Julia Walochnik
Michael Duchêne *Editors*

Molecular Parasitology

Protozoan Parasites and their Molecules

Springer

Molecular Parasitology

Julia Walochnik • Michael Duchêne
Editors

Molecular Parasitology

Protozoan Parasites and their Molecules

Springer

Editors
Julia Walochnik
Institute of Specific Prophylaxis
and Tropical Medicine
Center for Pathophysiology, Infectiology
and Immunology
Medical University of Vienna
Vienna
Austria

Michael Duchêne
Institute of Specific Prophylaxis
and Tropical Medicine
Center for Pathophysiology, Infectiology
and Immunology
Medical University of Vienna
Vienna
Austria

ISBN 978-3-7091-1415-5 ISBN 978-3-7091-1416-2 (eBook)
DOI 10.1007/978-3-7091-1416-2

Library of Congress Control Number: 2016947730

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer-Verlag GmbH Wien

Contents

Contributors

Anton Aebischer Head of FG 16 Infections, Robert Koch-Institute, Berlin, Germany

Carsten Balczun Central Institute of the Bundeswehr, Medical Service Koblenz, Koblenz, Germany

Ulrike Boehme Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, UK

Christian Bogdan Mikrobiologisches Institut – Klinische Mikrobiologie, Immunologie und Hygiene, Friederich-Alexander-Universität (FAU) Erlangen-Nürnberg, Universitätsklinikum Erlangen, Erlangen, Germany

Reto Brun Medical Parasitology and Infection Biology, Swiss Tropical and Public Health Institute, University of Basel, Basel, Switzerland

Christine Clayton Group Leader in the DKFZ-ZMBH Alliance, ZMBH, Heidelberg, Germany

Alan F. Cowman Head, Division of Infection and Immunity, The Walter and Eliza Hall institute of Medical Research, Parkville, VIC, Australia

Department of Medical Biology, Faculty of Medicine, Dentistry and Health Sciences, The University of Melbourne, Parkville, VIC, Australia

Kathryn Crouch Wellcome Trust Centre for Molecular Parasitology, Glasgow, UK

Joel B. Dacks Department of Cell Biology, Canada Research Chair in Evolutionary Cell Biology, University of Alberta, Edmonton, Alberta, Canada

Daniele Dessì Department of Biomedical Sciences, University of Sassari, Sassari, Italy

Michael Duchêne Institute of Specific Prophylaxis and Tropical Medicine, Center for Pathophysiology, Infectiology and Immunology, Medical University of Vienna, Vienna, Austria

Pier Luigi Fiori Department of Biomedical Sciences, University of Sassari, Sassari, Italy

Friedrich Frischknecht Parasitology - Department of Infectious Diseases, University of Heidelberg Medical School, Heidelberg, Germany

Tim Gilberger, PhD Pathology and Molecular Medicine, McMaster University, Hamilton, ON, Canada

Sven Gould Institute for Molecular Evolution, Heinrich-Heine-University, Düsseldorf, Germany

Vladimír Hampl Department of Parasitology, Faculty of Science, Charles University in Prague, Prague, Czech Republic

Omar S. Harb Department of Biology, University of Pennsylvania, Philadelphia, PA, USA

Julie Healer Walter & Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia

Emily Herman Department of Cell Biology, Canada Research Chair in Evolutionary Cell Biology, University of Alberta, Edmonton, Alberta, Canada

Volker Heussler Institute of Cell Biology, University of Bern, Bern, Switzerland

Bernhard Nocht Institute for Tropical Medicine, Hamburg, Germany

Robert Hirt Institute for Cell and Molecular Biosciences, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, UK

Ivan Hrdy Department of Parasitology, Faculty of Science, Charles University in Prague, Prague, Czech Republic

Shanshan Hu Menzies Institute for Medical Research, University of Tasmania, Tasmania, Australia

Olukemi O. Ifeonu University of Maryland School of Medicine, Institute for Genome Sciences, Baltimore, MD, USA

Anna Karnkowska Department of Parasitology, Faculty of Science, Charles University in Prague, Prague, Czech Republic

Christen Klinger Department of Cell Biology, Canada Research Chair in Evolutionary Cell Biology, University of Alberta, Edmonton, AL, Canada

Martina Köhsler Institute of Specific Prophylaxis and Tropical Medicine, Center for Pathophysiology, Infectiology and Immunology, Medical University of Vienna, Vienna, Austria

Heinrich Körner Menzies Institute for Medical Research, Tasmania, Australia

Carsten G.K. Lüder Department of Medical Microbiology, Göttingen University Medical School, Göttingen, Germany

Pascal Mäser Medical Parasitology and Infection Biology, Swiss Tropical and Public Health Institute, University of Basel, Basel, Switzerland

Martin Mrva Department of Zoology, Comenius University in Bratislava, Faculty of Natural Sciences, Bratislava, Slovakia

Norbert Müller Institute for Parasitology, University of Berne, Bern, Switzerland

Joachim Müller Institute for Parasitology, Vetsuisse and Medical Faculty, University of Berne, Bern, Switzerland

Paola Rappelli Department of Biomedical Sciences, University of Sassari, Sassari, Italy

David S. Roos Department of Biology, University of Pennsylvania, Philadelphia, PA, USA

Günter A. Schaub Zoologie/Parasitologie, Ruhr-Universität-Bochum, Bochum, Germany

Frank Seeber 2 Robert-Koch-Institut, Berlin, Germany

Joana C. Silva University of Maryland School of Medicine, Institute for Genome Sciences, Baltimore, MD, USA

Fatima Silva-Franco Institute of Integrative Biology, University of Liverpool, Liverpool, UK

Tobias Spielmann Bernhard Nocht Institute for Tropical Medicine, Hamburg, Germany

Staffan Svärd Uppsala University, Uppsala University, BMC, Uppsala, Sweden

Jan Tachezy Department of Parasitology, Faculty of Science, Charles University in Prague, Prague, Czech Republic

Kyle Tretina University of Maryland School of Medicine, Institute for Genome Sciences, Baltimore, MD, USA

Patric Vogel Zoology/Parasitology Group, Ruhr-Universität, Bochum, Germany

Julia Walochnik Institute of Specific Prophylaxis and Tropical Medicine, Center for Pathophysiology, Infectiology and Immunology, Medical University of Vienna, Vienna, Austria

Jonathan Wastling Executive Dean, Faculty of Natural Sciences, Keele University, Keele, UK

Gareth Weedall Vector Biology Department, Liverpool School of Tropical Medicine, Liverpool, UK

Iain B.H. Wilson Department for Chemistry, University of Natural Resources and Life Sciences (BOKU), Vienna, Austria

Dong Xia Department of Infection Biology, Institute of Infection and Global Health, University of Liverpool, Liverpool, UK

Part I

The Molecules

Genomics

1

Omar S. Harb, Ulrike Boehme, Kathryn Crouch, Olukemi O. Ifeonu, David S. Roos, Joana C. Silva, Fatima Silva-Franco, Staffan Svärd, Kyle Tretina, and Gareth Weedall

Abstract

In the last decade, the rise of affordable high-throughput sequencing technologies has led to rapid advances across the biological sciences. At the time of writing,

O.S. Harb (✉) • D.S. Roos
Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA
e-mail: oharb@upenn.edu

U. Boehme
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus,
Hinxton, Cambridgeshire CB10 1SA, UK
e-mail: ucb@sanger.ac.uk

K. Crouch
Wellcome Trust Centre for Molecular Parasitology,
B6-28 SGDB, 120 University Place, Glasgow G12 8TA, UK
e-mail: kathryn.crouch@glasgow.ac.uk

O.O. Ifeonu • J.C. Silva • K. Tretina
Institute for Genome Sciences, University of Maryland School of Medicine,
BioPark II, Room 645, 801W. Baltimore St., Baltimore MD 21201, USA
e-mail: KAbolude@som.umaryland.edu; jcsilva@som.umaryland.edu;
KTretina@som.umaryland.edu

F. Silva-Franco
Institute of Integrative Biology, University of Liverpool, Liverpool L69 7ZB, UK
e-mail: F.Silva-Franco@liverpool.ac.uk

S. Svärd
Department of Cell and Molecular Biology, Uppsala University, BMC,
Box 596, Uppsala SE-75123, Sweden
e-mail: staffan.svard@icm.uu.se

G. Weedall
Vector Biology Department, Liverpool School of Tropical Medicine,
Pembroke Place, Liverpool L3 5QA, UK
e-mail: gareth.weedall@lstmed.ac.uk

J. Walochnik, M. Duchêne (eds.), *Molecular Parasitology*,
DOI 10.1007/978-3-7091-1416-2_1

annotated reference genomes are available for most clades of eukaryotic pathogens. Over 550 genomes, including unannotated sequences, are available in total. This has greatly facilitated studies in many areas of parasitology. In addition, the volume of functional genomics data, including analysis of differential transcription and DNA-protein interactions, has increased exponentially. With this unprecedented increase in publicly available data, tools to search and compare datasets are becoming ever more important. A number of database resources are available, and access to these has become fundamental for a majority of research groups. This chapter discusses the current state of genomics research for a number of eukaryotic parasites, addressing the genome and functional genomics resources available and highlighting functionally important or unique aspects of the genome for each group. Publicly accessible database resources pertaining to eukaryotic parasites are also discussed.

1.1 Introduction

Arguably the field of genomics began when Friedrich Miescher first isolated DNA in 1869 (Dahm 2005), paving the way for the work of many scientists in understanding the role of this material in heredity (Avery et al. 1944), discovering its double-helical structure (Watson and Crick 1953) and deciphering the genetic code (Nirenberg et al. 1965). However, the technological advance that the entire field of genomics rests on is sequencing (Gilbert and Maxam 1973; Sanger et al. 1977; Wu 1972). The ability to read the genetic code is relatively new, having only been developed in the last 50 years. Sanger sequencing, which relies on dideoxy chain termination, remained the method of choice for several decades; however, early implementations of dideoxy chain termination methods were not well parallelized, and analysis was initially a painstaking manual process. Later, data analysis was carried out computationally but limited by the processing capacity of computers of the era. These factors combined to limit early sequencing to individual genes, small genomic fragments, or the genomes of small viruses and organelles. The emergence of techniques such as fluorescence-based cycle sequencing and the polymerase chain reaction in addition to the increased use of computational power to automatically read and analyze results allowed larger-scale genome projects to be undertaken (Prober et al. 1987). Indeed within a few years of this marriage of techniques and fields, the first bacterial, protozoan, fungal, plant, and animal genomes were sequenced (Fleischmann et al. 1995; Fiers et al. 1976; Goffeau et al. 1996; Gardner et al. 2002; The Arabidopsis Genome Initiative 2000; The *C. elegans* Sequencing Consortium 1998). Despite these advances, sequencing of whole genomes remained relatively costly and time consuming. As an example, sequencing the human genome took roughly 10 years at a price tag of 3 billion US dollars (https://www.genome.gov/11006943) (Lander et al. 2001).

The first forays into high-throughput analysis of sequence data came in the form of microarrays. A microarray consists of a panel of oligonucleotide probes bonded to a solid surface such as a glass slide. Hybridization of nucleic acids from a specimen to individual probes is detected by the intensity of a fluorescent signal (Schena et al. 1995). This technique was the first to make querying of sequence polymorphisms,

transcript expression levels, and segmental duplications possible on a genomic level and cheap enough to be widely available. In addition, microarrays forced the development of computational tools and techniques to handle data on a genomic scale. However, an important limitation of microarrays is the requirement for prior knowledge of the genome and the coincident inability to make de novo discoveries (i.e., one can query the presence of known SNPs but not discover new SNPs). A large volume of functional genomics data has been obtained using microarray technologies, but with a small number of exceptions (such as diagnostics), microarrays have for the most part been superseded by next-generation sequencing technologies.

Two factors have been instrumental in enabling sequencing to be taken to the next level: continued growth of computer processing capacity following Moore's law (Moore 1998) and the development of "next-generation" sequencing (NGS) methods (also known as second-generation sequencing), which enable massively parallel sequencing of millions of fragments by synthesis (Margulies et al. 2005; Shendure et al. 2005). One of the major advantages of next-generation sequencing is that it can be applied to a wide variety of methodologies (*readers are directed to an excellent series of manuscripts*; http://www.nature.com/nrg/series/nextgeneration/index.html) and unlike microarrays, does not require any prior knowledge of the sample. Methodologies include:

- DNA sequencing: High-throughput technology makes sequencing for de novo assembly of new genomes ever more affordable. Comparison of re-sequenced isolates against a reference is a common technique for discovery of sequence polymorphisms, while analysis of coverage depth and mapping topology can reveal information about structural variations such as chromosomal translocations and segmental duplications.
- RNA sequencing: Sequencing of RNA can provide important information about gene structure such as the locations of UTRs and intron/exon boundaries and the presence of alternative or *trans*-splice variants. Analysis of RNA-seq coverage depth over a time course or under different experimental conditions reveals information about transcription of genes under differing conditions, and combination of this technique with ribosomal profiling enables identification of the translational status of the genome. Specialized sample preparation techniques enable the sequencing of noncoding RNA species such as those involved in the RNAi-mediated translational silencing.
- Epigenomics: Chromatin immunoprecipitation (ChIP) sequencing is a powerful technique that allows determination of the "footprint" of DNA-binding proteins. This can be used to examine promoter-binding sites; transcription, replication, and repair mechanisms; and factors such as histone modification that can affect transcription. Other techniques are available, such as bisulfite sequencing, which enables profiling of DNA methylation.
- Metagenomics: Sequencing of DNA extracted from samples that contain mixed populations of organisms can be used to survey populations in environmental samples (such as soil) or biological samples (such as gut microbiomes). Metagenomics techniques can be used to determine the makeup of populations and to survey how this changes over time or under different conditions. Metagenomics analysis is a fast-growing field in which the problems of analysis have not yet been solved.

It is not surprising that the dawn of large-scale sequencing projects necessitated an expansion in the field of bioinformatics and data management. As high-throughput sequencing has become cheaper, it has moved from being a specialist technique to a tool used daily in labs across the world. This has necessitated the development of user-friendly tools that can run on desktop machines and thrust the field of bioinformatics into the foreground. The expansion of massively parallel sequencing has also led to a revolution in the teaching of biology, with computational techniques for management and analysis of genomic-scale datasets now being taught in many undergraduate courses. Data warehousing is also becoming a priority, with data repositories such as the National Center for Biotechnology Information (NCBI) (Kodama et al. 2012) having to rethink both their submissions procedures and their approaches to storage.

1.2 Parasite Genomics

The field of parasite genomics has benefited tremendously from the sequencing revolution. While only a handful of parasite genomes were sequenced by 2005, the number has exploded to over 550 genomes (http://genomesonline.org) (Reddy et al. 2015) by 2015. This number reflects both annotated and unannotated genomes and will already be out of date by the time this chapter is in print. Besides the technological advances, this increase in sequences has been aided by a number of initiatives with parasitology components. These include projects supported by the Wellcome Trust Sanger Institute in the United Kingdom and a number of parasite-specific genome sequencing white papers supported by the National Institute of Allergy and Infectious Diseases (NIAID) Genomic Centers for Infectious Diseases (GCID) in the United States. Together these centers have generated sequence, assemblies, and annotation from many important human and veterinary parasites. All data from these projects are available via project-specific websites (i.e., GeneDB: http://genedb.org) (Logan-Klumpler et al. 2012) and/or through the International Nucleotide Sequence Databases (GenBank, EMBL Nucleotide Sequence Database, and the DNA Data Bank of Japan (Kulikova et al. 2006; Benson et al. 2015; Tateno et al. 1998)).

1.3 General Features of Protozoan Parasite Genomes

1.3.1 *Giardia*

Giardia duodenalis, also known as *Giardia intestinalis* or *Giardia lamblia*, is a unicellular protozoan parasite that infects the upper intestinal tract of humans and animals (Ankarklev et al. 2010). The disease, giardiasis, manifests in humans as an acute diarrhea that can develop to a chronic diarrhea, but the majority of infections remain asymptomatic (Ankarklev et al. 2010). Giardiasis has a global distribution with 280 million cases reported annually, with its impact being more pronounced in the developing world.

G. duodenalis is divided into eight morphologically identical genotypes or assemblages (A to H). Only assemblages A and B have been associated with human infections, and they are further divided into sub-assemblages: AI, AII, AIII, BIII, and BIV (Cacciò and Ryan 2008). Despite extensive efforts to associate specific assemblages to symptoms, conflicting results have been obtained, and there is to date no clear correlation between assemblage and symptoms.

Giardia, like the other diplomonads, has two nuclei and each nucleus is diploid, resulting in a tetraploid genome (Bernander et al. 2001). *G. duodenalis* has five different linear chromosomes with ribosomal DNA tandem repeats next to TAGGG telomeric repeats (Adam 2001). The study of the genome structure and architecture in *Giardia* using pulsed-field gel electrophoresis (PFGE) revealed differences in size of individual chromosomes within and between *G. duodenalis* isolates (Adam et al. 1988). The size differences were attributed to frequently recombining telomeric regions and differences in copy number of rDNA arrays (Adam 2001). Evidence of aneuploidy has been suggested in individual *Giardia* cells based on cytogenetic evidence (Tůmová et al. 2006), with the most common karyotype differing between different assemblage A and B isolates.

The genomes of six *G. duodenalis* isolates, representing three different assemblages (A, B, and E), are available to date (Adam et al. 2013; Jerlström-Hultqvist et al. 2010; Morrison et al. 2007; Ankarklev et al. 2015). The first genome to be sequenced was WB-C6 (assemblage A1), which has a haploid size of ~11.7 MB distributed over the five chromosomes (Morrison et al. 2007). The compact genome contains few introns, and promoters are short and AT rich. 6470 open reading frames (ORFs) were identified but only 4787 were later shown to be associated with transcription (Birkeland et al. 2010). Genes are placed on both DNA strands and sometimes even overlapping. Reduction of components in metabolic pathways, DNA replication, and transcription was also detected. Several genes had bacterial origin and are candidates of lateral gene transfer (Morrison et al. 2007). Variable surface proteins (VSPs) are involved in antigenic variation in *Giardia* and later analyses have shown that there are 186 unique VSP genes in the WB genome (Adam et al. 2013). Chromosome-wide maps have been established by optical mapping of the WB genome (Perry et al. 2011). The results resolved some misassemblies in the genome and indicated that the actual genome size of the WB isolate is 12.1 Mb, in close agreement with PFGE analyses. The major discrepancy was an underestimation of the size of chromosome 5, the largest of the *Giardia* chromosomes. Chromosome 5 contained an 819 kbp gap in the optical map, most likely rDNA repeats (Perry et al. 2011).

Shortly after publication of the WB genome, the genome of the GS isolate (assemblage B) was sequenced using 454 technology (Franzén et al. 2009). However, the genome was highly fragmented with 2931 contigs. 4470 ORFs were identified and the genomes show 78 % amino acid identity in protein-coding regions. The repertoire of *vsp* genes was very different compared to the WB isolate, but only 14 VSP genes were complete. The GS genome was later re-sequenced, resulting in 544 contigs and a much more complete repertoire of VSPs, totaling 275 genes (Morrison et al. 2007). Moreover, the GS genome had a much higher level of allelic sequence heterozygosity (ASH) compared to WB (0.5 % versus 0.01 %). ASH was

distributed differently into low and high ASH regions over the GS genomic contigs (Franzén et al. 2009).

The third genome represents the first isolate to be sequenced that was not obtained from a human host. The P15 isolate originates from a symptomatic pig (piglet no. 15) and belongs to assemblage E (Jerlström-Hultqvist et al. 2010). Assemblage E has been found to be more closely related to assemblage A than to assemblage B (Cacciò and Ryan 2008), and the identity of protein-coding sequences was 90 % between P15 and WB and 81 % between P15 and GS (Jerlström-Hultqvist et al. 2010), consistent with earlier results. Obtaining the sequence of three phylogenetically distinct *Giardia* groups (WB, P15, and GS) made it possible to assign lineage specificity to the genes identified in the three genomes. 91 % of the genes (~4500 protein-encoding genes) were found to be present in all three *Giardia* genomes (three-way orthologs). The rest of the genes (9 %) are variable, belonging mostly to four large gene families (the variant-specific surface proteins (VSP), NEK kinases, Protein 21.1, and high-cysteine membrane proteins (HCMPs)). The highest number of isolate-specific genes (38) was found in the P15 isolate, followed by GS (31) and WB (5). The P15 and GS isolates shared 20 proteins to the exclusion of WB, with 13 of these found in a cluster of 20 kbp in the P15 genome (Jerlström-Hultqvist et al. 2010). Interestingly the ORFs in this genomic cluster are not expressed in any of the conditions tested. The chromosomal architecture in *Giardia* shows core gene-rich stable regions with maintained gene order interspersed with non-syntenic regions harboring VSPs and other non-core genes. These regions often have a higher GC% and show nucleotide signatures that deviate from surrounding regions, in part due to the common occurrence of VSP and genes encoding high-cysteine membrane proteins (HCMPs) that are more GC rich than the genome on average. The level of ASH in the P15 isolate was lower than in the GS isolate, 0.0023 % (Jerlström-Hultqvist et al. 2010).

Three assemblage AII isolates have been sequenced (DH1, AS98, and AS175) (Adam et al. 2013; Ankarklev et al. 2015). The amount of genetic diversity was characterized in relation to the genome of WB, the assemblage A reference genome. The analyses showed that the divergence between AI and AII is approximately 1 %, represented by ~100,000 single nucleotide polymorphisms (SNP) distributed over the chromosomes with enrichment in the variable genomic regions containing VSPs and HCMPs (Ankarklev et al. 2015). The level of ASH in two of the AII isolates (AS98 and AS175) was found to be 0.25–0.35 %, which is 25–30-fold higher than in the WB isolate and tenfold higher than the assemblage AII isolate DH1 (0.037 %, (Ankarklev et al. 2015)).

There is a need for further genomic analyses of *Giardia* genomes. The assemblage A (WB) and B (GS) reference genomes can be improved, which will facilitate reference-based genome mapping of data from clinical and environmental isolates. More isolates from the A and B assemblages should be sequenced so that all the genetic differences between the human-infecting isolates can be identified. Genomic information from the remaining assemblages, C-D, F-H, can reveal species-specific genomic features. Sequence data from other *Giardia* species, like *Giardia muris*, will be important for further studies of the evolution of *Giardia* biology and virulence. In addition to the underlying genomic sequence and annotation, a number of functional datasets are available for the GiardiaDB.

1.3.2 Trypanosomatids

Trypanosomatids are a group of parasitic unicellular flagellate eukaryotes. Their range of hosts is diverse and includes humans as well as a wide variety of species from both the animal and plant kingdoms. Trypanosomatids belong to the Kinetoplastida, which is included in the phylum Euglenozoa, a branch that diverged early in the eukaryotic tree (Campbell et al. 2003; Simpson et al. 2006). While a number of Kinetoplastida are pathogenic parasites, most are free-living organisms found in soils and aquatic habitats. The name Kinetoplastida derives from the presence of large amounts of mitochondrial DNA, visible by light microscopy as a dense mass known as the kinetoplast with its contained DNA referred to as kDNA. Trypanosomatids are obligate parasites that can be monoxenous or dixenous (usually an insect vector and other animal or plant (Votýpka et al. 2015)).

1.3.2.1 Trypanosomatid Genomes

The nuclear genome of trypanosomatids has some unusual characteristics when compared with other eukaryotic genomes. Their genome is organized in polycistronic transcriptional units (PTUs) and the production of individual mRNAs from PTUs requires trans-splicing of a splice leader (SL) sequence (Martínez-Calvillo et al. 2010). PTUs are well conserved and exhibit a high degree of synteny between species. The kDNA has an unusual physical structure, being arranged in circles of DNA that are interlocked in a chain-mail-like network. These mitochondrial mRNAs require post-processing in the form of insertion and deletion of uridines before being translated into proteins, a process known as RNA editing (Aphasizhev and Aphasizheva 2014; Lukeš et al. 2002). Other peculiarities of trypanosomatid genomes include the almost complete lack of introns, kinetoplastid-specific histone modifications and histone variants, unique origins of replication in some genera, a special DNA base (Base J) (Maree and Patterton 2014), and the transcription of protein-coding genes by RNA pol I in African trypanosomes, a behavior unique among eukaryotes (Daniels et al. 2010). Although none of these unusual features seem to be exclusive of trypanosomatids and are also present, at least in some basic form, in other free-living kinetoplastids, they may be related to the development of parasitism in trypanosomatids (Simpson et al. 2006; Lukeš et al. 2014).

1.3.2.2 Regulation of Gene Expression in Polycistronic Transcriptional Units

One of the most striking characteristics of trypanosomatid genomes is the organization of their protein-coding genes into long polycistronic transcriptional units (PTUs) that contain tens to hundreds of genes in the same orientation. Individual mRNAs are produced from the precursor mRNA by the 5' trans-splicing of a capped mini-exon or splice leader (SL) sequence, followed by the polyadenylation of the 3' end. The 5' trans-splicing is linked to the polyadenylation of the upstream gene. Gene order within PTUs is highly conserved among trypanosomatids, and the main differences are usually in the regions between the PTUs and at the ends of the chromosomes (Martínez-Calvillo et al. 2010; Clayton 2014).

The genes included in a PTU are functionally unrelated and can be expressed at different times of the cell cycle or in different life stages. Nonetheless, each PTU is transcribed from a single transcriptional start site (TSS), severely limiting the amount of regulation that could be provided by the induction or repression of promoters. In some cases correlation between the location of a gene in a PTU and its expression level has been described. For example, in *T. brucei*, genes downregulated after heat shock tend to be closer to the transcriptions start site (TSS), while upregulated genes tend to be more distal. Also, the position of the genes along the PTUs correlates with gene regulation during the different cell cycle stages. However, most of the genes do not seem to be ordered depending on their transcriptional regulation (Campbell et al. 2003; Martínez-Calvillo et al. 2010; Kelly et al. 2012).

In most organisms, the start of transcription is a fundamental step in the regulation of gene expression. In trypanosomatids this layer is constrained, but a swift and specific regulation of gene expression is still needed. The medically relevant kinetoplastids are dixenous parasites with complex life cycles that require fast and extensive changes in morphology and metabolism. These changes depend, ultimately, on changes in gene expression. For example, the parasite has to quickly adapt to differences in temperature, energy sources, and host immune system (Daniels et al. 2010; Kelly et al. 2012). Besides the regulation at the start of transcription, it is possible to modulate other steps in the transcription and translation process. Additional levels of control include transcriptional elongation, mRNA processing (trans-splicing and polyadenylation), export from the nucleus, mRNA degradation (in the cytoplasm and nucleus), translation (start and elongation), and protein degradation (Martínez-Calvillo et al. 2010; Clayton 2014).

Both mRNA processing and the control of the mRNA stability are important regulatory steps in trypanosomatids. The stability of the mRNAs depends on elements present in the 3' UTRs, for instance, duplicated genes in tandem arrays can be differentially regulated due to differences in their 3' UTRs. In *T. brucei,* the range of half-lives of mature mRNAs is very diverse and is also determined by the life cycle stage. In addition, the half-life of a mRNA not only depends on the stability of the mature mRNA but also on the rates of destruction of the precursor mRNA. If a mRNA undergoes a late or delayed polyadenylation, it is more susceptible to being degraded, even before finishing maturation (Clayton 2014; Jackson 2015).

Trypanosomatids contain a large number of RNA-binding proteins (RBPs) that likely regulate expression levels by binding to regulatory elements in the 3' UTRs of the mRNAs. The amount of RBPs is high compared with the number of mRNAs. Consequently the current hypothesis proposes the binding of multiple RBPs to each 3' UTR, which would compete or cooperate dynamically with other RBPs. The mix of RBPs would determine the stability of the mRNA and could also modulate the translation process (Clayton 2014; Clayton 2013). The expression of protein-coding genes can also be regulated at the translational level. In ribosome profiling studies, it has been shown that there is a wide range in the density of ribosomes associated to mRNAs, with differences between life stages. In addition, trypanosome mRNAs can contain upstream open reading frames in their 5' UTRs, which decrease the translation of the main ORF (Clayton 2014; Vasquez et al. 2014; Jensen et al. 2014).

1.3.2.3 Multi-copy Families of Surface Proteins

Genome reduction is frequent in parasites with functions that are essential for a free-living organism becoming obsolete inside a host. Surprisingly, compared with other single-cell parasitic eukaryotes, trypanosomatid genomes do not appear to be specially reduced in size or function. On the contrary, in the evolution of parasitism in trypanosomatids, the gain of new competences seems to have been more important than the loss of functions (Jackson 2015). One example of this gain of functions is the presence of large multi-copy families that encode surface proteins. These families are specific to trypanosomatids and usually have a nonrandom distribution in the genome. A number of them have been implicated in pathogenesis and defense against the host immune system, such as the major surface protease (MSP) family of metalloproteases involved in pathogenesis and conserved in all trypanosomatids. Other well-known examples are the variant surface glycoprotein (VSG) and procyclin in *T. brucei*, delta-amastin and promastigote surface antigen (PSA) in *Leishmania*, and trans-sialidases in *T. cruzi* (Jackson 2015; Rogers et al. 2011; El-Sayed et al. 2005a).

1.3.2.4 Epigenetic Regulation

In eukaryotes, nuclear DNA is organized into a complex of DNA and proteins known as chromatin. The nucleosome is the basic unit of the chromatin, providing a sevenfold condensation. It comprises an octamer made of two copies of each of the core histones (H2A, H2B, H3, and H4) around which approximately 147 bp of DNA is wrapped. In addition, there is a histone (H1) in the DNA region between two nucleosomes that helps stabilize the chromatin. The chromatin is folded into a 30 nm chromatin fiber that can be further compacted, up to the level of the distinct chromosomes that can be visualized during the eukaryotic mitosis (Martínez-Calvillo et al. 2010; Maree and Patterton 2014). Although the nucleosomes are still the basic unit of chromatin in trypanosomatids, their histones are divergent from those found in yeast and vertebrates. DNA in trypanosomatids is not condensed into the 30 nm chromatin fiber nor do chromosomes condense during mitosis. However, some differences in the level of condensation between life cycle stages have been described (Martínez-Calvillo et al. 2010; Daniels et al. 2010).

Mechanisms that influence the structure of chromatin have been implicated in the regulation of gene expression. In trypanosomatids, as in other eukaryotes, specific modifications of the N-terminal tails of histones, or the presence of histone variants, correlate with regions of active or repressed transcription. As of yet, no conserved sequences have been identified in the transcription start sites (TSSs) of the PTUs. It has been proposed that TSSs could be determined by chromatin structure rather than the presence of conserved sequence motifs. Some of the histone modifications described in trypanosomatids are common in eukaryotes, but there are also some modifications and histone variations specific to trypanosomatids, such as H3V and H4V (probable markers of transcription termination sites) (Maree and Patterton 2014; Siegel et al. 2009).

1.3.2.5 Mitochondrial Genome: Architecture and RNA Editing

The kDNA is made up of circles of DNA that are interlocked in a chain-mail-like network and are of two types: maxicircles and minicircles. Maxicircles store information for classical mitochondrial genes and proteins, but their transcripts require RNA editing, the insertion or deletion of uridines, before being translated. Minicircles encode guide RNAs (gRNAs) that act as templates during the editing process. Unlike other eukaryotes, mitochondrial tRNAs are found in the nuclear genome and require specific target sequences to be transported into the mitochondria (Campbell et al. 2003; Lukeš et al. 2002). The mitochondrial genome contains a few dozens of maxicircles, with identical sequence and a size of 20–40 kb, and thousands of minicircles. Minicircles differ in sequence content, but their size is species specific and uniform, usually between 0.5 and 10 kb. Maxicircles are concatenated together and simultaneously interlinked with the minicircle network. The DNA network and associated proteins are organized in a dense disk visible by light microscopy. While all kinetoplastids contain maxi- and minicircles, the concatenated network is unique to trypanosomatids (Simpson et al. 2006; Marande et al. 2005). During the RNA editing, uridines are inserted or deleted from mitochondrial mRNAs fixing errors in the sequence and restoring a viable coding sequence. The sequences to be used as templates are stored in the gRNAs (50–60 nt). These encode only a small portion of the information needed to repair a mRNA; therefore, multiple gRNAs are required to edit each mRNA. RNA editing is catalyzed by the RNA editing core complex or editosome. Several modules can combine to build different versions of the editosome, each with different specificities (Aphasizhev and Aphasizheva 2014; Liu et al. 2005).

1.3.2.6 Base J

Base J, or the modification of thymine to beta-D-glucosyl-hydroxymethyluracil, is enriched at the ends of PTUs, at potential transcription terminal sites (TTSs) and in repetitive DNA elements, such as the telomeric repeats (Campbell et al. 2003; Maree and Patterton 2014).

1.3.2.7 Transposable Elements

The two main classes of transposable element are DNA transposons and RNA retrotransposons. DNA transposons move by "cut and paste" and depend on a DNA intermediate, while RNA retrotransposons use a "copy and paste" strategy, with a RNA intermediate. DNA transposons have not been found in trypanosomatid genomes, but RNA retrotransposons have been shown to be present. For example, several classes of potentially active retrotransposons have been identified in *T. brucei* and *T. cruzi*, some of which could be involved in the regulation of gene expression, such as SIDER2, which localizes to the 3'UTRs of mRNAs and affects its stability (Martínez-Calvillo et al. 2010; Bringaud et al. 2006; Bringaud et al. 2008).

1.3.2.8 Sequenced Genomes

The first trypanosomatids sequenced were *Trypanosoma brucei*, *Trypanosoma cruzi*, and *Leishmania major*, the causative agents of sleeping sickness, Chagas disease, and one form of cutaneous leishmaniasis in humans (Berriman et al. 2005; Ivens

et al. 2005; El-Sayed et al. 2005b). Since then, the genomes of other medically relevant trypanosomes have been published. *Leishmania* species that have been sequenced include *L. donovani* (Downing et al. 2011), *L. infantum*, *L. braziliensis* (Peacock et al. 2007), *L. mexicana* (Rogers et al. 2011), *L. panamensis* (Llanes et al. 2015), *L. peruviana* (Valdivia et al. 2015), and *L. amazonensis* (Real et al. 2013). *Trypanosoma* species include *T. rangeli* (Stoco et al. 2014). Apart from the reference genomes, multiple strains and hundreds of isolates have been sequenced and are available in the databases [TriTrypDB, NCBI]. The range of published genomes has expanded to other dixenous species and includes parasites of reptiles (*Trypanosoma grayi* (Kelly et al. 2014) and *Leishmania tarentolae* (Raymond et al. 2012)), parasites of livestock (*T. evansi* (Global 2015)), or parasites of plants (*Phytomonas serpens*, *Phytomonas* spp. (Kořený et al. 2012; Porcel et al. 2014)). In addition, the genomes of a few monoxenous trypanosomatids have been published (*Leptomonas seymouri* (Kraeva et al. 2015) and *Lotmaria passim* (Runckel et al. 2014)). Some of these species harbor symbiotic bacteria and have been used as a model to study the evolution of organelles (*Crithidia acanthocephali*, *Herpetomonas muscarum*, *Strigomonas oncopelti*, *Strigomonas galati*, *Strigomonas culicis*, *Angomonas desouzai*, and *Angomonas deanei*) (Alves et al. 2013). Additional genomes are available prepublication in the genome databases (TriTrypDB, NCBI) and include *Endotrypanum monterogeii*, *Leptomonas pyrrhocoris*, and *Crithidia fasciculata*; the *Leishmanias L. aethiopica*, *L. tropica*, *L. gerbilli*, *L. enrietti*, and *L. turanica*, and the *Trypanosomas T. congolense* and *T. vivax*.

1.3.3 *Toxoplasma* and Related Organisms

Toxoplasma gondii is a member of the tissue cyst-forming coccidian parasites, which include *Neospora caninum*, *Hammondia hammondi*, and *Sarcocystis* spp., among others (Dubey and Ferguson 2014; Dubey and Lindsay 1996; Levine 1986; Dubey et al. 1998). Of these *T. gondii* appears to be the most widely distributed both geographically and by host diversity and is able to infect virtually any warm-blooded animal. While the diversity of *T. gondii* is restricted to three clonal lineages in Europe and North America, isolates from the southern hemisphere exhibit much wider genetic variability (Sibley and Ajioka 2008). Amazingly, while *T. gondii* can infect a wide variety of warm-blooded organisms, it can only undergo sexual recombination in Felidae. Cats shed infective sporozoites containing environmentally resistant cysts, which can be transmitted orally to other organisms such as rodents (Dubey et al. 1998). Following oral infection, sporozoites cross the small intestine and can infect a variety of cells where they undergo a developmental switch to fast-growing tachyzoites (Sibley and Ajioka 2008). Tachyzoites replicate through a process called endodyogeny where two daughter cells are formed within a mother cell by a combination of de novo building of cytoskeletal and secretory components, replication and segregation of mother cell components (i.e., nucleus, mitochondria, and apicoplast), and recycling of mother cell components (Francia and Striepen 2014; Ouologuem and Roos 2014). Pressure from the host immune system forces

tachyzoites to undergo another developmental change into bradyzoites (Miller et al. 2009). These semi-quiescent cells form clusters called tissue cysts that settle in brain and/or muscle tissue where they may remain for the life of the host, although reactivation of bradyzoites can occur in immunocompromised individuals. Bradyzoites also serve as a reservoir of transmission if an infected host is eaten by another animal. Interestingly, the tissue cyst tropism varies markedly between hosts. The fast replicating tachyzoite stage is often asymptomatic but can cause acute morbidity or mortality in immunocompromised individuals. Placental transmission is known to cause fetal mortality or serious congenital defects.

T. gondii contains a ~65 Mb nuclear genome comprising 14 chromosomes (Khan et al. 2005; Reid et al. 2012; Lorenzi et al. 2016), a 35 Kb apicoplast genome (Köhler et al. 1997), and a mitochondrial genome. *T. gondii* genomic-scale data such as expressed sequenced tags, sequenced BAC clones, and whole-genome shotgun sequencing were first made available through ToxoDB beginning in 2001 (Kissinger et al. 2003). Since then, additional genomic-scale data have been generated including genome sequence and transcriptomic data from a large-scale population sequencing project (Lorenzi et al. 2016). The genome of the closely related *H. hammondi* and *N. caninum* is ~65 Mb and ~62 Mb in size, respectively, and not surprisingly also comprises 14 chromosomes each (Table 1.1) (Reid et al. 2012; Lorenzi et al. 2016; Walzer et al. 2013). The genome of the more divergent *S. neurona* is almost twice the size of those previously described at ~130 Mb, while a GC content of roughly 53 % is common across this group (Table 1.1) (Blazejewski et al. 2015). A high degree of genomic synteny is observed between *T. gondii*, *H. hammondi*, and *N. caninum.* This level of synteny is not maintained with between this group and *S. neurona* (Reid et al. 2012; Lorenzi et al. 2016; Blazejewski et al. 2015).

Apicomplexan parasites in general have evolved secretory systems that transport effector molecules into their host cells. These have a range of functions, including modification of the intracellular environment, promotion of immune evasion, and modulation of host-cell transcription (Hakimi and Bougdour 2015). Most information about secretory effectors in coccidian parasites comes from *T. gondii* where numerous studies have defined dense granule (Mercier and Cesbron-Delauw 2015), rhoptry (Boothroyd and Dubremetz 2008), microneme (Carruthers and Tomley

Table 1.1 Basic genome statistics for *T. gondii* and related organisms

	Toxoplasma gondii[a]	*Hammondia hammondi* H.H.34	*Neospora caninum* Liverpool	*Sarcocystis neurona*[b]
Genome size (Mb)	63	65	62	128
No. of chromosomes	14	14	14	ND
No. of genes	8707	8176	7266	7140
% of genes with introns	76	76	77	81

ND not determined

[a]Average statistics from three strains: ME49, VEG, and GT1

[b]Average statistics from two strains: SN1 and SN3

2008), and SAG1-related sequence (SRS) proteins (Wasmuth et al. 2012). Comparative genomic analysis revealed that one of the primary features differentiating both different species of coccidian parasite and different strains of *T. gondii* is sequence diversity and copy number variation (CNV) at secretory effector loci (Reid et al. 2012; Lorenzi et al. 2016; Walzer et al. 2013). A comparison of 62 isolates of *T. gondii* and one isolate of *H. hammondi* showed that secretory effectors are often found in genomic regions exhibiting tandem amplification (Lorenzi et al. 2016). A comparison of reference isolates from the 16 major *Toxoplasma* haplogroups showed that all possess a repertoire of secretory effectors with most diversity occurring in rhoptry and SRS genes. Further comparison of secretory effectors between *T. gondii*, *H. hammondi*, and *N. caninum* revealed additional diversity and a *T. gondii*-specific family (*Tg*FAMs) of effectors, which may be important for host range and definitive host preferences (Lorenzi et al. 2016). Interestingly, a number of the *Tg*FAMs are clustered in telomeric regions and contain a variable region, which may implicate them in immune evasion (Lorenzi et al. 2016; Dalmasso et al. 2014), but they also may play a role during sexual development since many are expressed in the cat and in oocysts (Behnke et al. 2014).

1.3.4 *Cryptosporidium*

Cryptosporidium spp. are protozoan parasites with significant impact to the health of humans and livestock. They infect the intestinal and gastric epithelium of a variety of vertebrates, causing a disease known as cryptosporidiosis. Human cryptosporidiosis is responsible for diarrhea-induced death of young children in developing countries, and in immune-compromised adults, it constitutes an acute, usually self-limiting, diarrheal illness that results in significant morbidity and sometimes death. A recent study found *Cryptosporidium* to be the second leading cause of moderate-to-severe diarrhea in developing countries, with diarrheal diseases being the second principal cause of death among children under 5, globally (Kotloff et al. 2013).

There are no licensed vaccines against *Cryptosporidium*, and the only FDA-approved drug (nitazoxanide) is only effective in immunocompetent patients. Thus, the development of alternative therapeutic agents and vaccines against this disease is urgently required and remains a high public health priority. The lack of a practical and reproducible axenic in vitro culture system for *Cryptosporidium* is a major limitation to the development of specific anti-cryptosporidial vaccines (Arrowood 2002; Karanis and Aldeyarbi 2011). Advances in next-generation sequencing technologies and in genome assembly and annotation methodologies (Niedringhaus et al. 2011; Nagarajan and Pop 2013; Martin and Wang 2011; Yandell and Ence 2012) have facilitated the generation of -omics data for *Cryptosporidium*, with genomics resources now available for multiple *Cryptosporidium* species (Table 1.2, (Heiges et al. 2006)). These developments prompted a shift to in silico studies aiming to identify a wide pool of potential vaccine targets, to be further filtered according to properties common to antigens (Manque et al. 2011). This approach is similar to reverse vaccinology studies that have led to licensed vaccines in other organisms

Table 1.2 *Cryptosporidium* species with completed or draft genomes

Species	Number of draft genomes	Natural host range	Predilection site
C. hominis	8	Human, primates	Intestinal
C. parvum	8	Human, bovine	Intestinal
C. meleagridis	1	Various vertebrates	Intestinal
C. baileyi	1	Birds	Respiratory
C. muris	1	Rodents	Gastric
C. sp. chipmunk LX-2015	1	Rodents, human	Intestinal

(Donati and Rappuoli 2013; Kelly and Rappuoli 2005) and is particularly promising in organisms that, like *Cryptosporidium*, are difficult to cultivate continuously in the laboratory.

Apart from human, *Cryptosporidium* species infect other vertebrates including fish, birds, and rodents, and some species are capable of zoonotic transmission (Xiao and Herd 1994; Bouzid et al.). Some have a somewhat restricted host range, such as *Cryptosporidium hominis*, a human parasite that infects the small intestine; *Cryptosporidium muris*, a gastric parasite of rodents; and *Cryptosporidium baileyi*, an avian parasite. *Cryptosporidium parvum* and *Cryptosporidium meleagridis* have a wider host range and are known to infect both avian and mammalian species, including humans. *C. parvum* and *C. hominis* are considered class B agent of bioterrorism and are significant causes of gastrointestinal infections worldwide.

1.3.4.1 *Cryptosporidium* Genomic Resources

Cryptosporidium genomes are compact, with >75 % consisting of protein-coding sequences, and have an average size of approximately 8.5–9.5 mega base pairs (Mbp), and each encodes ~4000 genes (Table 1.3). *C. parvum* (isolate Iowa II) was the first species for which a genome was published (Abrahamsen et al. 2004). The genome was found to be 9.1 Mbp in length, assembled into 13 supercontigs. Pulsed-field gel electrophoresis studies had shown the nuclear-encoded genome to consist of eight chromosomes, and therefore the assembly includes five unresolved gaps. About 5 % of the 3807 predicted protein-coding genes in this assembly contained introns, and the average gene length was 1795 base pairs (bp). At about the same time, the genome of *C. hominis* (isolate TU502) was published (Xu et al. 2004). Since the two species were known to be closely related, with about 95–97 % DNA sequence identity between them, the *C. hominis* genome was sequenced to a much lower depth of coverage. The primary goal was to identify differences relative to *C. parvum*, rather than reconstruct a gold-standard genome assembly. Consequently, this assembly is much more fragmented, with the likely eight chromosomes split among 1413 contigs, which are grouped into ~240 scaffolds.

There were some fundamental differences between the annotated gene sets in the two species. The average gene length of *C. hominis* was 1360 bp, about 500 bp less than that of *C. parvum*, and about 5–20 % of the *C. hominis* genes were predicted to contain introns, compared to 5 % in *C. parvum* (Abrahamsen et al. 2004; Widmer

Table 1.3 Genome statistics for representative *Cryptosporidium* species

Species	Isolate	GenBank accession	Assembly length (Mb)	No. of contigs	Largest contig (bp)	No. of protein-coding genes	Average gene length (bp)	Percent coding (%)
C. hominis	TU502 (2004)	AAEL00000000	8.7	1413	90,444	3886	1360	60.4
C. hominis	TU502_new (2014)	SUB482083	9.1	120	1,270,815	3745	1845	75.8
C. parvum	Iowa	AAEE00000000	9.1	13	1,278,458	3807	1795	75.3
C. parvum[a]	Iowa	AAEE00000000	9.1	13	1,278,458	3865	1783	75.7
C. meleagridis	UKMEL1	SUB482042	9.0	57	732,862	4326	1861	89.7
C. baileyi	TAMU-09Q1	SUB482078	8.5	153	702,637	3700	1776	77.3
C. muris	RN66	AAZY02000000	9.2	84	1,324,930	3934	1780	79.2

[a]2015 re-annotation