Polymorphism, Recombination and Alternative Unscrambling in the DNA Polymerase α Gene of the Ciliate *Stylonychia lemnae* (Alveolata; class Spirotrichea)

David H. Ardell,*,†,1,2 Catherine A. Lozupone†,3 and Laura F. Landweber†

**Department of Molecular Evolution, Evolutionary Biology Center, Uppsala University, SE-752 36 Uppsala, Sweden and* †*Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544*

Manuscript received February 26, 2003
Accepted for publication August 11, 2003

ABSTRACT

DNA polymerase α is the most highly scrambled gene known in stichotrichous ciliates. In its hereditary micronuclear form, it is broken into >40 pieces on two loci at least 3 kb apart. Scrambled genes must be reassembled through developmental DNA rearrangements to yield functioning macronuclear genes, but the mechanism and accuracy of this process are unknown. We describe the first analysis of DNA polymorphism in the macronuclear version of any scrambled gene. Six functional haplotypes obtained from five Eurasian strains of *Stylonychia lemnae* were highly polymorphic compared to Drosophila genes. Another incompletely unscrambled haplotype was interrupted by frameshift and nonsense mutations but contained more silent mutations than expected by allelic inactivation. In our sample, nucleotide diversity and recombination signals were unexpectedly high within a region encompassing the boundary of the two micronuclear loci. From this and other evidence we infer that both members of a long repeat at the ends of the loci provide alternative substrates for unscrambling in this region. Incongruent genealogies and recombination patterns were also consistent with separation of the two loci by a large genetic distance. Our results suggest that ciliate developmental DNA rearrangements may be more probabilistic and error prone than previously appreciated and constitute a potential source of macronuclear variation. From this perspective we introduce the nonsense-suppression hypothesis for the evolution of ciliate altered genetic codes. We also introduce methods and software to calculate the likelihood of hemizygosity in ciliate haplotype samples and to correct for multiple comparisons in sliding-window analyses of Tajima's *D*.

CILIATES are characterized by *nuclear duality*: they possess *micronuclei* serving primarily germ-line functions and *macronuclei* serving primarily somatic functions. Micronuclei divide meiotically before sexual exchange and are the developmental precursors of macronuclei, but are transcriptionally silent during vegetative growth. Macronuclei are the source of templates for transcription during vegetative growth but are destroyed during sexual reproduction and regenerated afterward. The genomic organization of the two types of nuclei is usually quite different. Macronuclear genomes in the class Spirotrichea that includes *Stylonychia lemnae* are partitioned into tens of thousands of different types of *gene-sized pieces*, so called because they are short, ~2 kb in average length, and contain only one or a small number of genes each. Gene-sized pieces are highly polyploid: tens of thousands of each type populate a macronucleus on average. Although they are capped on both ends by short telomeres, they lack centromeres and neither pair nor segregate in mitosis. Therefore, strictly speaking, gene-sized pieces are not "chromosomes." During vegetative growth the macronucleus divides, and gene-sized pieces are randomly partitioned to daughter cells in a process sometimes called *amitosis* (reviewed in Prescott 1994).

Micronuclear chromosomes are far larger and fewer in number than macronuclear gene-sized pieces. Micronuclear chromosomes are generally diploid. During sexual conjugation, haploid micronuclei are exchanged and preexisting macronuclei are destroyed. After conjugation, new macronuclei develop from a single micronucleus through a sequence of events that include extensive DNA rearrangements. During this process, called *macronuclear development*, ciliates eliminate large micronuclear chromosomal fractions (up to 95–98% of the *S. lemnae* micronuclear genome), excise the remainder into short fragments at chromosomal breakage sites, and amplify the remainder to macronuclear ploidy (Prescott 1994; Jahn and Klobutcher 2002). Prior to this, a polytenization of the micronuclear chromosomes generally occurs. DNA is eliminated not only between

Sequence data from this article have been deposited with the EMBL/GenBank/DDBJ Data Libraries under accession nos. AY243489–AY243496.

[1]*Corresponding author:* Microbiology, Box 596, Biomedical Center, Uppsala University, SE-751 24 Uppsala, Sweden.
E-mail: dave.ardell@icm.uu.se

[2]*Present address:* Institute for Cell and Molecular Biology, Department of Microbiology, Biomedical Center Box 596, Uppsala University, SE-751 24 Uppsala, Sweden.

[3]*Present address:* MCDB Department, University of Colorado, Boulder, CO 80309.

but also within genes, even from within protein-coding regions (Klobutcher *et al.* 1984). Proper gene functioning therefore requires the correct elimination of these internally eliminated sequences (IESs) by DNA rearrangements during macronuclear development.

IESs may be classified by their mechanism of excision. The types of IESs studied here are AT-rich stretches of DNA bounded by short repeats (2–18 bp; Prescott and DuBois 1996) with no common recognized motif. These types of IESs are excised by an unknown site-specific recombination mechanism that leaves one copy of each repeat at the point of splicing in the macronuclear gene. We have called these short repeats *pointer sequences*. We introduce the terms *headpointer* and *tailpointer* to distinguish the two pointer copies (see annotation notes in supplementary materials at http://www.genetics.org/supplemental/).

Macronuclear destined segments (MDSs) are the micronuclear segments that are spliced together and retained during macronuclear development. We also refer to these segments as MDSs after they have been assembled in macronuclear sequence. The MDSs of the majority of genes that have been studied are collinear in micronucleus and macronucleus. Prescott and Greslin (1992), in studying the micronuclear organization of actin in the stichotrich *Sterkiella nova* (formerly *Oxytricha nova*), made the remarkable discovery that MDSs in some stichotrich genes may be *scrambled*, *i.e.*, are not collinearly organized in the micronucleus and macronucleus—and may even be found on opposite strands of the same micronuclear chromosome.

The gene encoding the α subunit of DNA polymerase (DNA pol α) is the largest and most complex scrambled gene known (Hoffman and Prescott 1996). It is also the only known scrambled gene in which MDSs are found in two micronuclear loci separated by at least 3 kb, called the "major" and "minor" loci. Long PCR between these loci in Stylonychia failed to amplify any sequence, suggesting a distance of at least 3 kb, while direct sequencing in *O. nova* (now called *S. nova*) sets a minimum distance of 5 kb in that species (Hoffman and Prescott 1997a; Landweber *et al.* 2000). IES location, sequence, length, and number evolve rapidly in all orthologous stichotrich genes—scrambled and nonscrambled—that have been studied to date, and the number of MDSs and complexity of scrambling in scrambled genes appear to increase over evolutionary time (DuBois and Prescott 1997; Landweber *et al.* 2000; Hogan *et al.* 2001).

A schematic of the micro- and macronuclear organization of DNA pol α in *S. lemnae*, along with the terminology and notational devices introduced in this article, appears in Figure 1. The micronuclear version of MDS 6, only 12 bases long, has not been found. IES X-29 and MDS 30 share a homologous region of at least 199 bp that differ by seven mismatches or 3.5% (Landweber *et al.* 2000). It was speculated that this region may be a kind of "superpointer" that plays a role in bringing together the major and minor loci for splicing during macronuclear development.

We report here our study of macronuclear DNA polymorphism of DNA pol α from five Eurasian strains of *S. lemnae*. The primary aim of our study was to look for recombination between the major and minor loci to gauge their genetic distance. As this was the first analysis of polymorphism in a ciliate scrambled gene, we wanted to examine the effect of gene scrambling on molecular evolution and compare the level of polymorphism to that of genes in other species. We hoped we might also learn about the molecular mechanism of unscrambling, including how precise and regular the unscrambling process is, what kinds of errors occur, and at what rates. We looked for evidence of *alternative unscrambling*—whether a given micronuclear gene can unscramble in more than one way.

Two examples of alternative micronuclear processing have been studied in hypotrichs. In *O. fallax*, *S. histomuscorum* (formerly *O. trifallax*), and reportedly other stichotrichs, the 81-MAC locus occurs in three macronuclear versions, each known to come from the same micronuclear locus (Cartinhour and Herrick 1984; Herrick *et al.* 1987; Williams and Herrick 1991; Seegmiller *et al.* 1996). In both this example and the C1 + C2 + C3 + C4 example in *S. nova* (Klobutcher *et al.* 1984; Ribas-Aparicio *et al.* 1987, 1988), multiple products are made from the same precursor through alternative chromosomal fragmentation. Alternative DNA processing of any kind has not been reported in a scrambled gene.

We have found direct and indirect evidence of alternative unscrambling within paralogously duplicated MDSs, at least in MDSs 29 and 30 in the superpointer region. Thus, IES X-29 is in fact a functional duplicate of MDS 30. The major and minor loci have demonstrably different genealogies in our data, suggesting genetic recombination between them. Although this recombination could be developmental or meiotic in origin, because we have observed only two haplotypes per individual, we favor the latter interpretation. We report a reproducible example of unscrambling error. Finally we demonstrate that unscrambling directly contributes to the relatively large nucleotide diversity in ciliate macronuclear scrambled genes.

MATERIALS AND METHODS

Strains: Macronuclear DNA from five strains of *S. lemnae* was generously provided by the laboratory of Dr. Hans Lipps (Witten University, Witten, Germany). One strain ("RUS") was collected from St. Petersburg, Russia; one was from North Germany (Dornen, annotated as "NGR"); two were from South Germany (Entringen, denoted "SGR1" and "SGR2"); and the last was from Lake Federsee, in South Germany (denoted "FED"). The strains were derived from lab crosses of strains

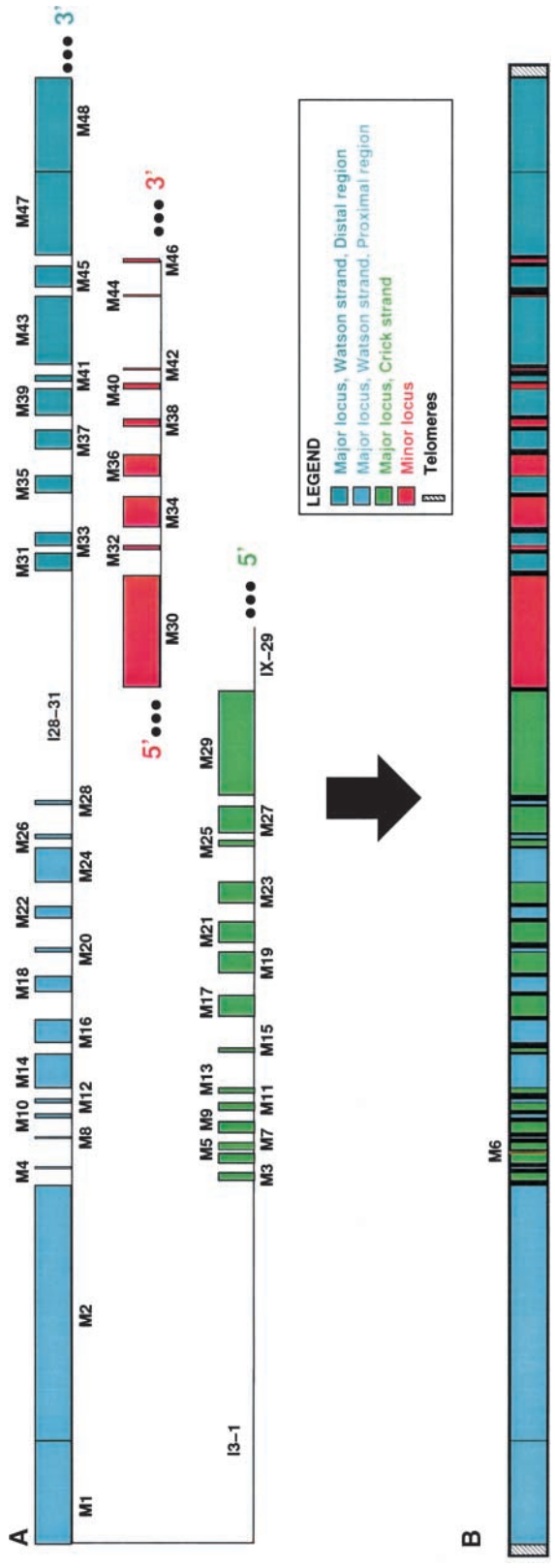

FIGURE 1.—Organization and nomenclature of the scrambled DNA polymerase α gene in *Stylonychia lemnae*. (A) Micronuclear organization from LANDWEBER *et al.* (2000), consisting of two loci, a major and a minor locus. MDSs (boxes) but not IESs (lines) are drawn approximately to scale. Only a few IESs are labeled to exemplify the annotation system used here. (B) Macronuclear organization after gene unscrambling and elimination of IESs during macronuclear development.

collected by the laboratory of Dieter Ammermann (University of Tübingen, Tübingen, Germany).

The numbers of generations after conjugation of the clones we analyzed were unknown. Two of the clones were closely related to the one from which we previously sequenced micronuclear and macronuclear versions of DNA pol α (Landweber *et al.* 2000). That clone was created from a cross between two strains from North and South Germany that were distinct from the clones we analyzed here.

PCR sequencing of haplotypes: A 4.8-kb region was amplified from macronuclear DNA of each strain, using the fewest number of cycles (usually 27–30) required to produce a band on an ethidium bromide-stained agarose gel. This approach reduced the chance of generating artifactual recombinants during PCR. The primers used were 5′-AGRAATAARAATT GAATAATRATACCGCG (forward) and 5′-GTCTTTGAGCA ACTGCCACATG (reverse). A hot-start PCR was used, followed by 94° denaturation (25 sec), 52° annealing (30 sec), and 72° extension (7 min). Final elongation was at 72° for 10 min.

PCR products were cloned using the TOPO-XL cloning kit (Invitrogen, San Diego). Clones were verified with colony PCR using the original primers. A small region from 2847 to 3870 (coordinates in GenBank sequence accession AF194338) from ~10 clones for each of the five strains was sequenced to type the number of haplotypes per strain. Two isolates of each haplotype were completely sequenced.

Bioinformatics: *Sequence and alignment manipulations:* Unless otherwise noted, the software used for this project were generalizable Perl scripts custom written by D. H. Ardell using BioPerl (v. 1.0; Stajich *et al.* 2002) and other freely available Perl libraries; these utilities in turn are freely and publicly available for inspection and use by contacting D. H. Ardell.

GenBank-feature-based subsetting and manipulation of alignments were done with SUBALIGN. SUBALIGN requires an input GenBank file and an alignment of the polymorphic data with the sequence in the GenBank file, in this case the previously published macronuclear sequence AF194338 (Landweber *et al.* 2000). This alignment was made with CLUSTALW (v.1.81; Thompson *et al.* 1994). Automated rescrambling of polymorphic alignments was done with RESCRAMBLE on the basis of GenBank files for micronuclear and macronuclear data annotated as described in the supplementary methods (http://www.genetics.org/supplemental/). Translation of gapped sequences according to the stichotrich genetic code was done with XL. Remapping ambiguous nucleotide states was done with SEQSTRICT. GenBank-feature-based postscript plotting was done with RENDER and other utilities were written to provide additional analysis of polymorphic sites.

Prediction of introns and initiation site: The true reading frame and initiation site of the DNA polymerase α gene has been the subject of some controversy (Landweber *et al.* 2000). We used patterns of synonymous and nonsynonymous polymorphism and divergence (this report; Hoffman and Prescott 1997; W.-J. Chang, D. H. Ardell and L. F. Landweber, personal communication), unpublished data concerning spirotrich intron acceptor and donor sites (D. M. Prescott, personal communication), and experimental data in another orthologous gene (W.-J. Chang, V. Addis, D. H. Ardell and L. F. Landweber, personal communication) to predict two introns and the true initiation site of the DNA pol α gene in *S. lemnae.* These results will be published separately.

Annotation of conserved regions: To relate the present data to functionally important domains of the polymerase α subunit protein, we used the annotation of protein-conserved regions from Mansour *et al.* (1994), modified from orthologous comparisons among bacteria and eukaryotes (Wong *et al.* 1988; Damagnez *et al.* 1991).

Analysis of mosaic structure in DNA polymorphism: We used two methods to detect mosaic structure in the Stylonychia macronuclear polymorphism data. TOPAL version 2.01b (McGuire and Wright 2000) slides a window across an alignment to calculate the difference in residual sum of squares of the left and right halves on the alignment on a least-squares distance tree estimated on one of the halves. This difference in residuals defines a "tree distance" D_{SS} that increases with differences in phylogenetic signal on the left and right halves of the window. The method allows the specific location of boundaries between regions with different phylogenetic signals. The significance of the values is calculated from a distribution of maximum D_{SS} calculated on 100 parametrically bootstrapped data sets, assuming no recombination.

TOPAL was run with a window size of 500 and an increment of 5 nucleotides. Because of the AT bias in the data (66% on average including all data), we used the "ML model" of DNA substitution, which is the "F84" model (see Felsenstein and Churchill 1996), allowing for heterogeneous equilibrium base composition and different rates of transition and transversion. The default transition-transversion ratio of 2 was used. Augmentation of the shell scripts used by TOPAL 2.01b was required; these Perl scripts are available by request from D. H. Ardell. As part of this modification, least squares was used for the simulated data sets instead of the neighbor-joining method, to match the calculation made with the actual data set (by default in TOPAL 2.01b, neighbor joining on the simulated data sets is forced regardless of the method used for the actual data set for reasons of time complexity). Although using least-squares in the parametric bootstrapping took longer than neighbor joining, on our data set the augmented analysis of simulated data was completed overnight on a 266-Mhz Pentium II.

We used RETICULATE (Jakobsen and Easteal 1996) to analyze compatibility among parsimoniously informative sites of the Stylonychia haplotypes. In compatibility analysis, a multiple pairwise "four-gamete test" is made over all nonsingleton variable sites, asking if all four combinations of two allelic states at two sites are present in the data. If all four combinations are present, the sites are said to be "incompatible"; otherwise they are said to be "compatible." RETICULATE was run on both the direct alignment and a "rescrambled" version of the alignment that restores the macronuclear polymorphism alignment data to an estimate of their micronuclear arrangement on the basis of known macronuclear data. This rescrambling was done automatically with the RESCRAMBLE program described above, with a small number of specific constraints added *post hoc.* For example, after initial exploration of the data, macronuclear pointer sequences were rescrambled as "tailpointers" as was most consistent with the data.

Gene genealogies and splits graphs: We calculated genealogies of the major and minor loci separately and after removing MDSs 29 and 30 and pointers. Neighbor-joining and parsimony trees were calculated in PAUP* (v.4.0b8; Swofford 2000), each with 100 bootstrap replicates. SPLITSTREE (v.3.1; Huson 1998) was used to calculate split decomposition on Hamming distances to assess and visualize support in the data for a unique tree.

Polymorphism statistics and neutrality tests: Polymorphic statistics and sliding-window analysis were calculated with a general-purpose command-line utility called PI (D. H. Ardell). Statistics were checked and examined for significance with DNASP (v.3.0; Rozas and Rozas 1999). To calculate the significance of Tajima's *D* for different regions of the data the infinite-sites coalescent simulation utility MS (Hudson 2002) was used, parameterized with Watterson's estimator as calculated under PI. The 10^5 sample ($n = 6$) genealogies were simulated and the Tajima's *D* values of the data (calculated with PI and

SUBALIGN) were compared with the 2.5th or 97.5th percentiles (two-sided test). Genealogies were simulated at constant population size and in the absence of recombination. DNASP (v.3.0; Rozas and Rozas 1999) was used to calculate the significance of windows in a sliding-window analysis under the assumption that the neutral distribution of Tajima's *D* is a β variate. This does not account for multiple comparisons. A utility called SCANMS was written by D. H. Ardell, which can be used to make parametric (coalescent) inferences of significant deviations of neutrality in a sliding-window analysis; the program calculates the maximum and minimum values over sliding windows of Tajima's *D* among a set of genealogies simulated by MS. SCANMS is available by contacting D. H. Ardell.

RESULTS

Stylonychia DNA pol α is highly polymorphic: Approximately 10 clones from each of the five strains were partially sequenced to screen the number of macronuclear haplotypes per strain. We failed to find more than two haplotypes per strain. Since one haplotype was shared between two apparently homozygous individuals, we obtained a total of seven macronuclear haplotypes. One of these seven contained nonsense and frameshift mutations and is believed to be inactive (see below). A total of 4734 nucleotides of overlapping sequence were obtained from the others, spanning from position 75 (in the 5′ leader) to position 4808 (before the end of the coding region) of the previously published macronuclear sequence AF194338 (Landweber *et al.* 2000). We denote these haplotypes by the geographic origin of their strains, hence *SGR1-RUS*, *NGR-a1*, *NGR-a2*, *SGR2-a1*, *FED-a1*, and *FED-a2*. We call the inactivated haplotype *SGR2-a2*.

Haplotypes *NGR-a1* and *SGR2-a1* were highly similar to AF194338. Excluding ambiguities, the per-site pairwise differences of these haplotypes from AF194338 were $\hat{\pi} = 2.37 \times 10^{-3}$ and $\hat{\pi} = 9.91 \times 10^{-3}$, respectively, while the average of other haplotypes ($\hat{\pi} = 1.58 \times 10^{-2}$) with AF194338 was comparable to the estimated nucleotide diversity of the sample ($\hat{\pi} = 1.74 \times 10^{-2}$). Because the haplotypes that contributed to the AF194338 sequence were not present in our sample, we could not directly use the method of Clark (1990) to determine them. Thus we excluded AF194338 in the following analyses where exact haplotype knowledge is assumed. Yet the close relatedness of the previously analyzed and present strains justified the use of published micronuclear data as a reference to interpret our results.

There was only one segregating insertion/deletion (indel) in the data, a singleton. A haplotype from the Federsee strain, *FED-a2*, had a single-base deletion of an A at position 210 relative to the others and to AF194338. This deletion lies in the first inferred intron (our unpublished data) and as such we deem it to be a true segregating indel.

Two strains, RUS and SGR1, from St. Petersburg, Russia and Entringen, South Germany, respectively, were apparently homozygous. The single haplotypes each presented were identical up to a small proportion of sequencing ambiguities ($3/4734 = 0.6\%$ in *RUS* and $5/4734 = 0.1\%$ in *SGR*). The positions of these ambiguities were completely nonoverlapping.

This apparent homozygosity and haplotype sharing may actually be hemizygosity and/or due to experimental error. Macronuclear hemizygosity may arise naturally through the stochastic partitioning of macronuclear molecules during growth or through fluctuations in the number of micronuclear chromosomes that develop in the macronucleus (reviewed in Prescott 1994). Micronuclear chromosome number can fluctuate widely in natural populations and within lineages over generations of laboratory culture (Ammermann 1987), and experimental hemizygosity of essential chromosomes can be viable (Ammermann 1965).

To examine the relative likelihood of hemizygosity we used the results of Karlin and McGregor (1972) concerning the probability of sampling an allele configuration under the infinite-alleles model (Ewens 1972) to calculate the likelihood that we truly observed the same haplotype more than once given the observed level of genetic variation. Following Equation 3.18 on p. 128 in Hartl and Clark (1989), the likelihood $L(m|\theta_w, n, k, \{a_i\})$ of observing an allele m times, $m \geq 1$, in a sample of size $n + m$ containing k alleles, given Watterson's estimator θ_w, is

$$L(m|\theta_w, n, k, \{a_i\}) = \frac{(n+m)!(\theta_w)^k}{\theta_w(\theta_w+1)\ldots(\theta_w+n+m-1)} \prod_{i=1}^{(n+m)} \frac{1}{i^{a_i} a_i!},$$

where a_i is the number of alleles present exactly i times in the sample. We use Watterson's estimator because it is invariant to allelic copy number and has relatively low variance. We implemented this calculation in a program called KM.

In our data, for the entire region we sequenced, the value of Watterson's estimator was approximately $\theta_w \sim 83$. With this value, the likelihood that we actually sampled *SGR1-RUS* only once in a sample of size six was >4 times that of sampling it twice (in a sample of size seven), >71 times that of sampling it three times, and 1000 times that of sampling it four times (assuming all other alleles are sampled once). Furthermore, when all other alleles are singletons ($k = n + 1$), we can show that θ_w must be less than $(k^2 - k)/2$ for the likelihood of sampling an allele twice (homozygosity) to exceed that of sampling it once (hemizygosity). Also, under these assumptions, for any value of θ_w, the maximum-likelihood copy number cannot exceed two.

We conclude that either one of the two strains was hemizygous or biased amplification of *SGR1-RUS* occurred, and that the apparent haplotype sharing may have been due to experimental error. We assumed that we sampled only six functional alleles, which we refer to as haplotypes because we have not strictly demonstrated

TABLE 1

Diversity statistics and nonneutrality tests in regions of DNA pol α

	Len (gf)[a]	S[b]	s[c]	$\hat{\theta}_W \pm$ SE[d]	$\hat{\pi} \pm$ SE[e]	Tajima's D
All	4734 (4710)	190	0.0403	0.0177 ± 0.0083	0.0174 ± 0.0088	−0.0893
Major locus[f]	3516 (3287)	144	0.0438	0.0192 ± 0.0091	0.0185 ± 0.0094	−0.2191
Watson strand	2936 (2709)	122	0.0450	0.0197 ± 0.0094	0.0192 ± 0.0098	−0.1572
Proximal region	1686 (1626)	73	0.0449	0.0197 ± 0.0094	0.0191 ± 0.0098	−0.1822
Distal region	1250 (1083)	49	0.0452	0.0198 ± 0.0096	0.0195 ± 0.0100	−0.1174
Crick strand	580 (578)	22	0.0381	0.0167 ± 0.0084	0.0152 ± 0.0082	−0.5435
Minor locus[g]	260 (259)	14	0.0541	0.0237 ± 0.0124	0.0239 ± 0.0133	0.0687
MDS 29	350 (348)	6	0.0172	0.0076 ± 0.0045	0.0096 ± 0.0057	1.5369*
MDS 30	380 (378)	20	0.0529	0.0232 ± 0.0118	0.0226 ± 0.0122	−0.1610
Pointers	429 (426)	5	0.0117	0.0051 ± 0.0031	0.0059 ± 0.0037	0.8781
CDS[h]	4575 (4523)	179	0.0396	0.0173 ± 0.0082	0.0173 ± 0.0087	−0.0161
Conserved[i]	729 (729)	32	0.0439	0.0192 ± 0.0095	0.0183 ± 0.0096	−0.3085
Introns	119 (116)	8	0.0690	0.0302 ± 0.0170	0.0247 ± 0.0151	−1.0718
5′ leader	72 (71)	3	0.0423	0.0185 ± 0.0128	0.0141 ± 0.0109	−1.2331**

*$P < 0.10$; **$P < 0.05$.
[a] Gap- and ambiguity-free length in nucleotides.
[b] Number of segregating sites.
[c] Proportion of segregating sites.
[d] Watterson's estimator per site, standard error assuming no recombination (Tajima 1993).
[e] Nucleotide diversity, standard error assuming no recombination (Tajima 1993).
[f] Excluding MDS 29 and pointers. See Figure 1 for definition of region names.
[g] Excluding MDS 30 and pointers.
[h] Protein-coding region.
[i] Conserved regions.

allelism. Except for the calculation of polymorphism statistics and neutrality tests, most of our results are indifferent to the true sample size (*i.e.*, including an extra copy of *SGR1-RUS*). For results comparing nucleotide diversity to the number of segregating sites, because we are emphasizing an excess nucleotide diversity, this treatment of the data is conservative.

The overall level of polymorphism measured in nucleotide diversity ($\hat{\pi}$) was 0.0174 ± 0.0088 with standard error assuming no recombination as calculated according to Tajima (1993). That of the protein-coding region was similar (Table 1). The only other available polymorphism data for a DNA pol α ortholog are six sequences from a protein-coding partial segment of *dnaE* from the bacterium *Vibrio cholerae* (Byun *et al.* 1999); there the nucleotide diversity is quite similar, at 0.014 ± 0.0075, as calculated by the program PI. The assumption of no recombination in calculating standard errors is conservative in this comparison. The standard error assuming free recombination for the ciliate CDS (protein-coding region) is 0.0013.

In comparison with the rough average nucleotide diversity of Drosophila and human protein-coding genes, the nucleotide diversity of DNA pol α in both Vibrio and Stylonychia is 2 and 10 times greater, respectively (Li and Sadler 1991). A rough average of nucleotide diversity in mostly X-linked coding and noncoding regions of humans is 8.1×10^{-4} (range 1.9×10^{-4} to 1.78×10^{-3}; Przeworski *et al.* 2000). Although the Esterase-6 locus in *Drosophila simulans* has a value of nucleotide diversity (2.2×10^{-2}) comparable to that we observed for DNA pol α in Stylonychia, this is considered hypervariable (Karotam *et al.* 1995). The average nucleotide diversity in autosomal *D. simulans* protein-coding genes is 7.83×10^{-3}, the highest average in Drosophilids (Moriyama and Powell 1996). Among the few surveys of polymorphism in ciliates available, the levels of polymorphism in the *SerH* alleles of the holotrichous ciliate *Tetrahymena thermophila* are higher than those reported here (averaging ~0.2), but the extracellularity, low amino acid complexity, and high, probably neutral, nonsynonymous variation of its encoded protein make this case probably exceptional (Gerber *et al.* 2002).

A naturally misunscrambled haplotype: The exceptional haplotype *SGR2-a2* was reproducibly obtained from one of the South German (Entringen) strains, SGR2. This haplotype had at least three indels that could be explained by incomplete unscrambling: (1) MDS 6, the 12-base MDS of unknown micronuclear location, had not been spliced in; (2) the first 5′ 12 bases of MDS 30 were missing; and (3) MDS 32 had not been spliced in, and instead IES 31-33 had been retained (Figure 2).

The failure in *SGR2-a2* of MDS 32 to be correctly spliced and of IES 31–33 to be excised may have been caused by two mutations at sites 3355 (G→T) and 3361 (C→A) in pointer 31–32. *SGR2-a2* also carried a unique base difference at site 3381 from the other haplotypes in pointer 32–33. As stated above and reinforced by

FIGURE 2.—Subalignment of the region around MDS 32 showing incomplete splicing in the *SGR2-a2* haplotype. "Major" and "Minor" refer to the major and minor loci micronuclear sequence data from another but closely related strain (LANDWEBER *et al.* 2000, see Figure 3 legend for accessions). Highlighted differences and site indexing are relative to AF194338, except for the IES 31–33 state (in color) where differences are relative to the major locus.

other results below, strain SGR2, from which this haplotype was taken, is closely related to the strain whose micronuclear sequence was previously published (LANDWEBER *et al.* 2000). The latter strain exhibits exactly the same micronuclear base difference between HP32-33 and TP32-33 at site 3381. Moreover, the state exhibited in *SGR2-a2* is identical to the headpointer state HP32-33 of the other strain, which is the state genetically linked to IES 31–33 in that strain (Figure 2). These results reinforce our interpretation that splicing of M32 and excision of I31–33 had not occurred in *SGR2-a2*.

Although the 12-base deletions of M6 and at the 5′ end of M29 did not introduce frameshifts in the gene, many other small frameshifting deletions and insertions and a large number of unique nonsynonymous or nonsense-inducing differences that were unrelated to known locations of splicing were present in *SGR2-a2*. In relation to the allelic consensus, the *SGR2-a2* coding region presented two frameshifting (1 base) insertions, four (1–2 base) deletions, 2 nonsense mutations, 25 nonsynonymous mutations, and 36 synonymous mutations. Noncoding regions contained 3 point mutations and one single-base deletion. The foregoing treats each mutation as an independent occurrence, that is, tallying synonymous and nonsynonymous substitutions disregarding the actual reading frame of the haplotype but rather analyzing each in the reading frame of the other haplotypes. We did not observe a tendency for compensatory frameshift mutations to restore frame. Rather, an abundance of stop codons appeared throughout its length. It was therefore surprising that among unique point mutations, in the original reading frame, nearly 1.5 times as many point mutations were synonymous as nonsynonymous. The relative base composition of the data was (A = 0.378, C = 0.154, T = 0.280, and G = 0.188). The expected ratio of synonymous to nonsynonymous mutations in a neutrally evolving sequence of this composition is 0.251, assuming infinite codons as calculated with the program NONSYN (D. H. Ardell), whereas the value would be 0.323 if all nucleotides were equally frequent. The observed ratio (1.44) is significantly different from its expectation (continuity-corrected $\chi^2 = 55.3$, $P < 0.001$).

In summary, some of the sequence disruptions in *SGR2-a2* corresponded to MDS-IES boundaries, suggesting partial or erroneous unscrambling. Many more did not correspond to MDS-IES boundaries. These suggested that *SGR2-a2* is inactivated. We therefore excluded this haplotype from our subsequent analyses of recombination and polymorphism. However, the excess of synonymous mutations in this haplotype demands further explanation, and we turn our attention to it again below.

Direct evidence of a paralogous MDS: We have scrutinized the sequence differences between IES X-29 from the previously published major locus (GenBank accession AF194337) and MDS 30 from the minor locus (GenBank accession AF194336) and show in Figure 3 evidence that IES X-29 is in fact an actively translated and transcribed paralogous copy of MDS 30.

If the major and minor loci were spliced together anywhere within IES X-29 as we know it rather than at its beginning, no disruption to the protein sequence would occur other than a deletion of the first five amino acids encoded by MDS 30 and a single conservative replacement of lysine with arginine caused by a transition at site 3022. All other differences between MDS 30 and IES X-29 are synonymous. Suggestively, AF194338 carries the IES X-29 state at site 3022 rather than the MDS 30 state, but the bulk of evidence suggests that AF194338 MDS 30 indeed derives from the minor locus MDS 30 in that strain.

The polymorphism data in the region are consistent with alternative unscrambling of MDS 30 and IES X-29, assuming that the corresponding micronuclear data from LANDWEBER *et al.* (2000) apply to these strains. Indeed, the previously studied strain is closely related to the NGR and SGR2 strains, the latter being the one with the misunscrambled haplotype *SGR2-a2*. Also, the level of variation (as measured in proportion of segregating sites) in this region is low (Table 1 and Figure 4). Thus, in the following paragraphs we assume that the previously published data apply.

Some aspects of the polymorphic data could be explained as alternative insertion of MDS 30 and IES X-29 into the macronucleus. Two haplotypes involving three

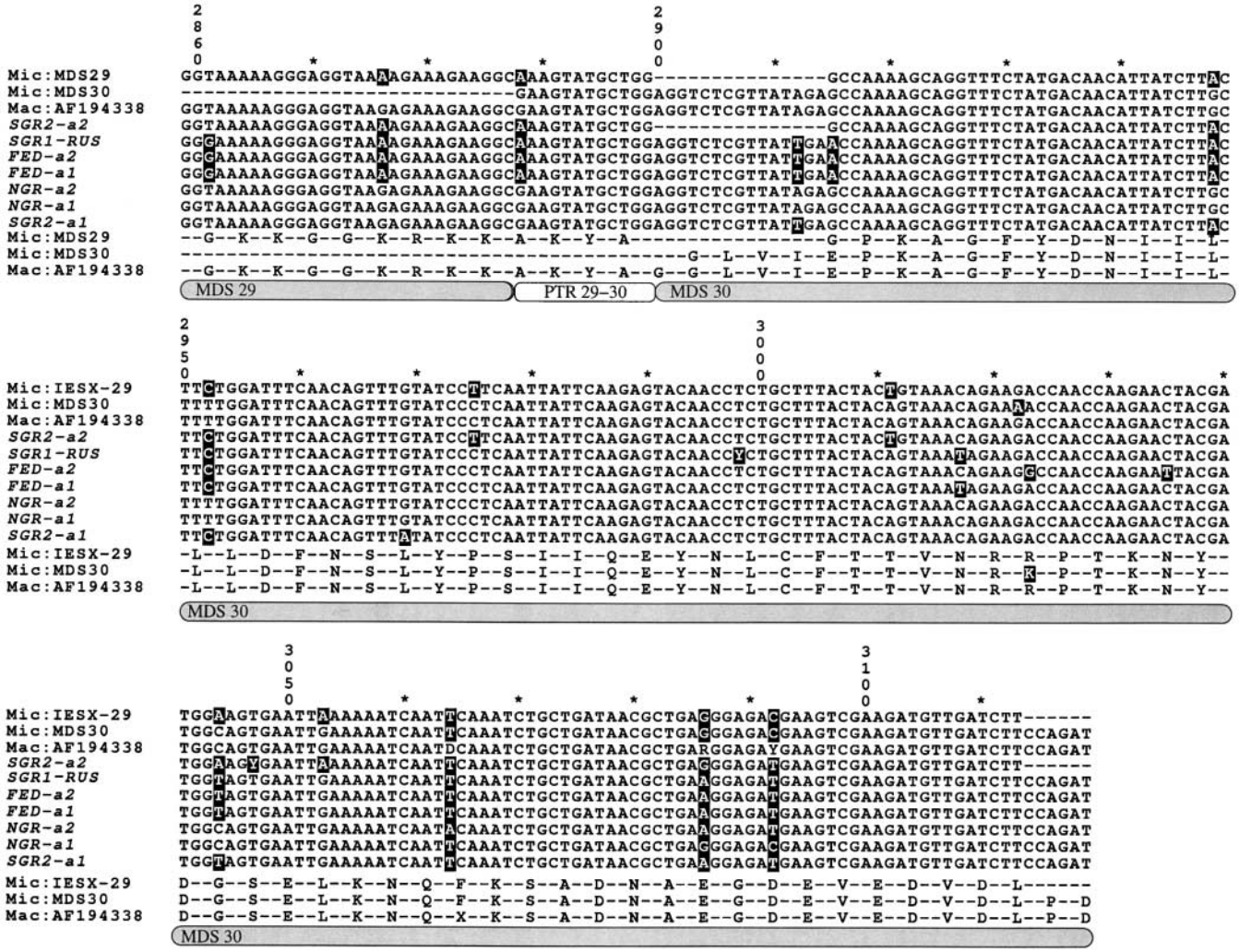

FIGURE 3.—Evidence that IES X-29 is a paralogous, redundant, and alternatively unscrambled MDS 30. An alignment of previously published micronuclear (Mic MDS 29 and IES X-29, accession no. AF194337; Mic MDS 30, accession no. AF194336) and macronuclear data (accession no. AF194338) with new polymorphic data at the transition region between MDSs 29 and 30 is shown. The entire known extent of IES X-29 is shown, while the sequence upstream of the minor locus MDS 30 is unknown. Differences from AF194338 in either of the micronuclear sequences (top two rows), their translations (bottom rows), or any of the polymorphic sequences are highlighted. All but one of the differences between IES X-29 and MDS 30 are both synonymous and polymorphic in this region. At site 3022, a nonsynonymous difference occurs; here AF194338 matches IES X-29 rather than MDS 30. The misunscrambled haplotype *SGR2-a2* additionally matches IES X-29 in AF194336 at sites 2975, 3011, and 3053.

synonymous segregating sites (2888, 2948, and 2952) could be explained as belonging to either the MDS 30 or the IES X-29 states as defined by AF194336 or AF194337, respectively, with the majority of the data by this interpretation coming not from MDS 30 but rather from IES X-29. Site 2888 is at the beginning of pointer P29–30, where the base differences had previously been observed in micronuclear data (LANDWEBER *et al.* 2000). By this interpretation *SGR2-a2*, *SGR1-RUS*, *FED-a1*, and *FED-a2* could be explained as being in the IES X-29 state, while *NGR-a1*, *NGR-a2*, and AF194338 would be in the MDS 30 state. *SGR2-a1* had regions in both states; for instance, it carried the IES X-29 state at sites 2948 and 2952, but the MDS 30 state at the more downstream sites 2975, 3011, and 3053.

The previously published sequence data provide evidence that not only MDS 30 but also MDS 29 exists in at least two alternatively unscrambled copies. A large number of discrepancies exist between the previously published micro- and macronuclear versions of MDS 29, and these differences are also reflected in a telltale way in our polymorphism data (supplementary Figure 1; http://www.genetics.org/supplemental/). Figure 3 shows one of these differences, at site 2876. There are 12 other such unambiguous mismatches between macronuclear and micronuclear MDS 29 from the same strain (AF194338 and AF194337). We cannot easily explain this discrepancy through allelic micronuclear differences because the experimental procedure that was used generally detects these as ambiguities (although

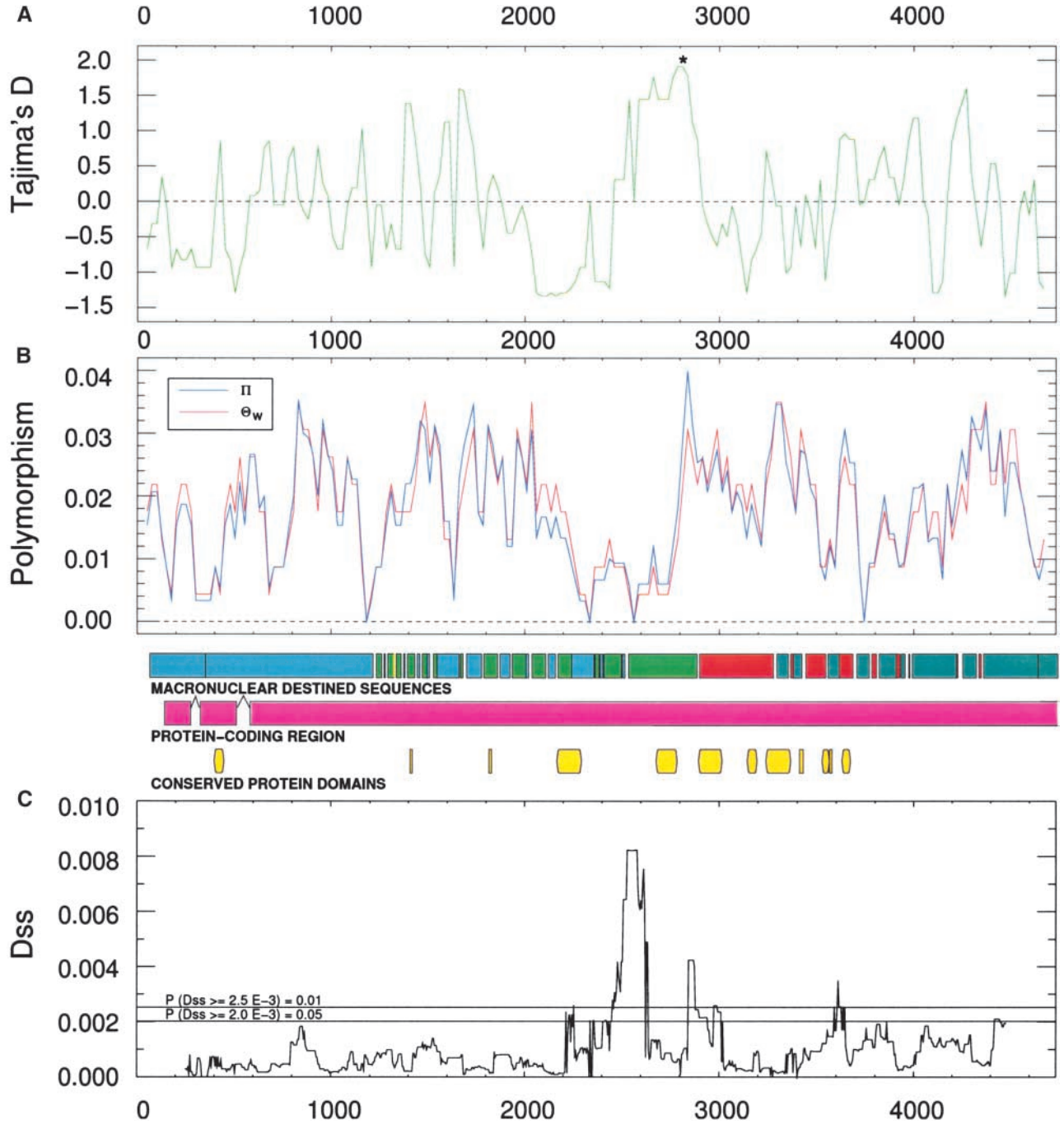

FIGURE 4.—Sliding-window analyses of recombination and polymorphism. Alignment of numbers with the feature legend is exact. The window sizes and increments are 100 and 25 nt (A and B) and 500 and 5 nt (C), respectively. (A) Tajima's *D* statistic. The starred peak is discussed in the text. (B) Two measures of polymorphism, nucleotide diversity π and per-site θ_w. (C) Mosaic structure in the data, plotting the D_{SS} statistic as computed with TOPAL 2.01b. The 95 and 99% significance levels (by parametric bootstrapping) are indicated.

PCR bias cannot be excluded). However, in 9 of the 12 mismatching sites AF194338 and all polymorphic data *except the misunscrambled haplotype SGR2-a2* are fixed in a state other than the observed micronuclear MDS 29 from AF194337 (supplementary Figure 1, http://www.genetics.org/supplemental/). The data therefore suggest that MDS 29 also exists in a paralogous micronuclear form, and this as yet unobserved paralogous MDS 29 is the source of most of the macronuclear data that have been observed to date.

Thus, the misunscrambled haplotype *SGR2-a2* complements this evidence for paralogous alternatively unscrambled MDSs 29 and 30. The association of variability with a haplotype already identified as misunscrambled strongly implicates the unscrambling process in generating macronuclear variability generally. *SGR2-a2* is the

only haplotype that matches micronuclear MDS 29 in AF194337 perfectly, and furthermore, *SGR2-a2* matches IES X-29 at all three of the mismatching sites in Figure 3 that are not otherwise reflected in polymorphic data (sites 2975, 3011, and 3053). Together with the 12-base deletion just after pointer 29–30, shared by micronuclear AF194337 and *SGR2-a2*, these results strongly suggest that both MDSs 29 and 30 in *SGR2-a2* derive mainly from a major locus sequence very similar to AF194337.

The previously published sequence AF194338, contrastingly, appears to derive almost entirely from the minor locus at MDS 30 and probably also at MDS 29. The minor locus is the most likely location for the hypothetical paralogous MDS 29, since homology in this region for MDS 30 is already known, and the sequence upstream of MDS 30 in the minor locus is unknown. Under this hypothesis MDSs 29 and 30 exist in two extensive and directly linked paralogous copies in both the major and minor loci, and one or more splicing events can occur at variable locations within these paralogous copies without disrupting the final gene. That is, splicing events between the major and minor loci within MDSs 29 and 30 may be multiple and they may vary in location from individual to individual. Note that if and when the major locus IES X-29 is the source for macronuclear MDS 30, 12 bases at the 5′ end of MDS 30 must be spliced in from somewhere else, but the precedent of a short splice from another part of the genome already exists in MDS 6.

To conclude, all the differences between IES X-29 and MDS 30 are synonymous and are reflected in polymorphism data, the misunscrambled haplotype, or both, suggesting that IES X-29 is an additional, paralogous and alternatively unscrambled MDS 30. Also, the majority of the data at MDS 29 seem to reject a major locus origin for this part of the gene in all but the misunscrambled haplotype *SGR2-a2*.

Intragenic mosaic structure: We can confirm the hypothesis of an alternative, dominantly minor locus origin of MDS 29 by analyzing the mosaic structure of the polymorphism data. Originally, our motivation was to check for genetic linkage of the major and minor loci. According to the Landweber *et al.* (2000) model for the micronuclear structure of this locus (*cf.* Figure 1), a significant signal of genetic recombination between MDSs 29 and 30 would have been expected if the major and minor loci were unlinked. According to the current model, this signal would occur at the beginning of MDS 29 rather than its end.

Analysis with an augmented version of TOPAL 2.01b (Figure 4C, see materials and methods) showed significant mosaic structural signals between MDSs 29 and 30, between MDSs 35 and 36, and in the middle of MDS 30. But by far the largest peak in D_{SS} came near the beginning of MDS 29. The shape and statistical significance of the D_{SS} peak at the 5′ end of MDS 29 were the same when a simpler Jukes-Cantor model was used for calculating D_{SS} (data not shown). The absence of other recombination spikes in the distal region (where major and minor locus MDSs alternate) is probably due to the shortness of these MDSs relative to the window size used to calculate the D_{SS} statistic [500 nucleotides (nt)].

The simplest model of meiotic recombination is not consistent with the concentrated multiple significant peaks in Figure 4C. Similarly, intermolecular unscrambling across homologous chromosomes during macronuclear development would be expected to create mosaic patterns of DNA polymorphism, but not concentrated in a particular region.

A more fitting explanation is that MDS 29 exists in a duplicated copy linked to the minor locus, as postulated in the previous section, and that variability in the location of splicing, and possible multiplicity in splice location within particular alleles, leads to a concentration of D_{SS} spikes within this extensively duplicated region at the border of the major and minor loci. The data are also consistent with partial gene conversion events within a paralogously duplicated region encompassing MDSs 29 and 30.

Variability in splice location between two extensive paralogs can lead to mosaic macronuclear data even when there is no variation segregating within the paralogs, so long as there are differences between them. For this reason we cannot conclude from the peaks in the MDS 29–MDS 30 region that the major and minor loci are unlinked. However, the significant peak between major locus MDS 35 and minor locus MDS 36, a region with no evidence of duplication, suggests that the major and minor loci may have independent genealogies.

We note in passing that the peaks in recombination signals seem to occur between the conserved protein domains, suggesting that either recombination or splicing during unscrambling or both is suppressed there (Figure 4C). In the micronucleus, this could be due to genetic hitchhiking of linked neutral sites in regions of purifying selection.

While D_{SS} reveals aggregate local variation in windows of genealogical signal indicative of a recombination-like event, a different level of resolution on the mosaic structure in the data is achieved with compatibility analysis, as shown in Figure 5. In compatibility analysis all segregating sites are individually and simultaneously compared in pairwise fashion. Extra spatial resolution is gained at the cost of loss in power to detect genealogical differences of sequence regions, because all four haplotypes of two alleles at two sites must be present to detect incompatibility, pairwise incompatible sites may be mutually compatible with a third site, and genealogical information from sites with more than two segregating nucleotides may not be fully used.

Figure 5 shows results of compatibility analysis on a rescrambled version of the macronuclear data, projecting it into our best estimate of its hereditary micronuclear structure using the program RESCRAMBLE. In this reordering, major locus MDSs on the Crick

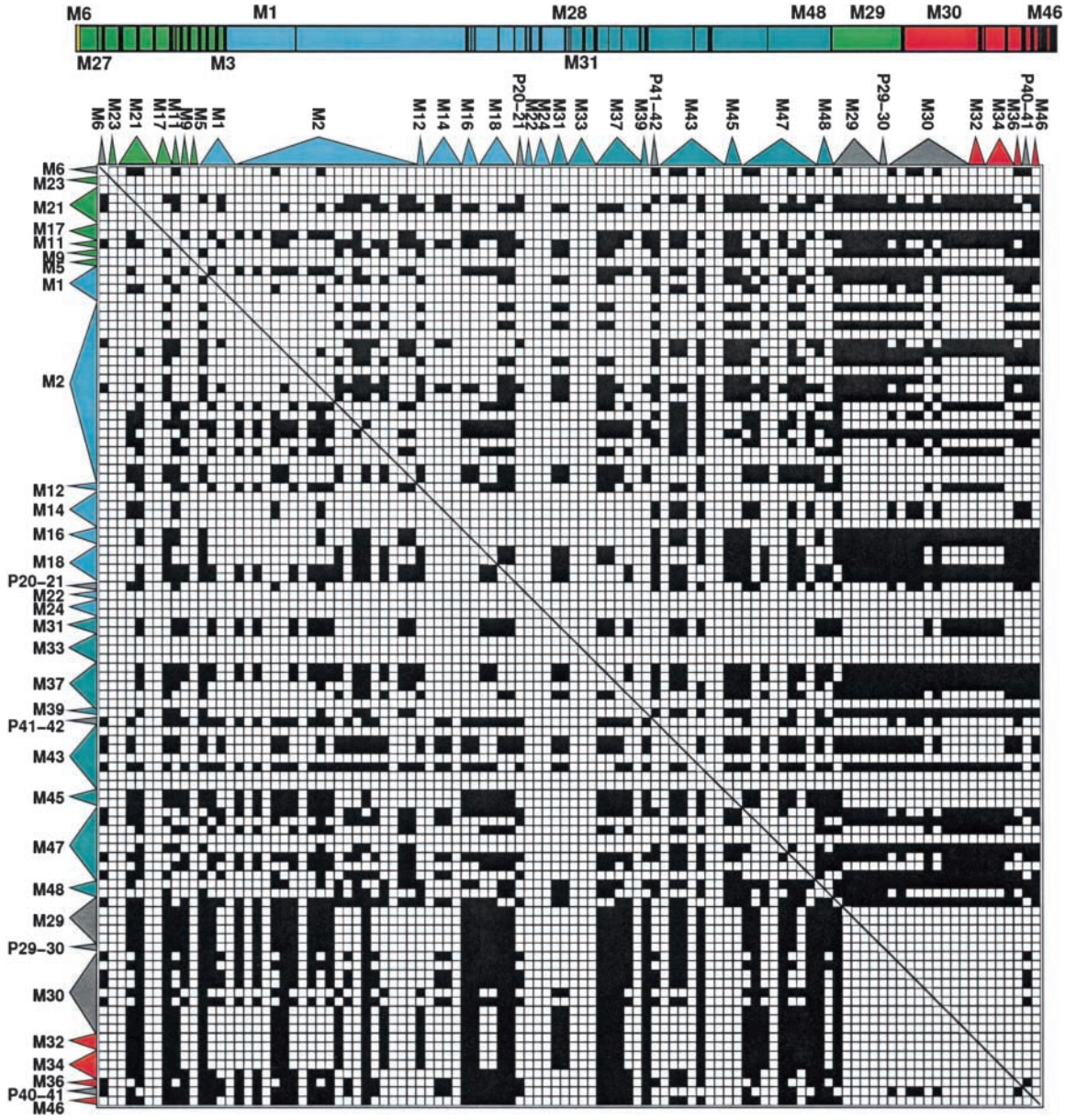

FIGURE 5.—Compatibility analyses of 104 "rescrambled" parsimoniously informative sites. Black (white) squares represent incompatible (compatible) pairs of sites (see MATERIALS AND METHODS for further explanation). Sites are grouped by MDS or pointer of origin; groupings are either colored as in Figure 1 or shaded gray. Gray indicates uncertainty or multiplicity of micronuclear origin. The alignment was projected into estimated micronuclear map order by RESCRAMBLE, allowing for putative linkage of MDS 29 to the minor locus. The top shows a key to this rescrambling.

strand are reordered in reverse complement, the minor locus is appended on the 3′ end of the major locus, and MDS 6, the location of which is unknown, is appended to the 5′ end. Departing from the previous model (Figure 1), but consistent with the D_{SS} scan (Figure 4) and polymorphism data, MDS 29 is appended to the 5′ end of the minor locus. Supplementary Figure 2 (http://www.genetics.org/supplemental/) shows our results with the raw data, along with the coordinates of sites analyzed in reference to the polymorphic alignment. The results are identical up to permutation, although the data pattern is easier to see in the form shown in Figure 5.

Several results stand out from compatibility analysis of the rescrambled macronuclear data:

1. MDS 29 is compatible with the minor locus and largely incompatible with the major locus. This is consistent with a dominantly minor locus origin of MDS 29.
2. The major and minor loci are largely mutually incompatible (but see below). This is consistent with recombination between loci.
3. The minor locus is internally compatible. This is consistent with meiotic and developmental linkage of the minor locus.

4. Some segments of the major locus appear to be compatible with the minor locus and incompatible with other segments of the major locus.
5. The first parsimoniously informative site of MDS 29 (site 2510 in supplementary Figure 2) has a unique pattern that is incompatible with most other sites. This may be due to variability in splice location across haplotypes in the MDS 29–MDS 30 region.
6. MDS 6, the location of which is still unidentified, has a similar compatibility pattern to parts of M1 and M2, M14, P20–21, P41–42, and P40–41. The pattern is more consistent with linkage of MDS 6 to the major locus.

The small-scale incompatibility within the major locus is very curious; its physical distance is much too small to display such extensive evidence of recombination. Although this signal could arise from meiotic recombination, it is more likely due to some combination of alternative unscrambling with variability in splice location, possibly among additional unobserved paralogous major locus MDSs, intragenic recombination during macronuclear development as has been observed in Tetrahymena (Deak and Doerder 1998), and partial gene conversion among paralogous gene segments in the micronucleus.

The interpretation of additional paralogous major locus MDSs might also explain the excess of synonymous mutations in the misunscrambled haplotype *SGR2-a2*. In that haplotype all of the synonymous mutations were in major locus MDSs, while nonsynonymous mutations were present in both major and minor locus MDSs (data not shown).

Gene genealogies and splits graphs of the major and minor loci: We then examined more explicitly the genealogical patterns in the data. In so doing we asked whether the data reflect the geographic origin of the haplotypes, whether there is any aggregate support for the major and minor loci sharing genealogical history, and, if not, what the minimum number of changes that separate the genealogies is. These last questions bear on the open issue of whether the major and minor loci are meiotically linked.

We used split decomposition to make genealogical networks of the major and minor loci because of the incompatibilities within the major locus. We excluded regions that we knew or suspected were duplicated in both the major and minor loci: MDSs 29 and 30 and pointer sequences with copies in both loci. Figure 6 shows that the minor locus split system was tree-like with a perfect fit of the split decomposition to the original distance matrix used, while the major locus genealogy was much less tree-like and the fit of the decomposition was worse. Even when broken up into Watson and Crick strands, with the Watson strand into distal and proximal regions, the genealogies were not tree-like and the split decomposition had poor fit, indicating some unknown level of recombinative structure in the data. This is consistent with the results of the compatibility approach in Figure 5. However, in none of these splits graphs of parts of the major locus did we observe any support for the minor locus tree. Both *FED-a1* and *NGR-a2* have different phylogenetic affinity at the major and minor loci, with very strong bootstrap support by all methods. Also, we note in passing that the minor locus genealogy appeared to reflect the geographic origin of the strains better than the major locus genealogy.

These results, particularly the absence of the minor locus genealogy in the major locus split system, suggest that the major and minor loci are meiotically recombining, that somatic intragenic recombination occurs after unscrambling preferentially between the major and minor loci, or that intermolecular unscrambling (across homologous chromosomes) occurs preferentially between the major and minor loci. Yet the latter two processes might generally be expected to produce up to four haplotypes per individual, where we have never observed more than two (see DISCUSSION). We assume in this argument that unscrambling, which is believed to occur at the polytene stage (D. M. Prescott, personal communication), is independent across different chromosomes or haploid sets of chromosomes. With this caveat, we tentatively suggest that our observations are more consistent with meiotic recombination.

Excess nucleotide diversity at MDS 29: A statistical summary of the polymorphism data broken down by region appears in Table 1. Statistical significance was calculated through parametric bootstrapping with coalescent simulations. Table 1 shows a borderline positive Tajima's *D* in MDS 29. Positive Tajima's *D* is consistent with a variety of demographic and evolutionary forces including subdivided population structure and the maintenance of polymorphism due to balancing selection as in the case of neutral sites linked to the segregating electromorph in Drosophila *Adh* (Kreitman 1983). Here the region in question is quite near a conserved protein region, and the corresponding levels of polymorphism are relatively low.

The data as a whole did not indicate that the sample was derived from a subdivided population. For instance, the average nucleotide diversity within individuals was approximately equal to if not greater than the diversity between individuals in a sliding-window analysis (considering distinct haplotypes, data not shown). Also, other features of the data had negative Tajima's *D*. Tajima's *D* was significantly negative for the 5′ leader sequence, consistent with purifying selection, perhaps to maintain *cis*-acting regulatory elements. Among 10 of 1500 amino acids segregating, none were in conserved protein regions. Watterson's estimator per amino acid residue for the entire protein sequence was $\hat{\theta}_w = 0.0029 \pm 0.0016$, Tajima's estimator per amino acid residue was $\hat{\theta}_\pi = 0.0025 \pm 0.0015$, and Tajima's *D* was $D = -0.89$. Standard errors here assume no recombination (Tajima

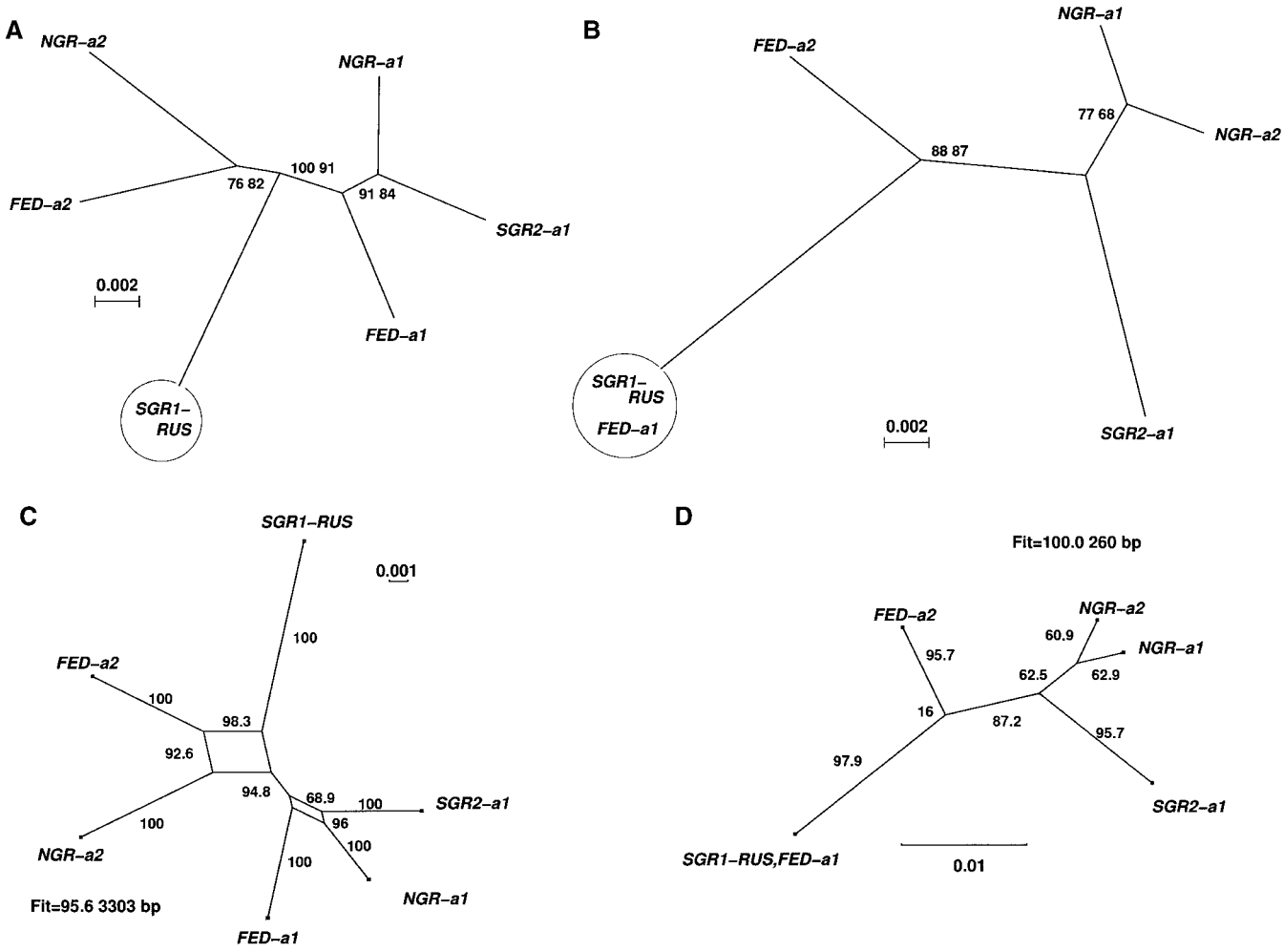

FIGURE 6.—Incongruent gene trees/networks of the major and minor loci. (A and C) Major locus minus MDS 29 and distal region pointer sequences (3511 nt). (B and D) Minor locus minus MDS 30 and pointer sequences (260 nt). The phylogenetic affinity of haplotypes is highly promiscuous over the major and minor loci. (A and B) Neighbor-joining phylograms of the major and minor loci, respectively, at the same scale. Bootstrap support is indicated at edges in order of neighbor joining and parsimony. Circled haplotypes indicate allelic identity over the subalignment. (C and D) Split decomposition of the major and minor loci. Bootstrap support of each split is shown (1000 replicates).

1993). These results suggest that subdivided population structure cannot explain the tendency toward positive *D* in MDS 29.

A sliding-window analysis of the data using PI, shown in Figure 4, A and B, shows a starred positive-value peak in Tajima's *D* (Figure 4A) corresponding to two windows in MDS 29 that were significantly positive using the β-distribution (TAJIMA 1989) as calculated with DNASP but not taking into account multiple comparisons. However, taking into account multiple comparisons, this result was not significant. Using SCANMS to model sliding-window analysis of neutral genealogies evolving with Watterson's estimator $\theta_w = 83$ for the region as a whole (no recombination, population growth, or migration) gave a 95th percentile of 2.1534 (one-sided test) among maximal window values of *D*. This does not detract from our result that *D* may be positive in MDS 29, but it does show that the results of sliding-window analyses of neutrality tests should be treated with caution if multiple comparisons are not taken into account.

That the excess nucleotide diversity occurs near the boundary between the major and minor loci suggests that the pattern may have more to do with the process of unscrambling than protein evolution does. A possible explanation is as follows: subdivided population structure violates the random-mating assumption of the Wright-Fisher model and can lead to a positive Tajima's *D* (see, *e.g.*, TAJIMA 1993). We suggest that nucleotide diversity is increased when the location of splicing is variable, because independently segregating paralogous sequences can then contribute to macronuclear variation at specific sites, and that this situation is analogous to sampling from a subdivided population.

Taking pointer sequences as an example, only one of the two repeats will be spliced into the macronucleus, but exactly where and how the cutting and splicing

occur is unknown. A macronuclear pointer could derive entirely from the headpointer, entirely from the tailpointer, or chimerically from both, depending on the splice location. During micronuclear meiosis, however, headpointers and tailpointers are distinct. They segregate and evolve independently. A headpointer haplotype could never fix in a population of tailpointers without some kind of conversion event—they are not at the same loci. This can be directly compared to a structured population where gene conversion acts as a weak migration force. Thus, if in forming the macronucleus, splicing occurs at variable locations within any region, a macronuclear sample would look like a sample from a subdivided population in just this region. Heterozygosity will be in apparent excess, for instance, because different allelic states in each subpopulation may be fixed but will be mixed at the macronuclear level. This excess heterozygosity leads to a positive Tajima's *D*.

Thus, the positive trend in Tajima's *D* in MDS 29 is consistent with alternative locations of splicing across haplotypes, which was also suggested by our other analyses. It is tempting to also ascribe the positive value of Tajima's *D* within pointer sequences to this effect. They share with MDS 29 a low absolute level of polymorphism. However, the region sampled includes only five segregating sites. More data are necessary to settle the tantalizing issue of whether variability in splice location within pointers increases the nucleotide diversity there or in adjacent regions.

DISCUSSION

IES excision, MDS splicing, and gene unscrambling have generally been regarded as deterministic processes. For instance, Prescott *et al.* (1998), Landweber *et al.* (2000), Jahn and Klobutcher (2002), and others discuss whether the length and complexity of pointer repeat sequences are sufficient to uniquely determine MDS splicing/unscrambling and IES excision. Consistent with this hypothesis, repeats at scrambled MDS-IES boundaries are longer than those of nonscrambled MDS-IES boundaries (Prescott and DuBois 1996).

Our results suggest that gene unscrambling may be fundamentally variable and perhaps best regarded as probabilistic in nature. Our data suggest at least three modes of variability in unscrambling: (1) unscrambling appears to facultatively incorporate one among a set of paralogous MDSs and pointers, (2) splicing location appears to vary within paralogous MDSs, and (3) entire MDSs can fail to be inserted and entire IESs can fail to excise.

Our observation of unscrambling variability raises issues in light of the hypothesis that macronuclear sequences may guide the unscrambling of micronuclear MDSs during development, as discussed by Prescott (1999) and Jahn and Klobutcher (2002) in relation to the experiments of Forney *et al.* (1996) and Duharcourt *et al.* (1998) in Paramecium and Chalker and Yao (1996) in Tetrahymena. Recent progress in Tetrahymena has started to build a mechanistic picture of how sequence information in the old macronucleus can shape the postconjugal developing micronucleus through epigenetic marking of chromatin (Taverna *et al.* 2002) and possibly the action of small RNAs (Mochizuki *et al.* 2002).

Suppose now that an error occurs in unscrambling, causing a misunscrambled macronuclear haplotype, such as *SGR2-a2*. Would this haplotype feed forward and increase the probability of error in future generations? Would it increase the chances of misunscrambling other alleles with which it is coinherited? If so, misunscrambled haplotypes could decrease long-term reproductive success, as exconjugant offspring of an individual carrying misunscrambled macronuclear haplotypes would be subsequently induced to misunscramble. Individuals that carry misunscrambled haplotypes arising even through environmental means, or lineages that carry mutations that increase the chance of unscrambling error (acting either *in cis* or *in trans*), would suffer a greater than expected long-term loss in reproductive success if there were no transfer of information from the macronucleus to the micronucleus in exconjugants. Further breeding and macronuclear injection experiments with the SGR2 strain could allow exploration of this "error catastrophe" issue.

Our data are consistent with a large genetic distance between the major and minor loci. This could possibly reflect a high meiotic recombination rate (short map distance) or a large physical distance between the loci. The data are also consistent with developmental recombination either in the unscrambling stage—that is, unscrambling across homologous chromosomes—or afterward during amplification as has been observed over very short distances ($<$1 kb) in Tetrahymena (Deak and Doerder 1998). These alternatives would also account for the incompatibilities we observed within the major locus, which, at 3.5 kb is $>$10 times longer than the minor locus (explaining its lack of internal incompatibility, Figure 5). Perhaps the extensive paralogy of the MDS 29–30 region, with a combined size of $>$500 bp, increases the specific occurrence of developmental recombination at just that location, accounting for both the mosaic structure at that location and the incongruent genealogies of the major and minor loci.

Yet, unlike in the case of Tetrahymena studied by Deak and Doerder, we did not find evidence for more than two macronuclear haplotypes per Stylonychia individual either in this study or in a separate analysis of $>$30 clones made from exconjugants of a laboratory cross of genetically distinct individuals (A. Goodman, G. Zilioli, D. H. Ardell and L. F. Landweber, unpublished data). This objection assumes that unscrambling occurs after polytene duplication and is in some measure independent across these duplicates with respect

to any variability in macronuclear development. For instance, one way in which development may not be independent is through a possible common dependency on macronuclear inheritance. A further uncertainty in the intermolecular unscrambling interpretation is whether or not homologous chromosomes are paired at the polytene stage in Stylonychia (Ammermann 1971; Williams *et al.* 2002).

It is also possible that the chance of intermolecular unscrambling in at least this region increases with physical distance on the chromosome, in which case we are still left with the interpretation of a relatively large physical distance between the major and minor loci. This is provocative because if unscrambling can indeed take place over large physical distances or between separate chromosomes, the entire genome could potentially become available for DNA-splicing interactions during macronuclear development.

In the case of DNA pol α, the variability of unscrambling that we inferred in the MDS 29–30 region was essentially silent at the protein level. There is, however, no reason why the development of other genes could not be much more dramatically affected by these same phenomena. In principle, alternative unscrambling could generate combinatoric diversity in macronuclear genes, analogous to that generated by alternative splicing or alternative DNA processing in the development of the vertebrate immune system. However, this adaptive potential comes at the cost of increased vulnerability to errors in the unscrambling process, for instance, through the increased chance of unscrambling at undesirable locations after random mutation creates additional matching pointer sequences.

Biochemical data suggest that the DNA elimination during IES removal occurs only within vesicles that subdivide and isolate polytenic chromosomes from one another (Ammermann 1971; Murti 1973; Prescott and Murti 1974). These structures might possibly restrict the potential domain of splicing interactions among micronuclear loci during gene unscrambling and IES excision, reducing the chance of error and fragility to mutation.

Nonscrambled ciliate genes might be affected by variability in macronuclear development, particularly IES excision. For instance, Bernhard (1999) made the highly unusual observation of >10 divergent macronuclear histone H3 genes in *S. lemnae*, with extensive size variation apparent in Southern analysis. Some of the sequenced variants contained short stretches of low-complexity sequence, suggesting to us that they may have retained IES sequence in some cases. Perhaps alternative IES excision underlies this macronuclear variation.

From a population genetic perspective, alternative unscrambling of paralogous MDSs will have mixed consequences on the molecular evolution of ciliate genes. This is because more than one micronuclear locus can contribute to the same macronuclear locus. In principle, perhaps an arbitrary number of paralogous MDSs could segregate independently yet contribute to the same macronuclear locus. Even if there were no genetic variation within each of these independently segregating paralogous loci, their alternative incorporation or variability in the location of splicing between them would disproportionately increase macronuclear nucleotide diversity as we have observed in MDS 29. One may speculate that the alternative incorporation of paralogous MDSs decreases the efficacy of selection at individual loci, for instance, by decreasing the effective population size of individual paralogs. On the other hand, the efficacy of weak directional selection could be enhanced by genetic distance between MDSs, paralogous or otherwise, through the reduction of Hill-Robertson interference (Hill and Robertson 1966; McVean and Charlesworth 2000). It has recently been claimed that even strong purifying selection could be made more effective by increased genetic distance between selected loci (Williamson and Orive 2002).

The significant excess of unique synonymous differences in *SGR2-a2* cannot be the consequence of allelic inactivation. It may be that this haplotype is more distantly related to the rest of the sample. Or perhaps this haplotype contains other paralogous MDSs that are alternatively unscrambled, but not otherwise present in the data. If so, perhaps their presence in *SGR2-a2* came from a disruption in the unscrambling process. It is also tempting to attribute the small-scale incompatibility in the major locus (Figure 5) to the presence of other paralogous major locus MDSs, but no other evidence supports this.

At least in some cases, paralogous MDSs and other features of scrambled genes could help protect against errors in the unscrambling process and other more general deleterious developmental and evolutionary processes. For instance, the extensive paralogy of MDS 30 and inferred paralogy of MDS 29 occur precisely at a point of major disruption in gene structure, right before a conserved protein region (Figure 4). Paralogy provides redundancy that, so long as multiple developmental pathways are expressed within individuals, might increase the chances of creating a functional macronuclear gene. The partly unscrambled haplotype that we observed, *SGR2-a2*, directly demonstrates that unscrambling can fail and also that other features of scrambled genes can protect against these failures. For instance, MDS 6 and the 5′-most 12 bases of MDS 30 failed to splice into the otherwise correctly scrambled haplotype. Perhaps the short length of these segments, and their length multiplicity of three, protects against misunscrambling in other haplotypes that are actively transcribed and translated. There is almost certainly more variability in the unscrambling process than we directly observed, which was presumably limited to that

which was neutral or nearly neutral in the laboratory-cultivated natural strains that we studied.

We suggest that ciliates endure rather frequent introductions of IESs or other noncoding sequence into their coding regions. This is proposed to occur both evolutionarily, through *cis*-acting mutations in pointer sequences, MDSs, and IESs, and developmentally, through chance errors in unscrambling and other DNA-splicing events during macronuclear development. Perhaps *trans*-acting mutations in the biochemical machinery of developmental DNA rearrangement also contribute to deleterious variability in the unscrambling process. However these errors arise, both the redundancy of paralogous MDSs and MDS-IES gene structure could play roles in protecting from loss of function as stated above. In this context we introduce the nonsense-suppression hypothesis for the origin of altered genetic codes in ciliates. All altered genetic codes in ciliates reduce the number of stop codons relative to the standard code, and such codes appear to have evolved multiple times in the evolution of ciliates (Lozupone *et al.* 2001). We propose that genetic codes in ciliates have coevolved with their developmental DNA processing, gene structure, and nuclear duality to increase tolerance to the spontaneous introduction of nonsense codons from A/T-rich IESs into macronuclear coding regions. The genomic scale of the interruption of protein-coding genes by noncoding DNA in ciliates supports this hypothesis.

Many of our interpretations and their uncertainties rest on inferring the micronuclear sequences from a related but distinct strain. Future studies that include micronuclear sequencing, searching for the paralogous MDS 29 we have inferred, and careful breeding experiments—for instance, with strains that present misunscrambled haplotypes—will vastly improve and refine our understanding of the genomic underpinning of gene unscrambling, its process, and its short- and long-term evolutionary consequences in ciliates and their genes.

The exceptional nature of ciliate genetics provides a potentially fertile ground for extending and testing molecular evolutionary theory and tools. For instance, gene scrambling and nuclear duality provide a unique opportunity to explore the interaction of recombination and selection on the evolution of protein-coding genes.

The authors thank Tom Doak, Matthew Webster, Alex Mira, Andrew G. Clark, Montgomery Slatkin, and two anonymous reviewers for helpful criticism of the manuscript. We also thank Michael Cummings and the Workshop on Molecular Evolution at the Marine Biological Laboratory at Woods Hole for hospitality and discussions at the beginning of this project. This material is based upon work supported by the National Science Foundation under a Postdoctoral Fellowship in Bioinformatics awarded to D.H.A. in 2000. L.F.L. acknowledges support from National Science Foundation grant 0121422 and National Institute of General Medical Sciences award GM59708.

LITERATURE CITED

Ammermann, D., 1965 Cytologische und genetische Untersuchungen an dem ciliaten *Stylonychia mytilus* Ehrenberg. Arch. Protistenkd. **108:** 109–152.

Ammermann, D., 1971 Morphology and development of the macronuclei of the ciliates *Stylonychia mytilus* and *Euplotes aediculatus*. Chromosoma **33:** 209–238.

Ammermann, D., 1987 Giant chromosomes in ciliates. Results Probl. Cell. Differ. **14:** 59–67.

Bernhard, D., 1999 Several highly divergent histone H3 genes are present in the hypotrichous ciliate *Stylonychia lemnae*. FEMS Microbiol. Lett. **175:** 45–50.

Byun, R., L. D. Elbourne, R. Lan and P. R. Reeves, 1999 Evolutionary relationships of pathogenic clones of *Vibrio cholerae* by sequence analysis of four housekeeping genes. Infect. Immun. **67:** 1116–1124.

Cartinhour, S. W., and G. A. Herrick, 1984 Three different macronuclear DNAs in *Oxytricha fallax* share a common sequence block. Mol. Cell. Biol. **4:** 931–938.

Chalker, D. L., and M. C. Yao, 1996 Non-Mendelian, heritable blocks to DNA rearrangement are induced by loading the somatic nucleus of *Tetrahymena thermophila* with germ line-limited DNA. Mol. Cell. Biol. **16:** 3658–3667.

Clark, A. G., 1990 Inference of haplotypes from PCR-amplified samples of diploid populations. Mol. Biol. Evol. **7:** 111–122.

Damagnez, V., J. Tillit, A. M. de Recondo and G. Baldacci, 1991 The POL1 gene from the fission yeast, *Schizosaccharomyces pombe*, shows conserved amino acid blocks specific for eukaryotic DNA polymerases alpha. Mol. Gen. Genet. **226:** 182–189.

Deak, J. C., and F. P. Doerder, 1998 High frequency intragenic recombination during macronuclear development in *Tetrahymena thermophila* restores the wild-type SerH1 gene. Genetics **148:** 1109–1115.

DuBois, M. L., and D. M. Prescott, 1997 Volatility of internal eliminated segments in germ line genes of hypotrichous ciliates. Mol. Cell. Biol. **17:** 326–337.

Duharcourt, S., A. M. Keller and E. Meyer, 1998 Homology-dependent maternal inhibition of developmental excision of internal eliminated sequences in *Paramecium tetraurelia*. Mol. Cell. Biol. **18:** 7075–7085.

Ewens, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **3:** 87–112.

Felsenstein, J., and G. A. Churchill, 1996 A Hidden Markov Model approach to variation among sites in rate of evolution. Mol. Biol. Evol. **13:** 93–104.

Forney, J. D., F. Yantiri and K. Mikami, 1996 Developmentally controlled rearrangement of surface protein genes in *Paramecium tetraurelia*. J. Eukaryot. Microbiol. **43:** 462–467.

Gerber, C. A., A. B. Lopez, S. J. Shook and F. P. Doerder, 2002 Polymorphism and selection at the SerH immobilization antigen locus in natural populations of *Tetrahymena thermophila*. Genetics **160:** 1469–1479.

Hartl, D. L., and A. G. Clark, 1989 *Principles of Population Genetics*, Ed. 2. Sinauer Associates, Sunderland, MA.

Herrick, G., D. Hunter, K. Williams and K. Kotter, 1987 Alternative processing during development of a macronuclear chromosome family in *Oxytricha fallax*. Genes Dev. **1:** 1047–1058.

Hill, W. G., and A. Robertson, 1966 The effect of linkage on limits to artificial selection. Genet. Res. **8:** 269–294.

Hoffman, D. C., and D. M. Prescott, 1996 The germline gene encoding DNA polymerase alpha in the hypotrichous ciliate *Oxytricha nova* is extremely scrambled. Nucleic Acids Res. **24:** 3337–3340.

Hoffman, D. C., and D. M. Prescott, 1997 Evolution of internal eliminated segments and scrambling in the micronuclear gene encoding DNA polymerase alpha in two Oxytricha species. Nucleic Acids Res. **25:** 1883–1889.

Hogan, D. J., E. A. Hewitt, K. E. Orr, D. M. Prescott and K. M. Muller, 2001 Evolution of IESs and scrambling in the actin I gene in hypotrichous ciliates. Proc. Natl. Acad. Sci. USA **98:** 15101–15106.

Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18:** 337–338.

Huson, D. H., 1998 SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics **14:** 68–73.

Jahn, C. L., and L. A. Klobutcher, 2002 Genome remodeling in ciliated protozoa. Annu. Rev. Microbiol. **56:** 489–520.

Jakobsen, I., and S. Easteal, 1996 A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. Comput. Appl. Biosci. **12:** 291–295.

Karlin, S., and J. McGregor, 1972 Addendum to a paper of W. Ewens. Theor. Popul. Biol. **3:** 113–116.

Karotam, J., T. M. Boyce and J. G. Oakeshott, 1995 Nucleotide variation at the hypervariable esterase 6 isozyme locus of *Drosophila simulans*. Mol. Biol. Evol. **12:** 113–122.

Klobutcher, L. A., C. L. Jahn and D. M. Prescott, 1984 Internal sequences are eliminated from genes during macronuclear development in the ciliated protozoan *Oxytricha nova*. Cell **36:** 1045–1055.

Klobutcher, L. A., M. E. Huff and G. E. Gonye, 1988 Alternative use of chromosome fragmentation sites in the ciliated protozoan *Oxytricha nova*. Nucleic Acids Res. **16:** 251–264.

Kreitman, M., 1983 Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature **304:** 412–417.

Landweber, L. F., T.-C. Kuo and E. A. Curtis, 2000 Evolution and assembly of an extremely scrambled gene. Proc. Natl. Acad. Sci. USA **97:** 3298–3303.

Li, W. H., and L. A. Sadler, 1991 Low nucleotide diversity in man. Genetics **129:** 513–523.

Lozupone, C. A., R. D. Knight and L. F. Landweber, 2001 The molecular basis of nuclear genetic code change in ciliates. Curr. Biol. **11:** 65–74.

Mansour, S. J., D. C. Hoffman and D. M. Prescott, 1994 A gene-sized DNA molecule encoding the catalytic subunit of DNA polymerase alpha in the macronucleus of *Oxytricha nova*. Gene **144:** 155–161.

McGuire, G., and F. Wright, 2000 Topal 2.0: improved detection of mosaic sequences within multiple alignments. Bioinformatics **16:** 130–134.

McVean, G. A., and B. Charlesworth, 2000 The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. Genetics **155:** 929–944.

Mochizuki, K., N. A. Fine, T. Fujisawa and M. A. Gorovsky, 2002 Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in Tetrahymena. Cell **110:** 689–699.

Moriyama, E. N., and J. R. Powell, 1996 Intraspecific nuclear DNA variation in Drosophila. Mol. Biol. Evol. **13:** 261–277.

Murti, K. G., 1973 Organization of genetic material in the macronucleus of hypotrichous ciliates, pp. 113–137 in *Molecular Genetics: Handbook of Genetics*, Vol. 5, edited by R. C. King. Plenum, New York.

Prescott, D. M., 1994 The DNA of ciliated protozoa. Microbiol. Rev. **58:** 233–267.

Prescott, D. M., 1999 The evolutionary scrambling and developmental unscrambling of germline genes in hypotrichous ciliates. Nucleic Acids Res. **27:** 1243–1250.

Prescott, D. M., and M. L. DuBois, 1996 Internal eliminated segments (IESs) of Oxytrichidae. J. Eukaryot. Microbiol. **43:** 432–441.

Prescott, D. M., and A. F. Greslin, 1992 Scrambled actin I gene in the micronucleus of *Oxytricha nova*. Dev. Genet. **13:** 66–74.

Prescott, D. M., and K. G. Murti, 1974 Chromosome structure in ciliated protozoans. Cold Spring Harbor Symp. Quant. Biol. **38:** 609–618.

Prescott, J. D., M. L. DuBois and D. M. Prescott, 1998 Evolution of the scrambled germline gene encoding alpha-telomere binding protein in three hypotrichous ciliates. Chromosoma **107:** 293–303.

Przeworski, M., R. R. Hudson and A. Di Rienzo, 2000 Adjusting the focus on human variation. Trends Genet. **16:** 296–302.

Ribas-Aparicio, R. M., J. J. Sparkowski, A. E. Proulx, J. D. Mitchell and L. A. Klobutcher, 1987 Nucleic acid splicing events occur frequently during macronuclear development in the protozoan *Oxytricha nova* and involve the elimination of unique DNA. Genes Dev. **1:** 323–336.

Rozas, J., and R. Rozas, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174–175.

Seegmiller, A., K. R. Williams, R. L. Hammersmith, T. G. Doak, D. Witherspoon *et al.*, 1996 Internal eliminated sequences interrupting the Oxytricha 81 locus: allelic divergence, conservation, conversions, and possible transposon origins. Mol. Biol. Evol. **13:** 1351–1362.

Stajich, J. E., D. Block, K. Boulez, S. E. Brenner, S. A. Chervitz *et al.*, 2002 The Bioperl toolkit: Perl modules for the life sciences. Genome Res. **12:** 1611–1618.

Swofford, D., 2000 *PAUP*: Phylogenetic Analysis Under Parsimony and Other Methods*, Version 4.0b8. Sinauer Associates, Sunderland, MA.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

Tajima, F., 1993 Measurement of DNA polymorphism, pp. 37–59 in *Mechanisms of Molecular Evolution*, edited by N. Takahata and A. G. Clark. Sinauer Associates, Sunderland, MA.

Taverna, S. D., R. S. Coyne and C. D. Allis, 2002 Methylation of histone h3 at lysine 9 targets programmed DNA elimination in Tetrahymena. Cell **110:** 701–711.

Thompson, J. D., D. G. Higgins and T. J. Gibson, 1994 Clustal-W—improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:** 4673–4680.

Williams, K. R., and G. Herrick, 1991 Expression of the gene encoded by a family of macronuclear chromosomes generated by alternative DNA processing in *Oxytricha fallax*. Nucleic Acids Res. **19:** 4717–4724.

Williams, K. R., T. G. Doak and G. Herrick, 2002 Telomere formation on macronuclear chromosomes of *Oxytricha trifallax* and *O. fallax*: alternatively processed regions have multiple telomere addition sites. BMC Genet. **3:** 16.

Williamson, S., and M. E. Orive, 2002 The genealogy of a sequence subject to purifying selection at multiple sites. Mol. Biol. Evol. **19:** 1376–1384.

Wong, S. W., A. F. Wahl, P. M. Yuan, N. Arai, B. E. Pearson *et al.*, 1988 Human DNA polymerase alpha gene expression is cell proliferation dependent and its primary structure is similar to both prokaryotic and eukaryotic replicative DNA polymerases. EMBO J. **7:** 37–47.

Communicating editor: M. Feldman