Potential gene flow in natural populations of the *Drosophila ananassae* species cluster inferred from a nuclear mitochondrial pseudogene

Kyoichi Sawamura[a,*], Kae Koganebuchi[b,#], Hajime Sato[c], Koichi Kamiya[d], Muneo Matsuda[c], Yuzuru Oguma[a]

[a]Graduate School of Life and Environmental Sciences, University of Tsukuba, Tsukuba, Ibaraki 305-8572, Japan, [b]College of Biological Sciences, University of Tsukuba, Tsukuba, Ibaraki 305-8572, Japan, [c]School of Medicine, Kyorin University, Mitaka, Tokyo 181-8611, Japan, [d]National Institute of Education, Nanyang Technological University, 637616, Singapore

*Corresponding author. Graduate School of Life and Environmental Sciences, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8572, Japan. Tel: +81 29 853 4669, Fax: +81 29 853 6614, *E-mail address*: sawamura@biol.tsukuba.ac.jp

[#]Present address. Graduate School of Frontier Sciences, University of Tokyo, Kashiwa, Chiba 277-8583, Japan

Abstract A pseudogene with 94% similarity to mitochondrial cytochrome c oxidase subunit I (*COI*) was identified and localized to chromosome 4 of *Drosophila ananassae*. Because this chromosome is believed to have reduced recombination, its history can be traced using the pseudo-*COI* sequence. Pseudo-*COI* sequences were obtained from 27 iso-female lines of six taxa belonging to the *D. ananassae* species cluster in which reproductive isolation is incomplete. The phylogenetic network constructed from seven recognized haplotypes (#0 – #6) indicated that different taxa inhabiting the same geographic area share the haplotypes: #1 from Papua New Guinean populations of *D. ananassae* and pallidosa-like-Wau; #2 from Papua New Guinean populations of *D. ananassae*, pallidosa-like, and papuensis-like; and #4 from South Pacific populations of *D. ananassae* and *D. pallidosa*. Taxon-K has a unique haplotype (#6), and 18 mutation steps separate it from the closest haplotype, #2. We discuss the possibility of chromosome 4 introgression beyond taxon boundaries.

Keywords: *Drosophila ananassae*, introgression, pseudogene, phylogeography

1. Introduction

The *Drosophila ananassae* species complex is suitable for molecular phylogenetic and evolutionary studies because the entire genome of a strain of *D. ananassae* was recently sequenced (*Drosophila* 12 Genomes Consortium, 2007). In fact, the population structure of *D. ananassae* and its sibling species has been analyzed using several nuclear/mitochondrial genes and microsatellites (Chen et al., 2000; Vogl et al., 2003; Baines et al., 2004; Das et al., 2004; Schug et al., 2004, 2007, 2008; C.S. Ng and A. Kopp, personal communication). The *D. ananassae* species complex is one of three species complexes recognized in the ananassae group, recently elevated from a subgroup of the melanogaster group (Kaneshiro and Wheeler, 1970; Lemeunier et al., 1986; Da Lage et al., 2007). Eleven species have been described in the *D. ananassae* species complex (Tobari, 1993), but two of them were recently removed from this complex (Da Lage et al., 2007); the complex now includes *D. ananassae*, *D. atripex*, *D. cornixa*, *D. ironensis*, *D. monieri*, *D. nesoetes*, *D. ochrogaster*, *D. pallidosa*, and *D. phaeopleura*.

Within the *D. ananassae* species complex, *D. ananassae* and *D. pallidosa* are very closely related. *D. ananassae* is widely distributed in the tropics and subtropics, presumably expanding its distribution with human migration (Tobari, 1993). Although

its place of origin is unclear, the center of *D. ananassae* diversity is in Southeast Asia (Vogl et al., 2003; Baines et al., 2004; Das et al., 2004; Schug et al., 2004, 2007, 2008). This species has two color morphs; one is the light cosmopolitan form, and the other is the dark form seen in the populations sympatric with *D. pallidosa* (Futch, 1966, 1973; Tobari, 1993; Tomimura et al., 1993). *D. pallidosa* is endemic to Samoa, Tonga, Fiji, and New Caledonia (Futch, 1966; Stone et al., 1966; Tobari, 1993), and was described as a new species mainly based on the reduced number of rows in the sex-comb and its lighter body coloration than *D. ananassae* (Bock and Wheeler, 1972).

D. ananassae and *D. pallidosa* may have several cryptic species, which together we refer to as the '*D. ananassae* species cluster'. The cryptic species are different in the chromosome configuration (i.e., carrying at most eleven specific inversions) and courtship songs (Futch, 1966; Tomimura et al., 1993; Yamada et al., 2002; H. Yamada, T. Sakai, and Y. Oguma, unpublished; see also Schug et al., 2008). Despite their morphological similarity, they can be distinguished by reproductive isolation (Doi, 1997; Y. N. Tobari and M. Matsuda, unpublished). The degree of sexual isolation varies depending on the pair of populations and the direction of the cross, and some hybrids are sterile.

Among the potential cryptic species of the *D. ananassae* species cluster, one

inhabiting Papua New Guinea, called ‘pallidosa-like’, is phenotypically similar to *D. pallidosa* and shares with it all chromosome inversions except one or two (Tobari, 1993; Tomimura et al., 1993). Because pallidosa-like populations exhibit phenotypic variation and are sometimes intermediate between *D. ananassae* and *D. pallidosa*, these populations may be of hybrid origin or represent a hybrid swarm. Some of the pallidosa-like flies first collected in Wau, Papua New Guinea, are apparently reproductively isolated from the others, and this population is called ‘pallidosa-like-Wau’. Futch (1966) recognized another cryptic species in Papua New Guinea and informally called it ‘papuensis’. Similar flies (‘papuensis-like’) were collected in later expeditions, but the chromosome configuration is slightly different (Tomimura et al., 1993). This may have been caused by recent introgression from *D. ananassae*. Papuensis-like also has been recorded in Cairns, Australia (Tobari, 1993). Another cryptic species, ‘Taxon-K’, has been found in Kota Kinabalu, Borneo, Malaysia (Tobari, 1993; Tomimura et al., 1993). Its distribution has been expanding in Southeast Asia, and the flies can now be collected in Taiwan (C.S. Ng and A. Kopp, personal communication) and Yaeyama Islands, Japan (M. Watada, Y. Tomimura, and M. Matsuda, unpublished).

Our interest was to determine the phylogenetic relationship of the taxa

belonging to the *D. ananassae* species cluster. Because some of them are sympatric (*D. ananassae* and *D. pallidosa* in the South Pacific; *D. ananassae*, pallidosa-like, pallidosa-like-Wau, and papuensis-like in Papua New Guinea; *D. ananassae* and Taxon-K in Southeast Asia) and reproductive isolation is incomplete (at least under laboratory conditions), they might share polymorphism because of incomplete lineage sorting or historical gene flow. We selected a nuclear mitochondrial pseudogene, pseudo-cytochrome c oxidase subunit I (pseudo-*COI*), for the present molecular analysis. This pseudogene has substantial similarity to mitochondrial *COI* and is also useful to study mitochondrial gene transfer to the nuclear genome. In the present report, the pseudogene is localized to chromosome 4, which is believed to have reduced recombination; its syntenic chromosome is virtually non-recombining in *D. melanogaster* (Hochman, 1975). Therefore, the history of the chromosome can be traced from this pseudogene. The present results will elucidate the evolutionary relationships among the taxa belonging to the *D. ananassae* species cluster.

2. Materials and methods

2.1. Flies and sequencing

The sequences of the AABBg1 strain were taken from the *D. ananassae* Aug.

2005 (droAna2) assembly available from UCSC Genome Informatics (http://genome.ucsc.edu/) and used as a reference. One male was selected from each iso-female line (30 lines; Table 1), and its genomic DNA was extracted using the DNeasy tissue kit (Qiagen). The sequences of interest were amplified by PCR from extracts using the primers 5'-CAAGCGGACTGCGACTCAAC-3' and 5'-GTGGTTGGCCACTGGATAGG-3' (Fig. 1). Amplification was conducted at 95°C for 4 min followed by 30 cycles of 95°C for 30 sec, 60°C (or 62°C) for 30 sec, and 72°C for 30 sec, with a final extension reaction at 72°C for 7 min. The PCR products (ca. 630 bp) were purified from an agarose gel, and both strands were sequenced directly with the 3730xl DNA Analyzer (Applied Biosystems) using the same amplification primers as for PCR with the BigDye Terminator v3.1 Cycle Sequencing kit (Applied Biosystems). The sequences were edited by Sequencher ver. 4.5 (Gene Codes Corporation).

2.2. Sequence alignment and construction of a haplotype network

The pseudo-*COI* sequences of the examined strains were aligned with the pseudo-*COI* sequence of AABBg1 using Genetyx ver. 6 (Genetyx Corporation), and nucleotide substitutions and insertion/deletions (indels) were searched by eye. The

results near the primers were ambiguous, so only sequences corresponding to nucleotide positions 56 to 577 of AABBg1 (Fig. 1) were considered in the present analysis. Because only seven haplotypes, each separated by a few mutation steps, were detected, the haplotype network was constructed manually.

2.3. Population genetic analysis of pseudo-COI

Nucleotide diversity within taxon (π; Nei, 1987), net average pairwise differences between taxa (D_a; Nei, 1987), and F_{ST} (Hudson et al., 1992) were calculated using the program SITES developed by J. Hey (available at http://lifesci.rutgers.edu/~heylab/HeylabSoftware.htm#SITES). To examine the possibility of gene flow, we analyzed the *D. ananassae* and *D. pallidosa* sequences under the "isolation with migration" (IM) model of divergence (Hey and Nielsen, 2004). We used the IM program (available at http://lifesci.rutgers.edu/~heylab/ HeylabSoftware.htm#IM) to estimate the marginal posterior probability distributions for six demographic parameters, including two-directional gene flow rates. Fifty million steps were sampled after a burn-in of 10,000 steps. Three independent runs revealed similar patterns for each demographic parameter.

2.4. Localization of pseudo-COI

To genetically localize pseudo-*COI* to chromosome 4, we used a recessive mutant, *sparkling*^{82} (*spa*^{82}) of *D. ananassae*, which contains a *Pvu*II site that results in a restriction fragment length polymorphism (RFLP) when compared with the reference strain, TNG. The *spa*^{82} mutation, characterized by a rough eye surface, is a derivative of HYD3 collected at Hyderabad, India, in 1981 by F. Hihara and O. Kitagawa, and resides on chromosome 4 (Moriwaki and Tobari, 1993). Females from the *spa*^{82} strain were crossed to TNG males, and the F_1 males were backcrossed to *spa*^{82} females (Fig. 2). The DNA from *spa/spa* and *spa/+* F_2 flies was isolated and amplified by PCR as described above and then digested with *Pvu*II (Takara Bio, Inc.) at 37°C for 2 hr. The fragments were electrophoresed on a 1% agarose gel or with the HDA-GT12 Genetic Analyzer (eGene, Inc.) and examined for RFLP.

3. Results

3.1. Sequence of pseudo-COI *and its nearby region*

The 630-bp sequence from the AABBg1 strain of *D. ananassae* (scaffold_13077:438334–438963) is shown as 'pseudo-COI' in Figure 1B. A gene, 13077.47, has been predicted near the pseudo-*COI* sequence by the GENSCAN

program (Burge and Karlin, 1997), although there is no evidence that this gene is actually transcribed. The 3' end of the first exon of 13077.47 corresponds to 56 bp of the reverse strand of the 3' region of the sequence shown in Figure 1B (positions 575–630; lower shaded region), and its predicted amino acid sequence is also indicated as '13077.47 a.a.'; the remainder of the sequenced region corresponds to the 13077.47 intronic sequence. At the 5' end of the sequenced region, 79 bp (upper shaded region) have substantial similarity to an interspersed sequence of *D. ananassae* existing in high copy number. The sequence shares several features with a *D. funebris* transposon, ISFUN-1 (Amador and Juan, 1999): sequence similarity, the well-conserved 14-bp inverted repeat in terminal and subterminal positions, no coding function, and dispersed genomic distribution (H. Sato and K. Sawamura, unpublished).

The internal 491-bp pseudo-*COI* sequence (positions 72–562) has 94% similarity (excluding gaps) to the middle region (positions 636–1133) of the 1536-bp mitochondrial *COI* of *D. ananassae* (scaffold_3895:2565–4100; Fig. 1A), and the predicted amino acid sequence of *COI* is indicated as 'COI a.a.' (Fig. 1B). The boundary between the interspersed sequence and pseudo-*COI* (positions 72–86; TTTCAATACATTTAG) is also similar to an upstream region of *COI* (positions 437–451; TTTCATTACATTTAG) (corresponding *COI* sequence not shown in Figure 1).

Because of frame shifts (a substitution of 3 bp with 10 bp and a 14-bp deletion) and stop codons, the pseudo-*COI* sequence does not appear to encode protein.

3.2. A haplotype network of pseudo-COI

The sequences of interest were amplified from 26 of the 30 strains examined; amplification failed in two pallidosa-like-Wau (WAU61 and WAU92) and two papuensis-like (LAE360 and LAE376) strains. Because several different primer sets that were designed based on the AABBg1 reference sequence failed to amplify the region, the nearby sequences of pseudo-*COI* might have diverged in these strains. Inspection of the resulting chromatograms indicated no heterozygosity of sequences in the 26 flies examined.

We recognized seven haplotypes (#0 – #6) in the sequences (Table 1 and 2) and constructed a haplotype network (Fig. 3). Taxon-K (haplotype #6) is quite different from the others, and of them, #2 is the most closely related; 18 mutation steps (17 substitutions and one indel) separate them. In Figure 3, the branch leading to haplotype #6 is attached to a hypothetical haplotype to minimize the mutation steps. Five haplotypes (#0 – #4) were recognized in *D. ananassae*, in which four of them (#0 – #3) are separated from each other by only one or two substitutions. Haplotypes of

pallidosa-like-Wau (#1), pallidosa-like (#2), and papuensis-like (#2) are included in this cluster. The other *D. ananassae* haplotype (#4) is separated from #2 by three substitutions, and is shared with *D. pallidosa*. One of the *D. pallidosa* strains showed one substitution from #4, and this haplotype is called #5 here.

3.3. Population structure based on pseudo-COI

Only one haplotype of pseudo-*COI* was detected in each taxon other than *D. ananassae* and *D. pallidosa*, although the sample size is very small (one to four). Nucleotide diversity of *D. ananassae* (π = 0.00586) was larger than that of *D. pallidosa* (0.00064) in which one strain exhibited one base substitution. Nucleotide divergence between taxa represented by D_a is small when Taxon-K is excluded (<0.01), whereas Taxon-K exhibits large nucleotide divergence from the others (>0.03; Table 3). Similarly, F_{ST} between *D. ananassae* and Taxon-K (0.917) is larger than F_{ST} between *D. ananassae* and the others (<0.5). Because haplotypes are almost monomorphic in taxa other than *D. ananassae*, F_{ST} of the remaining pairs are large (≈1; different haplotypes fixed) or undefined (the same haplotype fixed in pallidosa-like and papuensis-like).

To further examine the gene flow between *D. ananassae* and *D. pallidosa*, we used the IM program (Hey and Nielsen, 2004). The IM method revealed unidirectional

gene flow between *D. ananassae* and *D. pallidosa* (Fig. 4). The estimated migration parameter from *D. pallidosa* to *D. ananassae* was 5.585 (90 % highest posterior density (HPD) interval 2.695 – 9.995). The fact that the lower bound of the estimated 90 % HPD is 2.695 supports presence of post-divergence gene flow. On the other hand, there appeared to be little evidence of gene flow in the opposite direction, with the highest value of marginal posterior density essentially at zero. However, the right-hand tail of the distribution did not reach to zero, and thus we could not obtain the confidence intervals accurately for this parameter. Divergence time parameter was estimated as 9.805 (90 % HPD interval 2.445-9.995), but this estimate could not be accurate because the curve of the posterior density distribution was not declining at the upper limit (data not shown).

3.4. Genetic localization of pseudo-COI

The pseudo-*COI* gene is within scaffold_13077 of the *D. ananassae* genome (*Drosophila* 12 Genomes Consortium, 2007). Because this region of the genome includes genes homologous to *toy* (cytological position, 102D1), *plexA* (102D1), *ATPsyn-β* (102D1), and *Zyx102EF* (102D3) of *D. melanogaster* (Adams et al., 2000), the genomic region of interest may be syntenic with chromosome 4 of *D. melanogaster*.

Thus, we tested if pseudo-*COI* of *D. ananassae* is also on chromosome 4.

The pseudo-*COI* was amplified by PCR from the spa^{82} strain, and its sequence was shown to be identical to the reference sequence of AABBg1 (haplotype #0), in which a *Pvu*II restriction site (CAGCTG) exists at positions 102–107. Because the nucleotide substitution of C to T at position 102 in haplotypes #4 and #5 disrupts the restriction site, the TNG strain from #4 was selected for the RFLP analysis. The RFLP pattern co-segregated with the visible marker *spa* in 98 F_2 flies examined (Fig. 2): 27 *spa/spa* females and 23 *spa/spa* males were also homozygous for the allele from the spa^{82} strain, whereas 16 *spa/+* females and 32 *spa/+* males were heterozygous for the alleles from the spa^{82} and TNG strains. This confirms that pseudo-*COI* is on chromosome 4 of *D. ananassae*.

4. Discussion

4.1. Origin of pseudo-COI

Interestingly, the predicted intron of 13077.47 of *D. ananassae* contains not only an interspersed sequence but also a sequence with substantial similarity to mitochondrial *COI*. These sequences might have accumulated in this region as junk DNA. Transposition of mitochondrial genes to the nuclear genome is common in many

organisms but such nuclear mitochondrial (Numt) sequences are rare in *D. melanogaster*, presumably because pseudogenes are easily lost from the compact genomes (for reviews see Perna and Kocher, 1996; Zhang and Hewitt, 1996; Bensasson et al., 2001; Richly and Leister, 2004). The present case will provide an opportunity to study the mechanism of mitochondrial gene transfer to the nuclear genome. Although the precise mechanism of integration has not been described, Numts are frequently associated with repetitive elements (Mishmar et al., 2004). More Numts might exist in *D. ananassae* than in *D. melanogaster*, as the former has more interspersed sequences in the genome (Smith et al., 2007). Apparently, pseudo-*COI* described here does not encode any protein, although it may possibly function in RNA interference as has been indicated in other pseudogenes (Devor, 2006).

4.2. Historical population structure inferred from pseudo-COI

Because the pseudo-*COI* sequence of the spa^{82} strain, derived from an Indian population, was identical to that of the AABBg1 strain established from a Hawaiian population, this haplotype (#0) seems to be cosmopolitan. We detected four other haplotypes (#1 – #4) in *D. ananassae*, although only populations from Papua New Guinea and the South Pacific were analyzed in the present study. Interestingly,

pallidosa-like-Wau, pallidosa-like, and papuensis-like were shown to share haplotype #1 or #2 with *D. ananassae*, and *D. pallidosa* was shown to share haplotype #4 with *D. ananassae*. Apparently, the haplotypic clusters do not coincide with our taxonomic classification, except for Taxon-K (haplotype #6). Rather, haplotypes well represent the geographic origin of the strains: #1, #2, and #3 are mainly from Papua New Guinea, and #4 and #5 are from Samoa, Tonga, Fiji, and New Caledonia. Although #1 was described as a Papua New Guinean haplotype above, this haplotype is also seen in *D. ananassae* from Fiji (NAN84) and New Caledonia (NOU78). This might have resulted from migration from Papua New Guinea to the South Pacific.

Incongruence between taxon boundaries and haplotypic clusters is explained by incomplete lineage sorting or introgressions (Nichols, 2001; Machado and Hey, 2003; Bachtrog et al., 2006; Pollard et al., 2006; and references therein). The latter is more plausible in the present case for two reasons: (1) haplotypes are shared by sympatric populations of different taxa, and (2) reproductive isolation is incomplete, at least under laboratory conditions. Furthermore, the simulation based on the 'isolation with migration' model suggested potential gene flow from *D. pallidosa* to *D. ananassae*.

4.3. Gene flow, genetic drift, and natural selection in speciation

Because some haplotypes of mitochondria or the Y chromosome are shared between different taxa of the *D. ananassae* species cluster (C.S. Ng and A. Kopp, personal communication; see also Schug et al., 2008), gene flow between sympatric populations is also presumed, although it might be the result of incomplete lineage sorting. Recent population genetics analysis using microsatellite markers has suggested that divergence between some *D. ananassae* populations—even from the same locality—is much higher than that from the outgroup *D. pallidosa* (Schug et al., 2007). Although this can be explained by a hypothesis that *D. pallidosa* is derived from a local population of *D. ananassae*, this finding might also be the result of historical gene flow; the genetic distance between species of sympatric populations can be decreased by gene flow.

Importantly, the sequence analyzed in the present study is on a chromosome in which recombination is suppressed. The genealogy of the gene must correspond to that of the chromosome, and we can trace the history of the chromosome using the present data. In this context, it is interesting that the same haplotype (#4) is nearly fixed in the South Pacific populations of *D. ananassae* and *D. pallidosa*. Although this may have resulted from genetic drift (i.e., #4 is adaptively neutral), it is also plausible that #4 was fixed by natural selection for genes on chromosome 4 (i.e., selective sweep or

background selection) (Berry et al., 1991; Begun and Aquadro, 1992; Stephan and Mitchell, 1992; Hilton et al., 1994; Chen et al., 2000; Wang et al., 2002; Machado and Hey, 2003; Noor and Kliman, 2003; and references therein). The nucleotide diversity of *D. ananassae* populations from wider geographic range (π = 0.00586) is, however, similar to that of ten X-linked loci (on average, π = 0.0079; Das et al., 2004). On the contrary, the unique haplotype of Taxon-K (#6) might be the result of higher mutation rate and/or natural selection for genes on chromosome 4. Gene flow from sympatric *D. ananassae* to Taxon-K must be restricted, as hybrid males from the cross of *D. ananassae* females and Taxon-K males are sterile (Y. N. Tobari and M. Matsuda, unpublished).

In the present study, we examined only chromosome 4, a small portion of the genome. Apparently, more genes are necessary to elucidate the comprehensive history of the cryptic species belonging to the *D. ananassae* species cluster.

Acknowledgments We are grateful to Dr. Y. N. Tobari for providing us with useful advice and comments, and to members of the Dr. A. Kopp lab for sharing their unpublished results. The present research was supported by a Grant-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Culture, Sports, Science and

Technology (MEXT). M. M. also thanks the National Bio-Resource Project (NBRP) of MEXT for *Drosophila* resource support.

Figure 1. The DNA region analyzed in the present research. (A) Comparison between *D. ananassae* nuclear pseudo-*COI* on chromosome 4 (Chr. 4) and mitochondrial (Mt) *COI* (scaffold_3895:2565–4100). Both strands of chromosome 4 are shown, and the boxes below represent exons of a predicted gene, 13077.47, corresponding to scaffold_13077: 439044–438908 and 437110–437035. The region with substantial similarity to mitochondrial *COI* is shaded. Arrows indicate the location and direction of the PCR primers. The sequence of positions 72–86 of pseudo-*COI* is also similar to that of mitochondrial *COI* (positions 437–451). (B) The reference sequence of the AABBg1 strain of *D. ananassae*. The sequence of pseudo-*COI* is aligned with mitochondrial *COI* for the comparison, and the deduced amino acid sequences of 13077.47 and *COI* are also indicated as 13077.47 a.a. and COI a.a., respectively. Shaded sequences indicate those similar to an interspersed sequence (upper) and the predicted first exon of 13077.47 (lower). The sequences substituted or deleted in pseudo-*COI* compared with *COI* are bracketed. The *Pvu*II site used for the RFLP analysis is also indicated (boxed).

Figure 2. Genetic localization of pseudo-*COI* on chromosome 4. (A) Mating scheme. The TNG and spa^{82} strains of *D. ananassae* were crossed, and the DNA from *spa/spa* and *spa/+* F_2 flies was isolated. (B) RFLP pattern. The sequences of pseudo-*COI* were

amplified by PCR and then digested with *Pvu*II. The fragments were separated and visualized by a HDA-GT12 Genetic Analyzer and examined for RFLP.

Figure 3. Haplotype network of pseudo-*COI*. Numbered open circles represent recognized haplotypes connected by lines with hypothetical haplotypes (smaller shaded circles) indicating mutation steps. ANA, *D. ananassae*; PAL, *D. pallidosa*; PAL-like, pallidosa-like; PAL-like-W, pallidosa-like-Wau; PAP-like, papuensis-like; TK, Taxon-K. Geographic distribution of the haplotypes is also indicated (see text for exceptions).

Figure 4. The marginal posterior probability distributions for the IM model parameters for migration (scaled by the mutation rate) between populations of *D. ananassae* and *D. pallidosa*.

References

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D. et al., 2000. The genome sequence of *Drosophila melanogaster*. Science 287, 2185–2195.

Amador, A., Juan, E., 1999. Nonfixed duplication containing the *Adh* gene and a truncated form of the *Adhr* gene in the *Drosophila funebris* species group: different modes of evolution of *Adh* relative to *Adhr* in *Drosophila*. Mol. Biol. Evol. 16, 1439–1456.

Bachtrog, D., Thornton, K., Clark, A., Andolfatto, P., 2006. Extensive introgression of mitochondrial DNA relative to nuclear genes in the *Drosophila yakuba* species group. Evolution 60, 292–302.

Baines, J.F., Das, A., Mousset, S., Stephan, W., 2004. The role of natural selection in genetic differentiation of worldwide populations of *Drosophila ananassae*. Genetics 168, 1987–1998.

Begun, D.J., Aquadro, C.F., 1992. Levels of naturally occurring DNA polymorphism are correlated with recombination rates in *Drosophila melanogaster*. Nature 356, 519–520.

Bensasson, D., Zhang, D.X., Hartl, D.L., Hewitt, G.M., 2001. Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends Genet. 16, 314–321.

Berry, A.J., Ajioka, J.W., Kreitman, M., 1991. Lack of polymorphism on the *Drosophila* fourth chromosome resulting from selection. Genetics 129, 1111–1117.

Bock, I.R., Wheeler, M.R., 1972. The *Drosophila melanogaster* species group. Univ. Texas Publs. 7213, 1–102.

Burge, C., Karlin, S., 1997. Prediction of complete gene structure in human genomic DNA. J. Mol. Biol. 268, 78–94.

Chen, Y., Marsh, B.J., Stephan, W., 2000. Joint effects of natural selection and recombination on gene flow between *Drosophila ananassae* populations. Genetics 155, 1185–1194.

Da Lage, J.L., Kergoat, G.J., Maczkowiak, F., Silvain, J.F., Cariou, M.L., Lachaise, D., 2007. A phylogeny of Drosophilidae using the *Amyrel* gene: questioning the *Drosophila melanogaster* species group boundaries. J. Zoo. Syst. Evol. Res. 45, 47–63.

Das, A., Mohanty, S., Stephan, W., 2004. Inferring the population structure and demography of *Drosophila ananassae* from multilocus data. Genetics 168, 1975–1985.

Devor, E.J., 2006. Primate microRNAs miR-220 and miR-492 lie within processed pseudogenes. J. Heredity 97, 186–190.

Doi, M., 1997. Genetic studies on reproductive isolation and speciation in the *Drosophila ananassae* complex. Ph.D. dissertation submitted to University of Tsukuba, Tsukuba.

Drosophila 12 Genomes Consortium, 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450, 203–218.

Futch, D.G., 1966. A study of speciation in South Pacific populations of *Drosophila ananassae*. Univ. Texas Publs. 6615, 79–120.

Futch, D.G., 1973. On the ethological differentiation of *Drosophila ananassae* and *Drosophila pallidosa* in Samoa. Evolution 27, 456–467.

Hey, J., Nielsen, R., 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics 167, 747–760.

Hilton, H., Kliman, R.M., Hey, J., 1994. Using hitchhiking genes to study adaptation and divergence during speciation within the *Drosophila melanogaster* species complex. Evolution 48, 1900–1913.

Hochman, B., 1975. The fourth chromosome of *Drosophila melanogaster*. In: Ashburner, M., Novitski, E. (Eds.), The Genetics and Biology of *Drosophila*, vol. 1b. Academic Press, London, pp. 903–928.

Hudson, R.R., Slatkin, M., Maddison, W.P., 1992. Estimation of levels of gene flow from DNA sequence data. Genetics 132, 583–589.

Kaneshiro, K., Wheeler, M.R., 1970. Preliminary report on the species of the ananassae subgroup. *Drosophila* Inf. Serv. 45, 143.

Lemeunier, F., David, J., Tsacas, L., Ashburner, M., 1986. The melanogaster species group. In: Ashburner, M., Carson, H., Thompson Jr., J.N. (Eds.), The Genetics and Biology of *Drosophila*, vol. 3e. Academic Press, London, pp. 147–256.

Machado, C.A., Hey, J., 2003. The causes of phylogenetic conflict in a classic *Drosophila* species group. Proc. R. Soc. Lond. B, 270, 1193–1202.

Mishmer, D., Ruiz-Pesini, E., Brandon, M., Wallace, D.C., 2004. Mitochondrial DNA-like sequences in the nucleus (NUMTs): insights into our African origins and the mechanism of foreign DNA integration. Human Mut. 23, 125–133.

Moriwaki, D., Tobari, Y.N., 1993. Catalog of mutants. In: Tobari, Y.N. (Ed.), *Drosophila ananassae*: Genetical and Biological Aspects. Japan Scientific Societies Press, Tokyo, pp. 209–259.

Nei, M., 1987. Molecular Evolutionary Genetics. Columbia University Press, New York, pp. 1–512.

Nichols, R., 2001. Gene trees and species trees are not the same. Trends Ecol. Evol. 16,

358–364.

Noor, M.A.F., Kliman, R.M., 2003. Variability on the dot chromosome in the *Drosophila simulans* clade. Genetica 118, 51–58.

Perna, N.T., Kocher, T.D., 1996. Mitochondrial DNA: molecular fossils in the nucleus. Curr. Biol. 6, 128–129.

Pollard, D.A., Iyer, V.N., Moses, A.M., Eisen, M.B., 2006. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. Publ. Lib. Sci. Genet. 2, 1634–1647.

Richly, E., Leister, D., 2004. NUMTs in sequenced eukaryotic genomes. Mol. Biol. Evol. 21, 1081–1084.

Schug, M.D., Regulski, E.E., Pearce, A., Smith, S.G., 2004. Isolation and characterization of dinucleotide repeat microsatellites in *Drosophila ananassae*. Genet. Res. 83, 19–29.

Schug, M.D., Smith, S.G., Tozier-Pearce, A., McEvey, S.F., 2007. The genetic structure of *Drosophila ananassae* populations from Asia, Australia and Samoa. Genetics 175, 1429–1440.

Schug, M.D., Baines, J.F., Killon-Atwood, A., Mohanty, S., Das, A., Grath, S., Smith, S.G., Zargham, S., McEvey, S.F., Stephan, W., 2008. Evolution of mating isolation

between populations of *Drosophila ananassae*. Mol. Ecol. 17, 2706–2721.

Smith, C.D., Edgar, R.C., Yandell, M.D., Smith, D.R., Celniker, S.E., Myers, E.W., Karpen, G.H., 2007. Improved repeat identification and masking in Dipterans. Gene 389, 1–9.

Stephan, W., Mitchell, S.J., 1992. Reduced levels of DNA polymorphism and fixed between-population differences in the centromeric region of *Drosophila ananassae*. Genetics 132, 1039–1045.

Stone, W.S., Wheeler, M.R., Wilson, F.D., Gerstenberg, V.L., Yang, H., 1966. Genetic studies of natural populations of *Drosophila*. II. Pacific island populations. Univ. Texas Publs. 6615, 1–36.

Tobari, Y.N., 1993. Geographical distribution. In: Tobari, Y.N. (Ed.), *Drosophila ananassae*: Genetical and Biological Aspects. Japan Scientific Societies Press, Tokyo, pp. 19–22.

Tomimura, Y., Matsuda, M., Tobari, Y.N., 1993. Polytene chromosome variations of *Drosophila ananassae* and its close relatives. In: Tobari, Y.N. (Ed.), *Drosophila ananassae*: Genetical and Biological Aspects. Japan Scientific Societies Press, Tokyo, pp. 139–151.

Vogl, C., Das, A., Beaumont, M., Mohanty, S., Stephan, W., 2003. Population

subdivision and molecular sequence variation: theory and analysis of *Drosophila ananassae* data. Genetics 165, 1385–1395.

Wang, W., Thornton, K., Berry, A., Long, M., 2002. Nucleotide variation along the *Drosophila melanogaster* fourth chromosome. Science 295, 134–137.

Yamada, H., Sakai, T., Tomaru, M., Doi, M., Matsuda, M., Oguma, Y., 2002. Search for species-specific mating signal in courtship songs of sympatric sibling species, *Drosophila ananassae* and *D. pallidosa*. Genes Genet. Syst. 77, 97–106.

Zhang, D.X., Hewitt, G.M., 1996. Nuclear integrations: challenges for mitochondrial DNA markers. Trends Ecol. Evol. 11, 247–251.

Table 1 Iso-female lines analyzed for the haplotype of pseudo-*COI*

Taxon	Strain	Locality	Year	Collector	Haplotype
D. ananassae					
	AABBg1	Hawaii, USA	1945	*	#0
	NAN84	Lautoka, Fiji	1981	a	#1
	NOU78	Noumea, New Caledonia	1981	a	#1
	WAU138	Wau, Papua New Guinea	1981	b	#1
	POM431	Port Moresby, Papua New Guinea	1981	b	#2
	WAU120	Wau, Papua New Guinea	1981	b	#2
	LAE119	Lae, Papua New Guinea	1981	b	#3
	PPG183	Pago Pago, Samoa	1981	a	#4
	TBU136	Tongatapu, Tonga	1981	a	#4
	TBU209	Tongatapu, Tonga	1981	a	#4
	TNG	Tongatapu, Tonga	1966	**	#4
	VAV161	Vava'u, Tonga	1981	a	#4
D. pallidosa					
	NAN4	Lautoka, Fiji	1981	a	#4
	NAN64	Lautoka, Fiji	1981	a	#4
	NAN66	Lautoka, Fiji	1981	a	#4
	TBU155	Tongatapu, Tonga	1981	a	#4
	VAV92	Vava'u, Tonga	1981	a	#4
	NAN57	Lautoka, Fiji	1981	a	#5
pallidosa-like					
	LAE346	Lae, Papua New Guinea	1981	b	#2
	POM446	Port Moresby, Papua New Guinea	1981	b	#2
	POM458	Port Moresby, Papua New Guinea	1981	b	#2
	POM473	Port Moresby, Papua New Guinea	1981	b	#2
pallidosa-like-Wau					
	BLO79	Bulolo, Papua New Guinea	1979	c	#1
	WAU61	Wau, Papua New Guinea	1981	b	NA
	WAU92	Wau, Papua New Guinea	1981	b	NA
papuensis-like					
	WAU142	Wau, Papua New Guinea	1981	b	#2
	WAU145	Wau, Papua New Guinea	1981	b	#2
	LAE360	Lae, Papua New Guinea	1981	b	NA
	LAE376	Lae, Papua New Guinea	1981	b	NA
Taxon-K					
	B43	Kota Kinabalu, Malaysia	1971	d	#6
	T184	Kota Kinabalu, Malaysia	1979	e	#6

Collector:
a, Y. Fuyama, E. Takanashi (Matsuura), Y. N. Tobari
b, E. Takanashi (Matsuura), Y. N. Tobari
c, H. L. Carson, T. Okada
d, O. Kitagawa, K. I. Wakahama, T. Watanabe
e, Y. Fuyama, F. Hihara, T. K. Watanabe
* inbred line (the sequenced strain)
** donated by University of Texas

NA: not amplified (see text)

Table 2 Summary of variable sites in pseudo-*COI*

Haplotype	Nucleotide position																							
	61	69	79	101	102	141	146	147	206	214	243	246	269	349	360	375	384	387	389	391; 392	425	442	476	501	540
#0	C	C	A	C	C	T	T	T	A	A	G	C	T	T	A	C	C	G	A	-	A	C	A	A	G
#1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.
#2	T	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.
#3	T	T	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	T	.	.
#4	T	T	.	T	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	T	.	.
#5	T	T	.	T	T	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	T	T	.	.
#6	T	T	T	.	.	.	G	A	T	T	A	T	C	A	C	T	T	T	G	TCA	.	T	T	T	A

Table 3 Nucleotide divergence D_a (above diagonal) and F_{ST} (below diagonal)

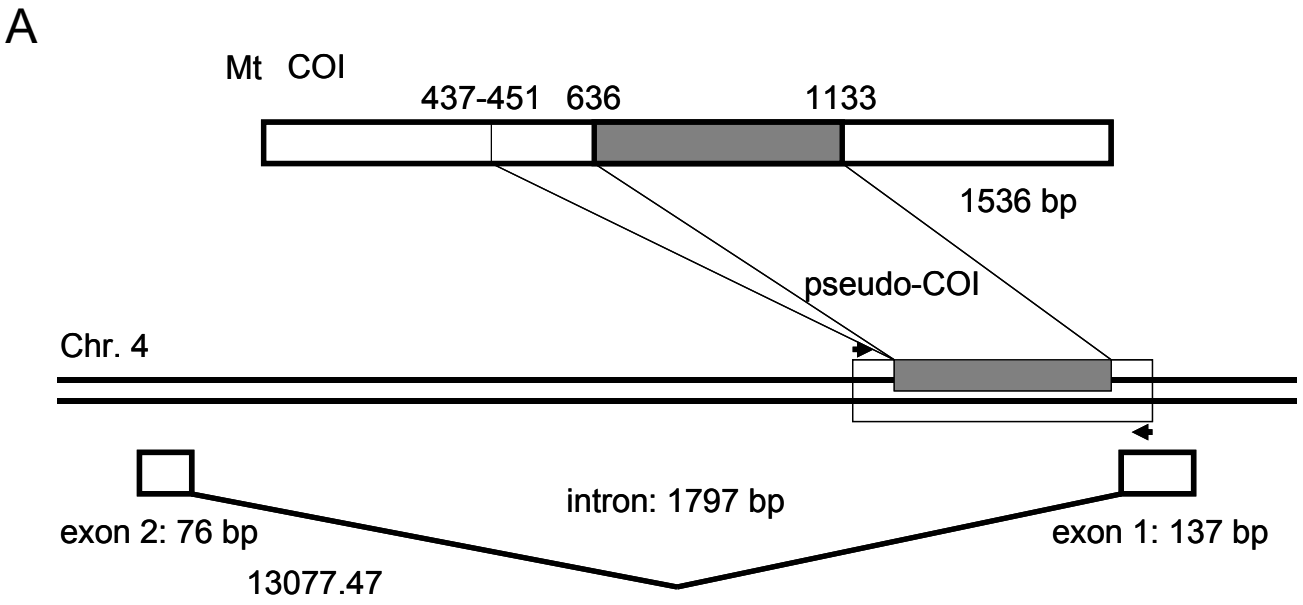
Taxon	1	2	3	4	5	6
1 *D. ananassae*		0.00218	0.00122	0.0025	0.00122	0.03219
2 *D. pallidosa*	0.401		0.00575	0.00958	0.00575	0.03416
3 pallidosa-like	0.294	0.947		0.00383	0	0.03257
4 pallidosa-like-Wau	0.46	0.968	1		0.00383	0.0364
5 papuensis-like	0.294	0.947	NA	1		0.03257
6 Taxon-K	0.917	0.991	1	1	1	

D_a: net average pairwise difference.

F_{ST}: a measure of population differentiation (Hudson et al., 1992).

NA: not available (see text).

Figure 1


A
Mt COI
437-451
636
1133
1536 bp
pseudo-COI
Chr. 4
intron: 1797 bp
exon 2: 76 bp
exon 1: 137 bp
13077.47

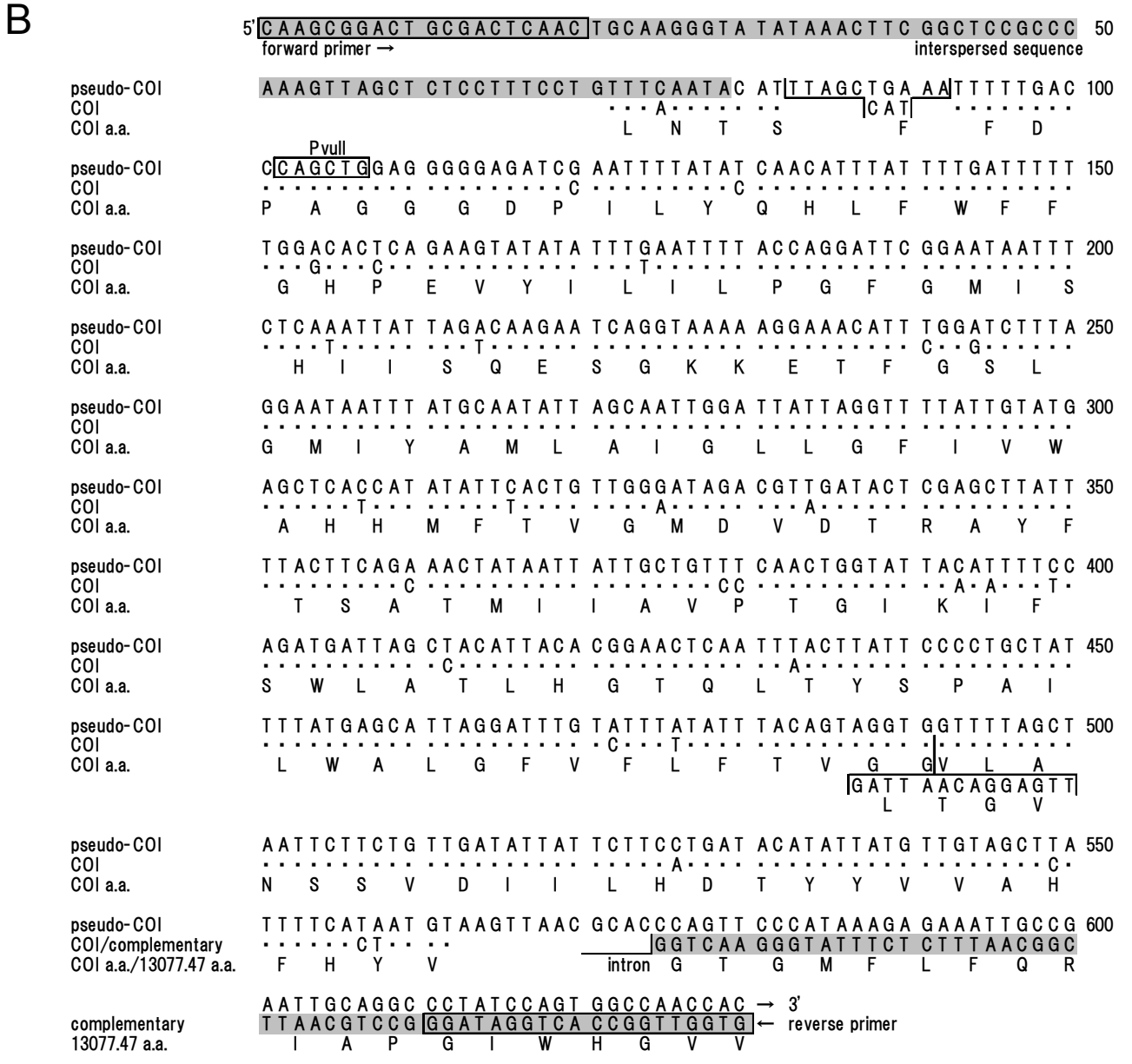

B
5' CAAGCGGACT GCGACTCAAC TGCAAGGGTA TATAAACTTC GGCTCCGCCC 50
forward primer →
interspersed sequence
pseudo-COI AAAGTTAGCT CTCCTTTCCT GTTTCAATAC ATTTAGCTGA AATTTTTGAC 100
COI ...A..... .. CAT ....
COI a.a. L N T S F F D
PvuII
pseudo-COI CCAGCTGGAG GGGGAGATCG AATTTTATAT CAACATTTAT TTTGATTTTT 150
COI .....C.....C.....
COI a.a. P A G G G D P I L Y Q H L F W F F
pseudo-COI TGGACACTCA GAAGTATATA TTTGAATTTT ACCAGGATTC GGAATAATTT 200
COI ...G...C.. .....T..... .....
COI a.a. G H P E V Y I L I L P G F G M I S
pseudo-COI CTCAAATTAT TAGACAAGAA TCAGGTAAAA AGGAAACATT TGGATCTTTA 250
COI ....T..... ...T..... .....C..G.....
COI a.a. H I I S Q E S G K K E T F G S L
pseudo-COI GGAATAATTT ATGCAATATT AGCAATTGGA TTATTAGGTT TTATTGTATG 300
COI .....
COI a.a. G M I Y A M L A I G L L G F I V W
pseudo-COI AGCTCACCAT ATATTCACTG TTGGGATAGA CGTTGATACT CGAGCTTATT 350
COI ......T... .....T.... ....A..... ...A..... .....
COI a.a. A H H M F T V G M D V D T R A Y F
pseudo-COI TTACTTCAGA AACTATAATT ATTGCTGTTT CAACTGGTAT TACATTTTCC 400
COI .....C..... .....CC..... ..A.A...T.
COI a.a. T S A T M I I A V P T G I K I F
pseudo-COI AGATGATTAG CTACATTACA CGGAACTCAA TTTACTTATT CCCCTGCTAT 450
COI .....C..... .....A.....
COI a.a. S W L A T L H G T Q L T Y S P A I
pseudo-COI TTTATGAGCA TTAGGATTTG TATTTATATT TACAGTAGGT GGTTTTAGCT 500
COI .....C...T..... .....
COI a.a. L W A L G F V F L F T V G G V L A
GATT AACAGGAGTT
L T G V
pseudo-COI AATTCTTCTG TTGATATTAT TCTTCCTGAT ACATATTATG TTGTAGCTTA 550
COI .....A..... .....C.
COI a.a. N S S V D I I L H D T Y Y V V A H
pseudo-COI TTTTCATAAT GTAAGTTAAC GCACCCAGTT CCCATAAAGA GAAATTGCCG 600
COI/complementary ......CT.. .. GGTCAA GGGTATTTCT CTTTAACGGC
COI a.a./13077.47 a.a. F H Y V intron G T G M F L F Q R
AATTGCAGGC CCTATCCAGT GGCCAACCAC → 3'
complementary TTAACGTCCG GGATAGGTCA CCGGTTGGTG ← reverse primer
13077.47 a.a. I A P G I W H G V V

Figure 2

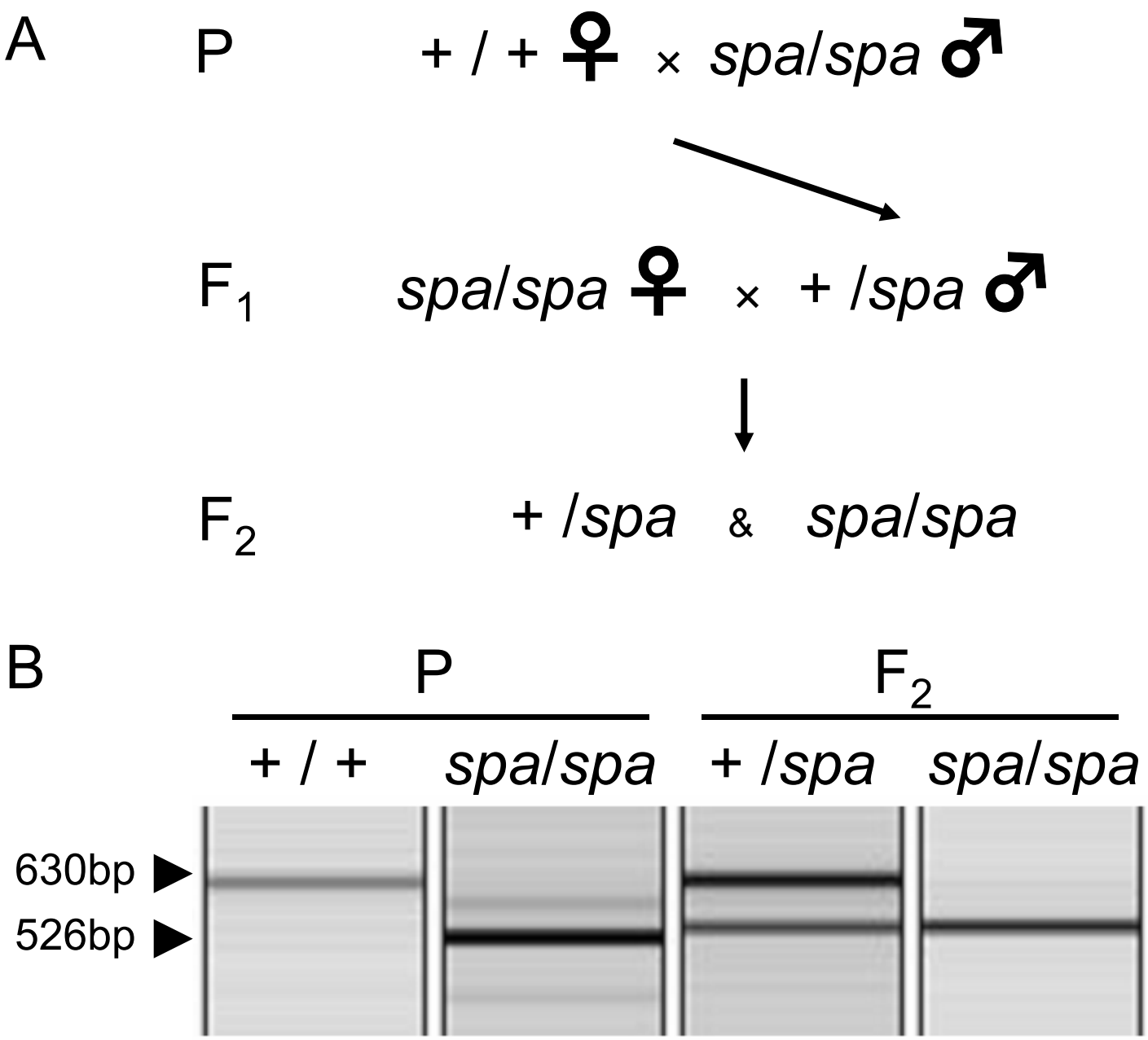

A
P
+ / + ♀ × spa/spa ♂
F1
spa/spa ♀ × + /spa ♂
F2
+ /spa & spa/spa
B
P
F2
+ / +
spa/spa
+ /spa
spa/spa
630bp
526bp

Figure 3

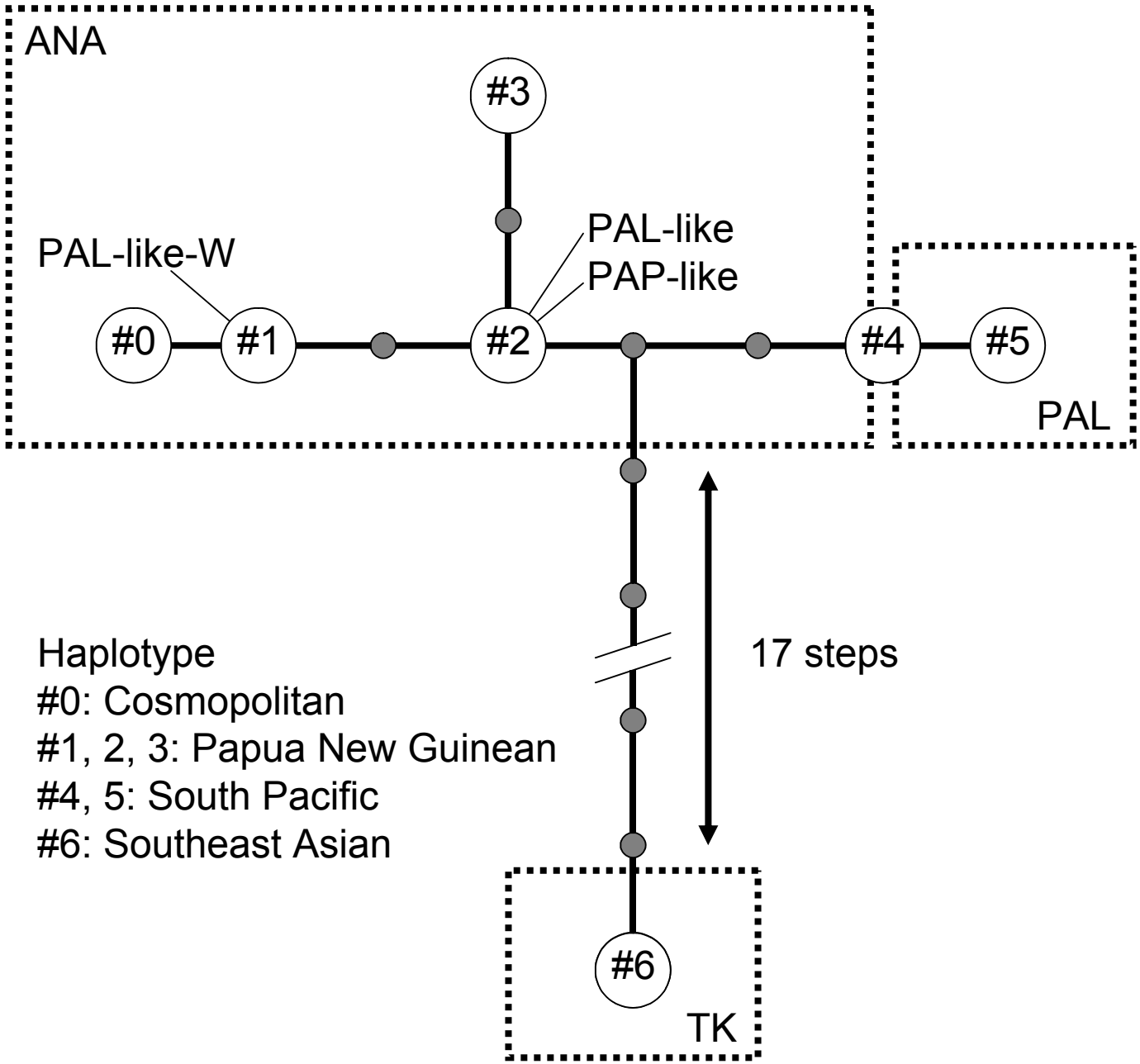
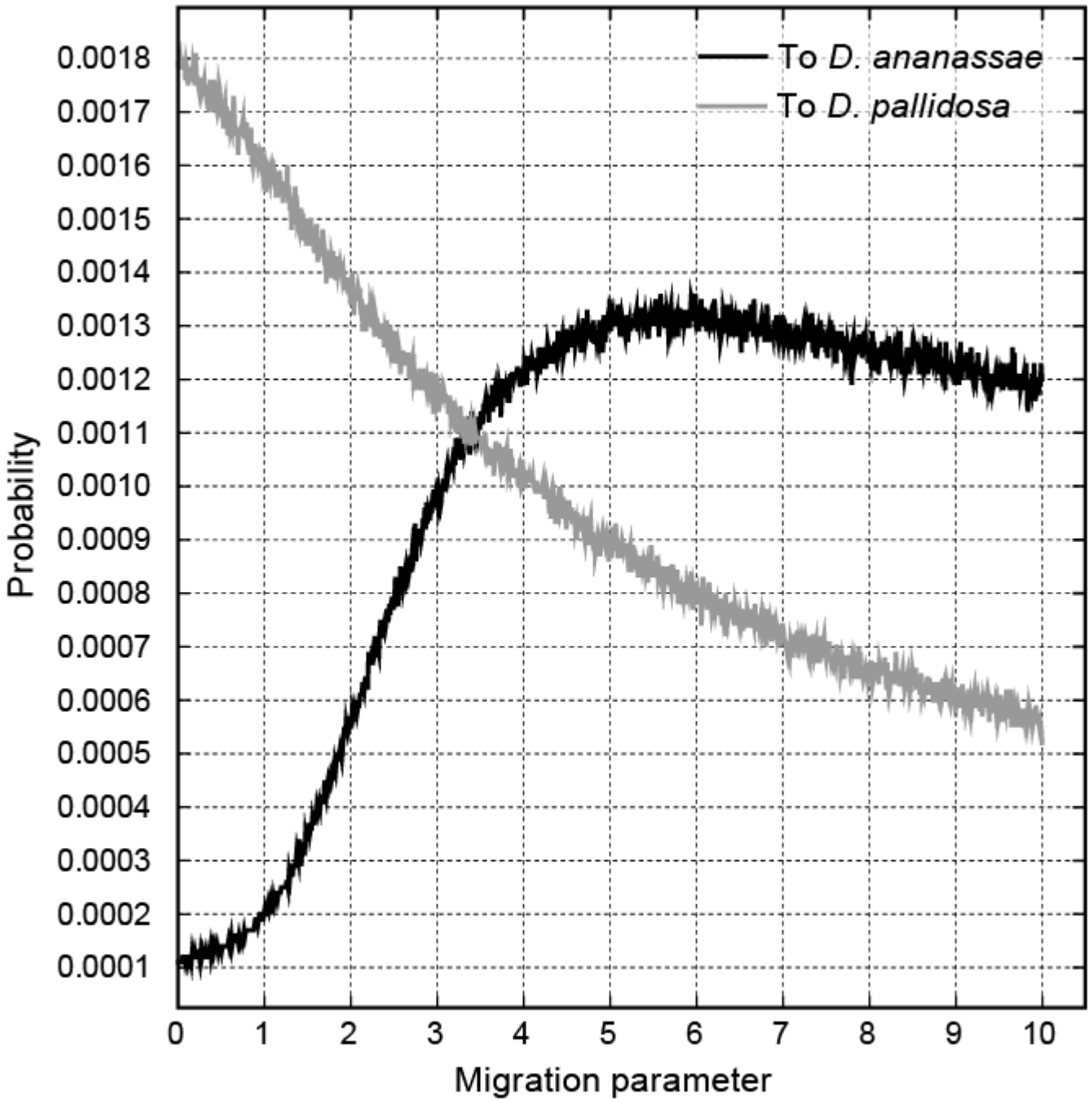

ANA
#3
PAL-like
PAP-like
PAL-like-W
#0
#1
#2
#4
#5
PAL
17 steps
Haplotype
#0: Cosmopolitan
#1, 2, 3: Papua New Guinean
#4, 5: South Pacific
#6: Southeast Asian
#6
TK

Figure 4


To *D. ananassae*
To *D. pallidosa*
Probability
0.0018
0.0017
0.0016
0.0015
0.0014
0.0013
0.0012
0.0011
0.0010
0.0009
0.0008
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0
1
2
3
4
5
6
7
8
9
10
Migration parameter