Published in final edited form as:
Infect Genet Evol. 2008 December ; 8(6): 901–906. doi:10.1016/j.meegid.2008.07.001.

Kinetoplastid genomics: the thin end of the wedge

Nancy R. Sturm[1,‡], **L. L. Isadora Trejo Martinez**[1], and **Sean Thomas**[2]

1*Department of Microbiology, Immunology & Molecular Genetics, David Geffen School of Medicine, University of California at Los Angeles, Los Angeles, CA 90095, USA*

2*Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA*

Abstract

The completion of the genome sequencing projects for major pathogens *Trypanosoma brucei*, *Trypanosoma cruzi* and *Leishmania major* has enabled numerous studies that would have been difficult or impossible to perform otherwise. New technologies in sequencing and protein analyses promise further rapid expansion in our capabilities. The keys to successful use of these new tools are recognizing the power and limitations of studies performed thus far, grasping the unrealized potential of new and developing technologies, and creating access to a multidisciplinary set of skills that will facilitate research, particularly in the bioinformatic analysis of the reams of data that will be forthcoming. In this Discussion, we will provide an overview of kinetoplastid genomics studies with emphasis on studies advanced through genomic data, and a preview of what may come in the near future.

Keywords

bioinformatics; *Leishmania*; *Trypanosoma*; trypanosome

Introduction

Trypanosomes, members of the order *Kinetoplastida*, are characterized by their overabundance of mitochondrial DNA, and are most notorious as the causative agents of African Sleeping Sickness, Chagas disease, and Leishmaniasis. As relatively ancient eukaryotes, the trypanosomes have maintained pathways that have not survived the test of time in most other lineages, while evolving mechanisms that are unique. Notable examples of these peculiarities include the widespread use of *trans*-splicing that permits translation of RNA polymerase I-generated transcripts that are viable mRNAs due to the acquisition of a RNA polymerase II-transcribed spliced leader (SL) sequence (Campbell et al., 2003), the insertion/deletion RNA editing mechanism of kinetoplastid mitochondrion (Stuart et al., 2005), and the elaborate structure of their mitochondrial DNA, composed of thousands of interlinked molecules (Lukeš et al., 2003).

In 2005 three kinetoplastid (a.k.a. ‘Tritryp’) genomes were published, along with the first genome-wide comparison showing extensive conservation of gene order, or synteny (El-Sayed

‡Address correspondence to: Nancy R. Sturm, Department of Microbiology, Immunology & Molecular Genetics, 609 Charles E. Young Drive East, University of California at Los Angeles, California 90095-1489. Tel.: +1 (310) 206-5556; Fax: +1 (310) 206-5231; E-mail: E-mail: nsturm@ucla.edu.

et al., 2005a). The individual reports focused on various biological questions, both specifics of the focus organism and highlighting features conserved throughout the Tritryps. The *Trypanosoma cruzi* analysis discussed the high level of multicopy genes, DNA metabolism pathways, signaling, and surface molecules (El-Sayed et al., 2005b); a complementary proteome analysis was also performed (Atwood et al., 2005). The spotlight centered on antigenic variation, cytoskeletal structures, and various physiological pathways in the *Trypanosoma brucei* presentation (Berriman et al., 2005). The *Leishmania major* report delved into transcription and RNA processing mechanisms (Ivens et al., 2005).

The Tritryp databases are a constant tool in most every laboratory examining the cellular biology, molecular biology, or evolution of kinetoplastids (a list of some useful web sites is provided in Table 1). We will give a few of the many examples of genome use, and discuss potential future directions using emerging technologies.

Extending and improving the genomes

Some limitations to the genome assemblies exist currently. Within repeated sequences insufficient data may be available to distinguish between polymerase errors, sequencing errors, true alleles, and genomic repeats. As a result repetitive portions of the genome are not represented accurately or may be collapsed to falsely appear singular. These issues continue to be addressed by a number of researchers.

SL RNA genes are found in multicopy arrays on the order of a hundred copies. The published genome sequences for the SL RNA gene arrays in all three Tritryps are incomplete in various ways due to their repetitive natures, although attempts were made to present the data in as much detail as possible. In *L. major*, for example, an arbitrary assembly was created in an effort to represent as much of the data as possible, filling the estimated length of the array. The high level of sequence conservation coupled with the head-to-tail tandem organization of the SL RNA gene array has made it an ideal target for kinetoplastid identification, or barcoding (Fernandes et al., 1994; Westenberger et al., 2004; Maslov et al., 2007). Typically a single PCR product is cloned and sequenced per identification, based on the assumption that the arrays are homogeneous. SL repeat units gathered from *L. major* and *T. cruzi* sequencing projects were used to perform an analysis of the multicopy SL RNA array (Thomas et al., 2005), challenging the basis of the SL marker. Unexpected variation was revealed among the sequences within a single array, including SL gene sequences in *T. cruzi*, however the use of single SL array representatives was validated as a taxonomic marker in both genera.

Unannotated shotgun sequences generated by the genome projects of two *T. cruzi* strains were used to assemble complete maxicircle sequences (Westenberger et al., 2006), including the problematic variable regions that contain high numbers of repeated sequences. A preliminary catalogue of guide RNA genes in the mitochondrial minicircle population was also generated using the shotgun sequencing reads (Thomas et al., 2007), taking advantage of the relatively low level of small fragments contaminating the 5-kb size cutoff for shotgun cloning; their proportional mass alone dictated that the 1.4-kb minicircles would be present in the sequenced population.

Kinetoplastid genomes are rife with multicopy genes. Over 20,000 new open reading frames in the *T. cruzi* genomic analysis were described recently using a modified criterion that took sequence coverage into account to combat artificial compression of the repeated alignments (Arner et al., 2007). The basic idea is that sequence coverage should have been proportional for all areas of the genome, thus if a particular gene was sequenced at a multiple of the 7.5-fold average coverage of the entire genome, additional copies are indicated as indicated by the sequence representation. This genome-wide bioinformatic study nearly doubles the number of

genes identified by the original report, emphasizing the importance of thorough *in silico* analysis.

Pathway-specific studies

While their unique pathways are often the focus of attention, kinetoplastids are true eukaryotes and share much of the same basic machinery with their vertebrate hosts and insect vectors. Sequences with unknown function within the kinetoplastid genome projects can be ascribed roles based on homology with other organisms. For example, part of the unique 'cap 4' hypermethylation of the *trans*-spliced universal mRNA leader sequence, or SL RNA, is catalyzed by three enzymes, one that is a member of a larger family found in other eukaryotes, and two of which are relatives of a viral methyltransferase. The cap 4 enzymes had been sought unsuccessfully using biochemical methods for decades, but emerged in an exhaustive bioinformatic study of methyltransferases (Feder et al., 2003). This comparative genomics study provided targets whose roles in important biological processes are being validated and further studied at the molecular level, thus far revealing three of the five or six methyltransferases involved (Zamudio et al., 2006; Zamudio et al., 2007; Mittra et al., 2008).

Using the shared and distinct characteristics of other organisms will provide clues for the roles of many hypothetical kinetoplastid genes. A clever screen for cytoskeletal elements involved in cellular motility revealed a group of 50 potential targets, 41 of which had not been characterized previously, by comparing motile-flagellar and non-motile-flagellar components the 41 proteins were challenged for motility function using inducible RNA interference (Baron et al., 2007). This study used functional parameters to generate a list of candidate proteins that are likely to include key structural and regulatory elements in kinetoplastid movement. Anticipated members include not only structural flagellar components, but also regulatory elements controlling motor function such as Dynein Regulatory Complex (DRC).

By generating sequences from more strains and species, more elaborate comparative genomics studies can be performed to distinguish differences between these groups and perhaps to shed light on differences in biology and pathogenesis. Regulatory gene families including the kinases (Parsons et al., 2005), phosphatases (Brenchley et al., 2007), and RNA binding proteins (De Gaudenzi et al., 2005) have been the focus of Tritryp-based studies. This approach can help distinguish the various species of *Leishmania*, for example, and has been broached on an organismal level (Peacock et al., 2007; Smith et al., 2007) as well as for particular gene families and genomic regions (Jackson et al., 2006; Liang et al., 2007; Puechberty et al., 2007). Genome data is being used to determine the differences between *T. brucei* and *T. cruzi* (Obado et al., 2007) and to determine differences between strains of *T. cruzi* (Westenberger et al., 2005).

Last, but certainly not least, the genome data serve as an invaluable reference in the daily lives of researchers in simple inglorious tasks such as the design of cloning strategies that would have required a great deal more effort in the past. It is rare to come across a kinetoplastid manuscript that does not take advantage of a genome project at some level, indicating that the effort was worthy and that the presentation is accessible to the research community. Furthermore, the presence of the kinetoplastid sequences in general databases such as Genbank will lead to their inclusion in biologically broad studies, such as the methyltransferase study (Feder et al., 2003), revealing unsuspected activities in the process.

Genomics-based tools

Microarrays have become less expensive, more reliable, increasingly customizable and, as a result, widely used. With the current technology, a single microarray can contain probes covering nearly the entire kinetoplastid genome sequence, allowing researchers to map a hybridizing molecule back to the genome with great precision. The most broadly accessible

services offer custom microarray design, and full-service labeling, hybridization, and scanning of samples. Such services remove the onus of developing in-house technical expertise from scratch and provide more reliability and quality control to the technique, allowing the researcher to focus on proper experimental design. While the analyses for some applications such as expression studies can be performed as a service, many advanced experimental questions such as time-courses or protein-binding and motif discovery are unique or difficult to streamline. For these applications laboratories must develop the bioinformatic infrastructure to manage and analyze the data, or initiate strategic collaborations.

For kinetoplastid studies microarrays have been used primarily to assay steady-state levels of RNA transcripts genome-wide. These arrays can be used to determine differences in expression between different lifecycle stages (Cohen-Freue et al., 2007; Holzer et al., 2006; Leifso et al., 2007; Saxena et al., 2007; Srividya et al., 2007), and to determine expression differences associated with drug resistance (Salotra et al., 2006; Singh et al., 2007). More recent studies have used microarrays and chromatin immunoprecipitations (ChIP-chips) to map the binding of transcription factors and epigenetic markers genome-wide (Peter Myler, personal communication). These studies revealed the first genome-wide predictions for sites of polycistronic transcription initiation, confirmed the observation of acetylations at divergent strand-switch regions (Respuela et al. 2008), and suggested that the acetylations may be life-cycle-dependent.

DNA sequencing: beyond Sanger

Eleven years and many millions of dollars were required between the selection of the strains used for genome sequencing and the completion of the three genome projects produced by whole-genome shotgun for *T. cruzi* (El-Sayed et al., 2005), large insert clones and whole chromosome shotgun sequencing for *L. major* (Ivens et al., 2005), and whole chromosome shotgun with bacterial artificial cloning walking strategies for *T. brucei* (Berriman et al., 2005). The *L. infantum* and *L. braziliensis* genomes were sequenced relatively rapidly for a much lower price (Peacock et al., 2007). The Sanger dideoxy chain termination method was used to generate the data in these genome projects, using cloned genomic DNA fragments, vector-specific oligonucleotide primers, and a DNA polymerase for elongation. Chain terminator nucleotides identify the nucleotide at the terminal end of each DNA strand, and individual sequencing ‘reads’ range from 400–900 nt.

The drive to develop efficient sequencing technologies in the race for the $1,000 dollar human genome has encouraged further innovation (Service, 2006). With newer sequencing platforms, cost and speed are improved dramatically (Pop and Salzberg, 2008; Shendure et al., 2008). Non-Sanger methodologies have given rise to a new set of jargon to describe the processes, including ‘emulsion PCR’, ‘cyclic array’ and ‘massively parallel sequencing’. The details of these methods may differ, but essentially they involve running sequencing reactions on a nanoscale, avoiding cloning altogether while relying on PCR to amplify entire fragmented genomes in isolated microenvironments. A bare bones description of a genomic sequencing reaction would proceed as follows: 1) purified DNA is sheared to a size range of 300–800 bp, followed by ligation of two adapter oligonucleotides; 2) each genomic fragment is isolated from the rest of the genome and amplified by PCR; 3) the PCR fragments are attached to a solid support and localized onto a microwell plate with DNA polymerase; 4) the microwell plate is flooded sequentially with each nucleotide, and elongations in individual wells are read and recorded. ‘Emulsion PCR’ describes the aqueous PCR reactions that occur within mineral oil based droplets prior to attachment to a bead. The microwell plating results in the ‘massively parallel’ sequencing reactions that occur by the nucleotide ACGT ‘cycling’ until the read is complete, and are read according to output of particular nucleotide markers or the generation of pyrophosphate in the elongation (hence, ‘pyrosequencing’). Some methods elongate only

one base at a time, while the pyrosequencing method can record accurately a maximum of 8 consecutive identical nucleotides. The read lengths of the various methodologies range from a few dozen to several hundred nucleotides, making the task of data assembly more of a challenge than with the traditional Sanger capillary sequencing approaches. The genome of model kinetoplastid *Crithidia fasciculata* has been sequenced using the Roche 454 platform, and is being assembled (Stephen Beverley, personal communication).

The next generation of DNA sequencing technologies is likely to be dominated by single-molecule, real-time (e.g. SMRT) techniques as opposed to methods that require amplification of DNA or where sequencing occurs in repetitive cycles with pausing at various stages. One implementation of this strategy uses Fluorescence Resonance Energy Transfer (FRET) to monitor fluorescent nucleotides in real-time as they are incorporated into individual elongating strands (Korlach et al., 2008). Relying on the processivity, speed, and fidelity of DNA polymerase coupled with thousands of simultaneous reactions, Pacific Biosciences anticipates that an entire human genome can be sequenced in 15 minutes for less than $500 by 2013. Also in the near future, nanopore methods (Shendure et al., 2008) may play an important role in DNA sequencing. By drawing individual DNA strands through a nanopore, its bases can be scanned as it flows through, generating a very rapid sequence with extremely long read lengths. In addition to its theoretical cost and speed benefits, this technology has the relatively unique potential to eliminate current problems in assembling repetitive sequences with little genetic variation.

Proteomics and post-genomic analyses

Proteomics analyses complement microarray studies (McNicoll et al., 2006; Leifso et al., 2007) with distinct data content since a subset of the genome project is represented with further modifications overlaid upon the basic amino acid sequence. As alternative splicing is not anticipated to be a common strategy to generate diversity in kinetoplastid gene expression due to the dearth of introns (Berriman et al., 2005), post-transcriptional protein modifications will account for the variety of products arising from single gene sources. The content of a proteome analysis can change rapidly, as during a switch between life stages, or vary according to the sub-cellular locale. While some kinetoplastid studies focus on particular sub-cellular regions (Foucher et al., 2006) or organelles such as the mitochondrion or glycosome (Colasante et al., 2006), the most ambitious studies tackle the whole organism (Atwood et al., 2005) or seek to compare species (Brobey et al., 2006). Parasite antigenicity (Dea-Ayuela et al., 2006; Forgber et al., 2006; Gupta et al., 2007), and drug resistance (Vergnes et al., 2007) are common foci in kinetoplastid proteomics. Analysis of the kinetoplast transcriptome and proteosome revealed unexpected diversity (Lukeš et al., 2005), which could be attributed to the phenomenon of differential RNA editing (Ochsenreiter et al., 2008).

The identification of particular genes marks the beginning of characterization and validation at the protein level. While most research groups will follow specific pathways, others have taken a wider approach, as through the systematic application of RNA interference to each of the 210 genes found on chromosome 1 of *T. brucei* (Subramaniam et al., 2006). Proteins of particular interest can be assessed for participation in complexes using a variant of the tandem affinity purification tag called PTP (Schimanski et al., 2005).

Mass spectrometry is one of most efficient tools to use for proteomic characterization. The RBP12 subunit of *T. brucei* RNA polymerase II was characterized by mass spec, and seven proteins that associate with the RNA polymerase were identified (Das et al., 2006). With the emergence of Multidimensional Protein Identification Technology (MudPIT), complex protein populations can be characterized without prior gel purification of the target proteins. Using sequential cation exchange and reverse phase chromatography, the tryptic peptides in a sample

are subjected to mass spectrometry. Post-transcriptional modifications will play a major role in kinetoplastid gene expression, thus a catalog of secondary protein modifications will be revealing. An example of what could be revealed in kinetoplastids using MudPIT is the identification of sumoylated proteins in *Saccharomyces cerivisiae* by employing nickel-column purification of His8-tagged SUMO (or small ubiquitin-related modifier) conjugated proteins: 271 proteins modified by this post-transcriptional pathway were validated by a minimum of two peptide hits per gene (Wohlschelegel et al., 2004). As in any mass spec-driven analysis, a complete genomic sequence is essential to protein identification.

Bioinformatics: beyond BLAST

Increasingly popular fee-based genomics services include analysis of microarray, sequencing and spectrometry data, however the details of kinetoplastid biology often necessitate the ability to tailor analyses to suit their peculiarities. The programming language Perl, introduced by Larry Wall in 1987, and the statistics package R (Ihaka and Gentleman, 1996) are *in silico* tools for extracting the most from genomic data. Both are available free for download for a variety of operating systems from sources like ActivePerl and CRAN, respectively. After developing the ability to read and write files in standard formats (fasta, gff, bed, wiggle, etc), a multitude of analytical possibilities are accessible. In general, Perl is used when complex algorithms and manipulation of genomic sequence data are involved, and R is especially good with statistics and with plots and graphs. As an example, we have written Perl scrips to identify and analyze sequencing reads of SL RNA genes (Thomas et al., 2005) and both Perl and R were used to cull minicircle reads from the *T. cruzi* sequence database and to develop a method of predicting the location of guide RNAs within those sequences using a simple Hidden Markov Model (Thomas et al., 2007). Likewise, the KISS database (Ochsenreiter et al., 2007), developed for *T. brucei* guide RNA analysis used an assortment of bioinformatic programs to generate a website tool for use by the RNA editing community. Simple stepping stones such as reading data from a file, performing simple analyses, and writing the results back to a file pave the way for the development of more sophisticated analyses like dynamic programming methods (Markov Chain processes, Metropolis/Hastings approaches, etc.).

Among our short list of popular bioinformatics tools, BioConductor for R is free source for a range of tools including those needed for microarray analysis available for download from the CRAN website. MEME and MAST (Bailey et al., 2006) can be used to find and map, respectively, any motifs found in a set of sequences. The aptly named ‘Cluster’ (Eisen et al., 1998) can perform a range of clustering techniques. For visualization of clustered data, Matrix2png can be run over the web (Pavlidis and Noble, 2003) as can WebLogo (Crooks et al., 2004), which is used to generate cartoons of sequence motifs. For mass spectrometry data, the best software available will likely be that owned by the unit performing the analysis, however a few free tools such as the Open Mass Spectrometry Search Algorithm (OMSSA) are available through NCBI (Geer et al., 2004).

Permutation methods offer another powerful set of tools. To determine if a sequence motif is enriched in a particular set of immunoprecipitated sequences, for example, the probability that the motif could be found at random from among sequences of identical base frequency and length as that used in the analysis must be determined. This probability can be determined without any statistical knowledge whatsoever, thus removing the boundary between the researcher and a homegrown tool. The dataset is shuffled randomly and the number of times that sequence is found in the permuted list is counted. When repeated infinitely (or several thousand times), the probability can be approximated. Three observations in 100,000 permutations can be interpreted directly as a 3 in 100,000 chance that the motif observed can occur randomly, given weight to any conclusion that such a motif is significant when found multiple times in a data set. Permutation tests are expensive computationally and do require

some coding, but sample scripts are available (Thomas et al., 2007) that can be adapted for particular uses.

Beyond analysis and implementation, a grasp of how to interpret properly genomics data is imperative. To generate potential drug targets from microarray data, for example, a vast number of statistical tests are performed and a few are selected based on specific criteria. Three popular ways to determine significance and control error when performing massive numbers of statistical tests include: 1) the standard single test approach (most relaxed), 2) the false discovery rate (FDR) (Benjamini and Hochberg, 1995) approach, and 3) the Bonferroni correction (Bonferroni, 1936) the most stringent of the three. A working knowledge of these methods should be sought out in order to grasp the challenges of whole-genome studies.

To determine whether certain data cluster in any meaningful way, such as when evolutionary trees are derived or when gene networks are inferred from microarray data, two general categories of clustering algorithms can be applied: hierarchical and partitioning. Evolutionary trees are an example of hierarchical clustering, where a number of objects are clustered by distance from one another and are linked in a relationship tree. Partition methods such as K-means clustering (MacQueen, 1967) break a group of objects into a preordained number of classes, 2 for example, when the goal is to identify genes affected by a drug versus those unaffected.

This short survey of concepts is a slice through the realm of skills useful in genomics research that provide new views of data not otherwise visible. The conduct of biology research has changed in dramatic ways with the introduction of the BLAST algorithm (Altschul et al., 1990), combined with an ever-increasing number of fully sequenced genomes. Advancements in genomics technology and analysis will continue to shape the broader field, and bioinformatic methodologies will have to rush to keep pace with the torrent of information that will fill databases in the coming years.

Catching the nuances

In genomics we need ways to study cryptic sub-populations and 'unculturable' specimens. In these respects, the development of single-cell genomics techniques (Walker and Parkhill, 2008) will aid researchers answer questions that are difficult to address currently, such as the assembly of genomes for unculturable organisms. Even in cells that have been ‘domesticated’ it is not certain that the genomes of these organisms truly represent the variety found in nature, and in genomics studies of pathogenicity or potential therapies, the behavior of these domestic strains may not always reflect the behavior of fresh isolates. The development of single-cell genomics techniques will greatly aid our understanding of these currently-opaque problems.

Up-and-coming technologies could generate 3 Mbp per second of sequence information. The sheer magnitude of this output would allow experiments that could only be imagined just a few years ago. Combined with single-cell genomics techniques, a fresh isolate containing a number of genetically or functionally unique unculturable cells could be sequenced as individuals, have their genomes assembled, be completely profiled for expression, epigenetic markers, DNA-binding, and RNA-protein binding all with amazing rapidity. In conjunction with proteomics technologies, researchers will be able to measure what is happening at all levels of kinetoplastid physiology. A critical bottleneck in the ability to utilize these technologies effectively will be access to mathematics, statistics, computer science, informatics, and engineering skill sets within the kinetoplastid research community.

Acknowledgements

The authors thank Stephen Beverley and Peter Myler for communication of unpublished results, and David Campbell for critical reading of the manuscript. The authors are supported by NIH grant AI056034.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J. Mol. Biol 1990;215:403–410. [PubMed: 2231712]

Atwood JA, Weatherly DB, Minning TA, Bundy B, Cavola C, Opperdoes FR, Orlando R, Tarleton RL. The *Trypanosoma cruzi* proteome. Science 2005;309:473–476. [PubMed: 16020736]

Arner E, Kindlund E, Nilsson D, Farzana F, Ferella M, Tammi MT, Andersson B. Database of *Trypanosoma cruzi* repeated genes: 20,000 additional gene variants. BMC Genomics 2007;8:391. [PubMed: 17963481]

Bailey TL, Williams N, Misleh C, Li WW. MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res 2006;34:W369–W373. [PubMed: 16845028]

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. Ser. B 1995;57:289–300.

Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, Lennard NJ, Caler E, Hamlin NE, Haas B, Böhme U, Hannick L, Aslett MA, Shallom J, Marcello L, Hou L, Wickstead B, Alsmark UC, Arrowsmith C, Atkin RJ, Barron AJ, Bringaud F, Brooks K, Carrington M, Cherevach I, Chillingworth TJ, Churcher C, Clark LN, Corton CH, Cronin A, Davies RM, Doggett J, Djikeng A, Feldblyum T, Field MC, Fraser A, Goodhead I, Hance Z, Harper D, Harris BR, Hauser H, Hostetler J, Ivens A, Jagels K, Johnson D, Johnson J, Jones K, Kerhornou AX, Koo H, Larke N, Landfear S, Larkin C, Leech V, Line A, Lord A, Macleod A, Mooney PJ, Moule S, Martin DM, Morgan GW, Mungall K, Norbertczak H, Ormond D, Pai G, Peacock CS, Peterson J, Quail MA, Rabbinowitsch E, Rajandream MA, Reitter C, Salzberg SL, Sanders M, Schobel S, Sharp S, Simmonds M, Simpson AJ, Tallon L, Turner CM, Tait A, Tivey AR, Van Aken S, Walker D, Wanless D, Wang S, White B, White O, Whitehead S, Woodward J, Wortman J, Adams MD, Embley TM, Gull K, Ullu E, Barry JD, Fairlamb AH, Opperdoes F, Barrell BG, Donelson JE, Hall N, Fraser CM, Melville SE, El-Sayed NM. The genome of the African trypanosome *Trypanosoma brucei*. Science 2005;309:416–422. [PubMed: 16020726]

Bonferroni C. Teoria statistica delle classi e calcolo delle probabilita. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 1936;8:3–62.

Brenchley R, Tariq H, McElhinney H, Szö?r B, Huxley-Jones J, Stevens R, Matthews K, Tabernero L. The TriTryp phosphatome: analysis of the protein phosphatase catalytic domains. BMC Gemonics 2007;8:434.

Brobey RK, Mei FC, Cheng X, Soong L. Comparative two-dimensional gel electrophoresis maps for promastigotes of *Leishmania amazonensis* and *Leishmania major*. Braz. J. Infect. Dis 2006;10:1–6. [PubMed: 16767307]

Campbell DA, Thomas S, Sturm NR. Transcription in the kinetoplastid protozoa: why be normal? Microbes Infect 2003;5:1231–1240. [PubMed: 14623019]

Cohen-Freue G, Holzer TR, Forney JD, McMaster WR. Global gene expression in *Leishmania*. Int. J. Parasitol 2007;37:1077–1086. [PubMed: 17574557]

Colasante C, Ellis M, Ruppert T, Voncken F. Comparative proteomics of glycosomes from bloodstream form and procyclic culture form *Trypanosoma brucei brucei*. Proteomics 2006;6:3275–3293. [PubMed: 16622829]

Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res 2004;14:1188–1190. [PubMed: 15173120]

Das A, Li H, Liu T, Bellofatto V. Biochemical characterization of *Trypanosoma brucei* RNA polymerase II. Mol. Biochem. Parasitol 2006;150:201–210. [PubMed: 16962183]

De Gaudenzi J, Frasch AC, Clayton C. RNA-binding domain proteins in kinetoplastids: a comparative analysis. Eukaryot. Cell 2005;4:2106–2114. [PubMed: 16339728]

Dea-Ayuela MA, Rama-Iñiguez S, Bolás-Fernandez F. Proteomic analysis of antigens from *Leishmania infantum* promastigotes. Proteomics 2006;6:4187–4194. [PubMed: 16791830]

Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA 1998;95:14863–14868. [PubMed: 9843981]

El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran AN, Ghedin E, Worthey EA, Delcher AL, Blandin G, Westenberger SJ, Caler E, Cerqueira GC, Branche C, Haas B, Anupama A, Arner E, Aslund L, Attipoe P, Bontempi E, Bringaud F, Burton P, Cadag E, Campbell DA, Carrington M, Crabtree J, Darban H, da Silveira JF, de Jong P, Edwards K, Englund PT, Fazelina G, Feldblyum T, Ferella M, Frasch AC, Gull K, Horn D, Hou L, Huang Y, Kindlund E, Klingbeil M, Kluge S, Koo H, Lacerda D, Levin MJ, Lorenzi H, Louie T, Machado CR, McCulloch R, McKenna A, Mizuno Y, Mottram JC, Nelson S, Ochaya S, Osoegawa K, Pai G, Parsons M, Pentony M, Pettersson U, Pop M, Ramirez JL, Rinta J, Robertson L, Salzberg SL, Sanchez DO, Seyler A, Sharma R, Shetty J, Simpson AJ, Sisk E, Tammi MT, Tarleton R, Teixeira S, Van Aken S, Vogt C, Ward PN, Wickstead B, Wortman J, White O, Fraser CM, Stuart KD, Andersson B. The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. Science 2005a;309:409–415. [PubMed: 16020725]

El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, Caler E, Renauld H, Worthey EA, Hertz-Fowler C, Ghedin E, Peacock C, Bartholomeu DC, Haas BJ, Tran AN, Wortman JR, Alsmark UC, Angiuoli S, Anupama A, Badger J, Bringaud F, Cadag E, Carlton JM, Cerqueira GC, Creasy T, Delcher AL, Djikeng A, Embley TM, Hauser C, Ivens AC, Kummerfeld SK, Pereira-Leal JB, Nilsson D, Peterson J, Salzberg SL, Shallom J, Silva JC, Sundaram J, Westenberger S, White O, Melville SE, Donelson JE, Andersson B, Stuart KD, Hall N. Comparative genomics of trypanosomatid parasitic protozoa. Science 2005b;309:404–409. [PubMed: 16020724]

Feder M, Pas J, Wyrwicz LS, Bujnicki JM. Molecular phylogenetics of the RrmJ/fibrillarin superfamily of ribose 2'-*O*-methyltransferases. Gene 2003;302:129–138. [PubMed: 12527203]

Fernandes O, Murthy VK, Kurath U, Degrave WM, Campbell DA. Mini-exon gene variation in human pathogenic *Leishmania* species. Mol. Biochem. Parasitol 1994;66:261–271. [PubMed: 7808476]

Forgber M, Basu R, Roychoudhury K, Theinert S, Roy S, Sundar S, Walden P. Mapping the antigenicity of the parasites in *Leishmania donovani* infection by proteome serology. PLoS ONE 2006;1:e40. [PubMed: 17183669]

Foucher AL, Papadopoulou B, Ouellette M. Prefractionation by digitonin extraction increases representation of the cytosolic and intracellular proteome of *Leishmania infantum*. J. Proteome Res 2006;5:1741–1750. [PubMed: 16823982]

Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH. Open mass spectrometry search algorithm. J. Proteome Res 2004;3:958–964. [PubMed: 15473683]

Gupta SK, Sisodia BS, Sinha S, Hajela K, Naik S, Shasany AK, Dube A. Proteomic approach for identification and characterization of novel immunostimulatory proteins from soluble antigens of *Leishmania donovani* promastigotes. Proteomics 2007;7:816–823. [PubMed: 17295358]

Holzer TR, McMaster WR, Forney JD. Expression profiling by whole-genome interspecies microarray hybridization reveals differential gene expression in procyclic promastigotes, lesion-derived amastigotes, and axenic amastigotes in *Leishmania mexicana*. Mol. Biochem. Parasitol 2006;146:198–218. [PubMed: 16430978]

Ihaka R, Gentleman R. R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics 1996;5:299–314.

Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, Sisk E, Rajandream MA, Adlem E, Aert R, Anupama A, Apostolou Z, Attipoe P, Baso N, Bauser C, Beck A, Beverley SM, Bianchettin G, Borzym K, Bothe G, Bruschi CV, Collins M, Cadag E, Ciarloni L, Clayton C, Coulson RM, Cronin A, Cruz AK, Davies RM, De Gaudenzi J, Dobson DE, Duesterhoeft A, Fazelina G, Fosker N, Frasch AC, Fraser A, Fuchs M, Gabel C, Goble A, Goffeau A, Harris D, Hertz-Fowler C, Hilbert H, Horn D, Huang Y, Klages S, Knights A, Kube M, Larke N, Litvin L, Lord A, Louie T, Marra M, Masuy D, Matthews K, Michaeli S, Mottram JC, Muller-Auer S, Munden H, Nelson S, Norbertczak H, Oliver K, O'Neil S, Pentony M, Pohl TM, Price C, Purnelle B, Quail MA, Rabbinowitsch E, Reinhardt R, Rieger M, Rinta J, Robben J, Robertson L, Ruiz JC, Rutter S, Saunders D, Schafer M, Schein J, Schwartz DC, Seeger K, Seyler A, Sharp S, Shin H, Sivam D, Squares R, Squares S, Tosato V, Vogt C, Volckaert G, Wambutt R, Warren T, Wedler H, Woodward J, Zhou S, Zimmermann W, Smith DF, Blackwell JM, Stuart KD, Barrell B, et al. The genome of the kinetoplastid parasite, *Leishmania major*. Science 2005;309:436–442. [PubMed: 16020728]

Jackson AP, Vaughan S, Gull K. Comparative genomics and concerted evolution of beta-tubulin paralogs in *Leishmania* spp. BMC Genomics 2006;7:137. [PubMed: 16756660]

Korlach J, Marks PJ, Cicero RL, Gray JJ, Murphy DL, Roitman DB, Pham TT, Otto GA, Foquet M, Turner SW. Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. Proc. Natl. Acad. Sci. USA 2008;105:1176–1181. [PubMed: 18216253]

Leifso K, Cohen-Freue G, Dogra N, Murray A, McMaster WR. Genomic and proteomic expression analysis of *Leishmania* promastigote and amastigote life stages: the Leishmania genome is constitutively expressed. Mol. Biochem. Parasitol 2007;152:35–46. [PubMed: 17188763]

Liang XH, Hury A, Hoze E, Uliel S, Myslyuk I, Apatoff A, Unger R, Michaeli S. Genome-wide analysis of C/D and H/ACA-like small nucleolar RNAs in *Leishmania major* indicates conservation among trypanosomatids in the repertoire and in their rRNA targets. Eukaryot. Cell 2007;6:361–377. [PubMed: 17189491]

Lukeš J, Guilbride DL, Votypka J, Zikova A, Benne R, Englund PT. Kinetoplast DNA network: evolution of an improbable structure. Eukaryot Cell 2002;1:495–502. [PubMed: 12455998]

Lukeš J, Hashimi H, Zikova A. Unexplained complexity of the mitochondrial genome and transcriptome in kinetoplastid flagellates. Curr. Genet 2005;48:277–299. [PubMed: 16215758]

MacQueen, J. Some methods for classification and analysis of multivariate observations; Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability; 1967.

Maslov DA, Westenberger SJ, Xu X, Campbell DA, Sturm NR. Discovery and barcoding by analysis of spliced leader RNA gene sequences of new isolates of Trypanosomatidae from *Heteroptera* in Costa Rica and Ecuador. J. Eukaryot. Microbiol 2007;54:57–65. [PubMed: 17300521]

McNicoll F, Drummelsmith J, Muller M, Madore E, Boilard N, Ouellette M, Papadopoulou B. A combined proteomic and transcriptomic approach to the study of stage differentiation in *Leishmania infantum*. Proteomics 2006;6:3567–3581. [PubMed: 16705753]

Mittra B, Zamudio JR, Bujnicki JM, Stepinski J, Darzynkiewicz E, Campbell DA, Sturm NR. The TbMTr1 spliced leader RNA cap 1 2'-*O*-ribose methyltransferase from *Trypanosoma brucei* acts with substrate specificity. J. Biol. Chem 2008;283:3161–3172. [PubMed: 18048356]

Obado SO, Bot C, Nilsson D, Andersson B, Kelly JM. Repetitive DNA is associated with centromeric domains in *Trypanosoma brucei* but not *Trypanosoma cruzi*. Genome Biol 2007;8:R37. [PubMed: 17352808]

Ochsenreiter T, Cipriano M, Hajduk SL. KISS: the kinetoplastid RNA editing sequence search tool. RNA 2007;13:1–4. [PubMed: 17123956]

Parsons M, Worthey EA, Ward PN, Mottram JC. Comparative analysis of the kinomes of three pathogenic trypanosomatids: *Leishmania major*, *Trypanosoma brucei* and *Trypanosoma cruzi*. BMC Genomics 2005;6:127. [PubMed: 16164760]

Pavlidis P, Noble WS. Matrix2png: a utility for visualizing matrix data. Bioinformatics 2003;19:295–296. [PubMed: 12538257]

Peacock CS. The practical implications of comparative kinetoplastid genomics. SEB Exp. Biol. Ser 2007;58:25–45. [PubMed: 17608236]

Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, Quail MA, Peters N, Adlem E, Tivey A, Aslett M, Kerhornou A, Ivens A, Fraser A, Rajandream MA, Carver T, Norbertczak H, Chillingworth T, Hance Z, Jagels K, Moule S, Ormond D, Rutter S, Squares R, Whitehead S, Rabbinowitsch E, Arrowsmith C, White B, Thurston S, Bringaud F, Baldauf SL, Faulconbridge A, Jeffares D, Depledge DP, Oyola SO, Hilley JD, Brito LO, Tosi LR, Barrell B, Cruz AK, Mottram JC, Smith DF, Berriman M. Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. Nat. Genet 2007;39:839–847. [PubMed: 17572675]

Pop M, Salzberg SL. Bioinformatics challenges of new sequencing technology. Trends Genet 2008;24:142–149. [PubMed: 18262676]

Puechberty J, Blaineau C, Meghamla S, Crobu L, Pages M, Bastien P. Compared genomics of the strand switch region of *Leishmania* chromosome 1 reveal a novel genus-specific gene and conserved structural features and sequence motifs. BMC Genomics 2007;8:57. [PubMed: 17319967]

Respuela P, Ferella M, Fada-Iglesias A, Åslund L. Histone acetylation and methylation at sites initiating divergent polycistronic transcription in *Trypanosoma cruzi*. J. Biol. Chem 2008;283:15884–15892. [PubMed: 18400752]

Salotra P, Duncan RC, Singh R, Subba Raju BV, Sreenivas G, Nakhasi HL. Upregulation of surface proteins in *Leishmania donovani* isolated from patients of post kala-azar dermal leishmaniasis. Microbes Infect 2006;8:637–644. [PubMed: 16469521]

Saxena A, Lahav T, Holland N, Aggarwal G, Anupama A, Huang Y, Volpin H, Myler PJ, Zilberstein D. Analysis of the *Leishmania donovani* transcriptome reveals an ordered progression of transient and permanent changes in gene expression during differentiation. Mol. Biochem. Parasitol 2007;152:53–65. [PubMed: 17204342]

Schimanski B, Nguyen TN, Günzl A. Highly efficient tandem affinity purification of trypanosome protein complexes based on a novel epitope combination. Eukaryot. Cell 2005;4:1942–1950. [PubMed: 16278461]

Service RF. Gene sequencing. The race for the $1000 genome. Science 2006;311:1544–1546. [PubMed: 16543431]

Shendure JA, Porreca GJ, Church GM. Overview of DNA sequencing strategies. Chapter 7. Curr. Protoc. Mol. Biol. 2008Unit 7 1

Singh N, Almeida R, Kothari H, Kumar P, Mandal G, Chatterjee M, Venkatachalam S, Govind MK, Mandal SK, Sundar S. Differential gene expression analysis in antimony-unresponsive Indian kala azar (visceral leishmaniasis) clinical isolates by DNA microarray. Parasitology 2007;134:777–787. [PubMed: 17306059]

Smith DF, Peacock CS, Cruz AK. Comparative genomics: from genotype to disease phenotype in the leishmaniases. Int. J. Parasitol 2007;37:1173–1186. [PubMed: 17645880]

Srividya G, Duncan R, Sharma P, Raju BV, Nakhasi HL, Salotra P. Transcriptome analysis during the process of in vitro differentiation of *Leishmania donovani* using genomic microarrays. Parasitology 2007;134:1527–1539. [PubMed: 17553180]

Stuart KD, Schnaufer A, Ernst NL, Panigrahi AK. Complex management: RNA editing in trypanosomes. Trends Biochem. Sci 2005;30:97–105. [PubMed: 15691655]

Subramaniam C, Veazey P, Redmond S, Hayes-Sinclair J, Chambers E, Carrington M, Gull K, Matthews K, Horn D, Field MC. Chromosome-wide analysis of gene function by RNA interference in the african trypanosome. Eukaryot. Cell 2006;5:1539–1549. [PubMed: 16963636]

Thomas S, Martinez LL, Westenberger SJ, Sturm NR. A population study of the minicircles in *Trypanosoma cruzi*: predicting guide RNAs in the absence of empirical RNA editing. BMC Genomics 2007;8:133. [PubMed: 17524149]

Thomas S, Westenberger SJ, Campbell DA, Sturm NR. Intragenomic spliced leader RNA array analysis of kinetoplastids reveals unexpected transcribed region diversity in *Trypanosoma cruzi*. Gene 2005;352:100–108. [PubMed: 15925459]

Vergnes B, Gourbal B, Girard I, Sundar S, Drummelsmith J, Ouellette M. A proteomics screen implicates HSP83 and a small kinetoplastid calpain-related protein in drug resistance in *Leishmania donovani* clinical field isolates by modulating drug-induced programmed cell death. Mol. Cell. Proteomics 2007;6:88–101. [PubMed: 17050524]

Walker A, Parkhill J. Single-cell genomics. Nat Rev Microbiol 2008;6:176–177. [PubMed: 18283727]

Westenberger SJ, Barnabé C, Campbell DA, Sturm NR. Two hybridization events define the population structure of *Trypanosoma cruzi*. Genetics 2005;171:527–543. [PubMed: 15998728]

Westenberger SJ, Cerqueira GC, El-Sayed NM, Zingales B, Campbell DA, Sturm NR. *Trypanosoma cruzi* mitochondrial maxicircles display species- and strain-specific variation and possess a conserved element in the non-coding region. BMC Genomics 2006;7:60. [PubMed: 16553959]

Westenberger SJ, Sturm NR, Yanega D, Podlipaev SA, Zeledón R, Campbell DA, Maslov DA. Trypanosomatid biodiversity in Costa Rica: genotyping of parasites from Heteroptera using the spliced leader RNA gene. Parasitology 2004;129:537–547. [PubMed: 15552399]

Wohlschlegel JA, Johnson ES, Reed SI, Yates JR 3rd. Global analysis of protein sumoylation in *Saccharomyces cerevisiae*. J. Biol. Chem 2004;279:45662–45668. [PubMed: 15326169]

Zamudio JR, Mittra B, Foldynová-Trantírková S, Zeiner GM, Lukeš J, Bujnicki JM, Sturm NR, Campbell DA. The 2'-*O*-ribose Methyltransferase for Cap 1 of Spliced Leader RNA and U1 small nuclear RNA in *Trypanosoma brucei*. Mol. Cell. Biol 2007;27:6084–6092. [PubMed: 17606627]

Zamudio JR, Mittra B, Zeiner GM, Feder M, Bujnicki JM, Sturm NR, Campbell DA. Complete cap 4 formation is not required for viability in *Trypanosoma brucei*. Eukaryot. Cell 2006;5:905–915. [PubMed: 16757738]

Table 1

List of useful web sites for the kinetoplastid genomics researcher

Kinetoplastid-specific	
GeneDB	www.genedb.org/
TrypanoFAN database	trypanofan.path.cam.ac.uk/trypanofan/main/
Structural Genomics of Parasitic Protozoa	www.sgpp.org
TDR drug discovery database	tdrtargets.org/
KISS RNA editing tool	rna.bmb.uga.edu/kiss/
Genome browsers	
UCSC genome browser	genome.ucsc.edu
Artemis genome browser	www.sanger.ac.uk/Software/Artemis
NimbleGen genome browser	www.nimblegen.com
Sequencing	
454 Sequencing techology	www.dkfz.de/gpcf/242.html
SMRT	www.pacificbiosciences.com/index.php
Nanopore	www.mcb.harvard.edu/branton/projects-NanoporeSequencing.htm
Bioinformatics	
Perl	www.activestate.com/Products/ActivePerl
R	cran.r-project.org
Cluster	rana.lbl.gov/EisenSoftware.htm
MEME (motif discovery)	meme.sdsc.edu/meme
Gibb's Sampler (motif discovery)	bayesweb.wadsworth.org/gibbs/gibbs.html
Sequence logos	weblogo.berkeley.edu
OMSSA (mass spec search)	pubchem.ncbi.nlm.nih.gov/omssa
matrix2png	www.bioinformatics.ubc.ca/matrix2png