Rapid evolution of smell and taste receptor genes during host specialization in *Drosophila sechellia*

Carolyn S. McBride*

Center for Population Biology, University of California, Davis, CA 95616

Edited by Douglas J. Futuyma, State University of New York, Stony Brook, NY, and approved February 1, 2007 (received for review September 24, 2006)

Our understanding of the genetic basis of host specialization in insects is limited to basic information on the number and location of genetic factors underlying changes in conspicuous phenotypes. We know nothing about general patterns of molecular evolution that may accompany host specialization but are not traceable to a single prominent phenotypic change. Here, I describe changes in the entire repertoire of 136 olfactory receptor (*Or*) and gustatory receptor (*Gr*) genes of the recently specialized vinegar fly *Drosophila sechellia*. I find that *D. sechellia* is losing *Or* and *Gr* genes nearly 10 times faster than its generalist sibling *Drosophila simulans*. Moreover, those *D. sechellia* receptors that remain intact have fixed amino acid replacement mutations at a higher rate relative to silent mutations than have their *D. simulans* orthologs. Comparison of these patterns with those observed in a random sample of genes indicates that the changes at *Or* and *Gr* loci are likely to reflect positive selection and/or relaxed constraint associated with the altered ecological niche of this fly.

comparative genomics | gustatory receptor | host adaptation | lineage-specific | olfactory receptor

Host specialization and host shifts in insects that feed on plants provide excellent opportunities to study the genetic basis of ecological adaptation. Until now, however, this endeavor has been limited to attempts to map factors responsible for conspicuous phenotypic changes that accompany the ecological shifts (e.g., refs. 1–5). We know nothing about genetic changes whose individual effects are subtle, but whose combined presence may leave a striking signature on the genomes of specializing or host-shifting insects.

For example, insects evaluate their environment largely by smell and taste, and we might therefore expect their chemical sensory systems to evolve during host specialization or shifts. The acquisition of a novel host may drive the adaptive divergence of sensory systems by positive selection, and the abandonment of an ancestral host may result in the deterioration of older sensory adaptations by genetic drift (or positive selection). Despite their potentially subtle phenotypic effects, such changes are likely to be pervasive (particularly because new host plants challenge insects with the task of recognizing and responding not only to a new food but often also to novel toxins, bacteria, fungi, predators, parasitoids, pupation sites, and mating environments) and are best detected by examining entire genomes or large groups of genes simultaneously.

The olfactory receptor (*Or*) and gustatory receptor (*Gr*) gene families encode a diverse group of transmembrane proteins that bind volatile and soluble chemicals from the environment and trigger nerve impulses to the brain (6). Individual receptor genes in insects are highly divergent (paralogous genes from a single species often sharing $<20\%$ of their amino acids), are expressed in narrow subsets of olfactory and gustatory neurons from well defined regions of smell and taste organs, and largely determine the odor response properties of the neurons in which they are expressed (e.g., odors to which the neuron is sensitive, spontaneous firing rate and signaling mode of the neuron) (6). The families were first described in *Drosophila melanogaster*, which has 60 *Or* and 60 *Gr* genes (encoding 62 and 68 proteins respectively by alternative splicing) (7–9), but have subsequently been found in other insects (6). Given their essential function in smell and taste, *Or* and *Gr* genes are likely to be involved in any broad evolutionary response of insect sensory systems to host shifts. The genomic data necessary for a comprehensive survey of molecular evolution at these loci is now available for *Drosophila sechellia*, an insect that has recently undergone a dramatic case of host specialization.

D. sechellia is endemic to the Seychelles archipelago in the Indian Ocean. Biogeographical and phylogenetic evidence suggests that this species evolved in isolation after colonization of these islands approximately half a million years ago by its sister species, *Drosophila simulans* (10, 11). Interestingly, whereas *D. simulans* is a quintessential generalist (12), *D. sechellia* feeds solely on fruit of the shrub *Morinda citrifolia* (13, 14) and has evolved a remarkable chemical preference for (and resistance to) toxins that occur in *Morinda* and strongly repel other vinegar flies (15–18). Although this novel preference may be related to an overabundance of two types of olfactory receptor neurons on *D. sechellia* antennae, the binding specificities and sensitivities of these neurons appear to remain unaltered in comparison to *D. simulans* (19, 20). We know nothing about further potentially less conspicuous changes in *D. sechellia*'s chemosensory system.

In a novel approach to the genetics of host specialization, I use publicly available genome sequences to examine the molecular evolution of *D. sechellia*'s entire suite of olfactory and gustatory receptor genes and thus characterize the potential genetic signature of host specialization on an insect chemosensory system. My strategy is to look for consistent differences in the rate and character of evolution at *Or* and *Gr* loci between the *D. sechellia* and *D. simulans* lineages, using *D. melanogaster* as an outgroup (i.e., compare evolution along branches a and b in Fig. 1).

Results

Gene Annotations. Using a combination of TBLASTN searches, GeneWise predictions, manual revision, and direct sequencing, I was able to identify *D. sechellia* and *D. simulans* orthologs for all known *D. melanogaster Or* and *Gr* genes and splice forms. The close relation among these three species (Fig. 1) made assignments of orthology unambiguous. All orthologous pairs were reciprocal best hits and shared an upstream and/or downstream neighbor (i.e., were microsyntenic) in all species. I also identified one *Or* gene, one *Or* splice form, and five *Gr* genes in *D. sechellia* and *D. simulans* that have been deleted in *D. melanogaster* (the

Author contributions: C.S.M. designed research, performed research, analyzed data, and wrote the paper.

The author declares no conflict of interest.

This article is a PNAS direct submission.

Abbreviation: LOF, lack-of-function.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. DQ990077–DQ990148).

*E-mail: cmcbride@ucdavis.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0608424104/DC1.

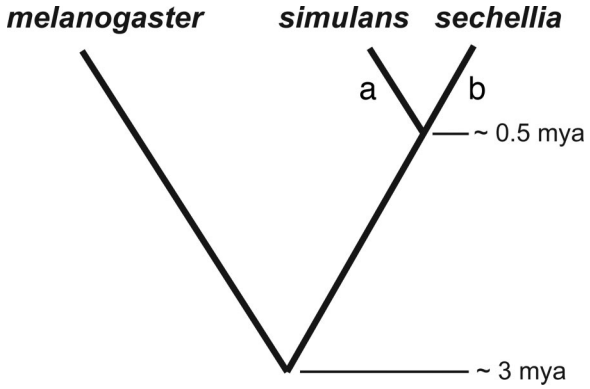


Fig. 1. Phylogenetic relationships among *D. sechellia*, *D. simulans*, and *D. melanogaster*. *D. sechellia* is a specialist, whereas *D. simulans* is a generalist. I compare the molecular evolution of *Or* and *Gr* genes along these two lineages (branches a and b) using *D. melanogaster* as an outgroup. Divergence times are from refs. 10 and 11.

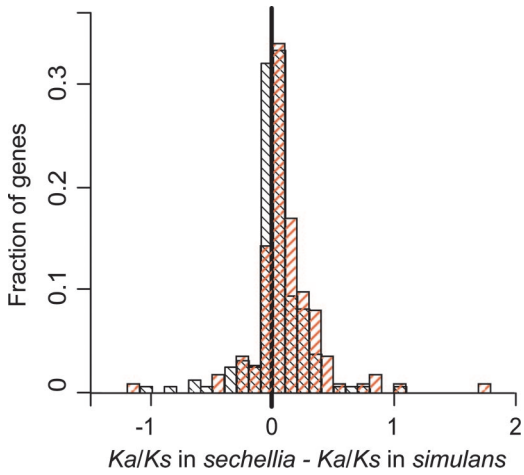


Fig. 2. Distribution of the difference in *Ka*/*Ks* between the *D. sechellia* and *D. simulans* lineages for each pair of orthologous receptor genes (red stripes) and control genes (black stripes). Note that both distributions are shifted to the right of the solid black line at zero (indicating that *Ka*/*Ks* tends to be higher in *D. sechellia* than in *D. simulans*), but that the distribution for receptors is shifted further to the right than that for controls.

remnants of five of the seven can be found in the *D. melanogaster* genome r4.1). I did not find any new duplicates in *D. sechellia*, although there may be one in *D. simulans* (ignored in this study). In total, the *D. sechellia* genome assembly has 60 *Or* genes and 65 *Gr* genes encoding 63 Or proteins and 73 Gr proteins by alternative splicing. I hereafter lump alternative splice forms together with independent loci and refer to them jointly as "genes."

Acceleration of Gene Loss in *D. sechellia*. Six of *D. sechellia*'s 63 *Or*s, and thirteen of its 73 *Gr*s exhibited lack-of-function (LOF) mutations that clearly render them pseudogenes (all of these LOF mutations were verified by direct resequencing; 15 additional genes exhibited LOF mutations that were found to be mistakes in the genome assembly). The majority of LOF mutations were large out-of-frame indels (≥5 bp), but three resulted from point mutations to premature stop codons, and two resulted from small out-of-frame indels (≤4 bp) [supporting information (SI) Table 5]. Data from *D. simulans* and *D. melanogaster* allowed me to infer by parsimony that all of these LOF mutations occurred along the *D. sechellia* lineage. In contrast, only two of the 73 *Gr* and none of the *Or* genes fixed LOF mutations along the *D. simulans* lineage (and eight receptors were deleted or fixed LOF mutations along the *D. melanogaster* lineage). Contingency tests showed that *D. sechellia* has fixed receptor pseudogenes (hereafter described as simply having "lost" receptor genes) at a more rapid pace than either *D. simulans* (χ^2, $P = 0.0001$; Table 1) or *D. simulans* and *D. melanogaster* combined (χ^2, $P = 0.0001$). This trend remained significant when *Or* and *Gr* genes were analyzed separately (Table 1). Note that a fraction of the LOF mutations that I consider "fixed" may actually be polymorphic, with functional alleles segregating in natural populations. This is more likely to be true for the *D. simulans* pseudogenes, however, because *D. simulans* is more polymorphic than *D. sechellia* (10), and because the *D. sechellia* LOF mutations were verified in two independent strains (including an outbred composite of five isofemale lines), whereas the *D. simulans* LOF mutations were verified in only one inbred strain.

Table 1. The number of receptor and control genes that became pseudogenes along the *sechellia* and *simulans* lineages

Genes	Status	*D. sechellia*, *n*	*D. simulans*, *n*	*P* value*
Or	Pseudo	6	0	
	Intact	57	63	0.012
Gr	Pseudo	13	2	
	Intact	60	71	0.003
Ors + *Grs*	Pseudo	19	2	
	Intact	117	134	0.0001
Controls†	Pseudo	9	2	
	Intact	181	188	0.03
Filtered *Ors*+*Grs*†	Pseudo	17	2	
	Intact	102	117	0.0003

Pseudo, pseudogene.
*P values are from χ^2 tests on each 2 × 2 table.
†Genes with indels ≤ 4 bp in *D. sechellia* are excluded (see *Materials and Methods*).

Elevation of *Ka*/*Ks* in *D. sechellia*. To investigate the pattern of molecular evolution of *Or* and *Gr* genes that remain intact in *D. sechellia*, I aligned each *D. sechellia* receptor to its orthologs from *D. simulans* and *D. melanogaster* (excluding the 22 genes that had fixed LOF mutations in one or more of the three taxa) and inferred lineage-specific silent substitution rates (*Ks*) and replacement substitution rates (*Ka*) for each branch in the unrooted three-species tree (raw data are included in SI Table 6). Both rates were significantly higher along the *D. sechellia* lineage than along the *D. simulans* lineage by a paired Wilcoxon rank sum test whether considering *Or* genes, *Gr* genes, or all receptors simultaneously (Table 3). Moreover, the increase in *Ka* was relatively greater than that in *Ks*, resulting in significantly higher *Ka*/*Ks* ratios in *D. sechellia* (paired Wilcoxon, $P < 0.0001$; Table 3). Fig. 2 illustrates the consistent nature of this effect: across all orthologous pairs, the difference in *Ka*/*Ks* between species (*Ka*/*Ks* of *D. sechellia* ortholog minus *Ka*/*Ks* of *D. simulans* ortholog) tended to be >0. The raw distributions of *Ka*/*Ks* inferred for receptor genes in *D. sechellia* and *D. simulans* (including the alignable portions of *D. sechellia* pseudogenes) appear in SI Fig. 3*A*.

Comparison with a Control Set of Randomly Selected Genes. To test whether gene deterioration and elevated *Ka*/*Ks* in *D. sechellia* are specific to the *Or*/*Gr* gene families or whether they reflect a genome-wide phenomenon, I annotated and repeated the above analyses in a control set of 190 randomly chosen genes. Nine genes had LOF mutations in *D. sechellia*, and only two had LOF

EVOLUTION

Table 2. Comparison of the rate of gene loss (proportion of genes that became pseudogenes vs. remained intact) among *Ors*/*Grs* with that among control genes within species

Species	Status	*Ors*+*Grs**	Controls*	*P* value†
sechellia	Pseudo	17	9	0.003
	Intact	102	181	
simulans	Pseudo	2	2	0.6
	Intact	117	188	

Pseudo, pseudogene.
*Genes with indels ≤ 4 bp in *D. sechellia* are excluded (see *Materials and Methods*).
†*P* values are from χ^2 tests on each 2 × 2 table.

mutations in *D. simulans* (rather than verify LOF mutations in control genes by direct resequencing, I used error rates estimated by resequencing of putative *Or*/*Gr* pseudogenes to filter the data; see *Materials and Methods*). Although this difference was marginally significant (χ^2, $P = 0.03$; Table 1), suggesting that *D. sechellia* has an elevated rate of gene loss in general, the trend was stronger among *Ors*/*Grs* than it was among the control genes (even when *Ors*/*Grs* were filtered in the same way as controls; Table 1). Moreover, a direct comparison of receptor genes to control genes within species showed that the rate of gene loss among *Ors* and *Grs* was significantly higher than that among controls within *D. sechellia* (χ^2, $P = 0.003$), but not within *D. simulans* (χ^2, $P = 0.6$, Table 2).

Ks, *Ka*, and *Ka*/*Ks* were higher along the *D. sechellia* lineage than along the *D. simulans* lineage for control genes as they were for receptors (Table 3; raw data are shown in SI Table 7). However, whereas *Ks* was elevated to the same degree in both gene sets (Wilcoxon, $P = 0.9$; Table 4), *Ka* and *Ka*/*Ks* were significantly more elevated among *D. sechellia* receptor genes than they were among *D. sechellia* control genes ($P \leq 0.0001$; Table 4). Fig. 2 illustrates this result. The mean difference in *Ka*/*Ks* between *D. sechellia* and *D. simulans* receptor gene orthologs is significantly larger (distribution is shifted to the right) than that between *D. sechellia* and *D. simulans* control gene orthologs. Also, although receptors had higher *Ka*/*Ks* ratios than control genes within both species (Wilcoxon, $P < 0.0001$ for *D. sechellia* and $P = 0.0008$ for *D. simulans*), the size of this effect was much greater in *D. sechellia* (Glass's Delta effect size = (receptor mean − control mean)/control SD = 0.62 compared with 0.18 in *D. simulans*). The raw distributions of *Ka*/*Ks* inferred for control genes in both species appear in SI Fig. 3*B*.

Table 3. Mean substitution parameters for *Or/Gr* genes and control genes along the *D. sechellia* and *D. simulans* lineages

Genes	Parameter	*sechellia*, mean	*simulans*, mean	Paired Wilcoxon *P* value
Ors	*Ks*	0.033	0.030	0.06
	Ka	0.005	0.003	<0.0001
	Ka/*Ks*	0.200	0.121	<0.0001
Grs	*Ks*	0.031	0.027	0.04
	Ka	0.007	0.004	<0.0001
	Ka/*Ks*	0.357	0.184	<0.0001
Ors+*Grs*	*Ks*	0.032	0.028	0.006
	Ka	0.006	0.003	<0.0001
	Ka/*Ks*	0.278	0.152	<0.0001
Controls	*Ks*	0.030	0.023	0.0001
	Ka	0.004	0.002	<0.0001
	Ka/*Ks*	0.145	0.117	0.001

The paired Wilcoxon *P* value tests the null hypothesis that the mean difference between the *D. sechellia* and *D. simulans* orthologs of each gene is zero.

Table 4. Mean difference in substitution parameters between the *D. sechellia* and *D. simulans* orthologs of *Or*/*Gr* or control genes

Parameter	*Ors*+*Grs*, mean difference	Controls, mean difference	Wilcoxon test *P* value
sec–sim Ks	0.0040	0.0073	0.9
sec–sim Ka	0.0030	0.0018	<0.0001
sec–sim Ka/*Ks*	0.125	0.028	0.0001

The Wilcoxon test *P* value tests the null hypothesis that the mean difference b/n *Or*/*Gr* orthologs is the same as that b/n control gene orthologs. *sec*, *sechellia*; *sim*, *simulans*.

Spatial Distribution of Amino Acid Substitutions Along Proteins. To test the null hypothesis that elevated *Ka*/*Ks* among *D. sechellia* receptor genes results from a complete relaxation of purifying selection on genes no longer of use to the fly, I examined the spatial distribution of amino acid substitutions along receptor proteins. In particular, I first derived the expected distribution along Or/Gr proteins that experience purifying selection by examining amino acid substitutions occurring along the lineages of *D. sechellia*'s generalist relatives (all branches except b in Fig. 1), because the vast majority of receptor genes have likely retained their functions along these lineages. A detailed description of the procedure used to derive this distribution can be found in *SI Methods* and SI Fig. 4. Briefly, I used an alignment of paralogous Or/Gr proteins to identify homologous sites from different proteins over which I then averaged rates of orthologous amino acid divergence (generating an overall estimate of divergence along the generalist lineages at each amino acid site in the alignment). I then reduced noise in the data by averaging these site-specific rates within a sliding window of 10 aa. This procedure resulted in a single spatial distribution of protein divergence along the generalist lineages for the set of aligned paralogs (Ors and Grs separately). In support of the idea that this distribution reflects purifying selection on important protein domains, mean divergence tended to be lower (for Ors but not for Grs) within putative transmembrane domains (negative correlation between the mean divergence rate of individual windows and the proportion of aligned paralogs with computationally predicted transmembrane domains in those windows; *Ors*: $r = -0.34$, $P = 0.017$; *Grs*: $r = -0.06$, $P = 0.7$). I then predicted (*i*) that the distribution of protein divergence along *D. sechellia* genes with the least elevated *Ka*/*Ks* should mirror this "generalist" distribution (because these genes are presumably also functional and under purifying selection) and (*ii*) that if the high *Ka*/*Ks* of *D. sechellia* orthologs with the most elevated ratios reflects a complete relaxation of purifying selection, then the spatial distribution of divergence for these genes should not mirror the expected. In accordance with the first prediction, mean protein divergence of *D. sechellia Ors* and *Grs* with the least elevated *Ka*/*Ks* was positively correlated with that of intact receptors in the generalist lineages across windows (*Ors*: $r = 0.66$, $P < 0.0001$; *Grs*: $r = 0.31$, $P = 0.008$; SI Table 8). In accordance with the second prediction, mean protein divergence of *D. sechellia Grs* with the most elevated *Ka*/*Ks* was not correlated with that of intact *Grs* in the generalist lineages across windows (for the 28 most elevated *Grs*, $r = 0.17$, $P = 0.09$; for the 10 most elevated *Grs*, $r = 0.02$, $P = 0.4$). The second prediction did not hold for *Ors*, however. The spatial distribution of protein divergence in the 29 *D. sechellia Ors* with the most elevated *Ka*/*Ks* did mirror that in the generalist lineages ($r = 0.43$, $P = 0.0001$). This was even true for a smaller subset of the 10 *D. sechellia Ors* with

the most elevated *Ka*/*Ks* ($r = 0.42$, $P = 0.0001$), suggesting that at least some of these genes are still useful on *D. sechellia*'s new host and that their high *Ka*/*Ks* ratios are not driven by a complete relaxation of purifying selection. SI Fig. 5 shows the locations of predicted transmembrane domains and the spatial distribution of amino acid divergence for Or/Gr proteins in the *D. sechellia* and generalist lineages.

Discussion

Insects rely heavily on their senses of smell and taste to recognize stimuli in their environment, such as resources, natural enemies, and mates. It is therefore likely that chemosensory genes are subject to novel evolutionary pressures when insects enter new niches during host shifts or host specialization events. I tested this hypothesis in *D. sechellia*, a host specialist that diverged from its generalist sister species *D. simulans* roughly half a million years ago, by comparing rates of gene loss and substitution along the *D. sechellia* lineage to those along the *D. simulans* lineage in these flies' entire repertoire of 136 olfactory and gustatory receptor genes. I found two striking patterns: (*i*) a surprisingly high fraction of *D. sechellia*'s receptors exhibited LOF mutations that clearly render them pseudogenes, resulting in a rate of gene loss 9–10 times higher than that in *D. simulans*; and (*ii*) those receptors that retain intact ORFs in *D. sechellia* have fixed amino acid replacement mutations at a consistently higher rate relative to silent mutations than their *D. simulans* orthologs (resulting in higher *Ka*/*Ks* ratios).

Low Effective Population Size. Several hypotheses may explain these observations. The first asserts that the low effective population size of *D. sechellia* (witnessed by reduced polymorphism and potentially attributable to a population bottleneck during initial colonization of the Seychelles or partial submergence of the Seychelles Bank ≈10,000 years ago) (12, 21) has weakened selection relative to drift and driven an increase in the frequency of slightly deleterious substitutions (including LOF substitutions in nonessential receptor genes, silent changes from preferred to unpreferred codons, and/or certain replacement substitutions). Indeed, others have already invoked this explanation for elevated *Ks* in *D. sechellia* (11). It is even possible to imagine a scenario in which low *Ne* could have driven the consistent increase in *Ka* relative to *Ks* that is responsible for *D. sechellia*'s high *Ka*/*Ks* ratios (e.g., a low level of initial codon bias could have made the slightly deleterious silent mutations less frequent than the slightly deleterious replacement mutations). However, if a prolonged bottleneck were the sole cause of gene deterioration and increased *Ka*/*Ks* among receptor genes, we would expect to see equivalent trends throughout the rest of the genome. Instead, the trends observed among receptor genes are significantly stronger than those observed in 190 randomly chosen genes. This result suggests that low effective population size may contribute to, but is not solely responsible for, the receptor-specific pattern.

Relaxed Purifying Selection and/or Positive Selection. Alternative hypotheses for the conspicuous increase in gene loss and *Ka*/*Ks* among receptor genes invoke changes in the selective environment experienced by *Or*s/*Gr*s along the *D. sechellia* lineage. First, a relaxation of purifying selection could explain the observed pattern. Because *D. sechellia* uses, and is required to recognize, only one type of fruit/microhabitat, it may no longer need many of the receptors used by its generalist ancestors to recognize and respond to a wide array of resources/microhabitats. If true, mutations that cause amino acid changes or premature stop codons in those superfluous receptors would no longer have been deleterious and would have become more frequent in *D. sechellia* than in *D. simulans*. Note, however, that the host range of specialized phytophagous insects appears to be shaped as much (if not more) by an aversion to nonhosts as by an attraction to hosts, and specialists tend to respond to a *wider* array of deterrent chemical stimuli than do generalists (22). Even so, relaxed purifying selection could have affected receptors involved in the assessment of stimuli that are associated with host plants but not directly involved in host selection (e.g., host-specific predators/pathogens).

Positive natural selection provides a second explanation for elevated *Ka*/*Ks* ratios and rate of gene loss among *D. sechellia* receptors. Because *D. sechellia* specializes on a novel host plant that is avoided by its close relatives (and presumably also by its generalist ancestor), amino acid replacement mutations that alter the selectivity and/or sensitivity of smell and taste receptors to this new host, to other aspects of the microhabitat provided by that host, or to aversive stimuli in nonhosts may have been favored. Moreover, just as *D. sechellia* receptors are challenged by a novel external environment, some may also be challenged by a novel internal environment. An *in vivo* electrophysiological examination of *D. sechellia* antennae showed that one type of sensillum (sensory hair housing the dendrites of olfactory receptor neurons and characterized by the specific *Or* genes expressed in those neurons) found on the antennae of all of *D. sechellia*'s eight closest relatives had effectively been replaced by additional "copies" of a different type of sensillum (housing neurons that express different *Or* genes) (20). This phenotypic change may have involved the expression of *Or*s in neurons/sensilla that they had not formerly experienced (e.g., containing a distinct suite of interacting proteins) and resulted in positive selection on these *Or*s for efficient function in a new cellular environment. In addition to explaining elevated *Ka*/*Ks*, positive natural selection may underlie *D. sechellia*'s elevated rate of gene loss. Selection may have favored LOF mutations disrupting receptors that put flies at a disadvantage in their new niche (e.g., mediate avoidance of *Morinda*, mediate attraction to non-*Morinda* resources, or occupy neurons that could be more "profitably" inhabited by other receptors) (23).

It is difficult to differentiate between the effects of relaxed purifying selection and positive natural selection by using rates of gene loss and substitution alone, and unfortunately, *D. sechellia*'s low level of polymorphism (10, 11) severely jeopardizes the utility of more powerful molecular population genetic methods that incorporate polymorphism data. As evidence of *D. sechellia*'s lack of variation, I found only 13 polymorphic sites in the process of resequencing >15 kb of partial *Or*/*Gr* coding regions from two independent strains.

One characteristic of the substitution rate that might at least help rule out the possibility that a complete relaxation of purifying selection underlies elevated *Ka*/*Ks* among *D. sechellia* receptors is its spatial distribution along proteins. For example, if purifying selection is completely relaxed, new mutations should fix at random positions, and the spatial distribution of amino acid substitutions should not mirror that in functional receptors. Interestingly, the distribution of amino acid substitutions along *Or* genes with the most elevated *Ka*/*Ks* ratios in the *D. sechellia* lineage did mirror that in functional *Or*s, suggesting that these genes still serve an important function on *D. sechellia*'s new host (note that it is also possible, although less parsimonious given the strength of the correlation, that the relaxation of constraint is recent enough that it has not had time to obscure the effects of purifying selection acting along the basal portion of the *D. sechellia* lineage, yet old enough that it has had time to significantly elevate *Ka*/*Ks*). The distribution of amino acid substitutions along *Gr* genes with the most elevated *Ka*/*Ks* in *D. sechellia*, on the other hand, did not mirror that in functional *Gr*s, leaving open the possibility that these *Gr* genes are indeed no longer useful to *D. sechellia*. It is also possible (and perhaps even probable), however, that paralogous *Gr* genes are so divergent in sequence and/or structural organization that a single expected distribution of *Ka* under purifying selection cannot usefully be derived (note that

amino acid divergence along *Gr*s in the generalist lineages was not correlated with the presence of putative transmembrane domains) and therefore that this analysis had little power to detect purifying selection on *D. sechellia Gr*s.

Similar Patterns in Human Olfactory Receptors. *D. sechellia* is not the only organism to be losing olfactory receptors at an accelerated rate. Humans also appear to be losing *Or*s more quickly than their closest relatives (24). And although there has been no comprehensive comparison of *Ka*/*Ks* between the human and chimpanzee lineages, at least a few genes have elevated ratios in humans (25). The spatial distribution of replacement substitutions along human *Or*s with high *Ka*/*Ks*, however, does not appear to be heterogeneous (26), and most studies find that human *Or* evolution is consistent with relaxed selective constraint in a species that no longer relies heavily on its sense of smell (but see refs. 25 and 27 for evidence of positive selection on a small number of *Or*s). Moreover, there is no evidence of altered evolutionary pressures on the few human gustatory receptors that have been studied (28).

***Gr* Evolution More Extreme than *Or* Evolution.** *D. sechellia Gr*s are, if anything, experiencing an even more dramatic change in their selective environment than are *D. sechellia Or*s. *D. sechellia* has lost 17.8% (13 of 73) of the *Gr*s present in its most recent common ancestor with *D. simulans*, whereas it has lost only 9.5% (6 of 63) of such *Or*s. Compared with the *D. simulans* reference values, the mean *Ka*/*Ks* of intact *D. sechellia Gr*s has increased by ≈94%, whereas the mean *Ka*/*Ks* of intact *D. sechellia Or*s has increased by only ≈67%. We know that *D. sechellia* uses its sense of smell to locate resources, but why might we expect host specialization to affect the evolution of gustatory receptor genes in this species? Many phytophagous insects use taste to assess plant quality and condition after locating a potential host (22). For example, sugar receptors may be used to assay nutritional value, and bitter receptors may be used to detect toxins, harmful bacteria, and plant secondary compounds with which harmful entities are associated. In addition to *D. sechellia* being a specialist, its host fruit contains compounds with antimicrobial activity (29, 30), which suggests that *D. sechellia* may be challenged by fewer food-borne pathogens than its generalist relatives (and therefore require fewer *Gr* genes to warn against these pathogens).

Does This Pattern Really Have Anything to do with Host Specialization? The cooccurrence of host specialization and rapid receptor evolution along the *D. sechellia* lineage does not prove that the former caused the latter. Nevertheless, it is clear that smell and taste receptor genes have experienced a unique selective environment along the *D. sechellia* lineage, and it makes sense that this should result from the dramatic ecological shift that the species has sustained. In support of this interpretation, at least one of the patterns documented here, accelerated gene loss, appears to be affecting other groups of genes thought to be involved in host adaptation in *D. sechellia* [e.g., odorant-binding proteins and genes involved in protein metabolism (I. Dworkin and C. Jones, personal communication); note that this may explain the marginally significant acceleration of gene loss in the control set from this study, because three of the only five *D. sechellia* control genes that both exhibited LOF mutations and have putative functions appear to be involved in protein metabolism]. Moreover, the idea that host specialization events are accompanied by the loss of traits that were important for survival and reproduction in the ancestral generalized niche traces back to the older observation that characters such as wings, eyes, and teeth are often reduced or lost in specialized groups (31).

Whether and to what degree the patterns described here will characterize smell and taste receptor genes in other insects undergoing host specialization events or shifts will likely depend on several factors, including (*i*) the extent of the difference between the ancestral and contemporary hosts (e.g., in chemistry and the associated community of natural enemies) and (*ii*) the intimacy of the relationship between the insect and its host (e.g., whether it feeds on its host as both a juvenile and an adult and whether it mates and/or rests on its host). Changes in the subset of receptor genes that are directly involved in host selection may additionally depend on the nature of the ecological change. As mentioned in *Relaxed Purifying Selection and/or Positive Selection*, specialized insects appear to have narrow host preferences largely because they are more sensitive than generalists to deterrent chemicals in nonhosts (22). One might therefore expect specialization on one of many former plants (or the abandonment of ancestral hosts in general) to be associated with amino acid substitutions that increase sensitivity to deterrents in abandoned hosts; i.e., elevated *Ka*/*Ks* (32, 33). Whereas loss of function may be restricted to insects that acquire/shift to novel hosts (because such losses provide one of many ways to disrupt receptor genes that respond to deterrents in the new hosts) or simply to the subset of receptors that are not directly involved in host selection.

D. sechellia has both specialized and shifted to a novel resource. That resource is quite distinct, at least chemically, from *D. sechellia*'s ancestral hosts, and it both nourishes the fly through all life stages and provides a site for resting and mating. Thus, although *D. sechellia* may use just a few *Or* and *Gr* genes to directly recognize *Morinda* fruit, in retrospect it is no surprise that the signatures of relaxed purifying and/or positive selection seem to be apparent in the *Or*/*Gr* gene superfamily as a whole. The accumulation of genome sequences for other insects of ecological interest should facilitate further research on the genomic signatures of host specialization and help determine the generality of the patterns observed in this study.

Materials and Methods

Gene Annotations. *D. sechellia Or* and *Gr* gene coding sequences were annotated by searching the publicly available CAF1 genome assembly (Broad Institute, Cambridge, MA, and http://rana.lbl.gov/drosophila) for orthologs of the *D. melanogaster* receptors described in ref (8). The following steps were repeated for each *D. melanogaster* receptor. First, I queried the *D. sechellia* genome with each protein, using TBLASTN (default parameters) (34). Second, I asked the program GeneWise (35) to identify an ortholog of the gene in the 40-kb region surrounding the best hit. Third, I filled any gaps in the assembly that fell within predicted coding sequences, using PCR. Fourth, I checked the predicted *D. sechellia* receptor by eye and made minor adjustments to ensure that the start, splice sites, and stop aligned as closely as possible to the *D. melanogaster* template. Fifth, I confirmed orthology by ensuring that the two genes were reciprocal best hits and verifying microsynteny (i.e., checking that the adjacent upstream and downstream neighbor of the *D. melanogaster* gene blasted to sequences upstream and downstream of the putative *D. sechellia* ortholog). I was also able to identify *D. sechellia* receptors without *D. melanogaster* orthologs by repeating the second and third steps (GeneWise and manual revision) for the remaining unexamined hits (e.g., second, third, forth best hits) from the original TBLASTN searches. *D. simulans Or* and *Gr* gene sequences were annotated by using a combination of the syntenic assemblies produced by the Drosophila Population Genomics Project (www.dpgp.org) and the mosaic assembly produced by the Washington University Genome Sequencing Center. The syntenic assemblies were used to annotate orthologs of *D. melanogaster* receptors by simply extracting sequences syntenic to the *D. melanogaster* genes, and the mosaic assembly was used to annotate *D. simulans* receptors

without orthologs in *D. melanogaster* with the same method used to identify such genes in the *D. sechellia* genome.

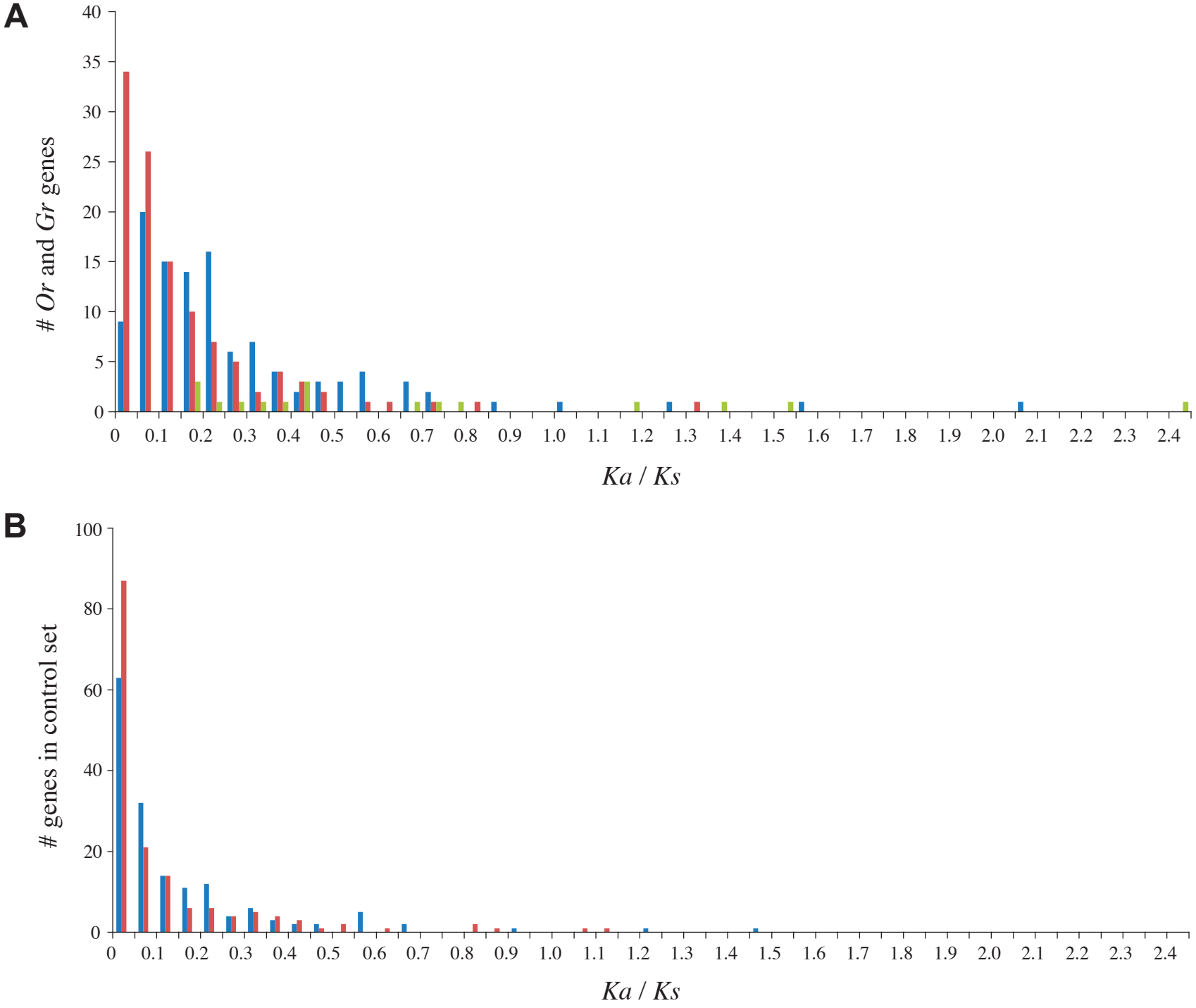
Pseudogene Analysis. To compare the rate of accumulation of pseudogenes along the *D. sechellia* and *D. simulans* lineages, I counted the number of receptors whose reading frames were disrupted by LOF mutations (that destroyed ≥20% of the original protein and ≥1 transmembrane domain) in one or the other species. LOF mutations came in three different forms: large out-of-frame indels (≥5 bp), small out-of-frame indels (≤4 bp), and point mutations to premature stop codons. All putative LOF mutations found in *D. sechellia* receptors were verified by direct resequencing from two different strains: the genome sequence strain (inbred nine generations, in culture since 1980) and a pool of five isofemale strains (individual strains in culture since 1985 and pooled in 1999); and all putative LOF mutations found in *D. simulans* receptors were verified by resequencing from one of the seven inbred genome sequence strains (w501). I then conducted a 2×2 contingency test on a table tallying the number of receptors in each species with verified LOF mutations (assumed to be pseudogenes) and the number of receptors with intact ORFs (assumed functional). Alternatively spliced transcripts [identified as such by orthology to alternatively spliced transcripts known from *D. melanogaster* (8)] were treated as independent genes, because LOF mutations were found only in regions specific to individual splice forms and never in shared exons.

Substitution-Rate Analysis. Replacement and silent substitution rates (*Ka*, *Ks*, *Ka*/*Ks*) specific to the *D. sechellia* and *D. simulans* lineages were estimated for each receptor by maximum likelihood, using a branch model implemented in the program PAML (36) (model = 1, NSsites = 0, *D. melanogaster* orthologs included as outgroups). I compared the mean *Ks*, *Ka*, and *Ka*/*Ks* ratio in the two species with paired Wilcoxon rank sum tests. These tests attempt to disprove the null hypothesis that the mean difference (in *Ks*, *Ka*, or *Ka*/*Ks*) between the two species across all orthologous gene pairs is zero. When testing for a mean difference in *Ka*/*Ks*, I excluded genes for which either species had *Ks* = 0.

Random Gene Analysis. I repeated the pseudogene and substitution rate analyses on a control set of genes randomly selected from those annotations of release 4.1 of the *D. melanogaster* genome that had empirical support (either an EST or cDNA). Orthologs for 235 such *D. melanogaster* genes were annotated in *D. sechellia* and *D. simulans* as described in *Gene Annotations* for receptor genes (except that microsynteny was not confirmed). I then conducted the pseudogene analysis on this set as described for receptors, except that I did not verify putative LOF mutations by resequencing. Instead, I used error rates estimated by resequencing of putative *Or*/*Gr* pseudogenes to infer the validity of these mutations. For *D. sechellia*, resequencing of receptors revealed that all 22 putative point mutations to premature stop codons and large out-of-frame indels (≥5 bp) were real, whereas 90% of 20 putative small out-of-frame indels (≤4 bp) reflected mistakes in the CAF1 assembly. I therefore categorized *D. sechellia* control genes with putative large indels as pseudogenes, and excluded from the analysis all control genes with small indels in *D. sechellia*. In *D. simulans*, direct resequencing always agreed with the syntenic assemblies, and I simply assumed that putative LOF mutations in *D. simulans* control genes were real. This filtering process resulted in a reduced set of 190 control genes ranging from 201 to 15,381 bp in length (mean = 1,675 bp; mean of receptors for comparison = 1,213 bp). I conducted the substitution-rate analysis on this reduced control set as described for receptors.

I thank Roman Arguello, Sergey Nuzhdin, Michael Turelli, David Begun, Corbin Jones, Mike Singer, and Matt Hahn for much discussion, advice, and exchange of ideas. I also thank three anonymous reviewers for comments that improved the manuscript substantially. Andrew Kern provided computational advice. This work was supported by a National Science Foundation predoctoral fellowship, and resequencing/computer time in the Nuzhdin lab was supported by National Institutes of Health Grant RO161773 (to Sergey Nuzhdin).

1. Dambroski HR, Linn C, Berlocher SH, Forbes AA, Roelofs W, Feder JL (2005) *Evolution (Lawrence, Kans.)* 59:1953–1964.
2. Hawthorne DJ, Via S (2001) *Nature* 412:904–907.
3. Jones CD (1998) *Genetics* 149:1899–1908.
4. Jones CD (2004) *Heredity* 92:235–241.
5. Sezer M, Butlin RK (1998) *Proc R Soc London Ser B* 265:2399–2405.
6. Hallem EA, Dahanukar A, Carlson JR (2006) *Annu Rev Entomol* 51:113–135.
7. Clyne PJ, Warr CG, Freeman MR, Lessing D, Kim J, Carlson JR (1999) *Neuron* 22:327–338.
8. Robertson HM, Warr CG, Carlson JR (2003) *Proc Natl Acad Sci USA* 100:14537–14542.
9. Vosshall LB, Amrein H, Morozov PS, Rzhetsky A, Axel R (1999) *Cell* 96:725–736.
10. Hey J, Kliman RM (1993) *Mol Biol Evol* 10:804–822.
11. Kliman RM, Andolfatto P, Coyne JA, Depaulis F, Kreitman M, Berry AJ, McCarter J, Wakeley J, Hey J (2000) *Genetics* 156:1913–1931.
12. Lachaise D, Sylvain, J-F (2004) *Genetica (Dordrecht)* 120:17–39.
13. Louis J, David JR (1986) *Acta Oecologica Oecologia Generalis* 7:215–230.
14. Tsacas L, Bachli G (1981) *Rev Fr Entomol* 3:146–150.
15. Farine, J-P, Legal L, Moreteau B, Le Quere, J-L (1996) *Phytochemistry* 41:433–438.
16. Legal L, Chappe B, Jallon JM (1994) *J Chem Ecol* 20:1931–1943.
17. Legal L, David JR, Jallon JM (1992) *Chemoecology* 3:125–129.
18. R'Kha S, Capy P, David JR (1991) *Proc Natl Acad Sci USA* 88:1835–1839.
19. Dekker T, Ibba I, Siju KP, Stensmyr MC, Hansson BS (2006) *Curr Biol* 16:101–109.
20. Stensmyr MC, Dekker T, Hansson BS (2003) *Proc R Soc London Ser B* 270:2333–2340.
21. Cariou ML, Solignac M, Monnerot M, David JR (1990) *Experientia (Basel)* 46:101–104.
22. Bernays EA, Chapman RF (1994) *Host-Plant Selection by Phytophagous Insects* (Chapman & Hall, New York).
23. Olson MV (1999) *Am J Hum Genet* 64:18–23.
24. Gilad Y, Man O, Paabo S, Lancet D (2003) *Proc Natl Acad Sci USA* 100:3324–3327.
25. Gilad Y, Man O, Glusman G (2005) *Genome Res* 15:224–230.
26. Gimelbrant AA, Skaletsky H, Chess A (2004) *Proc Natl Acad Sci USA* 101:9019–9022.
27. Gilad Y, Bustamante CD, Lancet D, Paabo S (2003) *Am J Hum Genet* 73:489–501.
28. Fischer A, Gilad Y, Man O, Paabo S (2005) *Mol Biol Evol* 22:432–436.
29. Hilgren JD, Salverda JA (2000) *J Food Sci* 65:1376–1379.
30. Viegas CA, Rosa MF, Sa-Correia I, Novais JM (1989) *Appl Environ Microbiol* 55:21–28.
31. Futuyma D, Moreno G (1988) *Annu Rev Ecol Syst* 19:207–233.
32. Menken SBJ, Roessingh P (1998) in *Endless Forms: Species and Speciation*, eds Howard DJ, Berlocher SH (Oxford Univ Press, New York), pp 145–156.
33. Olsson SB, Linn CE, Jr & Roelofs WL (2006) *Journal of Comparative Physiology A Neuroethology Sensory Neural and Behavioral Physiology* 192:289–300.
34. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D (1998) *Nucleic Acids Res* 25:3389–3402.
35. Birney E, Clamp M, Durbin R (2004) *Genome Res* 14:988–995.
36. Yang Z (1997) *Cabios* 13:555–556.


A
Or and Gr genes
0
5
10
15
20
25
30
35
40
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0 2.1 2.2 2.3 2.4
Ka / Ks
B
genes in control set
0
20
40
60
80
100
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0 2.1 2.2 2.3 2.4
Ka / Ks

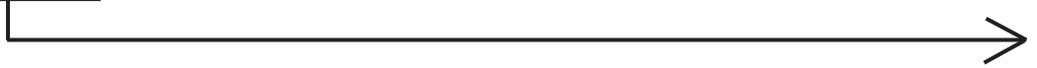
A

```
Or47b
 sechellia      MVDMMGISMFLQTALNLKLLCIEMRKLGDMEVSD-ERFHEEFCRVVRFHQHIIKLVGK
 simulans       MVDMVGISTFLQTALNLKLLCIEMRKLGEMEVSD-KRFHEEFCRVVRFHQHIIKLVGK
 melanogaster   MVDMVGISTFLQTALNLKLLCIEIRKLGDMEVSD-KRFHEEFCRVVRFHQHIIKLVGK
Or65a
 sechellia      VFENMGVSLFFELTSALRVLCIELRNLQELCLGDEDILYRELCRMTKFHQQIIILTDR
 simulans       VFENMGVSLFFELTSALRVLCIELRNLQELCLGDEDILYRELCRMTKFHQQIIILTDR
 melanogaster   VVENMGVSLFFELTSALRVLCIELRNLQELCLGDEDMLYRELCRMTKFHQQIILLTDR
Or65b
 sechellia      CIEGLSVFIYAEITFGIEVLCLELRHIHRHNHG-PKELRMETNRLVKLHQKIVEILDR
 simulans       CIEGLSVCIYAEITFGIEVLCLELRHIHRHNYG-PKELRMETNRLVKLHQKIVEILDR
 melanogaster   CIEGLSICIYAEITFGIEVLCLELRQIHRHNYG-LQELRMETNRLVKLHQKIVEILDR
```

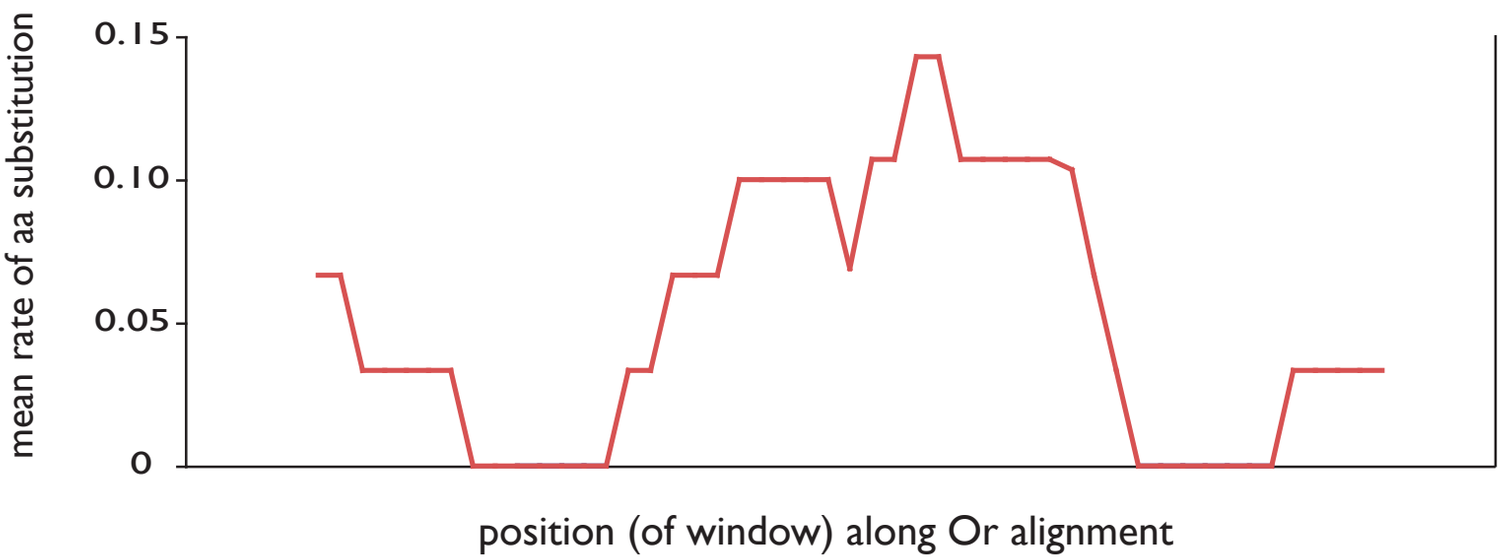
B

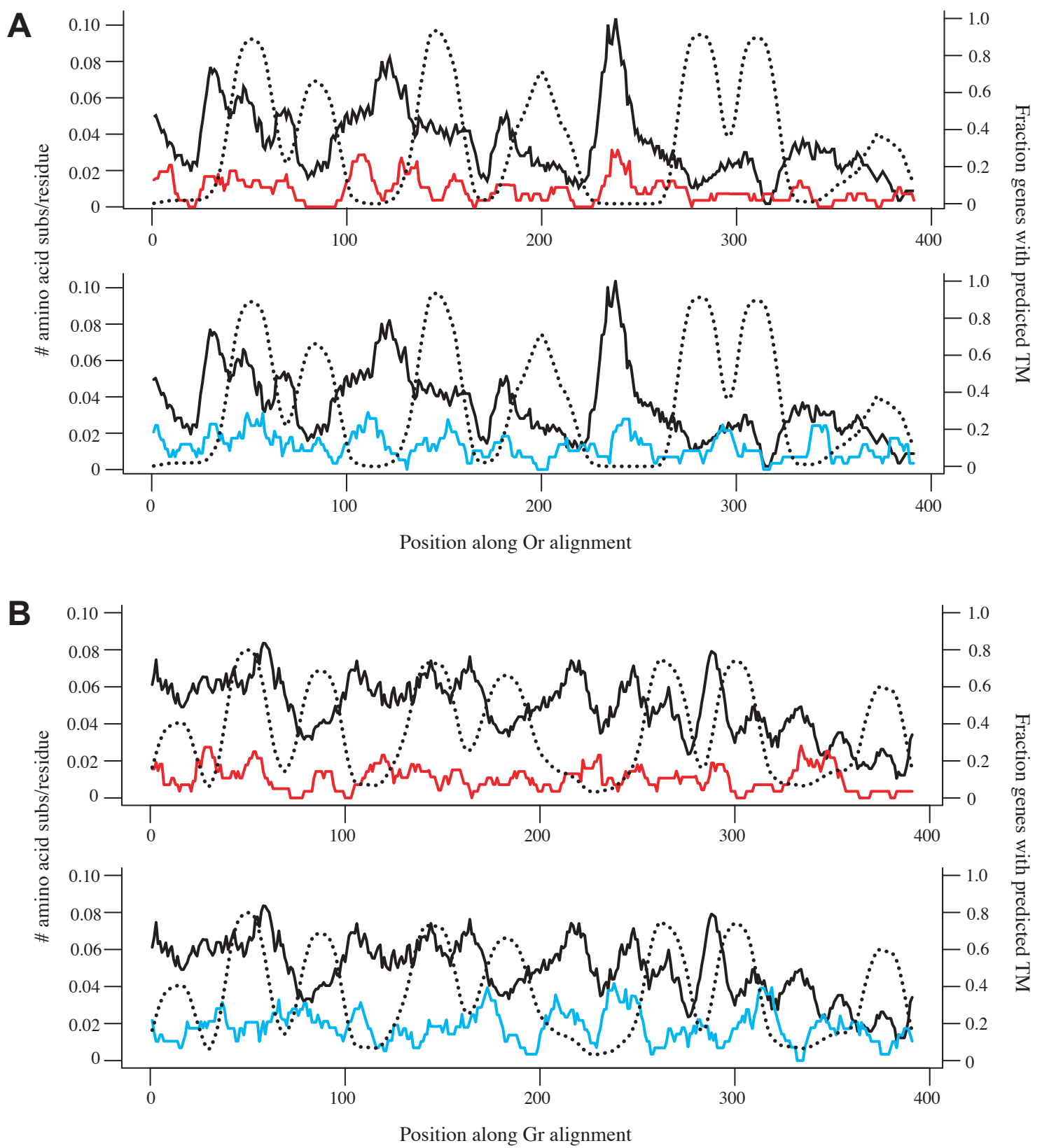
```
Or47b           MVDM■GIS■FLQTALNLKLLCIE■RKLG■MEVSD-■RFHEEFCRVVRFHQHIIKLVGK
Or65a           V■ENMGVSLFFELTSALRVLCIELRNLQELCLGDED■LYRELCRMTKFHQQII■LTDR
Or65b           CIEGLS■■IYAEITFGIEVLCLELR■IHRHN■G-■■ELRMETNRLVKLHQKIVEILDR

# sechellia subs 0000100110000000000000000000000010001000000000000000000000
# generalist subs 0100001000000000000000010100100000111000000000000000010000
# genes           3333333333333333333333333333333322333333333333333333333333
```



C




A
B
amino acid subs/residue
Fraction genes with predicted TM
0.10
0.08
0.06
0.04
0.02
0
1.0
0.8
0.6
0.4
0.2
0
0
100
200
300
400
Position along Or alignment
Position along Gr alignment