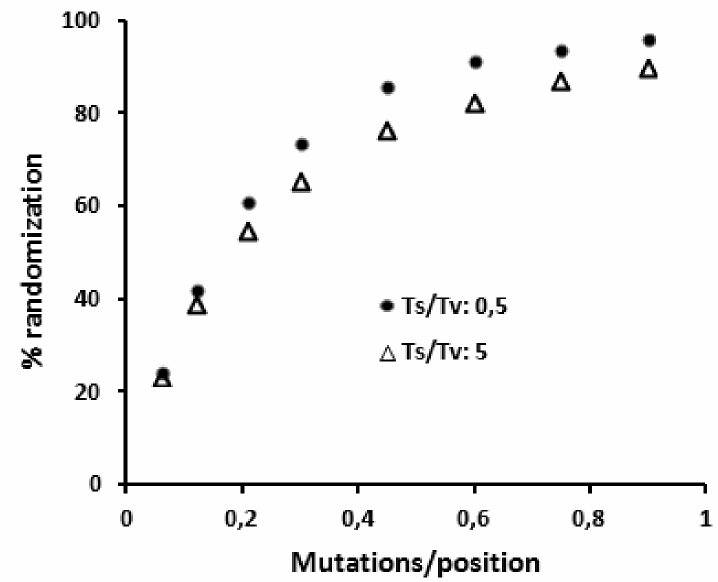

Figure 1. Randomization of the COI/COII third codon position. Randomization = Average divergence of evolving sequences/Divergence of random ones x 100. Sample: 50 offspring sequences born from the same progenitor. Ts/Tv: transition to transversion ratio.

The previous data and simulations strongly suggest that COI/COII provides no resolution for events located over a mutational 'horizon' to be evaluated more precisely. To locate the horizon, we generated three types of fictitious trees: type (a) was rooted at -30 My and split at regular intervals into two branches, one leading directly to present time and the other to the next divergence step further up (staircase tree, Figure 2A); type (b), of the bush-type (Figure 2B), was rooted at -20 to -28 My and springed up into 16 terminal sequences through four consecutive series of divergence events located each at identical levels on parallel branches; type (c) represented various hybrids between (a) and (b). The phylogenetic reconstruction from terminal sequences of the (b) type showed bootstrap support to fall below 50% at about -10 My and the frequent occurrence of erratic branching schemes at -15 to -20 My with reference to the input tree. The lengths of older branches also appeared much undervalued, a normal observation since they should comprise many undetected changes. For bush-type trees, a checkpoint located at the second earliest divergence helped to determine which critical distance (My) between the checkpoint and its adjacent nodes would be needed in order to maintain good resolution. When using a checkpoint at about -14 My and a critical distance of at least 3 My, the topology of the reconstructed tree exactly reproduced the input simulation in multiple and independent assays and the bootstrap support stayed over 80%. Moving the checkpoint at -17 My strictly preserved the right branching scheme in most cases, but at the cost of a bootstrap support sometimes as low as 25% for lower nodes even while recessing the root at -28 My. Erroneous rooting of lower nodes and additional loss of support were observed when the checkpoint was lowered further to about -20 My. Adding another divergence step (i.e. 32 terminal sequences) brought little improvement.

Genetic differentiation within tight groups

With the sole exception of the *Prosimulium* genus, molecular and morphological criteria showed congruence in the sense that all larvae from a same morphospecies formed a completely distinct COI/COII cluster (data not shown. See also Table 1 and comments). Moderate sequence homology and bootstrap support criteria revealed two cases of recently diverged species. In the first instance, molecular data fully agreed with the morphological identifications of *S. longistylatum* and *S. pictipes* as their COI/COII tree (Figure 3A) exhibited two well resolved branches coinciding with morphotypes and diverging from each other by an average of 4.9%. Comparable results were recorded for COII alone, as the 25 and 4 sequences from *S. longistylatum* and *S. pictipes* partitioned into the same distinct clades (data not shown). The second case concerned a sample of 23 specimens from the *Prosimulium* genus, which were collected during wintertime at the outlet of Lac Souris (46°35`N, 72°58`W). They were identified following the hypostomal teeth, postgenal cleft patterns and the head pigmentation design [40, 41]. A temporary classification into six nominal species is not congruent with molecular phylogeny as the various morphospecies appeared intermingled within a COI/COII tree comprising two major branches (Figure 3B). An average COI/COII sequence

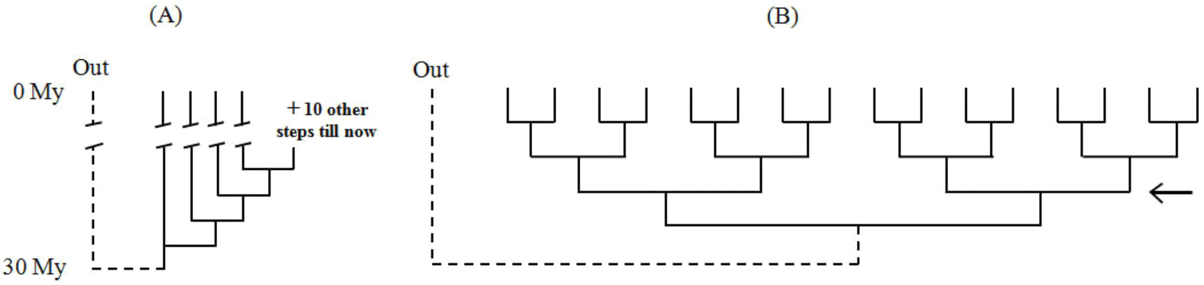

Figure 2. Shapes for fictitious trees. (A) Staircase-type, with an example of 2 My steps. (B) Bush-type. Out, outgroup sequence. Arrow: horizon checkpoint.

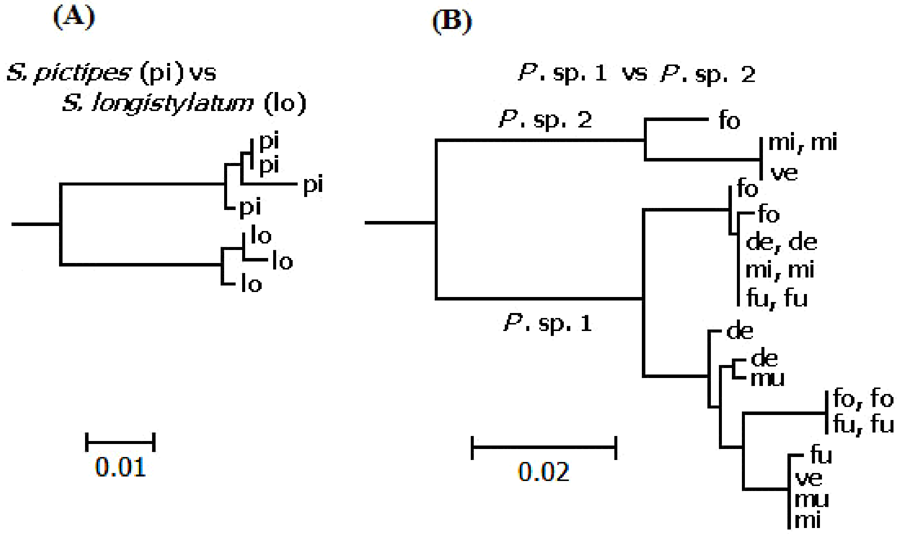

Figure 3. COI/COII sub-trees for (A) *S. pictipes*/*S. longistylatum* and (B) the *P*.sp.1/*P*.sp.2 group. *Prosimulium* nominal species according to larval morphology: fo (mi, fu, mu), *Prosimulium fontanum* (*mixtum, fuscum, multidentatum*); de (ve), *Helodon decemarticulatus* (*vernalis*). Scale bar: substitutions per site.

divergence of 7.8% between lineages (meaning an ancestor about 3.4 My old) hinted at the existence of at least two molecularly defined species, to be further designated as *P*.sp.1 (*Prosimulium* 1) and *P*.sp.2 (*Prosimulium* 2) for lack of formal names.

The 18S-5.8S rDNA cluster

Comparison with the *Drosophila melanogaster* Meigen, 1830, 18S rDNA show that 41 bases are presumably missing from the 5`end of our sequences. If we assume the existence of this segment, our black fly sequences varied in length from 1973 (*P*.sp.1, *P*.sp.2, *T. tibblesi*) to 1955 (*S. tuberosum*) nt while the remainder of those from the *Simulium* genus stayed constant at 1961 nt. The base composition, on the average of 28.5% A, 28.5% T, 24.5% G and 18.5% C, showed minimal interspecific variation. There was no evidence in sequencing chromatograms for dual-base positions in any species, which indicated virtual identity between most gene copies. The 5.8S region, consisting of 5.8S, spacer 2a and 2S rDNA, provided 174 nt sequences that could be aligned without any gap, except for a 2-nt insertion within the spacer 2a sequence from *S. tuberosum*. Due to its weak phylogenic signal [5], the short and highly conserved 5.8S region was joined to 18S before generating the tree shown in Figure 4.

Using two nematoceran outgroups, the 18S tree convincingly resolved the Prosimuliini (*Prosimulium* and *Twinnia* Stone and Jamnback, 1955) from the Simuliini (other Simuliidae) tribe, a basal divergence earlier recognized on both morphological [48] and molecular grounds [6]. *Simulium* appeared as a tight clade in spite of a too weak statistical evidence to root it precisely with respect to the *Cnephia* Enderlein, 1921, and *Stegopterna* Enderlein, 1930, genera. Several bootstrap values within the *Simulium* sub-tree appeared not significant, which likely reflected a lack of sequence variation as exemplified by a maximum of only 22 differences (average: 11.4) between the two most distant representatives of this genus. Consistent with this low resolving power, most of the 18S sequence was stringently conserved. Variability was largely concentrated inside two hotspot regions centered at about 760 and 1480 nt after the start of the gene (regions I and II in Figure 5A). Region I included a *Simulium*-specific 7-nt deletion (Figure 5B). Several positions from the hot spot regions mutated repeatedly, while mutations located elsewhere generally showed up as sporadic occurrences. Using the 18S phylogeny to reconstruct the mutation network, average densities of 1.4 and 0.8 mutations/site were recorded for the hypervariable cores (Figure 5B) of regions I and II, respectively.

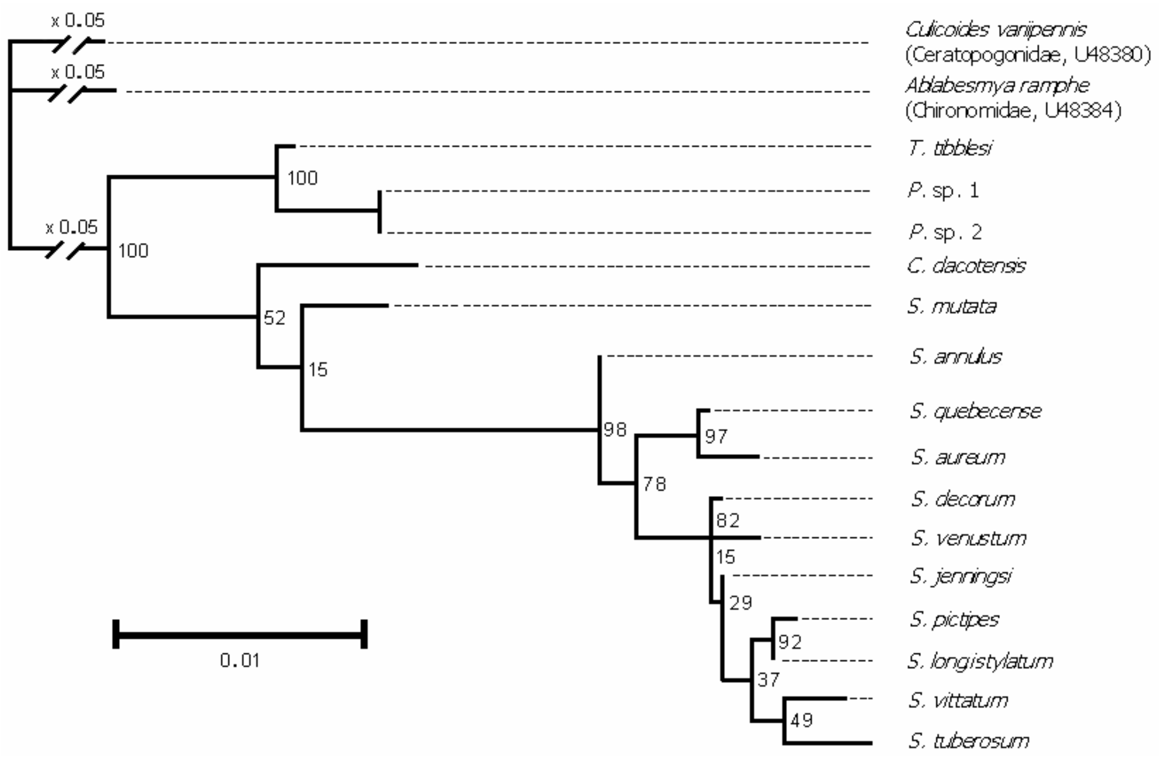

Figure 4. Rooted tree for rDNA (18S+5.8S regions). Bootstrap values (% over 500 replicates) are indicated. Scale bar: substitutions per site. Branch lengths are set proportional to mutational distance, except for outgroup sequences (20-fold size reduction).

Several diagnostic positions from the cores (and a few others scattered elsewhere) were also detected, in the sense that base identity or indels typified either Prosimuliini or Simuliini as a whole. As noticed before [10], similar correlations also occurred at positions 22 (Prosimuliini vs Simuliini: T vs A) and 172 (T to C substitution) of the 5.8S region.

A composite tree

Sequences from the 2D-aligned ITS2 [10] and the variable part of the 18S were concatenated to the COI/COII codons showing at least one amino acid substitution (Table 3), which represents 266, 104 and 65 variable positions. The selection of codons performed on COI/COII will hopefully eliminate much of the background noise due to sites undergoing purely synonymous mutations and will enhance relevant phylogenetic information. For both the 18S tree (Figure 4) and the one created from the concatenate (Figure 6), Prosimuliini (*Prosimulium* and *Twinnia*) clearly stand apart and *Cnephia*/*Stegopterna* diverges next. However, the composite exhibits a much better resolution of the terminal *Simulium* genus, which shows the same three major branches previously identified from ITS2 alone [10]. This observation was expected since ITS2 represents 61% of the variable positions from the concatenate and should weigh more heavily on the resolution of recent divergence events than the 24% contribution from the more conserved 18S. COI/COII could also add relevant information about terminal branches.

DISCUSSION

Simulated COI/COII evolution is implemented under a set of reasonable approximations and restrictions. We discarded the first and second codon positions from the analysis due to their negligible impact on the overall mutation pattern, which consists essentially of silent third position substitutions. Next, base composition varies little between species and the A + T bias appears as the only significant evolutionary constraint. These

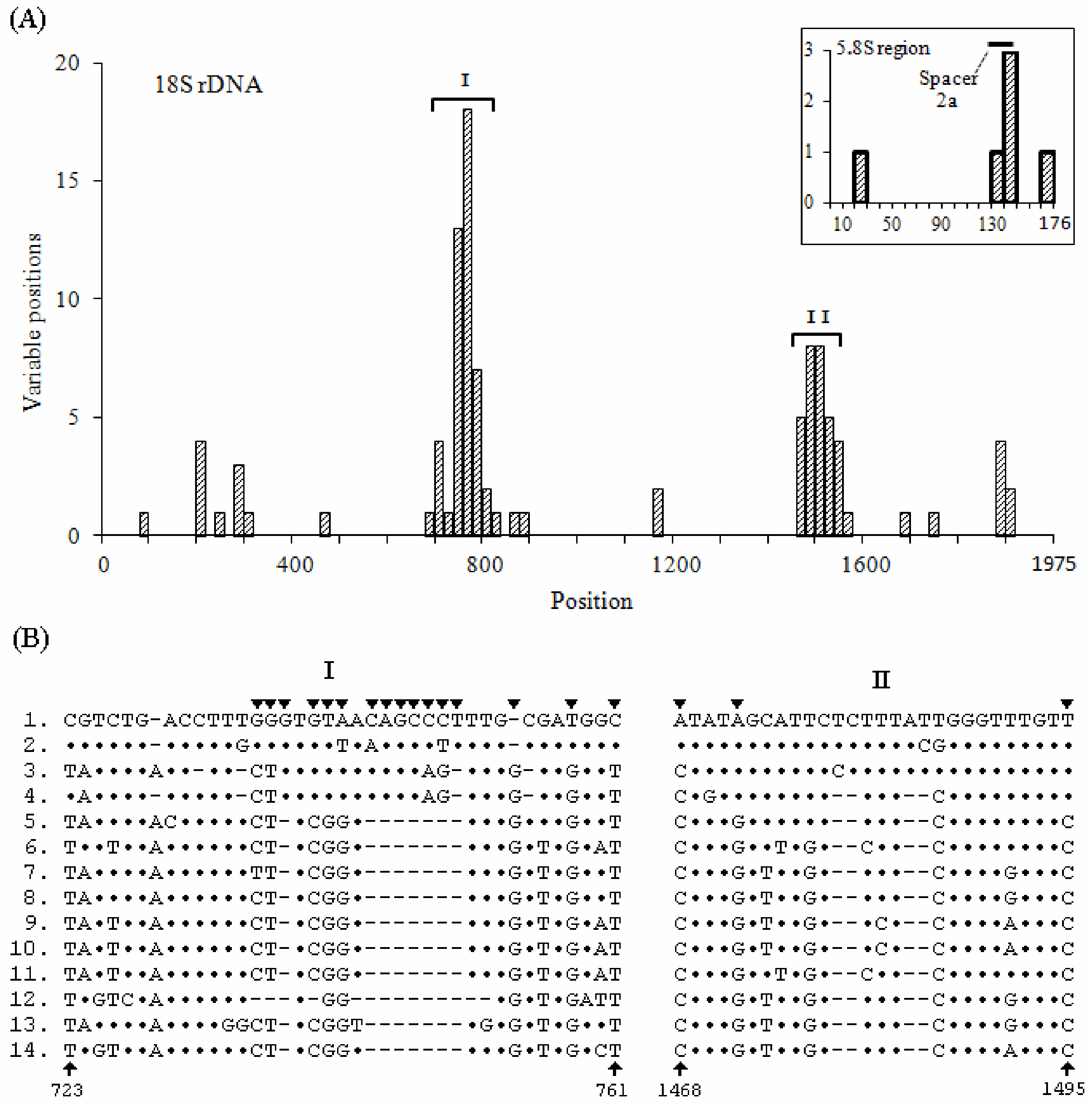

Figure 5. Mutation map for black fly rDNA sequences. (A) Frequency histogram (20-nt intervals) of the 101 variable positions of 18S rDNA. Position numbering includes both the 41-nt length presumably missing at the 5' end and the alignment-dictated gaps. Inset: the 5.8S region (10-nt intervals). (B) The hypervariable cores of hot spots I and II. Dot: identity with top sequence. Dash: deletion. Triangles: diagnostic positions for either Prosimuliini or *Simulium*. Species, from 1 to 14: *T. tibblesi*, *P.*sp.1/*P.*sp.2, *C. dacotensis*, *S. mutata*, *S. annulus*, *S. aureum*, *S. decorum*, *S. jenningsi*, *S. longistylatum*, *S. pictipes*, *S. quebecense*, *S. tuberosum*, *S. venustum*, and *S. vittatum*.

observations suggested mutation equilibrium, which led us to postulate, in order to simplify the evolution scheme, that each base interchange occurs at equal rates in both directions. Finally, relative mutation rates are computed by multiplying the frequencies of the two bases involved in a mutation under the rationale that bigger targets sustain more hits. With regard to the latter point and due to the low G + C content, a correction factor for T↔C and A↔G interchanges also allowed to modulate at will the Ts/Tv ratio. If we assume the general correctness

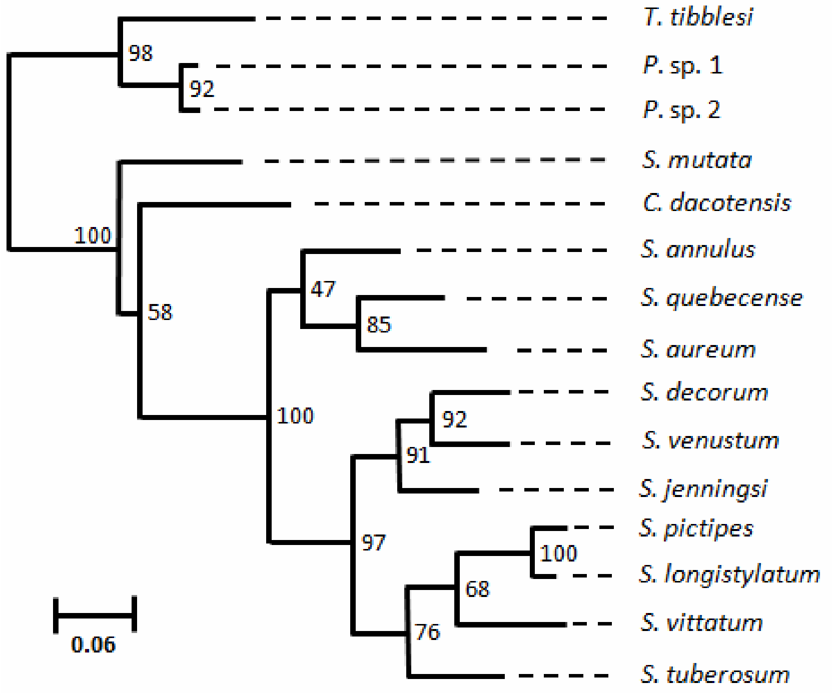

Figure 6. Composite tree for ITS2/18S/ COI+COII. Bootstrap values (% over 500 replicates) are indicated. Branch lengths are proportional to mutational distance. Scale bar: substitutions/site. For lack of adequate ITS2 outgroup sequence [10], the tree remains unrooted, but, for clarity purposes, is drawn to show the basal divergence between Simuliini and Prosimuliini.

of this model, our various simulations indicate that noise from the third codon position overwhelms the phylogenetic signal at about 17-20 My in the past. This time span covers at most 15-20% of the attested antiquity of Simuliidae [49, 50], such that even the relatively young and well documented *Simulium* genus cannot be adequately resolved through standard treeing procedures (the authors, unpublished). Even when adding the small contribution from the less mutable and presumably more informative first and second positions, the overall picture should not be substantially modified. This temporal horizon depends on the accuracy of the molecular clock rate. Since it relies upon on well documented geological events and several arthropod taxa, Brower's estimate [47] which we used here still appears the most credible. However, this uncorrected estimate does not consider multiple hits and parallel mutations. More recently, slightly higher or lower rates have been proposed for specific taxonomic groups and geological contexts [51].

By simply looking at the region close to the root tree, mutation saturation is often identified by the crowding of nodes and degraded bootstrap values. This behavior is quite apparent in our simulated trees and, although not noticed by the authors, also shows up in a study of 65 Nearctic black fly morphospecies [52]. The authors used a more densely populated network of species and genera than we did, but their neighbor-joining tree took a shallow appearance as soon as relatedness fell below subgenera. When considering that most of the base interchanges involve the third codon position, their average of 14.9% (all three positions, hence a rough estimate of 40% for the third one alone) for interspecific divergences indicates incoming mutation saturation with a concomitant loss of long-range phylogenetic inference. All these observations add significant weight to the common, yet empirical, practice in phylogenetic studies which should definitely aim at populations and recent speciation events except for the specific purpose of species barcoding [23]. Replacing DNA by COI/COII polypeptide sequences leads to the other extreme since their conservatism restricts applications to much higher taxonomic levels.

Concerning *S. pictipes* and *S. longistylatum*, these designations are no more considered valid since a reexamination of museum specimens led to the proposal of switching names from *S. longistylatum* to *S. pictipes* and equating the former *S. pictipes* with the already known *S. innoxium* Comstock and Comstock, 1895 [41], although morphological keys were not updated accordingly. The move is acknowledged in the last black fly inventory of Adler and Crosskey [24], but we still chose to be consistent with our previous publication [10] and to follow the accompanying full descriptions from Wood *et al.* [40] by keeping the former names. Unusually large larvae inhabiting swift current areas at the head of waterfalls characterize both *S. longistylatum* and *S. pictipes*. Although superficially quite similar, a few characters (*S. longistylatum*: darker antennae, rounder postgenal cleft and eyes spot region of lighter color) allow their morphological distinction. Our DNA data remove ambiguities with regard to the status of *S. pictipes* and *S. longistylatum* as distinct species and exclude the simple explanation of intraspecific variation between two populations from distant locations. First, COI/COII sequences diverge by about 5%, somewhat beyond the currently admitted threshold of about 3% for the operational definition of species [23, 53]. However, this figure neglects complexes of species which show maximum values as high as 6.5% in some black fly taxa [52]. Although this threshold issue remains moot, *S. longistylatum* and *S. pictipes* each exhibit a reduced genetic diversity while being neatly differentiated from the other, which indicates a clean separation of ancestral lines. The molecular clock points to an age of about 2.2 My for their last common ancestor. Convincing evidence from the nuclear genome about the existence of two distinct species is also provided by small differences in the highly conserved 18S and large, highly significant ones in ITS1 [10].

Prosimulium species were globally identified as members of the *P. hirtipes* Fries complex Rothfels, 1956, prior to 1956, after which they were distinguished by morphology, ecology and, later on, according to cytoforms that were often equated with true taxonomic units. In the *Prosimulium* genus, most of the useful morphological characters are embodied in the adult stage [41]. Snyder and Linton [54] even go to the point of suggesting that neither a single trait nor any given combination of morphological characters can reliably identify *Prosimulium* larvae at the species level. The COI/COII tree (Figure 3B) generated from our specimens, all collected at a unique location, lends support to this view as an initial classification into six morphospecies, checked separately by two of the authors, appears clearly at odds with the molecular phylogeny. Worse, it shows specimens from the *Helodon* Enderlein, 1921, and *Prosimulium* genera to intermingle. One could still rely on morphology while hypothesizing a high introgression rate of the maternally transmitted mtDNA, hence a common mtDNA pool for all six morphospecies. However, the 7.8% COI/COII divergence between *P*.sp.1 and *P*.sp.2, well above the threshold normally delineating species, would imply a gene flow transgressing reproductive isolation. This self-denying scenario for COI/COII also contradicts nuclear genome data since the actual 23 specimens split exactly along the same two major lines according to specific ITS1 features [10]. The erratic distribution of morphospecies along the molecular trees eliminates alternate explanations such as incomplete lineage sorting [55] and strongly suggests that *P*.sp.1 and *P*.sp.2 are distinct species as based on DNA evidence. Given the lack of caryotypes, we cannot yet equate *P*.sp.1 and *P*.sp.2 with known nominal species due to the inadequacies of morphology and the absence of close enough sequence matches from Genbank.

However, the reliability of caryotypes for defining true taxonomic units is far from absolute with regard to *P. mixtum* Syme and Davies, 1958, and *P. fuscum* Syme and Davies, 1958, with different caryotypes. These can hybridize with each other [56] even if morphological characters are attributed to slight differences in ecological requirements [54]. Molecular, morphological and cytological criteria can at times conflict with each other, as shown by the widespread *Prosimulium travisi* 1 Stone, 1952, and its sibling *P. travisi* 2, which inhabits only the Colorado mountain range and differs from the former by cytology and by a maximum COI divergence of 8.8% [52]. Surprisingly, the morphologically distinct *P. formosum* Shewell, 1959, and *P. froehni* Sommerman, 1958, cluster molecularly with *P. travisi* 1 and *P. travisi* 2, respectively, these relationships being tentatively interpreted as the result of mtDNA introgression. In another instance [57], genetic panmixia occurred in a contact zone of two siblings from the *Simulium arcticum* group Malloch, 1914. Furthermore, a tree based on four mitochondrial markers exhibits an inconsistent topology with respect to caryotype when six *S. arcticum* siblings are examined [58] and the authors put forward the view that the six cytospecies represent in fact six cytoforms of a unique species. To conclude, some black fly species complexes probably need a reappraisal by simultaneously including morphological, ecological, cytological and DNA (both nuclear and mitochondrial) data.

CONCLUSION

The conservatism of 18S rDNA allows the elucidation of phylogenies down to ordinal relationships in insects [59], but makes it inadequate to resolve close relationships, especially within the *Simulium* genus. Concatenation of several sequences (for instance, [60, 61]) appears as a solution to avoid this pitfall and improve bootstrap levels. This approach often neglects the idea that an ideal composite of DNA markers should comprise a right dosage of conserved and variable characters to resolve equally well all levels of a deep-ranging phylogenetic tree. Alternatively, we derived a composite tree made of the conserved 18S, the moderately variable 2D-aligned ITS2 [10] and only the COI/COII codons that undergo amino acid substitutions. Since there are few such codons (34/305), they should not create an excessive amount of overall noise at the root of the tree while enabling the resolution of recent divergence events. Here, both the composite tree and the one derived from ITS2 [10] from the same pool of species clearly show *Simulium* as the latest diverged branch within Simuliini and a better resolution of terminal clades. This includes a tight grouping of *S. jenningsi*, *S. venustum* and *S. decorum* and a neighboring cluster consisting of *S. vittatum*, *S. pictipes* and *S. longistylatum*. Also, S. *quebecense*, *S. aureum* and *S. annulus* all appear through either marker as early diverged *Simulium* lineages. Both trees also indicate a mutation rate slowdown in Prosimuliini. In particular, *P*.sp.1 and *P*.sp.2 18S sequences are 100% identical and exhibit only ten substitutions relative to *Twinnia*, although ITS1 data suggest that the two genera diverged about 30 My ago [10]. Finally, evidence from *C. dacotensis* and *S. mutata* is conflicting: when taken alone, ITS2 locates both species on a common early branch, while the composite tree shows *C. dacotensis* diverging last before the onset of the *Simulium* clade. Globally, bootstrap values are slightly higher and terminal branches somewhat longer for the composite tree. As concerning 18S alone, resolution is markedly improved, especially at low phylogenetic levels. In conclusion, sets of markers evolving on different time scales could generate trees encompassing multiple taxonomic levels at the same time.

ACKNOWLEDGEMENTS

We are grateful to the Université du Québec à Trois-Rivières who provided C. Gaudreau's scholarship and funding during her research.

CONFLICT OF INTEREST STATEMENT

All authors declare no conflict of interest.

REFERENCES

1. Folmer, O., Black, M., Hoeh, W., Lutz, R. and Vrijenhoek, R. 1994, Mol. Mar. Biol. Biotechnol., 3, 294-299.
2. Shepard, J. J., Andreadis, T. G. and Vossbrinck, C. R. 2006, J. Med. Entomol., 43, 443-454.
3. St-Onge, M., LaRue, B. and Charpentier, G. 2008, J. Invertebr. Pathol., 98, 299-306.
4. Crainey, J. L., Wilson, M. D. and Post, R. J. 2009, Med. Vet. Entomol., 23, 238-244.
5. Miller, B. R., Crabtree, M. B. and Savage, H. M. 1997, Insect Mol. Biol., 6, 105-114.
6. Moulton, J. K. 2000, Syst. Entomol., 25, 95-113.
7. Pace, N. R., Olsen, G. J. and Woese, C. R. 1986, Cell, 45, 325-326.

8. Schlötterer, C., Hauser, M. T., von Haeseler, A. and Tautz, D. 1994, Mol. Biol. Evol., 11, 513-522.

9. Coleman, A. W. 2003, Trends Genet., 19, 370-375.

10. LaRue, B., Gaudreau, C., Bagre, H. O. and Charpentier, G. 2009, Mol. Phylogenet. Evol., 53, 749-757.

11. Thanwisai, A., Kuvangkadilok, C. and Baimai, V. 2006, Genetica, 128, 177-204.

12. Rodríguez-Pérez, M. A., Núñez-González, C. A., Lizarazo-Ortega, C., Sánchez-Varela, A., Wooten, M. C. and Unnasch, T. R. 2006, J. Med. Entomol., 43, 701-706.

13. Gaudreau, C., LaRue, B., Charbonneau, V., Charpentier, G. and Craig, D. A. 2008, Invertebr. Syst., 22, 555-562.

14. Brockhouse, C. L., Vajime, C. G., Marin, R. and Tanguay, R. M. 1993, Biochem. Biophys. Res. Commun., 194, 628-634.

15. Krüger, A., Gelhaus, A. and Garms, R. 2000, Insect Mol. Biol., 9, 101-108.

16. Matsumoto, Y., Yanase, T., Tsuda, T. and Noda, H. 2009, J. Med. Entomol., 46, 1099-1108.

17. Gaudreau, C., LaRue, B. and Charpentier, G. 2010, Med. Vet. Entomol., 24, 214-217.

18. Caterino, M. S., Cho, S. and Sperling, F. A. 2000, Annu. Rev. Entomol., 45, 1-54.

19. Pramual, P., Wongpakam, K. and Adler, P. H. 2011, Genome, 54, 1-9.

20. Hernández-Triana, L. M., Crainey, J. L., Hall, A., Fatih, F., Mackenzie-Dodds, J., Shelley, A. J., Zhou, X., Post, R. J., Gregory, T. R. and Hebert, P. D. N. 2012, Zootaxa, 3514, 43-69.

21. Pramual, P. and Nanork, P. 2012, Entomol. Sci., 15, 202-213.

22. Conflitti, I. M., Pruess, K. P., Cywinska, A., Powers, T. O. and Currie, D. C. 2013, J. Med. Entomol., 50, 1250-1260.

23. Hebert, P. D. N., Cywinska, A., Ball, S. L. and deWaard, J. R. 2003, Proc. R. Soc. Lond., 270, 313-321.

24. Adler, P. H. and Crosskey, R. W. 2015, World blackflies (Diptera: Simuliidae): A comprehensive revision of the taxonomic and geographical inventory. https://biomia.sites.clemson.edu/pdfs/blackflyinventory.pdf (online, accessed 04/30/2015).

25. Currie, D. C. and Adler, P. H. 2008, Hydrobiologia, 595, 469-475.

26. Adler, P. H., Inci, A., Yildirim, A., Duzlu, O., McCreadie, J. W., Kúdela, M., Khazeni, A., Brúderová T., Seitz, G., Takaoka, H., Otsuka, Y., Takaoka, Y. and Bass, J. 2015, Biol. J. Linn. Soc., 114, 163-183.

27. Morales-Hojas, R., Post, R. J., Cheke, R. A. and Wilson, M. D. 2002, Med. Vet. Entomol., 16, 395-403.

28. Rothfels, K. H. 1979, Annu. Rev. Entomol., 24, 507-539.

29. Spironello, M., Hunter, F. F. and Craig, D. A. 2002, Can. J. Zool., 80, 1810-1816.

30. Day, J. C., Goodall, T. I. and Post, R. J. 2008, Med. Vet. Entomol., 22, 55-61.

31. Conceição, P. A., Crainey, J. L., Almeida, T. P., Shelley, A. J. and Luz, S. L. B. 2013, Acta Tropica, 127, 118-125.

32. Conflitti, I. M., Kratochvil, M. J., Sprironello, M., Shields, G. F. and Currie, D. C. 2010, Mol. Phyl. Evol., 57, 245-257.

33. Conflitti, I. M., Spironello, M. and Currie, D. C. 2012, Syst. Entomol., 37, 571-577.

34. Ekrem, T. and Willassen, E. 2004, Insect Syst. Evol., 35, 263-276.

35. Finn, D. S. and Adler, P. H. 2006, Freshw. Biol., 51, 2240-2251.

36. Yeh, W. B., Lee, H. M., Tu, W. C., Tang, L. C. and Lee, P. Y. 2009, J. Med. Entomol., 46, 249-256.

37. Joy, D. A. and Conn, J. E. 2001, Syst. Biol., 50, 18-38.

38. Joy, D. A., Craig, D. A. and Conn, J. E. 2007, Heredity, 99, 452-459.

39. Gaudreau, C. and Charpentier, G. 2011, Northeastern Naturalist, 18, 127-148.

40. Wood, D. M., Peterson, B. V., Davies, D. M. and Gyorkos, H. 1963, Proc. Entomol. Soc. Ont., 93, 99-129.

41. Adler, P. H., Currie, D. C. and Wood, D. M. 2004, The blackflies (Simuliidae) of North America. Cornell University Press, Ithaca, NY, USA.

42. Simon, C., Frati, F., Beckenbach, A., Crespi, B., Liu, H. and Flook, P. 1994, Ann. Entomol. Soc. Am., 87, 651-701.

43. Drummond, A. J., Ashton, B., Cheung, M., Heled, J., Kearse, M., Moir, R., Stones-Havas, S., Thierer, T. and Wilson, A. 2009, Geneious v4.7, available from http://www.geneious.com

44. Guindon, S. and Gascuel, O. 2003, Syst. Biol., 52, 696-704.
45. Posada, D. and Crandall, K. A. 1998, Bioinformatics, 14, 817-818.
46. Bos, D. H. and Posada, D. 2005, Immunol., 29, 211-227.
47. Brower, A. B. Z. 1994, Proc. Natl. Acad. Sci. USA, 91, 6491-6495.
48. Currie, D. C. 1988, Phylogeny of primitive Simuliidae (Insecta: Diptera: Culicimorpha). PhD dissertation, University of Alberta, Edmonton, Canada.
49. Bertone, M. A., Courtney, G. W. and Wiegman, B. M. 2008, Syst. Entomol., 33, 668-687.
50. Kalugina, N. S. 1991, Paleontol. J., 25, 66-77.
51. Papadopoulou, A., Anastasiou, I. and Vogler, A. P. 2010, Mol. Biol. Evol., 27, 1659-1672.
52. Rivera, J. and Currie, D. C. 2009, Mol. Ecol. Res., 9(Suppl. 1), 224-236.
53. Carew, M. E., Pettigrove, V., Cox, R. L. and Hoffmann, A. A. 2007, J. N. Am. Benthol. Soc., 26, 586-599.
54. Snyder, T. P. and Linton, M. C. 1983, Can. Entomol., 115, 81-87.
55. Funk, D. J. and Omland, K. E. 2003, Ann. Rev. Ecol. Evol. Syst., 3, 397-423.
56. Rothfels, K. H. and Freeman, D. M. 1977, Can. J. Zool., 55, 482-507.
57. Conflitti, I. M., Shields, G. F., Murphy, R. W. and Currie, D. C. 2015, J. Evol. Biol., 28, 1625-1640.
58. Conflitti, I. M., Shields, G. F., Murphy, R. W. and Currie, D. C. 2017, Syst. Entomol., 42, 489-508.
59. Kjer, K. M. 2004, Syst. Biol., 53, 506-514.
60. Cranston, P. S., Hardy, N. B., Morse, G. E., Puslednik, L. and McCluen, S. R. 2010, Syst. Entomol., 35, 636-648.
61. Ya'cob, Z., Takaoka, H., Low, V. L. and Sofian-Azirun, M. 2017, Acta Tropica, 167, 31-39.