Nucleic Acids Research, 1995, Vol. 23, No. 8 1279–1283

Macronuclear gene-sized molecules of hypotrichs

David C. Hoffman, Richard C. Anderson, Michelle L. DuBois and David M. Prescott*

University of Colorado, Department of Molecular, Cellular and Developmental Biology, Boulder, CO 80309-0347, USA

Received December 12, 1994; Revised and Accepted March 1, 1995

ABSTRACT

The macronuclear genome of hypotrichous ciliates consists of DNA molecules of gene-sized length. A macronuclear DNA molecule contains a single coding region. We have analyzed the many hypotrich macronuclear DNA sequences sequenced by us and others. No highly conserved promoter sequences nor replication initiation sequences have been identified in the 5′ nor in the 3′ non-translated regions, suggesting that promoter function in hypotrichs may differ from other eukaryotes. The macronuclear genes are intron-poor; ~19% of the genes sequenced to date have one to three introns. Not all macronuclear DNA molecules may be transcribed; some macronuclear molecules may not have any coding function. Codon bias in hypotrichs is different in many respects from other ciliates and from other eukaryotes.

The genomic DNA of the macronucleus of hypotrichous ciliates is organized very differently from the DNA in the hypotrich micronuclear genome, from the genomes of other ciliates, and from the genomes of other eukaryotes. The size of the DNA molecules in the macronucleus of hypotrichs is very small, ranging from a few hundred base pairs (bp) to ~15 kilobase pairs (kb) (1). The average size of the molecules differs from species to species: ~2200 bp in *Oxytricha nova*, ~2500 bp in *Stylonychia pustulata* and ~1800 bp in *Euplotes aediculatus* (1). Each different molecule is present in many copies per macronucleus; in *O.nova* the average copy number is ~1000 (2). An exception is the DNA molecule encoding 26S, 19S, and 4.5S rRNA (7400 bp long in *O.nova*); it is differentially amplified to ~10^5 copies/macronucleus (3). The number of DNA molecules of different sequence has been estimated as ~24 000 in *O.nova* (1). This is calculated by dividing the value of 1/2 C_0t (total sequence complexity of the macronuclear DNA) by the average size of the macronuclear DNA molecules (2). This number could be in error by as much as 50% because of the uncertainty in determining 1/2 C_0t. If each kind of molecule encoded a different gene, the total number of genes in the macronucleus would be somewhere between ~12 000 and ~36 000, although there is no evidence that every different molecule represents a different gene. Even the lower estimate of ~12 000 seems to be a high number of genes for a unicellular organism.

Since Martindale (4) examined codon usage in ciliates, the number of fully or partially sequenced hypotrich genes has increased from five to 83. This increase in information has supported a number of conclusions about hypotrich macronuclear gene structure and organization. Eighty-three DNA molecules selected from macronuclear genomic DNA libraries, cDNA libraries or obtained by PCR from the genomic DNA of several hypotrich species have been characterized by dideoxynucleotide sequencing and/or restriction nuclease mapping (Tables 1 and 2). With few exceptions the cloned macronuclear DNA molecules were selected with heterologous probes of known coding function. In every case examined so far, each macronuclear DNA molecule contains only a single gene, defined as a single open reading frame (ORF) or a single transcription unit. Each molecule possesses nontranslated regions that flank the ORF: a 5′ DNA leader and a 3′ DNA trailer. Because of the presence of a single ORF and the small size of the DNA, macronuclear DNA molecules are referred to as gene-sized molecules, but some of the different molecules may not encode genes (discussed below). Macronuclear DNA molecules are also sometimes described as minichromosomes. The ends of each gene-sized molecule are capped by repeats of the telomere sequence 3′-dG_4T_4-5′ with 16-base (*Oxytricha, Stylonychia, Onychodromus* and *Holosticha* species) (63,64) or 14-base (*Euplotes* species) (63), single-stranded tails. Figure 1 illustrates the generalized structure of *Oxytricha* and *Stylonychia* macronuclear DNA molecules.

Table 1 contains a list of protein-encoding macronuclear DNA molecules that have been partially or fully sequenced. Table 2 lists macronuclear DNA molecules encoding RNA products (rRNAs and the RNA molecules of telomerase).

ORFs begin with the methionine codon ATG and terminate with the stop codon TGA in *Oxytricha* and *Stylonychia* species and TAA (or rarely TAG) in *Euplotes* species. TAA and TAG encode glutamine in *Oxytricha* and *Stylonychia* species (43,48). TGA encodes cysteine in *Euplotes* species (65). Overall codon

Figure 1. The generalized structure of *Oxytricha* and *Stylonychia* macronuclear DNA molecules (3).

* To whom correspondence should be addressed

Table 1. Characteristics of 66 putative protein-encoding macromolecular gene-sized molecules in hypotrichs

Gene	Organism	Gene Length	5' LDR	ORF	3' TLR	Accession #	Ref.
actin	*E. crassus*	1247 bp	54 bp	1140 bp	53 bp	J04533	6
actin	*O. fallax*	1497 bp	183 bp	1128 bp	186 bp	J01163	7
actin I	*O. nova*	1532 bp	191 bp	1128 bp	213 bp	M22480	8
actin I	*O. trifallax (H)*	1504 bp	184 bp	1128 bp	192 bp	n.a.	9
actin I	*O. trifallax (WR)*	1502 bp	184 bp	1128 bp	190 bp	U18940	9
actin II[a]	*O. nova*	1353 bp	100 bp	1128 bp	125 bp	U06071	10
AS1[a]	*O. nova*	425 bp	87 bp	129 bp	209 bp	M57402	11
AS2	*O. nova*	443 bp	104 bp	129 bp	210 bp	M57403	11
calmodulin	*S. lemnae*	773 bp	130 bp	450 bp	193 bp	M76407	12
C2	*O. nova*	744 bp	82 bp	249 bp	413 bp	K02624	13
DNA pol α[a]	*O. nova*	4938 bp	329 bp	4479 bp	130 bp	U02001	14
DNA pol α[a]	*O. trifallax (H)*	4952 bp	254 bp	4536 bp	162 bp	n.a.	15
DNA pol II A	*S. lemnae*	>385 bp	n.d.	>385 bp	n.d.	Z11764	16
DNA pol II E	*S. lemnae*	>385 bp	n.d.	>385 bp	n.d.	Z11836	16
EF-1α	*S. lemnae*	1790 bp	183 bp	1341 bp	266 bp	X57926	17
81-MAC(IIIA)[e]	*O. fallax*	1570 bp	88 bp	1116 bp	149 bp	M15836	18
Er-2	*E. raikovi*	>280 bp	>21 bp	228 bp	>31 bp	X61174	19
Er-10	*E. raikovi*	>281 bp	>21 bp	228 bp	>32 bp	X61173	19
Er-11	*E. raikovi*	>281 bp	>21 bp	228 bp	>32 bp	X60453	20
HSP70	*O. nova*	2586 bp	394 bp	1821 bp	371 bp	n.a.	21
histone H1	*E. eurystomus*	1254 bp	636 bp	408 bp	210 bp	L15293	22
histone H4	*O. nova*	1619 bp	1153 bp	315 bp	151 bp	M24411	23
histone H4G	*S. lemnae*	>736 bp	196 bp	315 bp	>225 bp	X16019	24
histone H4K	*S. lemnae*	1633 bp	1187 bp	315 bp	131 bp	X16018	24
Ma52[a]	*S. lemnae*	3735 bp	2248 bp	1335 bp	152 bp	X73879	25
Ma68[a]	*S. lemnae*	1836 bp	151 bp	1524 bp	161 bp	X73880	26
memER-1	*E. raikovi*	>616 bp	>182 bp	393 bp	>41 bp	M86864	27
1.7 kb gene	*O. trifallax (H)*	1652 bp	n.d.	n.d.	n.d.	n.a.	28
ORF1[a]	*S. lemnae*	1176 bp	498 bp	582 bp	96 bp	M75100	29
PGK	*E. crassus*	1372 bp	76 bp	1257 bp	39 bp	n.a.	30
pheromone 3[b,d]	*E. octocarinatus*	>435 bp	n.d.	300 bp	n.d.	M69117	31
pheromone 4[b,e]	*E. octocarinatus*	1622 bp	134 bp	384 bp	123 bp	X58838	32
phospholipase C	*E. crassus*	524 bp	33 bp	426 bp	65 bp	M63336	33
POB4[a]	*S. lemnae*	1177 bp	454 bp	615 bp	108 bp	X16613	34
polyubiquitin	*E. eurystomus*	898 bp	140 bp	690 bp	68 bp	M57231	35
polyubiquitin[a]	*O. trifallax (H)*	1075 bp	179 bp	690 bp	206 bp	n.a.	36
R1	*O. nova*	989 bp	97 bp	201 bp	691 bp	M18657	37
RNA pol A1[b]	*E. octocarinatus*	>1980 bp	66 bp	>1914 bp	n.d.	X66450	38
RNA pol A2[b]	*E. octocarinatus*	3879 bp	352 bp	3501 bp	26 bp	X66451	38
RNA pol B1[b,d]	*E. octocarinatus*	>1771 bp	42 bp	>1434 bp	n.d.	X66452	38
RNA pol B2[b,c]	*E. octocarinatus*	3715 bp	62 bp	3584 bp	38 bp	X66453	38
RNA pol C1[b,c]	*E. octocarinatus*	>1784 bp	44 bp	>1533 bp	n.d.	X67660	39
RPL29	*E. crassus*	553 bp	36 bp	444 bp	73 bp	U13207	40
sER-1	*E. raikovi*	>593 bp	>258 bp	228 bp	>107 bp	J04141	41
719 (CaBP)[a]	*S. lemnae*	676 bp	66 bp	210 bp	400 bp	M90073	12
TBP[c]	*E. crassus*	1546 bp	40 bp	1341 bp	64 bp	M96818	42
TBP homolog[d]	*E. crassus*	1545 bp	54 bp	1383 bp	35 bp	M96819	42
α-TBP[a,c]	*O. nova*	2145 bp	159 bp	1488 bp	446 bp	M68930	43
α-TBP[a,c]	*S. mytilus*	2085 bp	155 bp	1482 bp	401 bp	X61749	44
β-TBP (A version)[c]	*O. nova*	1716 bp	162 bp	1158 bp	285 bp	M31310	45
β-TBP (S version)[c]	*O. nova*	1718 bp	162 bp	1158 bp	287 bp	M31309	45
β-TBP[c]	*O. trifallax (H)*	1802 bp	188 bp	1164 bp	364 bp	n.a.	9
β-TBP[a,c]	*S. mytilus*	1682 bp	177 bp	1179 bp	277 bp	X61748	44
3.3 kb gene	*O. nova*	3384 bp	n.d.	n.d.	n.d.	M81801	28
α-tubulin	*E. octocarinatus*	1531 bp	78 bp	1353 bp	100 bp	X69466	46
α-tubulin	*E. vannus*	1505 bp	86 bp	1350 bp	69 bp	Z11769	47
α-tubulin[a]	*O. granulifera*	1627 bp	173 bp	1353 bp	101 bp	Z11763	47
α-1-tubulin[a]	*S. lemnae*	1773 bp	192 bp	1338 bp	243 bp	X01746	48
α-2-tubulin[a]	*S. lemnae*	1691 bp	74 bp	1350 bp	267 bp	X12365	49
β-tubulin	*E. crassus*	1468 bp	49 bp	1341 bp	78 bp	J04534	6
β-tubulin	*E. octocarinatus*	1468 bp	63 bp	1335 bp	70 bp	X69467	46
β-1-tubulin	*S. lemnae*	1798 bp	126 bp	1329 bp	343 bp	X06653	50
β-2-tubulin[a]	*S. lemnae*	1783 bp	170 bp	1329 bp	284 bp	X06874	50
γ-tubulin[b,d]	*E. octocarinatus*	1580 bp	49 bp	1389 bp	80 bp	X71353	51
V2	*E. crassus*	814 bp	395 bp	216 bp	203 bp	M28500	52
V3[c]	*E. crassus*	1751 bp	36 bp	1419 bp	40 bp	n.a.	53

Sixty-six macronuclear gene-sized molecules of the hypotrichs *Euplotes crassus*, *E.eurystomus*, *E.octocarinatus*, *E.raikovi*, *E.vannus*, *Oxytricha fallax*, *O.granulifera*, *O.nova*, *O.trifallax* (WR), *O.trifallax* (H), *Stylonychia lemnae* and *S.mytilus* that have been sequenced or mapped with restriction endonucleases. Lengths of gene-sized molecules exclude telomeres. Open reading frame lengths include the translational stop codon but exclude introns. AS1, AS2, C2, 81MAC(IIIA), Ma52, Ma68, ORF1, POB4, R1, 719, V2, and V3 all have open reading frames of ≥129 bp encoding putative proteins of unknown function, although some (81MAC(IIIA), Ma52, Ma68 and 719) do display partial sequence identity with proteins of known function. The 1.7 and 3.3 kb molecules contain no convincing open reading frames and transcripts have not been detected. Sequence information for pheromone 3 from *E.octocarinatus*, and for memER-1 and sER-1 from *E.raikovi* was obtained from cDNA clones. GenBank accession numbers are included where applicable (5).

AS, amplified sequence; CaBP, calcium binding protein; Er, *Euplotes raikovi* mating pheromone; EF-1α, elongation factor-1 α; HSP70, heat-shock protein 70; LDR, DNA leader, or 5′ non-translated region; memEr, membrane bound isoform of *E.raikovi* mating pheromone; n.a., not available; n.d., not determined; ORF, open reading frame; PGK, phosphoglycerate kinase; pol, polymerase; RP, ribosomal protein; sER, soluble isoform of *E.raikovi* mating pheromone; TBP, telomere binding protein; TLR, DNA trailer, or 3′ non-translated region.

[a]Open reading frame containing TAA or TAG codons that specify Gln.

[b]Open reading frame containing TGA codons that specify Cys.

[c]Gene containing a single intron.

[d]Gene containing two introns.

[e]Gene containing three introns.

usage in hypotrichs is different from both holotrichous ciliates and other eukaryotes (see Table 3).

The nucleotide composition of the protein-encoding ORFs differs significantly from the composition of the 5′ and 3′ nontranslated regions. ORFs average ~54% AT, 5′ DNA leaders average ~76% AT, and 3′ DNA trailers of hypotrich genes average ~70% AT.

The 5′ DNA leaders presumably contain sequences that signal binding of RNA polymerase and initiation of transcription. Computer searches for the eukaryotic TATA-, CAAT-, and GC-box consensus sequences [5′-TATA($^{A}/_{T}$)A($^{A}/_{T}$)-3′, 5′-GG($^{T}/_{C}$)CAA-TCT-3′, and 5′-GGGCGG-3′ respectively] identified TATA-like sequences (66) in 48 of 66 genes and CAAT-like sequences (66) in 19 of 66 genes in Table 1; no GC-like sequences (67) were found. The TATA-like sequences are expected to occur randomly with higher frequency because leaders average ~76% AT; conversely CAAT- and GC-like sequences will occur by chance less frequently. Transcription start sites have been mapped for six

Table 2. Characteristics of 17 RNA-encoding macronuclear gene-sized molecules in hypotrichs

Gene	Organism	Gene Length	5' LDR	ORF	3' TLR	Accession #	Ref.
rRNA (5S)	*E. eurystomus*	846 bp	194 bp	120 bp	532 bp	X13718	54
rRNA (5S)	*E. woodruffi*	>120 bp	n.d.	120 bp	n.d.	K02347	55
rRNA (16S-like)	*E. aediculatus*	>1882 bp	n.d.	1882 bp	n.d.	X03949	56
rRNA (16S-like)	*O. quadricornatus*	>1771 bp	n.d.	1771 bp	n.d.	X53485	57
rRNA (16S-like)	*O. granulifera*	>1778 bp	n.d.	1778 bp	n.d.	X53486	57
rRNA (16S-like)	*O. nova*	>1771 bp	n.d.	1771 bp	n.d.	M14601	58
rRNA (16S-like)	*S. pustulata*	>1771 bp	n.d.	1771 bp	n.d.	M14600	58
rRNA (5.8S, 19S, & 25S)	*O. fallax*	~7400 bp	~1500 bp	~5500 bp	≤400 bp	n.a.	59
rRNA (5.8S, 19S, & 25S)	*O. nova*	~7400 bp	~1500 bp	~5500 bp	≤400 bp	n.a.	60, 61
rRNA (5.8S, 19S, & 25S)	*S. pustulata*	~8140 bp	~1540 bp	~6000 bp	~600 bp	n.a.	61
telomerase RNA	*E. aediculatus*	>430 bp	>49 bp	189 bp	192 bp	n.a.	61
telomerase RNA	*E. crassus*	609 bp	127 bp	191 bp	291 bp	M33461	62
telomerase RNA	*E. eurystomus*	>781 bp	>592 bp	189 bp	n.d.	n.a.	61
telomerase RNA	*O. nova*	>373 bp	n.d.	190 bp	183 bp	n.a.	61
telomerase RNA	*O. trifallax (H)*	>368 bp	>26 bp	186 bp	156 bp	n.a.	61
telomerase RNA	*S. lemnae*	>396 bp	>27 bp	189 bp	180 bp	n.a.	61
telomerase RNA	*S. mytilus*	669 bp	315 bp	189 bp	165 bp	n.a.	61

Seventeen macronuclear genes encoding RNA products of the hypotrichs *Euplotes aediculatus*, *E.crassus*, *E.eurystomus*, *E.woodruffi*, *Onychodromus quadricornatus*, *O.fallax*, *O.granulifera*, *O.nova*, *O.trifallax* (H), *S.lemnae*, *S.mytilus*, and *S.pustulata* (5). All clones except for the *E.eurystomus* 5S rRNA, *E.crassus* telomerase RNA, *O.fallax*, *O.nova*, and *S.pustulata* rRNA (5.8S, 19S and 25S molecules) were produced by PCR. The complete structures of the macronuclear DNA molecules encoding the 16S-like rRNA genes are not known because they are PCR clones.
n.d., not determined.

E.crassus genes and a single *O.nova* gene (8,68). In every case TATA sequences are absent upstream of the start site and CAAT- and GC-boxes are not present. These data suggest that either conventional eukaryotic promoter sequences are not required for transcription initiation or that the promoter sequence requirements for hypotrichs are less stringent than for other eukaryotes. A specific transcription regulatory site is present in at least one case; the gene encoding heat shock protein 70 in *O.nova* has two copies of the 14 bp eukaryotic consensus sequence that binds the heat shock factor (21). Also, the >1000 bp 5′ leaders in the histone H4 genes of *O.nova* and *S.lemnae* share extensive sequence commonalities that may represent sites for regulating transcription during the cell cycle (69). mRNA molecules in hypotrichs are polyadenylated, although the consensus eukaryotic poly(A) addition signal sequence of 5′-AATAAA-3′ (70) is present in only ~19% of the genes in Table 1 (see also 68).

Replication of macronuclear DNA molecules initiates at or very near one or both ends (71,72). Characterization of a DNA primase activity in *O.nova* that synthesizes 16-nt RNA primers templated by the 3′ telomeric overhang *in vitro* suggests that telomeres might serve as replication initiation sites (73). No other consensus sequence has been identified in *Oxytricha* or *Euplotes* species in the 5′ DNA leader or 3′ DNA trailer that might function as a specific initiation site. In *S.lemnae*, a moderately conserved palindrome [5′-($^{A}/_{T}$)ATTTAAAT($^{A}/_{T}$)-3′] has been identified in the region 40–70 bp from both the 5′ and 3′ telomere addition sites in 27 randomly selected macronuclear DNA molecules that may serve as an origin of replication (74). However, computer analysis of the genes in Table 1 only identified this sequence in both the leader and trailer of eight genes.

Three of the molecules in Table 1 were not selected with heterologous probes. One from *E.crassus* was studied because it possesses an unusually long telomere at one end (33). Sequencing showed that it possesses a single ORF, which encodes a protein with homology to a rat form-I phosphoinositide-specific phospholipase C. Two other non-selected molecules were a 3384-bp molecule from an *O.nova* macronuclear library and a 1652-bp molecule from a library of macronuclear DNA of *O.trifallax (H)* (28). Neither of these non-selected molecules contains a convincing ORF. The longest putative ORF in the 3384-bp molecule that is defined by an ATG and a TGA codon (TAA and TAG are not used as stop codons in *Oxytricha*—see Table 3) is 2029 bp, interrupted by two introns of 35 and 407 bp and encoding a polypeptide of 528 amino acids; other possible ORFs are much shorter. The longest putative ORF in the 1652-bp molecule is 1402 bp, interrupted by a single 535-bp intron, encoding a polypeptide of 288 amino acids. Neither of these two nonselected molecules contains a sequence with recognizable similarity to a known protein- nor RNA-encoding gene in other eukaryotes. Preliminary attempts to find transcripts of the two molecules by Northern hybridization have been inconclusive, and the putative ORFs in the two molecules do not conform to the ciliate codon bias (see Table 3). These observations imply that some, perhaps many, macronuclear DNA molecules possess no coding function. This hypothesis can be tested by analyzing a randomly chosen set of macronuclear DNA molecules.

The shortest macronuclear DNA molecule sequenced for any hypotrich so far is AS1, which is 425 bp and encodes a putative polypeptide of 43 amino acids in *O.nova* (11). The longest molecule sequenced is 4952 bp and encodes the large, catalytic subunit of DNA polymerase α in *O.trifallax (H)* (15). The coding functions, if any, of very long macronuclear DNA molecules (≤15 kb) are unknown.

Hypotrich macronuclear genes are intron-poor although the micronuclear (germline) versions of genes contain many interrupting sequences called internal eliminated sequences, or IESs, that are spliced out of the DNA during development of the macronuclear genome (somatic genome) from a micronuclear genome (3,13). Of the 83 macronuclear gene-sized molecules in Tables 1 and 2, 16 contain one to three introns. The 23 introns range in length from 24 to 772 bp, with an average length of ~118 bp. Hypotrich introns average ~74% AT, compared to ~54% AT for protein coding ORFs. All 23 introns begin with the highly conserved dinucleotide GT and end with the highly conserved dinucleotide AG; both the 5′ and 3′ splice sites are in good agreement with the eukaryotic consensus sequences of 5′-($^{A}/_{C}$)AG/GT($^{A}/_{G}$)AGT-3′ and 5′-($^{T}/_{C}$)AG/G-3′ respectively, where the slash indicates the boundary between exon and intron (75).

The frequencies with which the different codons for an amino acid or translational stop are used in hypotrichs are compared with their percentage usage in *Tetrahymena*, *Paramecium*, *Saccharomyces*, *Drosophila* and human in Table 3. Clearly, codon usage in hypotrich genes is strongly biased and in many respects is different from the codon bias in the other eukaryotes listed. For example, the arginine codon AGA is heavily used in hypotrichs while the arginine codons CGA and CGG are rarely used (Table 3). All three of these codons occur with different frequencies in the other eukaryotes. Ciliates use non-standard codons in some

Table 3. Codon frequencies

Amino Acid	Codon	Euplotes	Oxytricha	Stylonychia	Tetrahymena	Paramecium	S. cerevisiae	D. melanogaster	Human
Ala	GCT	0.40	0.42	0.45	0.67	0.50	0.44	0.20	0.28
	GCC	0.25	0.39	0.45	0.29	0.11	0.25	0.43	0.40
	GCA	0.32	0.18	0.10	0.04	0.38	0.23	0.18	0.22
	GCG	0.02	0.01	0.01	<.005	0.01	0.08	0.19	0.10
Arg	CGT	0.03	0.03	0.02	0.04	0.02	0.17	0.16	0.09
	CGC	0.01	0.05	0.03	0.01	0.01	0.04	0.29	0.19
	CGA	0.02	0.00	0.04	0.00	0.02	0.05	0.15	0.10
	CGG	0.01	0.00	0.02	0.00	<.005	0.02	0.15	0.19
	AGA	0.86	0.87	0.82	0.94	0.93	0.54	0.12	0.21
	AGG	0.07	0.05	0.06	0.02	0.01	0.17	0.14	0.22
Asn	AAT	0.51	0.36	0.24	0.40	0.82	0.54	0.46	0.44
	AAC	0.49	0.64	0.76	0.60	0.18	0.46	0.55	0.56
Asp	GAT	0.65	0.53	0.58	0.66	0.86	0.62	0.53	0.44
	GAC	0.35	0.47	0.42	0.34	0.14	0.38	0.47	0.56
Cys	TGT	0.42	0.20	0.22	0.43	0.68	0.68	0.31	0.42
	TGC	0.38	0.80	0.78	0.57	0.32	0.32	0.69	0.58
	TGA	0.21	stop	stop	stop	stop	stop	stop	stop
Gln	CAA	0.74	0.48	0.68	0.48	0.19	0.74	0.32	0.27
	CAG	0.26	0.24	0.18	0.04	0.02	0.26	0.68	0.73
	TAA	stop	0.18	0.11	0.36	0.70	stop	stop	stop
	TAG	stop	0.11	0.03	0.12	0.08	stop	stop	stop
Glu	GAA	0.68	0.52	0.48	0.94	0.86	0.74	0.32	0.41
	GAG	0.32	0.48	0.52	0.06	0.14	0.26	0.68	0.59
Gly	GGT	0.21	0.41	0.57	0.81	0.29	0.61	0.22	0.18
	GGC	0.06	0.15	0.12	0.05	0.05	0.15	0.40	0.33
	GGA	0.71	0.42	0.28	0.12	0.64	0.15	0.30	0.26
	GGG	0.02	0.01	0.04	0.01	0.02	0.09	0.09	0.23
His	CAT	0.54	0.32	0.34	0.29	0.77	0.60	0.41	0.41
	CAC	0.46	0.68	0.66	0.72	0.23	0.40	0.59	0.59
Ile	ATT	0.54	0.37	0.36	0.44	0.59	0.50	0.33	0.35
	ATC	0.28	0.56	0.54	0.50	0.17	0.30	0.50	0.52
	ATA	0.18	0.07	0.10	0.06	0.24	0.20	0.16	0.14
Leu	TTA	0.24	0.10	0.09	0.20	0.46	0.27	0.07	0.06
	TTG	0.19	0.15	0.14	0.27	0.24	0.36	0.16	0.12
	CTT	0.24	0.22	0.24	0.23	0.18	0.11	0.11	0.12
	CTC	0.20	0.39	0.44	0.26	0.06	0.04	0.16	0.20
	CTA	0.09	0.08	0.06	0.03	0.05	0.13	0.09	0.07
	CTG	0.06	0.05	0.04	0.01	0.01	0.09	0.41	0.43
Lys	AAA	0.47	0.28	0.24	0.27	0.72	0.51	0.30	0.40
	AAG	0.53	0.72	0.76	0.73	0.28	0.49	0.70	0.60

Amino Acid	Codon	Euplotes	Oxytricha	Stylonychia	Tetrahymena	Paramecium	S. cerevisiae	D. melanogaster	Human
Phe	TTT	0.44	0.22	0.12	0.19	0.44	0.53	0.35	0.43
	TTC	0.56	0.78	0.88	0.81	0.56	0.47	0.65	0.57
Pro	CCT	0.30	0.23	0.10	0.40	0.29	0.29	0.15	0.29
	CCC	0.03	0.33	0.16	0.53	0.08	0.13	0.30	0.33
	CCA	0.65	0.43	0.73	0.06	0.63	0.49	0.26	0.27
	CCG	0.02	0.01	0.01	0.01	0.00	0.09	0.28	0.11
Ser	TCT	0.24	0.19	0.13	0.38	0.25	0.32	0.10	0.18
	TCC	0.12	0.25	0.15	0.34	0.05	0.18	0.23	0.23
	TCA	0.36	0.30	0.49	0.13	0.41	0.19	0.10	0.15
	TCG	0.04	0.03	0.04	0.01	0.01	0.08	0.20	0.06
	AGT	0.13	0.08	0.08	0.08	0.20	0.14	0.13	0.14
	AGC	0.11	0.15	0.11	0.07	0.07	0.09	0.24	0.25
Thr	ACT	0.49	0.38	0.36	0.45	0.44	0.38	0.17	0.23
	ACC	0.22	0.49	0.52	0.47	0.08	0.25	0.36	0.38
	ACA	0.29	0.13	0.10	0.07	0.48	0.26	0.21	0.27
	ACG	0.01	<.005	0.02	0.01	0.01	0.12	0.26	0.12
Tyr	TAT	0.52	0.37	0.35	0.37	0.72	0.50	0.37	0.42
	TAC	0.48	0.63	0.65	0.63	0.29	0.50	0.63	0.58
Val	GTT	0.40	0.36	0.36	0.38	0.56	0.44	0.19	0.17
	GTC	0.29	0.38	0.51	0.53	0.13	0.24	0.26	0.25
	GTA	0.23	0.13	0.07	0.07	0.15	0.16	0.10	0.10
	GTG	0.08	0.13	0.06	0.03	0.05	0.15	0.44	0.48
Stop	TGA	Cys	1.00	1.00	1.00	1.00	0.34	0.53	0.61
	TAA	0.91	Gln	Gln	Gln	Gln	0.49	0.30	0.22
	TAG	0.09	Gln	Gln	Gln	Gln	0.17	0.17	0.17

Codon frequencies (expressed in decimals) for *Euplotes, Oxytricha, Paramecium, Stylonychia* and *Tetrahymena* macronuclear genes were tabulated from the open reading frames (ORFs) specified in the GenBank entries using MacVector version 4.0 (IBI/Kodak) (5). Macronuclear clones of unknown function were included if they contained an identifiable ORF of ≥100 bp, as defined by an ATG and an appropriate translation termination codon. Codon frequencies for *Paramecium* and *Tetrahymena* were tabulated as described above using a representative sample of macronuclear gene sequences deposited in GenBank. Non-hypotrich sequences examined and their GenBank accession numbers include: *P.primaurelia*—168G surface protein (X52133), G surface protein (X03882); *P.tetraurelia*—immobilization antigen 51A (M65163), immobilization antigen 51B (L04795), immobilization antigen 51C (M65164), β-tubulin (X67237), calmodulin (M34540), protein phosphatase I (X67492); *T.pyriformis*—actin (X05195), α-tubulin (X12767), β-1-tubulin (X12768), β-2-tubulin (X12769), calmodulin (D10521), EF-1α (D11083), phosphoenolpyruvate mutase (M85236), polyubiquitin (X61053), polyubiquitin–ribosomal protein fusion (X56693); *T.thermophila*—actin (M13939), α-tubulin (M86723), β-1-tubulin (L01415), β-2-tubulin (L01416), calmodulin (X52242), citrate synthase (D90117), cnjB (L03710), cnjC (X62317), cysteine protease (L03212), histone H1 (M14854), histone H2A.1 (L18892), histone H2B.1 (M31332), phosphoglycerate kinase (X63528), ribosomal protein L29 (M76719), ribosomal protein L37 (M59428), 23 kDa calcium binding protein (J05227), 25 kDa calcium binding protein (J05109). Codon frequencies for *Drosophila melanogaster, Saccharomyces cerevisiae* and human were tabulated with the program GCG, using 412, 1952 and 435 genes respectively, from GenBank release 63. These tables were retrieved via anonymous ftp from nic.funet.fi, in the directory /pub/sci/molbio/databases/codon.

cases as well. In *Euplotes octocarinatus*, the codon TGA specifies cysteine rather than translation termination (65). Of the 10 *E.octocarinatus* genes from which the data in Table 3 was compiled, eight contain TGA codons; 46 of 129 cysteine codons in those eight genes (~36%) are non-standard. In *Oxytricha* and *Stylonychia*, the codons TAA and TAG specify glutamine rather than translation termination (43,48). Of the 17 *Oxytricha* and 14 *Stylonychia* genes from which the data in Table 3 was compiled, seven and 10 contain TAA or TAG codons respectively; 104 of 270 glutamine codons (~39%) in the seven *Oxytricha* genes and 33 of 193 glutamine codons (~17%) in the 10 *Stylonychia* genes are non-standard. The origin and significance of codon bias and non-standard codon usage are not known.

In summary, the hypotrich macronuclear genome has unique features that distinguish it sharply from other eukaryotes. Genes are spliced out of the micronuclear, germline chromosomes during macronuclear development to create short, telomere-bounded DNA molecules that encode single genes. Hypotrich genes generally lack TATA-, CAAT- and GC-boxes, and initiation of transcription is apparently signalled in a different way than most genes in other eukaryotes. Replication origins have not been identified for hypotrichs, although replication might initiate at telomeres, and a putative origin sequence has been reported recently in *Stylonychia lemnae* (74). These features of the hypotrich genome represent the evolution of ways of processing,

organizing, transcribing and replicating a genome that are quite different from those of other eukaryotes.

ACKNOWLEDGEMENTS

The authors wish to thank Larry Klobutcher and Dorothy Shippen for sharing unpublished data, and Gayle Prescott for assistance with manuscript preparation. This work was supported by NIGMS grant GM19199 to DMP.

REFERENCES

1 Swanton, M.T., Heumann, J.M. and Prescott, D.M. (1980) *Chromosoma*, **77**, 217–227.
2 Lauth, M.R., Spear, B.B., Heumann, J. and Prescott, D.M. (1976) *Cell*, **7**, 67–74.
3 Prescott, D.M. (1994) *Microbiol. Rev.*, **58**, 233–267.
4 Martindale, D. W. (1989) *J. Protozool.*, **36**, 29–34.
5 Benson, D. A., Boguski, M., Lipman, D. J. and Ostell, J. (1994) *Nucleic Acids Res.*, **22**, 3441–3444.
6 Harper, D.S. and Jahn, C.L. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 3252–3256.
7 Kaine, B. and Spear, B. (1982) *Nature* (London), **295**, 430–432.
8 Greslin, A.F., Loukin, S.H., Oka, Y. and Prescott, D.M. (1988) *DNA*, **7**, 529–536.
9 DuBois, M.L. and Prescott, D.M. (1995) *Proc. Natl. Acad. Sci. USA*, **92**, in press.
10 Mitcham, J.L. and Prescott, D.M. (1994) *Gene*, **144**, 119–122.
11 Harper, D. S., Song, K. and Jahn, C.L. (1991) *Gene*, **99**, 55–61.
12 Gaunitz, C., Witte, H. and Gaunitz, F. (1992) *Gene*, **119**, 191–198.
13 Klobutcher, L.A., Jahn, C.L. and Prescott, D.M. (1984) *Cell*, **36**, 1045–1055.
14 Mansour, S.J., Hoffman, D.C. and Prescott, D.M. (1994) *Gene*, **144**, 155–161.
15 Gray, E.A., Hoffman, D.C. and Prescott, D.M. (unpublished).
16 Plagge, A. and Gaunitz, F. (unpublished).
17 Bierbaum, P., Dönhoff, T. and Klein, A. (1991) *Mol. Microbiol.*, **5**, 1567–1575.
18 Herrick, G., Cartinhour, S. W., Williams, K. R., and Kotter, K. P. (1987) *J. Protozool.* **34**, 429–434.
19 Miceli, C., LaTerza, A., Bradshaw, R. and Luporini, P. (1991) *Eur. J. Biochem.*, **202**, 759–764.
20 La Terza, A., Miceli, C. and Luporini, P. (unpublished).
21 Anderson, R.C., Lindauer, K.R. and Prescott, D.M. (unpublished).
22 Hauser, L., Treat, M. and Olins, D. (1993) *Nucleic Acids Res.*, **21**, 3586–3591.
23 Harper, D.S. and Jahn, C.L. (1989) *Gene*, **75**, 93–107.
24 Wefes, I. and Lipps, H.J. (1990) *J. DNA Sequencing Mapping*, **1**, 25–32.
25 Maercker, C. and Lipps, H.J. (1994) *Gene*, **141**, 145–146.
26 Maercker, C. and Lipps, H.J. (1994) *Gene*, **141**, 147–148.
27 Miceli, C., LaTerza, A., Bradshaw, R. and Luporini, P. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 1988–1992.
28 Greslin, A.F. and Prescott, D.M. (unpublished).
29 Kreyenberg, H. (1993) Thesis Abteilung Zellbiologie. Universität Tübingen, Tübingen, Germany.
30 Pearlman, R.E., Hoppe, K. and Klobutcher, L.A. (unpublished).
31 Bruenen-Nieweler, C., Schmidt, H. J. and Heckmann, K. (1991) *Gene*, **109**, 233–237.
32 Meyer, F., Schmidt, H.J. and Heckmann, K. (1992) *Dev. Genet.*, **13**, 16–25.
33 Klobutcher, L.A., Turner, L.R. and Peralta, M.E. (1991) *J. Protozool.*, **38**, 425–427.
34 Wegner, M., Helftenbein, E., Müller, F., Meinecke, M., Müller, S. and Grummt, F. (1989) *Nucleic Acids Res.*, **17**, 8783–8802.
35 Hauser, L.J., Roberson, A.E. and Olins, D.E. (1991) *Chromosoma*, **100**, 386–394.
36 Dodge, J.K. and Prescott, D.M. (unpublished).
37 Ribas-Aparicio, R.M., Sparkowski, J.J., Proulx, A.E., Mitchell, J.D. and Klobutcher, L.A. (1987) *Genes Dev.*, **1**, 323–336.
38 Kaufmann, J. and Klein, A. (1992) *Nucleic Acids Res.*, **20**, 4445–4450.
39 Kaufmann, J., Florian, V. and Klein, A. (1992) *Nucleic Acids Res.*, **20**, 5985–5989.
40 Jahn, C.L., Erbeznik, M., Jaraczewski, J.W., Melek, M. and Shippen, D.E. (1994) *Gene*, **151**, 231–235.
41 Miceli, C., LaTerza, A. and Melli, M. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 3016–3020.
42 Wang, W., Skopp, R. Scofield, M. and Price, C. (1992) *Nucleic Acids Res.*, **20**, 6621–6629.
43 Gray, J.T., Celander, D.W., Price, C.M. and Cech, T.R. (1991) *Cell*, **67**, 807–814.
44 Fang, G. and Cech, T.R. (1991) *Nucleic Acids Res.*, **19**, 5515–5518.
45 Hicke, B.J., Celander, D.W., MacDonald, G.H., Price, C.M. and Cech, T.R. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 1481–1485.
46 Liang, A., Schmidt, H.J. and Heckmann, K. (1994) *J. Eukaryot. Microbiol.*, **41**, 163–169.
47 Gaunitz, F. (1990) Thesis Abteilung Zellbiologie. Universität Tübingen, Tübingen, Germany.
48 Helftenbein, E. (1985) *Nucleic Acids Res.*, **13**, 415–433.
49 Helftenbein, E. and Müller, E. (1988) *Curr. Genet.*, **13**, 425–432.
50 Conzelmann, K.K. and Helftenbein, E. (1987) *J. Mol. Biol.*, **198**, 643–653.
51 Liang, A. and Heckmann, K. (1993) *Gene*, **136**, 319–322.
52 Baird, S.E., Fino, G.M., Tausta, S.L. and Klobutcher, L.A. (1989) *Mol. Cell. Biol.*, **9**, 3793–3807.
53 Hale, C. and Klobutcher, L.A. (unpublished).
54 Roberson, A. E., Wolffe, A.P., Hauser, L.J. and Olins, D.E. (1989) *Nucleic Acids Res.*, **17**, 4699–4712.
55 Kumazaki, T., Hori, H. and Osawa, S. (1983) *J. Mol. Evol.*, **2**, 411–419.
56 Sogin, M.L., Swanton, M.T., Gunderson, J.H. and Elwood, H.J. (1986) *J. Protozool.*, **33**, 26–29.
57 Schlegel, M., Elwood, H. J. and Sogin, M. L. (1991) *J. Mol. Evol.*, **32**, 64–69.
58 Elwood, H.J., Olsen, G.J. and Sogin, M.L. (1985) *Mol. Biol. Evol.*, **2**, 399–410.
59 Spear, B.B. (1980) *Chromosoma*, **77**, 193–202.
60 Swanton, M.T., Greslin, A.F. and Prescott, D.M. (1980) *Chromosoma*, **77**, 203–215.
61 Lingner, J., Hendrick, L. L. and Cech, T. R. (1994) *Genes Dev.*, **8**, 1984–1998.
62 Shippen-Lentz, D. and Blackburn, E. H. (1990) *Science*, **247**, 546–552.
63 Klobutcher, L.A., Swanton, M.T., Donini, P. and Prescott, D.M. (1981) *Proc. Natl. Acad. Sci. USA*, **78**, 3015–3019.
64 Prescott, D. M. (unpublished).
65 Meyer, F., Schmidt, H.J., Plümper, E., Hasilik, A., Mersman, G., Meyer, H.E., Engström, A. and Heckmann, K. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 3758–3761.
66 McKnight, S. L. and Kingsbury, R. (1982) *Science*, **217**, 316–324.
67 Briggs, M. R., Kadonaga, J. T., Bell, S. P. and Tjian, R. (1986) *Science*, **234**, 47–52.
68 Ghosh, S., Jaraczewski, J.W., Klobutcher, L.A. and Jahn, C.L. (1994) *Nucleic Acids Res.*, **22**, 214–221.
69 Herrick, G. (1992) *J. Protozool.*, **39**, 309–312.
70 Proudfoot, N. J. and Brownlee, G. G. (1976) *Nature*, **263**, 211–214.
71 Allen, R.L., Olins, A.L., Harp, J.M. and Olins, D.E. (1985) *Eur. J. Cell Biol.*, **39**, 217–223.
72 Murti, K.G. and Prescott, D.M. (1983) *Mol. Cell Biol.*, **3**, 1562–1566.
73 Zahler, A.M and Prescott, D.M. (1988) *Nucleic Acids Res.*, **16**, 6953–6972.
74 Maercker, C. and Lipps, H.J. (1993) *Dev. Gen.*, **14**, 378–384.
75 Mount, S.W. (1982) *Nucleic Acids Res.*, **10**, 459–472.