(Mol. Phylog. Evol.)

The utility of indels in population genetics: the *Tpi* intron for host race genealogy of *Acrocercops transecta* (Insecta: Lepidoptera)

Issei Ohshima 1 & Kazunori Yoshizawa 2*

1 Division of Evolutionary Biology, National Institute for Basic Biology, Okazaki 444-8585, Japan

2 Systematic Entomology, Graduate School of Agriculture, Hokkaido University, Sapporo 060-8589, Japan. E-mail: psocid@res.agr.hokudai.ac.jp (*correspondence)

ABSTRACT

We investigated the utility of indel data for genealogical and population genetic analyses using the *Tpi* intron of the leaf mining moth *Acrocercops transecta* (Insecta: Lepidoptera). Genealogical analyses revealed that indel data were less homoplasious than DNA sequence data and that indel data contained a sufficient signal to provide a high resolution tree that was highly congruent with the tree estimated from DNA sequences. Although some conflicts were identified in the distributions of multi-residue indels, such conflicts were especially useful for the unambiguous detection of recombinations. For the first time, we adopted a Bayesian clustering method for indel characters to infer genetic structure of the moth. We concluded that indel characters have the potential to be a powerful tool in the analysis of population genetics and population structure as well as in the detection of gene flow.

Keywords: insertions/deletions, genealogy, population genetics, *Tpi* intron, host race

Introduction

Many genealogical analyses of closely related species or intra-species of animals have utilized sequence data from mitochondrial protein-coding genes (e.g., Ohshima and Yoshizawa 2006; Abellán et al., 2009; Valade et al., 2009; see Galtier et al., 2009 for review). Advantages of mitochondrial markers, compared to nuclear ones, include a faster substitution rate, little or no recombination, and faster fixation of mutation. In addition, mitochondrial protein-coding genes rarely contain length-mutational events (insertions and deletions: indels) because of their lack of introns, which also make sequence alignment easier. However, recent genealogical studies utilize rapidly evolving nuclear markers as well, especially sequences of non-coding nuclear regions or introns. Different from mitochondrial genes, intron markers frequently include gaps that are inferred to be caused by indel events. Generally, the presence of indels in sequence data is considered to be a nuisance because it makes sequence alignment difficult. If indel rich regions are unable to be aligned reliably, then the regions might need to be eliminated from the analysis, causing a reduction in the available data. However, like nucleotide substitution, length-mutational events are heritable, thus, they have the potential to be used as phylogenetic/genealogical markers.

The importance of indel mutation data has long been recognized (e.g., Lloyd and Calder 1991). To date, the methodology of indel coding for phylogenetic analyses is well established (e.g., Simmons and Ochoterena 2000; Müller 2006), and the performances of different coding methods have been compared based on simulation data (Ogden and Rosenberg 2007; Simmons et al., 2007). In particular, multi-residue indels are regarded as rare genomic changes (Rokas and Holland 2000) and are expected to experience few homoplasies (Simmons et al., 2001) because homoplasious multi-residue indels require several independent events to be matched (starting point, length and, for insertion, sequence: Keeling and Palmer, 2001). In contrast, indel data are also sensitive to hidden paralogy, lineage sorting, recombination, and lateral gene transfer (Keeling and Palmer 2001). Therefore, it has also been pointed out that indel markers are not unconditionally reliable markers of phylogenetic analyses (Keeling and Palmer 2001; Bapteste and Philippe 2002).

As previously mentioned, the presence of indels in sequence data is generally regarded as a nuisance for phylogenetic studies because multiple indel events make sequence alignment difficult and reduce the amount of data available for the analysis. Additionally, in phylogenetic studies, indel-rich regions are frequently excluded from analyses (Dopman et

al., 2005; Narita et al., 2006; Malausa et al., 2007; Kronforst 2008; Ohshima and Yoshizawa 2010). However, disadvantages of indel-rich regions for phylogenetic studies (i.e., estimation of species-tree) may not cause serious difficulty in genealogical analyses among closely related populations with a relatively small numbers of indel events (Beltrán et al., 2002). Moreover, multi-residue indels have been considered to be reliable markers to detect lateral gene transfer (Keeling and Palmer 2001). Homoplasies caused by introgression and recombination can even provide very useful information for population genetic studies, although such homoplasies are generally recognized as noise for the estimation of a species tree in which most components of an organism are assumed to share a common evolutionary history. Therefore, indel data could be a very powerful tool for genealogical analyses.

Introns of the *Tpi* gene are one of the most widely used nuclear gene markers for the genealogical analysis of closely related populations of butterflies and moths (Lepidoptera). Advantages of this marker include its rapid substitution rate, its location on the Z chromosome (females of lepidopterans have a single Z chromosome because of ZW or ZO sex-chromosome combinations in females, whereas males have two Z chromosomes), and faster fixation rate due to its smaller effective population size (3/4 compared to the expected fixation time for neutral autosomal genes). In contrast, introns of the *Tpi* gene are known to contain many indels and, in most cases, regions containing indels are excluded from the analyses (Dopman et al., 2005; Narita et al., 2006; Malausa et al., 2007; Kronforst 2008; Ohshima and Yoshizawa 2010). Only in a few cases are the indel data incorporated into the analyses, but these data are usually appended to DNA sequences without detailed examination of the signals contained in the indel data (Beltrán et al., 2002; Bull et al., 2006).

In this paper, we examine the utility of indel markers for genealogical and population genetic analyses using data scored from *Tpi* intron sequences of a leaf mining moth, *Acrocercops transecta*. Two host races, the *Juglans* and *Lyonia* races, are known for the species (Ohshima and Yoshizawa, 2006; Ohshima 2008), and our previous study based on sequence data detected significant gene flow between the two host races (Ohshima and Yoshizawa 2010). By using indel data, we conclude that indel data contain clear information to resolve genealogical history and population genetic structure, with a much lower level of homoplasy. In addition, we show that indel markers are especially useful in clearly detecting gene flow and recombination.

Materials and methods

Data preparation

A modified version of the *Tpi* data sets presented in Ohshima and Yoshizawa (2010) were used. In Ohshima and Yoshizawa (2010), data from different populations (Sendai and Niimi) were analyzed separately; however, both data sets were combined in a single data matrix here, and individuals with an identical haplotype were represented by one individual. Our previous analyses detected many multi-residue indels and possible recombinations within the data. Many alignment softwares which employ a progressive alignment method (e.g., Clustal, on which alignment of Ohshima and Yoshizawa [2010] was based) generate a guide tree based on all pairwise alignments between all the sequences. Then, the multiple alignment is performed by following the branching order in the tree. However, application of a single guide tree for entire sequences can be problematic when recombinations are evident because each single sequence can include multiple historical backgrounds. Especially, direct optimization method as implemented in POY is problematic to align such sequences because alignment is optimized along with one and the same tree topology (Simmons, 2004; Yoshizawa, 2010). For example, application of Clustal, MAFFT, and POY to the present data all failed to align the region around indel 50 (19-bp deletion: Fig. 2) correctly, although three sequences containing this deletion event exhibit exactly identical sequence for more than 30 bp around the indel region, probably because of heterogeneous signals contained in these sequences (Fig. 2; see also online supplement). In contrast, a local alignment method as implemented in the software program Dialign-TX (Morgenstern et al., 1998; Subramanian et al., 2008) does not use guide tree and constructs multiple alignments from local pairwise sequence similarities. Therefore, the approach is especially successful in finding local homologies (Simmons et al., 2008) and thus is expected to perform better to align sequences with recombinations. For example, only this software correctly aligned the region around the indel 50. One more possible advantage of this approach is that there is no necessity to use gap opening/extension cost parameters which also affect to the resulting alignments greatly but there is no objective criterion to select the optimal parameter setting (Pons and Vogler, 2006). Therefore, we used the alignment generated by Dialign-TX for the present analyses, with a few obvious mis-alignments corrected by eye using the similarity criterion (Simmons, 2004). Aligned data consisted of 605 aligned sites. Data matrices (both original Dialign-TX

alignment and the edited version) are available as online supplements from the journal's web site or at http://kazu.psocodea.org/data/indel.

Indel characters were coded using the simple indel coding method (SIC: Simmons and Ochoterena 2000) implemented in the software program SeqState (Müller 2005). Simulation tests by Simmons et al. (2007) showed that this coding method performed better than other indel coding methods (e.g., 5th state coding: Bena et al., 1998) or as well as the modified complex indel coding method (MCIC: Müller 2006). SIC scored a total of 52 indel characters for our dataset. Because quite a few conserved blocks were preserved in the alignments (Fig. 2), homology of most indel characters appeared to be unambiguous, although slight sifts of the starting and ending positions were plausible for several indels. In contrast, homology of a few single-residue indels might be questionable (Fig 2 and online supplement). However, all the indel characters scored by SIC method were accepted for the analyses, and genealogical values of such indels were evaluated on the basis of the resulting trees (Simmons et al., 2008). We also adopted MCIC, which is known to perform equally as well as SIC (Simmons et al., 2007). The analysis based on the MCIC matrix provided highly congruent trees with SIC trees. However, the matrix produced by the MCIC cannot be used for STRUCTURE analyses (see below); thus, results from the MCIC matrix are not presented here.

Based on aligned DNA data, the possibility of recombinations was detected using IMgc (Woerner et al., 2007). To our knowledge, IMgc is a only program available to produce the optimal recombination-filtered datasets from recombining input data. Two parameter settings, determining the most data-rich recombination-filtered block (default) and detecting maximum number of putative recombining sites (commands -w and -s, respectively), were applied for the aligned data. IMgc infers recombination sites on the basis of the four-gamete rule (Hudson and Kaplan 1985). However, this rule is valid only under infinite sites; therefore, the -s mode is likely to be too strict for the current case because the K81uf+G model, which allows for repeated changes at the same site, was selected for the aligned dataset by Modeltest 3.06 (Posada & Crandall, 1998: data not shown). Thus, we accepted the results from the default IMgc setting for the following analyses, while results from the -s mode were also used to supplement as the maximum estimation of the number of putative recombination sites. IMgc analysis with the default setting identified two possible recombinations within the aligned sequences. Next, the longest non-recombinant region identified by IMgc (5-340 of the aligned sequence data: indel characters 1-36) was subjected

to the following analyses. However, the remaining sequences (341-605) and indels 37-52 were also used as a supplementary analysis to examine incongruences of genealogical signals between these regions. Aligned DNA sequences 1-4 were not analyzed because of a lack of data and indels. IMgc with the -s command detected 8 putative recombining sites within the longest non-recombinant region obtained by the default setting, and 14 within the remaining region.

Genealogical analyses

The maximum parsimony criterion was adopted for both DNA and indel data sets using PAUP* 4b10. A heuristic search with 100 replicates of TBR was performed. Branches were collapsed if the maximal branch length was zero. All character changes were equally weighted. For the analysis of DNA data, gapped regions were included, and gaps were treated as missing data. Bootstrap supports for branches were calculated for the DNA trees using 100 replicates with a TBR branch swapping. Because of the many polytomies in the resulting trees (many haplotypes differentiated only by parsimony-uninformative characters and the presence of many gaps because of the inclusion of indel regions for analysis), Maxtree was set to 10,000 for each bootstrap replicate. The most parsimonious reconstruction of ancestral states for indel characters was performed using MacClade (Maddison & Maddison, 2001). A Spearman's rank correlation coefficient between accumulations of nucleotide substitutions and length-mutation events was calculated by R 2.10.1 (R Development Core Team 2009).

Inference of genetic structure based on indel data

The indel data were further analyzed by a Bayesian clustering method implemented in STRUCTURE 2.3.2 (Faluch et al., 2003; Pritchard et al., 2010) to infer genetic structures. This method infers the number of clusters of individuals (*K*) in a way that maximizes Hardy-Weinberg equilibrium and minimizes linkage disequilibrium within clusters without models assuming particular mutation processes (Pritchard et al., 2000). Applications of STRUCTURE against the current indel data violate a requirement of the method because it is not designed to deal with tightly linked markers. Therefore, the results from the analyses must be interpreted with some caution. However, eight putative recombinations were detected within the focal longest non-recombinant region (out of a total of 22 for the full length dataset), indicating that these indels could be regarded as not very tightly linked when

a substantial time scale is assumed. In addition, even if these markers are very tightly linked, this will only reduce the power for detecting population structures, and individuals from different populations will be admixed symmetrically (Falush et al, 2003). Therefore, if STRUCTURE analysis detects significant population structures from the indel data, this result can be interpreted as positive evidence, not as false-positive artifact. We estimated the probability from one to ten clusters (*K*), and all runs for each single *K* were replicated 20 times. We used the linkage model (Faluch et al., 2003) with a default setting and assumed that allele frequencies were correlated among populations. Physical distances (i.e., base pairs) between indel characters were used for inter-marker distances. Individual simulations were run for 100,000 steps following 20,000 burn-in steps. We calculated ΔK from the LnP(D) values to estimate the number of *K* (Evanno et al., 2005).

Results

Genealogical analyses

The maximum parsimony analysis of DNA data resulted in a huge number of equally parsimonious trees (length=201, CI=0.83 and RC=0.80), and the initial analysis was interrupted after 24 hours had elapsed (with 2,570,572 equally parsimonious trees saved). When Maxtree was set to 10,000 and 100 replicates of the TBR search each with different random starting tree were performed, the analysis also resulted in trees of length=201. The 50% majority consensus tree of the 2,570,572 trees corresponded to one of the best trees that is presented here as the most consistent (Fig. 1, left). Although not shown, Bayesian analysis of DNA data also yielded identical topology. The maximum parsimony analysis of indel data yielded 1,180 equally parsimonious trees (L=39, CI=0.92, RC=0.90), and Fig. 1 (right) shows one of them. The other equally parsimonious trees differed from it only by the presence of virtually zero-length branches caused by missing data in the matrix and the placement of clade III (in alternative topologies, indel characters 10 was considered to be non-homoplasious, but this interpretation is less plausible as discussed below). Trees estimated from DNA and indel data sets were highly congruent, and clades I-VIII, highlighted with gray in Fig. 1, were supported by all equally parsimonious trees estimated from both DNA and indel data sets. Three samples of the *Juglans* race (CJ45, 90 and NJ38: indicated by arrows in Fig. 1) were embedded within the *Lyonia* clade by both DNA and

indel data (Clade I). Indel data recovered a clade composed of seven samples (top end of Clade I: Fig. 1) supported by indel 30 (26-bp deletion) and a clade of CJ31+NJ40 supported by indel 1 (12-bp deletion). These clades were unresolved by DNA data. Homoplasy indices (ensemble consistency and retention indices: Kluge and Farris, 1969; Farris, 1989) calculated from these data showed that both DNA and indel data included a relatively low amount of homoplasies, with indel data having less homoplasy. Combined DNA+indel data yielded 2,576 equally parsimonious trees (tree not shown: L=242, CI=0.84, RC=0.81) that were congruent with DNA and indel trees. Spearman's rank correlation test of total amount of pairwise differences of indel data against DNA data showed that there was strong positive correlation between accumulations of nucleotide substitutions and length-mutation events (r=0.7102, P<0.0001).

Some conflicts of indel distribution (indicated by indel character numbers in Fig. 1 right) were evident within the non-recombinant region (indels 1-36). Homology of indel 10 provoked some question (a single nucleotide deletion overlapped with indels 8, 9 and 11: see online supplemental data for detail), but conflicts of indel distributions were evident even between unambiguously aligned multi-residue indels. Distributions of two overlapping indels (4 and 5) contradicted those of indels 15, 26, and 33 (Figs 1, 2). Although not shown here (available online), MP analysis of indels 37-52 resulted in 25 equally parsimonious trees which were completely incongruent to those estimated from indels 1-36 and DNA sequences 5-340 (Fig. 1). Except for parsimoniously uninformative indels (42, 43, 48, and 52) and three indels separating group VIII from the others (41, 44, and 45), the distribution of all indels within characters 37-52 conflicted with all parsimoniously informative characters from indels 1-36.

Genetic structure analysis of indel data

Results from a Bayesian clustering analysis of indel 1-36 were very similar to the findings from the maximum parsimony analysis (Fig. 3). Evanno's method revealed the highest ΔK value at K = 2, while the highest log probability (LnP[D]) was detected at K = 3. In both two and three K clusters, STRUCTURE clearly distinguished the *Lyonia*-race + gene flow samples (clade I) from the remaining *Juglans*-race samples in all replicates (Fig. 3A). By K = 3, an additional cluster was identified only in the *Juglans*-race samples. The application of K = 4-10 also revealed further clusters only in the *Juglans*-race samples (not

shown).

Investigating further subclustering for the *Juglans*-race samples recovered the highest Δ*K* and LnP(D) values at *K* = 3 (Fig. 3B). Members of clade II in Fig. 1 were distinguished from the others and were clustered as a very highly estimated membership, as shown in black (>.90, Fig. 1B). The remaining samples showed a mixed estimated membership of gray/white/black clusters. Samples with conflicting indel 4 demonstrated to have similar membership patterns with each other (indicated by broken lines in Fig. 3B), and one sample with conflicting indel 5 (CJ38) had a substantial membership with the black cluster (≈.10). We also analyzed the *Lyonia*-race + gene flow samples (clade I) with *K* = 1-10, but the highest Δ*K* and LnP(D) values were recovered at *K* = 1, indicating that no further clusters were identified in this group.

Analyses of indels 37-52 detected the highest Δ*K* and LnP(D) values at *K* = 1 for all-sample dataset, and no further clusters were detected in the *Juglans*-race samples or in the *Lyonia*-race + gene flow samples.

Discussion

Indel mutation data are generally omitted from phylogenetic and genealogical analyses (Dopman et al., 2005; Narita et al., 2006; Malausa et al., 2007; Kronforst, 2008; Ohshima and Yoshizawa 2010). Even in the few genealogical studies which incorporate indel data, indels are analyzed with DNA sequence data (e.g., Beltrán et al., 2002; Bull et al., 2006). Therefore, potential of indel data for genealogical studies has not been examined critically. The present analyses show that indel data can contain very useful genealogical information that is highly congruent with DNA sequence data. As shown in Fig. 1, trees estimated from DNA and indel data are highly concordant and well-resolved. Most importantly, three samples of the *Juglans* race (CJ45, 90 and NJ38) were unambiguously embedded within the *Lyonia* clade using indel data and DNA sequences (Clade I: indicated by arrows in Fig. 1). The Bayesian clustering analysis of indel characters also grouped these *Juglans* samples with the *Lyonia* cluster, indicating the indels have substantial population genetics signals (Fig. 3).

Our previous coalescent analyses based on DNA sequences clearly showed that there was significant gene flow from the *Lyonia* to *Juglans* races in this moth, and these three *Juglans* samples possess *Lyonia* type Z-linked genes (*Tpi*, *Period*, and/or *Ldh* genes: Ohshima and Yoshizawa 2010). In the study, indel character and indel sequences were

omitted from the analyses. Therefore, the indel data analyzed here provided further support for the introgression. In addition, DNA sequence data placed CJ45 at the most basal branch of the *Lyonia* clade (Fig. 1) and suggested the possibility that an ancestral polymorphism in the *Juglans* race could not be excluded (Ohshima and Yoshizawa 2010). However, possession of three indels unique to the *Lyonia* race by CJ45, coupled to its high membership with the *Lyonia*-race cluster (K=2, >.95; K=3, >.90) in the Bayesian clustering analysis (Fig. 3), cannot be explained by ancestral polymorphism. Thus, it is evident from the indel data that the placement of CJ45 in the *Lyonia* clade is due to gene flow from the *Lyonia* to *Juglans* races. These findings demonstrate that indel data are very valuable pieces of information to infer genealogical history and the genetic structure of closely related populations, which are even less homoplasious when compared to DNA sequence data. Furthermore, indel data recovered several clades that were not recovered by DNA sequences alone (CL54-NJ40 and CJ31+NJ40: Fig. 1). Each clade was supported by an unambiguously aligned multi-residue deletion, thus the indel characters supporting these clades are highly reliable. Indel data also contributed greatly to stabilize tree estimation by combining them with DNA data (i.e., numbers of equally parsimonious trees reduced from far more than 2.5 million to 2,567). It shows that indels have the potential to provide additional genealogical information that is difficult to extract from DNA sequences alone.

Further support for the usefulness of indel data was provided by the comparison of two data partitions (indels 1-36 vs. 37-52) to detect recombinations. When indels 37-52 were analyzed, a clade composed of CJ39, 43 and NJ40 was strongly supported by indels 47 (4-bp insertion) and 50 (19-bp deletion: Fig. 3). This strongly contradicts with clade VII (Fig. 1: CJ43 + NJ43), which is supported by indels 13 (6-bp deletion), 27 (6-bp deletion) and DNA sequences (5-340bp: 100% bootstrap support value). Furthermore, CJ107 (clade VI in Fig. 1) was placed at the sister of the *Lyonia* + gene flow clade (clade I) by indel 49 (9-bp insertion), although it was placed at the sister of CJ102 by indel 11 (3-bp deletion) and the DNA sequence data (100% bootstrap support value). Shared possession of indel 49 between the *Lyonia* and *Juglans* races may provide another example of gene flow between two races. However, indel 49 may actually represent a plesiomorphic condition because a similar sequence (indel 45) was also observed in group VIII (see online supplement). Even if indel 49 represents a plesiomorphic condition, shared plesiomorphic condition between clades VI and VIII contradicts the results from sequences 5-340 and indels 1-36. Results from the

Bayesian clustering analysis of two data partitions also contradicted. In either case, the distribution of all indels provided further support for the presence of recombination between them as identified by the default IMgc analysis.

Multi-residue indels are known to be rare genomic changes (Rokas and Holland, 2000) and, generally, distributions of such indels match exactly to phylogenetic history, especially for closely related species (Kawakita et al., 2003; Lavoué et al., 2003; Liu et al., 2009). In the present case, however, some conflicts between indel distribution are evident even within the possible non-recombinant region (sequences 5-340, indels 1-36: Fig. 1). Except for a single-site deletion for which homology is highly doubtful (indel 10 overlapping with indels 8, 9 and 11), all conflicting indel characters are multi-residue. For instance, indel 4 is a 8-bp deletion which contradicts indels 26 (a 12-bp insertion) and 33 (a 11-bp deletion: Figs 1, 2). There are no ambiguities in their homology, and their independent origins or secondary reversals are hardly plausible. In contrast, the occurrence of recombination can reasonably explain these conflicting indel distributions (Beltrán et al., 2002; Bull et al., 2006). Contradictions between indels 4/5 and indels 15, 26, and 33 (Fig. 2) can be explained by two recombinations occurring between indels 4/5 and 15. This explanation is more reasonable because occurrences of independent indel events at exactly identical positions and lengths at two different sites are less likely than recombinations (Keeling and Palmer 2001).

Sequence-based detection of recombinations is very sensitive to parameter settings. When the four-gamete rule was applied strictly to the sequence data, IMgc detected a total of 22 recombinations within the aligned data, compared to two recombinations identified under the default setting. Contradictions between indels 4/5 and indels 15, 26, and 33 can be eliminated by accepting a possible recombination at the aligned DNA sites 39 and 40 identified by IMgc under the strict four-gamete-rule setting (Fig. 2). Acceptance of this recombination does not contradict the distributions of other indels. In addition, the Bayesian clustering method revealed that samples with the conflicting indel 4 show similar membership patterns (Fig. 3), suggesting a sharing of gene fragments between the samples due to recombination. Thus, incongruence of indel distributions can help clearly identify the occurrence of recombinations.

Several sequence based identification methods of recombination are available (including distance- [RAT by Etherington *et al*., 2005], phylogenetic- [SimPlot by Lole *et al*., 1999] and substitution-based methods [GeneConv by Sawyer 1989]) but none of them can

detect all recombinations correctly and, as mentioned above, such methods are very sensitive to parameter settings (Posada and Crandall 2001; Posada *et al*., 2002). In contrast, by using indel character, the occurrence of recombination can be identified clearly if the incongruence of indel distribution is evident. Furthermore, indel-based parsimony analysis can aid to identify not only the recombinant position within the sequence but also the occurrence of gene flow among populations. For instance, possession of indel 5 in CJ38 and clade II suggests the occurrence of recombination between these populations (Fig. 3B). Using the strict four-gamete rule setting, IMgc also identified two recombinations between indels 33 and 34, and one between indels 34 and 35. However, there are no conflicts among these indels, suggesting that these recombinations are false-positive due to an overly strict application of the four-gamete-rule. Thus, indel information would be useful to distinguish important (probable) recombinations from non-important (false-positive) recombinations.

For the first time, a Bayesian clustering method as implemented in the software program STRUCTURE is adopted for indel data. However, the method is not designed for using tightly linked markers so the results should be interpreted with some caution. Tight linkage reduces the power for detecting population structures due to the stable linkage disequilibrium (background LD, Falush et al. [2003]). However, the present analysis recovered clear subpopulation structures in the *Juglans* race, indicating that admixture events have been occurred even among the tightly linked markers. Also, Falush et al. (2003) showed that when background LD is a problem, STRUCTURE infers that all individuals from both populations are admixed symmetrically. In contrast, our present result clearly shows asymmetrical admixture events (i.e. gene flow from the *Lyonia* race to the *Juglans* race). Thus, background LD due to using tightly linked markers should not hamper the current conclusion.

In summary, indels have the potential to be a powerful tool in genealogical and population genetics analyses, including the estimation of genealogical history, detection of gene flow, detection of recombinations, and the analysis of population structure. Specifically, less homoplasious indel characters aid in demonstrating gene flow and recombination events more clearly than sequence data. Therefore, the disadvantages of indel character for estimating a species tree could actually turn out to be advantages for genealogical analyses. However, indels should also be used with some caution. As shown by Spearman's test, indel events accumulate according to substitution events. Usually, the accumulation of

substitutions does not cause serious problem for sequence alignment, but the accumulation of many indel events will make sequence alignments extremely difficult and homology assessments of indel characters obscure (Beltrán et al., 2002). Therefore, indel character of indel rich regions, such as the *Tpi* intron, should only be used when a reliable alignment can be produced. If a primary homology statement among sequences cannot be justified, such regions should be excluded from the analyses (Kjer et al., 2009; Morgan and Kelchner, 2010; Yoshizawa, 2010).

Acknowledgments

We thank T. Nakamura for IMgc installing; M. Simmons, A. Vogler, and an anonymous reviewer for their valuable comments; and J. Pritchard for kind suggestion on the appropriate use of the software program STRUCTURE. Supported by JSPS 20-5555 to IO and JSPS 18770058 to KY.

References

Abellán P., Millán A., Ribera I., 2009. Parallel habitat-driven differences in the phylogeographical structure of two independent lineages of Mediterranean saline water beetles. Mol. Ecol. 18, 3885-3902.

Bapteste E., Philippe H., 2002. The potential value of indels as phylogenetic markers: Position of Trichomonads as a case study. Mol. Biol. Evol., 19, 972-977.

Bena, G., Prosper, J.-M., Lejeune, B., Olivieri, I., 1998. Evolution of annual species of the genus *Medicago*: a molecular phylogenetic approach. Mol. Phylogenet. Evol. 9, 552-559.

Beltrán M., Jiggins C.D., Bull V et al., 2002. Phylogenetic discordance at the species boundary: comparative gene genealogies among rapidly radiating *Heliconius* butterflies. Mol. Biol. Evol., 19, 2176–2190.

Bull V., Beltrán M., Jiggins C.D., McMillan W.O., Bermingham E., Mallet J., 2006. Polyphyly and gene flow between non-sibling *Heliconius* species. BMC Biol. 4, 11.

Dopman E.B., Pérez L., Bogdanowicz S.M., Harrison R.G., 2005. Consequences of reproductive barriers for genealogical discordance in the European corn borer. Proc. Natl. Acad. Sci. USA 102, 14706-14711.

Etherington G.J., Dicks J., Roberts I.N., 2005. Recombinaiton analysis tool (RAT): a

program for the high-throughput detection of recombination. Bioinformatics 21, 278-281.

Evanno G., Regnaut S., Goudet J., 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol. Ecol. 14, 2611-2620.

Farris J.S., 1989. The retention index and the rescaled consistency index. Cladistics 5, 417-419.

Falush D., Stephens M., Pritchard J.K., 2003. Inference of population structure using multilocus genotype data: linked loci and correlated alleles frequencies. Genetics 164, 1567-1587.

Galtier N., Nabholz B., Glémin S., Hurst G.D.D., 2009. Mitochondrial DNA as a marker of molecular diversity: reappraisal. Mol. Ecol. 18, 4541-4550.

Hudson R.R., Kaplan N.L., 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111, 147-164.

Kawakita A., Sota T., Ascher J.S., Ito M., Tanaka H., Kato M., 2003. Evolutionary and phylogenetic utility of alignment gaps within intron sequences of three nuclear genes in bumble bees (*Bombus*). Mol. Biol. Evol., 20, 87-92.

Keeling P.J., Palmer J.D., 2001. Lateral transfer at the gene and subgenic levels in the evolution of eukaryotic enolase. Proc. Natl. Acad. Sci. USA 98, 10745–10750.

Kjer K.M., Roshan U., Gillespie J.J., 2009. Structural and evolutionary consideration for multiple sequence alignment of RNA, and the challenges for algorithms that ignore them. In: Rosenberg, M.S. (Ed), Sequence alignment: methods, concepts, and strategies. University of California Press, Barkeley, CA, pp. 105-149.

Kluge A.G., Farris J.S., 1969. Quantitative phyletics and the evolution of Anurans. Syst. Zool. 18, 1-32.

Kronforst M.R., 2008. Gene flow persists millions of years after speciation in *Heliconius* butterflies. BMC Evol. Biol. 8, 98.

Lavoué S., Sullivan J.P., Hopkins C.D., 2003. Phylogenetic utility of the first two introns of the S7 ribosomal protein genes in African electric fishes (Mormyroidea: Teleostei) and congruence sith other molecular markers. Biol. J. Linn. Soc. 78, 273-292.

Lloyd D.G., Calder V.L., 1991. Multi-residue gaps, a class of molecular characters with exceptional reliability for phylogenetic analyses. J. Evol. Biol. 4, 9–21.

Liu K., Raghavan S., Nelesen S., Linder C.R., Warnow T., 2009. Rapid and accurate

large-scale coestimation of sequence alignments and phylogenetic trees. Science 324, 1561-1564.

Lole K.S., Bollinger R.C., Paranjape R.S., Gadkari D., Kulkarni S.S., Novak N.G., Ingersoll R., Sheppard H.W., Ray S.C., 1999. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J. Virol. 73, 152-160.

Maddison D.R., Maddison W.P., 2001. MacClade 4. Sinauer Assoc., Sunderland, MA.

Morgan M.J., Kelchner S.A., 2010. Inference of molecular homology and sequence alignment by direct optimization. Mol. Phylogenet. Evol. 56, 305-311.

Morgenstern B., Frech K., Dress A., Werner T., 1998. DIALIGN: finding local similarities by multiple sequence alignment. Bioinformatics 14, 290-294.

Marausa T., Leniaud L., Martin J-F., Audiot P., Bourguet D., Ponsard S., Lee S-F., Harrison RG., Dopman E., 2007. Molecular differentiation at nuclear loci in French host races of the European corn borer (*Ostrinia nubilalis*). Genetics 176, 2343-2355.

Müller K., 2005. SeqState: primer design and sequence statistics for phylogenetic DNA data sets. Appl. Bioinform. 4, 65–69.

Müller K., 2006. Incorporating information from length-mutational events into phylogenetic analysis. Mol. Phylogenet. Evol. 38, 667–676.

Narita S., Nomura M., Kato Y., Fukatsu T., 2006. Genetic structure of sibling butterfly species affected by Wolbachia infection sweep: evolutionary and biogeographical implications. Mol. Ecol. 15, 1095–1108.

Ogden T.H., Rosenberg M.S., 2007. How should gaps be treated in parsimony? A comparison of approaches using simulation. Mol. Phylogenet. Evol. 42, 817–826.

Ohshima I., 2008. Host race formation in the leaf-mining moth *Acrocercops transecta* (Lepidoptera: Gracillariidae). Biol. J. Linn. Soc. 93, 135-145.

Ohshima I., Yoshizawa K., 2006. Multiple host shifts between distantly related plants, Juglandaceae and Ericaceae, in the leaf-mining moth *Acrocercops leucophaea* complex (Lepidoptera: Gracillariidae). Mol. Phylogenet. Evol. 38, 231-240.

Ohshima I., Yoshizawa K., 2010. Differential introgression causes genealogical discordance in host races of Acrocercops transecta (Insecta: Lepidoptera). Mol. Ecol., 19, 2106-2119.

Pons J, Voglar A.P., 2006. Size, frequency, and phylogenetic signal of multiple-residue indels

in sequence alignment of introns. Cladistics 22, 144-156.

Posada D., Crandall K.A., 1998. Modeltest: testing the model of DNA substitution. Bioinformatics 14, 817-818.

Posada D., Crandall K.A., 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc. Natl. Acad. Sci. USA 98, 13757-13762.

Posada D., Crandall K.A., Holmes EC., 2002. Recombination in evolutionary genomics. Ann. Rev. Genet. 36, 75-97.

Pritchard J.K., Stephens M., Donnelly P., 2000. Inference of population structure using multilocus genotype data. Genetics155, 945-959.

Pritchard J.K., We X., Falush D., 2010. Documentation for structure software: version 2.3. Available from http://pritch.bsd.uchicago.edu/structure_software/release_versions/v2.3.3/html/structure.html.

R Development Core Team., 2009. A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

Rokas A., Holland PWH., 2000. Rare genomic changes as a tool for phylogenetics. Trends in Ecol Evol. 15, 454-459.

Sawyer S., 1989. Statistical tests for detecting gene conversion. Mol Biol Evol. 6, 526-538.

Simmons M. P., 2004. Independence of alignment and tree search. Mol. Phylogenet. Evol. 31, 874-879.

Simmons M.P., Ochoterena H., 2000. Gaps as characters in sequence-based phylogenetic analyses. Syst. Biol. 49, 369–381.

Simmons M.P., Ochoterena H, Carr, T.G., 2001. Incorporation, relative homoplasy, and effect of gap characters in sequence-based phylogenetic analyses. Syst. Biol. 50, 454-462.

Simmons M.P., Müller K., Norton A.P., 2007. The relative performance of indel-coding methods in simulations. Mol. Phylogenet. Evol. 44, 724-740.

Simmons M.P., Richardson D., Reddy S.N., 2008. Incorporation of gap characters and lineage-specific regions into phylogenetic analyses of gene families from divergent clades: and example from the kinesin superfamily across eukaryotes. Cladistics 24, 372-384.

Subramanian A.R., Kaufmann M., Morgenstern B., 2008. DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. Algor. Mol.

Biol. 3, 6.

Valade R., Kenis M., Hernandez-Lopez A., Augustin S., Mari Mena N., Magnoux E., Rougerie R., Lakatos F., Roques A., Lopez-Vaamonde C., 2009. Mitochondrial and microsatellite DNA markers reveal a Balkan origin for the highly invasive horse-chestnut leaf miner *Cameraria ohridella* (Lepidoptera, Gracillariidae). Mol. Ecol. 18, 3458-3470.

Woerner A.E., Cox M.P., Hammer M.F., 2007. Recombination-filtered genomic datasets by information maximization. Bioinformatics 23, 1851-1853.

Yoshizawa K., 2010. Direct optimization overly optimizes data. Syst. Entomol. 35, 199-206.

Figure captions

Fig. 1. Maximum parsimony trees estimated from DNA (left) and indel (right) data. Branch lengths are proportional to the number of substitution/indel events present on the branch. Numbers associated with the DNA tree indicate bootstrap values higher than 70%. Bars associated with the indel tree indicate indel characters presented on the branches. Widths of the bars correspond to lengths of the indel base pairs. Gray bars accompanied by indel numbers indicate the presence of conflicts in the distribution of indel characters. Clades consistently recovered by both data sets are highlighted by a shadow. The clade supported by character 26 of the indel data is in agreement between trees presented here, but this clade is not supported by some equally parsimonious trees estimated from indel data (see text for detail).

Fig. 2. (Top) MP tree estimated from indel data (left) and corresponding data matrix in bird eye's view (right). In the matrix, different nucleotides are indicated by different darknesses, and white regions indicate gaps. Conflicting indel data are surrounded by dotted line with indel character number. Recombination sites identified by default IMgc setting are indicated by black arrow heads with site positions below the matrix, and an additional recombination site identified by IMgc under the strict four-gamete-rule setting and suggested by indel distribution conflicts is indicated by a white arrow head with site positions.
(Bottom) Conflicts between indel data. Sequences followed by Ans. indicate ancestral conditions, and sequences followed by numbers indicate indel data corresponding to that in the data matrix and on the parsimonious tree (Fig. 1). Gaps are indicated by dash. Double headed arrows indicate conflicts between indels and a broken line indicates a possible recombination event.

Fig. 3. Results of the Bayesian clustering analysis using STRUCTURE for indel dataset 1-36. Each vertical bar corresponds to one sample and the proportion of each of the colors indicates the posterior probability of membership to their respective color ancestries. A) Dataset including all samples. B) Dataset except for clade I. The tree is from Fig.1 and the order of samples in the clustering results is identical to those of the tree. The gray band corresponds to clade II - VIII in Fig. 1. Samples with conflicting indel 4 are indicated by horizontal broken lines.

Figure1

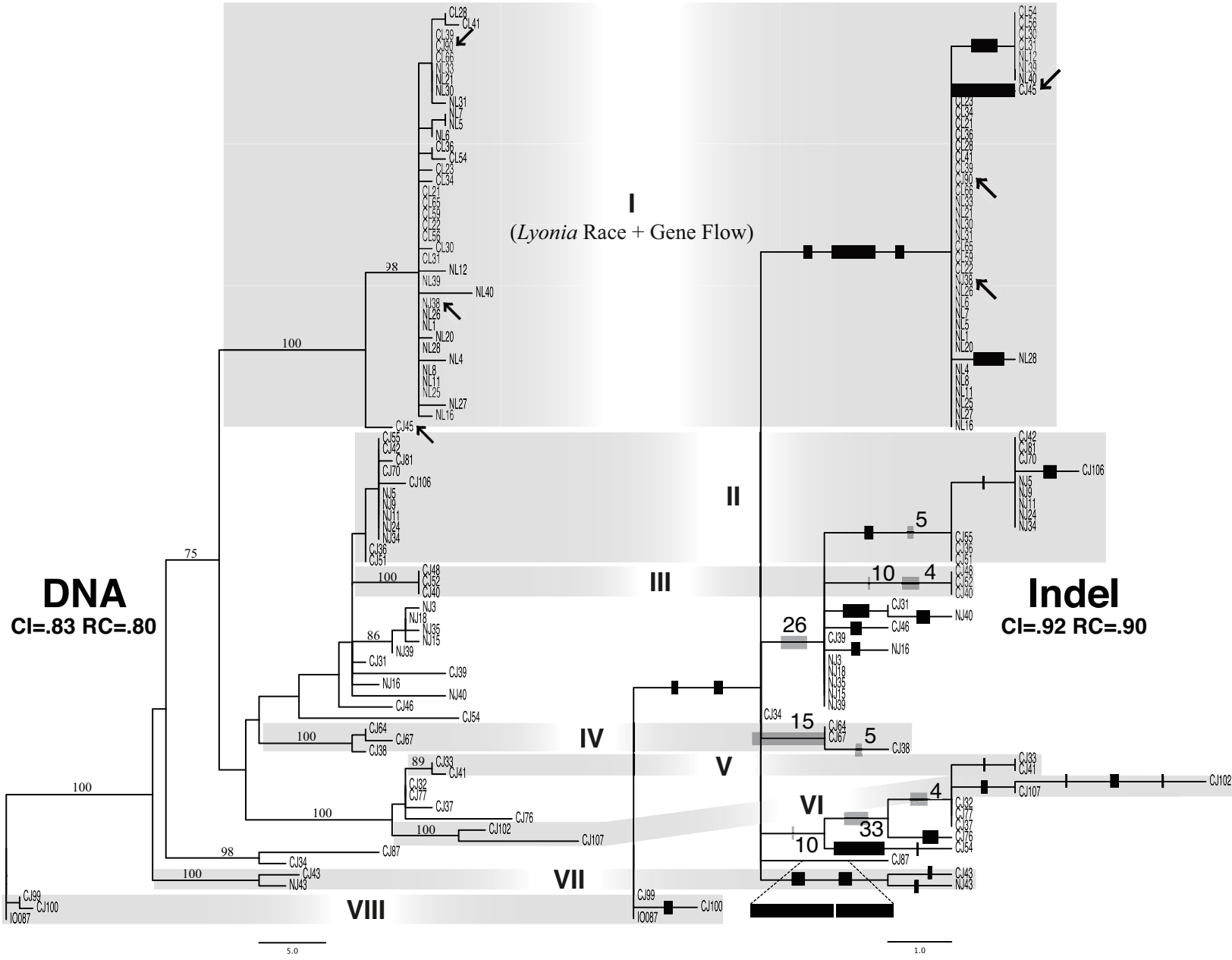
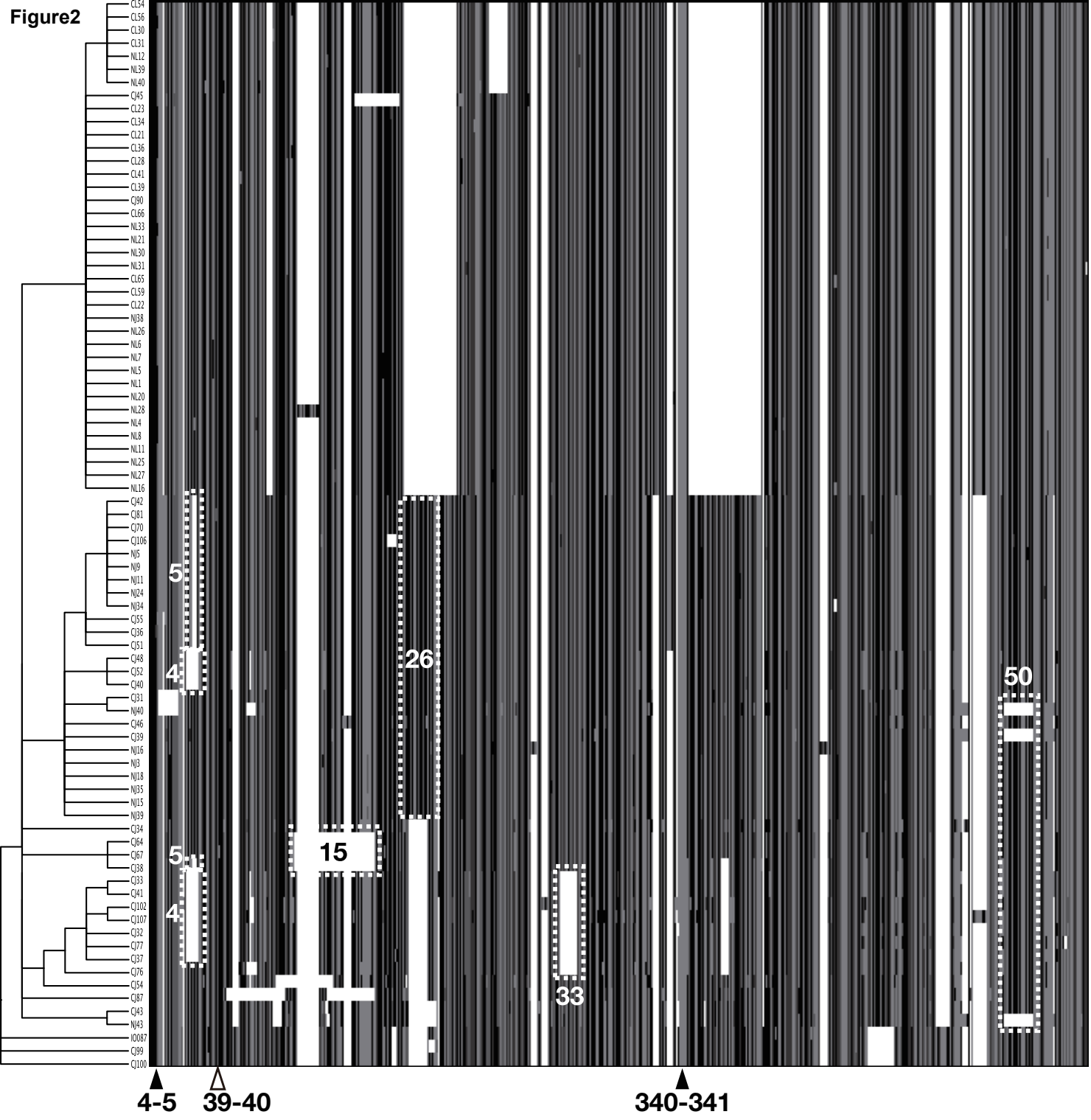

I
(*Lyonia* Race + Gene Flow)
II
III
IV
V
VI
VII
VIII
DNA
CI=.83 RC=.80
Indel
CI=.92 RC=.90
5.0
1.0

Figure2



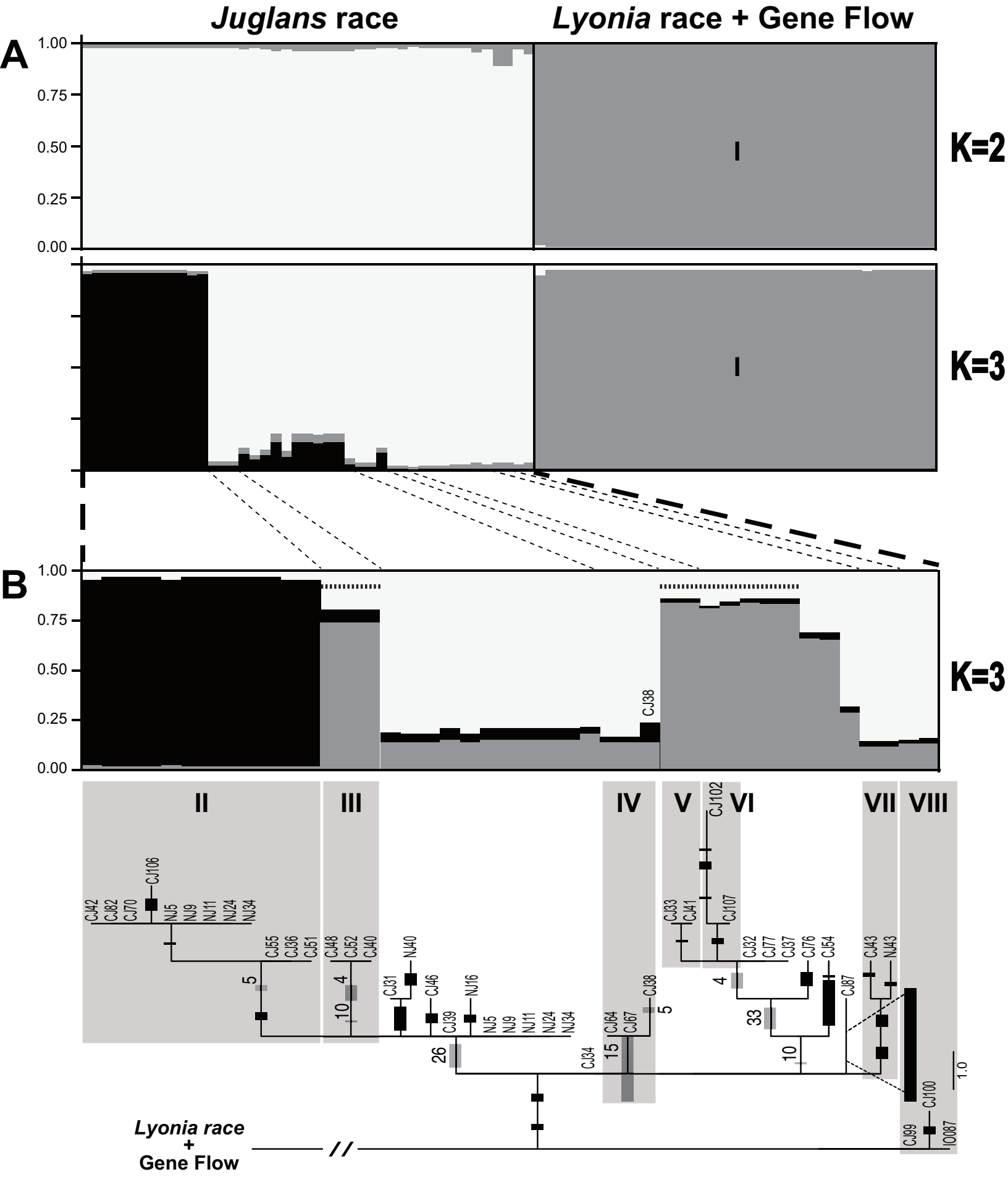
CTTTGTATACACATTGAA: Anc.
CTTTG--------TTGAA: 4
CTTTGTATT---ATTGAA: 5

GATTAATGCTGTAAAACTACTTTAAAATATTTTTTTTATTGAAA: Anc.
GATTT----------------------------------TATTC: 15

ATTCC------------TACAT: Anc.
ACTAAACATTACAACAATACAT: 26

CCAATGCGCTACCGTAGTCGT: Anc.
CGTGT-----------GTCGT: 33

Figure3


Juglans race
Lyonia race + Gene Flow
A
1.00
0.75
0.50
0.25
0.00
I
K=2
I
K=3
B
1.00
0.75
0.50
0.25
0.00
CJ38
K=3
II
III
IV
V
VI
VII
VIII
CJ106
CJ42
CJ82
CJ70
NJ5
NJ9
NJ11
NJ24
NJ34
CJ55
CJ36
CJ51
5
CJ48
CJ52
CJ40
4
10
CJ31
NJ40
CJ46
CJ39
NJ16
NJ5
NJ9
NJ11
NJ24
NJ34
26
CJ34
CJ64
CJ67
15
CJ38
5
CJ102
CJ33
CJ41
CJ107
4
CJ32
CJ77
CJ37
CJ76
CJ54
33
10
CJ87
CJ43
NJ43
CJ100
CJ99
IO087
1.0
Lyonia race
+
Gene Flow
//