Review Article

Gene Expression in Trypanosomatid Parasites

Santiago Martínez-Calvillo,1 Juan C. Vizuet-de-Rueda,1 Luis E. Florencio-Martínez,1 Rebeca G. Manning-Cela,2 and Elisa E. Figueroa-Angulo1

1 Unidad de Biomedicina, Facultad de Estudios Superiores Iztacala, Universidad Nacional Autónoma de México, Av. De los Barrios 1, Col. Los Reyes Iztacala, Tlalnepantla, Edo. de México, CP 54090, Mexico
2 Departamento de Biomedicina Molecular, Centro de Investigación y de Estudios Avanzados del IPN, Apartado Postal 14-740, 07360 México, D.F., Mexico

Correspondence should be addressed to Santiago Martínez-Calvillo, scalv@campus.iztacala.unam.mx

Received 19 August 2009; Accepted 4 November 2009

Academic Editor: Luis I. Terrazas

Copyright © 2010 Santiago Martínez-Calvillo et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The parasites Leishmania spp., Trypanosoma brucei, and Trypanosoma cruzi are the trypanosomatid protozoa that cause the deadly human diseases leishmaniasis, African sleeping sickness, and Chagas disease, respectively. These organisms possess unique mechanisms for gene expression such as constitutive polycistronic transcription of protein-coding genes and trans-splicing. Little is known about either the DNA sequences or the proteins that are involved in the initiation and termination of transcription in trypanosomatids. In silico analyses of the genome databases of these parasites led to the identification of a small number of proteins involved in gene expression. However, functional studies have revealed that trypanosomatids have more general transcription factors than originally estimated. Many posttranslational histone modifications, histone variants, and chromatin modifying enzymes have been identified in trypanosomatids, and recent genome-wide studies showed that epigenetic regulation might play a very important role in gene expression in this group of parasites. Here, we review and comment on the most recent findings related to transcription initiation and termination in trypanosomatid protozoa.

1. Introduction

The process by which an RNA molecule is synthesized from a DNA template is known as transcription. All cells must constantly produce RNA molecules that are directly or indirectly involved in life processes like reproduction, growth, repair, and regulation of metabolism. Eukaryotic cells have three distinct classes of nuclear RNA polymerases (Pol): Pol I, II, and III. Each class of polymerase is responsible for the synthesis of a different kind of RNA. Pol I is involved in the production of 18S, 5.8S and 28S ribosomal RNAs (rRNAs), and Pol II participates in the generation of messenger RNAs (mRNAs) and most of the small nuclear RNAs (snRNAs). Pol III synthesizes small essential RNAs, such as transfer RNAs (tRNAs), 5S rRNA and some snRNAs. Most organisms control the expression of their genes at the level of transcription initiation. However, the regulation of gene expression can also be achieved during either transcription elongation (at the level of chromatin structure), RNA processing, RNA stability or transport, or translation. A large number of transcription factors help the RNA polymerases produce RNA. Studies on gene expression in eukaryotes have focused mainly on animals, fungi and plants; whereas a relatively small amount of information is available for parasitic protozoa.

The flagellated protozoa Leishmania, Trypanosoma brucei and Trypanosoma cruzi are trypanosomatid parasites (from the order Kinetoplastida) that produce devastating human diseases. Together, these pathogens cause millions of deaths in developing countries (in the tropical and subtropical regions of the world). They exhibit complex life cycles, with different developmental stages that alternate between vertebrate and invertebrate hosts. Leishmania species cause a spectrum of diseases, known as leishmaniasis, which range from self-resolving skin ulcers to lethal infections of the internal organs [1]. The World Health Organization (WHO) has estimated that there are over two million new cases of leishmaniasis each year in the world, with 367 million
people at risk. The infection with *Leishmania* starts with the introduction of the infective form, the metacyclic promastigote, into the skin by the bite of an infected sandfly. Once inside the mammalian host, the infective promastigotes invade the macrophages and differentiate into amastigotes, which are the proliferative forms within the vertebrate host. In the insect vector, the parasite replicates as a non-infective procyclic promastigote [2]. *T. brucei*, the African trypanosome, is the causative agent of sleeping sickness in humans and nagana in animals. Approximately 500,000 people, in the least developed countries of Central Africa, are affected by the disease every year. The parasite is transmitted among mammalian hosts by the tsetse fly. The procyclic form of *T. brucei* multiplies in the gut of the insect vector and differentiates into a bloodstream form that is found in the blood and tissue fluids of mammalian hosts. *T. brucei*, unlike *T. cruzi* and *Leishmania*, does not present any intracellular forms [3]. *T. cruzi* is the etiological agent of Chagas disease, which affects several million people in Latin America. It is normally transmitted by reduvid insects via the vector feces. The parasite replicates as an epimastigote in the midgut of the insect, and transforms into an infective metacyclic trypanosomatid in the hindgut. Amastigotes are the proliferative form in the vertebrate host [4]. Trypanosomatids have also attracted the attention of molecular biologists because they possess unique mechanisms for gene expression, such as polycistronic transcription, trans-splicing, the involvement of Pol I in the synthesis of mRNA and RNA editing [5–10]. This work will review the current knowledge on Pol I in the synthesis of mRNA and RNA editing.

2. Organization of the Nuclear Genome

The 32.8 megabases (Mb) of DNA constituting the nuclear genome from *L. major* is distributed among 36 relatively small chromosomes that range from 0.28 to 2.8 Mb in size [11, 12]. *T. cruzi* possesses a genome of 60.3 Mb organized into 41 small chromosomes [13, 14], whereas *T. brucei* (genome of 26 Mb) has 11 large chromosomes [15, 16]. The genomes of trypanosomatids are organized into polycistronic gene clusters (PGCs), that is, tens-to-hundreds of protein-coding genes arranged sequentially on the same strand of DNA (Figure 1). This unusual gene organization was first observed on *L. major* chromosome 1 (the first entirely sequenced chromosome in trypanosomatids), which contains 85 genes organized into two divergent PGCs, with the first 32 genes clustered on the bottom strand and the remaining 53 genes grouped on the top strand [17]. The publication of the complete genomes for *L. major* [12], *T. brucei* [14] and *T. cruzi* [16], showed that the majority of genes in all the trypanosomatid chromosomes are organized into large PGCs. Tandem arrays of rRNA genes are present between PGCs. Most tRNA genes are organized into clusters of 2 to 10 genes, on either top or bottom strand, which may contain other Pol III-transcribed genes; most of the clusters are located at the boundaries of PGCs [12, 18]. In contrast to other organisms, the distribution of tRNA genes in the genomes of *L. major* and *T. brucei* does not seem to be totally random, as these genes are confined to a subset of chromosomes [12, 18]. The 55 rRNA genes in *T. brucei* and *T. cruzi* are organized into tandem arrays [19, 20], whereas in *L. major* they are dispersed throughout the genome and are always associated with tRNA genes [12]. Despite the fact that these species diverged more than 200 million years ago, the strong conservation of gene order (synteny) observed in the genomes of trypanosomatids for protein-coding genes is remarkable [21]. In contrast, the majority of the tRNA clusters do not show synteny [18]. Also, the vast majority of protein-coding genes in trypanosomatids lack introns; in fact, cis-splicing has only been demonstrated for the gene encoding the poly(A) polymerase [22]. Similarly, only one isotype of tRNA genes, tRNA-Tyr, contains an intron in trypanosomatids [18, 23]. These organisms are diploid, even though some chromosomes are aneuploid [24]. In addition, the ends of the chromosomes in trypanosomatids contain the telomeric repeat GGTTA, while the subtelomeric regions are composed of variable repetitive elements, which are responsible for a major part of the size polymorphisms observed between homologous chromosomes [25, 26].

3. Processing of mRNA

Unlike the majority of eukaryotic organisms, transcription in trypanosomatids is polycistronic (Figure 1) [27–29]. Most chromosomes contain at least two PGCs, which can be either divergently transcribed (towards the telomeres) or convergently transcribed (away from the telomeres). Genes from a polycistronic unit in trypanosomatids generally do not code for functionally related proteins [7]. This is entirely different from how operons function in bacteria and nematodes. Mature nuclear mRNAs are generated from primary transcripts by trans-splicing and polyadenylation (Figure 1) [6]. Trans-splicing is a process that adds a capped 39-nucleotide minixxon or spliced leader (SL) to the 5’ termini of the mRNAs [30, 31]. Like cis-splicing, trans-splicing occurs via two transesterification reactions, but it involves the formation of a Y structure instead of a lariat intermediate [32]. An AG dinucleotide at the 3’ splice site and an upstream pyrimidine-rich region are the most conserved sequences required for this process [33–35]. The trans-splicing and polyadenylation of adjacent genes are apparently linked, as the selection of a splice site for a gene influences the choice of a polyadenylation site for the upstream gene [36]. All the genes that are part of a PGC are transcribed at the same level, as a consequence of polycistronic transcription. However, the mature mRNAs of adjacent genes might show very different concentrations and/or stage-specific expression. This is because gene expression in trypanosomatids is mainly regulated posttranscriptionally at the level of mRNA processing and stability [6, 37]. Sequences in the 3’ untranslated region (3’-UTR) of an mRNA play a key role in gene expression. For example, the
Figure 1: Transcription and processing of mRNAs in trypanosomatids. The top part of the figure represents a hypothetical chromosome with three polycistronic gene clusters (PGC1-3). Pol II transcription initiates upstream of the first gene of the PGCs (arrows) [29, 38]. The G-run usually present at divergent strand-switch regions (SSR) is indicated. Nucleosomes located at the vicinity of transcription initiation regions contain histone variants H2AZ and H2BV [39]. The N-terminal tail of histone H3 in such nucleosomes are acetylated at K9/K14 (labeled as K9ac in the figure) and tri-methylated at K4 (K4me) [40, 41]. The N-terminal tail of histone H4 is acetylated at K10 (K10ac) [39], and at K5/K8/K12/K16 [41] (not shown in the figure). The bromodomain factor BDF3 [39], and transcription factors TRF4 and SNAP50 [40] also bind at transcription initiation regions. Pol II transcription of some PGCs terminates near tRNA genes (convergent strand-switch region between PGC2 and PGC3) [38], in regions of DNA that contain nucleosomes with histone variants H3V and H4V [39]. Transcription of a PGC produces a primary transcript (shown only for PGC2) that is processed by trans-splicing and polyadenylation to generate the mature mRNAs. By trans-splicing, a capped SL RNA (yellow box) is added to the 5′ end of every mRNA. In the spliced leader locus (located in a different chromosome) each gene possesses a Pol II promoter region (arrows). The cap in the SL RNA is indicated by an asterisk at the 5′ end of the RNA. The polycistronic transcript contains pyrimidine-rich regions (indicated by a striped box in the intergenic regions) that are needed for both trans-splicing and polyadenylation. The pyrimidine-rich regions are also present in the DNA, but they are not shown to simplify the figure. The four As located at the 3′ end of the mature mRNAs represent the poly-A tail.

3′-UTR from the amastin mRNA in L. infantum has a 450 bp region that confers amastigote-specific gene expression by a mechanism that increases the mRNA translation [42]. Also, the mRNA of the phosphoglycerate kinase PGKB in T. brucei contains a regulatory AU-rich element in the 3′-UTR that destabilizes the mRNA in bloodstream forms but not in procyclic forms [43]. Moreover, the 3′-UTR region of the EP procyclin mRNA contains 16-mer and 26-mer elements that contribute to mRNA stability and translation efficiency [44, 45]. Interestingly, a growing number of reports have shown that mRNA and protein abundance do not always correlate, and that translational and post-translational control play an important role in trypanosomatids [46–48].

4. Promoter Regions

4.1. Initiation of Transcription by Pol II. Precise transcription initiation of eukaryotic genes is controlled by a segment of DNA, the promoter region, which includes the transcription start site (TSS, +1) and the immediate flanking sequences. Promoter regions for Pol II typically comprise about 40 bases and contain functional subregions called core promoter elements. These elements include the TATA box, the initiator (Inr) and the downstream promoter element (DPE). The core promoter elements direct the recruitment and assembly of the preinitiation complex (PIC), which is composed of Pol II and the general transcription factors TFIIA, TFIIB, TFII D, TFIIE, TFIIF and TFIH [49, 50].

Presently, the only Pol II promoter that has been extensively characterized on trypanosomatids is the one driving the expression of the SL RNA [51–53]. In Leishmania tarentolae it consists of two domains: the −60 element (from −67 to −58, relative to the TSS) and the −30 element (from −41 to −31) (Figure 2). In Leptomonas seymouri (a trypanosomatid that infects insects), an initiator sequence at the TSS is additionally required for Pol II to synthesize the SL RNA.
Identification of the Pol II sequences that direct the expression of protein-coding genes in trypanosomatids has proven to be an elusive aim, complicated by factors such as relatively low transcriptional activity and rapid processing of the primary transcripts. However, a run-on analysis of chromosome 1 from \textit{L. major} showed that Pol II transcription of the entire chromosome initiates in the strand-switch region (between the two divergent PGCs) and proceeds bidirectionally towards the telomeres [29]. Several TSSs were mapped for both PGCs within a \(<100\)-bp region that contains long G-tracts (or C-tracts), but do not contain a TATA box or any other typical Pol II elements that reside exclusively upstream of the coding region [63]. More than 80% of human protein-coding genes are driven by TATA-less promoters [54, 55] that possess several transcription initiation sites spread over 50–100 bp [56, 57]. While promoters with a TATA box and a unique TSS are generally related to tissue-specific expression, TATA-less promoters usually occur within a CpG island and drive expression of ubiquitously expressed genes. \textit{In vitro} transcription analyses showed that transcription from TATA-less promoters does not require the complete TFIID (TBP and about 15 TAF subunits) and instead requires only TBP [55, 58]. It is possible that other components of the PIC are not required either. This indicates that TATA-less promoters may need simpler initiation complexes. Also, recent studies indicate that bidirectional promoters are common in the human genome [59]. They normally lack TATA box sequences, contain several TSSs and have a high GC content. There is evidence of GC anisotropy (more guanines on the plus strand) in the region around the major TSSs; it is believed that the predominance of guanines on the plus strand could contribute to promoter orientation [56, 60]. Therefore, these genome-wide studies indicate that TATA-less promoters in humans and other mammals share several characteristics with the strand-switch region from \textit{L. major} chromosome 1 (and probably with all the strand-switch regions in trypanosomatids): they lack a TATA box and have an array of closely located TSSs that span over 50–100 bp; they contain G and C tracts that might direct bidirectional transcription; and they might direct constitutive transcription. Interestingly, some genes with TATA-less promoters in human and mouse have weak TSSs scattered over the majority of exons and 3′-UTRs [56]. The function of the resulting RNAs is uncertain at present. Likewise, the transcriptional analysis of \textit{L. major} chromosome 1 indicated that low levels of nonspecific transcription (10 times lower than transcription starting at the strand-switch area) probably take place all along the chromosome [29].

4.2. Pol III Promoters. An unusual feature of Pol III promoter regions is that most of them require sequence elements located downstream of the TSS, within the transcribed region [61]. The majority of Pol III promoters fall into one of three different categories, depending on the location and type of cis-acting elements. Type I promoters, characteristic of 5S rRNA genes, consist of three internal domains: Box A, an intermediate element (IE), and box C. These elements span a region of approximately 50 bp starting at position +45. Type II promoters are present on tRNA genes and consist of two conserved internal elements: boxes A and B. While box A is normally positioned close to the TSS, the location of box B is variable, partly because some tRNAs have short introns within their coding regions. Type III promoters, characteristic of U6 snRNA genes, consist of elements that reside exclusively upstream of the coding sequence. They contain a TATA box near position −30, a proximal sequence element (PSE) at position −60, and a distal sequence element (DSE) further upstream of the TSS [62, 63].

Pol III synthesizes all the snRNAs in trypanosomatids, in addition to the 5S rRNA, 7SL RNA and tRNAs. The snRNA- and 7SL-gene promoters have been characterized in \textit{T. brucei} and \textit{Leptomonas} [64, 65]. Interestingly, these genes have a divergently oriented tRNA gene in their 5′-flanking region, and boxes A and B from the neighboring tRNA genes are essential for expression of the 7SL RNA and snRNAs [66] (Figure 2). Some of the snRNA genes also require intragenic regulatory elements to achieve an optimal level of expression. Boxes A and B are also needed for expression of the tRNA itself [64]. Gel-retardation assays on \textit{T. brucei} tRNA genes showed the specific binding of nuclear proteins to both boxes [67]. 5S rRNA genes in trypanosomatids contain boxes A and C and IE sequences (Figure 2) [19, 20, 68], but they have not been functionally characterized.

4.3. Transcription by Pol I. In most eukaryotes, the genes encoding the 18S, 5.8S, and 28S rRNA molecules occur as tandem repeats that are clustered at one or several loci. rDNA repeat units are separated from each other by an intergenic spacer [63]. Each cell may contain 100 to several thousand rDNA units. Transcription of the rDNA unit by
Figure 2: Schematic representation of promoter regions in trypanosomatids. Transcription start sites are indicated by arrows. DNA sequence elements required for transcription initiation are denoted by dark blue boxes. Coding regions are shown as light blue boxes. The spliced leader RNA and chromosome 1 (transcribed by Pol II) correspond to L. tarentolae and L. major, respectively. The tRNA/U6snRNA locus and the 5S rRNA (transcribed by Pol III) correspond to T. brucei. The VSG, procyclin and rRNA transcription units (transcribed by Pol I) also correspond to T. brucei. The spliced leader RNA promoter contains the −30 and −60 elements [51]. In the divergent strand-switch region (SSR) from chromosome 1 several TSSs were mapped, in a ∼70 bp region, for both PGCs [29]. The tRNA gene possesses boxes A and B and the U6 snRNA an initiator element (Inr) [66]. The 5S rRNA gene contains boxes A and C and an intermediate element (IE). Domains I and II regulate transcription of the VSG expression site [69]. The VSG gene is located ∼47 kb downstream of the expression-site-associated-gene 7 (ESAG7). The procyclin promoter consists of four domains (I–IV) [70]. Procyclin genes EP1 and EP2 are followed by procyclin-associated genes (PAGs) 1 and 5. The rRNA promoter contains a bipartite core element (domains I and II), a distal element (domain III) and an upstream control region (UCR) [71]. Coding regions of the 18S, 5.8S and 24Sα rRNA genes are shown. Maps are drawn to scale. Please note that the scales are different upstream and downstream of the TSS in the VSG, procyclin and rRNA transcription units.
Pol I produces a large precursor (35–45S) that includes both internal and external transcribed spacers. Subsequent elimination of the transcribed spacers produces the mature rRNA molecules. The precursors of rRNA are synthesized in the fibrillar centers of the nucleoli [72]. In spite of the sequence divergence, promoters in most species have a common structural organization, because they contain two essential domains: an upstream domain located at the 5′ boundary near position −150 and a core promoter domain near the site of transcription initiation at +1. Maintenance of the correct spacing between the two domains is critical [72].

A distinctive property of rRNA genes in trypanosomatids is the fragmentation of the 28S-like rRNA into multiple independent molecules: 24Sα, 24Sβ, S1, S2, S4 and S6 [73–75]. In L. major there are two copies of the S6 gene per rRNA repeat [76]. The T. brucei rRNA promoter contains a bipartite core element (domains I and II in Figure 2) and a distal element (domain III), which resembles the typical eukaryotic Pol I promoter [77]; the promoter also contains an upstream control region (UCR) that extends to roughly −250, but it only has a minor influence on transcription efficiency [69, 71]. In Leishmania and T. cruzi the rRNA promoters are apparently smaller, as they lack upstream control elements [78–83]. Interestingly, a down regulating region of about 200 bp was located downstream of the TSS in the rRNA genes from T. cruzi [84].

Pol I not only produces rRNA in T. brucei, but also has the remarkable capacity to synthesize the mRNAs of two of the most abundant proteins in the parasite: the variant surface glycoproteins (VSG) and the EP/GPEET procyclins [85, 86]. VSGs are expressed in the bloodstream form of the parasite, and they participate in the process of antigenic variation, a survival strategy that allows the parasite to escape host cell immunological attacks [87, 88]. T. brucei has ∼1000 VSG genes, but only the one located at the active expression site (ES) is expressed. VSG switching occurs by turning off the active ES and turning on another one (in situ switch), or by replacing the VSG gene within the active ES. The replacement of the VSG gene can be achieved by copying a new VSG gene into the active ES by duplicative gene conversion, or by reciprocal translocations between two expression sites [87–91]. It was recently demonstrated that antigenic switching by gene conversion is triggered by a DNA double-stranded break within the 70-bp repeats located upstream of the transcribed VSG gene [92].

The VSG coat is replaced by a new surface coat composed of EP/GPEET procyclins when bloodstream forms differentiate into the procyclic forms found in the tsetse fly gut [93]. The expressed VSG gene is located at the end of a ∼50 kb polycistronic unit that contains up to twelve genes known as the expression-site-associated-genes (ESAGs) [5, 94] (Figure 2). The procyclin genes are organized into 5–10 kb polycistronic transcription units on chromosomes VI and X. Each procyclin locus has two procyclin genes followed by several procyclin-associated genes (PAGs) [95] (Figure 2). The promoter regions of the genes encoding procyclins are very similar to the rRNA gene promoter in T. brucei, as they are composed of four domains that extend to nucleotide −246 (Figure 2) [70, 77, 96]. In contrast, the VSG promoter contains only a bipartite core domain that extends to position −67 (Figure 2) [69, 97, 98]. Further work is required to comprehend the functional implications of the differences among Pol I promoters in T. brucei.

5. RNA Polymerase Subunits and Transcription Factors

In S. cerevisiae, Pol II contains 12 subunits, whereas Pol I and Pol III are composed of 14 and 17 subunits, respectively [49, 63]. Five of the subunits are common to all three RNA polymerases (ABC27, ABC23, ABC14.5, ABC10α and ABC10β) while two are shared between Pol I and Pol III (AC40 and AC19), with homologues in Pol II (B44 and B12.5, resp.). Another five are homologous subunits (A190/B220/C160, A135/B150/C128, A43/B16/C25, A14/B32/C17, and A12/B12.6/C11). Five subunits are exclusive to Pol III (C82, C53, C37, C34 and C31), and two subunits are Pol I-specific (A49 and A34) [61, 62, 99].

All of the common and most of the homologous RNA polymerase subunits were identified in trypanosomatids by in silico analysis; however, some of the specific subunits were not found [12]. Interestingly, trypanosomatids have two or three different genes encoding ABC10β, ABC23, and ABC27. The sequences of such genes are widely divergent, which suggests that the subunits they encode may not be common to all three RNA polymerases, as they are in other eukaryotes [12]. In fact, Pol II transcription complexes isolated by the TAP-tag protocol in T. brucei and L. major contained only one of the isoforms of ABC10β, ABC23 and ABC27 (as discussed below, isoforms ABC10βz, ABC23z and ABC27z are part of Pol I) [100, 101]. The other Pol II subunits (B220, B150, B44, B32, B16, ABC14.5, B12.6, B12.5 and ABC10α) were identified in trypanosomatids by either in silico analysis or biochemical characterization [100–102]. Another difference between trypanosomatids and other eukaryotes is that the C-terminal domain (CTD) on the largest subunit (B220) of Pol II in trypanosomatids does not contain the characteristic heptapeptide repeats [103] that are phosphorylated at specific amino acids and play important roles in the regulation of transcriptional initiation, elongation and termination in yeast and vertebrates [49].

A few general transcription factors have been identified in trypanosomatids [104, 105]. While some of them show clear sequence identity to their orthologs in yeast and vertebrates, others present a very low degree of similarity. Regarding Pol II transcription, several general transcription factors that participate in SL RNA synthesis have been identified in T. brucei. These include the TBP-related protein 4 (TRF4) [106–108], a very divergent TFIIB ortholog [109, 110] and SNAPc [107, 111]. In humans, SNAPc is essential for the synthesis of small nuclear RNAs, transcribed by either Pol II or Pol III. In T. brucei, SNAPc binds to the upstream domain of the SL RNA gene promoter and consists of three subunits (SNAP50, SNAP2 and SNAP3) [107]. Other proteins that have been identified as part of the Pol II complex that transcribes the SL RNA gene, are the two subunits of TFIIA [107, 108] and TFIIH [112, 113].
In *T. brucei*, TFIIH consists of nine different subunits, including two essential trypanosomatid-specific subunits [114]. Complex PB2 is also required for transcription of the SL RNA in *Leptomonas* [115]. Thus, these recent findings indicate that trypanosomatids possess more general transcription factors than initially estimated from *in silico* studies.

In *L. major*, Pol III transcription complexes were purified using the TAP-tag procedure with ABC23 as the target [101]. Analysis of the purified complexes revealed 12 Pol III subunits: C160, C128, C82, C53, C37, C34, C17, AC40, AC19, ABC27, ABC23 and ABC14.5. The rest of the 17 subunits, with the exception of C31, have been identified by *in silico* analysis. Eight Pol II subunits were also co-purified with TAP-tagged ABC23. However, no single Pol I-specific subunit co-eluted with ABC23, which showed that this isoform of ABC23 is restricted to Pol II and Pol III, while the other isoform (ABC23z) is limited to Pol I. This result was confirmed in *T. brucei* by co-immunoprecipitation experiments [116]. Other proteins that were purified with TAP-tagged ABC23 are: two RNA binding proteins, a putative transcription factor, the splicing factor PTS-R-1, four helicases and several proteins of unknown function [101]. BRF1 and B””, two subunits of the Pol III transcription factor TFIIB, have also been identified in trypanosomatids [12]. However, neither TFIIB nor TFIIC have been found in this group of parasites. Interestingly, ChIP-chip studies indicated that TRF4 and SNAP50 bind to all tRNA, snRNA and 5S rRNA gene clusters in *L. major* [40].

Analysis of the Pol I complex in *T. brucei* led to the identification of ten subunits: A190, A135, A12, ABC23z, ABC27z, ABC14.5, ABC10βz, ABC10α, AC19 and AC40. These subunits were identified by *in silico* analysis and by TAP-tagging A12 [117], A190 [116] and ABC23z [118]. It was recently shown that B16 (RPB7), a Pol II-specific subunit, is also associated with Pol I in *T. brucei* [119]. Also, a novel trypanosomatid-specific Pol I subunit (p31) was identified in *T. brucei* [118].

Several transcription factors take part in the synthesis of rRNAs in vertebrates and yeast, including UBF and SL1, which interact with each other in the promoter region to allow the binding of Pol I to the initiation complex. UBF is not present in yeast, which have a different factor called UAF. Another protein involved in Pol I transcription initiation is RRN3 (also known as TIF-IA) [63, 72]. None of these transcription factors have been identified in trypanosomatids. However, a new Pol I protein, named class I transcription factor A (CITFA), was recently purified and characterized in *T. brucei* [120]. It consists of a dynin light chain and six proteins that are conserved only among trypanosomatids. CITFA specifically binds to VSG, procyclin and tRNA promoters [120]. Unexpectedly, it was demonstrated in *L. major* that TFR4 and SNAP50 bind to the tRNA coding regions, but not to the promoter sequences [40].

### 6. Transcription Termination

Termination of transcription is a process that has received little attention in trypanosomatids. It has been reported that a T tract, similar to the T-rich regions that are involved in Pol III transcription termination, located downstream of the SL RNA gene directs Pol II transcription termination. At least six Ts are required for efficient termination *in vivo* in *L. tarentolae* [121]; the mature 3’ end of the SL RNA is generated by nucleolytic processing. Interestingly, Pol II transcription of protein-coding genes does not stop at T-rich sequences, as such sequences are very common to intergenic regions in PGCs. Therefore, there must be functional differences between the Pol II complexes that synthesize the SL RNA and those that transcribe protein-coding genes; alternatively, epigenetic regulation might cause the observed differences. Termination of transcription has also been analyzed on chromosome 3 from *L. major*, which contains two convergent PGCs separated by a tRNA gene [122]. Nuclear run-on and RT-PCR assays indicated that Pol II-mediated transcription of both PGCs terminates within the tRNA-gene region [38]. The presence of a termination region between two convergent PGCs on *L. major* chromosome 3 suggests that *Leishmania*, like yeast, may require the separation of adjacent Pol II transcription units by proper termination signals to avoid polymerase collisions [123]. Because several convergent PGCs in trypanosomatids are separated by tRNA genes (or other genes transcribed by Pol III) [18], the involvement of tRNA genes in transcription termination may not be exclusive to *L. major* chromosome 3. Interestingly, the tRNA gene region on chromosome 3 can also terminate Pol I and Pol III transcription [38]. A ChIP-seq study showed that histone variants H3V and H4V are present at convergent strand-switch regions and other parts of the *T. brucei* genome where transcription probably ends, suggesting that chromatin structure plays a significant role in transcription termination in trypanosomatids [39].

In most cases, Pol III ends transcription at simple clusters of four or more T residues, normally flanked by G+C-rich sequences [124]. In human and mice, tRNA genes need four Ts to end transcription, whereas in *S. pombe* and *S. cerevisiae* tRNA genes require six and seven Ts, respectively [125, 126]. For a particular species, termination efficiency tends to increase with the length of the T run. The proteins that stimulate Pol III transcription termination in human cells include Nuclear Factor 1 (NF1), Positive Factor 4 (PC4), and the La antigen, an UUU-OH-terminus-binding protein [61]. Subunit TFIIIC2 of TFIIIC participates in transcription termination and in transcription reinitiation. Moreover, it has been reported that Pol III subunits C11, C37 and C53 form a subcomplex that is also involved in transcription termination and reinitiation [61].

In trypanosomatids, as in higher eukaryotes, T runs function as Pol III termination signals. It was reported that transcription of the *T. brucei* U2 snRNA terminates at several T's located downstream of the gene [127]. A cluster of Ts of variable length was found on every single tRNA gene in trypanosomatids. The distance between the end of the gene and the run of Ts varies from zero to seven bases [18]. In *L. major*, the average length of the T run is 4.87 bases, with a minimum of four and a maximum of nine Ts. In *T. brucei* the mean T-run length is 4.89 bases (ranging from four to ten Ts). The stretches of Ts are longer in tRNA genes from *T. cruzi*.
DNA binding [133]. Interestingly, it was shown that Pol III transcription of the rRNA gene located on chromosome 3 terminates within a tract of four Ts [38]. To date, no single protein involved in Pol III transcription termination has been identified in trypanosomatids.

In eukaryotic cells, transcription termination elements for Pol I are located downstream of the 28S rRNA gene and upstream of the transcription start site. In mammals, factor TTF-I binds the termination elements at the 3′ end of the transcribed region, forcing Pol I to pause, and cooperates with the transcript-release factor PTRF in conjunction with a T-rich DNA region to induce transcription termination [72, 128]. Regarding termination of transcription in rRNA genes in trypanosomatids, in *L. infantum* it was found that transcription ceases downstream of the 3′ end of the rDNA unit in an area that contains short sequences with the potential to form stem-loop structures, which are reminiscent of the bacterial rho-independent transcriptional terminators [129]. *L. major* rRNA genes contain a similar sequence in a region where run-on assays indicated that transcription terminates [76]. In the GPEET/procyclin locus from *T. brucei*, three sequence elements that are located downstream of the last gene of the cluster act synergistically to terminate Pol I transcription in an orientation-dependent way [130]. Also, it was recently reported that Pol I transcription of the EP/procyclin locus ends within the PAG1 gene (Figure 2). Transient and stable transfections showed that sequence elements on both strands of the gene can inhibit Pol I transcription [131].

### 7. Transcription of Transposable Elements

Although trypanosomatids do not seem to contain DNA transposons, analysis of their genomes confirmed the presence of abundant long terminal repeat (LTR) and non-LTR retrotransposons [14, 132]. They account for ~5% and 2% of the *T. cruzi* and *T. brucei* genomes, respectively. In *T. cruzi*, one of the most abundant retrotransposons is L1Tc, which encodes a protein that contains several domains, including reverse transcriptase, endonuclease, RNase H and DNA binding [133]. Interestingly, it was shown that the first 77 bp of L1Tc behave as a promoter region that activates transcription of the retrotransposon [134]. Run-on experiments indicated that transcription of L1Tc is driven by Pol II. It is worth noting that the 77 bp region is also present in other transposable elements in trypanosomatids [14]. SLACS is a retrotransposon from *T. brucei* that integrates exclusively at nucleotide 11 of the SL RNA gene [135]. It was reported that transcription of SLACS starts at the +1 nucleotide of the interrupted SL RNA gene, but in this case transcription is directed by the upstream SL RNA promoter (carried out by Pol II) [136].

In contrast to *T. cruzi* and *T. brucei*, *Leishmania* species do not contain active retrotransposons. However, they have remnants of extinct ingi/L1Tc-like retroposons called DIREs [137]. Recently, two new families of degenerated retrotransposons were identified in *Leishmania*: SIDER1 and SIDER2 [138, 139]. These sequences are predominantly located within the 3′-UTR of *Leishmania* mRNAs. It was shown that SIDER2 acts as an instability element, since SIDER2-containing mRNAs are generally expressed at lower levels compared to the non-SIDER2 mRNAs [138]. Interestingly, a significant number of SIDER elements map to divergent strand-switch regions [140], which may contain Pol II transcription initiation sites (see Section 4.1). Thus, it is feasible that SIDER sequences participate in the control of transcription in *Leishmania* [140]. Supporting this possibility, a large fraction of binding sites for transcription factors are embedded in distinctive families of transposable elements in mammals [141, 142].

### 8. Epigenetics

Nuclear DNA in eukaryotic cells is organized in a complex DNA-protein structure called chromatin. The fundamental subunit of chromatin is the nucleosome core, composed of an octamer of small, basic proteins named histones around which ~147 bp of DNA are wrapped. The histone octamer consists of two copies each of H3, H4, H2A and H2B, known as the core histones. A different histone, H1, binds to the “linker DNA” region (~80 bp) between two nucleosomes, and helps to stabilize the chromatin. The nucleosomal array, which imparts about a sevenfold condensation of the DNA molecule, is compacted another sixfold into a 30-nm chromatin fiber. Dynamic changes in chromatin structure play a very important role in the regulation of DNA-dependent processes such as transcription. Several mechanisms regulate chromatin structure, including: nucleosome remodeling by ATP-driven complexes, covalent modifications of the N-terminal tails of the core histones, replacement of one or more of the core histones by their variants, and nucleosome eviction [143, 144]. In general, the nucleosome works as a transcriptional repressor that prevents the binding of transcription factors to promoter regions. However, correctly positioned nucleosomes can bring remote DNA sequences into close proximity to activate transcription [145]. At least eight different types of modifications have been found on histones [146]. Small covalent modifications such as acetylation and methylation have been extensively studied; acetylation is almost invariably associated with transcription activation, while methylation can be related to either activation or repression [146]. Histone variants, which are non-allelic isoforms of canonical histones, can be incorporated into nucleosomes. As a result, the structure and function of the nucleosome are modified [147]. The combination of histone modifications and histone variants generates a vast variety of nucleosomes.

Little is known about chromatin structure and epigenetic regulation in trypanosomatids [148–150]. These organisms have several copies of the genes encoding histones H1, H2A, H2B, H3, and H4 [12]. Histones in trypanosomatids are, however, extremely divergent from those found in other organisms. Nevertheless, as in other organisms, nucleosomes constitute the basic structural unit of chromatin in trypanosomatids [151, 152]. Micrococcal nuclease (MNase) digestions of chromatin showed the typical nucleosome ladder, with a monomer of ~200 bp in trypanosomes.
of polycistronic transcription in trypanosomatids do not condensate during mitosis; actually, chromatin does not even fold into the 30-nm fibers that are commonly found in higher eukaryotes [157].

Nucleosomal ladders analyzed by Southern-blot, with different regions of the SL RNA gene (transcribed by Pol II) from L. tarentolae, indicated that the promoter and transcribed regions are not organized into nucleosomes [158]. However, a consistently positioned nucleosome was found within the non-transcribed intergenic region [158]. Similar experiments performed in L. major indicated that while the promoter region of the rRNA unit is devoid of nucleosomes, the rRNA genes are packed into nucleosomes (Vizuet-de-Rueda and Martinez-Calvillo, unpublished results). Also, tRNA and 5S rRNA genes, which contain internal Pol III promoters, showed marked smearings in the MNase digestion profiles, suggesting an “open” structure of chromatin. Protein-coding genes and intergenic regions presented a nucleosomal organization (Vizuet-de-Rueda and Martinez-Calvillo, unpublished results). Thus, the presence and position of nucleosomes along DNA sequences most likely play an important role in controlling transcription initiation by the three different types of RNA polymerases in trypanosomatids, as has been reported in other eukaryotes.

A substantial number of post-translational modifications in histones have been recently identified in trypanosomes. These include several modifications commonly found in eukaryotes, such as acetylation and methylation at several lysine residues on histone H4 [159, 160]. N-methylalanine is a novel modification found at the N-termini of histones H2A, H2B, and H4 in T. brucei [161]. The functional relevance of this and other modifications has yet to be determined. In silico analysis indicates the presence of a number of enzymes implicated in histone modifications in trypanosomatids, such as acetyltransferases, methyltransferases and histone deacetylases [12]. In T. brucei, histone acetyltransferase 2 (HAT2) is required for H4-K10 acetylation [162], whereas HAT3 is responsible for acetylation at H4-K4 [163]. Histone modifications create a “histone code” that is “read” by effector complexes that will mediate subsequent functional outcomes. Components of the effector complexes possess different domains to bind to specific modified histones, including bromodomains, chromodomains and SANT domains [164]. Although several proteins with such domains are present in trypanosomatids, only the Bromodomain Factor 2 (BDF2) from T. cruzi has been proven to associate with acetylated histones [165]. Many histone variants have been identified in trypanosomatids [39]; these include H2AZ [166], which has a role in the maintenance of silent chromatin boundaries in higher eukaryotes.

A recent report revealed that acetylated (H3-K9/K14 and H4-K5/K8/K12/K16) and methylated (H3-K4) histones are enriched at divergent strand-switch regions in the T. cruzi genome [41]. A genome-wide study showed the presence of H3 histone acetylated at K9/K14 at the origins of polycistronic transcription in L. major, together with TRF4 and SNAP50 (Figure 1) [40]. A very small number of peaks (184) from the acetylated histone were found in the complete genome. While most of them were present at divergent strand-switch regions, some (54) were located at chromosome ends and within PGCs. Sixteen peaks were found in the vicinity of clusters of tRNA and snRNA genes [40]. Similarly, a ChIP-seq analysis in T. brucei demonstrated that histone H4 acetylated at K10, histone variants H2AZ and H2BV, and the bromodomain factor BDF3 are enriched at probable Pol II TSSs (Figure 1) [39]. As in L. major, most peaks were found upstream of PGCs, but 61 of them were located within PGCs (some peaks were associated with tRNA genes). It was demonstrated that H2AZ/H2BV-containing nucleosomes are less stable than canonical nucleosomes [39]. Thus, posttranslational histone modifications and histone variants might generate an open chromatin structure that is required for initiating transcription in trypanosomatids [167, 168]. Additionally, it was reported that histone variants H3V and H4V, only found in trypanosomatids presently, are enriched in regions where transcription seems to end [39].

9. Concluding Remarks

At the present time, we have limited knowledge about the way gene expression is regulated in these parasites will help us comprehend a number of important processes, such as differentiation, virulence and antigenic variation. It may also help in discovering key targets to control the infections caused by these organisms [9]. Moreover, the study of transcription and epigenetic regulation in early-branching eukaryotes, like trypanosomatids, will help us comprehend the evolution of the transcription machinery and the histone code [167]. Several lines of evidence indicate that transcription of protein-coding genes in trypanosomatids initiates upstream of the characteristic PGCs. Interestingly, the transcription initiation region from L. major chromosome 1 shares several features with TATA-less promoters from humans and other mammals. Likewise, recent functional studies have indicated that there are more general transcription factors in trypanosomatids than originally estimated. Thus, transcription of protein-coding genes in trypanosomatids might not be as atypical as initially thought. The identification of histone modifications and histone variants at the origins of polycistronic transcription revealed that chromatin structure plays an important role in transcription initiation in trypanosomatids, as it does in other organisms. Similarly, termination of transcription is likely to be influenced by chromatin-mediated epigenetic regulation. In the near future, genome-wide studies may provide more interesting data that will help us understand the complex mechanisms of transcription initiation and termination in this remarkable group of eukaryotes.

Acknowledgements

This work was supported by Grants 5 R01 TW007255-04 from The Fogarty International Center of NIH, 47543
from CONACyT, and IN203606 and IN203909 from PAPIIT (UNAM) to S. Martinez-Calvillo.

References


J.-P. Ruan, G. K. Arhin, E. Ullu, and C. Tschudi, “Functional
R. Hernandez-Rivas, L. E. Florencio-Martinez, M. Martinez-
B. Schimanski, T. N. Nguyen, and A. Gunzl, “Characteriza-
A. Gunzl, L. Vanhamme, and P. J. Myler, “Transcrip-


