Evolution of Programmed DNA Rearrangements in a Scrambled Gene

Li Chin Wong* and Laura F. Landweber†

*Department of Molecular Biology, Princeton University; and †Department of Ecology and Evolutionary Biology, Princeton University

Gene unscrambling in spirotrich ciliates involves massive genome-wide DNA deletion and rearrangement events during development. During each sexual cycle, the somatic nucleus (macronucleus) regenerates from the germ line nucleus (micronucleus). Development of the polyploid somatic genome requires programmed DNA deletion of micronuclear-limited intragenic noncoding sequences and permutation and amplification of the protein-coding regions. Recent studies suggest that, despite novel insertions of endogenous transposon or foreign DNA into the germ line genome, ciliates possess a whole-genome surveillance system that guides the recapitulation of a functional somatic genome. This renders the germ line genome an extremely dynamic structure over evolutionary time. Here we describe the germ line and somatic architectures of the gene encoding α-telomere–binding protein in three early-diverging species (Holosticha sp., Uroleptus sp., and Paraurostyla weissei) and trace the natural history of DNA rearrangements in this gene in six species, including three previously studied oxytrichids. Comparisons of homologous coding regions between earlier and later diverging species provide evidence for fusion of scrambled germ line fragments as small as 24 bp during evolution, as well as simultaneous fragmentation and scrambling of the germ line locus and shifting of the boundaries between coding and noncoding DNA, leading to distinct gene architectures in each species. We infer an evolutionary recombination pathway that passes through identified intermediate species and gives rise to the observed patterns in all known species, capitalizing on their unique DNA rearrangement machinery and germ line flexibility.

Introduction

Gene unscrambling in spirotrich ciliates is an extreme example of genomic DNA microarrangements during development. Nuclear duality in ciliates permits the partitioning of germ line (micronucleus) and somatic (macronucleus) functions of the genome (for a review, see Prescott 1994). During sexual conjugation, haploid micronuclei exchange to form zygotic nuclei. A copy of the zygotic nucleus then develops into a new macronucleus. This developmental process involves massive DNA rearrangement events that are thought to depend upon maternal effects via communication between the old macronucleus and new micronucleus (reviewed in Meyer and Garnier 2002; Yao, Fuller, and Xi 2003). Noncoding micronuclear-limited internally eliminated sequences (IESs) are excised. The remaining macronuclear destined sequences (MDSs), are rearranged and reassembled into gene-sized molecules capped at both ends with telomeres. This DNA processing involves elimination of 95% or more of the micronuclear sequence complexity. The MDSs of some genes in a subset of spirotrich ciliates are coded in a nonlinear order in the micronucleus, relative to the functional macronucleus. Unscrambling of these genes involves programmed DNA rearrangements that restore the functional order of MDSs in the somatic genome. Two genetic peculiarities that could have facilitated the evolution of gene scrambling in spirotrichs include the presence of nuclear dimorphism and gene-sized “chromosomes” in the somatic genome.

Assembly of scrambled genes may involve homologous recombination, guided by direct repeats, termed pointers, at the 3′ end of one MDS and the 5′ end of the next contiguous MDS, relative to the somatic order. In the micronuclear genome, these pointer sequences form the boundaries between MDSs and IESs. Only one copy of each pointer sequence pair is retained in the macronuclear genome (Prescott 1994; DuBois and Prescott 1997).

There are three known scrambled genes in spirotrichs—actin I, DNA polymerase α (DNA pol-α), and α-telomere-binding protein (α-TBP)—with the recent discovery of three more described in Kuo, Chang, and Landweber (2006) and Chang, Kuo, and Landweber (2006). All three previously studied genes have strikingly different scrambling patterns. Actin I is scrambled in a less organized pattern and is fragmented into 8–11 MDSs (Hogan et al. 2001), whereas DNA pol-α (Hoffman and Prescott 1997; Landweber, Kuo, and Curtis 2000) and α-TBP are scrambled in a distinctly nonrandom pattern with odd-numbered MDSs segregated from even MDSs (J. D. Prescott, DuBois, and D. M. Prescott 1998) and the genes fragmented into 45–51 and 14–17 MDSs, respectively, in oxytrichids.

Our study presents a model for how germ line recombination shapes the genome over evolutionary time. Recent studies suggest that ciliates have a whole-genome surveillance system which, aided by communication between the somatic and germ line genomes, guides the formation of a new somatic genome (Yao, Fuller, and Xi 2003). Therefore, many germ line variations are phenotypically silent as they are absent from the somatic genome. In effect, nuclear dimorphism provides evolutionary capacitance where the genotypic polymorphisms in the germ line are buffered and phenotypically silent in the somatic nucleus (Rutherford and Lindquist 1998; Bergman and Siegal 2003). Our study demonstrates the dynamic nature of the germ line genome by showing that the DNA rearrangement and recombination events involved in the formation of the somatic genome are flexible over evolutionary time, despite evolutionary constraint on the encoded protein. Here we investigated the scrambling patterns and complexity in orthologs of the α-TBP gene in six related species to further understand the evolution and complexity of scrambling across these species. One hypothesis would be that the level of germ line scrambling and IES excision between earlier and later diverging species arose gradually, with a wide range of...
scrambling complexity across species. An alternative hypothesis proposes that the existence of unscrambling machinery makes scrambling complexity a binary decision, with comparable levels of scrambling in all scrambled species. Our data suggest that both are true for the α-TBP gene, which has comparable levels of global scrambling across species but local scrambling differences that arose gradually in different species.

Materials and Methods
Macronuclear and Micronuclear DNA

All macronuclear and micronuclear DNA used in this study was a gift from Wei-Jen Chang (Princeton, N.J.) and Mann Kyoon Shin (Ulsan, Korea). Holosticha sp., Uroleptus sp., and P. weissei were locally isolated in the Princeton, N.J. area. Morphological comparisons of these species and DNA extraction methods are further described in Chang et al. (2005).

Polymerase Chain Reaction Protocols

The macronuclear α-TBP genes in the three species were amplified using a set of forward degenerate primers against MDS 3, 4, and 6 (in Stylonychia mytilus) and reverse degenerate primers against the nonscrambled regions of the gene. Telomere suppression polymerase chain reaction (PCR) (Curtis and Landweber 1999; Chang et al. 2004), using nested gene-specific and telomere-specific primers, was then used to obtain the whole macronuclear gene, including the telomeric ends of the α-TBP macronuclear chromosomes. Micronuclear α-TBP loci were first amplified using primers designed against the species-specific macronuclear versions of the gene. When a micronuclear fragment is recovered, micronuclear-specific primers were designed to obtain the full micronuclear sequence.

PCR of both macronuclear and micronuclear DNA was performed with a thermal cycler (Biometra, Göttingen, Germany) using the following conditions: 94°C for 2 min, 35 cycles of 94°C for 30 s, 50°C for 1 min, and 72°C for 2 min, followed by 72°C for 10 min using DNA Taq polymerase (Roche, Indianapolis, Ind.) and EasyStart PCR Mix-in-a-Tube (Molecular BioProducts, Inc., San Diego, Calif.) which contains wax barriers that prevent nonspecific primer annealing. All macronuclear and micronuclear PCR products were gel purified, cloned using a Topo2.1 vector (Invitrogen, Carlsbad, Calif.), and sequenced. Multiple clones were sequenced to obtain the consensus sequence used in the data analysis. Sequences that are newly reported in this study are as follows: (1) both the macronuclear and micronuclear α-TBP genes for Uroleptus sp., Holosticha sp., and P. weissei (GenBank accession numbers:AY493350–AY493355) and (2) the macronuclear α-TBP genes for Eschaneustyla sp. and Urostyla grandis (GenBank accession numbers: DQ238011–DQ238012).

Sequence Analysis

Bioedit was used to generate nucleotide alignments, conceptual translation of the nucleotide sequences, and amino acid alignments. Gene Unscrambler (Cavalcanti and Landweber 2004), a program that utilizes the Blast2seq algorithm (Tatusova and Madden 1999), was used to compare the macronuclear to the corresponding micronuclear forms of the α-TBP gene. The computer-generated alignments of MDSs were then manually verified.

Genealogy Reconstruction

To determine the phylogenetic relationships among α-TBP homologs, we aligned the DNA or protein sequences for 10 homologs using ClustalX v1.81 (Thompson, Higgins, and Gibson 1994) and then refined the alignment manually. Using Euplotes crassus as an outgroup, phylogenetic analysis was performed using quartet puzzling with the program Tree-Puzzle 5.0 (Schmidt et al. 2002) using the BLOSUM 62 amino acid substitution matrix (S. Henikoff and J. G. Henikoff 1992), gamma-distributed rate variation model, and 10,000 bootstrap replicates. A similar tree topology was obtained using the maximum parsimony algorithm in PAUP* 4.0 (Swoford 2002) with default parameters and 10,000 bootstrap replicates.

Results and Discussion

In this study, we report the scrambled micronuclear and unscrambled macronuclear forms of the α-TBP gene from three spirorichous species, Holosticha sp., Uroleptus sp., and P. weissei. We also report the macronuclear forms of the α-TBP gene from Eschaneustyla sp. and U. grandis; micronuclear forms of these genes could not be determined due to the unavailability of micronuclear DNA from these species, which are difficult to culture in the laboratory. To determine the MDS/IES gene architecture and whether a gene is scrambled, we compared the micronuclear and macronuclear forms of the gene in the three species for which we recovered both versions. We integrated these data with macronuclear and macronuclear structures of orthologs in three previously studied Oxytrichidae: Oxytricha nova (also called Sterkiella nova), Oxytricha trifallax (also called Sterkiella histriomuscorum), and S. mytilus. Therefore, our study is the most comprehensive comparison thus far for the α-TBP gene. The comparison of homologous MDS and IES locations allowed us to trace the evolution of gene scrambling in these six species and to propose an evolutionary scrambling pathway for the α-TBP gene. The nonrandom and relatively compact scrambling pattern of the α-TBP gene, relative to more complex scrambled genes like DNA pol-α, allowed us to extract information at the sequence level and to construct an informative step-wise model that proposes how this gene became scrambled in the different lineages over evolutionary time. Furthermore, this model provides a testable hypothesis for how the molecular mechanism of developmental unscrambling might occur at the whole-gene level.

Germ Line and Somatic Organization of α-TBP

The micronuclear gene structures of α-TBP in the previously studied and closely related species O. nova, O. trifallax, and S. mytilus are divided into 14–17 MDSs that are generally scrambled in the odd-even order of 1-3-5-7-9-11-2-4-6-8-10-12-13-(14,...,17) (Mitcham, Lynn, and Prescott 1992; J. D. Prescott, DuBois, and D. M. Prescott 1998).
Here we examined putative intermediate species that diverged before the Oxytrichidae with known scrambled genes, but after the earlier diverging lineage \textit{U. grandis}, where only nonscrambled genes have been found (Hogan et al. 2001; Chang et al. 2005). The phylogeny of these species has been inferred from trees based on small- and large-subunit rDNA (Hewitt et al. 2003) and concatenated protein sequences (Chang et al. 2005). We found that the unscrambled \(\alpha\)-TBP gene in \textit{Holosticha} sp. is fragmented into six MDSs (fig. 1). The \(\alpha\)-TBP gene is scrambled in both \textit{Uroleptus} sp. (fig. 1) and \textit{P. weissei} (fig. 1) and fragmented into 16 MDSs in each species. The larger number of MDSs in scrambled genes suggests a possible trend of germ line scrambling to increase levels of fragmentation concomitant with rearrangement of coding regions.

Evolution of the Somatic Gene

We compared the evolution of the macronuclear versions of the \(\alpha\)-TBP gene in nine different species—the six species described above plus \textit{U. grandis}, \textit{Eschaneustyla} sp and \textit{S. mytilus}. This allowed us to compare the gene genealogy and evolution of the macronuclear \(\alpha\)-TBP protein (fig. 2) to the natural history of scrambling of the \(\alpha\)-TBP gene.

The genealogies constructed using full-length \(\alpha\)-TBP amino acid sequences (fig. 2), rDNA sequences (Hewitt et al. 2003; Chang et al. 2005), or concatenated proteins (Chang et al. 2005) are congruent and provide support for a branch that separates \textit{Uroleptus} sp. from the clade containing \textit{P. weissei} and the Oxytrichidae, with strongest support for this branch and highest resolution in the concatenated protein tree (Chang et al. 2005). Because recombination can generate different genealogies for different regions of a gene (reviewed in Posada, Crandall, and Holmes 2002), it is not surprising that certain regions of the \(\alpha\)-TBP gene bear slightly different genealogies, which increases noise in phylogenetic analysis based on a single gene. For example, some portions of the sequence suggest that \textit{P. weissei} and \textit{Uroleptus} may share a common ancestor (see below).

Evolution of Pointer Locations

The spiral model for intramolecular unscrambling of the \(\alpha\)-TBP gene (Mitcham, Lynn, and Prescott 1992) suggests that the molecule forms a DNA structure during development that may aid assembly of the gene, although multiple copies of the gene could also align and permit unscrambling via intermolecular rearrangements. The initial alignment of pointer pairs may guide formation of such a structure by placing adjacent pointers in close proximity with their cognate repeat, thus facilitating correct alignment of subsequent pointer pairs (Landweber, Kuo, and Curtis 2000).

Alignment of the amino acid and nucleotide sequences of macronuclear \(\alpha\)-TBP genes reveals that some pointer locations have shifted across species, while others are tightly conserved (fig. 1, tables 1–3, and Figure S1, Supplementary Material online). Comparison to the \textit{O. nova} protein crystal structure reveals that functionally important residues of the protein (Horvath et al. 1998) lie both in

\[ \text{FIG. 1.—The micronuclear } \alpha \text{-TBP gene is nonscrambled in } \text{Holosticha} \text{ sp. and scrambled in } \text{Uroleptus} \text{ sp. and } \text{Paraurostyla weissei. Processing of the odd-even nonrandomly scrambled micronuclear segments produces the macronuclear genes shown below. Numbers at the start of each MDS indicate sizes in base pairs. MDSs, but not IESs (represented by straight or curved lines), are drawn to scale and approximately aligned. Colors highlight homologous scrambled/nonscrambled regions. Macronuclear organization of } \alpha \text{-TBP orthologs are schematically aligned. In all the Oxytrichidae and } \text{P. weissei}, \text{ the region homologous to MDS 2 (blue) is not fragmented, whereas the region homologous to MDS 12-14 (green) is fragmented. The fragmentation pattern of the homologous MDSs in } \text{Uroleptus} \text{ sp., MDS 2-4 and MDS 14, is the reverse of the fragmentation pattern in the Oxytrichidae species and } \text{P. weissei}. \text{ Scrambled and non-scrambled regions of the genes are indicated, and these correspond to the N- and C-terminal domains of the protein that interact with DNA and } \beta \text{-TBP, respectively (Horvath et al., 1998).} \]
conserved pointer locations and outside of pointer locations entirely and that pointers fall in both structured and unstructured regions of the protein (Horvath et al. 1998) as well as in the boundaries between structured and unstructured regions (Figure S1, Supplementary Material online, annotated by H for α-helices and S for β-sheets). We found no significant correlation between pointer locations and the functional and structural (α-helices and β-sheets) domains of the O. nova α-TBP crystal structure.

Most sequences of pointer pairs are consistent with formation by a single ancestral event, followed by mutational sliding, as noted previously (DuBois and Prescott 1997). In this study, we consider MDSs with pointers that overlap by at least 1 nt with a pointer in one of the other four species (as aligned in Figure S1, Supplementary Material online) as homologous. This class of sequences does not shift by more than 25 bp in this study, from the first position of the most upstream pointer to the last nucleotide of the furthest downstream pointer. In all scrambled α-TBP genes, the locations of scrambled pointers are considerably more conserved than nonscrambled pointers (fig. 1, boundaries of boxes representing scrambled MDSs, as compared to boxes representing nonscrambled MDSs), consistent with greater constraint on scrambled pointers and production of some nonscrambled IESs by independent DNA insertion events (fig. 3). Similarly, the pointer locations that are conserved across the five scrambled orthologs are generally not conserved in the nonscrambled Holosticha gene, suggesting independent IES insertions in this nonscrambled lineage (fig. 3).

Pointer sliding allows the pointer sequence and location to vary across species, despite constraint on the coding region (DuBois and Prescott 1997). Diversity in pointer sequences (tables 1–3) implies that the core machinery for IES excision and DNA rearrangement, relatively conserved across spirotrichous species, may act upon approximately 60,000 (Prescott 1994) to 200,000 (DuBois and Prescott 1997) sets of direct repeats endogenous to a genome, responding to myriad sliding sets of direct repeats. This suggests that (1) there is a component of the unscrambling machinery that evolves rapidly, as rapidly as and to accommodate the observed substitutions reflected in pointer sliding (it has been suggested that ciliates evolve rapidly; Lynch and Conery 2003), (2) there is an unrecognized set of features common to all IESs or pointers, (3) recognition of paired pointers is sequence independent, or (4) the pointers are not sufficient to guide recombination, and instead the unscrambling machinery includes another nucleic acid component, a result of somatic-germ line genome communication, that dictates the specificity of rearrangements, as others have suggested, and evidence implicates small RNAs (Yao, Fuller, and Xi 2003). We favor a combination of (2) and (4) because the sample size of known IESs is still extremely small, compared to the size of the respective ciliate genomes, and there is increasing evidence that small RNAs guide excision of IESs in different ciliate species. In any case, the net result is an excision and unscrambling machinery that supports and is highly tolerant of a dynamic germ line genome. A supporting piece of evidence for the ability of this machinery to join any pair of direct repeats in the germ line genome, even at noncanonical pointers, lies in the discovery of a partially unscrambled α-TBP molecule in Oxytricha trifallax with cryptic pointers joining MDS 1–MDS 16 and all intervening MDSs and IESs excised (L. C. Wong and L. F. Landweber, unpublished data).

Evolution of Germ Line Organization in Six Species

Comparison of IES locations and homologous MDSs of the five scrambled α-TBP genes reveals that species with more similar evolutionary histories (inferred by phylogenetic analysis based on both small- and large-subunit rDNA [Hewitt et al. 2003] and α-TBP or concatenated protein sequences) tend to share local scrambling patterns despite accumulation of comparable global scrambling complexity between earlier and later diverging species.

For example, in this study, we find three distinct MDS-splitting events across the species (represented in figs. 1 and 3 by different colors). Two of these germ line–scrambling events distinguish the Oxytrichidae species plus P. weissei from Uroleptus, thus presenting evidence that these MDSs (figs. 1 and 3, blue and green MDSs) became fragmented and scrambled independently in the two different lineages, Paraurostyla weissei and Uroleptus, however, show more similarity in the third scrambled region (fig. 1, purple), and also in their IES sequences in this general region, suggesting that the split architecture in the purple region is either a shared derived character, possibly reflecting a paralogous evolutionary history of this region derived from an ancient segregating allele, or a shared ancestral character that subsequently diverged in the Oxytrichidae. This is also one of the few regions that distinguish P. weissei from the
Oxytrichidae. The second model requires germ line fusion of two of the purple MDSs in an ancestral oxytrichid, leading to the pattern in *O. trifallax*, and then finally fusion of the last two purple MDSs in an ancestor of *S. mytilus* and *O. nova* (fig. 3). Because the first of these fusions occurs between two scrambled MDSs, it requires translocation of the middle purple MDS, with concomitant creation of the nonscrambled IES in *O. trifallax* by recombination between two IESs. Because the first model would require insertion of a non-scrambled IES in *O. trifallax* at coincidentally the same location as a scrambled pointer in *P. weissei* and *Uroleptus*, we favor the model with fusions illustrated in figure 3.

The germ line architectures of the blue and green regions are consistent with published phylogenies of sprotichrous ciliates (Hewitt et al. 2003). In the green region, the later diverging group of the Oxytrichidae plus *P. weissei* are more scrambled compared to the earlier diverging *Uroleptus*, with MDS 14 in *Uroleptus* split into three MDSs in the Oxytrichidae and *P. weissei* (fig. 1). The blue region presents a reverse case, with MDS 2 in the Oxytrichidae and *P. weissei* split into three MDSs 2–4 in *Uroleptus* (fig. 1). The scrambled architectures of these two regions suggests that the Oxytrichidae and *P. weissei* share a common ancestor, after diverging from *Uroleptus* sp. Thus, while more closely related species share similar local scrambling complexity, global scrambling complexity and the resulting number of MDSs remain comparable across distantly related species. This is true of the three previously published sets of scrambled genes to date, with scrambled orthologs of α-TBP fragmented into 14–17 MDSs (J. D. Prescott, DuBois, and D. M. Prescott 1998), actin into 8–11 MDSs (Hogan et al. 2001), and DNA pol-α into 35–51 MDSs (Hoffman and Prescott 1997; Landweber, Kuo, and Curtis 2000; Chang et al. 2005).

While comparison of the sequences and locations of pointers and MDSs provided a wealth of information about the evolutionary history of scrambled genes, comparisons of IESs are confounded by their rapid rate of divergence, making it difficult to align and identify homologous IESs or nucleotides within an IES. Furthermore, IESs can vastly differ in their sizes and sequences across species (DuBois and Prescott 1997). Therefore, the similarity between the IES sequences in the general region of MDS 10–12 in *Uroleptus* and MDS 8–10 in *P. weissei* suggests that fusion of the homologous MDSs in the oxytrichids is a possible recent evolutionary event.
A Scrambling Pathway for α-TBP

Based on comparisons of α-TBP gene structures in Holosticha sp., Uroleptus sp., O. nova, O. trifallax, and S. mytilus, we propose a possible evolutionary scrambling pathway for this gene (fig. 3). It is possible that the developmental unscrambling pathway is a reverse of the germ line recombination in the evolutionary scrambling pathway. A nonscrambled gene that may have been a close germ line recombination in the evolutionary scrambling of the developmental unscrambling pathway is a reverse of the scrambling pathway for this gene (fig. 3). It is possible that forms of the gene observed in the extant species

Conclusions

Molecular evolutionary studies of five scrambled and one nonscrambled α-TBP orthologs suggest that de novo splitting and fusing of coding regions occur independently in lineages after divergence from a common ancestor.
Scrambled $z$-TBP genes show comparable global scrambling complexity among the five lineages but have local differences in scrambling patterns. The DNA excision and unscrambling machinery in spirotrichous ciliates lets them regenerate a functional somatic genome from an encrypted germ line, and it makes them both flexible and apparently robust to DNA rearrangements that would cripple the genomes of most other organisms. Germ line fragmentation and rearrangement of scrambled genes in spirotrichous ciliates appear to be very fluid processes which, like most evolutionary and biological processes, do not follow a universal path but exhibit diverse peculiarities, as illustrated by the unique germ line architectures of the $z$-TBP genes in each species in this study.

**Supplementary Material**

Figure S1 and Table S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

**Acknowledgments**

The authors wish to thank Wei-Jen Chang and Mann Kyoon Shin for the generous gift of DNA used in this study and Wei-Jen Chang for technical advice; Andre R. O. Cavalcanti for help in annotating the data and use of software prior to publication; Danny van Noort for help with illustration; Nicholas A. Stover and Thomas G. Doak for critical reading of the manuscript and discussions; Han Liang, Qing Bao, and Thomas H. Clarke for help with phylogenetic, protein, and DNA sequence analysis; and two anonymous referees for their comments. This work was supported by National Institute of General Medical Sciences grant GM59708 and National Science Foundation grant 0121422.

---

**Literature Cited**


Martin Embley, Associate Editor
Accepted December 14, 2005