1-1-2014

Genome Structure Drives Patterns of Gene Family Evolution in Ciliates, a Case Study Using *Chilodonella uncinata* (Protista, Ciliophora, Phyllopharyngea)

Feng Gao
Smith College

Weibo Song
Smith College

Laura A. Katz
Smith College, lkatz@smith.edu

Published in final edited form as:
Evolution. 2014 August ; 68(8): 2287–2295. doi:10.1111/evo.12430.

Genome structure drives patterns of gene family evolution in ciliates, a case study using *Chilodonella uncinata* (Protista, Ciliophora, Phyllopharyngea)

Feng Gao[1,2], **Weibo Song**[1], and **Laura A. Katz**[1,3,*]

[1]Department of Biological Sciences, Smith College, Northampton, MA 01063, USA

[2]Laboratory of Protozoology, Institute of Evolution & Marine Biodiversity, Ocean University of China, Qingdao 266003, China

[3]Program in Organismic and Evolutionary Biology, UMass-Amherst, Amherst, MA 01003, USA

Abstract

In most lineages, diversity among gene family members results from gene duplication followed by sequence divergence. Because of the genome rearrangements during the development of somatic nuclei, gene family evolution in ciliates involves more complex processes. Previous work on the ciliate *Chilodonella uncinata* revealed that macronuclear β-tubulin gene family members are generated by alternative processing, in which germline regions are alternatively used in multiple macronuclear chromosomes. To further study genome evolution in this ciliate, we analyzed its transcriptome and found that: 1) alternative processing is extensive among gene families; and 2) such gene families are likely to be *C. uncinata*-specific. We characterized additional macronuclear and micronuclear copies of one candidate alternatively processed gene family -- a protein kinase domain containing protein (PKc) -- from two *C. uncinata* strains. Analysis of the PKc sequences reveals: 1) multiple PKc gene family members in the macronucleus share some identical regions flanked by divergent regions; and 2) the shared identical regions are processed from a single micronuclear chromosome. We discuss analogous processes in lineages across the eukaryotic tree of life to provide further insights on the impact of genome structure on gene family evolution in eukaryotes.

Keywords

genome evolution; gene family; alternative processing; ciliate; macronucleus; micronucleus

Introduction

Gene families, functionally related genes formed by duplication, are very important for organismal complexity (Ohta 1987, 2000; Demuth and Hahn 2009). In most lineages,

*Correspondence: Laura A. Katz, Address: Department of Biological Sciences, Smith College, 44 College Lane, Northampton, MA, 01063. lkatz@smith.edu.

Note: All the sequences newly reported in this study were deposited in GenBank database under accession numbers KJ000261 – KJ000272, KJ626299 and KJ626300.

diversity among gene family members results from gene duplication followed by pseudogenization, neofunctionalization (Ohta 1991; Walsh 1995) and /or subfunctionalization (Lynch and Conery 2000). Under these scenarios, genes duplicate, mutate and then are retained or lost depending on their resulting function. Such processes can be extensive in some genomes including ciliates, a diverse clade of microbial eukaryotes. Ciliates have among the largest gene families, with an estimated 1,565 potassium channel genes in the completed genome of *Tetrahymena thermophila* (Eisen et al. 2006).

The evolution of gene family members in ciliates has to be interpreted in light of their dual genomes, the presence of both germline and somatic genomes within each cell. During sexual conjugation, a meiotic product of the micronucleus is exchanged between two mating cells to form a genetically novel zygotic nucleus. The new zygotic nucleus divides by mitosis and then develops into either a micronucleus or a somatic macronucleus. The macronucleus is transformed through a series of chromosomal rearrangements, including fragmentation, elimination of internal excised sequences, and amplification (Prescott 1994; Katz 2001; Riley and Katz 2001; Katz et al. 2003; Chalker 2008; Heyse et al. 2010; Chalker and Yao 2011; Nowacki et al. 2011; Goldman and Landweber 2012). Gene scrambling, the presence of fragmented coding domains (termed macronuclear destined sequences) in non-canonical order in the micronucleus, has been described in ciliates from two classes, Spirotrichea (Prescott and Greslin 1992; Curtis and Landweber 1999; Nowacki and Landweber 2009) and Phyllopharyngea (represented by *Chilodonella uncinata*, Katz and Kovner 2010).

In the phyllopharyngean ciliate *Chilodonella uncinata*, the focus of this paper, gene family members evolve through complex processes involving gene scrambling plus alternative processing of shared micronuclear regions. For example, within the macronucleus of *C. uncinata*, three of the five β-tubulin macronuclear gene family members share regions identical at the nucleotide level that are flanked by more variable regions. Analyses of corresponding micronuclear loci reveal that these three β-tubulin macronuclear sequences are assembled by alternative processing of scrambled micronuclear loci, with some regions being used in multiple macronuclear chromosomes (Katz and Kovner 2010).

In the present study, we used the combination of analyses of high throughput *C. uncinata* transcriptome data and polymerase chain reaction (PCR) to assess patterns of gene family evolution. We identified candidate alternatively processed gene families from the *C. uncinata* transcriptome and find that alternative processing is extensive among gene families within this ciliate. We then explored one of these examples by characterizing the macronuclear and micronuclear protein kinase domain containing protein (PKc) family members generated from two geographically isolated strains of *C. uncinata*. We also find that there are multiple alternatively processed members of the PKc gene family, some of which share nine identical nucleotide regions flanked by more variable regions. We present two models to explain alternative processing in *C. uncinata*.

Material and Methods

Ciliate culturing and DNA extraction

We maintain two previously characterized strains of *C. uncinata*, Pol = ATCC PRA-256, USA = USA-SC2, following protocols in Katz et al. (2011). To isolate DNA, cultures were treated overnight with antibiotics and cells were pelleted by spinning at 5,000 rpm for 20 min. Genomic DNA was extracted using phenol/chloroform following standard protocols (Ausubel et al. 1993). Micronuclear DNA was isolated according to Katz and Kovner (2010). Briefly, micronuclear DNA was isolated by gel electrophoresis using Low Melt UltraClean™ Agarose (Mobio15005-50, Carlsbad, CA) after digesting with Bal-31 Nuclease (New England Biolabs M02135, Ispwich, MA) to enrich micronuclear DNA. Gel isolated micronuclear DNA was purified using β-agarase (New England Biolabs M03925).

Transcriptome data analysis

Transcriptome data of the Pol strain of *C. uncinata* came from Grant et al. (2012). After assembly, 9029 contigs and single reads were passed to custom python scripts that used a BLAST all against all strategy to generate sequence pairs using default BLASTN parameters. Pairs with e-value higher than 0.01 were selected while pairs united by unprocessed linkers were removed. The resulting 3172 pairs (1288 sequences) that shared identical regions were binned into clusters, which resulted in 448 clusters. All the clusters were assessed further by eye using Megalign (DNAStar) to explore whether they are candidate alternatively processed gene families, identified by the sharing of two more ≥25 base pair regions of identity. We then performed BLASTX analysis on the longest sequences of the clusters that show strong signatures of alternative processing patterns to assess their function. DNAsp (Librado and Rozas 2009) was used to perform sliding window analysis to calculate average pairwise differences (π) of the longest two sequences in the clusters that show strong signatures of alternative processing patterns. Sliding window analyses were performed with a 20 base pair window and a 5 base pair step.

Traditional PCR and cloning

We choose one candidate alternatively processed gene family, protein kinase domain containing protein (PKc), to explore in two *C. uncinata* strains. Primers for macronuclear PKc gene family members were designed from the highly conserved regions shared among PKc contigs from the transcriptome data of the Pol strain (Figure S1). The PKc gene of Pol strain and USA strain was then amplified using Phusion Hot Start High Fidelity DNA Polymerase (Finnzymes F 540L, Finland). Amplified products were cloned using Zero Blunt TOPO kits (Invitrogen, CA), and screened using the polymerase TaqGold (Applied Biosystems, CA).

Genome walking PCR and cloning

Micronuclear sequences of PKc for USA strain were amplified using Seegene's DNA Walking SpeedUp™ kit (K1052; Seegene, Rockville, MD). PCR amplification was performed following Seegene kit protocol using kit primers and gene-specific primers designed for this study (Figure S2, Table S1). Genome walking PCR products were cloned

using TA TOPO cloning kits (Invitrogen 45–0641), and screened using the polymerase TaqGold (Applied Biosystems, CA).

Sequencing and data analysis

Sequences were generated using the BigDye terminator v3.1 cycle sequencing kit from PE Applied Biosystems (4337455, Wellesley, MA). Reactions were cleaned using gel filtration columns from Edge Biosystems (42453, Gaithersburg, MD) and analyzed on a PerkinElmer ABI-3100 automated sequencer at the Center for Molecular Biology (Smith College, Northampton, MA). Additional sequencing reactions and sequencing were performed at the Penn State Genomics Core Facility (University Park, PA). Contigs were assembled in Seqman (DNAStar) and all polymorphisms were confirmed by eye. All the sequences newly reported in this study were deposited in GenBank database under accession numbers KJ000261 – KJ000272, KJ626299 and KJ626300.

Seaview v. 4.2.4 (Gouy et al. 2010) and Megalign (DNAStar) were used to create alignments. Genealogies based on nucleotide alignments were estimated using RAxML-HPC2 v7.2.8 (Stamatakis 2006; Stamatakis et al. 2008) on CIPRES Science Gateway (Miller et al. 2010). Support came from a majority rule consensus tree of 1000 bootstrap replicates. Average pairwise differences (π) of the sequences were calculated using DNAsp (Librado and Rozas 2009). Sliding window analyses were performed with a 20 base pair window and a 5 base pair step.

Results

Transcriptome data analyses

We set out to assess alternatively processed gene families in the transcriptome data from the Pol strain of *C. uncinata*, starting from assembled 454 data generated for Grant et al. (2012). Using the custom python script, 1288 out of 9029 sequences (contigs + individual reads) generated 448 clusters at 70% similarity (Table 1). A total of 142 clusters (662 sequences) are putative gene families and 69 of the gene families contain at least three representatives. The remaining 306 clusters (667 sequences) contain sequences that differ by indels representing either unprocessed micronuclear-limited regions or canonical eukaryotic introns (most of which are about 19–25 bp long), or sequences differing by one or few nucleotides (i.e. alleles or the results of sequence error). Of the 142 clusters that are putative gene families, 48 clusters (325 sequences) show strong signatures of alternative processing patterns, defined as more than one region of identity (at least 25 bp) shared by the sequences (Table S2, Figs 1, S3–50), 64 clusters (209 sequences) show weak signatures (e.g. there is only one region of identity shared by the sequences), while 30 clusters (87 sequences) show no clear signature of alternative processing. Using a sliding window analysis, we compared the longest two sequences of the 48 clusters that show strong signatures of alternative processing patterns. Pairwise comparisons of these sequences reveal islands of identity nested among highly divergent regions (Figs 1, S3–50).

We performed BLASTX analysis on the longest sequences of the 48 clusters that show strong signatures of alternative processing. Nine clusters hit named proteins and five clusters

hit hypothetical proteins, with the best hit generally coming from one of the four completed ciliate genomes (e.g. *Tetrahymena thermophila*, *Paramecium tetraurelia*, *Ichthyophthirius multifiliis*, or *Oxytricha trifallax*, Table 2). The remaining 34 clusters have no significant BLAST hit, suggesting they may be *C. uncinata* or phyllopharyngean specific genes.

Macronuclear PKc sequence analyses

We chose one of the candidate alternatively processed gene families, a protein kinase domain containing protein (PKc), to characterize from the macronuclear (somatic) genome of two strains of *C. uncinata* (Pol = ATCC PRA-256, USA = USA-SC2) since previous analyses have indicated these strains may represent cryptic species (Katz et al. 2011). We had identified eight PKc gene family members in the Pol strain transcriptome data with variable read numbers from one to 31 (Table 3). Using primers designed from the highly conserved regions shared among PKc contigs from the transcriptome data (Figure S1), we analyzed 50 clones from the Pol strain and 159 clones from the USA strain. For the Pol strain, 50 clones represent seven haplotypes (Pol1-7), six of which are present in the Pol transcriptome data. We did not detect either Pol8 or Pol9 by PCR, suggesting either primer bias or low copy number for these gene family members. For the USA strain, 159 clones represent five haplotypes (USA1-5).

The PKc family members show two patterns in terms of the number and location of shared identical regions (Figs. 2, S3). The first pattern is found in haplotypes Pol1-5 and USA1-3, all of which share nine identical regions nested among more divergent regions (Pattern I, Figs. 2A, 2C, S3). The second pattern is in haplotypes Pol6-7 and USA4-5 that share only two identical regions (Pattern II, Figs. 2B and S3). We performed genealogical analyses using only the unique regions from each family member, as these regions have unique histories compared to the shared regions. The genealogy based on nucleotide data reveals that the haplotypes from the same patterns group together (Fig. 3). The topology of pattern II haplotypes is consistent with a single duplication event prior to the divergence of the Pol and USA strains while for pattern I, the topology suggests a complex history of duplication, divergence and loss (Fig. 3). Interpretation of pattern I is further complicated by the fact that we may have missed rare haplotypes in the transcriptome sequencing and PCR survey of the macronuclear genomes.

Because each gene resides on its own chromosome in the macronucleus of *C. uncinata* (Riley and Katz 2001; Juranek and Lipps 2007), it is possible to compare the number of copies of any given sequence in the macronucleus with its expression level (Bellec and Katz 2012). Intriguingly, our estimate of copy number of haplotypes in the macronucleus of the Pol strain, approximated by number of clones using traditional PCR, does not simply correlate to read number in the transcriptome. For example, within pattern I, the most (14) clones of Pol5 are represented by a single read in the Pol strain transcriptome, while the haplotype Pol8 that was the most frequent (14 reads) in the transcriptome was not found using traditional PCR (Table 3).

Micronuclear PKc sequence analyses

Using walking PCR, we characterized a micronuclear region corresponding to PKc haplotype USA1 that is scrambled with one region being inverted and separated from the other region by a 170 bp internal excised sequence (Fig. 4). We also characterized part of the micronuclear copy corresponding to haplotype USA2; this region contains only a portion of USA2-specific regions, some of which are reversed and scrambled (Fig. 4). Walking PCR yielded an additional micronuclear PKc locus (MIC USA6) that did not correspond to any of the haplotypes determined by PCR or transcriptome analyses. We subsequently found the putative MAC copy of this locus (MAC USA6) using traditional PCR with specific primers. This locus appears to be a pseudogene since it has a frame shift mutation. Additional sequences were found in only a single (walking) PCR that included some of the unique regions of either USA3 or USA6; since we were not able to confirm these sequences by multiple PCRs, we interpret them as either partially processed chromosomes and/or PCR recombinants (data not shown).

Discussion

Alternative processing is extensive within *C. uncinata* and may involve many lineage specific genes. Analyses of Pol strain transcriptome data indicate that 112 out of 142 clusters (~ 80%) of putative gene family members show a signature of alternative processing. About 70% of the alternatively processed gene families that show strong signatures have no significant BLASTX hit (Table 2), not even to the fully sequenced ciliate genomes of *Paramecium*, *Ichthyophthirius* and *Tetrahymena* (Oligohymenophorea) and *Oxytricha* (Spirotrichea). The lack of homologs in other eukaryotes suggests that these genes are limited to *C. uncinata* and its relatives. Our interpretation of patterns here are limited by the relatively low coverage of Pol transcriptome data (Grant et al. 2012) and the evidence that copy numbers of the expressed gene family members vary considerably (Bellec and Katz 2012). Based on these limitations, we predict that many more alternatively processed gene families may exist in *C. uncinata*.

Polymerase chain reaction based analyses of one candidate alternatively processed gene family, a protein kinase domain containing protein (PKc), identified multiple alternative processed members present in the macronuclei of the Pol and USA strains of *C. uncinata* (Table 3). The PKc haplotypes fall into two patterns, one consistent with a single duplication event prior to divergence of the strains (Pattern II; Fig. 3) and one with a more complex history of duplication, divergence and loss (Pattern I; Fig. 3). Interestingly, for pattern I, copy number in the macronucleus and transcriptome does not appear to be simply correlated as the more abundant macronuclear PKc members based on PCR have lower levels of expression based on number of reads in the transcriptome. Though uncertainty exists in both estimates, this observation is consistent with qPCR analyses of the β-tubulin gene family in the same strains where more abundant macronuclear chromosomes have lower expression levels and suggests the potential epigenetic regulation mechanism of copy number in *C. uncinata* (Bellec and Katz 2012).

We hypothesize that PKc gene family members in *C. uncinata* are generated through gene scrambling and alternative processing of shared micronuclear regions as was found for

macronuclear β-tubulin haplotypes (Katz and Kovner 2010). The macronuclear PKc gene family members share identical regions nested among highly divergent regions. Compared to the amplified and gene-sized macronuclear genome, the micronuclear genome is diploid and has more complex structures (e.g. includes internally eliminated sequences and scrambled regions), making it difficult to characterize micronuclear genes. We have tried many times but failed to characterize all the micronuclear regions of all of the PKc gene family members. However, our current data of micronuclear copies of haplotypes USA1 and USA2 show that the shared regions are processed from a single micronuclear locus (Fig. 4).

We propose that the generation of PKc macronuclear gene family members come from multiple micronuclear loci, with the shared regions processed from a single micronuclear region (Fig. 5). Such a model is consistent with the pattern observed in micronuclear regions that encode for the alternatively processed macronuclear copies of β-tubulin: regions of identity among β-tubulin gene family members are processed from one micronuclear locus while more divergent sequences are from multiple, divergent micronuclear loci (Katz and Kovner 2010). Alternatively, the sharing of identical PKc regions could be due to non-allelic gene conversion, which would lead to the prediction that paralogs resulting from gene conversion exist in the micronucleus. Instead the micronuclear loci we characterized require alternative processing of scrambled regions to generate the observed macronuclear gene family members. We do recognize the limitation of our PCR-based approach as full genome sequencing might reveal additional micronuclear loci that offer a different explanation for the unusual pattern of shared and divergent regions present among PKc macronuclear gene families.

The generation of alternatively processed gene family members relies on the unique genome structure of ciliates like *C. uncinata* in which gene-sized macronuclear chromosomes are generated from long micronuclear chromosomes. The presence of “giant” highly polytenized chromosomes during macronuclear development (Ammermann 1987) may underlie the use of identical germline micronuclear regions to form the transcriptionally-active macronucleus. This is because multiple copies of micronuclear sequences present in giant chromosomes can be alternatively processed to yield macronuclear chromosomes for gene family members that vary in the number and location of shared, alternatively processed micronuclear regions. The subsequent extensive fragmentation of micronuclear chromosomes breaks down the linkage between genes, and may enable selection to act on individual gene family members (Zufall et al. 2006). Combined, these mechanisms enable ciliates to explore protein space in novel manners.

The unusual patterns of chromosome processing in ciliates do have analogs in lineages across the eukaryotic tree of life. For example, the switching of variant surface glycoprotein (VSG) to generate antigenic variation in *Trypanosoma brucei* uses DNA rearrangements of >1,000 VSG genes (Stockdale et al. 2008). Similarly, recombination of V(D)J regions generate diversity in immunoglobulins in human and other vertebrates (Nemazee 2006). Another example might be exon shuffling, a natural process of creating new combinations of exons by intronic recombination, which is the most efficient way of constructing modular proteins (Patthy 1996). In ciliates, protein construction can happen during the process of macronuclear development through alternative processing because of its special dual

genomes. This study adds to the growing literature on gene family evolution and yields important insights into the impact of genome structure on protein evolution in eukaryotes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank Jessica R. Grant, Jie Huang and Gladys Palaguachi for technical help and data analyses. This work is supported by an NIH grant to LAK (1R15GM097722).

Reference

Ammermann D. Giant chromosomes in ciliates. Results Probl. Cell Differ. 1987; 14:59–67. [PubMed: 3112879]

Ausubel, FM.; Brent, R.; Kingston, RE.; Moore, DD.; Seidman, JG.; Smith, JA.; Struhl, K. Current protocols in molecular biology. New York: Wiley-Liss; 1993.

Bellec L, Katz LA. Analyses of chromosome copy number and expression level of four genes in the ciliate *Chilodonella uncinata* reveal a complex pattern that suggests epigenetic regulation. Gene. 2012; 504:303–308. [PubMed: 22588027]

Chalker DL. Dynamic nuclear reorganization during genome remodeling of *Tetrahymena*. Biochim. Biophys. Acta. 2008; 1783:2130–2136. [PubMed: 18706458]

Chalker DL, Yao MC. DNA elimination in ciliates: transposon domestication and genome surveillance. Annu. Rev. Genet. 2011; 45:227–246. [PubMed: 21910632]

Curtis EA, Landweber LF. Evolution of gene scrambling in ciliate micronuclear genes. Ann. N. Y. Acad. Sci. 1999; 870:349–350. [PubMed: 10415495]

Demuth JP, Hahn MW. The life and death of gene families. BioEssays. 2009; 31:29–39. [PubMed: 19153999]

Eisen JA, Coyne RS, Wu M, Wu D, Thiagarajan M, Wortman JR, Badger JH, Ren Q, Amedeo P, Jones KM, Tallon LJ, Delcher AL, Salzberg SL, Silva JC, Haas BJ, Majoros WH, Farzad M, Carlton JM, Smith RK Jr, Garg J, Pearlman RE, Karrer KM, Sun L, Manning G, Elde NC, Turkewitz AP, Asai DJ, Wilkes DE, Wang Y, Cai H, Collins K, Stewart BA, Lee SR, Wilamowska K, Weinberg Z, Ruzzo WL, Wloga D, Gaertig J, Frankel J, Tsao CC, Gorovsky MA, Keeling PJ, Waller RF, Patron NJ, Cherry JM, Stover NA, Krieger CJ, del Toro C, Ryder HF, Williamson SC, Barbeau RA, Hamilton EP, Orias E. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. PLoS Biol. 2006; 4:e286. [PubMed: 16933976]

Goldman AD, Landweber LF. *Oxytricha* as a modern analog of ancient genome evolution. Trends Genet. 2012; 28:382–388. [PubMed: 22622227]

Gouy M, Guindon S, Gascuel O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. 2010; 27:221–224. [PubMed: 19854763]

Grant JR, Lahr DJG, Rey FE, Burleigh JG, Gordon JI, Knight R, Molestina RE, Katz LA. Gene discovery from a pilot study of the transcriptomes from three diverse microbial eukaryotes: *Corallomyxa tenera*, *Chilodonella uncinata*, and *Subulatomonas tetraspora*. Protist Genomics. 2012; 1:3–18.

Heyse G, Jonsson F, Chang WJ, Lipps HJ. RNA-dependent control of gene amplification. Proc. Natl. Acad. Sci. USA. 2010; 107:22134–22139. [PubMed: 20974970]

Juranek SA, Lipps HJ. New insights into the macronuclear development in ciliates. Int Rev Cytol. 2007; 262:219–251. [PubMed: 17631190]

Katz LA. Evolution of nuclear dualism in ciliates: a reanalysis in light of recent molecular data. Int J Syst Evol Microbiol. 2001; 51:1587–1592. [PubMed: 11491362]

Katz LA, DeBerardinis J, Hall MS, Kovner AM, Dunthorn M, Muse SV. Heterogeneous rates of molecular evolution among cryptic species of the ciliate morphospecies *Chilodonella uncinata*. J. Mol. Evol. 2011; 73:266–272. [PubMed: 22258433]

Katz LA, Kovner AM. Alternative processing of scrambled genes generates protein diversity in the ciliate *Chilodonella uncinata*. J. Exp. Zool. Part B. 2010; 314:480–488.

Katz LA, Lasek-Nesselquist E, Snoeyenbos-West OL. Structure of the micronuclear alpha-tubulin gene in the phyllopharyngean ciliate *Chilodonella uncinata*: implications for the evolution of chromosomal processing. Gene. 2003; 315:15–19. [PubMed: 14557060]

Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009; 25:1451–1452. [PubMed: 19346325]

Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science. 2000; 290:1151–1155. [PubMed: 11073452]

Miller, MA.; Pfeiffer, W.; Schwartz, T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. New Orleans, LA: Proceedings of the Gateway Computing Environments Workshop (GCE); 2010. p. 1-8.

Nemazee D. Receptor editing in lymphocyte development and central tolerance. Nat. Rev. Immunol. 2006; 6:728–740. [PubMed: 16998507]

Nowacki M, Landweber LF. Epigenetic inheritance in ciliates. Curr. Opin. Microbiol. 2009; 12:638–643. [PubMed: 19879799]

Nowacki M, Shetty K, Landweber LF. RNA-mediated epigenetic programming of genome rearrangements. Annu. Rev. Genomics Hum. Genet. 2011; 12:367–389. [PubMed: 21801022]

Ohta T. Simulating evolution by gene duplication. Genetics. 1987; 115:207–213. [PubMed: 3557113]

Ohta T. Multigene families and the evolution of complexity. J. Mol. Evol. 1991; 33:34–41. [PubMed: 1909373]

Ohta T. Evolution of gene families. Gene. 2000; 259:45–52. [PubMed: 11163960]

Patthy L. Exon shuffling and other ways of module exchange. Matrix Biol. 1996; 15:301–312. [PubMed: 8981326]

Prescott DM. The DNA of ciliated protozoa. Microbiol. Rev. 1994; 58:233–267. [PubMed: 8078435]

Prescott DM, Greslin AF. Scrambled actin I gene in the micronucleus of *Oxytricha nova*. Dev. Genet. 1992; 13:66–74. [PubMed: 1395144]

Riley JL, Katz LA. Widespread distribution of extensive chromosomal fragmentation in ciliates. Mol. Biol. Evol. 2001; 18:1372–1377. [PubMed: 11420375]

Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006; 22:2688–2690. [PubMed: 16928733]

Stamatakis A, Hoover P, Rougemont J. A rapid bootstrap algorithm for the RAxML web servers. Syst. Biol. 2008; 57:758–771. [PubMed: 18853362]

Stockdale C, Swiderski MR, Barry JD, McCulloch R. Antigenic variation in *Trypanosoma brucei*: joining the DOTs. PLoS Biol. 2008; 6:e185. [PubMed: 18666832]

Walsh JB. How often do duplicated genes evolve new functions? Genetics. 1995; 139:421–428. [PubMed: 7705642]

Zufall RA, McGrath CL, Muse SV, Katz LA. Genome architecture drives protein evolution in ciliates. Mol. Biol. Evol. 2006; 23:1681–1687. [PubMed: 16760419]

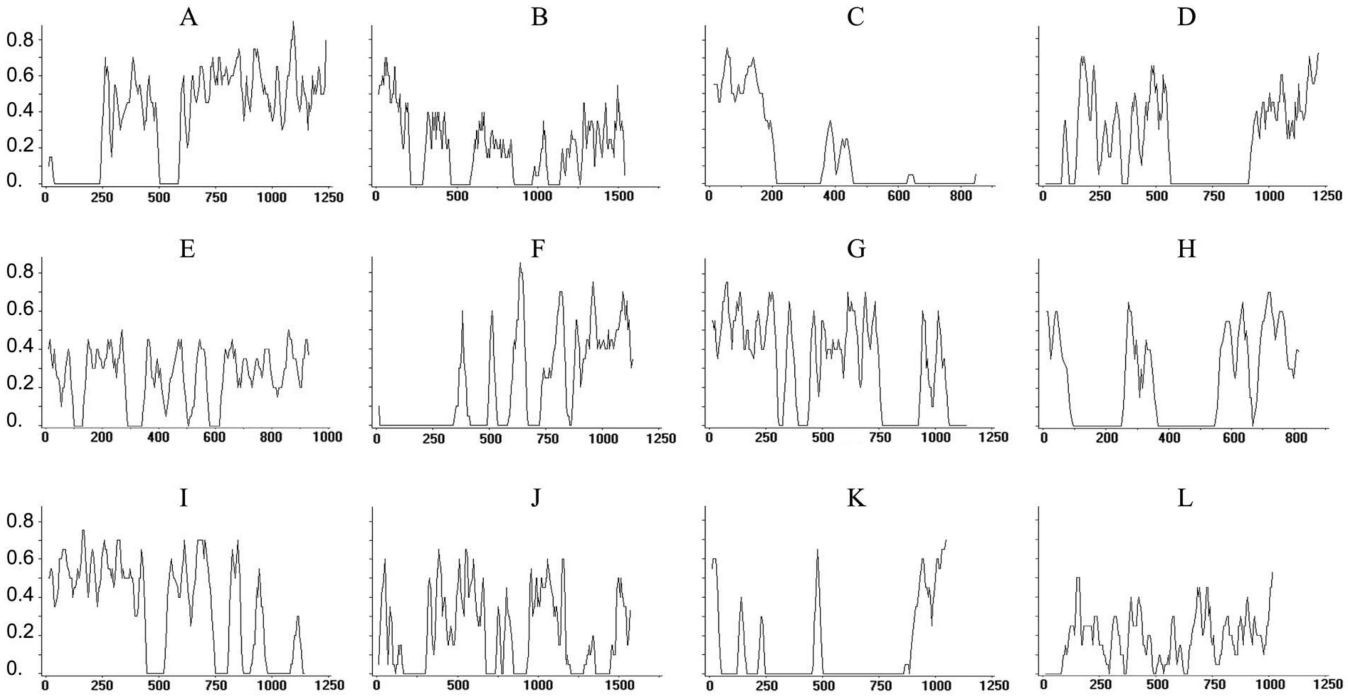

Figure 1.

Twelve examples of sequence comparisons between putative alternatively processed genes show islands of identity nested among highly divergent sequences. Graphs represent sliding window comparisons of the two longest sequences in gene family clusters (overlap > 800 bp) that have strong signatures of alternative processing. Candidates for alternative processing have at least two regions of identity that are ≥25bp. Y-axis =π, all drawn to same scale and calculated using DNAsp (Librado and Rozas, 2009); X-axis = position in base pair. A, C, D, F, and J, hypothetical proteins; B, Leishmanolysin family protein; E, H, K, and L, No significant BLAST hit; G, von Willebrand factor type A domain containing protein; I, Protein kinase domain containing protein.

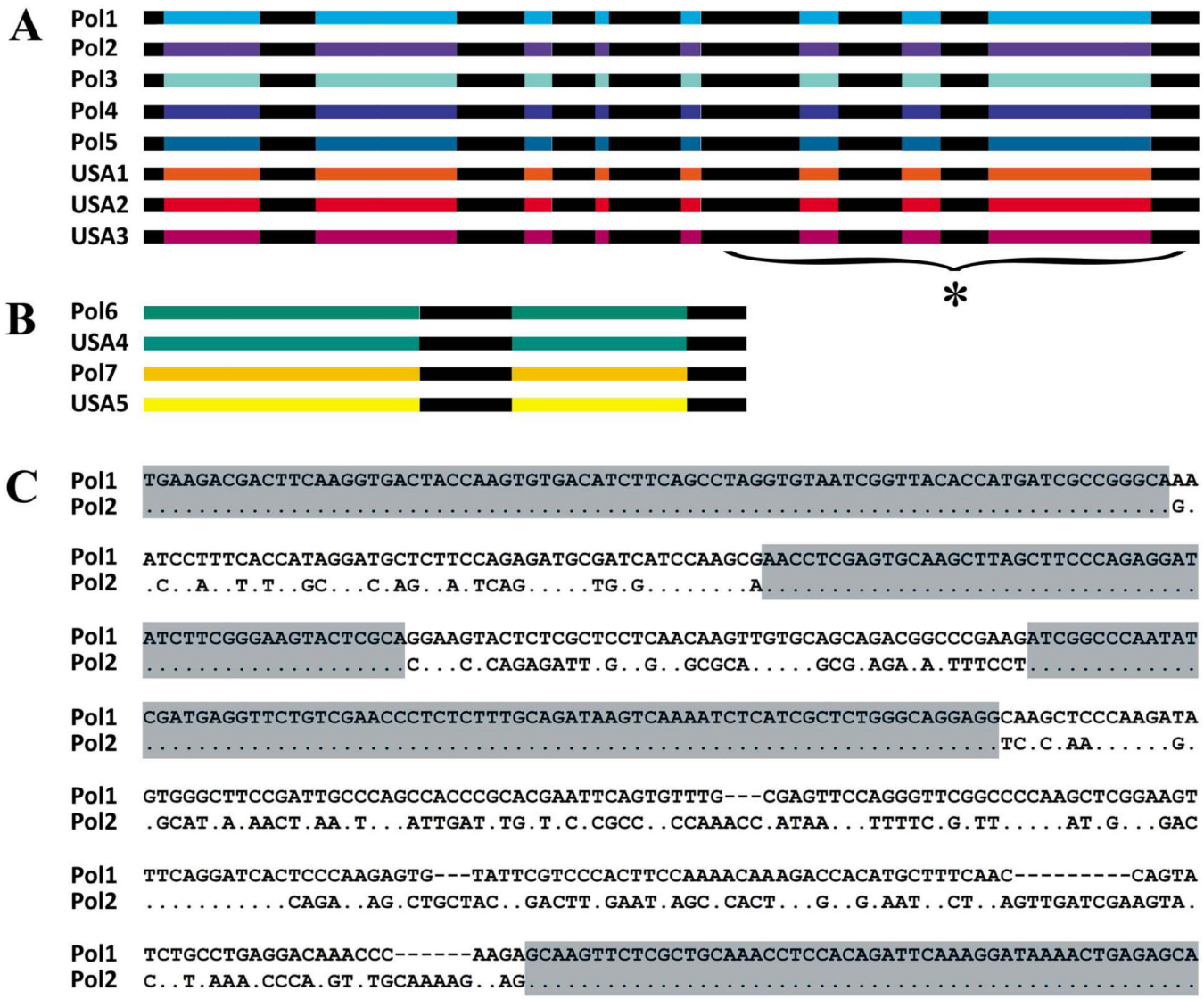

Figure 2.

Two patterns of PKc haplotypes from macronucleus that share identical regions (A, B) and sequence alignment of haplotypes Pol1 and Pol2 of the region indicated by the bracket and asterisk in A (C). Regions in black at identical positions correspond to shared sequences, and colors correspond to divergent regions. Matched sites are represented by dots (.). The identical regions in C are highlighted.

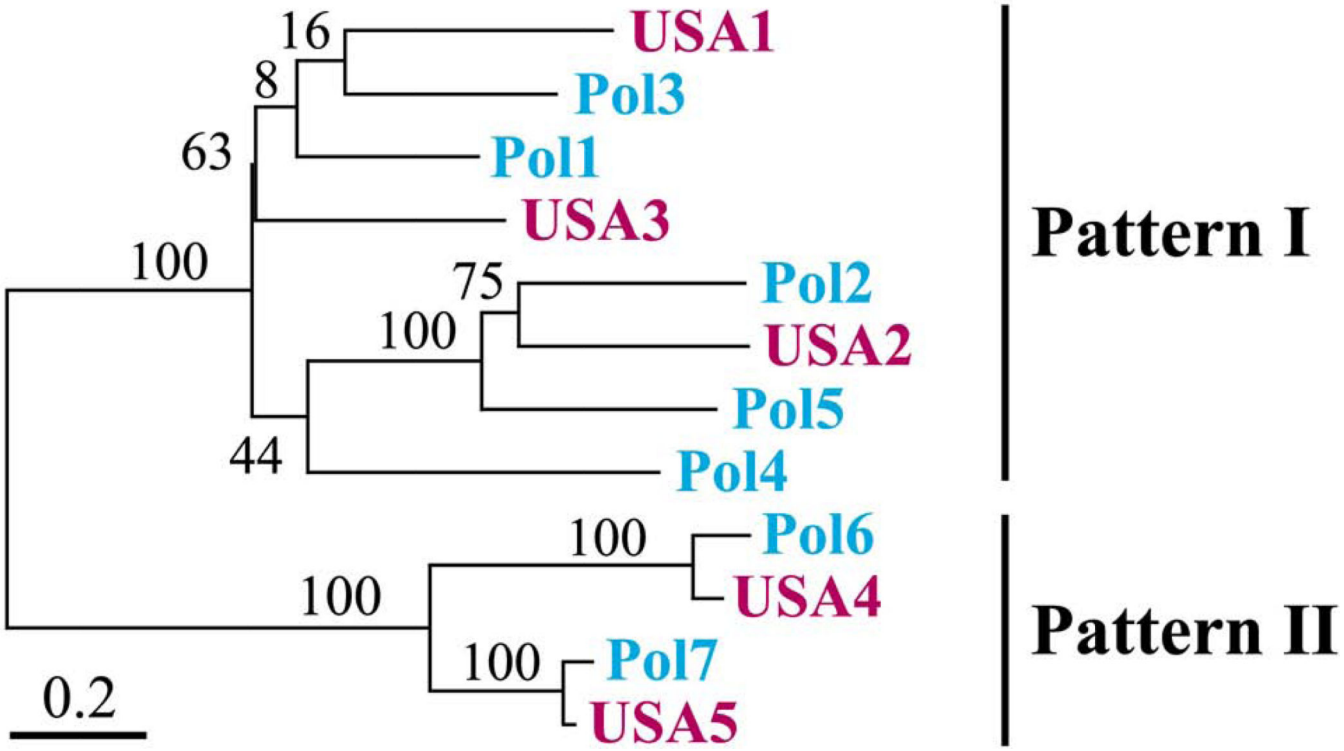

Figure 3.

Phylogenetic relationships of unique regions in the PKc gene family members based on nucleotide sequences. Pattern I shows a complex history of duplication, divergence and possible loss while pattern II is consistent with a single duplication even prior to the separation of the Pol and USA strains, followed by divergence of DNA sequence regions. Numbers at nodes represent the bootstrap values of ML out of 1,000 replicates. The scale bar corresponds to 20 substitutions per 100 positions.

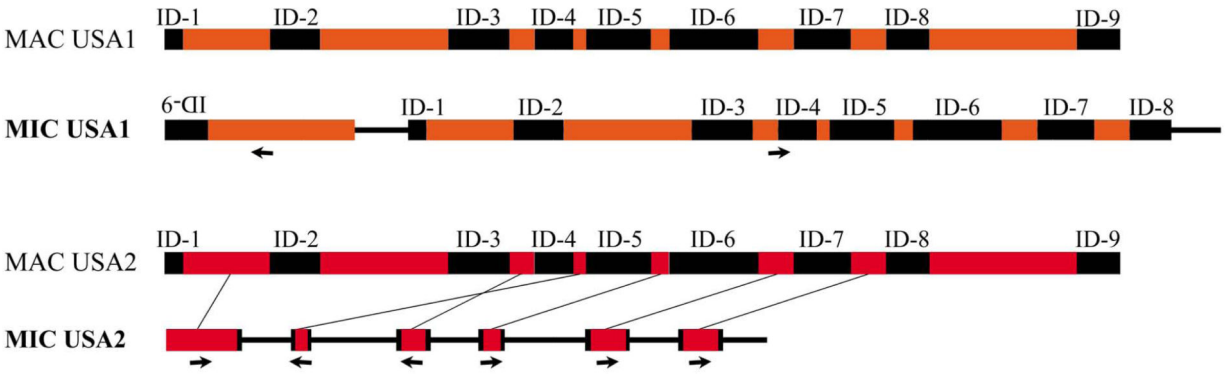

Figure 4.

Schematic maps of the macronuclear (MAC) and corresponding micronuclear (MIC) sequences of haplotypes USA1 and USA2. Black regions (ID 1–9) are identical in sequences across macronuclear (MAC) gene family members, and colors correspond to divergent regions. Arrows indicate the directions of the macronucleus destined sequences compared to the macronuclear sequences.

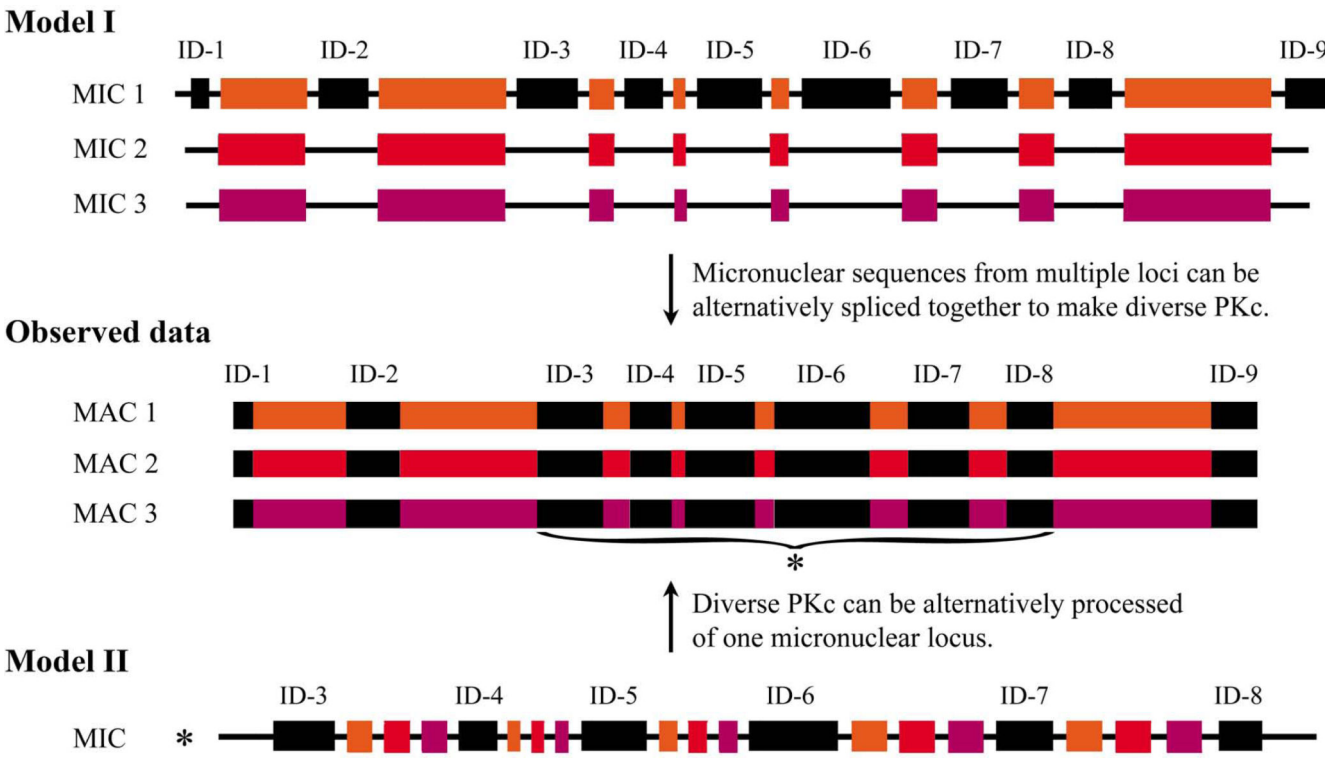

Figure 5.

Two models showing possible arrangements for PKc in the germline micronucleus. PKc may be alternatively processed from multiple micronuclear loci (Model I) or from a single micronuclear locus (Model II). Black regions (ID 1–9) are identical in sequences across macronuclear (MAC) gene family members, and colors correspond to divergent regions. In micronuclear (MIC) version, boxes indicate macronuclear destined sequences and black lines are internal excised sequences.

Table 1

Candidates of alternatively processed gene families in the transcriptome of the Pol strain of *C. uncinata*.

	# clusters	# seqs	>2 seqs	# Introns	# IESs
Strong signature	48	325	36	1	0
Weak signature	64	209	22	22	7
No signature	30	87	11	6	1
Others	306	667	35	188	12
In sum	448	1288	104	217	20

Clusters generated based on similarity contained 1288 sequences (contigs + single reads). Strong signature of alternative processing is defined as sharing more than one ≥25bp region. Weak signature is defined as sharing only one region. No signature clusters are putative gene families that show no sign of alternative processing. 'Others' contain sequences with either unprocessed micronuclear-limited regions, canonical eukaryotic introns, or sequences differing by one or few nucleotides (i.e. alleles or the result of sequence error). >2 seqs = the number of clusters that have more than two sequences; # introns = the number of unprocessed introns that are contained in the sequences; # IESs = the number of unprocessed internal excised sequences (IES) that are contained in the sequences. Transcriptome data are from Grant et al. (2012).

Table 2

Best BLASTX hit of gene families that have strong signatures of alternative processing from *C. uncinata* Pol strain transcriptome.

# clusters	# seqs	Best BLASTX hit	Taxon	E-value
34	209	No significant BLAST hit		>1.0e-15
5	78	Hypothetical proteins	*Paramecium/Tetrahymena*	<7.0e-18
1	10	Protein kinase domain containing protein	*Tetrahymena*	1.0e-67
1	7	von Willebrand factor type A domain containing protein	*Tetrahymena*	2.0e-63
1	4	Copine family protein	*Tetrahymena*	3.0e-58
1	3	Histidine acid phosphatase family protein	*Tetrahymena*	7.0e-25
1	3	Leishmanolysin family protein	*Tetrahymena*	4.0e-40
1	3	ab-hydrolase associated lipase region family protein	*Tetrahymena*	4.0e-47
1	3	Danj domain containing protein	*Tetrahymena*	8.0e-43
1	3	Cysteine proteinase	Plant	2.0e-34
1	2	Glutathione S-transferase, N-terminal domain containing protein	*Tetrahymena*	8.0e-21

Table 3

PKc sequence analyses from two *C. uncinata* strains of Pol (transcriptome + macronuclear genome) and USA (macronuclear genome).

Pattern	PKc haplotype	Pol transcriptome		Pol MAC		USA MAC	
		Reads	Length	Clones	Length	Clones	Length
I	Pol1	7	1251 bp	3	1358 bp		
I	Pol2	10	1214 bp	5	1376 bp		
I	Pol3	7	973 bp	2	1358 bp		
I	Pol4	2	471 bp	11	1373 bp		
I	Pol5	1	359 bp	14	1379 bp		
I	USA1					46	1376 bp
I	USA2					38	1358 bp
I	USA3					59	1367 bp
II	Pol6/USA4	31	2171 bp	11	797 bp	5	797 bp
II	Pol7/USA5			4	794 bp	11	794 bp
I	Pol8	14	966 bp	--	--		
II	Pol9	4	1334 bp	--	--		

MAC = macronuclear genome, reads = the number of reads of each haplotype in the transcriptome (Grant et al. 2012), clones = the number of clones from PCR reactions, which is a very rough proxy of copy number in macronuclear genome (Bellec and Katz 2012).