Comparative Genomics and Molecular Characterization of N-alpha Acetyltransferase in Trypanosomes for Drug Target Identification

Stephen O Ochaya

Stockholm, 2013
To my family
(I love you all)
In memory of my late father Alfred Abuneri Ochaya
and to my mother Atuniata Atto Ochaya, my children,
my dear brothers and sisters, and others.
ABSTRACT

*Trypanosoma cruzi* and *Trypanosoma brucei* are flagellated protozoan parasites that cause Chagas disease and African trypanosomiasis in Latin American and African countries, respectively. Currently, over 8 million people are infected with *T. cruzi* and about 25 million more are at risk. About half a million people are affected by *T. brucei*. Trypanosome species share many peculiar biological and biochemical features, such as RNA editing. In contrast, they exhibit profound differences at the level of host-parasite interaction and disease pathology. Both parasites are transmitted to their host via different insect vectors. There are no available vaccines, and the current treatments have severe adverse effects. We were involved in sequencing the *T. cruzi* genome, an initiative launched by the WHO to increase our knowledge of the molecular basis of the parasite.

The aim of this thesis was to participate in the sequencing and analysis of the *T. cruzi* genome, and use the data to investigate acetyltransferase enzymes, presumably linked to important metabolic pathways, as possible drug targets. In Papers I and II, we describe genome sequencing and analysis of two distinct *T. cruzi* strains. One of the selected strains, CL Brener, was found to be a genetic hybrid of two divergent strains; and it contains about 22 000 genes, encoded on 700 scaffolds with a total genome size of 110 Mb. About 50% of the genes are of unknown function, and lack homology to other sequenced eukaryotes. Large numbers of members of surface molecule gene families, such as trans-sialidase, mucin, mucin-associated protein, and GP63 were found. Comparative analyses revealed that TcI had a smaller genome by up to about 11 Mb. The genome size difference was linked to genes encoding surface molecules and to other repeats and repeated genes. Additionally, six reading frames present in TcVI were not detected in TcI. Genetic polymorphisms such as, indels, microsatellites and SNPs were identified and analyzed. Many genes were found to be under different selective pressures in *T. cruzi*, indicating differential evolutionary rates, signifying their importance to parasite biology. Within syntenic regions, the two genomes have the same gene complement. Identified features warrant sequencing of further *T. cruzi* strains, and findings from our studies offer opportunities for more targeted functional studies as well as tools for epidemiology.

In the second part of this thesis, Papers III to V, a *Trypanosoma cruzi* acetyltransferase gene family, identified in the genome project, was chosen for functional characterization as a first step to evaluate its potential as drug target.

Acetyltransferases are responsible for protein acetylation, where an acetyl molecule is transferred from acetyl-Coenzyme A to lysine residues in a protein sequence, N-epsilon acetylation, and to N-termini of proteins or peptides, protein N-alpha acetylation. N-alpha acetylation is linked to many metabolic pathways, influences protein stability, protein-protein interaction, localization to organelles and acts as degradation signals. The impact of this post-translational modification in parasite is not known.

We have identified *T. cruzi* NatC and A, and show that they are expressed in the three life cycle stages (epimastigote, trypomastigote, and amastigote). The catalytic and auxiliary subunits form a complex in vivo. Additionally, they partially co-sediment with the ribosome and may have both co-translational and post-translational protein acetylation functions. In epimastigote, the catalytic subunit of *T. cruzi* NatA was localized both in the nuclear periphery and in cytoplasm, whereas NatC was predominantly assigned to the cytoplasm. The auxiliary subunit of NatA was mainly confined to the cytoplasm with cytoskeletal-like labelling, whereas NatC showed a punctate profile. Interestingly, the staining patterns of the different subunits analysed
for NatA and NatC differ between the life cycle stages, which suggests differential regulation and expression.

The native substrates for NatC and predicted NatA, are similar to those described in yeast and humans, suggesting evolutionary conserved functions. The proteins appear to acetylate a large number of proteins N-terminally, suggesting that manipulation of the enzymes may simultaneously affect many cellular functions and thereby could interfere with or abolish infection. Additionally, our data indicate that NatC and A, may have both N-alpha and N-epsilon acetylation potential.

Collectively, the genome analyses presented here have provided more molecular insights into the parasite's biology, and have narrowed the gaps between scientific communities working on parasite research. The identification of Nats and native substrates has hopefully laid a solid foundation for future study of Nats, which could provide chemotherapeutic targets for parasitic diseases.
LIST OF PUBLICATIONS

This thesis is based on the following list of publications referred to in the text by their roman numerals:

I. **The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease.**


II. **Shotgun sequencing analysis of Trypanosoma cruzi I Sylvio X10/1 and comparison with T. cruzi VI CL Brener.**

   Franzén O, **Ochaya S**, Sherwood E, Lewis MD, Llewellyn MS, Miles MA, and Andersson B.


III. **Characterization of a Trypanosoma cruzi acetyltransferase: cellular location, activity and structure.**


   *Molecular Biochemical Parasitology*. 2007 Apr;152(2):123-31

IV. **Identification and Molecular Characterization of N-alpha acetyltransferase C Protein Complex in Trypanosomes.**

   **Ochaya S**, Buhwa DA, Franzén O, Foyh H, Arnesen T, Matovu E, Åslund L, and Andersson B.

   *Manuscript*

V. **Cloning, Characterization and Expression Analysis of the N-alpha acetyltransferase A protein complex in Trypanosoma cruzi.**

   **Ochaya S**, Franzén O, Foyh H, Stov S, Arnesen T, Åslund L, and Andersson B.

   *Manuscript submitted*
Other publications

During the time of my doctoral studies I also contributed to the following studies.

- **Comparative karyotyping as a tool for genome structure analysis of** *Trypanosoma cruzi*. Branche C, **Ochaya S**, Åslund L, and Andersson B. *Molecular Biochemical Parasitology*. 2006 Jan 31 147, 30-38

# List of abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ac-coA</td>
<td>Acetyl-Coenzyme A</td>
</tr>
<tr>
<td>AT</td>
<td>Acetyltransferase</td>
</tr>
<tr>
<td>bp</td>
<td>Base pairs</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>DGF-1</td>
<td>Dispersed gene family 1</td>
</tr>
<tr>
<td>DTU</td>
<td>Discrete typing unit</td>
</tr>
<tr>
<td>EST</td>
<td>Expressed sequence tag</td>
</tr>
<tr>
<td>GCN5</td>
<td>General control non-depressible 5</td>
</tr>
<tr>
<td>GP63</td>
<td>Surface glycoprotein 63</td>
</tr>
<tr>
<td>GPI</td>
<td>Glycosylphosphatidylinositol</td>
</tr>
<tr>
<td>GST</td>
<td>Glutathione S-transferase</td>
</tr>
<tr>
<td>HPLC</td>
<td>High-performance liquid chromatography</td>
</tr>
<tr>
<td>kb</td>
<td>Kilobase</td>
</tr>
<tr>
<td>kDNA</td>
<td>Kinetoplast deoxyribonucleic acid</td>
</tr>
<tr>
<td>MASP</td>
<td>Mucin associated surface protein</td>
</tr>
<tr>
<td>Mb</td>
<td>Megabase</td>
</tr>
<tr>
<td>MMR</td>
<td>Mismatch repair</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger ribonucleic acid</td>
</tr>
<tr>
<td>NAT</td>
<td>N-terminal acetyltransferase</td>
</tr>
<tr>
<td>NatA-F</td>
<td>N-terminal acetyltransferase A-F</td>
</tr>
<tr>
<td>Nt-acetylation</td>
<td>Protein N-α-acetylation</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>PFGE</td>
<td>Pulsed field gel electrophoresis</td>
</tr>
<tr>
<td>TPR</td>
<td>Tetra-tricopeptide repeat</td>
</tr>
<tr>
<td>PTM</td>
<td>Post translational modification</td>
</tr>
<tr>
<td>PTU</td>
<td>Polycistronic transcription unit</td>
</tr>
<tr>
<td>TcAT</td>
<td><em>Trypanosoma cruzi</em> acetyltransferase</td>
</tr>
<tr>
<td>TcNaa10</td>
<td><em>T. cruzi</em> N-alpha acetyltransferase protein 10</td>
</tr>
<tr>
<td>T/S</td>
<td>Trans-sialidase</td>
</tr>
<tr>
<td>WHO</td>
<td>World health organization</td>
</tr>
</tbody>
</table>
Table of Contents

1. **Introduction** .................................................................................................................. 1
   1.1. Kinetoplastida ............................................................................................................... 3
   1.2. Transmission, epidemiology and life cycle of trypanosomes ...................................... 3
   1.3. *Trypanosoma cruzi* and Chagas disease ................................................................... 3
   1.4. *Trypanosoma brucei* and Sleeping sickness .............................................................. 5
   1.5. Some unique biological features in kinetoplastids ......................................................... 6
   1.6. Genetic diversity among *Trypanosoma cruzi* and clinical implications ....................... 7
   1.7. Persistence and pathogenesis of *T. cruzi* infection ..................................................... 8
   1.8. Control and chemotherapy of trypanosomiasis ............................................................. 9
   1.9. Karyotype and genome sequence of *T. cruzi* ............................................................ 9
       1.9.1. Karyotype ............................................................................................................... 10
       1.9.2. Genome ................................................................................................................. 10
   1.10. Comparative genomics .............................................................................................. 11
   1.11. Regulation of gene expression in trypanosomes ......................................................... 12
   1.12. Proteomics .................................................................................................................. 13
   1.13. Protein acetylation ...................................................................................................... 14
   1.14. Protein N-α-acetylation and N-α-acetyltransferases .................................................. 16

2. **AIMS** .............................................................................................................................. 18

3. **PRESENT INVESTIGATION** ............................................................................................ 19
   3.1. METHODOLOGICAL CONSIDERATIONS .................................................................. 19
       3.1.1. Genome sequencing .............................................................................................. 19
       3.1.2. Molecular functional characterization of an acetyltransferase gene family .......... 19
   3.2. RESULTS AND DISCUSSION ...................................................................................... 20
   3.3. Preliminary result ......................................................................................................... 28
       3.3.1. *In vitro* acetylation assay ...................................................................................... 28
   3.4. CONCLUDING REMARKS AND PERSPECTIVES ..................................................... 29

4. Acknowledgements ............................................................................................................. 32

5. REFERENCES ...................................................................................................................... 35
1. Introduction
In my childhood environment, I was exposed to the risk of acquiring parasite infections such as malaria, trypanosomiasis, leishmaniasis, amoebiasis, schistosomiasis, onchocerchiasis (river blindness), elephantiasis (filariasis) and parasitic nematodes. I used to see young children dying of high fever-related diseases and having worms in their faeces. People of different age groups complained of stomach-ache and passed away. Additionally, I used to graze hundreds of animals and could see them dying at appalling rates. Moreover, one of my colleagues all of a sudden fell asleep and woke up after some hours.

The diseases above have in common that they are caused by parasites. As a child, I asked why people and domestic animals fall sick and die at a young age, and what could be done to rescue them. At that time, I was told that a doctor, commonly known in the local LUO language as “daktar”, knows the answers and treats sick people and animals. It is thus possible that my interest in tropical diseases started at this stage in life.

Altogether, neglected tropical diseases affect one billion people at any given time and are the leading cause of death [1] (WHO 1999) in developing countries and have tremendous economic impact. The less developed nations constitute about 70% of the world population and one third of all the global deaths are due to infectious diseases caused by parasites and the situation is seriously aggravated by HIV co-infection [2]. Despite this, studies performed in 2002 showed that out of 1400 drugs developed between 1975 and 1999, only 16 were against tropical diseases and tuberculosis [3]. The interest by pharmaceutical companies to develop drugs for diseases, which are not viewed as public health threats to the developed world, is limited, as it may not be profitable [3]. The European Union has been quite instrumental in funding malaria research [4]. Recently, the financial support towards this research has been reduced [5], and yet, the disease kills over one million people every year, the majority of which are children below the age of five [4-6].

The World Health Organization has designated several tropical diseases as neglected, due to insufficient funding in relation to their global impact. Genome sequencing can aid the identification of new drug targets, or antigens for vaccine development and for diagnostic purposes. Possibly, drugs that are currently used for the treatment of other pathologies, such as cancer, could be effective against parasites and constitute alternative ways to control tropical diseases. Thus, homologs or orthologs of proteins used as chemotherapeutic targets are candidates for functional molecular characterization for their potential as candidates for new drug targets.

Parasites are scientifically interesting to study because of their unique biological characteristics, including mechanisms to evade the immune system, thereby allowing them to successfully colonize the host. Moreover, many Kinetoplastid organisms have unusual biological features, e.g., RNA-editing, polycistronic transcription and trans-splicing. Despite extensive basic research into the biology of most parasites for many years, there are still no vaccines and only a few proper drugs available.

Controlling insect vectors associated with infectious diseases is becoming problematic due to their resistance to insecticides. In order to address an effective long-term public health response to deadly parasitic diseases, the WHO launched parasite genome projects in 1994, concerning five pathogenic parasites causing health hazards in the developing world; Trypanosoma cruzi, Trypanosoma brucei, Leishmania, Schistosoma and Brugia malayi. The aims were to increase knowledge on the molecular biology of these parasites and to identify new genes involved in key cellular functions, which could be eligible as targets for new therapy.
To identify the genes of an organism, genome sequencing, using for example the shotgun technique (Figure 1), where genomic DNA is sheared into pieces and subsequently assembled, is commonly performed.

![Image of Genome Organisation and Sequencing](image-url)

**Figure 1:** Genome organisation and sequencing. Isolated genomic DNA is sheared into small fragments and a consensus sequence is produced. Sequences are then assembled into contigs, which are organized into supercontigs (scaffolds). Scaffolds can then be mapped to chromosomes. A genomic map is then constructed. Image credits: http://www.cdhgenetics.com/genetic-overview.cfm and http://what-when-how.com/insect-molecular-biology-and-biochemistry/insect-genomics-part-1/.

The genome sequencing of the first *T. cruzi* genome, which is the first part of this thesis, was an international effort where our laboratory at Karolinska Institute was involved. The results were published together with the genomes of *Leishmania major* and *T. brucei*, in the so called “Trypanosomatid (Tritryp) genome papers” [7-9]. The elucidation of the molecular karyotype (actual numbers and sizes of chromosomes in a cell) of *T. cruzi*, was a part of the genome project. The purpose of karyotyping was to facilitate chromosome-by-chromosome sequencing, which later in the project turned out to be technically difficult.

The *T. cruzi* genome was instead sequenced using a shotgun strategy (Figure 1). The genome sequence of *T. cruzi* [7] facilitated comparative genomics with other strains, such as *T. cruzi* Sylvio X10 (Tc1) [10]. The sequences produced have resulted in the discovery of many new genes, which could be exploited for drug or vaccine development. The perfect drug should be effective, safe and affordable, used orally and easy to administer. Finding drugs that target parasitic diseases in man and those threatening animals could be attractive measures to solve public health problems in low-income countries.

The genome sequences alone, however, only provide the genomic blueprint of an organism, but do not describe how it can function. Protein (proteomics), or post-translational modification studies as well as analyses of protein complexes will facilitate the understanding of parasite cellular processes.
In the second part of this thesis, a putative \textit{T. cruzi} drug target, the acetyltransferase gene family, with a focus on N-alpha acetyltransferases, was characterized. We wanted to understand the functions of these proteins in the parasite and evaluate their possible potential as drug targets in trypanosomes, and possibly other organisms causing infectious diseases.

1.1. Kinetoplastida
\textit{Trypanosoma cruzi}, \textit{Trypanosoma brucei} and \textit{Leishmania} belong to the order \textit{Kinetoplastida}, the name derives from their mitochondria-like organelle, called the kinetoplast. They are unicellular, flagellated parasites that infect a broad range of species, including plants, animals and humans [11]. \textit{T. cruzi} causes Chagas disease in 21 Latin- and South American countries [12]. \textit{T. brucei} is the causative agent of sleeping sickness in humans and Nagana in livestock, prevalent in sub-Saharan Africa [13]. \textit{Leishmania} has world-wide prevalence and causes visceral and cutaneous leishmaniasis [14]. In contrast to \textit{T. brucei}, which is extracellular, both \textit{T. cruzi} and \textit{leishmania} enters host cells. Diseases caused by these organisms affect mainly people living in rural areas.

1.2. Transmission, epidemiology and life cycle of trypanosomes
Trypanosomatids, here referred to as \textit{T. cruzi} and \textit{T. brucei}, have many biological and structural similarities, such as unique gene regulation and organelle composition, for example, the kinetoplast. Despite these, they have different vectors, transmission cycles and mechanisms of antigenic variation. Furthermore, their life cycle and disease pathologies are strikingly different [11].

1.3. \textit{Trypanosoma cruzi} and Chagas disease
\textit{T. cruzi}, the etiological agent of Chagas disease, was described for the first time in a two-year old patient by the Brazilian physician Carlos Chagas in 1909. It was suggested by Hotez \textit{et al} [15] as equivalent to diseases caused by HIV in the Americas, though this may be exaggerated [16]. About 10 million people are infected and 100 million more are at risk [12]. As there is no effective treatment, the economic loss and medical and social consequences are substantial (table 1). Due to population movement of infected people (immigrants) from Latin America to non-endemic regions and countries of the world such as USA, Canada, Europe, Australia, China and Japan, Chagas disease has become a problem outside of South- and Central America [17-19]. In Sweden, approximately 1000 individuals are infected with \textit{T. cruzi} [20].
<table>
<thead>
<tr>
<th>Disease</th>
<th>Pathogenic agent</th>
<th>Vector</th>
<th>People infected (inf), and at risk in millions (M)</th>
<th>Death, and new cases/year</th>
<th>Economic loss/year and disabilities (Dis)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malaria</td>
<td><em>Plasmodium falciparum</em></td>
<td>Female Anopheles mosquito</td>
<td>40% of the world population at risk.</td>
<td>3 M death, 0.5 M cases</td>
<td>Reduced economic growth by 1.3%</td>
</tr>
<tr>
<td>Schistosomiasis</td>
<td><em>Schistosoma sp.</em></td>
<td>Snail species</td>
<td>200 M (inf), 600 M Risk</td>
<td>0.2 M death</td>
<td>1.7 M Dis</td>
</tr>
<tr>
<td><em>Chagas disease</em></td>
<td><em>T. cruzi</em></td>
<td>Reduviid bugs triatomines</td>
<td>About 10 M (inf), over 25 M Risk</td>
<td>0.2 M death</td>
<td>Over US$ 1.2 billion, 0.43 M Dis</td>
</tr>
<tr>
<td><em>Leishmaniasis</em></td>
<td><em>Leishmania sp.</em></td>
<td>Phlebotomine sand flies</td>
<td>12 M (inf), 350 M Risk</td>
<td>0.6 M death, 0.2 M cases</td>
<td>1.97 M Dis</td>
</tr>
<tr>
<td><em>Sleeping Sickness</em></td>
<td><em>T. brucei</em></td>
<td>Tsetse flies</td>
<td>0.07 M (inf), 70 M Risk</td>
<td>0.5 M death</td>
<td>US$ 1.5 billion, 1.7 M Dis</td>
</tr>
<tr>
<td>River Blindness</td>
<td>Filarial parasite</td>
<td>Simulium blackflies</td>
<td>37 M (inf), 120 M Risk</td>
<td>0.27 M cases</td>
<td>US$ 50 by year 2020, 0.39 M Dis</td>
</tr>
<tr>
<td>Onchocerchiasis</td>
<td><em>Onchocerca volvulus</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


*T. cruzi* is transmitted by haematophagus insects of the family Reduviidae, (Figure 2) via (i) the classical route of transmission [21]. (ii) contaminated blood during transfusions or organ transplants [22, 23]. In the USA, this kind of transmission has become substantial due to immigration from endemic countries [24], (iii) congenital (mother to child) transmission, where infected mothers, during pregnancy and birth can pass the infection to the fetus, which could lead to miscarriage, premature birth and low birth weight, (iv) accidental transmission in the laboratory and ingestion of contaminated foods are rare but nevertheless possible [25]. An example of the latter is contaminated fruit and sugar cane juice, which has caused several outbreaks.

The *T. cruzi* life cycle is complex and can be divided into four main developmental forms: Two forms occur in the invertebrate (epimastigote and metacyclic tryomastigote), and two in the vertebrate (tryomastigote and amastigote) [26, 27] and (Figure 2).

**Figure 2.** Life cycle of *T. cruzi*: 1 and 2. The insect deposits feces contaminated with metacyclic tryomastigotes on the skin of the host. Infection occurs when the host for example scratches the site of the bite. Acute infection is
characterized by swelling around the infection site. 3. The trypomastigotes transform into amastigotes (round and oval shape without or with short flagella), which can infect any nucleated cell and proliferate inside the cells. 4. In about 2-3 days, the amastigotes differentiate into trypomastigotes and are released into the bloodstream when cells rupture [28]. Trypomastigotes can subsequently infect new cells. 5. The cycle is completed when the insect ingests new parasites from the infected host during feeding. 6. The trypomastigotes transform into epimastigotes inside the insect midgut and multiply. 7. Parasites transform into infective metacyclic trypomastigotes and move to the hindgut, ready to be excreted during defecation and infect new hosts. Cardiac disease, megacolon and megaesophagus may result from T. cruzi infection.

1.4. Trypanosoma brucei and Sleeping sickness

T. brucei was first described in South African livestock in 1894 [29]. A decade later, in 1901-1902, the parasite was observed in human blood [30]. Human African trypanosomiasis is caused by the subspecies T. b rhodesiense and T. b gambiense. T. conglobense, T. vivax and the subspecies T. brucei brucei causes Nagana in domestic animals. The diseases are vector-borne and transmitted by tsetse flies (Glossina sp.) to their hosts. Millions of people and animals, mostly in the rural areas and peri-urban centres in 36 African countries south of Sahara, are at risk of being infected [31, 32]. About 70,000 are infected (Table 1), though the numbers could be underestimated. Outside Africa, the parasite is not endemic but transmitted via migration and by travellers [33, 34]. The life cycle alternates between the two hosts, vertebrate and invertebrate, and is displayed in (Figure 3).

![Figure 3](image-url)

**Figure 3.** T. brucei life cycle. 1) The infective metacyclic epimastigotes in the insect are injected into the host by the tsetse fly. 2) In the host, they multiply in the circulatory system. 3) The parasite transforms into the stumpy non-dividing stage. 4) The parasite, especially, the stumpy form, is ingested by the tsetse fly from the infected host during the bite. 5) In the fly midgut, it transforms into the procyclic epimastigote stage and divides. 6) The parasite moves to the salivary glands and transform into metacyclic trypomastigotes that are ready to infect new hosts [13].

Depending on the subspecies of the parasite, sleeping sickness takes two forms: i- the acute phase, with symptoms appearing within weeks, is caused by T. b. rhodesiense (East and South African form); and ii- the chronic stage, with signs of disease appearing after some months or
years, is caused by *T. b gambiense* (West and Central African form). Both species have been detected in Uganda [13, 35]. At the late stage of infection, after about 6 months and 3 years for *T. b rhodesiense* and *T. b. gambiense*, respectively, the parasite crosses the blood- brain barrier, enters the central nervous system and causes neuropsychiatric disorders that lead to death [13, 36].

*T. brucei* uses an intricate mechanism of antigenic variation to escape the host immune system. It circumvents immune challenge by alternating its variant surface glycoprotein molecules (VSG) [13]. Efforts such as vector surveillance, training off field officials, political good will and new organisations has significantly reduced the parasite prevalence and burden [37]. However, a complete eradication may be hampered due to non-effective drugs to treat those infected and the absence of proper diagnostic tools to monitor disease pathogenesis [37].

**1.5. Some unique biological features in kinetoplastids**

Kinetoplastids have morphological, biochemical and genetic traits that are unusual, making them both exciting and difficult to study. In addition, several cellular features that are also found in other organisms were discovered in trypanosomatids. Some features are: 1. Unique organelles such as: i) the kinetoplast, a mitochondria-like structure, where RNA-editing occurs. RNA-editing is an RNA maturation step, where uridine residues are inserted or removed, directed by small guide RNAs (gRNA) [38]. Structurally, two cellular compartments (organelles); the nucleus and the kinetoplast, contain DNA. Kinetoplast DNA (kDNA), which constitutes approximately 25% of the *T. cruzi* total DNA [39, 40], consists of concatenated circular DNA molecules known as minicircles and maxicircles. About 1500 minicircle genes, making up to over 95% of the kDNA are present in the genome. They encode guide RNAs, involved in RNA-editing [41, 42], whereas, the maxicircle contains fewer genes, and ranges from 20 to 40kb in size, encoding mitochondrial enzymes and ribosomal RNAs [42]; ii) the acidocalcisome, an organelle containing calcium, though recently shown to be conserved from bacteria to human [43]; iii) the flagellar pocket, where endocytic activities occur [44]; iv) the glycosome, an organelle which harbors glycolysis enzymes [45] in contrast to mammalian glycolytic enzymes, which are located in the cytoplasm. The parasites, especially the blood stage of *T. brucei*, depend entirely on the energy produced by glycolysis and there is evidence of cell death when glycosomal biogenesis is impaired [46, 47]. This structure has been proposed as a possible drug target for kinetoplastids [48]; v) contractile vacuoles, suggested to be involved in the process of osmoregulation [49]; vi) endocytic and secretory pathways, which are restricted to the flagellar pocket and cytoplasmic regions [50]. 2. Regulation of nuclear gene expression, where the genes are transcribed together as a polycistrionic transcription unit and are processed into individual mRNAs through addition of a 3’ poly-A tail and 39 nucleotides at the 5’ structure known as the splice leader (SL) through trans- splicing [51-55] (Figure 4). 3. Antigenic variation of surface glycoproteins, especially in *T. brucei* [56]. 4. RNA polymerase I mediated transcription of protein-coding genes [57]. 5. GPI anchors of membrane proteins [58] and GPI biosynthetic pathways [59]. 7. Contraction and expansion of subtelomeric DNA repeats [60]. 8. Lysosome repair of plasma membranes [61-63].
1.6. Genetic diversity among *Trypanosoma cruzi* and clinical implications

*T. cruzi* is mainly clonal, with infrequent genetic exchange, which has resulted in a heterogeneous population structures [64]. Initially, using ribosomal RNA, the *T. cruzi* isolates were grouped into lineage I, which can be associated with both sylvatic and domestic cycles [65], and lineage II to the domestic cycle [65, 66]. Lineage II was further divided into five phylogenetic subgroups, named Iia-Ile, using the multilocus enzyme electrophoresis (MLEE)/RAPD technique [67, 68]. The genome sequencing reference strain *T. cruzi* CL Brener (lineage Ile) [7] was chosen as a representative of lineage II, whereas the *T. cruzi* Sylvio X10 strain, sequenced later [10] was a representative of lineage I. More recently, new nomenclature categorized *T. cruzi* into six discrete typing units (DTU) designated TcI-TcVI [69, 70]. The six DTUs infect humans and occur across the Americas to different extent.

*T. cruzi* can undergo genetic exchange [71]. The DTU TcVI, consists of hybrid lineages [72-74]. Lineages TcIII-TcIV may have arisen due to ancestral recombination between TcI and TcII. Similarly, TcV and TcVI may have arisen due to recombination between TcII and TcIII [69, 71]. Hence, genetic exchange appears to have taken place but it is sufficiently rare to allow partition of *T. cruzi* into clonal lineages.

Further complexity can be added to the *T. cruzi* population structure by the fact that many strains may be regarded as multiclonal [75]. During infection and disease, the parasites may undergo clonal selection. Thus, parasites collected from the insects could be genetically different compared to those isolated from the patients [75]. The heterogeneity within *T. cruzi* has made it difficult to explain the epidemiology and pathology of Chagas disease, since it is unclear to what extent the *T. cruzi* genotype influences infection and disease. Several studies indicate a possible association between parasite lineage, geography and clinical aspects of Chagas disease, which may be attributed to both parasite and host genetic variation [69, 75, 76].

Comparatively, TcI is most common in countries north of the Amazons, and is associated with heart disease (cardiomyopathy) [65, 77], whereas, TcVI is predominant in southern and central regions of South America, and contributes to cardiac complications, megaesophagus and megacolon [65, 78, 79]. The mismatch repair (MMR) machinery, which controls high mutation rates during DNA metabolism in the parasite, as evidenced by cloning one of the genes (MSH2) and SNP analysis [80, 81], was partly linked to the clinical differences between *T. cruzi* lineages. The efficiency of the MMR machinery in *T. cruzi* is directly associated with the mutation rate.

**Figure 4.** *Trans*-splicing and polycistronic transcription. Image credit: http://www.library.csi.cuny.edu.
and is more efficient in TcI compared to TcVI [81]. TcVI strains display high infectivity in animal models, while the TcI strain causes the chronic stage of the disease with 100-fold higher inoculums [82]. It is also possible that the genetic status of the host influences the final phenotype [83, 84].

The mechanisms behind *T. cruzi* genetic diversity and its importance for immune evasion have recently been discussed [85]. In contrast to *T. brucei*, which uses antigenic variation to escape the host immune system, *T. cruzi* makes use of its heterogeneity to adapt to the host environment. For example, various surface glycoprotein molecules such as, trans-sialidase and trans-sialidase like molecule known as parasite-derived neurotrophic factor (PDNF) [86] have hundreds of expressed variants in each strain. Also, the parasite can evade the immune system by binding to the receptor tyrosine kinase (Trka receptor) of the host and enter the cell [87]. Several genes in the host involved in key cell functions, such as immune response, are affected during *T. cruzi* infection [88], suggesting that, the parasite induces mechanisms to protect its own survival in the host cell [89].

The sophisticated strategies employed by the parasite to avoid immune-detection, is likely a result of several million years of co-evolution with various animal hosts. Several aspects of the *T. cruzi* life cycle are as yet unknown. Understanding the genetic diversity of *T. cruzi* is essential for the study of Chagas disease, since it may facilitate understanding of the epidemiological aspects and pathogenesis of the disease.

### 1.7. Persistence and pathogenesis of *T. cruzi* infection

The acute phase of *T. cruzi* infection, which lasts four to eight weeks, is characterized by high parasitaemia in the blood and several other tissues and organs. Despite an immune challenge from the host, parasites are not completely eliminated [90, 91]. The chronic stage is characterized by low parasitaemia, and preferential infection of cardiac muscle, colon and esophagus [12]. Only about 30% of infected individuals develop chronic symptoms after one to four decades [12].

The rest, about 70% remain asymptomatic carriers of *T. cruzi*. Since its discovery, Chagas disease pathogenesis has been linked to the level of parasitaemia [92]. However, evidence of tissue inflammation and a lack of detectable moving parasite when analysed by histochemical methods, indicate that autoimmunity plays a role in the etiology of the disease [92]. This notion has had serious negative connotations, since it may make the disease more difficult to treat [93]. In a later study, PCR was used to detect parasites in latent and chronic stages [90], which was not possible using histochemical methods. A reduction in parasite burden was later shown to correlate with a reduction of the symptoms of Chagas disease [94]. There is also a strong correlation between parasite levels and the immune status of the host [95]. Taken together, the etiology of the disease is parasite dependent, and autoimmunity is a secondary factor [93].

Recent studies have shown that parasite persistence could depend on the amastigote’s potential to enter cells [96]. The ability of *T. cruzi* to hide in adipocyte tissues and resurface decades later may be some of the factors leading to parasite persistence and disease pathology [97].

Collectively, three main theories may explain Chagas disease: i) latent parasite infection causing an inflammatory response and tissue destruction [98, 99]; ii) neurogenic dysfunction as a result of the action of toxin induced by the parasite, and autoimmune responses triggered by the infection; and iii) the controversial concept of autoimmunity caused by the vertical transfer of
parasite DNA into the host genome. The retained DNA in the host could be passed to their progenitors (lateral transfer), and may explain the perpetuating infection [100, 101].

1.8. Control and chemotherapy of trypanosomiasis
African trypanosomiasis was almost eliminated by the year 1960, before the colonial powers left Africa after independence. This achievement was obtained by various control programmes set up by colonial administrations [37]. By 1998, the prevalence of sleeping sickness was as high as 1929, partly due to deterioration of vector control and disease surveillance [102]. Current epidemiological data suggest a drastic decline of T. brucei infection, thanks to control measures being adopted, which may lead to eradication of especially, T. b. gambiense, in contrast to T. b. rhodesiense, which has substantial wild life reservoirs [13, 102].

Containment of Chagas disease in the Americas is problematic due to the zoonotic nature of the parasite. Public health education and killing the vector by spraying with insecticides have been somewhat successful with a decrease in prevalence of Chagas disease [103, 104]. The emerging insecticide resistance, however, could make this interruption of transmission difficult. Screening blood and organ donors, minimizing congenital transmission by testing pregnant women and treating those found to be positive, are some useful measures used to control transmission [105, 106].

At the moment, no vaccine is available towards any parasitic disease, and while T. cruzi vaccines are being investigated, these are far from the market [107]. Thus, the only treatment option is chemotherapeutics.

Treatment: Currently, there are only seven drugs approved to fight human trypanosomiasis and five of these were developed more than 30 years ago. For T. brucei, pentamidine (a diamidine compound) is used for the control of the first stage of T. b. gambiense and suramin, a trypan red derivative for acute T. b. rhodesiense treatment [13, 108]. Elfnorithine, an inhibitor of ornithine decarboxylase and melarsoporol, are the first line drugs recommended to control chronic sleeping sickness, though elfnorithine is less effective for the treatment of T. b. rhodesiense infections [13]. Combination of elfnorithine and Nifurtimox is currently advised as a therapy for the second stage of T. b. gambiense infection [109].

For T. cruzi, two drugs are available to treat Chagas disease: i) Nifurtimox, a derivative of nitrofuran, which inhibits trypaniothione reductase and causes production of free radicals, toxic to the trypanosome; and ii) benznidazole, a nitroimidazole, which likely blocks parasite DNA synthesis [110, 111].

The available drugs have different efficacy for the different life cycle stages [12, 112], and adverse side effects can be very significant [113]. The current drugs for T. brucei are far from ideal, especially in chronic stage [108, 114]. However, flexindiazole may be a promising new candidate drug for treating African trypanosomiases, [108]. Studies are on-going to find new drugs for the treatment of Chagas disease, most notably, cruzipain inhibitors, aimed at inhibiting ergosterol biosynthesis [115]. Additionally, two antifungal triazole derivatives, posaconazole and ravuconazole, may be promising drugs for Chagas treatment [48, 116]. In conclusion, there is an urgent need to characterize new drug targets.

1.9. Karyotype and genome sequence of T. cruzi
Sequencing the parasite genome and making the data available to the scientific community is an important step in the fight against parasite infections.
1.9.1. Karyotype
The karyotype of an organism refers to the number of chromosomes, and their sizes and structures. In most protozoans, the chromosomes are poorly condensed during cell division, which makes analysis by conventional cytogenetic techniques difficult as distinct chromosomes cannot be visualized [117]. This phenomenon seems to be more pronounced in T. cruzi, which may relate to the unusual structure of T. cruzi histone 1 [118]. Another approach to establish the karyotype is to use pulsed-field gel electrophoresis (PFGE) [119], which allows separating chromosomes by size. This technique has been used to determine the molecular karyotype of several parasites such as T. brucei [120, 121], Leishmania [122, 123], Plasmodium spp. [124], Giardia intestinalis, and other microorganisms like Saccharomyces [125].

The chromosome structure of T. cruzi is rather complex with a relatively large number of chromosomes. Contrary to Leishmania and T. brucei, there is a significant difference in signal intensity of the PFGE-separated chromosomal bands of T. cruzi when stained by ethidium bromide. This suggests that two or more chromosomes co-migrate or that there is aneuploidy. There is remarkable chromosomal size polymorphisms in different strains and among clones originating from the same strain [126]. These polymorphisms could possibly be attributed to expansion/contraction of repetitive sequence of the parasites, and to recombination in subtelomeric regions [126]. In one set of PFGE conditions, about 20 chromosomal bands could be visualized when T. cruzi intact chromosomes were separated by PFGE [127-129]. In our hands, the distribution of ethidium bromide for PFGE separated chromosomes bands was not the same, moreover, we observed extensive size differences between homologous chromosomes of the three different T. cruzi strains analysed [130] (Figure 5). These results confirmed previous studies suggesting extensive chromosomal variation in the parasite.

Figure 5. Comparison of the molecular karyotype of homologous chromosomes of CL-Brener (TcVI), Sylvio X10 (TcI) and CAI (TcI). The size of the circle is proportional to the number of chromosome pairs sharing the same degree of difference; the smallest circle represents one chromosome pair. (The figure was reproduced with permission [130]).

1.9.2. Genome
The completion of the sequence of the human genome in July 2000 has made a tremendous impact regarding the understanding of human genetics and biology and enabled a move towards individual medicine. Post genomic or functional genomics studies of what causes a disease or disease outcome can now be directly investigated by analyzing genes of interest. The first sequenced genome of a protozoan parasite, Plasmodium falciparum, was made available in 2002 [131]. To date, several important pathogen genomes have been sequenced, for example, Giardia [132], Kinetoplastids [7-9] and Schistosoma japonicum [133]. All point towards a better
understanding of the molecular biology of these organisms and the knowledge gained is being translated into new modes for intervention.

Based on the chromosome analysis, the total size of the *T. cruzi* CL Brener genome was estimated to be around 87 Mb [129]. Compared to other kinetoplastids, the genome of *T. cruzi* is more complex, as it varies extensively among strains and within clones of the same strains [71, 128, 134-136]. A comparison of the genome size estimates of *T. cruzi* lineages I and VI using different techniques [69, 71, 128, 136], suggests that the total DNA content of *T. cruzi* lineage VI is considerably larger than that of TcI.

The *T. cruzi* reference genome CL Brener (TcVI) was chosen for sequencing due to its clinical relevance, and it is well studied in different laboratories [137]. During the genome project, it was discovered that TcVI is a hybrid strain containing two different haplotypes, designated Esmeraldo-like (TcII-derived) and non-Esmeraldo-like (TcIII-derived) [7, 34, 137].

Recently, we generated a draft genome sequence of *T. cruzi* Sylvio X10/1 (Tc1) and compared it with TcVI (see below). The results indicate that TcVI has about 6 Mb more sequence related mainly to genes encoding surface proteins such as trans-sialidase [10].

The sequencing of the genome has provided information and resulted in the discovery of many genes, which have been submitted to the public database. Moreover, some unique genetic features in the parasite have been identified, which may facilitate the understanding of the fundamental cellular process in this organism.

### 1.10. Comparative genomics

Comparative genomics of *T. cruzi*, *T. brucei* and *Leishmania* [11], revealed 6200 conserved genes, 94% of which are organized in syntenic directional gene clusters [11]. The genome size and density varies, 50Mb (about 12000 genes), 26Mb (about 9000 genes) and 33Mb (about 8,300 genes) for *T. cruzi*, *T. brucei* and *L. major*, respectively [11]. Although the organisms diverged 200 to 500 million years ago [138], the kinetoplastid genomes show conservation of gene order. About 1600 protein domains have been identified in the Tritryp genomes [11]. Only 5% of the domains are specific for each parasite occurring on non-syntenic chromosomes and are mainly macrophage migration inhibition domains in *Leishmania*, VSG expression site associated domains in *T. brucei*, and serine carboxypeptidase in *T. cruzi* [11, 139]. At the amino acid level, clustering of orthologous gene across the three organisms revealed an average of 57% identity between *T. cruzi* and *T. brucei*, and as expected, 44% identity between *T. cruzi* and *leishmania* in agreement with phylogenetic relationships [139, 140]. The high percentage of domain conservation within Tritryp organisms has provided information for comprehensive analyses for chemotherapeutic interventions, highlighting the possibility of designing drugs that could potentially be used against all the three parasites. However, significant intraspecific variation does exist, and there is need to sequence and compare strain genomes in order to further understand the population structures and phenotypic variation.

It is estimated that 20 leishmania species and sub-species infect humans and cause disease in 88 countries in five different continents. The clinical outcomes vary depending on the species. Visceral leishmaniasis is caused by *L. infantum*. *L. major* contributes to cutaneous lesions and *L. braziliensis* is the main causative agent of mucosal leishmaniasis [141]. A comparison between the genome of *L. major* with the *L. infantum* and *L. braziliensis* genomes revealed that the Dicer and Argonaute proteins, required for RNAi, were identified in *L. braziliensis* and are absent in the other two strains, as in *T. cruzi* [141]. Furthermore *L. braziliensis* contains 47 genes not present in the other two parasite genomes. Twenty-seven and five specific genes for *L. infantum* and *L. major*, respectively, were identified.
In contrast, a comparison of the *T. brucei* reference genome with that of *T. gambiense* (strain DAL 972), isolated from a patient at Daloa, Ivory Coast in 1986 [142], revealed more similar gene content and synteny [143]. Moreover, the estimated sequence identity in the coding regions between the two genomes was 99.2% [143]. No specific gene was found that could explain why *T. gambiense* infects human, whereas, *T. b. brucei* does not. Thus, other factors such as genetic polymorphisms may explain phenotypic variations [143].

New sequencing technology (the next generation sequencing platforms) has drastically decreased the cost of genome sequencing. We used the new methodology to sequence a *T. cruzi* TclI lineage and compared it with the reference hybrid strain (TcVI) [10]. Consistent with comparative genomics between intra *T. brucei* and leishmania species, the genome architectures of the two *T. cruzi* strains were very similar, containing the same gene complement within the syntenic regions [10]. In agreement with the karyotyping study, the TcVI genome is larger [10, 130], but unlike *T. brucei*, there is a high degree of sequence polymorphism between strains.

Clearly, as evidenced by the data produced in the Tritryp genome project and the comparative analyses [11] including intra-species sequence investigation [10, 37, 141, 143], there is high conservation and synteny of genes. In general, there are no immediate explanations regarding epidemiology and pathogenesis of these organisms based on genome sequences alone, although a few species-specific genes and many SNPs were identified. Thus, the genome has provided a genetic road map for functional characterization of genes and cellular features in the parasites. Protein-coding genes shared by the three parasites, suspected to participate in similar metabolic pathways might be interesting for functional laboratory investigations.

**1.11. Regulation of gene expression in trypanosomatids**

Due to their complex life cycle, trypanosomatids go through several transformations in order to adapt to different environments. In addition, they face harsh immune challenges, temperature cues, variation in nutrient availability and drug pressure. Additionally, in the same host, they transform from dividing to non-proliferative stages. Thus, there is a need to reprogram their cellular machineries and coordinate gene expression. Mechanisms for regulation of gene expression are therefore required. Compared to most other eukaryotes, trypanosomatids have unusual mechanisms of gene expression [144]. Genes are organized in long polycistronic transcription units and are transcribed by RNA polymerase II (Pol II) to yield polycistronic pre-mRNAs [144]. Intriguingly, in *T. brucei*, some VSG genes are transcribed by RNA polymerase I [57].

In the mitochondria, the primary pre-mRNA in trypanosomatids undergoes extensive editing processes, involving posttranscriptional insertion and removal of uridine residues directed by guide-RNA encoded from minicircles [57, 145]. To achieve mature mRNA from polycistronic pre-mRNA, two essential reaction events need to occur: *trans*-splicing and polyadenylation (Figure 4).

*Trans*-splicing is a process, which ensures that the mature mRNA has identical sequence of 39 nucleotides, the so-called spliced leader (SL) at its 5’ end. This process is accompanied by addition of multiple adenosines at the 3’-end (polyadenylation), resulting in monocistronic mature mRNAs [145]. The two processes are coupled and appear to be guided by polypyrimidine-rich tracts present in intergenic regions [53, 145]. There is some evidence that alternative splicing and polyadenylation may occur, which could alter gene expression [146, 147].
Information about proteins or DNA sequences that are associated with transcriptional initiation and termination is limited, which indicates that genes are mainly post-transcriptionally regulated in these organisms. The use of trans-splicing and polyadenylation of poly-cistronic transcripts suggests that transcription is constitutive [144], perhaps as a result of the parasite’s inability to control gene expression at the transcription initiation stage [55]. Gene expression could thus be regulated by changes in mRNA turnover and translation [144], which in the latter case seems to be regulated by ribosomal binding proteins (RBPs), for example TbPUF9 [148]. The emergence of new technologies such as, RNA-seq and ChIP-seq (chromatin immunoprecipitation) followed by DNA sequencing have greatly improved our knowledge of transcriptome in trypanosomatids [86, 147, 149-151]. The current evidence suggests that there seems to be chromatin-mediated epigenetic control of gene expression in kinetoplastids [152-154]. Furthermore, an origin of replication initiation site has recently been mapped, and it appears to be associated with transcription [155].

Interestingly, several studies have revealed differential expression of genes during trypanosomatid life cycles [156-159]. However, as analysed in T. brucei, only 6% of the membrane transport system, accounting for about 10% of the total genome, was found to be differentially regulated [160], whereas 7% and 9% of genes in the total genome of L. infantum and L. major, respectively, were found to differ in their expression between the promastigote and amastigote stages [161], suggesting that large numbers of genes in trypanosomatids are constitutively expressed during the life cycle. As published by Minning et al [156], for T. cruzi there is significant differences in the abundance of transcripts, up to about 50%, in the four distinct life cycle stages of the parasites. Moreover, certain genes were preferentially enriched depending on the host (vertebrate or invertebrate), or dividing stage/non proliferative stage. The published results indicate that the variation of the amount of transcripts for most genes between the different developmental stages are small, and do not correlate with the amount of proteins that is expressed [144, 162].

Some examples of PTMs in parasites; Acetylation, phosphorylation, methylation, glycosylation, N-terminal pyroglutamylation, deamidation, and tryptophan oxidation, S-nitrosylation, ubiquitination and proteolytic processing

Figure 6. The concept of studying DNA, RNA and Protein.
The picture was partly credited from https://sites.google.com/a/nyu.edu/vogel-lab/research

1.12. Proteomics
As already discussed, trypanosomtids differ in species and tissue tropism and in the types of disease that they cause. The integration of different types of data, for example genomics, transcriptomics and proteomics (Figure 6), may lead to an improved understanding of how these parasites live and how they interact with the host. In particular, proteomics studies may be important. Therefore, studying proteins and post-translational modifications (PTMs) in
Kinetoplastids may facilitate drug or vaccine development. We therefore, chose to analyse PTM, in this case acetylation. The achievements in genomic sequencing of T. cruzi and other parasites have provided data and new avenues to gain direct biological insights into the processes leading to human disease. However, characterizing the encoded proteins and understanding their functions are crucial. Proteomics is the study of the entire proteome using high-throughput techniques. In many cases the cellular locations of proteins are indicative of their putative function.

For kinetoplastids, protein studies may be particularly important, since the control of gene expression in general, appears to be posttranscriptional, and this may render nucleic acid based studies less informative [163]. Furthermore, most drugs target proteins and not directly nucleic acids.

Already, T. cruzi proteomics studies have revealed proteins that are abundantly expressed at the surface, significant differential expression of proteins between developmental stages and some distinct stage-specific pathways, which could be used for vaccine development or as targets for drug intervention [164-167]. To further gain some insights into unknown proteins, subcellular proteomics of the parasites at organelle levels has contributed to annotating many of the genes where a function has not yet been assigned [165]. For example, a subcellular proteomics study of ribosomes from T. cruzi by Ayub et al [166], indicated that about 50% of T. cruzi ribosomal genes have homologs in yeast. Moreover, in the same study, many of the T. cruzi proteins annotated, differ from yeast at the C or N-terminal sequences, which may be useful for developing therapeutically interventions [166].

One of the mechanisms of gene regulation is the use of post-translational modifications (PTMs), which may regulate and affect protein stability, activity, localization or molecular interactions.

1.13. Protein acetylation

As mentioned, the functionality of proteins in a cell may be determined by PTM. When a protein is released from the ribosome, it often undergoes PTM, such as glycosylation, which has significant effects on, for example, protein folding. Another PTM is phosphorylation, which takes place mainly on serine, threonine and tyrosine residues. There are also other PTMs, such as ubiquitination, methylation (on arginine and lysine residues) and acetylation. In plasmodium and kinetoplastids, the findings from proteomics, comparative proteomics and PTM studies, have highlighted the importance of PTM studies for therapeutics [165, 168].

Acetylation is a modification where an acetyl moiety is transferred from acetyl-CoA to the N-terminal or an internal residue of a protein/peptide by acetyltransferase enzymes (Figure 7), and it has been described as a regulatory modification as important as phosphorylation [169, 170]. It is present in all kingdoms of life [171]. The biological importance of histone acetylation has attracted the most attention thus far, since it is correlated to chromatin structure; the nucleosome, which consists of a histone octamer, and approximately 147 bp of DNA. Histones are important for chromatin structure. N-terminals of histones that protrude from the nucleosome octamer are subjected to posttranslational modifications, such as acetylation of conserved lysine residues [172]. Other histone domains can also be modified.
Figure 7. Protein acetylation. (A) Histone acetyltrasferase (Hats), where acetyl is transferred to a lysine residue in a protein sequence. It enhances transcription. This process is reversible by histone deacetylases (HDs/HDACs), which suppress transcription. (B) NATs, transfer acetyl moiety from acetyl-CoA to the primary α-amines of N-termini of proteins/polypeptides, and can determine the fates of many proteins. Picture 8A was credited from http://www.web-books.com/MoBio/Free/Ch4G.htm.

The enzymes responsible for the transfer of acetyl to lysine residues are known as lysine acetyltransferases (LATs)/histone acetyltrasferases (HATs), which partly contribute to epigenetic signalling in histones. It is reversible by deacetylases (DCs)/HDACs. The imbalance between the HAT and HDAC activities is reported to be associated with significant changes in cellular metabolism, that is, from normal cell function to, for example, carcinogenesis [173]. Some of the LATs and HDACs are present as protein complexes and modify multiple targets, both histones [170] and non-histone proteins [174]. In cancer cells, the expression of HDACs is elevated, prompting increased cell proliferation, which is a hallmark of tumours. Fortunately, this process can be reversed by inhibiting the enzymes through histone deacetylase inhibitors (HDACi). Even though the different HDACs specificities could not be clearly defined for a particular cancer cell type [175], many small inhibitors that target the enzymes are in use or in clinical trial as anti-cancer compounds, for example, vorinostat [176].

Studies of HATs/LATs in parasites have lagged behind, but the field has accelerated after genome sequences became available. The most well characterized family of human acetylases is the so called GCN5 N-terminal acetyltransferase (GNAT) super family. GNATs contain an Acetyl-CoA binding region/motif that is conserved in all organisms. Members of this family are present in the genomes of all different parasites sequenced thus far [9, 177, 178].

In malaria, many LATs have been recognized to be key players in gene regulation, for example, in chromatin remodeling, cell cycle control and DNA repair [179-181]. Several classes of HDACs have also been identified in Schistosoma mansoni [182]. Recent proteomics data detected over 400 new acetylation sites in Toxoplasma gondii. This involved many different proteins that could possibly be used for therapeutic intervention [183]. Six histone acetyltransferases (HAT1 to HAT4, ELP3a and ELP3b), and seven histone deacetylases have been found in T. cruzi, T. brucei and Leishmania [9, 184]. These enzymes have been linked to many cellular functions, including transcription [184, 185]. Bromodomain factors, which may be necessary for protein-protein interaction involving these enzymes or necessary for the recognition of lysine residues, were detected [9, 184, 186]. The alignment of Triryp LATs with homologs from other organisms identified unique characteristics, which makes parasite LATs interesting for possible drug design [184].
Possibly, major parasitic infectious diseases such as malaria, schistosomiasis, Toxoplasmosis and those caused by kinetoplastids, could be controlled by HDAC inhibitors [182, 184], compounds used in the treatment of cancer. However, much work remains in order to determine the therapeutic potential of these compounds in parasitic diseases.

1.14. Protein N-α-acetylation and N-α-acetyltransferases
Compared to Nε-acetylation, protein N-α-acetylation (Nt-acetylation) is a permanent modification, where an acetyl group is transferred from Acetyl CoenzymeA (Ac-CoA) to the alpha-amino group of N-terminal residues. It is a co- and post-translational modification, carried out by N-α-acetyltransferases (Nats). Nt-acetylation is linked to various biological processes, such as cell cycle and cell differentiation and to diseases like cancer. Particularly, the modification can determine the fate of a cell by influencing protein folding/stability, activity and localization to, for instance the endoplasmic reticulum organelle, or act as a degradation signal [187].

Thus far, the acetylation state has been defined by proteomics tools in many organisms, and it has been found to be variable. About 18 %, 60 %, 75% and 90% of proteins in archea, yeast, plant and human, respectively, are thought to undergo Nt-acetylation [188-190]. The co-translational modification occurs after about 20-50 amino acids protrude from the ribosome during protein synthesis [191-193]. The capability of Nats to post-translationally acetylate proteins is not well-documented [194].

Nats are categorized depending on the amino acid sequence of the N-termini that they acetylate. In humans, NatA to Nat F has been defined [187, 188]. The six different Nats and their known native substrates preferences are as shown in Table 2.

To date, no efforts have been made to identify Nats and corresponding substrates in any parasite. The question is, if Nats are present in these organisms, how many of them exist, and what proteins do they acetylate, which cellular pathways and functions do the acetylated proteins participate in, and what are the biological consequences of inhibiting Nats? The answer to these questions may open a new avenue for parasite control.

<table>
<thead>
<tr>
<th>NatA</th>
<th>NatB</th>
<th>NatC</th>
<th>NatD</th>
<th>NatE</th>
<th>NatF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>Met-Asp</td>
<td>Met-Leu</td>
<td>Ser</td>
<td>Met-Leu</td>
<td>Met-Lys</td>
</tr>
<tr>
<td>Ser</td>
<td>Met-Glu</td>
<td>Met-Ile</td>
<td>Met-Leu</td>
<td>Met-Lys</td>
<td>Met-Ala</td>
</tr>
<tr>
<td>Thr</td>
<td>Met-Asn</td>
<td>Met-Phe</td>
<td>Met-Ala</td>
<td>Met-Val</td>
<td></td>
</tr>
<tr>
<td>Gly</td>
<td>Met-Trp</td>
<td>Met-Met</td>
<td>Met-Met</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Val</td>
<td>Asp ‡</td>
<td>Glu ‡</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Different NATs in human and their substrates as described in [188, 195]. NatA acetylates the indicated substrates after the removal of methionine by methionine amino peptidase enzymes. ‡ NatA can also post-translationally acetylate γ and β actin, which has acidic N termini, MEEEIAAL and MDDDIAAL, respectively. NatB, C, E and F acetylate the N-terminal methionine of substrates with the respective N termini sequences. The retention of methionine could be due to the larger size of the amino acid in the second position compared to those of NatA. Histone 2A and histone 4 (H2A/H4) are thus far, the only two substrates identified for NatD.
In yeast and humans, NatA to C are most well studied, and they show similar composition and substrate preferences [171, 196]. No NatF has been detected in yeast, raising speculation that, Nats perhaps contribute to the complexity of higher eukaryotes. Generally, Nats have one catalytic subunit (containing the Ac-CoA structure) and one or more auxiliary subunits, with different molecular weights [196]. They form a stable complex in vivo, and usually co-sediment with the ribosome, though a fraction is in most cases also detected outside the ribosome [197]. In cancer biology, the significance of Nats has been increasingly recognized [197].

Prior to this work, no detailed investigation of Nt-acetylation in pathogenic protozoa, had been published. I have therefore carried out molecular characterization of enzymes that contribute to this post-translational modification, and assessed the biological roles played by acetylation in parasites and in their ability to infect the host. Moreover, Nt-acetylation may partly explain gene expression regulation in kinetoplastids. The results are presented in the second part of this thesis.
2. AIMS

The general aims of this thesis were to take part in the genome sequencing of *T. cruzi* strain CL-Brener (TcVI) and a subsequent comparison with the genome of another strain, Sylvio X10 (TcI). The comparative genomics study was partly aimed at identifying strain-specific features that may be used in epidemiological and functional studies. In the second part of this thesis, the objective was to use genome data to select genes of particular interest for functional investigation and evaluation of their potential as possible drug targets. The genes chosen were the *T. cruzi* N-alpha acetyltransferase gene family.

Specific aims

**Paper I and II:** The genome sequence of the *Trypanosoma cruzi* TcVI strain CL-Brener, and comparison with TcI SylvioX10.

- To participate in the genome annotation and analysis of *T. cruzi* TcVI and TcI strains and publicly release the genome data.
- To participate in a comparative analysis of the genomes of the two strains TcVI CL-Brener and TcI Sylvio X10.
- To assess what could be behind the genome size differences between the two strains.
- To identify candidate genes for diagnosis, vaccine and drug development.

**Paper III and IV:** Characterization of an acetyltransferase of *Trypanosoma cruzi* displaying autoacetylation activity, and analyses of the N-alpha acetyltransferase C Protein Complex in trypanosomes.

- Investigate the drug target potential of *T. cruzi* N-alpha acetyltransferase C protein (NatC).
- Identify *T. cruzi* NatC protein subunits and analyze whether they form a complex.
- Investigate the subcellular location of the proteins and their association with the ribosome.
- Assess NatC expression across the *T. cruzi* life cycle.
- Determine the activity and substrate-specificity of the NatC enzyme.
- Predict the NatC acetylation state in trypanosomes.
- Investigate functional significance of the NatC catalytic subunit in *T. brucei* by knockdown of this gene using RNAi.

**Paper V:** Analysis of the N-alpha acetyltransferase A protein complex in *T. cruzi*.

- Identify *T. cruzi* NatA subunits and characterize the gene.
- The other specific aims are identical to Paper III and IV.
3. PRESENT INVESTIGATION

3.1. METHODOLOGICAL CONSIDERATIONS
The full details of the methods used in this study are outlined in papers I to V. In this section, the techniques used are summarized.

3.1.1. Genome sequencing
The initial approach to sequence the *T. cruzi* TcVI was a clone-by-clone strategy, *i.e.*, sequencing the genome using overlapping bacterial artificial chromosome (BAC) clones. Unfortunately, the BAC-clone approach was not successful due to uneven BAC coverage, most likely due to the repetitive nature of the genome. Instead, a whole genome shotgun approach was adopted. In this method, the whole genome was sheared into small pieces (fragments) and each fragment was sequenced from both ends. Thus, each DNA fragment gave rise to two sequencing reads. The large amount of random sequences were subsequently processed and assembled *in silico*. Bioinformatics tools were used to assemble the reads into contigs, which reflect chromosomal regions and subsequently into scaffolds (supercontigs). The open reading frames were subsequently annotated using a comparative approach.

The genome of the Tcl Sylvio X10 strain was also sequenced using shotgun sequencing, but using a second-generation sequencing instrument (Roche 454). Roche 454 allows greater throughput in terms of sequence data at the cost of shorter sequencing reads. The genome was subsequently assembled and annotated. Manual annotation was performed, erroneous frame shifts were corrected and the genome was analyzed using various bioinformatics tools and compared with the CL Brener genome.

3.1.2. Molecular functional characterization of an acetyltransferase gene family
Open reading frames of the genes of interest were amplified from genomic DNA by PCR and cloned into an appropriate plasmid expression vector, for example, pGEX5-1. Recombinant proteins were used for the production of antibodies and semi-purified proteins were used for *in vitro* acetylation enzymatic assays and GST-pulldown. The enzymatic reaction and substrate screening was monitored by HPLC. Immunoprecipitation was used to assess the physical association of proteins *in vivo*. The cellular distribution of the proteins of interest was analyzed using immunofluorescence. Western blotting was used to determine protein expression throughout the life cycle of the parasite. The functional significance of the catalytic subunit of *T. brucei* NatC was investigated by RNAi-mediated gene knockdown.
3.2. RESULTS AND DISCUSSION
The population structure of *T. cruzi* has to a great extent resulted from clonal propagation. A high degree of variability of the genome size between different strains, and among different clones of the same strain [134, 135, 198], as well as a high degree of sequence variation are notable features of *T. cruzi*. The chromosomes are poorly condensed during cell division and this initially hampered genomic studies in that the identification of chromosomes by conventional cytogenetics was difficult. Differences in the capability of the parasite to infect different mammals and insect species and invade different tissues, as well as clinical outcome warrants sequencing of multiple strains for comparative purposes. Such information, coupled with environmental factors, will provide a better picture of parasite evolution and the epidemiology of Chagas disease.

**Paper I:** This paper describes the first complete genome analysis of a *T. cruzi* strain. This work was carried out by our group at Karolinska Institutet together with collaborators in the USA: (Seattle Biomedical Research Institute (SBRI)) and (The Institute for Genomic Research (TIGR)). We found that the sequenced strain had a high degree of allelic variation, which was attributed to a recent hybrid origin of the CL Brener strain. We thus described two different haplotypes, designated Esmeraldo-like and Non-Esmeraldo-like, for most of the genome [74, 199-201]. The haplotypes have almost identical gene content, and homologous coding sequences differ by on average 2.2 % in sequence identity. The final genome assembly, with 7-14X coverage was assembled into 5489 scaffolds, generated from 8740 contigs. The diploid genome size was estimated to be approximately 110 Mb containing approximately 12,000 genes per haploid genome.

As is the case for kinetoplastids [8, 9], *T. cruzi* protein-coding genes are organized in long directional polycistronic transcription units (PTU), similar to bacteria operons [7, 202]. About 20 or more of these proteins are present on one strand. PTUs are separated by strand switch regions, which have been postulated to, in many cases be involved in transcriptional initiation and termination as well as regulation of DNA replication [7, 202-204]. Surprisingly, no DNA motifs have been identified in strand-switch regions, suggesting that transcription-level control of gene expression is not common in these organisms. An extensive genome comparison of *T. cruzi*, *T. brucei* and *L. major*, published at the same time [11], indicated that 95% of the gene order in core regions is conserved. Sequencing of TriTryps and coordinating the analyses has bridged the gap between scientific communities working with the three pathogens, raising the possibility for common strategies for drug design against all the organisms.

Interestingly, about 55 % of TcVI genes encode proteins of unknown function, suggesting that they may be specific to the parasite. Additionally, around 50% of the genome is repeated. This includes regions encoding genes for surface proteins, such as, trans-sialidases (1430 members), mucin-associated surface proteins (1377 members), mucins (863 members), DGF-1 (565 members) and surface glycoproteins GP63 protease (425 members) [7], and other repeated genes. The significance of these surface molecules in parasite biology, which includes the ability of the parasite to use them to invade the host, has recently been reviewed in [205]. Also, about 5% of the genome is made up of retrotransposons. Some transposable elements are active and may cause mutations and genomic rearrangements [206]. They are categorized as long terminal repeats (LTR) retrotransposons and non-LTR elements [7, 202]. In conclusion, while the TriTryp genomes share synteny for most genes, the *T. cruzi* genome appears to be significantly more
repeated. It can be speculated that the *T. cruzi* genome is more prone to instability, which may represent an evolutionary adaptation.

To improve the *T. cruzi* genome annotation, Weatherly *et al* [207] organized the scaffolds and contigs into 82 diploid chromosomes using the sequence data from the genome project. The proposed 41 chromosome pairs predicted is, while slightly lower in number possibly due to chromosomes missed because of assembly issues in repetitive regions [207] consistent with our experimental data generated by PFGE [130].

To date, no technique has been able to completely characterize the *T. cruzi* karyotype. Thus, the precise chromosome count of the CL-Brener strain is not known [137].

In conclusion, several unique genetic features of the parasite have been identified, and this has increased our understanding of parasite biology. Thus, the reference genome is an essential tool, which has laid a foundation for sequencing of other *T. cruzi* strains for comparative genome analyses and it has accelerated the entire field of *T. cruzi* cell biology and Chagas disease research.

**Paper II:** Since the completion of the first genome, we have published two additional strains, and more are currently being sequenced and analyzed. To build on the knowledge of the sequenced reference strain TcVI, we sequenced the genome of the TcI strain Sylvio X10. The two strains occupy different geographical areas: TcI is mainly confined to the area north of Amazon; and TcVI predominate the southern corn countries of Latin America. Furthermore, compared to TcVI, TcI is generally accepted as characterized by low parasitaemia, less acute infection profile, associated with chronic infections and different disease tropisms [140, 208]. Due to the phenotypic differences, we decided to produce a TcI reference genome sequence in order to explore genetic differences between clades.

In contrast to the Sanger sequencing technology used for TcVI, the TcI genome was sequenced using the 454 platform and assembled into 7092 contigs with genome coverage of 11X. As a result of the comparative analysis, CL Brener genome annotation was improved by the addition of 169 genes previously annotated as pseudogenes due to the presence of frame shifts, likely as a consequence of sequencing errors, which were not observed in TcI. The genome size was estimated using sequence read alignments, which indicated that the haploid TcI genome is approximately 44 Mbp in size, i.e., 11 Mbp smaller than TcVI. The surface molecules gene containing regions are 5.9 Mbp shorter in TcI, due to the presence of fewer copies of multi copy gene family members, such as MASP, mucin, DGF and GP63. The fact that the TcVI genome was found to be larger than that of TcI, confirmed our previous results from *T. cruzi* comparative karyotyping of three different *T. cruzi* genomes [130] (Figure 5).

Six reading frames were identified in TcVI, that were not detected in TcI. All are hypothetical proteins ranging from 155 to 337 amino acids in size, (Figure 8). Genes 1 to 6 are located on chromosome 40, 40, 6, 41, 13 and 23, respectively. The subcellular distribution of proteins 1 to 5 were assigned by CELLO version 2 [209] to plasma membrane and transmembrane regions or helices were predicted. Protein 6 on the other hand, was predicted to be located in the nucleus. We can speculate that these hypothetical proteins may possibly communicate with host surface proteins and this may contribute to the higher TcVI infectivity. However, this needs to be tested experimentally.
The molecular differences between TcVI and TcI strains also included genetic polymorphisms in terms of SNPs and indels. Specifically, TcI usually lacked certain sequences and microsatellites were longer in TcVI, suggesting either insertions in TcVI or deletions in TcI during evolution.

<table>
<thead>
<tr>
<th>CL Brener gene id</th>
<th>Product description</th>
<th>Length (aa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tc00.1047053510275.90</td>
<td>Hypothetical protein</td>
<td>337</td>
</tr>
<tr>
<td>Tc00.1047053510279.240</td>
<td>Hypothetical protein</td>
<td>300</td>
</tr>
<tr>
<td>Tc00.1047053510073.24</td>
<td>Hypothetical protein</td>
<td>305</td>
</tr>
<tr>
<td>Tc00.1047053503973.10</td>
<td>Hypothetical protein</td>
<td>155</td>
</tr>
<tr>
<td>Tc00.1047053511807.125</td>
<td>Hypothetical protein</td>
<td>229</td>
</tr>
<tr>
<td>Tc00.1047053508811.55</td>
<td>Hypothetical protein</td>
<td>305</td>
</tr>
</tbody>
</table>

Figure 8. The identified TcVI-specific sequences, and their corresponding GeneDB identifiers. The numbers on the left correspond to the respective genes, described in the text.

Comparison of TcI with each TcVI haplotypes at the nucleotide levels, suggests that the nucleotides differences is 1.8 % and 2.5 % for non-Esmeraldo and Esmeraldo, respectively. This genome analysis result indicates that non-Esmeraldo is more similar to TcI. The functional significance of this result cannot be known from the current data alone.

Most genes were found to be under purifying selection, while a few are under positive selection pressure, that is, they have dN/dS greater than one, suggesting that, they are evolving more rapidly. This included gene Tc00.1047053509769.70, annotated as a DNA polymerase and also hypothetical proteins: gene id TcVI (Tc00.1047053510299.70, Tc00.1047053511909.30 and Tc00.1047053431849.20) with dN/dS 1.263, 2.007 and 1.312, respectively. These could be of interest for functional investigation, as they contain features such as trans-membrane domains and signal peptides.

Collectively, the genetic differences between TcI and TVI may help explain differences in pathogenesis and clinical outcome.

The completion and availability of the genomes and subsequent comparative genomics of pathogenic parasites such as Plasmodium [178], Schistosoma [177], Giardia [210] and kinetoplastids [11] have provided the full gene complement and a lot of information regarding genetic variation in these parasites, and may provide candidates and strategies to facilitate drug and vaccine development as well as control: i) One of the best options could be the mining of the genome for parasite-specific proteins, absent in the host, for specific drug design against the pathogen while minimizing the risk for side effects, ii) Identification of proteins that have been described to be possible drug targets in other organisms that can be functionally characterized to ascertain their drug target potential, iii) Orthologs of proteins or classes of proteins that are being investigated as targets in humans and are linked to important pathways in the parasite [211]. An example of the latter is drugs designed for the treatment of cancer or other pathologies, and that may be used to treat parasite infections. Notably, the anti-cancer drug imatinib that targets protein kinases, has been shown to inhibit Schistosoma mansoni in vitro and may be an alternative chemotherapy [211]. Vorinostat, a compound targeting acetylation, used in the
treatment of cutaneous T-cell lymphoma, could be a new drug to control malaria [212]. In another example, aspartic peptidases, which are enzymes that are present in all kingdoms of life and are important in metabolic pathways, are under study as a target for many pathologies [213]. Inhibitors against these enzymes are being used in the treatment of HIV, and the compounds have been found to be potent against kinetoplastids [213].

As mentioned above, N-alpha acetyltransferases are possible candidates and I have therefore studied these further.

**Paper III:** Substrates of N-acetyltransferases have been poorly analysed or described in parasite. Therefore, in this study, we have begun to elucidate the molecular characteristics of predicted acetyltransferases (AT) in *T. cruzi*. We cloned a *T. cruzi* acetyltransferase (*TcAT*) annotated in our laboratory during the genome project and investigated its function.

An *in vitro* acetylation assay using acetyl CoA was performed in order to determine the *TcAT* enzymatic activity. The data indicated that the purified recombinant protein had an autoacetylation activity, suggesting that it most likely carries out Nε acetylation. We performed an immunofluorescence experiment to assess the sub-cellular distribution of the protein in the parasite by incubating CL Brener epimastigotes strain with purified anti-*TcAT* antibodies. The *TcAT* protein was found to localise in the cytoplasm (Figure 9).

![Figure 9. Cellular localisation of N-acetyltransferase.](image)

1. Nucleus (N) and kinetoplast (K) were stained with DAPI (blue colour). 2. Staining with anti-*TcAT* antibody. 3. Overlay of DAPI and antibody staining. 4. Phase contrast image of the parasite. Scale bar=5µm.

The heterogeneous population structure of *T. cruzi* prompted us to determine stage expression of *TcAT* by Western blotting in different *T. cruzi* strains. These included TcVI (CL-Brener), TcII (Y strain) and TcI (CAI strain). The results showed that the protein was constitutively expressed in the three different life-cycle stages (epimastigote, trypomastigote and amastigote) in these parasites strains.

*In silico* analyses indicated that the *TcAT* protein motifs matched those previously described in members of the GCN5-related acetyltransferase (GNAT) family. Orthologous proteins were found to be present in *T. brucei* and *L. major*. The protein appears to be more closely related to N-terminal ATs than to HATs. Moreover, the 3D topology model prediction results for *TcAT* are typical of the GNAT super family of ATs. The N-terminal of the most closely related human protein (HYPO_HUMAN) is extended and there are three additional alpha helices predicted, compared to Kinetoplastids. The functional site residues where Acetyl-CoA binds were conserved except for glycine and Isoleucine (Yeast and human) in positions 94 and 95 (MAK3_YEAST) to instead Arginine and leucine, respectively, for *TcAT*, *T. brucei* and *L.
major. The predicted structural differences between the Kinetoplastid and human enzymes as well as the amino acid differences at the active site could be exploited and used in drug design. These results warrant further study, including the investigation of proteins that interact with TcAT, native substrate preferences and metabolic pathways that are affected by TcAT. These efforts are described in paper VI, and are discussed below.

**Paper IV**: The closest human hypothetical protein to TcAT identified in paper III, shares 44 % identity at the amino acid level. About two years after paper III was published, Arnesen et al characterized the human protein, which turned out to be the catalytic subunit of human NatC (hNaa30), based on the native substrate profile [214]. Moreover, the study identified three proteins that constituted the components of human NatC. The catalytic subunit was functionally assigned to cell differentiation and viability in a cancer cell line, and it was suggested that this enzyme could be a possible target for cancer therapy [214]. These results raised the question whether the TcAT we partially analysed before, could be *T. cruzi* NatC.

We investigated the substrate preference of the enzyme by purifying the recombinant protein and subjecting it to synthetic oligopeptides in the presence of Acetyl-CoA. The peptides had different amino acids at their N-termini and the acetylation of each peptide was assessed. As expected, the protein acetylated methionine when the second residue was a hydrophobic amino acid. No acetylation was observed when NatA substrate (STPD/EEEIA) or NatB (MDEL) were tested. This result confirmed that the protein under study was indeed the *T. cruzi* NatC catalytic subunit, and that the TcNatC catalytic subunit is evolutionarily conserved from parasite to humans with respect to substrate preference. Thus, we named the *T. cruzi* protein the catalytic subunit of NatC (TcNaa30). The native substrate signature of TcNaa30 identified in this study, and the auto acetylation activity detected previously [215], suggested that the TcNaa30 may have both Nε and Nα acetylation functions, as described for human NatA [216].

Based on previous studies in other organisms, we predicted the possible TcNatC auxiliary subunits and identified protein-coding gene Tc00.1047053511755.119 located on chromosome 17 and Tc00.1047053507209.10 assigned to chromosome 23, as the likely subunits.

We named the proteins TcNaa35 and TcNaa38, respectively, in accordance with the current recommendation [217]. TcNaa38 is approximately 12.5 kDa in size. It was previously annotated as small nuclear ribonucleoprotein. These are components of the spliceosome, raising speculation, if the annotation is correct, that it may also be involved in spliceosome function. The BLAST searches also identified a *T. cruzi* protein Sm-like LSM8; GI Tc00.1047053510091.70 (Esmeraldo-like) located on chromosome 40, suggested to be associated with RNA degradation and spliceosome pathways. Compared with hNaa38, both proteins have the same sequence identity of 38 %, suggesting that Tc00.1047053510091.70 may possibly be part of the TcNatC complex. However, a protein interaction assay using a bioinformatics tool indicated that TcNaa38 is more likely one of the subunits of TcNatC. We can, however, not completely rule out the hypothesis that the LSM8 may be a subunit of TcNatC without experimental verification.

Recombinant proteins were produced for the predicted auxiliary subunits and a polyclonal antibody was generated against TcNaa38. GST-pull down and immunoprecipitation experiments suggested that the three proteins might physically form a stable complex *in vivo*. One of the characteristics of Nats is the potential to co-sediment with the ribosome. In our assay, TcNaa30 and TcNaa38 both co-sedimented with the ribosome, but a fraction of the protein was detected outside the ribosomal fraction, indicating that TcNatC may have ribosomal and non-
ribosomal functions as suggested previously [214]. The proteins were found to be expressed in the epimastigote, trypomastigote and amastigote stages of *T. cruzi*. In this experiment, expression of TcNaa38 was not detected in metacyclic trypomastigotes. If confirmed, this may suggest that protein translation is inactive at this stage. Another possibility is that the enzymatic activity of TcNaa30, if functional at the metacyclic trypomastigote stage, may not always require auxiliary subunits.

We examined the cellular locations of TcNaa30 and TcNaa38 in the different parasite stages using immunofluorescence. In epimastigotes and amastigotes, TcNaa38 exhibited a punctate distribution while there was a predominantly cytoskeletal staining profile in trypomastigotes. There was no TcNaa38 staining in metacyclic trypomastigotes, which is in agreement with the Western blot result. Interestingly, as shown in (Figure 10), TcNaa30 was mainly located in the cytoplasm in the epimastigote stage, whereas in metacyclic and culture derived trypomastigotes, the staining showed a cytoskeletal like structure. In the amastigote stage, TcNaa30 was found mainly in the cell periphery, suggesting that it might shuttle between the cytoplasm and the nucleus.

![Figure 10. Localisation of TcNaa30 in four developmental stages of the parasite. Cells were immunolabelled with anti-TcNaa30. The nucleus and kinetoplast were visualized using DAPI stain. Scale bar=5µm.](image)

In addition, we examined the staining pattern of TcNaa30 and Tc38 in amastigotes inside cells. TcNaa38 showed a cytoskeletal profile as in the trypomastigote stage. TcNaa30, on the other hand, exhibited a cytoplasmic pattern, similar to the stationary epimastigote stage. The differences in the TcNatC staining profiles suggest that they differ depending on the developmental stage, suggesting differential localisations and perhaps functions. The cell state, environmental factors or possibly phosphorylation, may influence the protein localisation. Indeed, in this study, multiple phosphorylation sites were predicted for TcNaa30. Variable cellular distribution of Nat proteins in *e.g.*, has previously been reported by Hole *et al* [218]. Additional controls will be needed in order to verify these patterns.

Several hundred annotated proteins from trypanosome and leishmania genomes were found to contain the NatC N-terminus substrate sequences (Met-Leu-, Met-Ile-, Met-Phe-, Met-Trp and Met-Tyr). The abundance of substrates predicted indicated that NatC might be heavily involved in various cellular processes in kinetoplastids. For *T. cruzi*, the majority of the proteins predicted are specific to the parasite, or linked to key metabolic pathways. Hypothetical proteins and trans-sialidases (TSs) were overrepresented. The TSs are proteins used by the parasite to
invade the host and are candidates for vaccine and drug developments [219]. Similarly, important T. brucei surface molecules such as receptor-like adenylate cyclases and variant surface glycoproteins were predicted to be potential substrates.

Assessing gene function in T. cruzi using RNAi, is not possible, since DICER and the Argonuate proteins essential for this pathway are not present. We therefore, carried out RNAi knockdown of the orthologous gene in T. brucei. The initial results indicated that the protein is important for parasite growth. This is work in progress, however, additional controls must be included, especially, measurements of the RNA and protein levels during the knockdown.

Though related, T. cruzi and T. brucei are different parasites with unique life cycles, and different mechanisms of pathogenesis. It may thus in the future be better to study each species individually for in-depth understanding of the biological roles played by closely related proteins of interest. This is likely to be the case also for Nats, as inhibition of the enzyme may simultaneously affect multiple different proteins, of which many are species-specific. Probably, the knockdown might interfere with infection. This, and the high degree of difference between parasite and human Nats justify interrogation of these enzymes as potential drug targets.

**Paper V:** NatA is the most well-characterized Nat in human and yeast, and it is evolutionarily conserved regarding to substrates that it acetylates, *i.e.*, Ser, Ala, Thr, Gly, Val and Cys at the N-terminal [187, 193] (see table 2 in introduction). For acetylation to occur, the first methionine must first be removed by methionine amino peptidase enzymes. The human NatA catalytic subunit (hNaa10) also known as (ARD1p), and the auxiliary subunit hNaa15 (Nat1p), form a complex that co-sediments with the ribosome [220]. Moreover, both proteins are in some case required for NatA activity associated with translation [220]. In higher eukaryotes, NatA function is more complex due to gene duplication or splice variants, which results in different functional proteins [187]. In addition, many proteins have been shown to interact with NatA, for example hypoxia inducible factor (HIF) [187, 221], further expanding the functional possibilities of the NatA. A wide range of substrates is affected by Nt-acetylation mediated by NatA. The biological significance of these, including involvement in cancer biology is reviewed in [187, 196, 197].

The first functional analysis of a trypanosomatid NatA was carried out by homologous recombination in T. brucei by Ingram, Cross *et al* [222]. The study showed that the protein is essential for parasite viability in both the mammalian and insect stage. No information was available in the literature about T. cruzi NatA, and its biological significance. We identified the genes *TcCLB.506227.230* and *TcCLB.504163.110* and named them *T. cruzi* NatA catalytic (TcNaa10) and auxiliary (TcNaa15) subunit, respectively, with molecular weights of approximately 30 and 83 kDa, respectively. Both proteins have orthologs in other trypanosomatids. For functional analyses, we generated recombinant protein (GST-TcNaa10) and (GST-TcNaa15) in bacteria. In comparison with the auxiliary subunit, the catalytic subunit showed poor solubility, even when we changed the vector from pGEX5-1 to pETM41 containing a maltose binding protein (MBP). This problem made it difficult to produce sufficient amounts of the protein for native enzyme substrate screening. However, unconfirmed data, discussed under preliminary result section 4, indicated typical NatA substrate preferences. The active sites responsible for auto acetylation activity, as was shown in human [216], were mapped by sequence comparison. TcNaa10 showed a high sequence identity, about 60%, to hNaa10. These results suggest that the NatA catalytic subunit in *T. cruzi* may be similar in function to the human enzyme.
The recombinant (GST-TcNaa10) protein and specific synthetic peptides for TcNaa15 were used to produce polyclonal antibodies in rabbit. The purified antibodies were able to recognize the three main developmental stages of *T. cruzi*: epimastigote, trypomastigote and amastigote, indicating that, both subunits are constitutively expressed during parasite development. Moreover, anti-TcNaa10 was able to detect the *T. brucei* protein whereas anti-TcNaa15 was not, which confirmed the specificity of anti-TcNaa15. An immunoprecipitation experiment demonstrated that the proteins interact, suggesting they may be physically associated *in vivo*. A typical hallmark of Nat proteins is their ability to co-migrate with the ribosome, indicating ribosomal function. Indeed, recent finding in yeast associated Nt-acetylation with ribosomal biogenesis [223]. In our hands, the *T. cruzi* subunits were present in both the ribosome fraction and outside the ribosome.

We investigated the subcellular location of the TcNatA subunits by incubating the parasite developmental stages with the respective antibodies. The majority of TcNaa15p showed cytoplasmic-cytoskeletal like staining profile in all the different cells tested, with minor variation, for example, the disappearance of cytoplasmic staining in metacyclic trypomastigotes and diffuse cytoplasmic exhibition in the trypomastigote stage. The TcNaa15 cytoskeletal distribution is similar to that reported for the human auxiliary NatA protein (hNaa15) [220]. As shown in Figure 11, in epimastigotes, TcNaa10 was predominantly found to be located in the nuclear surface with a small fraction in the cytoplasm. It is possible that the protein moves between the nucleus and the cytoplasm. A small amount of TcNaa10 was observed in metacyclic trypomastigotes and we saw a reduction in tissue derived trypomastigotes compared with epimastigotes. Amastigote staining showed a cytoskeletal distribution of the protein. As for TcNatC, the staining patterns of TcNatA proteins do vary, suggesting differential regulation and expression. Taken together, it appears that *T. cruzi* Nat proteins are regulated in order to direct specific cellular events. The mechanisms behind the regulation of the proteins’ expression as well as their differential localization remain speculative.

![Subcellular localisation of TcNaa10 in four developmental stages of the parasite.](image)

**Figure 11.** Subcellular localisation of TcNaa10 in four developmental stages of the parasite. *T. cruzi* cells were immunolabelled with anti-TcNaa10, the nucleus and kinetoplast were visualized using DAPI stain. Scale bar = 5µm.

Human NatA contributes approximately 40% of all Nt-acetylated proteins and this includes key players in various cellular pathways [187]. Based on the typical NatA substrate; Met-Ala, Met-Ser, Met-Thr, Met-Gly, Met-Val or Met-Cys N-termini, we extracted protein
sequences from the CL-Brener genome sequence and selected potential substrate proteins using the Terminator algorithm. Many proteins were predicted to be N-terminally acetylated by NatA, more than those predicted for *T. cruzi* NatC. Interestingly, hypothetical proteins were overrepresented, followed by MASP. This differs from TcNatC, which was predicted to acetylate many trans-sialidases. This may suggest that each Nat in *T. cruzi* is assigned a specific role in parasite biology and in infection. An experimental test of the TcNatA substrate profile will be needed to verify the biological role of NatA in *T. cruzi*.

### 3.3. Preliminary result

#### 3.3.1. *In vitro* acetylation assay

NatA acetylates proteins with specific peptide sequences, as determined in yeast and human [224]. We performed an *in vitro* NAT assay to assess whether TcNaa10 has acetyltransferase activity and if the substrate-specificity is similar to that of human Naa10. The following conditions were used to prepare recombinant (GST-TcNaa10) protein.

*E. coli* cells harboring the expression plasmid pGEX5-1-TcNaa10 were grown at 37 °C in LB medium containing the appropriate amounts of ampicillin. Expression was induced at an OD600 of about 0.5 by the addition of 0.3 mM IPTG and growth was continued for another 25 hrs at 17 °C, 189 rpm. Protein was prepared using glutathione-Sepharose 4B (GE Healthcare). The amount of protein on the beads was estimated from Coomassie staining of SDS-PAGE gels.

The purified recombinant protein (GST-TcNaa10) was incubated with Acetyl-CoA and synthetic peptides suggested to be the substrates for NatA. The activity of the enzyme was stopped after 30 minutes and the results were analyzed using HPLC. As shown in (Figure 12), the unconfirmed data indicated that TcNaa10 preferentially acetylated the synthetic peptide sequences STPD and EEEIA. The STPD peptide represents the N-terminus of a classical NatA substrate [189], while the acidic EEEIA-peptide represents the gamma-Actin N-terminus that is likely to be a non-canonical substrate post-translationally acetylated by Naa10 [225]. The N-terminal peptides representing a semi-optimal NatA substrate (AVFA), a NatB-substrate (MDEL) or a NatC/E/F-class substrate (MLGP) were not significantly acetylated. These data tentatively indicated that the TcNaa10 is a true orthologue of hNaa10 and yNaa10 and that NatA substrate specificity is conserved between trypanosomatids, yeast and higher eukaryotes. However, this signal was weak overall and an attempt to reproduce the results was not successful, partly due to the difficulty to generate sufficient amounts of soluble protein. More work is needed for confirmation.
Figure 12. In vitro acetyltransferase activity of recombinant GST-TcNaa10. GST-TcNaa10 (TcNaa10p) was incubated with acetyl-CoA (300 mM) and selected oligopeptides (300 mM) for 30 min at 37 °C. dH2O was used as negative control. The amount of acetylated peptide was determined with reverse phase HPLC. Oligopeptide names indicate the first four amino acids from the N-terminus.

3.4. CONCLUDING REMARKS AND PERSPECTIVES

Chagas disease, sleeping sickness and leishmaniasis are categorized as neglected tropical diseases by WHO. They have great economic impact and health consequences for humans and domestic animals. For T. cruzi, many strains infect humans, and these often show different disease outcomes, perhaps as a consequence of host and parasite genetics. It is difficult to morphologically distinguish them, though they may be genetically different. There is no vaccine for these diseases. The current drugs are far from ideal due to their toxicity and relative inefficiency. This makes genome sequencing of the parasites, in order to find new drug targets and biologically important variation, useful.

Before this study, no T. cruzi or any other kinetoplastid genome was sequenced. In the present work, the first T. cruzi genome CL Brener genome (Paper I), which was subsequently compared with T. cruzi Sylvio X 10 (Paper II) strain, has provided insight into the biology of the two parasites. Differences and genetic polymorphisms between the two genomes may contribute to the phenotypic differences in term of parasite infectivity. Thus, more strains need to be sequenced in order to identify functional variants.

- We have published a third T. cruzi genome, derived from the bat. This subspecies does not infect humans [226]. Interestingly, a unique acetyltransferase gene was identified. Three additional genomes have been sequenced and are under analysis. More T. cruzi genome sequences are in progress, and will provide resources to the scientific laboratories working with this parasite.
- Relevant post-genomic analyses have been reviewed elsewhere [177, 202].
- A potential next step would include to functional studies of hypothetical proteins using for example RNAi. This could help identify new drug targets. In addition, extensive proteomics work will make it possible to identify differentially expressed genes and abundant proteins.
Cross-species comparisons of, for example *T. cruzi* surface molecules or any other molecules of interest and that of the host could now be possible. This approach has been used for *T. brucei* as described in [227].

The relatively fragmented nature of *T. cruzi* genome sequences, in large part caused by the repetitive nature of the genome, can possibly be improved by using additional sequence data sets and improved bioinformatics methods. The genome sequences have greatly stimulated the field and they form the basis for functional characterization of genes of interest for vaccine or drug development, and specific markers or elements identified could be used for diagnostic purposes.

**Paper II-V:** In molecular biology, the cellular functions of a protein may partly depend on where and when the proteins are expressed. This allows the correct sequence of cellular events. Nats are involved in diverse cellular events, such as cell division, cell survival/death, movement or differentiation. Individual proteins are regulated by PTMs, such as those mediated by Nats, in order to accurately carry out biological functions.

In this thesis, we showed that Nats are present in *T. cruzi*. Acetylation, especially Nt-acetylation in kinetoplastids is poorly studied. In contrast, it is well characterized in mammals and yeast, which provided the opportunity to extend the knowledge to trypanosomatids. As evidenced by sequence alignments, there are significant differences between the acetyltransferases of kinetoplastids and human. Our analyses showed that *T. cruzi* NatC and A together with their respective subunits are expressed in the three main distinct life cycle stages of the parasite and, intriguingly, display different subcellular localizations, suggesting differential regulation. The predicted acetylated proteins were predominantly parasite specific. It is possible that NATs exert their effects on key proteins essential to circumvent the host immune system. Some specific surface proteins involved in host-parasite communication were predicted to be acetylated. Typical examples are trans-sialidases (TS) and mucins for *T. cruzi*, Variant Surface Glycoproteins (VSGs) for *T. brucei* and Gp63 surface molecules in *Leishmania*, highlighting the potential significance of Nats in parasite biology.

The high degree of conservation of Nat substrate specificity may result in drugs that are effective against more than one human pathogen. This phenomenon has been seen for other drugs. For example the treatment of schistosomiasis with artemisinin, a malaria drug, is under consideration [228]. The analyses of Nats presented in this thesis raise more questions, which require scientific verification and follow up:

- The differential localization of *T. cruzi* Nats as assessed by immunofluorescence could be confirmed using another technique, for example, tagging the protein with GFP for live imaging, or investigating co-localization with known proteins assigned to particular compartments of interest. The effect of phosphorylation if any should be analysed to assess whether it influences the observed staining profiles.
- It is necessary to confirm the interaction of the different Nats or identify more protein interaction partners by, for example, immunoprecipitation and analysing the result using mass spectrometry.
- If possible, identify all the Nats in the parasite. Our *in silico* analyses not reported here, indicate the presence of NatB in *T. cruzi* and it is predicted to acetylate many proteins in the parasite.
- To further understand the significance of Nats in trypanosomes, reverse genetic analyses or mutation studies for all the subunits will be important to ascertain phenotypic
differences and whether the individual auxiliary subunit influences the catalytic subunit. In addition, comparative proteomic studies will define the substrate profiles.

- Given the high identity of the trypanosome Nat catalytic subunits with those of humans, it is possible that, specific inhibition will be difficult to achieve. The approach could be redirected to the auxiliary subunits instead, on condition that they influence the enzyme activity and parasite growth.
- The acetylation states in trypanosomes are not known. To confirm in silico prediction, an acetylome investigation using a proteomics approach should be carried out.
- Knowledge of the acetylation state will make it possible to investigate the downstream or upstream effect of acetylation in T. cruzi.
- We need to attempt to prohibit the parasites the capability to use acetyl.
- It is important to identify proteins that are involved in the pathway leading to the production of acetyl-coA in the parasites, also as potential drug targets.

In summary, this thesis has provided information for improved understanding of the parasite genome, which can be used for both epidemiology and cell biology studies. In addition, I hope that this study will facilitate hypothesis that parasite survival depends on N-terminal acetylation. Thus, Nats may turn out to be useful drug targets for the control of parasitic diseases.
4. Acknowledgements
This work has been carried out at KAROLINSKA INSTITUTET, Department of Cellular and Molecular Biology, Stockholm, Sweden. I did some of the genome mapping (karyotyping) at the beginning, at Rudbeck Laboratory, Uppsala University, Department of Genetics and Pathology where my supervisor was working at that time.

I wish to express my sincere gratitude to Professor Christer Höög and study director Matti Nikkola for opening their doors and corridors to me.

My deepest heartfelt thanks go to my main supervisor PROFESSOR BJÖRN ANDERSSON for having trust in me and accepting me as his graduate student and introducing me to the real world of science and research. This rare opportunity has immensely improved my academic standing and I thank you for your tolerance and patience, for supporting me all this time and for guidance throughout this work. I really value and admire your intellectual skills and potential. I wish you all the very best in your future career and family.

Many thanks to my second supervisor, Associate professor Lena Åslund, Uppsala University, for intellectual input to my work. Lena, your academic skills and endless new ideas have tremendously contributed and guided me to successful accomplishment of this work. Your support and constructive criticisms are greatly acknowledged. Along the way, I have realized that the philosophy of science and research are not the skewed pictures preached in secondary schools by reading textbooks.

I am also in debt to my third supervisor, Professor Martin Rottenberg, Karolinska Institutet for supporting me throughout this work.

I am truly indebted to Dr. Esteban Bontempi, a visiting researcher from Argentina, who at the beginning of my work, gave me inspiration and stimulation. I learned a lot of cloning techniques and other laboratory work from you.

To my mentor, Professor Mats Wahlgren. I am very grateful for all your support, and for helping me with contacts with other scientific laboratories and for honest advice, which will guide me for the rest of my life.

Many thanks to co-authors and collaborators for fruitful sharing of scientific experiences: Dr. Johan Eriksson and his group, previously at Uppsala University, Professor Thomas Arnessen and his group at University of Bergen, Norway, Professor Énock Matovu and his group, Makererer University, Uganda, Professor Kevin Tyler and Dr Claire Butler, Biomedical Research Centre, Norwich Medical School, University of East Anglia, Norwich, Norfolk, UK, for all your contributions.

Professor Ulf Hellman, Ludwing Institute in Uppsala, BMC, for connecting me to Dr. Eriksson’s lab where I was able to perform the in vitro acetylation assay, Dr. John Andersson, Karolinska Institutet, for teaching me how to handle parasites. Many thanks to all co-authors for your contributions to the papers presented in this thesis.
The Andersson group at CMB: Though some of you are no longer in the group, you are all remembered for your inspiration. **Oscar Franzén** for all your help and support, for the scientific discussions and jokes during lunch time and beyond. Oscar, I am really going to miss that for a long time to come. Your bioinformatics skills have greatly helped me throughout this work. Moreover, you are always willing to share scientific thoughts with me. **Stefanie Prast-Nielsen**, **Johannes Luthman**, **Fredrik Lysholm** and **Carlos Talavera-Lopez** for all scientific discussions.

**Carole Branche**, who made a huge contribution to the karyotyping paper. **Marcela** and **Shane** for all experiences, discussions, jokes and for a good working environment. **Ellen, Erik, Martti, Anh-Nhi** and especially, **Daniel** for Bioinformatics contributions and many good ideas. **Hamid, Yumi, Daryouch, Alan, Sindy, Kimberly Andersson, and Delal**, for all your assistance and support to me, and to the project.

I really appreciate my fellow graduate students at CMB and those at KI at large, whom I associated with and all those graduate international students I met during various courses, for the stimulating scientific discussions. What a wonderful career to be in medical science, where one can meet people from all walks of life from across the globe and discuss a common language of understanding life in the context of molecular biology to improve human health.

With great appreciation to those researchers and staff at CMB, including administrators, who have given me help and support in different ways.

To you Emma, Pauline, Fredrick Mashili, Reginald Austin, Sidney Carter, Daniel Hagey, Suman, and others, for scientific discussion. Many thanks to you Agneta Söderstedt and members of the pathology department, at Karolinska Hospital, Huddinge for all the jokes and support.

There is no need to mention all your names, but you dear friends of mine in Sweden and elsewhere; I really admire your relentless and wonderful support, consolation and encouragement for all these years. I wish all of you the very best in life. Julius and Jarek for a good conduct.

My gratitude also goes to **Rose Alobo, Rose A O, Yan Swerup, Elisabeth Eriksson Berit Vistrand** and all of you my friends for welcoming me and supporting me throughout my study, and when I first landed in Sweden. Your hospitality, encouragement, assistance and other things are deeply acknowledged. Again, to you Yan S, Elisabeth E and Berit V, without you, I would not have managed the many challenges surrounding me in different circumstances.

My special thanks go to all **my children, my entire family**, for understanding me, endurance, sacrifice, constant support and relentless love during all these long periods when I was absent from home. Keep in mind that, I am with you in my heart. To you **Kristina Ochaya**, I have no words of appreciation for all the support during good and bad times. Thank you for helping me to fine tune the parasite life cycle drawing. For now, I can say that the best reward I can give to you is health.

Once again, to you my children and many of you, whom I have taken care of as a parent: **Concy A, Susan A, Julia A, Jenny A, Christopher O, Oscar O, Particia A, Shida F, Agnes A, Samson O,**
Emmanuel O, Okot, Odoch, Ochen, Odong, Felix, Jouyce A, and you the rest, thank you for your obedience and being responsible people. I have seen you grow to where some of you are today, which has inspired me, given me the conscience of proudness, happiness and strength to complete this work. I am confident and encourage you or anybody else to do better than what I have accomplished.

To MY MOTHER, Atuniata Atto Ochaya, I tell you that you deserve all my respect for your remarkable ability, for having given me life, and for the confidence you instilled in me and for all that we went through together on this planet. I am with you whole-heartedly until the very end. To my late father Alfred Abuner Ochaya, although you are not with us anymore and could not live to see us grow because you fell ill when I was barely 15 years old and passed away shortly thereafter, I appreciate and will never forget all you told me when I was a little boy. I am where I am today because of the dignity and values you instilled in me. Just rest there in peace.
5. REFERENCES


