An *Oxytricha trifallax* micronuclear BAC library

Thomas G. Doak*, Glenn Herrick*, Laura F. Landweber†

*Department of Pathology, University of Utah, Salt Lake City, Utah 84132; †Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544.

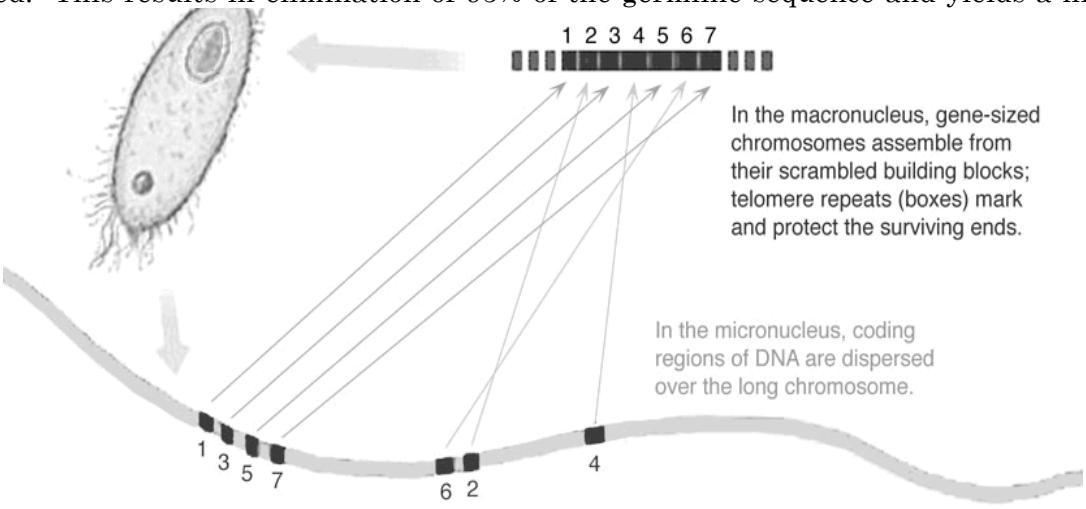
We propose construction of a BAC library for the germline micronuclear genome (MIC) of the ciliate protist *Oxytricha trifallax* (recently renamed *Sterkiella histriomuscorum*; alveolate; class Spirotrichia). This library will primarily be used as a tool to characterize the structure and function of the ciliate micronucleus and its relationship to that of the somatic macronucleus, rather than for complete genome sequencing. Micronuclear differentiation into macronuclei requires specialized DNA replication, cleavage and splicing, telomere addition, specific DNA destruction, and highly dramatic DNA rearrangement events, including the unscrambling of protein-coding gene fragments dispersed in the MIC. As many as 25-30% of the 20,000-30,000 genes in the MIC appear to be scrambled in complex patterns ranging from few to over 50 fragments, present at one or more micronuclear loci. These fragments become properly re-ordered and joined into linear gene-sized "chromosomes" in the MAC. Studying these DNA processing events requires the detailed comparison of micronuclear with macronuclear sequences, which will be greatly aided by the availability of a micronuclear BAC library.

Ciliates have been important experimental organisms for over 100 years, contributing to the discovery and understanding of many essential cellular processes—including self-splicing RNA, telomere biochemistry, and transcriptional regulation by histone modification—with *Oxytricha* representing the lineage with the surprising discoveries of gene-sized macronuclear mini-chromosomes and scrambled genes. Among Spirotrichs, *O. trifallax* also appears to have the highest density of scrambled genes in its micronucleus.

• INTRODUCTION TO CILATE BIOLOGY

Ciliates diverged from other microbial eukaryotes quite late in the history of life, as part of a monophyletic lineage (alveolates) with apicomplexans (e.g. *Plasmodium*) and dinoflagellates (Wright & Lynn 1997; Baldauf *et al*. 2000). Therefore, as phylogenetic outgroups, ciliates provide a foil to studies of the crown eukaryotes plants, animals, and fungi. Molecular biologists have focused on several ciliates, of one of two classes, the Oligohymenophorans *Tetrahymena* and *Paramecium*, and the Spirotrichs, with *Euplotes* and several Stichotrichs (*Stylonychia* and *Oxytricha*); each has its experimental virtues.

Ciliates have a typical diploid, meiotic germ-line genome in the micronucleus (MIC), except that its genes are not directly expressed, and those that are scrambled in the MIC have an exceptional organization (see figure). A highly specialized expression organelle, the macronucleus (MAC), develops from a mitotic copy of the MIC immediately after cells mate and exchange gametic nuclei. Stichotrich MAC differentiation is the premier showcase of somatic genome alterations (Jahn & Klobutcher 2002). As the MAC differentiates: MIC chromosome segments are differentially amplified as polytene chromosomes, DNA sequences interrupting genes are spliced out (internally eliminated segments, or IESs), MIC chromosomes are fragmented, scrambled fragments are joined together and re-ordered, telomeres added to the new ends, and the new MAC mini-chromosomes (sometimes called "gene-sized pieces") are amplified. At the same time the old parental MAC is apoptotically destroyed. This results in elimination of 95% of the germline sequence and yields a highly



polyploid MAC nucleus consisting of nearly pure coding DNA packaged on small fragments. The MAC is entirely responsible for vegetative growth. The MIC genome consists of MIC-limited and MAC-destined sequences.

In *Oxytricha* the MAC genome consists of genes deployed on a collection of ~20,000 different miniature (often gene-sized) chromosomes each at a ploidy of ~1000/MAC. Their total sequence complexity is ~50 Mb–thus an attractive genome size for complete sequencing. These chromosomes range from ~250 bp to ~40 kb (Maercker *et al* 1999). These chromosomes range from ~250 bp to ~40 kb (Maercker *et al* 1999). The MIC genome is ~10^9 bp (Lauth *et al*. 1976), 40% of which is repetitive.

Qualitatively, MAC differentiation is similar in most ciliates (Jahn & Klobutcher 2002), but the number of DNA splicing-events, break sites, and chromosome kinds are orders of magnitude higher in Spirotrichs; Oligohymenophoran MAC chromosomes contain hundreds of genes each, whereas Spirotrich MAC chromosomes contain only one or a few genes each. This Spirotrich pattern seems to be shared with other less-studied classes of ciliates (Riley & Katz 2001), suggesting that massive fragmentation may be both ancestral and more representative of ciliate biology than the modest fragmentation in Tetrahymena and Paramecium. Genome reorganization has been a major focus of ciliate research since the discovery of these dramatic processes.

> "No doubt ciliates are not quite the same as most other eucaryotes, but then the phages T4 and lambda are not the same as a typical procaryote. This has not meant that they have proved of no value for fundamental biological research. Ciliates may prove to be equally useful."
>
> Francis Crick in a letter to Hans Lipps, 1977.

• IMPORTANCE OF *O. TRIFALLAX* TO BIOMEDICAL AND BIOLOGICAL RESEARCH

Ciliates belong to a large group of deeply-diverging microbial eukaryotes, the alveolates, which include the parasites *Plasmodium*, *Toxoplasma*, and *Theileria*. While there are genome projects for these parasites, there are no projects for free-living alveolates. We expect that characterizing the genome of free-living protists will be important for understanding the genome organization, gene loss, and specializations of the medically important alveolates.

Current ciliate sequence projects have focused on the MAC genome or cDNA and ESTs. History suggests–and we and others have argued (see Orias 2000)–that knowing the complete proteome sequence of several ciliates will have immediate and far-ranging applications to both an understanding of basic eukaryotic biology and of pathways essential to human health and treatment.

Just as interesting and intrinsic to human biology are many of the processes required for DNA differentiation of the MAC genome (Jahn & Klobutcher 2002). Fragmented genes exist in a variety of taxa, including other Alveolates (e.g., split ribosomal RNAs in *Plasmodium* and *Theileria*). However, rarely are they actually "sewn" back together at the level of DNA, as in *Oxytricha*, to create a contiguous gene from all pieces. A notable exception is V(D)J splicing in the vertebrate immune system. To understand the processes of DNA rearrangement in ciliates, comparisons of MIC and MAC sequence are essential, but it is currently a challenging process to obtain the MIC cognate of a MAC sequence: the fragments that assemble from scrambled genes can be encoded on both strands, separated by long intervening sequences, or even on potentially unlinked loci (Landweber *et al*. 2000). Deciphering the organization of genes in the MIC will be greatly facilitated by the availability of a micronuclear BAC library, and this would pave the way towards partial MIC genome sequencing to understand the genetic architecture of a Spirotrich MIC. Especially, we would like to define the structures and sequences in the MIC that guide and direct these well-orchestrated events.

These include:

– Polytene chromosomes. MIC chromosomes in Stichotrichs form classic polytene chromosomes early in MAC differentiation, with bands, interbands, and puffs. Indeed, it should be easy to isolate biochemical amounts of polytene chromosomes. Signals for polytenization may well be MIC-limited. Most MIC DNA is eliminated as polytene chromosomes are fragmented, to create individual MAC chromosomes. Ultrastructural studies show that this elimination is associated with protein vesicles that enclose polytene bands, leaving interbands outside the vesicles (Prescott & Murti 1974). These structures

may be homologous to *Tetrahymena* "dumposomes" (Smothers *et al*. 1997). Understanding polytenization, fragmentation and elimination requires understanding the large-scale features of the MIC genome and how it maps to the polytene bands, interbands and MAC-destined sequences.
– Internally eliminated sequences (IESs; reviewed in Jahn & Klobutcher 2002). During and immediately after polytenization of the MIC genome, sequences interrupting genes are removed by precise excision: the interrupting sequences are spliced out and the flanking sequences rejoined to reconstitute the functional gene. Almost all Spirotrich and *Paramecium* genes carry multiple IESs. There are two classes of IES: short (two dozen to a few hundred base pairs) unique sequences, and large families of cut-and-paste transposons (4-5 kb long, with thousands of copies per haploid genome). In MAC development each transposon is removed, precisely reverting the germline insertional mutations. To identify and characterize IESs it is necessary to compare the MAC sequence to the MIC sequence. A BAC library would be very helpful for this.
– Scrambled genes (reviewed in Prescott 2000). In this case IESs interrupt genes, but contiguous gene segments are no longer adjacent, nor necessarily on the same strand or even at the same locus in the MIC (Landweber *et al*. 2000), and the order of gene segments is permuted in both random and strikingly nonrandom patterns, such as 1-3-5-7-2-4-6-8. In one known example a single gene in *O. trifallax* is scrambled into 51 MIC segments located in both orientations on two separate loci (Hoffman and Prescott 1997). IES removal somehow links coding segments in the correct, translatable, order and orientation. As many as 20-30% of *Oxytricha* genes are estimated to be scrambled in the MIC genome. We expect to find more cases of this, and need to be able to map these loci in the genome.
– Transposons. The biology of ciliate transposons is proving to be very interesting. The transposases of Mariner/Tc1 and Pogo-like elements–and their relationship to retroviral integrases–was first recognized because of new transposons discovered in ciliates (Doak *et al*. 1994). In addition to transposase, some ciliate transposons carry novel types of protein kinase and tyrosine recombinase. Also, the population dynamics of transposons in ciliates seems to be uniquely different than in metazoans: transposon genes exist under a positive selection for protein function (Klobutcher & Herrick 1997). This could be unique to ciliates, but more likely the ciliate case an extreme on a continuum of different transposon-host relationships. Understanding the ciliate case will illuminate the coevolution of hosts–including humans–and their transposon parasites.

• USES FOR THE O. TRIFALLAX BAC LIBRARY, IN ADDITION TO GENOMIC SEQUENCING

We do not intend this BAC library to be used for whole-genome sequencing (at this time); hence coverage even less than 5X would still be helpful. The library would be used immediately to:
– Isolate MIC loci that represent cognate MAC chromosomes, for both sequencing and to generate long constructs for transformation of *Oxytricha* or *Stylonychia*. The latter will allow us to test the *cis* sequence requirements for IES removal and unscrambling. Transformation of *Stylonychia* with short (kb) regions flanking macronuclear-destined DNA has proved insufficient to promote full processing of exogenous DNA constructs, but does allow correct telomere addition (Wen *et al*. 1995).
– Map MAC-destined sequences relative to each other.
– Identify and sequence repetitive MIC genome elements.
– Characterize chromosome number, size, and large-scale features of the MIC genome.

• THE SIZE OF THE RESEARCH COMMUNITY THAT COULD POTENTIALLY USE THE O. TRIFALLAX LIBRARY AND THE COMMUNITY'S INTEREST IN AND SUPPORT FOR HAVING A BAC LIBRARY

Ciliate molecular biologists and geneticists form an active and united community. There has been a biennial meeting of this group since 1982, recently as either Gordon or FASEB conferences, with an attendance of ~150. In addition to US workers, both European and Asian researchers are well represented at these meetings. An international group of Stichotrich researchers meets annually as part of Landweber's 5-year NSF ITR grant on their molecular genetics, evolution, and bioinformatics. The entire community is looking forward to ciliate genome information.

Major community efforts are in progress to advance the genomics of the two Oligohymenophorans, *T. thermophila* (Orias 2000) and *P. tetraurelia* (Dessen *et al*. 2001).

The *Tetrahymena* community (www.lifesci.ucsb.edu/~genome/Tetrahymena) has focused on EST projects at the outset (A. Turkewitz, pers. comm.), while procuring funding for sequencing the MAC genome. The *Paramecium* community is generating an equalized cDNA library for sequencing, coordinated by Jean Cohen of the CNRS.

Numerous researchers interested in the evolution of eukaryotes or the various parasitic alveolates will also benefit.

•WILL *O. TRIFALLAX* BE PROPOSED TO A PUBLICLY FUNDED AGENCY FOR BAC-BASED GENOMIC SEQUENCING. We are not aware of proposed BAC-based genomic sequencing for a ciliate.

•THE STRAIN OF *O. TRIFALLAX* PROPOSED AND RATIONALE FOR ITS SELECTION

O. trifallax represents the primary clade of focus in the Spirotrichs and is represented by many healthy strains in use in a number of labs. The various Stichotrichs are all similar enough that the genome sequence of any one of them will be informative for other species. Comparative approaches between the various Stichotrichs are already standard practice in various labs (e.g., Ammermann, Herrick, Landweber, Prescott).

O. trifallax is easily stored as frozen cysts–which is not true of many Stichotrichs. Without such storage, vegetative lines senesce, leading to sterility and death of the clone. Clonal senescence has plagued most Stichotrich research. Senescent cells do not mate, so studying MAC development has often required inferring developmental sequence changes by comparing MIC and MAC sequences. But non-senescent *O. trifallax* cultures are fully capable of fertile mating, allowing Mendelian genetic experiments, and mass synchronous matings to study biochemical intermediates and activities. We will use the isolate JRB310. It is one of a pair of isolates (of different mating types) that have been most used.

•THE SIZE OF THE *O. TRIFALLAX* GENOME

The MIC genome size of *O. trifallax* is approximately 5 x 10^8 bp, based on estimates of *O. fallax* (Herrick *et al.* 1985), and as much as one-third may be repetitive. There are on the order of 100 MIC chromosomes (Raikov 1982).

•THE AVAILABILITY OF A SOURCE OF *O. TRIFALLAX* DNA FOR CONSTRUCTION OF THE BAC LIBRARY (EVIDENCE OF ITS QUALITY FOR THIS PURPOSE)

There are several protocols for the purification of intact micronuclei in the presence of nuclease inhibitors, and pure MIC DNA has been prepared to examine MIC chromosomes by pulse-field gels (Maercker *et al.* 1999), so it should be relatively straight-forward to obtain intact MIC DNA for BAC cloning.

•SPECIFICATIONS AND RATIONALE FOR THE LIBRARY (E.G., LIBRARY DEPTH, BAC INSERT SIZE).

Since the BAC library will not be used to generate whole-genome sequence a modest coverage of ~5x should be sufficient. However, a library with greater coverage could be used in the future to sequence the entire MIC. Particularly large inserts are not necessary: we expect inserts of ~100 kb to be large enough to describe large scale features of the MIC genome, while being relatively easy to work with.

•THE TIME FRAME IN WHICH THE *O. TRIFALLAX* LIBRARY IS NEEDED. A BAC library could be put to use at any time by a number of labs, but there is no required time frame. If sequencing of the MAC genome is funded, there will sharply increased interest in a MIC library.

•OTHER SUPPORT AVAILABLE OR REQUESTED FOR CONSTRUCTION OF THE LIBRARY

No other support has been requested for a BAC library. In a separate NHGRI white paper we have requested that the *O. trifallax* MAC genome be sequenced, and also propose exploratory MIC sequencing, focused on identifying certain regions containing MAC-destined sequences. These MIC sequences are proposed to come from a lambda or cosmid library; this would allow us to sequence small clones containing specific MAC-destined segments. But such small clones are not ideal for general MIC genome characterization.

The Landweber lab has an active 5-year NSF ITR grant to study the unscrambling of MIC genes during MAC development, using molecular biology, comparative biology and a range of computational approaches. This intriguing example of biocomplexity and biocomputation has stimulated both national and international collaboration of Stichotrich and computational researchers. This NSF grant could provide support for analyses of BAC-generated results.

In a separate NSF proposal the Herrick lab has proposed a set of experiments that will: generate an archived and indexed 9.2-hit plasmid library of MAC chromosomes; screen the library for examples of alternatively processed MAC chromosome families; obtain end sequence for these examples; and completely sequence some of these clones. An inexpensive project, focused on a specific phenomenon, it will generate a high-quality library of MAC clones to pair with the proposed MIC BAC library.

•**THE NEED FOR AN ADDITIONAL BAC LIBRARY IF ONE OR MORE ALREADY EXISTS.** We are not aware of a BAC library for any ciliate MIC or MAC, nor of other plans to make one.

LETTERS OF SUPPORT. We have enclosed letters of support from Larry Klobutcher and John Logsdon, since they make particularly salient points. We have also exuberant letters of support from Germany, France, Canada, and the Netherlands, as well as from researchers in this country.

REFERENCES

Ammermann D, Muenz A. 1982. DNA and protein content of different hypotrich ciliates. Eur J Cell Biol. 27:22-24

Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. Science. 290:972-977.

Doak TG, Doerder FP, Jahn C, Herrick G. 1994. A family of transposases genes in transposons found in prokaryotes, multicellular eukaryotes and ciliated protozoans. Proc Natl Acad. Sci USA. 91:942-946.

Dessen P, Zagulski M, Gromadka R, Plattner H, Kissmehl R, Meyer E, Betermier M, Schultz JE, Linder JU, Pearlman RE, Kung C, Forney J, Satir BH, Van Houten JL, Keller AM, Froissard M, Sperling L, Cohen J. 2001. *Paramecium* genome survey: a pilot project. Trends Genet. 17:306-308.

Herrick G, Cartinhour S, Dawson D, Ang D, Sheets R, Lee A, Williams K. 1985. Mobile elements bounded by C4A4 telomeric repeats in *Oxytricha fallax*. Cell. 43:759-768.

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

Jahn CL, Klobutcher LA. 2002. Genome remodeling in ciliated protozoa. Ann. Rev. Microbiol. 56. In press.

Klobutcher LA, Herrick G. 1997. Developmental genome reorganization in ciliated protozoa: the transposon link. Prog Nucleic Acid Res and Mol Biol. 56:1-62.

Landweber LF, Kuo TC, Curtis EA. 2000. Evolution and assembly of an extremely scrambled gene. Proc Natl Acad Sci U S A. 97:3298-3303.

Lauth MR, Spear BB, Heumann J, Prescott DM. 1976. DNA of ciliated protozoa: DNA sequence diminution during macronuclear development of *Oxytricha*. Cell. 7:67-74.

Maercker C, Kortwig H, Lipps HJ. 1999. Separation of micronuclear DNA of *Stylonychia lemnae* by pulsed-field electrophoresis and identification of a DNA molecule with a high copy number. Genome Res. 9:654-661.

Orias E. 2000. Toward sequencing the *Tetrahymena* genome: exploiting the gift of nuclear dimorphism. J Eukaryot Microbiol. 47:328-333.

Prescott DM. 2000. Genome gymnastics: unique modes of DNA evolution and processing in ciliates. Nat Rev Genet. 1:191-198.

Raikov IB. 1982. The protrozoan nucleus. Springer-Verlag (Vienna).

Riley JL, Katz LA. 2001. Widespread distribution of extensive chromosomal fragmentation in ciliates. Mol Biol Evol. 18:1372-1377.

Smothers JF, Mizzen CA, Tubbert MM, Cook RG, Allis CD. 1997. Pdd1p associates with germline-restricted chromatin and a second novel anlagen-enriched protein in developmentally programmed DNA elimination structures. Development. 124:4537-4545.

Wen JP, Eder C, Lipps HJ. 1995. The processing of macronuclear-destined DNA sequences microinjected into the macronuclear anlagen of the hypotrichous ciliate *Stylonchia lemnae*. Nucleic Acids Res. 23:1704-1709.

Wright ADG, Lynn DH. 1997. Maximum ages of ciliate lineages estimated using a small subunit rRNA molecular clock: Crown eukaryotes date back to the Paleoproterozoic. Archiv Protistenkd. 148: 329-341.