Microbiology (2005), **151**, 121–133 DOI 10.1099/mic.0.27353-0

Molecular analysis of the anaerobic rumen fungus *Orpinomyces* – insights into an AT-rich genome

Matthew J. Nicholson,[1,2]† Michael K. Theodorou[1] and Jayne L. Brookman[1]

[1]Institute of Grassland and Environmental Research, Plas Gogerddan, Aberystwyth, Ceredigion SY23 3EB, UK

[2]School of Biological Sciences, University of Manchester, 1.800 Stopford Building, Oxford Road, Manchester M13 9PT, UK

Correspondence
Michael K. Theodorou
mike.theodorou@bbsrc.ac.uk

Received 21 May 2004
Revised 22 September 2004
Accepted 29 September 2004

The anaerobic gut fungi occupy a unique niche in the intestinal tract of large herbivorous animals and are thought to act as primary colonizers of plant material during digestion. They are the only known obligately anaerobic fungi but molecular analysis of this group has been hampered by difficulties in their culture and manipulation, and by their extremely high A+T nucleotide content. This study begins to answer some of the fundamental questions about the structure and organization of the anaerobic gut fungal genome. Directed plasmid libraries using genomic DNA digested with highly or moderately rich AT-specific restriction enzymes (*Vsp*I and *Eco*RI) were prepared from a polycentric *Orpinomyces* isolate. Clones were sequenced from these libraries and the breadth of genomic inserts, both genic and intergenic, was characterized. Genes encoding numerous functions not previously characterized for these fungi were identified, including cytoskeletal, secretory pathway and transporter genes. A peptidase gene with no introns and having sequence similarity to a gene encoding a bacterial peptidase was also identified, extending the range of metabolic enzymes resulting from apparent trans-kingdom transfer from bacteria to fungi, as previously characterized largely for genes encoding plant-degrading enzymes. This paper presents the first thorough analysis of the genic, intergenic and rDNA regions of a variety of genomic segments from an anaerobic gut fungus and provides observations on rules governing intron boundaries, the codon biases observed with different types of genes, and the sequence of only the second anaerobic gut fungal promoter reported. Large numbers of retrotransposon sequences of different types were found and the authors speculate on the possible consequences of any such transposon activity in the genome. The coding sequences identified included several orphan gene sequences, including one with regions strongly suggestive of structural proteins such as collagens and lampirin. This gene was present as a single copy in *Orpinomyces*, was expressed during vegetative growth and was also detected in genomes from another gut fungal genus, *Neocallimastix*.

INTRODUCTION

The anaerobic gut fungi are an unusual group of organisms belonging to the Chytridiomycetes, family *Neocallimastigaceae* (http://www.indexfungorum.org). There are six recognized genera, namely *Neocallimastix*, *Piromyces*, *Orpinomyces*, *Anaeromyces*, *Caecomyces* and the recently described *Cyllamyces* (Munn *et al.*, 1988; Ozkose *et al.*, 2001). They occupy a unique niche in the gastrointestinal tract of large herbivorous animals, including ruminants such as cows and sheep and the hindgut-fermenting animals such as elephants and horses (Theodorou *et al.*, 1996). The gut fungi are thought to be the primary colonizers of plant material in the rumen, and together with rumen bacteria and protozoa they are responsible for the degradation of ingested plant biomass that would be otherwise indigestible to the host animal.

The vast repertoire of potent plant cell-wall degrading enzymes secreted by these fungi has been characterized more fully than any other area of their biology. They produce a complex supramolecular structure, similar to the clostridial cellulosome, to process and degrade ingested plant cell walls (Bayer *et al.*, 1998). Furthermore, a significant number of these enzymes appear to be the result of trans-kingdom transfer from bacteria as identified by analysis of codon bias

†Present address: AgResearch Grasslands, Private Bag 11008, Palmerston North, New Zealand.

Abbreviations: ETS, external transcribed spacer; LTR, long terminal repeat.

The GenBank/EMBL/DDBJ accession numbers for the sequences determined in this work are given in Supplementary Table S1 with the online version of this paper at http://mic.sgmjournals.org.

Downloaded from www.microbiologyresearch.org by
IP: 93.91.26.109
On: Fri, 27 Nov 2015 19:02:49

(Garcia-Vallvé *et al.*, 2000). The energy-generation machinery of the gut fungi has also been investigated. The organelle responsible for energy generation in the absence of any mitochondria, the hydrogenosome, has been used as a model to understand the metabolic processes of obligately anaerobic eukaryotes (Embley *et al.*, 2003). Progress in understanding the ecology of the rumen ecosystem has been enhanced by the development of molecular phylogenetic approaches for the fungi, together with more advanced rumen bacterial methodologies (Brookman *et al.*, 2000; Kocherginskaya *et al.*, 2001).

Relatively little information is available on the genomes of these fungi due to their extreme nature; the AT content of the anaerobic fungal genome is approximately 80–85 mol% and is amongst the highest reported in any organism (Billon-Grand *et al.*, 1991; Brownlee, 1989). This extreme nucleotide bias is reflected in both the coding and non-coding regions of the genome, with codon usage tending towards more AT-rich codons (Garcia-Vallvé *et al.*, 2000). The non-coding regions of the anaerobic fungal genome are known to be extremely AT-rich, with many regions expected to be near or above 95 mol% AT content (Brookman *et al.*, 2000; Brownlee, 1989; Durand *et al.*, 1995; Harhangi *et al.*, 2003a; Steenbakkers *et al.*, 2002).

The AT content of the gut fungal genome and the consequent technical problems associated with genome manipulation of DNA that is often unstable in normal bacterial cloning hosts (see Gardner, 2001), particularly when compounded by the poor yields of material from anaerobically growing organisms, has meant that few studies have been performed using genomic DNA and most of the molecular data are derived from cDNA. Hence, little is known of the structure and organization of the anaerobic fungal genome. Paradoxically, rather then seeing it as a problem, the study reported herein takes advantage of the biased nucleotide content and begins to address some fundamental questions about genome organization in the gut fungi. Plasmid-based genomic libraries have been made, exploiting the atypical nucleotide composition of the gut fungal genome to influence the composition of the individual libraries. Clones from these libraries have been sequenced and a characterization of the resultant population presented. These data enable an early description of some basic parameters of genome organization, such as intron–exon boundary sequences, ribosomal sequence variation and patterns of AT content. The libraries describe putative orphan genes, including a gene that is expressed during vegetative growth, is present in the genome of more than one genus of gut fungi and appears to encode a protein which may be of considerable interest to structural biologists. The libraries also indicate a preponderance of retrotransposon sequences in the *Orpinomyces* genome. We believe our analysis provides insight into the genome content and structure of these biologically interesting and agronomically important micro-organisms.

METHODS

Fungal strains and maintenance. The polycentric fungus used in this study (OUS1) was isolated from the rumen of UK sheep and characterized as an *Orpinomyces* strain (Brookman *et al.*, 2000). The *Neocallimastix* strains used for characterization of the orphan gene were NMZ4, NMW3 and NMW5 (Brookman *et al.*, 2000). Fungal cultures were maintained in Orpin's medium (Orpin, 1975) with either cellobiose (2·5 g l^{-1}) or milled wheat straw (5 g l^{-1}) as the major source of carbon, under anaerobic conditions, at 39 °C without shaking (Lowe *et al.*, 1985).

Genomic DNA extraction. DNA was extracted from biomass harvested by centrifugation (3000 ***g***, 10 min) from 48 h cultures grown on cellobiose using the method described by Brookman *et al.* (2000).

Plasmid genomic library production and sequencing. Genomic DNA was fully digested with either *Eco*RI or *Vsp*I and digested fragments cloned into the commercial pBluescript (using the *Eco*RI cloning site) or pGEM 5zf+ (using the *Nde*I cloning site) vectors respectively. *Eco*RI recognizes the sequence G↓AATTC and was selected to cut throughout the genome with a possible bias towards cloning intergenic fragments. *Vsp*I recognizes the sequence AA↓TATT and was selected to cut within the high AT content non-coding regions of the genome and hence generate gene- or exon-sized fragment clones. The *Vsp*I-digested genomic DNA was first separated on an agarose gel and five different-sized fragment ranges were excised and gel purified (using a Qiagen PCR purification kit) for cloning separately. The fragments cloned were <0·7 kbp, 0·7–1 kbp, 1 kbp, 1–2 kbp and 5 kbp. All ligation reactions were performed using T4 DNA ligase. All sequencing was performed by MWG Biotech. Clones were routinely cloned to a minimum of 99 % accuracy (Phred 20 score) on a single strand plus a minimum of 40 % overlap between sequences to give partial second-strand information. Traces were observed manually for contentious calls and possible errors giving stop codons, inappropriately placed intron boundaries and/or frameshifts from a known sequence; in such cases, clones were repeat sequenced to ensure accuracy.

Sequence analysis. The proportion of A+T nucleotides within a 19 nt moving window was plotted against the nucleotide position to form a 'pAT plot' and potential coding regions were identified by visual analysis. Sequences were compared to the databases using BLASTN, BLASTX and BLASTP for potential coding regions using GenomeNet (http://blast.genome.ad.jp/). Sequences were translated using the Protein Machine website (http://www2.ebi.ac.uk/translate/). Codon usage was calculated with the aid of the Countcodon program (http://www.kazusa.or.jp/codon/countcodon.html). The following genes with their accession numbers were used for codon usage analysis. Glycosyl hydrolase genes: *Piromyces* sp. endo-1,4-β-mannanase (X97520), mannanase A (X91857), endo-1,4-β-mannanase (X97408) and xylanase A (X91858); *Orpinomyces* sp. cellulase B (U57818), cellulase A (U63837), cellulase C (U63838), xylanase A (U57819), lichenase A (U63813) and cyclophilin B precursor (U17900); *Orpinomyces joyoni* cellulase B29 (AF015248), cellulase B2 (AF015249) and endoglucanase precursor (U59432). Metabolic and housekeeping genes: *Neocallimastix frontalis* enolase gene (X80474), phosphoenolpyruvate carboxykinase (M59372), malic enzyme precursor (U62041) and β-succinyl-CoA synthetase precursor (X84222); *Piromyces* sp. hydrogenosomal adenylate kinase (AJ224660), isocitrate dehydrogenase (Y16751), malate dehydrogenase (Y16748), aconitate hydratase (Y16747), ketol-acid reductoisomerase (Y16743) and formate C-acetyltransferase (Y16739).

The data were analysed using a Poisson log-linear model in which the gene totals were fixed, resulting in a multinomial model with k codon

Downloaded from www.microbiologyresearch.org by
IP: 93.91.26.109
On: Fri, 27 Nov 2015 19:02:49

classes (GenStat for Windows, Release 7.1, 7th edition; vsN International). Whether the proportions of codons over all genes were equal was tested by fitting Codon as a main effect, and whether the distribution of codons was the same for each gene-type was tested by fitting the interaction between Codon and Gene-type. Codons were combined when necessary to ensure that the expected values for all Codon–Gene-type combinations were greater than five, as required, except when there were only two codons for an amino acid, in which case Fisher's exact 2×2 tests were used instead.

RNA extraction and RT-PCR. RNA was extracted from *Orpinomyces* OUS1 cultures grown in Orpin's medium with cellobiose as sole carbon source for 20, 42 and 66 h, using the aminosalicyclate extraction, lithium chloride precipitation method described by Sambrook *et al.* (1989). Samples were treated with RQ1 RNase-free DNase (Promega) before RT-PCR. Primers MN7 ($^{5'}$GGA CCA GAA TAT GGA ATG CCT$^{3'}$, forward) and MN13 ($^{5'}$TTG TGG TAC CAT ACC AGG ACT$^{3'}$, reverse), designed from the sequence of clone V3-7 were used for first-strand cDNA synthesis (MN13) and subsequent PCR amplification of the cDNA sequence. All RT-PCR reactions contained 300 ng RNA. One-step RT-PCR reactions used ABgene 2× pre-aliquoted PCR master mix with the addition of 0·5 µl Stratascript reverse transcriptase (Stratagene). Amplification was achieved under the following conditions: first-strand synthesis at 42 °C for 30 min, denaturation at 95 °C for 5 min then 40 cycles (90 °C, 2 min/55 °C, 1 min/68 °C, 1 min) and a final extension phase at 68 °C for 10 min.

PCR and Southern blot analysis. Primers MN17 and MN18 were used to determine the presence of a similar gene product to the V3-7 gene fragment in two *Neocallimastix* strains. MN17 was designed as a forward primer from the middle of the clone's coding region with a *Bam*HI linker for cloning, $^{5'}$CGC CGG ATC CCC AAT ACA AGT ACC AAT G$^{3'}$ and the reverse primer MN18 complementary to the 3′ terminus of V3-7 with an *Eco*RI linker, $^{5'}$GCG CGA ATT CTT ATT GAG ATG GTC TTG GAC$^{3'}$, giving a 500 bp fragment size from genomic DNA.

A Southern blot was prepared with *Orpinomyces* genomic DNA and probed with a 350 bp *Xba*II/*Sal*I restriction fragment corresponding to the 5′ region of the clone V3-7 gene fragment. The probe was labelled with digoxigenin-dUTP and hybridized with the Southern membrane according to the manufacturer's protocols (Boehringer Mannheim).

RESULTS AND DISCUSSION

Sequence analysis

Two plasmid libraries were constructed from *Orpinomyces* (OUS1) genomic DNA for sequencing and subsequent analysis. The libraries were designed to make use of the biased AT content of the gut fungal genome: the enzyme *Vsp*I recognizes a 100 % AT 6-base sequence and is thus more likely to digest intergenic/non-coding DNA, whereas the 66 % AT 6-base recognition site of *Eco*RI makes it more likely to digest the *Orpinomyces* DNA within the lower AT content genic regions (see Fig. 1); however, the clones analysed show no strong bias for the coding versus intergenic sequences, except that the *Vsp*I library gave a large number of ribosomal clones (6/20) and fewer retrotransposons (3/20) compared with the *Eco*RI library (5/91 and 21/91, respectively). The 5 kbp band visible on the *Vsp*I digest (Fig. 1, lane 2) was cloned separately to enable contiguous representation of the rDNA within the

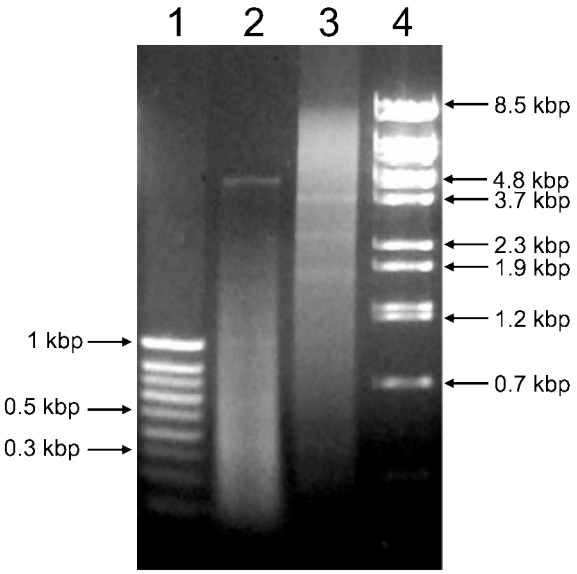

Fig. 1. *Orpinomyces* (OUS1) genomic DNA digested with restriction enzymes *Vsp*I, recognition sequence ATTAAT (lane 2), and *Eco*RI, recognition sequence GAATTC (lane 3). Note the prevalence of low-molecular-mass fragments produced by *Vsp*I; this unusual restriction pattern is due to the extremely high AT content of anaerobic fungal genomic DNA. Lane 1, Bioline hyperladder IV. Lane 4, Amresco low-range molecular mass marker.

sequenced clones. A total of 101 sequences from the plasmid genomic libraries (20 from the *Vsp*I library and 81 from the *Eco*RI library) were analysed and the clones assigned to functional groups as shown in Table 1.

For each nucleotide sequence, putative coding regions were identified by the use of a 'pAT plot' in which the proportion of A+T within a 19 nt moving window is plotted against the nucleotide position. This approach to sequence analysis is similar in principle to hydrophobicity plots used to analyse protein sequences. Potential coding regions can be identified as areas of decreased AT content (Fig. 2); regions with an overall AT content below 70 % and reading frames with at least one of the six reading frames devoid of putative stop codons, apart from any putative introns, were designated possible coding sequences.

The use of the pAT plot technique for identification of potential gene products and intron boundaries is shown in Fig. 2 with the *N. frontalis* enolase gene plot for demonstration purposes in Fig. 2(a), with the open reading frames, intron and up- and downstream regions indicated as identified by Durand *et al.* (1995). Fig. 2(b) shows the plot for clone E2-98, with 5′ upstream sequence, five exons and four introns of the malate dehydrogenase gene (Table 2; Akhmanova *et al.*, 1998). Repetitive sequences were also identifiable using the visual information from the pAT plot as seen for clone E2-50, a gypsy-like retrotransposon

Downloaded from www.microbiologyresearch.org by
IP: 93.91.26.109
On: Fri, 27 Nov 2015 19:02:49

Table 1. Functional groups of genomic clones

Clones containing	Number	Details
Coding sequence homologous to known proteins	12	A range of housekeeping proteins and extracellular enzymes (see Table 2)
Coding sequence of unknown function (hypothetical proteins or orphan sequences)	14	Sequences which appear coding (sensible reading frame: low AT content and devoid of stops) but show only low or no sequence similarity to known proteins
Non-coding sequence	20	Contain no sensible reading frame, often AT-rich
Ribosomal sequence	11	Clones include coverage from the ETS to the 28S gene
Retrotransposon sequence	24	16 different gypsy-like sequences, 4 copia-like and 4 non-LTR retrotransposon sequences
Unassigned	20	No clear assignation could be made for sequence type

sequence, where the regular pattern of the long terminal repeat can be seen as a region of repeating spikes between nucleotides ~75 and 325 (Fig. 2c). The increase in AT content downstream of coding regions at the 3′ end of genes can also be seen as shown in Fig. 2(d) for a xylanase Y clone (E2-15, Table 2). Complete sequences and potential coding regions identified by pAT plots were compared with database sequences by BLAST analysis.

Coding sequences

The genes identified in this study encode a wide variety of proteins, some of which would be expected in a fungal eukaryote and others that are more surprising (see Table 2). For example, the proteins involved in the secretory pathway and cytoskeleton such as clones E2-34 and E1-9 – the putative UDP-glucose phosphotransferase and dynein heavy chain respectively – would be expected within any

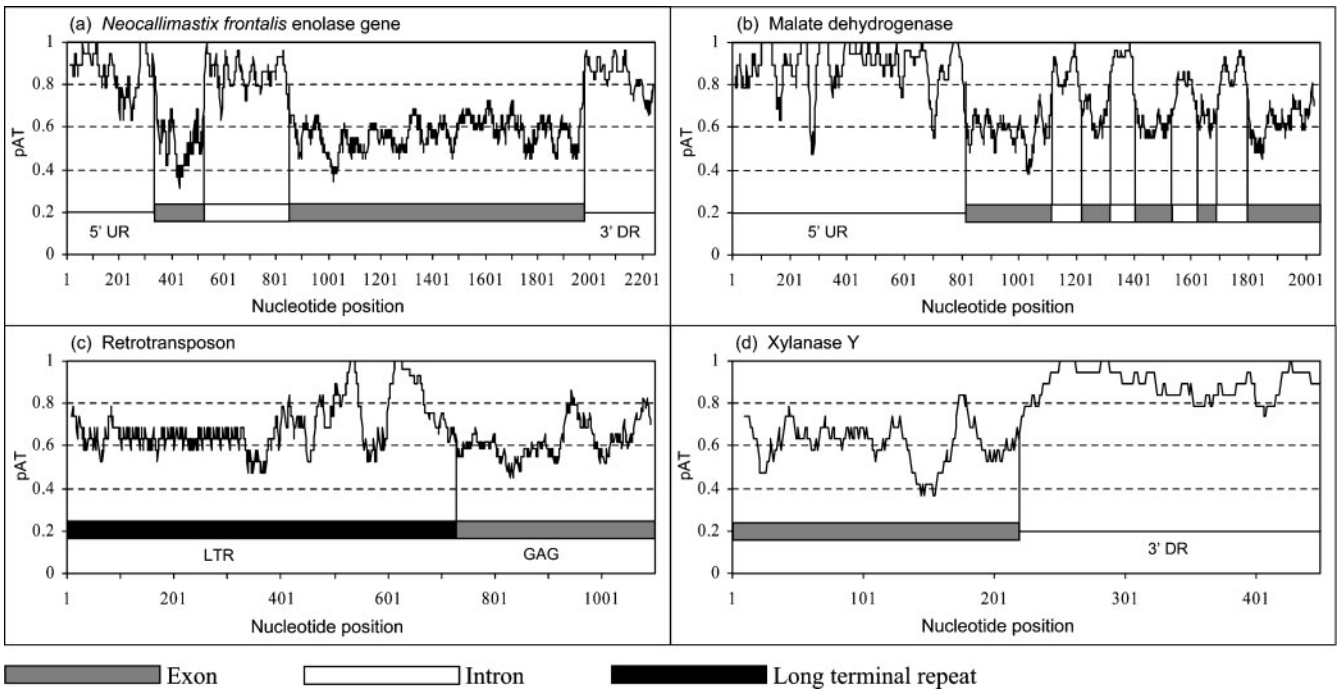

Fig. 2. The proportion of A+T over a moving window of 19 nt is plotted against the nucleotide position to form a 'pAT plot'. pAT plots are useful for identifying coding sequences due to the difference in AT content between coding and non-coding regions of the genome. (a) The previously published *N. frontalis* enolase gene, containing upstream and downstream regions (UR and DR) and one intron (Durand *et al.*, 1995; Fischer *et al.*, 1995), illustrates the use of this technique in identifying coding regions within the sequence. (b) Clone E2-98 encoding part of the malate dehydrogenase gene, four complete introns and 805 bp of the upstream region. (c) Clone E2-50 encoding part of a gypsy-like retrotransposon with upstream long terminal repeat (LTR). Note that the LTR contains a 17mer imperfectly repeated 15 times. (d) Clone E2-15 encodes the 3′ end of the bacterial-type xylanase Y gene and downstream region.

Downloaded from www.microbiologyresearch.org by
IP: 93.91.26.109
On: Fri, 27 Nov 2015 19:02:49

Table 2. Clones containing sequences for which the deduced amino acid sequences are similar to known or hypothetical proteins

Clone	Size (kbp)	Homologue	Details	Amino acid identity	BLAST score	*E* value
E2-3	0·240	Pyruvate formate-lyase (*Piromyces* sp.)	Hydrogenosomal enzyme	78/80 (97 %)	168	10^{-41}
E2-34	0·376	UDP-glucose phosphotransferase (*Schizosaccharomyces pombe*)	Contains one complete intron	60/76 (78 %)	138	10^{-32}
E2-98	2	Malate dehydrogenase (*Piromyces* sp.)	90 % of gene with four complete introns	85/144 (59 %)	400	10^{-111}
V4-9, V4-8	1·671	Aminoacyl-histidine dipeptidase (*Vibrio cholerae*)	96 % of complete gene sequence; intronless	109/292 (37 %)	280	4×10^{-74}
E2-70	1	AZR1 protein (*Schizosaccharomyces pombe*)		53/143 (37 %)	102	5×10^{-21}
E1-9	0·248	Dynein heavy chain (*Arabidopsis thaliana*)		28/72 (38 %)	99	8×10^{-21}
E2-6	0·699	Hypothetical protein (*Fugu rubripes*)	Contains introns	45/99 (45 %)	105	2×10^{-22}
E2-312	0·399	Hypothetical protein (*Plasmodium falciparum*)		33/110 (30 %)	45	3×10^{-4}
E3-231	0·128	Multidrug resistance gene (*Plasmodium falciparum*)	ABC transporter	26/42 (61 %)	52	10^{-6}
E2-104	0·575	Nucleoporin (*Saccharomyces cerevisiae*)	Same protein as E2-91 (upstream)	41/142 (28 %)	58	7×10^{-8}
E2-15	1·3	Xylanase B (*Ruminococcus albus*)	Terminator and 3′ downstream sequence	55/72 (76 %)	157	4×10^{-37}
E2-85	2	Xylanase B (*Neocallimastix patriciarum*)	2 kb covering the central region of the gene	111/125 (88 %)	400	10^{-111}
E2-35	0·143	Golgi UDP-GlcNAc transporter (*Kluyveromyces lactis*)	Contains intron	12/24 (50 %)	32	1·5

fungal genome. Similarly, the nucleoporin protein (clone E2-104) is likely to be present in all eukaryote forms although none of the genes above have been identified in the gut fungal family to date.

Genes encoding proteins previously observed in the anaerobic fungi include the metabolic enzymes pyruvate formate-lyase (clone E2-3, Table 2; Akhmanova *et al.*, 1999) and malate dehydrogenase (clone E2-98, Table 2; Akhmanova *et al.*, 1998), and the plant cell-wall degrading enzyme xylanase B (clone E2-85, Table 2; Black *et al.*, 1994). The energy production machinery and plant cell-wall degrading enzymes from the gut fungi have been well characterized in *Piromyces* and *Neocallimastix* isolates (e.g. Harhangi *et al.*, 2003a, b; Steenbakkers *et al.*, 2001, 2002; van der Giezen *et al.*, 1997).

Proteins of no known function or gene products designated hypothetical proteins in other organisms were also found (Table 2). For example, clone E2-6 appears to be coding and contains putative introns. When these are removed from the sequence the predicted translation is very similar to hypothetical proteins in other organisms including *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe* and *Fugu rubripes* (32–45 % identity).

There are several putative genes that appear to be of bacterial origin; they have no observable intron sequences and are similar to known bacterial genes, such as aminopeptidase D from a number of prokaryotes including *Vibrio cholerae* and *Escherichia coli* for clone V4-9 (Table 2; Henrich *et al.*, 1990) and the *Clostridium thermocellum* and *Ruminococcus albus*-like xylanase Y (B) for clone E2-15 (Table 2; Fontes *et al.*, 1995).

The presence of bacterial-type genes encoding plant cell-wall degrading enzymes within the anaerobic gut fungi is thought likely to be the result of a trans-kingdom transfer from the anaerobic bacteria within the rumen (Zhou *et al.*, 1994). Analysis of codon bias has shown that the known glycosyl-hydrolase genes of the anaerobic fungi are more similar to their rumen bacterial homologues than to aerobic fungal genes (Garcia-Vallvé *et al.*, 2000). Neither of the two xylanase genes revealed in this study (clones E2-15 and E2-85, Table 2) contains any introns within the regions sequenced and clone E2-15 is very similar, with 76 % amino acid identity, to a rumen bacterial xylanase. Further phylogenetic analysis of this group of proteins was not possible as only bacterial examples of this protein were found with BLASTP and TBLASTX searches of the sequence databases, suggesting that these genes are both of bacterial origin.

Downloaded from www.microbiologyresearch.org by
IP: 93.91.26.109
On: Fri, 27 Nov 2015 19:02:49

One clone, V4-9, by comparison with its nearest neighbour sequence from *Vibrio cholerae*, contained an almost complete intronless amino acid dipeptidase encoding gene. Alignment of a range of similar sequences showed that the gut fungal and bacterial proteins could be aligned over the entire 1383 base coding sequence with a small gap between nucleotides 1054–1125. No eukaryotic peptidases were capable of alignment with these proteins and hence no further phylogenetic analysis was possible. This observation of an apparent bacterially derived metabolic protein in the anaerobic gut fungal genome, together with similar observations from others (Akhmanova *et al.*, 1999; Harhangi *et al.*, 2003c; Voncken *et al.*, 2002), suggests that this may be a broader phenomenon than the transfer of cell-wall degrading enzymes alone. Indeed, genome studies are likely to confirm that trans-kingdom transfer of both central metabolic and auxiliary enzymes may be more widespread than originally thought. Two recent studies of the anaerobic protist *Giardia lamblia* have suggested lateral transfer of metabolic enzyme genes from prokaryotes, although these appear to be ancient transfers (Andersson & Roger, 2002; Nixon *et al.*, 2002).

Codon usage patterns for the gut fungi were analysed using our *Orpinomyces* sequences together with published *Orpinomyces*, *Neocallimastix* and *Piromyces* sequences. The samples were divided into housekeeping/metabolic sequences, glycosylhydrolases (bacterial-type), and gypsy-like retrotransposon sequences (5113, 6347 and 5219 codons, respectively). The overall AT content of the groupings was 56·9, 61·2 and 67·5 %, and for all codons the AT content increased from the first through to the third or wobble base of each codon (see Fischer *et al.*, 1995). This pattern was most pronounced in the least AT-rich set, i.e. the housekeeping genes, and least so in the gypsy-like sequences, which displayed the highest AT content.

For each amino acid the codon usage patterns were compared and statistically tested for each of the three different gene types (data not shown). The housekeeping/metabolic, glycosylhydrolase and gypsy groups displayed distinctly different codon usage patterns ($P<0{\cdot}001$, with P values as low as 10^{-89}) for most amino acids except those discussed below. Aspartate and cysteine showed no difference in their codon usage for all three gene types whilst glutamine and histidine had similar codon biases for the gypsy sequences and glycosylhydrolase genes. For tyrosine codons, a lack of statistical significance was seen when the housekeeping genes were compared with both other gene types; however, this was probably due to a much smaller sample size for tyrosine-encoding codons in this gene type (see below). More detailed examination of the statistical significance data revealed that in most cases the gypsy sequences and housekeeping codon usage patterns displayed the greatest dissimilarity whereas the greatest similarity was between the glycosylhydrolases and the housekeeping genes.

Codon selection, as characterized in bacteria, appears to be the result of a variety of processes acting on the genome of an organism: highly expressed genes exhibit bias towards codons that are optimal for translation; genes that are present in the genome as a result of horizontal transfer show a second stream of bias; and finally the position of the gene sequence on the leading or lagging DNA strand provides a third route of bias (see Grocock & Sharp, 2002). The small-scale analysis reported here appears to concur with the first two of these and we await with interest the conclusion from a genome-wide analysis from any genome sequencing effort in the future. Similarly the actual usage of amino acids between the groups varied, with a very strong bias in favour of tryptophan residues in glucosylhydrolase proteins (205 compared with 45 and 55), considerably lower usage of glycine and alanine by gypsy gene sequences (200 versus 615 and 437, and 225 versus 459 and 421 respectively) and lower asparagine residues for the housekeeping gene set (31 versus 246 and 302).

Thirteen clones appear to contain exons but gave no significant results for BLAST searches; we have designated these as orphan sequences. The number of orphan or poorly matched genes identified from genome sequences is quite variable, e.g. *Schizosaccharomyces pombe* contained very few orphans, with 99·7 % genes showing homology to known proteins and 76·5 % of its gene complement assigned to functional categories (http:/mips.gsf.de/proj/yeast/CYGD/hemi) whereas the sequence from the *Plasmodium falciparum* genome, which is approximately twice the physical size of the *Schiz. pombe* genome but with a similar number of genes, provided many new gene products with approximately 60 % genes appearing to be unique to this organism (Gardner *et al.*, 2002a). One of the orphan sequences from *Orpinomyces* with interesting amino acid sequence characteristics is characterized further below.

Preliminary characterization of an orphan gene

One of the cloned sequences identified in this study was of particular interest as it contained, for the gut fungi, an usually GC-rich nucleotide sequence with repetitive features in the pAT plot flanked by shorter regions of high AT content (Fig. 3a). Further analysis showed the clone to contain the terminal exon of a gene encoding a protein rich in amino acids encoded by GC-rich codons, in particular glycine, proline and alanine. The translated sequence showed an unusual, eight-residue, repetitive motif with the consensus GPPGAPPM imperfectly repeated seven times. The repetitive motif was followed by a further 183 amino acid residues, also rich in glycine and proline (Fig. 3b). The symmetry and amino acid composition, together with the presence of a GP-rich tail to the predicted protein, suggested a possible functional/structural similarity with structural proteins such as collagens and lampirin found in the extracellular matrix of higher eukaryotes (Kielty *et al.*, 1993).

An assessment of RNA expression in mid-exponential, late-exponential and early stationary-phase *Orpinomyces* cultures using RT-PCR showed expression that may be growth related. The RNA product was detectable at 20 h and at 42 h

Downloaded from www.microbiologyresearch.org by
IP: 93.91.26.109
On: Fri, 27 Nov 2015 19:02:49

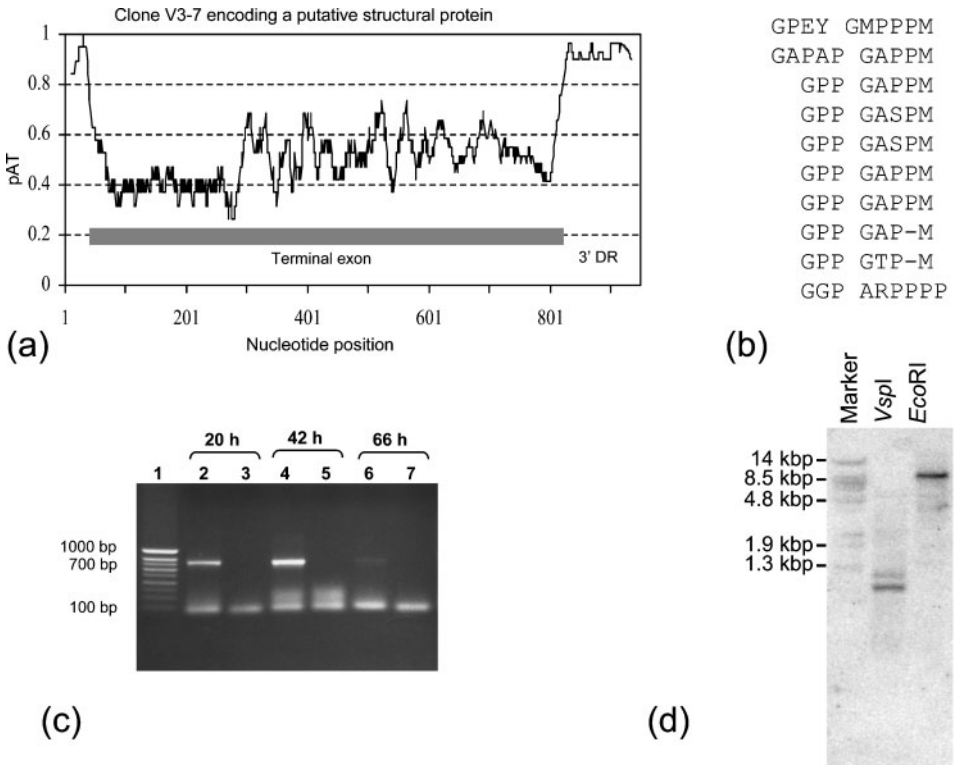

Fig. 3. (a) pAT plot anlaysis of a 934 bp clone V3-7 shows it to be a terminal exon with an extremely GC-rich repetitive region at the 5′ end of its coding region (nt 80–260). (b) The deduced amino acid sequence of this region shows a glycine- and proline-rich repeat sequence. (c) RT-PCR analysis of RNA samples from mid-exponential (20 h, lane 2), late-exponential (42 h, lane 4) and early stationary phase (66 h, lane 6) shows varying levels of expression of the V3-7 gene product. Negative controls are in lanes 3, 5 and 7. (d) Southern analysis of the *Orpinomyces* sp. genomic DNA shows the V3-7 gene to be present as a single copy.

but was very much reduced by 66 h (Fig. 3c) despite no obvious degradation in the starting RNA used for amplification (data not shown). Several unsuccessful attempts were made to clone a full-length cDNA from the 42 h RNA sample using 5′ RACE. The presence of this gene in the genome of another member of the gut fungal family was confirmed by PCR on three separate *Neocallimastix* samples. The sequence of the 500 bp fragment amplified from the *Neocallimastix* isolates was identical to the *Orpinomyces* sequence at the nucleotide level. Finally, Southern hybridization analysis revealed that this gene was carried as a single copy in the *Orpinomyces* sp. genome (Fig. 3d).

The best-known glycine/proline-rich structural proteins, the collagens, are found in the extracellular matrices of mammalian tissues and contain a tandem GXY repeat, where X and Y are often proline. Individual collagen chains form homo- or heterodimers which self-aggregate to produce long chains (Kielty *et al.*, 1993). Lampirin is an important matrix protein in the annular cartilage of lamprey (an eel-like fish), and contains a GGLGY repeating motif (Bochicchio *et al.*, 2001). There is also evidence for GP-rich structural proteins in the smut fungus *Microbotryum violaceum*. Internal and N-terminal amino acid sequencing of the protein component of fimbriae from the surface of this fungus revealed that this protein is at least in part composed of a collagen-like triplet repeat, with a glycine residue in every third position (Celerin *et al.*, 1996). The fimbriae are essential for surface interactions of *Microbotryum violaceum* during mating.

The predicted amino acid sequence encoded by the gene fragment in clone V3-7 did not have a typical collagen-like triplet motif. However, this sequence did contain an alternating 3-residue–5-residue repeating motif with a glycine residue in the first position of each of the 3 and 5 repeats (Fig. 3b) suggestive of a structural role for the hypothetical protein. The growth-associated expression of the gene mRNA together with its conservation in *Neocallimastix* suggests that it may have an important function in the fungus and its unusual amino acid sequence may be of keen interest to structural biologists.

Non-coding DNA

Fig. 1 demonstrated the overall AT-richness of the *Orpinomyces* genomic DNA, with the AT-rich cutter *Vsp*I giving small DNA fragments after restriction compared with

Downloaded from www.microbiologyresearch.org by
IP: 93.91.26.109
On: Fri, 27 Nov 2015 19:02:49

a more normal range of restriction fragments achieved by digestion with *Eco*RI. This observation suggests that to achieve an overall AT content exceeding 80 % as reported for these fungi, a substantial proportion of the genome is made up of non-coding AT-rich DNA.

Analysis of the non-coding segments of the *Orpinomyces* plasmid library clones sequenced has revealed these sequences to have an AT content of 81·3 mol%. This calculation only considered a single copy of the ribosomal genes in the calculation as it is unclear how many repeats there are in the genome. This value is consistent with previous cloning and physical studies (Billon-Grand *et al.*, 1991; Brownlee, 1989; Durand *et al.*, 1995; Steenbakkers *et al.*, 2002).

The prevalence of longer AT-rich fragments in the libraries was not as great as we had expected, although it is likely that clones containing long stretches of AT-rich DNA have been selected against during the cloning process because of their instability in the bacterial host. The sequencing of the similarly AT-rich *Plasmodium falciparum* genome has been affected by cloning bias and strategies had to be developed to provide adequate coverage over the problematic extremely AT-rich regions of the genome (Gardner *et al.*, 2002a, b; Hall *et al.*, 2002).

The same pattern of elevated AT content in the non-coding DNA intergenic regions and in introns has been observed in the AT-rich genomes of both *Plasmodium falciparum* and *Dictyostelium discoideum*, with just under a doubling (1·75/1·95-fold increase) in GC content of the genic regions compared with non-coding areas of DNA (Gardner *et al.*, 2002a; and see http://dictygenome.bcm.tmc.edu). Introns in filamentous fungi tend to be more AT-rich than coding regions but the difference is much less pronounced (see http://www.broad.mit.edu/annotation/fungi). The sequences from the anaerobic gut fungi in this and other studies, although from a much smaller dataset, suggest an approximately 1·8-fold increase in GC content in exons.

Promoter sequence

Based on nucleotide sequence similarity, clone E2-98 encodes approximately 90 % of the *Piromyces* malate dehydrogenase gene (Akhmanova *et al.*, 1998) and also contains 805 bp of sequence upstream of the putative start codon (Table 2). To date, the only promoter sequence characterized from the anaerobic fungi is that of the enolase gene from *N. frontalis* (Fischer *et al.*, 1995; Akhmanova *et al.*, 1998). Harhangi *et al.* (2003a) have described CAAT and TATA boxes in upstream regions from *Piromyces* hemicellulase genes in agreement with eukaryotic rules. Examination of the putative *Orpinomyces* malate dehydrogenase promoter showed a possible TATA box at −69 to −64 and a CAAT box at −154 to −151 relative to the translation start codon. Unlike the enolase promoter sequence reported by Fischer *et al.* (1995), no regulatory motifs, characteristically found upstream of genes encoding enzymes of the glycolytic and gluconeogenic pathways of filamentous fungi, were observed.

Intronic sequences

Seven probable introns were identified from four clones in this study; the AT content, size, positioning and intron boundary sequences of these are shown in Table 3 together with the only other introns previously described from this family (Durand *et al.*, 1995; Steenbakkers *et al.*, 2002). The probable introns have not been verified by cDNA sequence but in all cases the intron position ensures continued similarity to the comparable protein, which in six of the seven instances is well characterized. The extreme bias of AT content shown by the anaerobic gut fungi also facilitates finding putative intron boundaries as the introns described here and by Durand *et al.* (1995) and Steenbakkers *et al.* (2002) all show extremely high AT content (range 82–96 mol%, Table 3).

The size of intron observed varies from 64 bp in clone E2-35 to 331 bp in the enolase gene, and all begin with the trinucleotide sequence GTA and end in either TAG or AAG, conforming with the GT...AG consensus for eukaryotic intron boundaries. This size range is consistent with data from other filamentous fungi, with mean intron sizes across entire genomes of 135, 143, 93 and 101 for *Neurospora crassa*, *Magnaporthe grisea*, *Fusarium graminearum* and *Aspergillus nidulans* respectively (see http://www.broad.mit.edu/annotation/fungi).

The introns identified modify the consensus for exon–intron boundaries and lariat sequences proposed for the gut fungal genome by Steenbakkers *et al.* (2002). It seems likely that there is considerably more variability in the size of introns, and hence the distances between the lariat branch site and the 5′ and 3′ boundaries, as is also seen for other fungi (http://www.broad.mit.edu/annotation/fungi). There also appears to be greater variability in the sequences observed at the 5′ and 3′ boundaries, reducing the consensus at these sites to three residues from the eight suggested. The lariat consensus is also considerably broader but still consistent with other fungi (Bon *et al.*, 2003; Deutsch & Long, 1999). We suggest the modified consensus below for introns in the gut fungal family:

$^{5'}$GTA.....(A/T)(A/G/T)(C/T)TA$\underline{A}$(C/T).....(T/A)AG$^{3'}$

where the adenosine residue thought likely to be involved in lariat formation is underlined.

Ribosomal sequences

Eleven of the clones contain rDNA, representing three non-identical fragments of the ribosomal repeat region (as illustrated in Fig. 4). A 5 kb clone from the *Vsp*I library (V5-1) extends from the 18S to the 28S subunit sequence. Five separate 1·9 kbp clones from the *Eco*RI library (clones E2 and E4, Fig. 4) extend from the external transcribed spacer (ETS) into the 18S gene sequence and overlap the

Downloaded from www.microbiologyresearch.org by
IP: 93.91.26.109
On: Fri, 27 Nov 2015 19:02:49

Table 3. Comparison of characteristics of putative introns with published intron sequences from anaerobic gut fungi

Some introns had several lariat possibilities (lower-case letters), which conformed to the consensus sequence (A/T)(A/G/T)(T/C)TA**A**(T/C) with the bold **A** thought to be involved in lariat formation. The size of the sequences between the lariat and 6 bp intron boundary is given with the corresponding guanine nucleotide content of that sequence.

Clone	Protein encoded		Intron start	Lariat sequence	Intron stop	Intron size (bp)	AT content (mol%)
	N. frontalis enolase	*Possibility 1*	**GTA**AGT	tatta**a**t	AAA**TAG**	331	86
		Possibility 2		tatta**a**t			
	Piromyces sp. E2 cel9A	Intron 1	**GTA**AGT	TATTA**A**T	TAA**TAG**	127	89
		Intron 2	**GTA**AGT	TATTA**A**T	AAA**TAG**	93	85
		Intron 3	**GTA**AAT	TATTA**A**C	ATA**TAG**	90	87
		Intron 4	**GTA**AAG	TGTTA**A**T	AAA**TAG**	113	86
E2-98	Malate dehydrogenase	Intron 1	**GTA**ATT	TTCTA**A**T	TTA**AAG**	99	88
		Intron 2	**GTA**AAT	TATTA**A**T	TTA**AAG**	85	94
		Intron 3	**GTA**AGT	TGTTA**A**T	GTT**TAG**	73	82
		Intron 4	**GTA**TTT	ATTTA**A**T	ATA**TAG**	102	96
E2-34	UDP-glucose phosphotransferase	*Possibility 1*	**GTA**TGT	ttta**a**t	TTT**TAG**	141	94
		Possibility 2		ttta**a**t			
		Possibility 3		ttta**a**t			
E2-35	UDP-GlcNAc transporter	*Possibility 1*	**GTA**ATA	ttta**a**t	TAT**TAG**	64	94
		Possibility 2		tacta**a**t			
E2-6	Hypothetical gene	Intron 1	**GTA**CAA	TGTTA**A**C	TTA**AAG**	105	92
		Intron 2	**GTA**TAT	TGTTA**A**T	TAT**TAG**	163	91
V3-7	Possible structural protein		–	–ATTA**A**T	ATA**TAG**	>46	89

whole length of five shorter (1 kbp) *Vsp*I clones (clones V2, V3 and V4, Fig. 4). The cloned part of the non-functional ETS sequences have 83 mol% AT content compared with 58 mol% AT for the functional rRNA gene sequences. The sequences illustrated by the E and V clones in Fig. 4 appear to be polymorphs of the ribosomal repeat. The area of overlapping sequence shows 52 % identity between the two types over 100 nt, whereas each type shows 95 % identity for the E clones and 99 % identity for the V clones, and is therefore unlikely to be as a result of any sequencing errors for these clones.

The variation in ribosomal repeat sequences suggests a contradiction of the concerted evolution dogma proposed by van Nues *et al.* (1995). However, the presence of two or more differing rDNA copies has been observed in other fungi to date, e.g. O'Donnell & Cigelnik (1997) found two ITS2 types in several *Fusarium* isolates; this appears to be the result of interspecific hybridization. Several ITS1 types were observed in the basidiomycete *Trichaptum abietinum* (Ko & Jung, 2002) and we have found multiple types of ITS1 sequences in *Orpinomyces* and other gut fungal genera (Brookman *et al.*, 2000; Nicholson, 2003).

Retrotransposons

Retrotransposons were found to be highly represented (24/101) in the sequences cloned, although the differences in these sequences suggest this is not a simple cloning artefact. The long terminal repeat (LTR)-containing gypsy- and copia-type elements were found together with sequences showing homology to non-LTR retrotransposons such as TRE3-A from the AT-rich genome of the slime mould *Dictyostelium discoideum*.

Sixteen clones showed similarity to MAGGY, a gypsy-like retrotransposon from the rice blast fungus *Magnaporthe grisea* (Farman *et al.*, 1996) or a similar retrotransposon from the puffer fish *Fugu rubripes* (Poulter & Butler, 1998). Conserved functional domains, such as the YLDDI reverse transcriptase motif and DAS RNase H motif previously identified in gypsy-like elements were present in these sequences (Fig. 5), although seven of the clones did not appear to be intact as they contained mutations causing frameshifts and/or stop codons within the sequence or in the case of clone E2-84 a deletion of approximately 900 bp. The GAG translation start site was identified for six clones and for three of these there was sufficient coverage to identify repeats within the LTR region. Clones E2-30 and E2-50 contain almost identical nucleotide sequences except that a 17mer (TTTCTTGGTGAAACTCT) repeated 3 times in the LTR of E2-30 is repeated 15 times in the LTR of E2-50. In clone E2-101 a different sequence (TACTCTTTTGGAT) is repeated 20 times. These differences between the LTR repeat sequences suggest that there may be more than one lineage for this MAGGY-like retrotransposon.

The prevalence of stop codons in the reading frames of

Downloaded from www.microbiologyresearch.org by
IP: 93.91.26.109
On: Fri, 27 Nov 2015 19:02:49

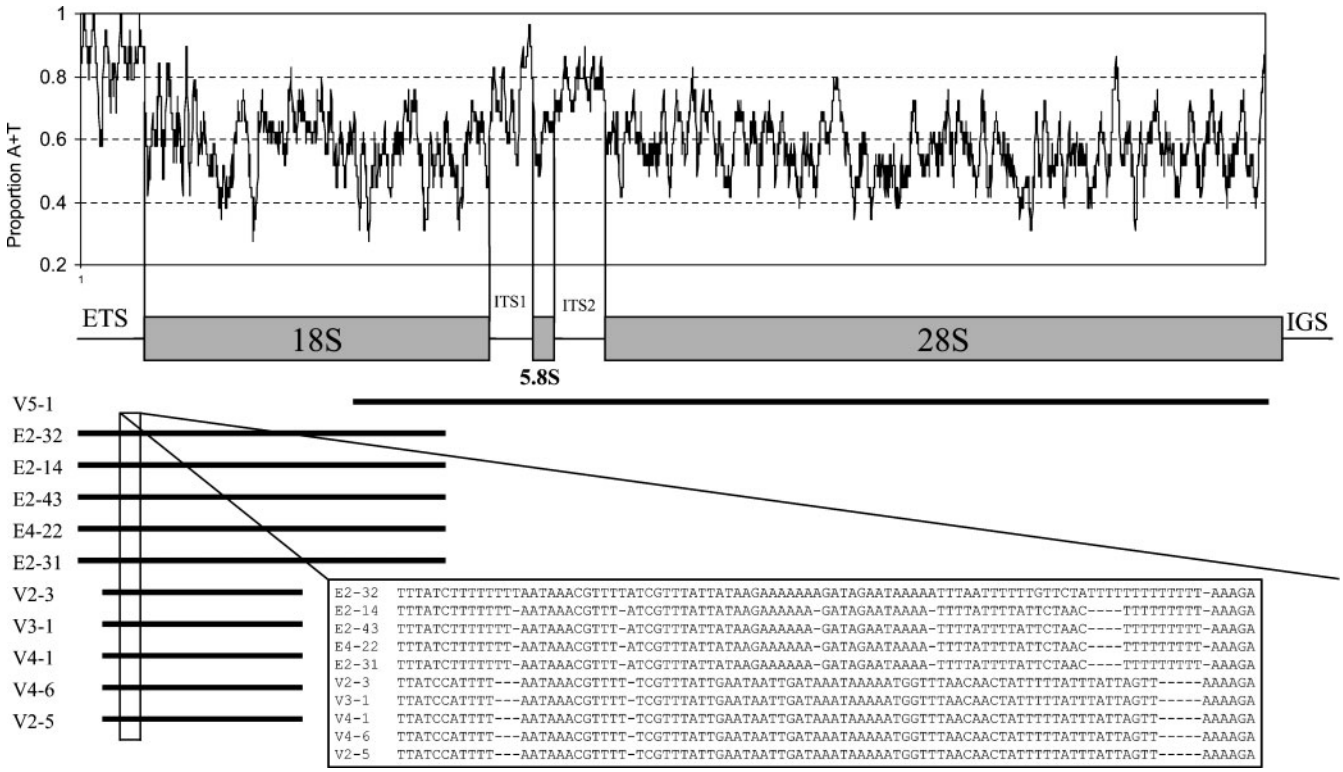

Fig. 4. Alignment of clones encoding rDNA. Eleven clones cover approximately 6·2 kbp of the ribosomal complex from 350 nt upstream of the 18S gene to approximately 140 nt before the 3′ end of the 28S gene. The proportion of A+T nucleotides is plotted against nucleotide position, illustrating the increased proportion of A and T nucleotides in the external transcribed spacer (ETS) and internal transcribed spacers (ITS1 and ITS2) compared with the ribosomal genes. The *Vsp*I clones V2-3, V3-1, V4-1, V4-6 and V2-5 are >99 % identical to one another whereas the *Eco*RI clones are >95 % identical to one another. However, the two clone sets are distinct, with only 78 % identity over the 1·4 kbp of overlapping sequence. The most variable region is shown (boxed); this ~100 bp of sequence immediately upstream of the 18S gene shows only 52 % similarity between the clone sets.

the gypsy-like sequences and the high level of divergence suggests that there may be some kind of host defence mechanism causing mutations within these sequences. A similar observation was noted in the genome of *M. grisea*. In addition to the many copies of MAGGY there was also found a degenerate inactive form of the sequence. Analysis of the degenerate MAGGY sequence revealed a high level of G : C to A : T transitions, thought to be the result of a repeat induced point mutation (RIP) host defence mechanism (Nakayashiki *et al.*, 1999).

The slime mould *Dictyostelium discoideum* has an AT-rich genome that contains considerable numbers of transposable elements, mainly LTR- and non-LTR-retrotransposons. These elements often contain AT-rich sequences and the genome sequence suggests that many copies may be defective, e.g. the DIRS-1 element is present at 200 copies per genome but most copies appear to be defective due to recombination events (see review by Winckler *et al.*, 2002). Conversely the other example of a sequenced AT-rich genome, *Plasmodium falciparum*, showed no transposable elements whatsoever (Gardner *et al.*, 2002a).

Conclusion

Considerable advances have been made over recent years in our understanding of the physical characteristics of individual genomes. The advent of entire sequenced genomes from a variety of prokaryotic and eukaryotic organisms has required sophisticated methods of sequence and comparative genomic analysis. These can been used in gene finding (Fickett & Tung, 1992) and to identify generic rules that may govern genome structure, function and expression; e.g. relationships have been observed between codon usage and gene length (Moriyama & Powell, 1998), intron length (Vinogradov, 2001), intragenic position, level of expression (Bulmer, 1988), and differences in codon usage at conserved versus non-conserved amino acid positions (Diaz-Lazcoz *et al.*, 1995). The data presented here have given a brief glimpse of some of the rules followed by these organisms and enabled consideration of how well from a genomic perspective these unusual fungi, with their atypical lifestyle and nucleotide composition, fit in to the wider eukaryotic kingdom. Furthermore, the completion of the genome sequence of the malaria parasite, *Plasmodium*

Downloaded from www.microbiologyresearch.org by
IP: 93.91.26.109
On: Fri, 27 Nov 2015 19:02:49

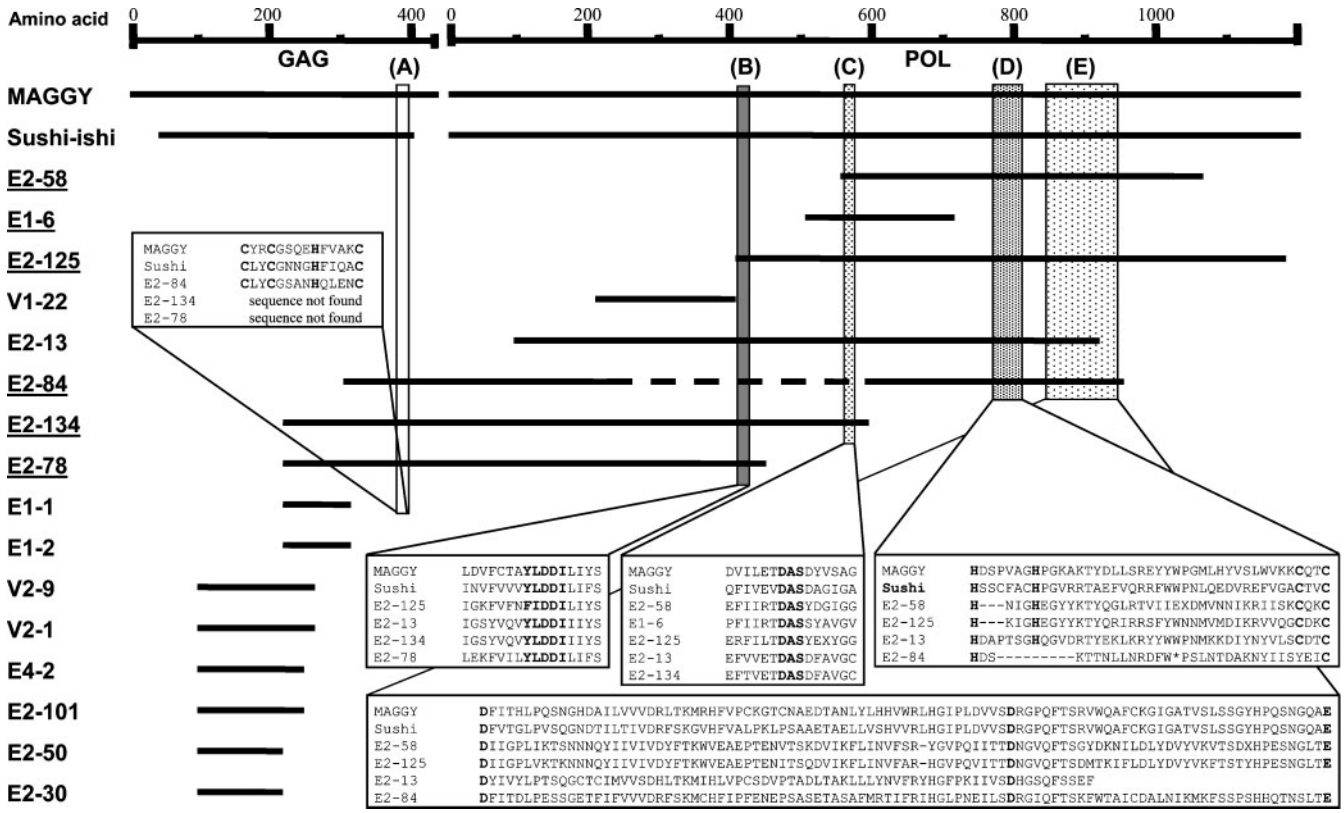

Fig. 5. Alignment of gypsy-like clones from the *Orpinomyces* sp. genome. MAGGY, a retrotransposon from the rice blast fungus *Magnaporthe grisea*, and a similar retrotransposon (sushi-ishi) from the puffer fish *Fugu rubripes* are shown as reference sequences. Whilst ten of the gypsy-like clones appear to be functionally intact, the other six contain frameshifts and/or stop codons interrupting their reading frames. Clone E2-84 contains a large deletion of approximately 1 kbp (broken line). Highly conserved domains, characteristic of gypsy elements, are shown: (A) C-C-H-C RNA binding motif. (B) YLDDI reverse transcriptase motif. (C) RNaseH DAS motif. (D) H-H-C-C integrase DNA binding motif. (E) D-D-E integrase motif.

falciparum, has removed the major structural hurdles to a successful genome sequence programme for this biologically interesting and agronomically important group of fungi.

ACKNOWLEDGEMENTS

Matthew Nicholson was funded as a BBSRC CASE student with IGER, Aberystwyth. We are grateful to Professor Tony Trinci, Dr Geoff Robson, Dr Danny Tuckwell and Professor Geoff Hazlewood for useful discussions during this work. The authors are grateful to Zaneta Park-Ng for assistance with the statistical analyses.

REFERENCES

Akhmanova, A., Voncken, F. G., Harhangi, H., Hosea, K. M., Vogels, G. D. & Hackstein, J. H. (1998). Cytosolic enzymes with a mitochondrial ancestry from the anaerobic chytrid *Piromyces* sp. E2. *Mol Microbiol* **30**, 1017–1027.

Akhmanova, A., Voncken, F. G., Hosea, K. M., Harhangi, H., Keltjens, J. T., Op den Camp, H. J., Vogels, G. D. & Hackstein, J. H. (1999). A hydrogenosome with pyruvate formate-lyase, anaerobic chytrid fungi use an alternative route for pyruvate catabolism. *Mol Microbiol* **32**, 1103–1114.

Andersson, J. O. & Roger, A. J. (2002). A cyanobacterial gene in nonphotosynthetic protists – an early chloroplast acquisition in eukaryotes? *Curr Biol* **12**, 115–119.

Bayer, E. A., Shimon, L. J. W., Shoham, Y. & Lamed, R. (1998). Cellulosomes – structure and ultrastructure. *J Struct Biol* **124**, 221–234.

Billon-Grand, G., Fiol, J. B., Breton, A., Bruyere, A. & Oulhaj, Z. (1991). DNA of some anaerobic rumen fungi, G+C content determination. *FEMS Microbiol Lett* **66**, 267–270.

Black, G. W., Hazlewood, G. P., Xue, G. P., Orpin, C. G. & Gilbert, H. J. (1994). Xylanase B from *Neocallimastix patriciarum* contains a non-catalytic 455-residue linker sequence comprised of 57 repeats of an octapeptide. *Biochem J* **299**, 381–387.

Bochicchio, B., Pepe, A. & Tamburro, A. M. (2001). On (GGLGY) synthetic repeating sequences of lamprin and analogous sequences. *Matrix Biol* **20**, 243–250.

Bon, E., Casaregola, S., Blandin, G. & 8 other authors (2003). Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns. *Nucleic Acids Res* **31**, 1121–1135.

Brookman, J. L., Mennim, G., Trinci, A. P., Theodorou, M. K. & Tuckwell, D. S. (2000). Identification and characterization of

Downloaded from www.microbiologyresearch.org by
IP: 93.91.26.109
On: Fri, 27 Nov 2015 19:02:49

anaerobic gut fungi using molecular methodologies based on ribosomal ITS1 and 18S rRNA. *Microbiology* **146**, 393–403.

Brownlee, A. G. (1989). Remarkably AT-rich genomic DNA from the anaerobic fungus *Neocallimastix*. *Nucleic Acids Res* **17**, 1327–1335.

Bulmer, M. (1988). Codon usage and intragenic position. *J Theor Biol* **133**, 67–71.

Celerin, M. J. M., Ray, J. M., Schisler, N. J., Day, A. W., Stetler-Stevenson, W. G. & Laudenbach, D. E. (1996). Fungal fimbriae are composed of collagen. *EMBO J* **15**, 4445–4453.

Deutsch, M. & Long, M. (1999). Intron-exon structure of eukaryotic model organisms. *Nucleic Acids Res* **27**, 3219–3228.

Diaz-Lazcoz, Y., Henaut, A., Vigier, P. & Risler, J. L. (1995). Differential codon usage for conserved amino acids, evidence that the serine codons TCN were primordial. *J Mol Biol* **250**, 123–127.

Durand, R., Fischer, M., Rascle, C. & Fevre, M. (1995). *Neocallimastix frontalis* enolase gene, *enol* – first report of an intron in an anaerobic fungus. *Microbiology* **141**, 1301–1308.

Embley, T. M., van der Giezen, M., Horner, D. S., Dyal, P. L. & Foster, P. (2003). Mitochondria and hydrogenosomes are two forms of the same fundamental organelle. *Philos Trans Roy Soc Lond B Biol Sci* **358**, 191–203.

Farman, M. L., Tosa, Y., Nitta, N. & Leong, S. A. (1996). MAGGY, a retrotransposon in the genome of the rice blast fungus *Magnaporthe grisea*. *Mol Gen Genet* **251**, 665–674.

Fickett, J. W. & Tung, C. S. (1992). Assessment of protein coding measures. *Nucleic Acids Res* **20**, 6441–6450.

Fischer, M., Durand, R. & Fevre, M. (1995). Characterization of the promoter region of the enolase encoding gene *enol* from the anaerobic fungus *Neocallimastix frontalis* – sequence and promoter analysis. *Curr Genet* **28**, 80–86.

Fontes, C. M., Hazlewood, G. P., Morag, E., Hall, J., Hirst, B. H. & Gilbert, H. J. (1995). Evidence for a general role for non-catalytic thermostabilizing domains in xylanases from thermophilic bacteria. *Biochem J* **307**, 151–158.

Garcia-Vallvé, S., Romeu, A. & Palau, J. (2000). Horizontal gene transfer of glycosyl hydrolases of the rumen fungi. *Mol Biol Evol* **17**, 352–361.

Gardner, M. J. (2001). A status report on the sequencing and annotation of the *P. falciparum* genome. *Mol Biochem Parasitol* **118**, 133–138.

Gardner, M. J., Hall, N., Fing, E. & 42 other authors (2002a). Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498–511.

Gardner, M. J., Shallom, S. J., Carlton, J. M. & 34 other authors (2002b). Sequence of *Plasmodium falciparum* chromosomes 2, 10, 11 and 14. *Nature* **419**, 531–534.

Grocock, R. J. & Sharp, P. M. (2002). Synonymous codon usage in *Pseudomonas aeruginosa* PA01. *Gene* **289**, 131–139.

Hall, N., Pain, A., Berriman, M. & 77 other authors (2002). Sequence of *Plasmodium falciparum* chromosomes 1, 3–9 and 13. *Nature* **419**, 527–531.

Harhangi, H. R., Akhmanova, A., Steenbakkers, P. J. M., Jetten, M. S. M., van der Drift, C. & Op den Camp, H. J. M. (2003a). Genomic DNA analysis of genes encoding (hemi-)cellulolytic enzymes of the anaerobic fungus *Piromyces* sp. E2. *Gene* **314**, 73–80.

Harhangi, H. R., Freelove, A. C., Ubhayasekera, W. & 8 other authors (2003b). Cel6A, a major exoglucanase from the cellulosome of the anaerobic fungi *Piromyces* sp. E2 and *Piromyces equi*. *Biochim Biophys Acta* **1628**, 30–39.

Harhangi, H. R., Akhmanova, A., Emmens, R. & 6 other authors (2003c). Xylose metabolism in the anaerobic fungus *Piromyces* sp. strain E2 follows the bacterial pathway. *Arch Microbiol* **180**, 134–141.

Henrich, B., Monnerjahn, U. & Plapp, R. (1990). Peptidase D gene (*pepD*) of *Escherichia coli* K-12, nucleotide sequence, transcript mapping, and comparison with other peptidase genes. *J Bacteriol* **172**, 4641–4651.

Kielty, C. M., Hopkinson, I. & Grant, M. E. (1993). Collagen. In *Connective Tissue and its Heritable Disorders*, pp. 103–147. Edited by P. M. Royce & B. Steinmann. New York: Wiley-Liss.

Ko, K. S. & Jung, H. S. (2002). Three nonorthologous ITS1 types are present in a polypore fungus *Trichaptum abietinum*. *Mol Phylogenet Evol* **23**, 112–122.

Kocherginskaya, S. A., Aminov, R. I. & White, B. A. (2001). Analysis of the rumen bacterial diversity under two different diet conditions using denaturing gradient gel electrophoresis, random sequencing, and statistical ecology approaches. *Anaerobe* **7**, 119–134.

Lowe, S. E., Theodorou, M. K., Trinci, A. P. J. & Hespell, R. B. (1985). Growth of anaerobic rumen fungi on defined and semi-defined media lacking rumen fluid. *J Gen Microbiol* **131**, 2225–2229.

Moriyama, E. N. & Powell, J. R. (1998). Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. *Nucleic Acids Res* **26**, 3188–3193.

Munn, E. A., Orpin, C. G. & Greenwood, C. A. (1988). The ultrastructure and possible relationships of four obligate anaerobic chytridiomycete fungi from the rumen of sheep. *Biosystems* **22**, 67–81.

Nakayashiki, H., Nishimoto, N., Ikeda, K., Tosa, Y. & Mayama, S. (1999). Degenerate MAGGY elements in a subgroup of *Pyricularia grisea*, a possible example of successful capture of a genetic invader by a fungal genome. *Mol Gen Genet* **261**, 958–966.

Nicholson, M. J. (2003). *Molecular characterisation of anaerobic gut fungi and their colonisation of plant material in the rumen.* PhD thesis, Faculty of Science & Engineering, University of Manchester, UK.

Nixon, J. E., Wang, A., Morrison, H. G., McArthur, A. G., Sogin, M. L., Loftus, B. J. & Samuelson, J. (2002). A spliceosomal intron in *Giardia lamblia*. *Proc Natl Acad Sci U S A* **99**, 3701–3705.

O'Donnell, K. & Cigelnik, E. (1997). Two divergent intragenomic rDNA ITS2 types within a monophyletic lineage of the fungus *Fusarium* are nonorthologous. *Mol Phylog Evol* **7**, 103–116.

Orpin, C. G. (1975). Studies on the rumen flagellate *Neocallimastix frontalis*. *J Gen Microbiol* **98**, 423–430.

Ozkose, E., Thomas, B. J., Davies, D. R., Griffith, G. W. & Theodorou, M. K. (2001). *Cyllamyces aberensis* gen. nov. sp. nov., a new anaerobic gut fungus with branched sporangiophores isolated from cattle. *Can J Bot* **79**, 666–673.

Poulter, R. & Butler, M. (1998). A retrotransposon family from the pufferfish (fugu) *Fugu rubripes*. *Gene* **215**, 241–249.

Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning: a Laboratory Manual*, 2nd edn. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.

Steenbakkers, P. J. M., Li, X.-L., Ximenes, E. A., Arts, J. G., Chen, H., Ljungdahl, L. G. & Op den Camp, H. J. M. (2001). Noncatalytic docking domains of cellulosomes of anaerobic fungi. *J Bacteriol* **183**, 5325–5333.

Steenbakkers, P. J., Ubhayasekera, W., Goossen, H. J., van Lierop, E. M., van der Drift, C., Vogels, G. D., Mowbray, S. L. & Op den Camp, H. J. (2002). An intron-containing glycoside hydrolase family 9 cellulase gene encodes the dominant 90 kDa component of the cellulosome of the anaerobic fungus *Piromyces* sp. strain E2. *Biochem J* **365**, 193–204.

Downloaded from www.microbiologyresearch.org by
IP: 93.91.26.109
On: Fri, 27 Nov 2015 19:02:49

Theodorou, M. K., Mennim, G., Davies, D., Zhu, W.-Y., Trinci, A. P. J. & Brookman, J. (1996). Anaerobic fungi in the digestive tract of mammalian herbivores and their potential for exploitation. *Proc Nutrition Society* **55**, 913–926.

van der Giezen, M., Rechinger, K. B., Svendson, I., Durand, R., Hirt, R. P., Fèvre, M., Embley, T. M. & Prins, R. A. (1997). A mitochondrial-like targeting signal on the hydrogenosomal malic enzyme from the anaerobic fungus *Neocallimastix frontalis*: support for the hypothesis that hydrogenosomes are modified mitochondria. *Mol Microbiol* **23**, 11–21.

van Nues, R. W., Venema, J., Rientjes, J. M., Dirks-Mulder, A. & Raué, H. A. (1995). Processing of eukaryotic pre-rRNA: the role of the transcribed spacers. *Biochem Cell Biol* **73**, 789–801.

Vinogradov, A. E. (2001). Intron length and codon usage. *J Mol Evol* **52**, 2–5.

Voncken, F. G. J., Boxma, B., van Hoek, A. H. A. M., Akhmanova, A. S., Vogels, G. D., Huynene, M., Veenhuis, M. & Hackstein, J. H. P. (2002). A hydrogenosomal [Fe]-hydrogenase from the anaerobic chytrid *Neocallimastix* sp. L2. *Gene* **284**, 103–112.

Winckler, T., Dingermann, T. & Glöckner, G. (2002). *Dictyostelium* mobile elements – strategies to amplify in a compact genome. *Cell Mol Life Sci* **59**, 2097–2111.

Zhou, L., Xue, G. P., Orpin, C. G., Black, G. W., Gilbert, H. J. & Hazlewood, G. P. (1994). Intronless *celB* from the anaerobic fungus *Neocallimastix patriciarum* encodes a modular family A endoglucanase. *Biochem J* **297**, 359–364.

Downloaded from www.microbiologyresearch.org by
IP: 93.91.26.109
On: Fri, 27 Nov 2015 19:02:49