

## GENOME WATCH

### A feast of protozoan genomes

Christiane Hertz-Fowler, Matthew Berriman and Arnab Pain



The recent completion of five new protozoan parasite genomes should lead to an improved understanding of their biology and, through comparative genomics, shed light on protozoan genome diversity and evolution.

*Entamoeba histolytica* is a human gut parasite that causes amoebiasis, a significant health problem in developing countries. The 23.7-Mb draft *E. histolytica* genome contains 9,938 predicted genes and a striking abundance of tandemly repeated transfer-RNA-containing arrays<sup>1</sup>.

The genomes of *E. histolytica* and the amitochondrial protist pathogens *Giardia lamblia* and *Trichomonas vaginalis* share several metabolic adaptations. These include reduced or eliminated mitochondrial metabolic pathways. Indeed, the genome data are consistent with the lack of a mitochondrial genome. Tricarboxylic-acid-cycle and mitochondrial-electron-transport-chain enzymes are lacking. Secondary gene loss and lateral gene transfer, mainly from prokaryotes, seem to have shaped *E. histolytica* metabolism. This parasite has several bacterial-like fermentation enzymes. More than half of the 96 genes identified as being acquired by lateral gene transfer encode metabolic enzymes for carbohydrate and amino-acid metabolism.

In the human gut, *E. histolytica* can access many bacterial and host-derived metabolites and has lost amino-acid biosynthetic pathways, only retaining those for serine and cysteine. Purine, pyrimidine and thymidylate *de novo* synthesis pathways are absent, and *E. histolytica* must rely on salvage pathways. This parasite cannot synthesize fatty acids, folate-dependent enzymes and folate

transporters, isoprenoids or sphingolipid headgroup aminoethylphosphonate, but can produce phospholipids.

In contrast to its metabolic deficits, *E. histolytica* has the most varied set of signal-transduction-related proteins described in a single-celled eukaryote, including 270 protein kinases, which represent all seven families of the eukaryotic protein-kinase superfamily. There is also redundancy in virulence-factor genes, including the multi-subunit GalGalNAc lectins, cysteine proteases and pore-forming peptides (amoebapores).

Extensive horizontal gene transfer has also been noted following the recent genome sequencing of the trypanosomatid parasites *Leishmania major*<sup>2</sup>, *Trypanosoma brucei*<sup>3</sup> and *Trypanosoma cruzi*<sup>4</sup>, all of which are important human and animal pathogens.

Despite using different insect vectors, causing different pathologies and employing diverse immune-evasion strategies, the genomes share several features, although genome characteristics vary (TABLE 1). Of approximately 8,000–12,000 genes, 6,200 are common to all three parasites and are found in polycistronic clusters, which might reflect the link of transcription and subsequent RNA processing in these kinetoplasts. These parasites have a radically different transcription machinery to higher eukaryotes, with very few transcription factors, consistent with a reliance on post-transcriptional gene regulation<sup>2</sup>. A large number of RNA-binding proteins are also present. A mechanism of gene duplication and amplification seems to increase gene expression. Diversification of gene families is prominent in the *T. brucei* and *T. cruzi* genomes, in which 10–18% of the genomes

comprise multigene families of surface proteins. These families are usually located either at the chromosome ends or in the chromosome-internal regions, although this varies among the parasite species<sup>3</sup>. While *T. brucei* encodes the variant surface glycoprotein gene (VSG) family, *T. cruzi* has several surface-protein families, including a new family with at least 1,300 members. In both organisms these families are predominantly pseudogenes. In *T. brucei* these pseudogenes might form an information pool that contributes to the diversity of the antigen repertoire instead of simply representing defunct genes<sup>3</sup>. Both species also contain retroelements that increase recombination frequency and contribute to genome plasticity but have distinct mechanisms for retroelement silencing<sup>4</sup>. The conservation of synteny is consistent with the existence of a common ancestor that contained a fragmented genome, with the subsequent divergence of the *Leishmania* genus and the monophyletic *Trypanosoma* genus. Despite reports of plant-like sequences in the trypanosomatid genomes<sup>5</sup>, it is thought unlikely that the common ancestor harboured an endosymbiont<sup>6</sup>.

Breakpoints in regions of otherwise conserved synteny between the three genomes seem to be preferentially associated with both strand-switch regions and with expansions of multigene families, retroelements and structural RNAs. Also, some species-specific genes are inserted at these synteny breaks, such as the subunits of the *T. brucei* heterodimeric transferrin receptor, which is required for host adaptation<sup>6</sup>.

In addition to identifying the 'core' proteome, El-Sayed *et al.*<sup>6</sup> showed that more

Table 1 | General features of three Kinetoplastid parasite genomes

Feature	<i>Leishmania major</i>	<i>Trypanosoma brucei</i>	<i>Trypanosoma cruzi</i> *
Disease	Leishmaniasis	Human African trypanosomiasis (sleeping sickness), Ngana in cattle	Chagas disease
Ploidy	Diploid (polyploid for some chromosomes)	Diploid	Diploid
Haploid genome size (Mb)	33	26	55
Chromosomes	36	11 <sup>†</sup>	~28 <sup>§</sup>
Genes	8,272	9,068	~12,000
Pseudogenes	39	904	3, 590 <sup>  </sup>
Non-coding RNA genes	911	>556	1,994 <sup>  </sup>

\*Draft genome only. †Excludes an unspecified number of intermediate and mini chromosomes. ‡Exact number is not known. §The *T. cruzi* genome sequenced is a hybrid with significant allelic variation. Data from both haplotypes are included.

orthologues were shared between the intracellular parasites *L. major* and *T. cruzi*, whereas the two *Trypanosoma* genomes have more species-specific genes, which mostly seem to encode members of surface-antigen families. Examination of species-specific proteins and domains can shed light on parasite–host interactions. For instance, a tandemly duplicated, putative macrophage-migration inhibitory factor in the *L. major* genome might contribute to parasite survival in macrophages. Likewise, a *T. cruzi*-secreted peptidase might facilitate entry into host cells. Positive selection pressure — estimated from non-synonymous codon changes between orthologues — seems to be greatest for genes of unknown function, which might imply that some of the most rapidly evolving trypanosomatid genes have yet to be characterized.

As for metabolism, *L. major* is the most complex parasite, with pathways present to deal with the nutrient-poor mammalian macrophage environment as well as those to metabolize the sugars present in the nectar diet of the sandfly vector. By contrast, *T. brucei*, an extracellular parasite, obtains nutrients from the host blood or from the midgut of the tsetse fly vector. Evidence of horizontal gene transfer from bacteria, which might increase metabolic versatility, was also found.

Infection with *Theileria* spp. affects livestock development. Recently, the genomes of two closely related tick-borne apicomplexan cattle haemoparasites, *Theileria parva* and *Theileria annulata*, were sequenced<sup>7,8</sup>. Both organisms cause severe lymphoproliferative disorders in cattle — East Coast fever (*T. parva*) and tropical theileriosis (*T. annulata*). Disease progress is linked to the ability of the parasite to induce uncontrolled proliferation of the infected host leukocyte. Both parasites cause high rates of mortality in cattle in disease-endemic regions. Although these cattle diseases have features in common, the parasites

are transmitted by different tick species, invade different host cell types and have different clinical symptoms.

Gardner *et al.*<sup>7</sup> compared the metabolic potential of the nuclear and organellar genome sequences of *T. parva* with those of *Plasmodium falciparum*. In a related report, Pain *et al.*<sup>8</sup> compared the genome of *T. annulata* with *T. parva*. Both nuclear genomes are compact (~8.3 Mbp), with only four haploid chromosomes, and are approximately one third the size of the *P. falciparum* genome. Moreover, *T. parva* and *T. annulata* have a higher gene density (4,035 and 3,792 genes, respectively) and more spliced genes. Despite conserved gene sequences and synteny — 3,265 orthologous gene pairs between the two species — several species-specific genes were identified. Most genes without orthologues are members of unequally expanded gene families with only a small proportion present as single copies (34 in *T. annulata*, 60 in *T. parva*). The chromosomes of *T. annulata* and *T. parva* have syntenic regions with only a few inversions of small blocks and no movement of blocks between chromosomes. Short breaks in synteny correspond to gene insertions or deletions and often involve members of large gene families such as the *Tpr* genes (*T. parva* repeat) and their counterparts in *T. annulata*, the *Tar* (*T. annulata* repeat) genes.

Like many parasitic protozoa, both *Theileria* species have tandem arrays of genus-specific, hypervariable gene families that are located in the subtelomeres and are predicted to encode secreted proteins. The overall arrangement of subtelomeric genes is conserved with one (or more) ABC-transporter gene(s) marking the boundaries between subtelomeric gene families and the house-keeping genes. Many *Theileria*-specific protein family members incorporate one or more copies of a polymorphic FAINT (frequently associated in *Theileria*) domain. Over 900 copies of the FAINT domain are present in ~166 proteins of both genomes.

Like the trypanosomatids, evidence of positive immune selection was found in the macroschizont and merozoite stage-expressed genes. Also, several candidate genes for host-cell transformation have been identified, by assuming that these genes are expressed in macroschizonts and that their products are released into the host cell cytoplasm or expressed on the parasite surface.

Compared with *P. falciparum*, the metabolism of *Theileria* spp. is streamlined. Several biosynthetic pathways are absent, including those for haem, type II fatty acids, polyamines and shikimic acid. *Theileria* spp. have lost the ability to salvage purines and have limited ability to interconvert amino acids, but isoprenoid biosynthesis is, however, present. This reduced metabolism suggests substantial dependence on the host cell for many substrates.

*Christiane Hertz-Fowler, Matthew Berriman and Arnab Pain are at the Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. e-mail: microbes@sanger.ac.uk*

doi:10.1038/nrmicro1237

- Loftus, B. *et al.* The genome of the protist parasite *Entamoeba histolytica*. *Nature* **433**, 865–868 (2005).
- Ivens, A. C. *et al.* The genome of the kinetoplastid parasite, *Leishmania major*. *Science* **309**, 436–442 (2005).
- Berriman, M. *et al.* The genome of the African trypanosome *Trypanosoma brucei*. *Science* **309**, 416–422 (2005).
- El-Sayed, N. M. *et al.* The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. *Science* **309**, 409–415 (2005).
- Hannaert, V. *et al.* Plant-like traits associated with metabolism of *Trypanosoma* parasites. *Proc. Natl Acad. Sci. USA* **100**, 1067–1071 (2003).
- El-Sayed, N. M. *et al.* Comparative genomics of trypanosomatid parasitic protozoa. *Science* **309**, 404–409 (2005).
- Gardner, M. J. *et al.* Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes. *Science* **309**, 134–137 (2005).
- Pain, A. *et al.* Genome of the host-cell transforming parasite *Theileria annulata* compared with *T. parva*. *Science* **309**, 131–133 (2005).

## Online links

### DATABASES

The following terms in this article are linked online to:

Entrez: <http://www.ncbi.nlm.nih.gov/Entrez>

*Entamoeba histolytica* | *Giardia lamblia* | *Trichomonas vaginalis*

| *Leishmania major* | *Trypanosoma brucei* | *Trypanosoma cruzi*

| *Theileria parva* | *Theileria annulata* | *Plasmodium falciparum*

Access to this interactive links box is free online.

**Online links:**

*Entamoeba histolytica*:

[http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj&cmd=Retrieve&dopt=Overview&list\\_uids=9532](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj&cmd=Retrieve&dopt=Overview&list_uids=9532)

*Giardia lamblia*:

[http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj&cmd=Retrieve&dopt=Overview&list\\_uids=9531](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj&cmd=Retrieve&dopt=Overview&list_uids=9531)

*Trichomonas vaginalis*:

[http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj&cmd=Retrieve&dopt=Overview&list\\_uids=12676](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj&cmd=Retrieve&dopt=Overview&list_uids=12676)

*Leishmania major*:

[http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj&cmd=Retrieve&dopt=Overview&list\\_uids=9528](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj&cmd=Retrieve&dopt=Overview&list_uids=9528)

*Trypanosoma brucei*:

[http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj&cmd=Retrieve&dopt=Overview&list\\_uids=9529](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj&cmd=Retrieve&dopt=Overview&list_uids=9529)

*Trypanosoma cruzi*:

[http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj&cmd=Retrieve&dopt=Overview&list\\_uids=9530](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj&cmd=Retrieve&dopt=Overview&list_uids=9530)

*Theileria parva*:

[http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj&cmd=Retrieve&dopt=Overview&list\\_uids=9544](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj&cmd=Retrieve&dopt=Overview&list_uids=9544)

*Theileria annulata*:

[http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj&cmd=Retrieve&dopt=Overview&list\\_uids=9543](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj&cmd=Retrieve&dopt=Overview&list_uids=9543)

*Plasmodium falciparum*:

[http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj&cmd=Retrieve&dopt=Overview&list\\_uids=9538](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj&cmd=Retrieve&dopt=Overview&list_uids=9538)