From Department of Cell and Molecular Biology
Karolinska Institutet, Stockholm, Sweden

# *TRYPANOSOMA CRUZI* GENOME PLASTICITY AND EVOLUTION

Carlos N. Talavera-López


Karolinska Institutet

Stockholm 2016

Cover artwork: Synteny map of two *Trypanosoma cruzi* genomes
All previously published papers were reproduced with permission from the publisher.
Published by Karolinska Institutet.
Printed by AJ E-Print AB

ISBN 978-91-7676-370-4

# *Trypanosoma cruzi* genome plasticity and evolution

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By


## Carlos N. Talavera-López

*Principal Supervisor:*
**Prof. Björn Andersson Ph.D**
Karolinska Institutet
Department of Cell and Molecular Biology

*Co-supervisor(s):*
**Lena Åslund Ph.D**
Uppsala Universitet
Department of Immunology, Genetics and Pathology

*Opponent:*
**Prof. Peter Myler Ph.D**
University of Washington - Seattle
Department of Global Health
Center for Infectious Disease Research

*Examination Board:*

**Anders Andersson Ph.D**
Kungliga Tekniska Högskolan - KTH
School of Biotechnology
Science for Life Laboratory

**Teresa Frisán Ph.D**
Karolinska Institutet
Department of Cell and Molecular Biology

**Susanne Nylén Ph.D**
Karolinska Institutet
Department of Microbiology, Tumor and Cell Biology

*A mi familia*

*“The greatest and noblest pleasure which men can have in this world is to discover new truths; and the next is to shake off old prejudices”*

**Frederick II of Prussia**

# ABSTRACT

*Trypanosoma cruzi*, a protozoan from the *Kinetoplastidae* family, is the etiologic agent of Chagas disease, a major public health problem affecting mostly the poorest areas of Latin America. Due to the complex nature of the parasite's genome it has been impossible to produce a complete reference genome sequence, thus hampering the implementation of post-genomic approaches to unveil the mechanisms of generation of antigenic variation and the identification of new drug targets. My doctoral studies have focused on the application of combined genome sequencing and computational methods to produce a complete reference *T. cruzi* genome sequence and perform comparative analyses to better understand the mechanisms that allow *T. cruzi* to evade the mammalian host immune system and to briskly adapt to novel environments.

In **paper I** and **II**, different genome assembly strategies and second generation sequencing technologies were implemented to perform comparative analyses to identify elements of virulence between *T. cruzi* and two trypanosomatids that are non-pathogenic to humans: *Trypanosoma cruzi marinkellei*, a bat-restricted sub-species of the *T. cruzi* clade and the human avirulent species *Trypanosoma rangeli*. The studies reveal the expansion of *T. cruzi*-specific genomic traits specialised in the invasion of mammalian cells.

In **paper III**, using third-generation, PacBio sequencing data it was possible to assemble the complete reference genome sequence of a *Trypanosoma cruzi* isolate from the DTU-I clade. This breakthrough allowed us - for the first time - to explore in detail the genome architecture of the subtelomeric areas where many parasite virulence factors are encoded. One of the most interesting discoveries was the overrepresentation of interspersed retrotransposons and microsatellites in tandem gene arrays coding for surface molecules, hinting at a retrotransposon-driven mechanism of recombination for generating new sequence variants. Whole genome sequencing of 35 *T. cruzi* DTU-I isolates, collected from different locations in the American continent, made possible to identify and characterise the mechanisms of adaptability employed by the parasite.

Finally, **paper IV** analyses the mechanisms of genomic hybridisation in *T. cruzi* and the evolution over time of the hybrid offspring. The analysis revealed that during hybrid formation, the parasite integrates genetic material from each parental strains with the aid of retrotransposons and microsatellites, and the genome of these hybrid isolates moves quickly from a tetraploid to a diploid state. As a result, the hybrid strain has more genetic material,

mostly in the subtelomeres, providing the parasite with a pool of new surface molecule genes with the potential to possibly increase its fitness in a new environment.

In conclusion, the work presented here has advanced the understanding of parasite biology and provided a genomic resource to be exploited for the identification of drug targets and vaccine candidates.

# LIST OF SCIENTIFIC PAPERS

I. Oscar Franzén, **Carlos Talavera-López**, Stephen Ochaya, Claire E. Buttler, Louisa A. Messenger, Michael D. Lewis, Martin S. Llewellyn, Cornelis J. Marinkelle, Kevin M. Tyler, Michael A. Miles, Björn Andersson. **Comparative genomic analysis of human infective *Trypanosoma cruzi* lineages with the bat-restricted subspecies *T. cruzi marinkellei*.** *BMC Genomics* (2012) **13**:531.

II. Patricia Hermes Stoco, Glauber Wagner, **Carlos Talavera-López**, Alexandra Gerber, Arnaldo Zaha, Claudia Elizabeth Thompson, Daniella Castanheira Bartholomeu, Débora Denardin Lückemeyer, Diana Bahia, Elgion Loreto, Elisa Beatriz Prestes, Fabio Mitsuo Lima, Gabriela Rodrigues-Luiz, Gustavo Adolfo Vallejos, José Franco da Silveira Filho, Sérgio Schenkman, Karina Mariante Monteiro, Kevin Morris Tyler, Luiz Gonzaga Paula de Almeida, Mauro Freitas Ortiz, Miguel Angel Chiurillo, Milene Höehr de Moraes, Oberdan de Lima Cunha, Rondon Mendoça-Neto, Rosane Silva, Santuza Maria Ribeiro Teixeira, Silvane Maria Fonseca Murta, Thais Cristine Marques Sincero, Tiago Antonio de Oliveira Mendes, Turán Peter Urmenyi, Viviane Grazielle Silva, Wanderson Duarte DaRocha, Björn Andersson, Álvaro José Romanha, Mario Steindel, Ana Tereza Ribeiro de Vasconcelos, Edmundo Carlos Grisard. **Genome of the avirulent human-infective trypanosome – *Trypanosoma rangeli*.** *PLoS Neglected and Tropical Diseases* **8(9)**: e3176.

III. **Carlos N. Talavera-López**, Louisa A. Messenger, Michael D. Lewis, Juan D. Ramírez, Felipe Guhl, Henán Carrasco, Sofía Ocana, Jaime A. Costales, Edmundo C. Grisard, Daniella C. Bartholomeu, Santuza M. R. Teixeira, María E. Bottazzi, Peter J. Hotez, Barbara Burgleigh, Michael A. Miles, Björn Andersson. **Genome analysis of the *Trypanosoma cruzi* DTU-I clade reveals mechanisms to generate antigenic diversity.** *Submitted.*

IV. **Carlos N. Talavera-López**, Michael D. Lewis, Louisa A. Messenger, Matthew Yeo, Michael A. Miles, Björn Andersson. **Comparative genomic analyses of Trypanosoma cruzi experimental hybrids reveal mechanism of genetic exchange.** *Manuscript.*

# OTHER PUBLICATIONS

- Björn Nystedt, Nathaniel R. Street, Anna Wetterbom, Andrea Zuccolo, Yao-Cheng Lin, Douglas G. Scofield, Francesco Vezzi, Nicolas Delhomme, Stefania Giacomello, Andrey Alexeyenko, Riccardo Vicedomini, Kristoffer Sahlin, Ellen Sherwood, Malin Elfstrand, Lydia Gramzow, Kristina Holmberg, Jimmie Hällman, Olivier Keech, Lisa Klasson, Maxim Koriabine, Melis Kucukoglu, Max Käller, Johannes Luthman, Fredrik Lysholm, Totte Niittylä, Åke Olson, Nemanja Rilakovic, Carol Ritland, Josep A. Rosselló, Juliana Sena, Thomas Svensson, **Carlos Talavera-López**, Günter Theißen, Hannele Tuominen, Kevin Vanneste, Zhi-Qiang Wu, Bo Zhang, Philipp Zerbe, Lars Arvestad, Rishikesh Bhalerao, Joerg Bohlmann, Jean Bousquet, Rosario Garcia Gil, Torgeir R. Hvidsten, Pieter de Jong, John MacKay, Michele Morgante, Kermit Ritland, Björn Sundberg, Stacey Lee Thompson, Yves Van de Peer, Björn Andersson, Ove Nilsson, Pär K. Ingvarsson, Joakim Lundeberg & Stefan Jansson. **The Norway spruce genome sequence and conifer genome evolution**. *Nature* **(2013)** 497:579 – 584.

- Edmundo C. Grisard, Santuza Maria Ribeiro Teixeira, Luiz Gonzaga Paula de Almeida, Patricia Hermes Stoco, Alexandra Gerber, **Carlos Talavera-López**, Oberdan Cunha Lima, Björn Andersson, Ana Tereza Ribeiro de Vasconcelos. ***Trypanosoma cruzi*** **clone Dm28c draft genome sequence**. *Genome Announcements* **2(1)**: e1114-13.

- João Luís Reis-Cunha, Gabriela F. Rodrigues-Luiz, Hugo O. Valdivia, Rodrigo P. Baptista, Tiago A. O. Mendes, Guilherme Loss de Morais, Rafael Guedes, Andrea M. Macedo, Caryn Bern, Robert H. Gilman, **Carlos Talavera-López**, Björn Andersson, Ana Tereza Vasconcelos and Daniella C. Bartholomeu. **Chromosomal copy number variation reveals differential levels of genomic plasticity in distinct *Trypanosoma cruzi* strains.** *BMC Genomics* (2015) **16**:499.

# CONTENTS

# LIST OF ABBREVIATIONS

| | |
|---|---|
| SNV | Single Nucleotide Variants |
| CNV | Copy Number Variant |
| Mbp | Megabasepair |
| InDel | Insertion/Deletion |
| SV | Structural Variant |
| NGS | Next Generation Sequencing |
| IPE | Illumina Paired End library |
| IMP | Illumina Mate Paired library |
| HLA | Human Leukocyte Antigen |
| RNA-Seq | RNA sequencing data |
| RAM | Random Access Memory |
| PCR | Polymerase Chain Reaction |
| dNTP | Deoxynucleotide triphosphate |
| HR | Homologous Recombination |
| NAHR | Non-allelic Homologous Recombination |
| NHEJ | Non-homologous End Joining |
| MMEJ | Microhomology-mediated End Joining |
| LD | Linkage Disequilibrium |
| MARK | *Trypanosoma cruzi marinkellei* |
| RANG | *Trypanosoma rangeli* |

# CHAPTER 1 - INTRODUCTION

## 1.1 A biologist's field guide to genome sequencing technologies:

The chain terminator sequencing method described by Frederick Sanger in 1977 [1] was a major breakthrough that allowed molecular biologists to better understand the information encoded by DNA. Further technological advances, such as PCR and the incorporation of fluorochromes, made it possible to read the nucleotide sequence of complete genes. A gene sequence database was soon available for the analysis of their molecular functions and potential interactions, but soon it became clear that to have a complete picture of all the molecular processes of an organism it was necessary to go "genome-wide".

In the 1980's, with the first automated sequencing machines, the idea of sequencing the entire human genome was proposed, and in 1988 the U.S congress funded the National Institutes of Health (NIH) and the Department of Energy (DoE) and in 1990 the initial research plan for the Human Genome Project (HGP) was released [2]. The early stages of the HGP were dedicated to the development of new methods and technologies to analyse the genomic sections in a cheap and efficient way [3].

In 1995, *Haemophilus influenzae* became the first organism to have its genome completely sequenced [4]. This was later followed by the genome of other majors organisms such as *Saccharomyces cerevisiae* (1996)[5], *Caenorhabditis elegans* (1998)[6], *Drosophila melanogaster* (2000)[7], and *Homo sapiens* (2001)[8,9]. The first draft genome for *Trypanosoma cruzi* was produced in 2005[10]. These genomes were sequenced using somewhat automated versions of the first generation Sanger sequencing machines that produced a low amount of sequencing data in short time but were very expensive to maintain and involved a lot of time consuming manual input[11]. New technological advances in sequencing methods have reduced the operational costs while allowing a much higher level of production of data, these methods are usually referred as Next Generation Sequencing (NGS) technologies. The large amount of data produced by the second generation sequencing platforms have the shortcoming of having high error rates and producing significantly shorter sequence reads (35 - 150 bp) compared with their first generation counterparts (600 - 1000 bp) [12].

### 1.1.1 Second generation sequencing technologies: 454 and Illumina.

The first NGS platform was the 454 pyrosequencing®. Pyrosequencing is considered a variant of the sequencing by synthesis (SBS) method. Genomic DNA is randomly shredded into small fragments of the desired size, usually ranging between 150 - 550 bp, and attached to a small bead where the sequencing reaction takes place. The controlled addition of dNTPs produce a luminous signal that is detected by a charge-couple device (CCD) camera. In this way, it is possible to sequentially detect which dNTP has been incorporated in the reaction and thus determine the nucleotide sequence [13]. The reads produced by the 454 pyrosequencing® platform ranged only between 250 - 450 bp but produced about 1 million reads per run which constituted a considerable advantage over the low yield Sanger sequencing machines [14].

The most widely used sequencing platform to date is the Illumina HiSeq®. Based on the cyclic reversible termination variant of the SBS method, the system works in a similar way as the Sanger sequencing technique. Following the random shredding of the genomic DNA, the fragments are organised into clusters and ligated to sequence adapters that serve as primers for a DNA polymerase reaction. This reaction is terminated once a labelled dideoxynucleotide triphosphate (dNTP) is incorporated into the sequence template. The dNTPs used in the reaction are labelled with a specific fluorochrome and these are recognised by a set of four lasers, each one with a specific wavelength per fluorochrome. Once the lasers have imaged the reaction, the labelled fluorochromes are washed away and a new sequencing cycle begins [15]. In this way is possible to determine the nucleotide sequence of each fragment. The Illumina platforms produce short reads in the range of 75 - 300 bp with different sequencing data outputs ranging from 15 million (Illumina MiSeq®) to 6 billion (Illumina HiSeq X Ten®).

454 and Illumina, were the first platforms to produce high throughput sequencing data, and they made it necessary to further improve the algorithms required to process and analyse the new kind of data. Illumina is the largest supplier of sequencing machines to date, and part of the success has been the ability to produce equipment tailored for specific needs such as bench top sequencing (Illumina MiSeq®) to population-scale whole genome sequencing (Illumina HiSeq X Ten®), as well as the relatively low error rate (> 0.5 %)[11]. This has not been the case for the 454 pyrosequencing® platform, where the costs of the sequencing reagents and the relatively low throughput compared to Illumina forced the platform to be officially retired in early 2016.

### 1.1.2 Third generation sequencing: Long sequencing reads:

Second generation sequencing technologies provide researchers with large data sets that can be exploited for the analysis of models organisms with good reference genome sequences. Two examples of the application of second generation, short read sequencing in model organisms are the analysis of human variation in the 1000 Human Genomes Project[16] and the ENCODE Project [17]. The scenario is different for non-model organisms, as not all of them have been sequenced and the few that have are still not complete [18]. The short length nature of the reads from second generation sequencing platforms makes it impossible to completely assemble highly repetitive regions (i.e: centromeres, telomeres, short tandem repeats, retrotransposons, etc) and high complexity regions (i.e: HLA loci).

Recently, a third generation of sequencing machines have been able to produce long sequence reads on the order of several kilobases. One of these platforms, from Pacific Biosciences (PacBio), is Single-Molecule Real-Time (SMRT) sequencing. PacBio SMRT sequencing fixes a DNA polymerases at the bottom of a micro well known as zero-mode waveguides [19]. The polymerase allows a single molecule of the template DNA to be processed at a time. Labelled dNTPs are incorporated as the DNA molecule passes the DNA polymerase, while a laser excites a fluorochrome and a camera detects the emitted signal. Subsequently, the marked dNTP is detached by the polymerase to allow the incorporation of a new nucleotide until the entire molecule has been processed [20]. Alternatively, the process can take place in a circular consensus sequencing (CCS) fashion, where the single DNA molecule is circularised and read several times in shorter segments and the consensus of these shorter segments is used to produce a longer read[21]. The reads produced by the PacBio machines are on average 12.5 Kb in length, but can be as long as 50 Kb [15].

The length of these reads makes them suitable for reconstruction of complex regions in *de novo* genome assembly projects [22] or single molecule mRNA sequencing [23]. Unfortunately, the error rate of these long reads can be as high as a 15% of the total length [24]. These errors can be corrected using medium to high coverage (25 - 80 X) of PacBio data[25] or high coverage (> 50 X) of Illumina data[26], but the process is highly demanding in terms of computational resources and time-consuming; a factor that should be taken into account in the experimental design of a given genome project. The production of PacBio data is also more expensive than Illumina data with an estimate of €885 per Gb for PacBio compared to €133 per Gb for Illumina [15], as of 2016.

## 1.2 Introduction to genome assembly:

Suppose for a moment that you receive a very important letter in the post and before you have the chance to read it, this document is thrown into a paper shredder. Twice. The task to assemble a genome sequence from scratch is similar to trying to put together the double-shredded letter with the hope that you can read it and that this message makes sense. The process to assemble sequencing data into a complete reference genome is called *de novo* genome assembly. To date, due to technological limitations, it is not possible for any sequencing method to read an entire genome at once. The fact that sequencing reads are usually shorter than the length of an entire genome sequence implicates the need for algorithms to order all the reads together into a genome as if they are a jigsaw puzzle. The problem of genome assembly has led to multiple innovations regarding the data produced and the algorithms to process the data.

Below, I give a brief description of the most popular sequencing data and assembly methods used to date.

### 1.2.1 *Bestiarum vocabulum* of sequencing data:

Different types of sequencing data that can be generated from second and third generation sequencing platforms. The generation of these data depend on the method used to build the sequencing library and the purpose of the sequencing experiment.

Libraries using long genomic fragments - usually between 50 and 100 Kb - can be produced from a highly purified DNA preparation, as required for the PacBio sequencing protocol [15] or for the production of Bacterial Artificial Chromosomes (BAC), and Fosmids. The advantage provided by sequencing the latter relies on the distance information from the long inserts, which can be used for fragment ordering during a *de novo* assembly [27].

Mate paired libraries, also known as jumping libraries, are produced by fragmenting DNA into pieces between 1 and 20 Kb in length. The long pieces of DNA are submitted to a circularisation process where both extremes of the DNA molecule are joined together by biotinylation and later sheared into smaller fragments whose ends are sequenced [28]. The sequencing data are used to resolve long repetitive regions in *de novo* genome assemblies [29].

The preparation of genome sequencing libraries with insert sizes between 180 - 1000 bp often starts with the enrichment of genomic material via cloning or amplification by PCR.

However, this enrichment step is optional and it will depend on the quantity of the starting genomic material, as some protocols require as little as 10 ng/uL [30,31]. Later, the genomic material is randomly sheared by means of physical (e.g: sonication), enzymatic (e.g: DNAse I or transposases) or chemical (e.g: metal cations) methods to the desired size [32]. The resulting fragments are used as template for sequencing with second generation sequencing platforms. Two types of sequencing data can be produced with this kind of genomic library: singled and paired ends.

Singled end data results from the sequencing of a fragment of one of either of the ends of a genomic fragment. Singled end data is currently mostly used for transcript quantification in RNA-Seq experiments [33].

Paired end data is the most widely used type of sequencing data and is produced by sequencing both ends of a given fragment. Besides the sequence information contained in the reads, the data also provides distance information between both ends of the sequenced fragment. This information is useful for *de novo* genome assemblies [28] and the detection of genomic structural changes [34].

### 1.2.2 Genome assembly algorithms:

The general idea of a *de novo* assembly experiment is to transform the millions of short sequencing reads into a single consensus that contains all the information encoded by the genome at a high level of accuracy. To achieve this, computer scientists have implemented different algorithms to be able to analyse the large amount of data in an organised way to determine the best possible reconstruction of a genome sequence [35].

There are two categories of algorithms used to achieve this task: Overlap Layout Consensus (OLC) and de Bruijn Graph (DBG). The usefulness of these algorithms depends on the type of sequencing data produced and the available computational resources.

OLC-based assemblers work the best using long sequence reads such as those produced by the Sanger method and PacBio. Examples of this kind of genome assembler are the CELERA assembler used for the first versions of the *D. melanogaster* and *H. sapiens* genome assemblies, among others [36]. Given a set of relatively long reads, an OLC based algorithm will first hash the reads into smaller fragments of an arbitrary length - known as K-mers - and perform an all-against-all comparison of these to find overlaps. The overlaps are incorporated into a data structure called a graph, from where consensus sequences called contigs are

inferred. The nature of the algorithm, especially the all-against-all search step, makes these assemblers slow and impractical when working with large amounts of sequencing data [35,37]; but recent advances have made it possible to assemble genomes to a high level of completion using OLC-based assemblers and long PacBio reads [25]. Hybrid approaches have been developed to take advantage of high sequence coverage in combination with long sequences with limited success [38].

DBG-based assemblers work best with short reads from second generation sequence platforms such as Illumina and 454. The first assembler to implement a DBG approach was EULER [39]. The DBG algorithm relies on the lower error rate in short reads to build a graph based on K-mers in a similar way to OLC-based assemblers. Overlaps are stored in the graph and the consensus assembly results from finding the most accurate Eulerian path in the graph [37]. However, the algorithm tends to collapse repetitive sequences and to breaking the consensus when highly heterozygous sequences are encountered, which yields a fragmented assembly [29]. Assemblers based on DBG are faster, but since the graph of K-mer overlaps is stored in the RAM of the computer, they can be computationally demanding. Despite these drawbacks, several implementations of DBG-based assemblers have been successfully applied to reconstruct the genomes of several organisms [40,41].

The final contigs, from both approaches, will contain the best representation of the genome with the given sequencing data and they can be ordered into scaffolds using long insert size mate paired libraries [42,43], fosmid pools [27,44] and long PacBio reads [45]. Additionally, the scaffolded genome assembly could be further improved by the incorporation of additional data such as optical and linkage maps [46].

### 1.2.3 A recipe for the perfect *de novo* assembly:

There are many factors involved in the successful reconstruction of a genome sequence, such as the presence of repetitive sequences, the level of heterozygosity, the ploidy of the organism, as well as the estimated genome size. Some studies have tried to benchmark the behaviour of different genome assemblers with different types of data from a set of species [47,48], and although these evaluations are an invaluable resource, they do not provide an universal protocol for the perfect assembly. In my opinion, each genome has its own optimal assembly protocol that uses the techniques and algorithms available at a given time.

Nevertheless, thanks to the benchmarking studies, there are some parameters that can be fine-tuned at the beginning of an assembly experiment. For example, in DBG-based assemblers, the first step should be the selection of the K-mer size for graph construction and for OLC-based assemblers, the minimum overlap length between two reads.

Most importantly, the generation of the adequate sequencing data is key for the success of the experiment. For an Illumina-only approach it is suggested to use at least three paired end libraries with insert sizes between 180 - 750 bp as well as at least three mate pair libraries with insert sizes between 2 - 15 Kb [29,49]. With the recent introduction of third generation long reads, it is possible to sequence and *de novo* assemble small to medium genomes with a high level of success using these and other ordering data such as linkage and optical maps.

Ideally, a combination of different types of data is required, but this does not guarantee an easy reconstruction of the genome of an organism [50].

### 1.3 Methods for comparative genomic analyses:

The genome sequence has been compared to a "blueprint" for life, implying that it is a rigid structure encoding genes for all the biological functions of an organism and without much change beyond point mutation in coding sequences [51]. Recent studies have shown that genomes can be highly dynamic and that their evolution involves a wide range of genomic variants and adaptive evolutionary processes [52–54]. The generation of *de novo* mutations that allow for sequence diversity and adaptability is closely linked to the architecture of the genome: retrotransposons, simple repeats, gene clusters, etc [55–59].

The first comparative studies used protein and gene sequences to identify sequence changes that may produce be functional. At that time, when the number of sequences was small and their length was usually short, the comparisons were often performed by hand, but once the number of available gene sequences increased from tens to thousands, the first sequence alignment algorithms were implemented to perform comparative gene analyses [60,61]. With the production of complete genome sequences and introduction of second and third generation sequencing platforms the field of comparative genomics was born, together with the difficulties of handling large amounts of sequencing data and mapping sequences with precision to their specific locations in genomes for comparative purposes.

In this section I will describe the basic methods used for comparative genomics using second generation sequencing data.

### 1.3.1 Alignment of sequencing reads:

The problem of correct placement of short reads in a genome is known as read mapping and there has been several implementations of methods for highly accurate mapping.

To be able to search the genome for matches with the reads in a rapid, yet accurate way, the genome sequence has to be indexed. A widely used indexing strategy is based on Burrows-Wheeler Transform (BWT), implemented in the BWA [62] and Bowtie2 [63] aligners. In a BWT-based aligner, the reads are mapped in a seed-and-extend fashion, where segments of the short reads - referred as seeds - are aligned against the indexed genome and the placement is evaluated based on the maximum number of tolerated mismatches (i.e: diversity between individuals of the same species). This process continues until the correct position for the entire read has been found, or not found, in the genome. In the case of paired end reads, the process is extended to both reads in a pair, and if one of the reads cannot be mapped, the other one could still be included in the final alignment [64]. Genomic regions rich in repetitive sequences constitute a considerable problem for read mapping and some tools have been developed to map reads in these regions correctly [65]. One of these tools is Stampy[66], which uses a statistical model that takes into account not only a per-base match but also the surrounding area of the read to identify potential erroneous placements in repetitive regions. This method makes it possible to in many cases map reads in complex areas of a given genome and even to map sequencing data from other species with a sequence divergence up to 15% [65].

To store the large amount of alignment information generated by these mappers, a new file format was created: The Sequence/Alignment Map or SAM [67] and its binary version BAM. Currently, BAM files are the standard format to store read alignment data since they are smaller than the human readable SAM files, but due to the rapid increase in the amount of sequencing data and larger, population-scale genomic studies, new versions of the format, such as the CRAM format used in the 1000 Human Genomes Project, have been created to reduce the storage footprint of large genomic datasets [16,68].

### 1.3.2 Genomic variant calling in a nutshell:

Genomic variant calling is usually achieved by mapping reads or assembled contigs against the complete reference genome sequence of a given organism [69].

The most widely used approach for variant calling is the mapping of reads from second or third generation sequencing platforms. Once the reads have been mapped to a reference sequence, the resulting mapping file has to be processed prior to variant calling. The first step is to sort all the alignments based on their specific coordinates in the genome. Subsequently, reads that are redundant, i.e: PCR and optical duplicates, are removed. Finally, labels to allow the identification of the sample should be added to the BAM file, e.g: sample name, sequencing platform, sample population, etc. Several tools have been produced to streamline the preprocessing of mapping files for variant calling. These include Picard Tools [70], SAMtools [67] and Sambamba [71]. After these preprocessing steps, the mapping files are ready to be used for the identification of different types of genomic changes. Additional steps can be incorporated into the preprocessing, such as base quality recalibration, but these depend on the variant calling algorithm to be used and the reader is referred to the GATK Best Practices manual [72] for more details. After these preprocessing steps, the mapping files are ready to be used for the identification of different types of genomic changes.

Several tools have been developed to use mapped reads for genomic variation analysis and, based on their internal algorithm, they perform better in certain genomic regions than in others [73–75]. In my experience, a set of high quality, reliable variant calls can be obtained by integrating the output from multiple methods. For instance, the approach taken in the 1000 Human Genomes Project used several variant callers and a consensus was created from all the results [76]. In principle, variants are identified by searching for supported mismatches in the reads that can be linked to biology, such as sample diversity, rather than sequencing errors [69]. To achieve this, the tool used uses mapping parameters to support a statistical model to weight the confidence of the identified variant [77]. The statistical model usually relies on how many bases from the mapped reads support a given variant, the sequence quality of that base, the mapping quality of that base, among others [78]. Once the genomic variants have been identified, it is possible to filter more complex situations and this will depend on the purpose of the experiment and the genomic characteristics of the studied organism [79].

The identified genomic variants are stored in a format called Variant Calling Format (VCF). This format is structured such that a given genomic change can be localised quickly in the

genome as well as displaying all the characteristics of the variant. Additional information, such as allelic frequency and genotype are also included [80].

These genomic variants are stored in a format called Variant Calling Format (VCF). This format is structured in a way that a given genomic change can be localised quickly in the genome as well as displaying all its possible characteristic. Additional information of the variant in a given context such as allelic frequency and genotype are also included.

Several tools that provide a high level of accuracy have been developed. The Genome Analysis ToolKit (GATK) has been developed for organisms such as human or mouse [77,81], but it can be adapted to work with other species [69]. A different approach has been used in FreeBayes, where variants are identified taking both haplotypes into account directly from the reads [82], and a recent re-implementation has been designed for the rapid analysis of genomic variation in human genomes [83]. An alternative approach to the study of whole genome variation is based on sequence or genome assembly, where instead of comparing reads directly, they are first assembled into contigs [84]. The advantage of the assembly methods is that it can detect variation in regions that are not well resolved in the reference sequence, such as the HLA locus in the human genome, and they can be much faster than methods that rely on read mapping [85,86]. In the future, this approach will be useful for the study of genomic variation in organisms with a high degree of genomic complexity.

## 1.4 The biology of *Trypanosoma cruzi*:

*Trypanosoma cruzi* is a unicellular protozoan from the family *Kinetoplastida*, which contains other unicellular protists such as *Leishmania* sp., *Trypanosoma brucei*, *Crithidia* sp among others [87], and is the etiological agent of American trypanosomiasis, clinically known as Chagas disease [88]. The disease and the parasite were first described by Carlos Chagas in 1907 while he was working as a medical doctor for a malaria eradication program in the Brazilian state of Minas Gerais [89]. It is estimated that there are 6 - 7 million people infected with T. cruzi just in Latin America[90] but with many more at risk, particularly in the poorest regions [91].

In this section I will provide a very brief introduction to the biology of *T. cruzi* and its clinical features.

### 1.4.1 *T. cruzi* life cycle:

*Trypanosoma cruzi* has a complex, dual life cycle with four main stages [92,93]. As an epimastigote, the parasite has the ability to survive in the insect intestine without causing any detectable pathology. In this stage, T. cruzi undergoes binary fission. When the parasite exits the gut in the vector feces, it transforms into metacyclic trypomastigotes. The vector takes a bloodmeal from the mammalian host by perforating the skin with its proboscide and locally anesthetizes the perforated area with its saliva. Once the vector finishes the meal, it deposits fecal material near the wound area. When the parasite exits the gut in the vector feces, it transforms into metacyclic trypomastigotes. After the anesthetizing effect expires the host auto-inoculates the metacyclic trypomastigotes through skin abrasions at the bite site or other mucosas. When the metacyclic trypomastigotes enter the bloodstream, they transform into the slender trypomastigote form that evades the host immune system and invades host cells, with a preference for myocytes [94]. Once the parasite is inside the target cell, it transforms into a rounded form called amastigote, which proliferates quickly inside the host cell forming clusters which disrupt the cell membrane, followed by transformation of the parasites into trypomastigotes that can infect other cells or be ingested by another vector to continue the life cycle [95,96].

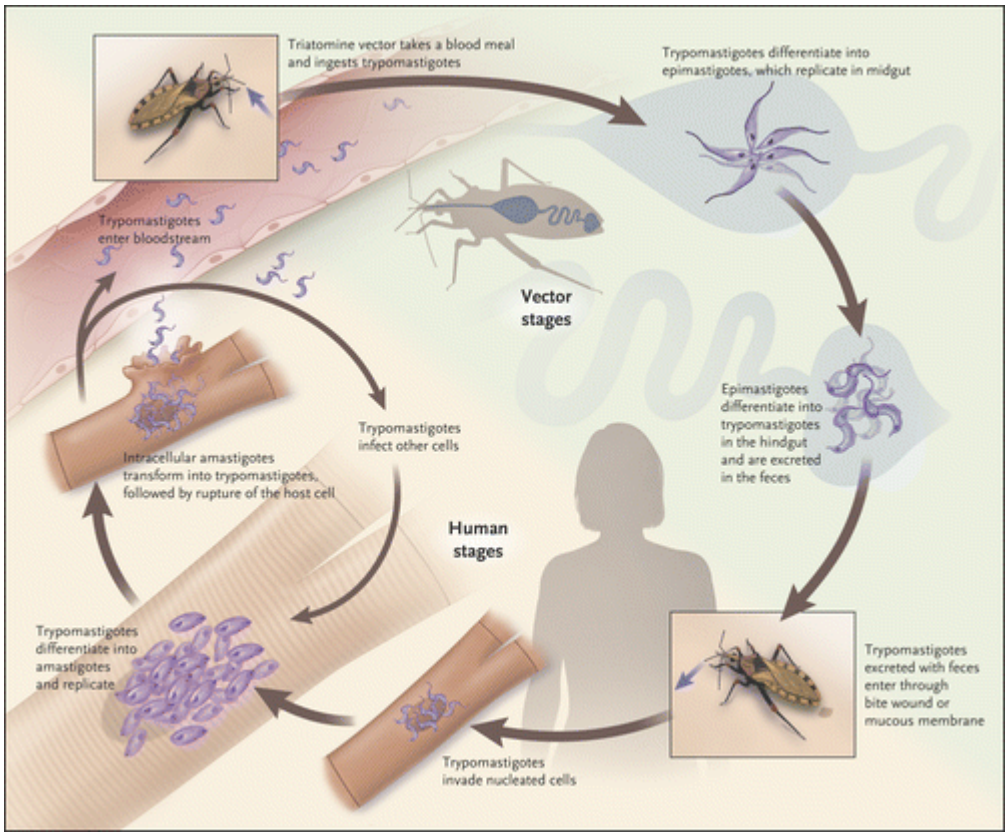


**Figure 1**: *Trypanosoma cruzi* life cycle. Image credit: **Bern C. (2015)** Chagas' disease. *N Engl J Med* 373:456 - 466.

### 1.4.2 Mechanisms of cellular invasion and immune evasion by *T. cruzi*:

*Trypanosoma cruzi* establishes chronic infection by successful evasion of the innate and adaptive immune responses, a strategy shared with other kinetoplastids. However, instead of entering the phagocytic immune cells, as is the case for *Leishmania sp.*, *T. cruzi* invades other nucleated cells [97,98].

The complement pathway and its serum components is the most important innate immune response challenge for *T. cruzi* [99,100]. Antigenic molecules present on the surface of the parasite trigger the proteolytic cleavage of C3, resulting in C3b molecules that attach to the parasite surface antigens and promote the accumulation of the complement elements involved in pathogen cellular lysis [99,101]. To avoid being lysed by the complement pathway, *T. cruzi* releases trans-sialidases and mucins that specifically attach to C3a and C3b and block them.

After *T. cruzi* penetrates the skin or mucosa, it invades non-professional phagocytic cells by attaching itself using surface molecules, such as trans-sialidases and mucins [102]. Once

attached, *T. cruzi* can be internalised into the cell by two mechanisms: Ca++-dependent lysosomal recruitment at the parasite entry site [103], a pathway shown to be used by isolates from the TcI clade; or parasite infolding into the plasma membrane with followed by lysosomal fusion, a pathway favored by isolates from the TcII - TcVI clades [94,104,105]. To date, it is not known if the parasite exploits a specific host cell receptor to promote the internalisation [104].

Lysosomes attached to the vacuole containing the parasite and the low pH environment - produced by oxidative radicals - facilitates the transition of the trypomastigote to the amastigote [106,107]. Inside the acidic phagolysosome, the parasite releases trans-sialidases that remove the sialic acid from lysosome-associated membrane proteins (LAMP1 and LAMP2), which appear to act as structural scaffolds for the phagolysosome [108,109]. Additionally, the parasite also releases pore-forming proteins that allow the parasite to escape to the cytosol, where the trypomastigote transforms into an amastigote and proliferates inside the cell [110]. After a variable number of replications, the amastigotes convert to trypomastigotes that disrupt the cellular membrane and are released to the bloodstream, ready to invade new cells.

The intracellular stage of the parasite protects itself from the host innate immune response by hijacking cellular signalling pathways and reshaping the cellular structure [101], but the viability of *Trypanosoma cruzi* inside the host depends on its ability to evade the innate and adaptive immune systems. It is possible to detect different sequence variants of surface molecules involved in immune evasion between metacyclic trypomastigotes and amastigotes [111].

This ability has been linked to specific gene families, most of them coding for surface molecules [112]. These molecules have been identified by studying the behaviour of the parasite in different culture conditions and other modified environments [113,114]. Two of the first gene families to be associated with virulence were trans-sialidases [115] and mucins [116].

Trans-sialidases were first described in *T. cruzi* in as a modified version of a sialidase enzyme involved in the incorporation of sialic acid on the parasite cellular surface [117]. Since *T. cruzi* is not able to synthesize sialic acid *de novo*, it is taken from the surface of host cells and transferred to acceptor molecules located on the parasite surface [118].

It has been estimated that the reference *T. cruzi* TcVI CL Brener strain contains about 1400 genes that code for trans-sialidase family members [119]. Based on a sequence clustering analysis, it was concluded from the sequence diversity of these genes that they split into four different groups with individual expression patterns and specific chromosomal locations [120]. Members of the trans-sialidases gene family are actively involved in the cell invasion

process, where they participate in the cell adhesion process, thus facilitating the entry of the parasite into a specific host cell. Additionally, it has been observed that *T. cruzi* sheds trans-sialidases to the bloodstream as a way to modify the surface of target cells to facilitate invasion [121]. Trans-sialidases play a fundamental role in parasite immune evasion by suppressing the triggering of the complement pathway. Early studies demonstrated that the removal of sialic acid attached to the parasite surface made it more vulnerable to the host complement system and phagocytosis [122]. A specific group of trans-sialidases genes, described as Complement Regulatory Proteins (CRP), binds to the complement elements C3b and C4b and prevents them from recognizing the parasite [97,123]. Trans-sialidases may also influence the regulation of T-cells by modifying the sialylation of the cell surface and preventing the activation of these cells against *T. cruzi* [118]. It has also been shown that trans-sialidases can re-sialylate the surface of CD8+ T-cells, and thereby hampering the response of these cells to the parasite invasion [124]. These characteristics, despite the variability, could make trans-sialidases suitable for immunotherapy development and possibly targets for testing new enzymatic inhibitors. More studies are required to understand the complete interaction with the host immune system and the mechanisms involved in the generation of sequence diversity in members of this gene family.

Mucins are heavy glycoproteins that can be found covering epithelial cells of the digestive and respiratory tract of mammals [125]. They are also found in the surface membrane of *T. cruzi* and other protozoans as acceptors of sialic acid. In this way a negatively-charged sugar coat is formed as a means to evade the immune response of the host and the vector [126,127]. This gene family, with approximately 900 gene copies [10,119], is one of the most diverse in the *T. cruzi* genome and the genes are variable in sequence [116] and structure between isolates [127].

The TcMUC mucins are expressed in the mammalian host while TcSMUG mucins are expressed in the vector and both are post-translationally modified in a different way [128]. Previous studies have shown that mucins have an important role in parasite immune evasion by stimulating dendritic cells via TLR2, and this activation has been found to delay the rapid response of other elements of the immune system, such as CD8+ T-cells [129]. A novel gene family, Mucin-associated surface proteins, was identified after the first draft genome sequence was published. These genes comprise 6% of the CL Brener genome and are found near Mucin tandem gene arrays and they are expressed during the trypomastigote stage [130], which makes them possible vaccine candidates [131]. The expression patterns of MASPs in the trypomastigote have been observed to be tissue-specific, and they seem to be more actively

expressed when the parasite is in the bloodstream, compared with other host locations, suggesting a differential expression mechanism linked to the surrounding environment [132,133].

Protozoan genes coding for specific molecules, such as mucins with glycosilphosphatidylinositol (GPI) anchors, are recognised by the innate system via Toll-like receptors (TLR) and promote the activation of innate effectors like macrophages and dendritic cells [97,134]. After the GPI-anchored molecule is recognised by TLR4 or TLR2, members of the mitogen-activated protein kinase (MAPK) and nuclear factor kB (NF-kB) are activated and trigger the release of IL-12, TNF and INF- [129,134]. Other receptors, such as TLR-9 can trigger the same response by recognition of nucleic acids [135]. *T. cruzi* use innate immune response pathways, specifically TLR-2 and TLR-4, to establish chronic infection. One example is the induction of protein phosphatase 2A by *T. cruzi,* which leads to the deactivation of MAPK and NF-kB molecules, thus promoting immune tolerance to secondary exposure [136].

The slow initial response of the host cellular immunity is the direct result of the evasion of innate immunity by *T. cruzi* [98,137]. The adaptive immune response is triggered by cell death after the parasite has disrupted the cellular membrane, which results in a strong, CD8+-mediated immune response which targets the antigens exposed on the parasite surface at the late amastigote and early trypomastigote stages [138]. However, this parasitic antigenic coat is very variable and by the time the adaptive immune response has been mounted to target these antigens, a new antigenic combination of trans-sialidases and mucins is expressed on the parasite surface [139].

### 1.4.1 The population structure of *T. cruzi*:

Early studies using a set of biochemical [140] and molecular [141] markers indicated that *T. cruzi* has an exclusively clonal population structure [142]. This clonal theory is based on using different sets of markers that show a clonal behaviour based on population genetic tests, such as high levels of linkage disequilibrium (LD), the presence of 'near clades' and overrepresentation of multilocus genotypes [143]. Analyses of genetic signatures in the kinetoplast minicircle (kDNA) indicated that these clonal groups had a single origin, thus reinforcing the clonal model for *T. cruzi* [144]. Later meta-analyses using additional markers such as isoezymes, Random Amplified Polymorphic DNA (RAPD), microsatellites, among others identified signatures of long-term clonal evolution and the clustering of the current strains into subgroups denominated as Discrete Typing Units (DTUs)[145]. Further

characterisation using a wide panel of markers in selected reference isolates established the subdivision of *T. cruzi* into six DTUs [146,147]. These clonal DTUs have been observed to have different patterns of virulence and tissue tropism in the mammalian host [148].

However, these observations were challenged when genetic interchange to produce hybrid strains was demonstrated [149,150]. Analyses of different nuclear and mitochondrial markers have demonstrated that hybridisation events have occurred [151] and that these hybridisation events have created three major groups [10,152]. Moreover, active genetic exchange in field isolates have also been demonstrated using nuclear and mitochondrial markers [153]. Population genetic studies of isolates derived from humans, indicated the existence of natural hybrids and strain variability linked to anthroponotic dispersal [154]. Furthermore, evidence of mitochondrial introgression has been observed in field and *in vitro* isolates [155].

As a response to this challenging observations, the clonal model was further developed into the Predominant Clonal Evolution (PCE) model [156,157], which states that despite the observation of genetic exchange in Trypanosoma cruzi, and other parasites, the exchange is rare and the main model of propagation is clonal, with random cases of selfing and parthenogenesis [158]. Thus far, the population genetics analyses of field isolates does not agree with this model [159], and large studies have confirmed that there is more divergence within a single DTU than what would be expected under a clonal model [160].

To date, the population structure of *T. cruzi* is far from completely understood.

# CHAPTER 2 – PRESENT RESEARCH

## 2.1 **Paper I** and **Paper II**: Comparative genomics of *Trypanosoma cruzi*: Insights into mechanisms of virulence from other trypanosomatids.

The first draft genome assembly for *Trypanosoma cruzi* was produced using first-generation, Sanger sequencing data [10] and was used to performed the first comparative analysis with the other kinetoplastid species *Trypanosoma brucei* and *Leishmania major* [161]. This first assembly provided an adequate reconstruction of the core regions where the housekeeping genes are located, despite a high degree of polymorphism in the CL Brener reference strain, but the highly repetitive subtelomeric regions, where most of the surface molecule gene families are encoded, were only partially assembled and highly fragmented [162,163]. In **paper I** and **paper II** we sequenced and assembled a less polymorphic TcI strain of *T. cruzi* from high sequence coverage of different types of second-generation reads, and compared this genome against two kinetoplastids that do not cause disease in the human host: *Trypanosoma cruzi marinkellei* (MARK) which is restricted to bats [164] and *Trypanosoma rangeli* (RANG) which is highly prevalent in opossums but also common in humans [165].

In **paper I** we used high coverage of Illumina and 454 short reads and a hybrid assembly approach to reconstruct the genome sequence of *T. cruzi* TcI Sylvio X10/cl1 and MARK and performed comparative analyses. This strategy made use of the assemblers developed for early second-generation sequencing technologies with reads ranging between 50 - 75 nucleotides (nt) in length, such as the Velvet assembler [166]; and the Celera assembler for the, relatively, long (~ 350 nt) 454 reads [167]. Since genome assemblers perform better in different genomic regions and with different datasets [168], merging the individual assemblies produced a more contiguous and correct *de novo* draft sequence for these parasite genomes. This resulted in improved assemblies, producing the most contiguous *T. cruzi* genome sequence yet, when compared with previous attempts using only 454 short reads [169]. This new assembly made it possible to perform intra-species comparative genomics for the first time. The initial analysis revealed that the two sub-species displayed a coding sequence divergence of nearly 7.5 % and MARK was found to have a smaller genome compared to Sylvio X10/cl1. Both protozoans shared the same core genome components but the Sylvio X10/cl1 genome has more genes coding for surface protein families involved in infection and possible host specificity (See **paper III**).

In **paper II** high coverage of 454 paired end data was used to assemble the genome of *T. rangeli.* The lower repeat content of RANG made it possible to reconstruct the complete genome using the 454 short reads, which combined with the linking information produced long scaffolds. This would not have been possible for a more complex genome, where only this type of sequencing data is not enough to produce such contiguity [167,168]. We could assemble the complete genome of RANG into only 259 scaffolds with an average length of ?, with an mean scaffold size of 202 kilobases. The low repetitive content of RANG allowed the genome assembler to reconstruct the complete genome using the 454 short reads, which combined with the linking information produce long scaffolds; unlike more complex genomes where these sequencing data is not enough to produce such contiguity [170,171]. A drastic reduction in the copy numbers of surface molecule gene families was observed when compared with the *T. cruzi* CL Brener strain. The reduction of genes such as trans-sialidases and mucins could indicate an adaptive evolution process by genome reduction [59,172], and it can be speculated that the absence of these genes could be involved in the inability of RANG to cause disease in mammals [173]. The levels of simple repeats in RANG were much lower than those observed in Sylvio X10/cl1 (See **paper III**), but the overall reduction of genome size seems to be almost entirely related to the contraction of repetitive gene families, which is in line with selection to reduce paralog content to redirect resources to other biochemical pathways that may confer an adaptive advantage [174]. On the other hand, the large numbers of members of these gene families in *T. cruzi* could indicate their importance for the the invasion of particular cells and tissues [175].

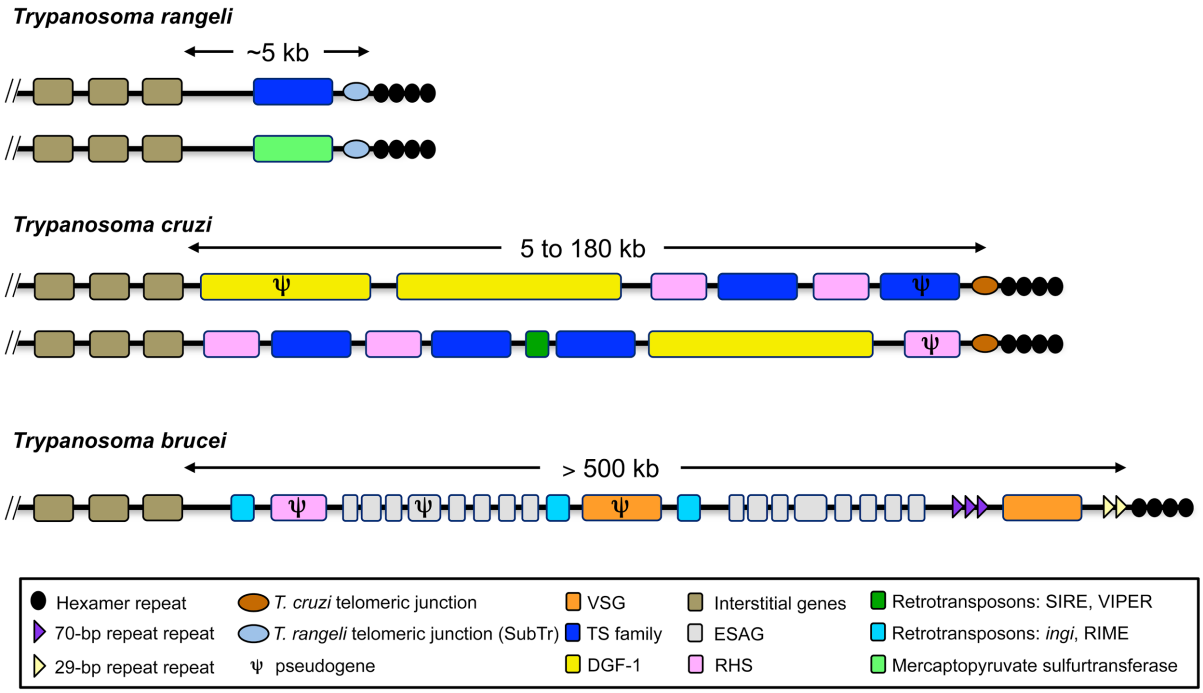


**Figure 2**: Structure of *T. rangeli* telomeres showing reduction of surface molecule gene families in comparison with *T. cruzi* and *T. brucei*. (See **paper II**).

Subsequent recent comparative analysis carried out using Illumina reads (the methodology is described in **paper III**) from these species against the Sylvio X10/cl1 reference confirmed the stable nature of the core genome in these species, but a low mappability rate in the subtelomeric regions, which indicates that these regions have a more rapid rate of evolution than the core genome. Similar characteristics have been noticed in fungi, where genomic areas containing genes coding for virulence factors have a more rapid evolutionary rate compared to regions containing housekeeping genes, to the point of rendering them sample specific [176–178]. Large genomic rearrangements were observed across the entire genome of these two protozoans when compared against the Sylvio X10/cl1 reference, including core regions, implying an active genomic restructuring process.

In the case of MARK, interspersed duplications were observed in chromosomal regions rich in mucins and retrotransposons while tandem duplications occurred in smaller regions within tandem arrays of trans-sialidases, mucins and MASPs. Multiple InDels with a slight bias towards short insertions, between one and five nucleotides were observed in the subtelomeric regions of MARK, accompanied by high levels of sequence diversity (7.24 SNV/Kb) in these areas compared with the core regions (1.1 SNV/Kb) when using Sylvio X10/cl1 as a reference; which together confirms that the surface proteins of MARK are diverged from those of *T. cruzi*, in response to the adaptation to different niches.

For RANG, deletions ranging from one to four kilobases in the subtelomeric regions were the most common genomic rearrangement. Interestingly, segmental duplications of regions containing trans-sialidase and mucin tandem arrays that were observed in MARK were absent in the same regions for RANG. Additionally, large genomic inversions in the subtelomeres and core regions were present exclusively for RANG in chr4 and chr16. Unlike the InDels observed in the MARK genome, it was not possible to detect an insertion or deletion bias in RANG, which could suggest a different mechanism of recombination in these regions compared to MARK or *T. cruzi*. Another characteristic of RANG was the lower sequence diversity observed in the subtelomeres compared with that of MARK (3.4 SNV/Kb).

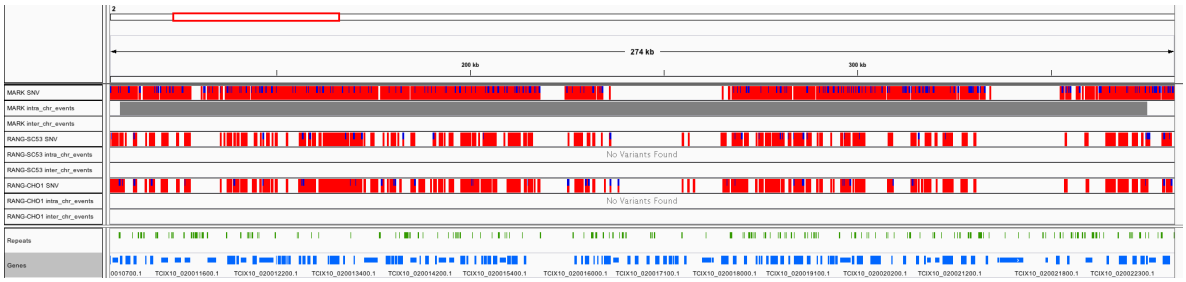


**Figure 3**: Comparative analysis of *T. rangeli* and *T. cruzi marinkellei* against chromosome 2 of the *T. cruzi* Sylvio X10/cl1 reference strain. The large grey box represents a large (> 200 Kb) duplication in chromosome 2 for *T. cruzi marinkellei*. Green boxes represent retrotransposons, clear blue represent genes, red and dark blue segments represent SNV.

In conclusion, the expanded comparative analysis of these two trypanosome species with *T. cruzi* emphasised the specialised role of the subtelomeric gene families in the adaptive evolution of the parasite to new niches. These gene families can be expanded or eroded from the genome depending on their adaptive role and the extant paralogs are used to develop new specialised function tailored to invade cells and evade the immune system of a given host.

## 2.2 Paper III: Genome analysis of the *Trypanosoma cruzi* TcI clade reveals mechanisms to generate antigenic diversity.

In this study, using high coverage of third generation sequencing data, it was possible to reconstruct the complete genome sequence of a *Trypanosoma cruzi* TcI strain at the whole chromosome level, revealing the entire genomic architecture of the elusive subtelomeric regions of this protozoa for the first time. This new reference sequence and whole-genome data from 35 *T. cruzi* TcI isolates and clones from different regions of the American continent were used to identify mechanisms involved in antigenic diversity generation in the parasite.

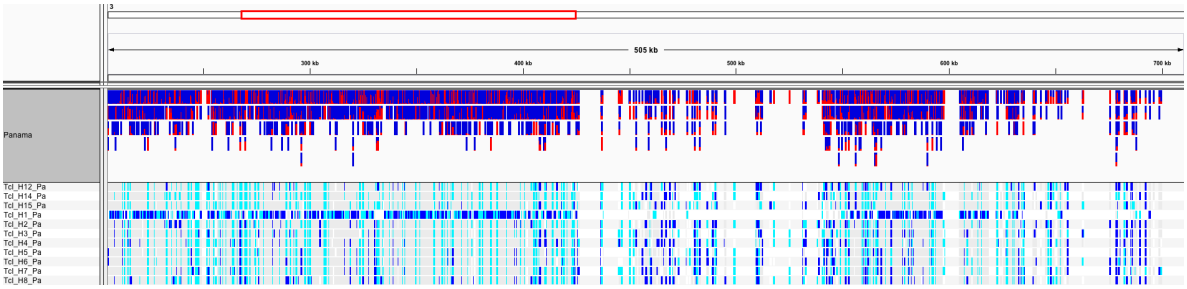


**Figure 4**: Genomic diversity of Panamanian *T. cruzi* DTU-I isolates from patients. A 505 Kb segment from chromosome 3 shows the levels of sequence diversity in subtelomeric areas coding for trans-sialidases (far left) and core regions (centre).

### 2.2.1 The *T. cruzi* genome architecture:

The genome of the *T. cruzi* Sylvio X10/cl1 reference isolate contains a large amount of retrotransposons (3.06 Mbp) and microsatellites (1.4 Mbp) interspersed throughout the genome, but with a significantly higher density in subtelomeric regions. Previous estimates of repetitive content - using second generation sequencing data - were much lower [169], most likely due to the inability of genome assemblers to reconstruct these complex areas using second generation data[29]. The subtelomeric gene families are structured in tandem arrays consisting of two to three complete gene copies and a variable number of pseudogenes separated by microsatellite segments of variable length. Retrotransposons of the VIPER and R1 class were observed within or surrounding the tandem gene arrays. The presence of retrotransposons close to surface molecule genes and their pseudogenes generated the hypothesis that these elements are involved in the generation of new sequence variants. One

possible mechanism for this could be an RNA-guided DNA insertion mechanism [179], as has been reported in other eukaryotes [180]. However, other types of data, such as RNA-Seq are required to confirm this.

### 2.2.2 Mechanisms for antigenic diversity generation:

By studying the pattern of InDels and genomic rearrangements between TcI strains in the subtelomeric regions we identified two possible mechanisms to give rise to new sequence variants in the surface molecule repertoire. These mechanisms appear to be linked to the genomic architecture of the subtelomeres.

The analysis of InDels in the 35 DTU-I isolates revealed a strong bias towards one to three nucleotide insertions in genomic segments rich in microsatellites and retrotransposons, a bias that has been previously associated with recombination hotspots [181,182] and gene conversion [183,184]. These InDels were found within microsatellites containing tandem repeats with motifs such as n(A), n(AT) and n(AG) or within the LTR region of VIPER retrotransposons. The microsatellites and retrotransposons provide a source of homology that can be exploited by a double strand break (DSB) mechanism for recombination that in turn can generate sequence diversity [185–187].

The microhomology provided by LTRs and microsatellites suggests two possible mechanisms: Non-Allelic Homologous Recombination (NAHR)[188] and Microhomology-Mediated End Joining (MMEJ)[189]. MMEJ has been observed in *T. brucei* in the presence of sequence repeats [190] and a recent study, which applied CRISPR-Cas9 to *T. cruzi,* revealed signatures of DSB repair by MMEJ [191]. On the other hand, NAHR has been associated with retrotransposons in higher eukaryotes [192] and the generation of *de novo* structural changes during meiosis [193], which has not been reported in *T. cruzi* but in other protozoans [194].

Subsequently, we analysed genomic rearrangements in the 35 DTU-I genomes to identify actual recombination events in subtelomeric regions. In **paper III** we have identified large structural variants such as deletions, tandem and interspersed duplications, genomic inversions larger than ten kilobases, as well as translocation-associated break ends occurring

in subtelomeric regions in *T. cruzi*. These types of large genomic events have been associated with NAHR [195] but this is not the only possible mechanism, since NHEJ has also been linked to the generation of sequence translocations in AT-rich areas in humans, but at a much lower frequency than NAHR [196]. The breakpoints of the interchromosomal rearrangements were detected using an improved mapping strategy and a consensus of different methods to detect genomic rearrangements (See **paper III**). The breakpoints were frequently located within the coding sequences of surface molecules genes, short microsatellite segments and LTRs of VIPER retrotransposons.

Copy Number Variation (CNV) analysis in these isolates revealed a dynamic pattern of expansion and contraction of surface molecule gene family gene clusters. These were unique to each isolate regardless of their geographic origin or sample source, e.g: isolated from vectors or humans. Based on the results of the interchromosomal breakpoint mapping, it could be assumed that these expansions and contractions of specific gene families are the result of active intra- and inter-chromosomal reshuffling of subtelomeric segments for adaptive purposes, similar to what has been observed in certain fungi [176], which may be retrotransposon-driven [178,197].

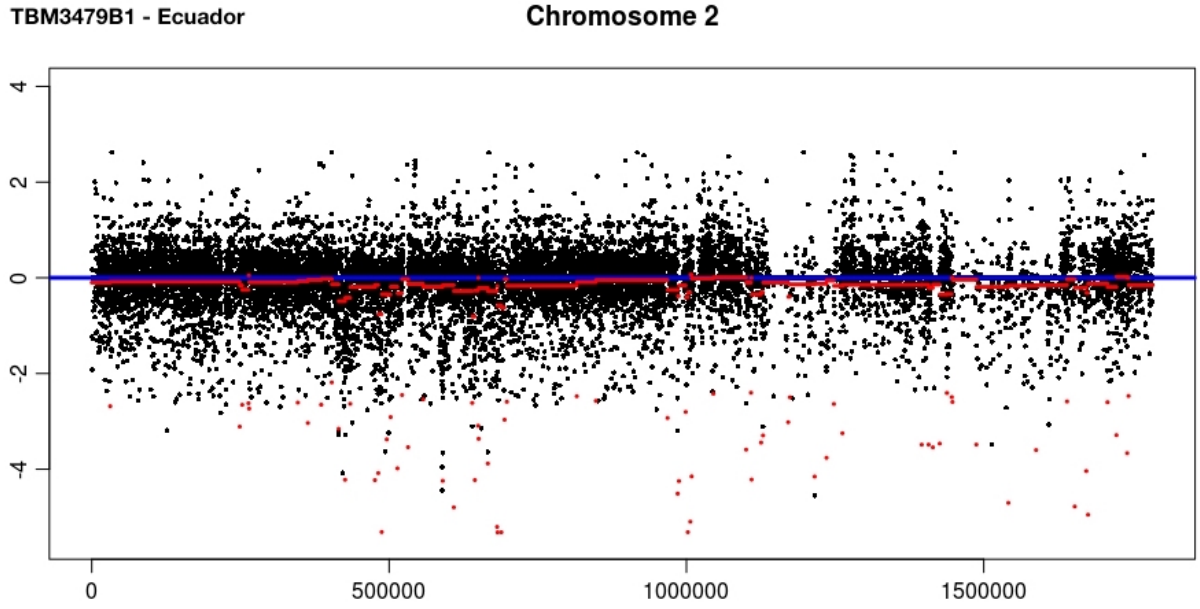


**Figure 5:** Copy Number Variation (CNV) profile of chromosome 2 of a DTU-I isolate from Ecuador. The blue line represents the reference sequence. Each black dot depicts a coverage sliding window. The red line represents

the sample genome in comparison with the reference. The X-axis represents the position in the chromosome in base-pairs; the y-axis represents fold-change.

### 2.2.3 The population genetics of the *T. cruzi* TcI clade:

WGS data from 35 *T. cruzi* DTU-I isolates from different geographic locations of the American continent was used to study the genomic diversity of this clade. A PCA analysis using InDels revealed that these isolates mostly formed well defined clusters based on their geographic location, a pattern that may be a reflection of adaption to the specific environment and selective pressure, similar to that observed in many other parasitic species [198,199]. Linkage Disequilibrium (LD) was scanned genome-wide using the r2 statistic, which revealed that subtelomeric regions exhibit r2 values close to zero while core regions had values close to one [200]. The lower r2 values and the patterns of balancing selection observed in the subtelomeric regions indicate that gene the families in these areas evolve rapidly, while the core genome changes more slowly, implying a genomic duality similar to the one observed in certain fungi [177,201].

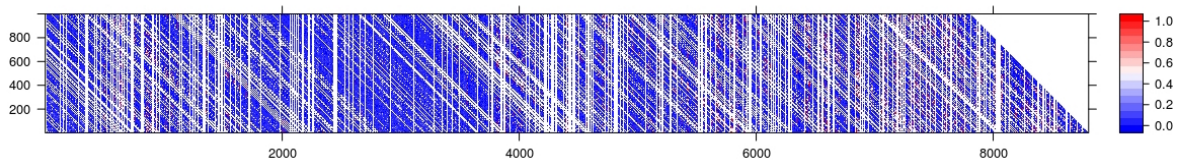


**Figure 6**: TcI linkage disequilibrium $r^2$ matrix for chromosome 16. X-axis represents the number of markers; y-axis represent the number of 10 Kb sliding windows. The $r^2$ values close to 0 indicate recombination, while $r^2$ values close to 1 indicate clonality.

Multiple diverse genotypes, reflected in the patterns of positive Tajima's D, were observed in the subtelomeric gene families in parasites from different geographic locations. Samples isolated from vectors showed high levels of balancing selection in the subtelomeric regions whereas parasites isolated from mammalian hosts showed signatures of selective sweeps in the same regions. The balancing selection observed among Colombian clones derived from the same primary isolate, indicated that the parasite is constantly generating sequence diversity in the subtelomeres. The average Fst value for samples isolated from vectors was $Fst = \sim 0.12$ whereas for human isolates it was $Fst = \sim -0.05$ suggesting little to moderate genetic differentiation, respectively [202].

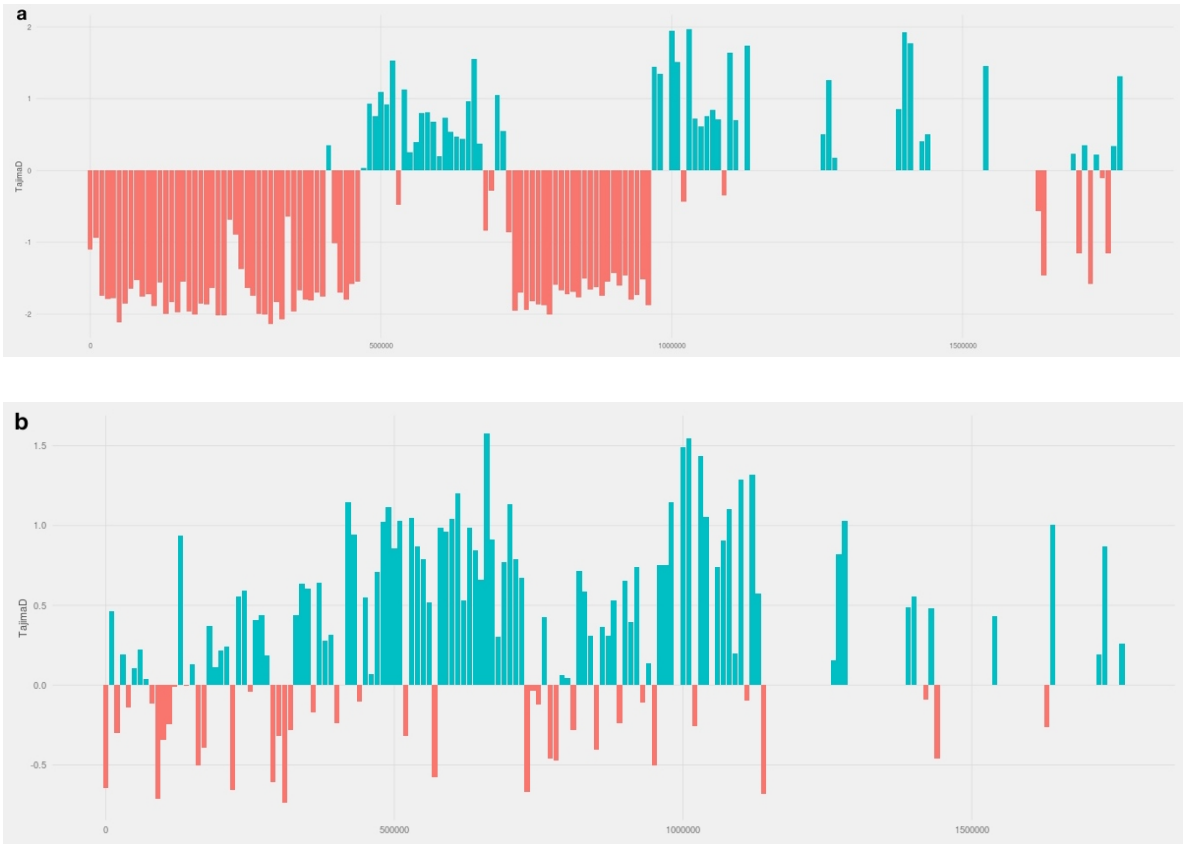


**Figure 7**: Distribution of Tajima's D in 10 Kb sliding windows for chromosome 10 in Panama samples, where **a)** represents isolates derived from human patients and **b)** represents isolates derived from vectors. Red areas show values with negative Tajima's D values (i.e: selective sweeps) and cerulean areas indicate positive Tajima's D values (i.e: balancing selection).

In conclusion, these data indicate that *T. cruzi* generates new antigenic variants via an active process of subtelomeric recombination - possibly mediated by MMEJ - while the core genome remains evolutionarily stable. Balancing selection in subtelomeric regions has resulted high levels of polymorphism for the surface protein gene families present in these areas. New alleles in these genes may confer an adaptive advantage [203], since they can be exploited by the parasite for immune evasion or rapid adaptation to the immune system of a new host, as reported in other eukaryotes [204].

## **2.3 Paper IV:** Comparative genomic analyses of *Trypanosoma cruzi* experimental hybrids reveal mechanisms of genetic exchange.

The identification of hybrid parasitic strains and genetic exchange is of great public health importance for the control of Chagas disease and the development of new drug targets [205]. Earlier studies using biochemical [206] and single-locus molecular markers [207] suggested that *Trypanosoma cruzi* has a clonal population structure. However, more recent studies using molecular markers and genome sequencing (ref) as well as the formation of hybrids *in vitro* have shown that genetic exchange occurs in *T. cruzi* [150]. These finding were later corroborated in multiple isolates [151], which led to the restructuring of the *T. cruzi* population history [146]. In **paper IV** we extended the original analysis of in vitro hybrid strains using high coverage, whole genome sequencing of the parental and hybrid strains at different time-points.

### **2.3.1 Genome analysis of the parent strains**:

Two IPE libraries and a single IMP library with an average insert size of eight kilobases from both parental strains were sequenced at a total coverage depth of 140 X per each sample. These data were used to create *de novo* assemblies of the two parental strains for comparative purposes. Despite the high sequence coverage and the long-insert library, the repetitive nature of the *T. cruzi* genome made it impossible to completely reconstruct these genomes. As mentioned earlier, it is clear that either high coverage of single molecule long reads, or multiple data sets are needed to decipher the repetitive regions of the *T. cruzi* genome [171,208,209]. We were able to completely assemble the core regions, where most of the housekeeping genes are located, and to evaluate the conservation of core gene synteny between the parent strains and our TcI reference sequence, as previously described [210,211]. The synteny was found to be well conserved, which is in agreement with previous analyses of TcI strains. Since the parental assemblies did not cover the entire genomes, we decided to proceed with the comparative analysis using the sequencing reads and taking advantage of the newly assembled and closely related Sylvio X10/1 reference genome.

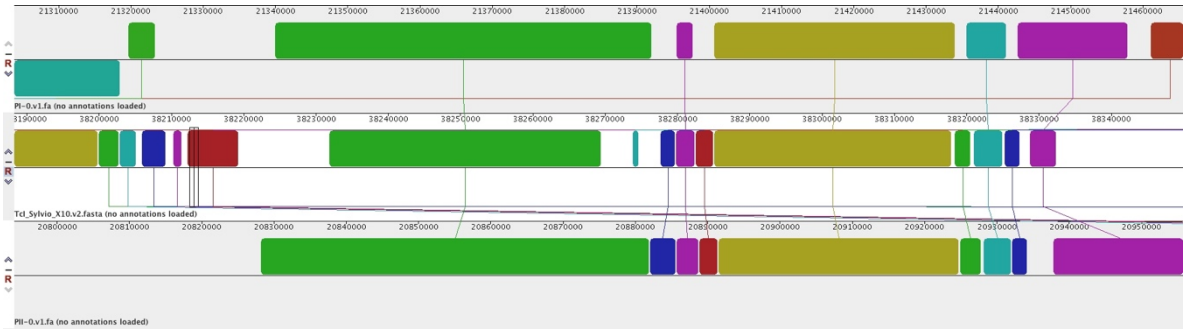


**Figure 8**: Synteny in the core regions between the parental strains PI-0 and PII-0 and the Sylvio X10/cl1 reference genome. Each coloured block represents a syntenic block in each sample. It can be noted that some blocks have been expanded in the parental strains (clear green and plum) or missing (dark blue).

A systematic analysis of genomic variation in the parental strains showed that the vast majority of the sequence and structural changes were limited to the subtelomeric regions, which is in agreement with the analyses of other *T. cruzi* TcI strains (See **paper III)**. Multiple genome rearrangements were detected and the most common were sequence break-ends similar to the unbalanced translocations observed in higher eukaryotes [212]. Copy number variation (CNV) analyses showed that the subtelomeric regions, and certain loci in the core regions were expanded compared with the Sylvio X10/1 reference strain, and these expansions accounted for the higher repetitive content of the parents and this made it even more difficult to assemble the parent strains *de novo*. This observation also agrees with the variability of the genome size in *T. cruzi* field isolates from the same clade [153,213–215] as well as differences in the karyotype of isolates from the same clade [216,217].

Interestingly, the subtelomeric gene families in the parental strains were found to have lost gene copies after 800 generations of growth in culture. The eroded regions contained genes involved in cell invasion and immune system evasion such as trans-sialidases and mucins. This may indicate that *T. cruzi* evolves in culture by selecting for the removal of genes that are not necessary in this environment, to allocate resources to other pathways.This mechanism has been widely observed in other species [172,218]. This observation has implications for future evolutionary studies of the *T. cruzi* surface molecule repertoire and requires further analysis and experimental follow-ups.

Figure 2

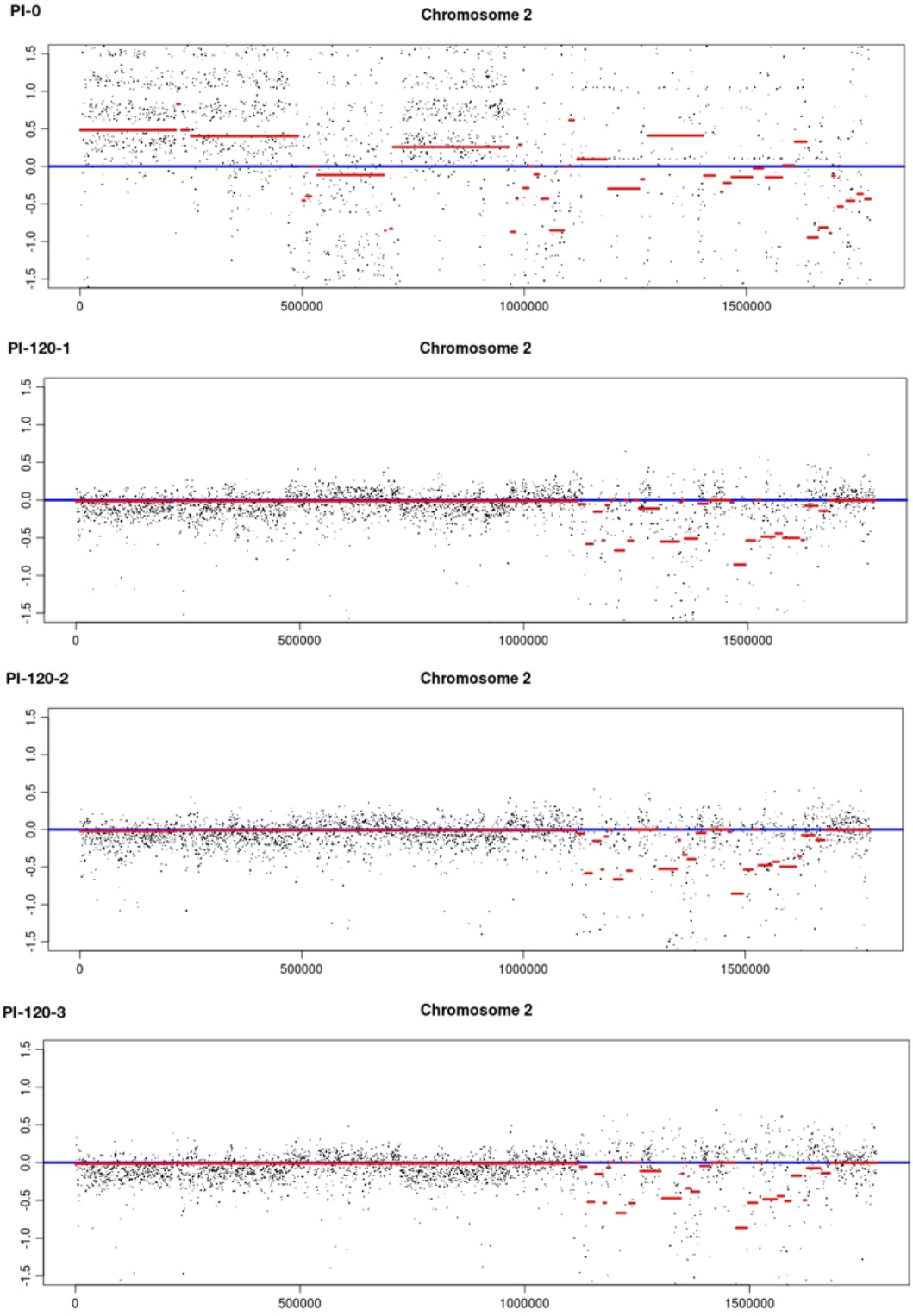


**Figure 9**: CNV evolution of a parental strain over time. PI-0 is the sample at the initial state, and PI-120 are isolates after 800 generations. Each black dot depicts a coverage sliding window. The red line represents the sample genome in comparison with the reference. X-axis represents the position in the chromosome in base-pairs; y-axis represents fold-change.

### 2.3.2 Genome analysis of the hybrid strains:

Strain-specific genomic variants, such as SNPs and InDels, were identified in the parental clones to make it possible to identify parent-specific genetic material in the hybrid offspring. We compared the reads from the hybrid genomes against the *T. cruzi* Sylvio X10/cl1 reference and detected parent-specific signatures distributed throughout the genome as well as new mutations. Parent-specific genomic blocks in the hybrid isolates were surrounded by short InDels, principally insertions ranging in size between one and five nucleotides, found within long stretches of simple repeats in intergenic regions and within the LTR segments of VIPER retrotransposons. As expected, the hybrid strains had a higher level of polymorphism in both housekeeping and surface molecule genes compared to the parental strains, due to the presence of both parental haplotypes for much of the genome . It is tempting to speculate that these genomic signatures are the result of microsatellite instability associated with defective DNA mismatch repair [219] and recombination associated with cell division [220–222]. While the meiotic machinery has been described in *T. cruzi* and other protozoan parasites [194] and, although the presence of gametes has been reported in other kinetoplastids [223], there is no experimental evidence of gametes in *T. cruzi*. It is therefore still more likely that the recombination occurs during mitosis.

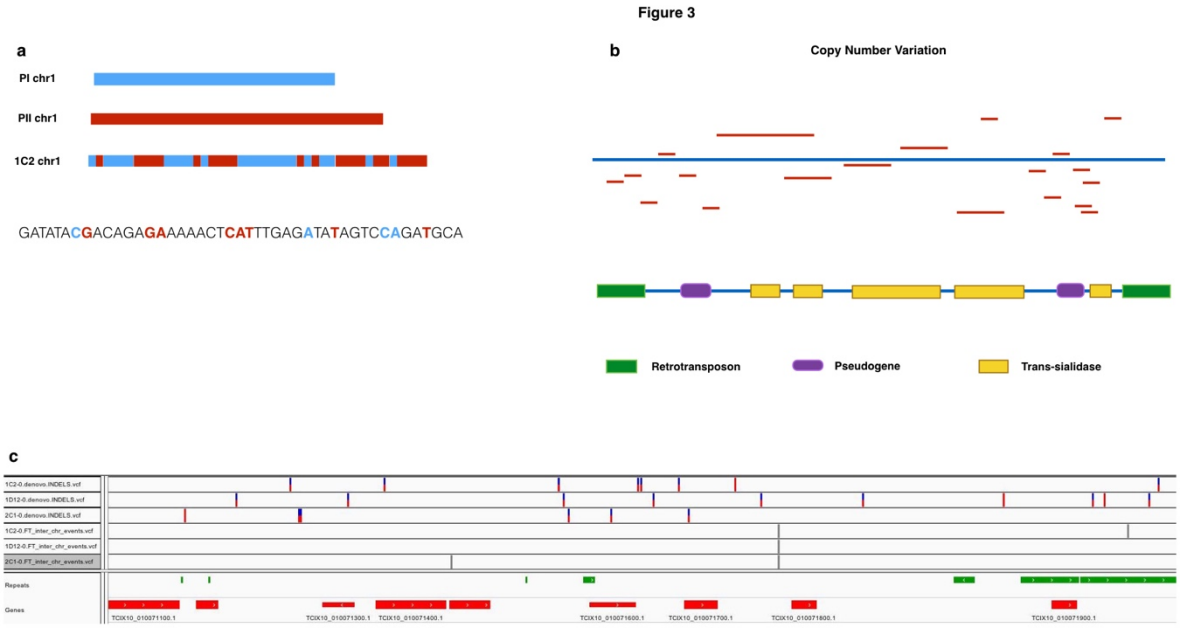


**Figure 10**: a) Patterns of parent-specific genomic material in a hybrid. b) CNV in a trans-sialidase tandem array of a hybrid strain (red lines) compared with the reference (blue line). c) InDels (red and blue bars) around breakpoints (grey bars) in the hybrid offspring. Red blocks show genes and green blocks show retrotransposons.

By analysing the sequence coverage distribution and the pattern of structural changes in the hybrid offspring, it was possible to identify, almost entirely subtelomeric, regions that were expanded or eroded after the hybridisation event. These sequence expansions increased the genome sizes of the hybrids between 34.1% and 48.6 % compared to the parents, which seems to be why hybrid strains, such as TcVI CL Brener have larger subtelomeres [10], compared to non-hybrid strains such as TcI Sylvio X10/1[169]. It is currently unclear whether the expansion occurs through a 4n hybrid intermediate or through a different mechanism. The extensive expansion of surface molecule genes and the genomic rearrangements observed in these samples are very similar to the ones observed in pathogenic fungi, where they have been associated with increased virulence [176]. It is possible to hypothesise that the combination of new surface molecule genes from both parents provides the hybrid strains with a pool of new sequences as substrates for the MMEJ-like recombination mechanism, as observed in other TcI strains, to create new antigenic variants in order to increase the potential for forming viable hybrid offspring (See **paper III**). After 800 generations of growth in culture, the hybrid genomes showed a similar pattern of genome erosion as in the parent strains.

# CHAPTER 3 – FUTURE PERSPECTIVES

Unlike other protozoan parasites, such *Plasmodium falciparum* and *Trypanosoma brucei*, the lack of a complete reference sequence has hampered the implementation of post-genomic studies in *Trypanosoma cruzi*.

The present work describes the complete genome assembly and analysis of a *T. cruzi* TcI strain, furnishing the parasitology research community with a valuable resource to better understand the biological aspects of Chagas disease. The new reference sequence will serve as the scaffold for the implementation of large scale population genomic studies, gene expression analysis in different stages of the parasite life cycle, the characterisation of transcription dynamics of the parasite virulence factors, just to mention a few.

In this work, using this genome sequence and an integrative genomic data analysis approach, it was possible to characterise a retrotransposon-driven mechanism that could be involved in the generation of antigenic variation in *T. cruzi* which could potentially be targeted with chemotherapeutic agents. However, this is just a small step in the application of post-genomic methods quest for new therapeutic interventions for Chagas disease and new analytical strategies should be implemented.

There is an urgent need for the discovery of new, more effective drug targets and vaccine candidates for Chagas disease and other parasitic malaises, but the identification of those, with the current methodologies, is an expensive and time-consuming process [224]. Genomic approaches have opened a new way to analyse complete gene families of protozoan genomes and their potential interaction with the host, providing a new alternative to identify drug targets and vaccine candidates more efficiently and cost effective.

Omics datasets - composed by genomics, transcriptomics, proteomics, among others - provide an attractive source of biological information for different pathogenic organisms stored in public databases, such as the NCBI Short Reads Archive (SRA) [225]. Some initiatives are already planning to use large public and private genomic datasets to identify links between genes and diseases in a personalised way [226]. These large datasets could be mined with the latest data analytics [227,228] for the identification of vaccine candidates, new drug targets and the synergy between these molecules and the human host, in a cheaper, faster and effective way.

# ACKNOWLEDGEMENTS

These four and a half years of doctoral studies have been a fantastic experience of professional and personal development. I would like to express my sincere gratitude to the people who, in one way or another, have contribute to this experience. I am very grateful to:

My supervisor **Björn Andersson**, for giving me the opportunity to do my PhD studies in his group, for his patience and support during the bumpy road of becoming a researcher, for the stimulating conversations; and overall, thank you for your strong belief in my scientific skills to take on any task – no matter how ambitious or wild this seemed to be – and for the freedom to pursue my own ideas.

My dear friend **Hamid Darban**, the Sha of Labbet! Thank you very much for teaching me so many sequencing skills in so little time during my first visit to the lab, for your friendship, the wise advice in matters of work and life, and for always being ready to show me the bright side of any situation.

Moving from a pure clinical environment to the computational biology research was a rather intimidating experience but, fortunately, I had the privileged opportunity to learn and work alongside the best people while working in the Spruce Genome Project. I am very grateful to:

**Anna Wetterbom** for her help and advice in the early days of the project.

**Francesco Vezzi** for teaching me all I know about genome assembly and how to conduct bioinformatics research properly; thank you for being such a good teacher and friend.

**Lars Arvestad** for sharing your knowledge, discussing ideas and for the great company.

While working at **CMB**, I also had the great luck to learn from and enjoy the company of wonderful people. I am particularly grateful to:

**Mauricio Barrientos Somarribas** for being such a good friend, teacher, flat-mate, and comrade in PhD-related torment. Thank you for always being willing to discuss scientific ideas, encouraging me to think like a computer scientist, providing critical comments on my research and your advice on how to find high-quality, educational YouTube® clips.

**Christian Pou González** for his work advice and such fun company and **Fabian Nordenskjöld** for his help with conifer genomics and for fun 'awk' tricks.

My former colleagues, **Stephen Ochaya**, **Oscar Franzén**, **Stefanie Prast-Nielsen** and **Johannes Walter Luthman** for providing a nice working environment.

I am very grateful to our collaborators at the London School of Hygiene and Tropical Medicine, **Louisa Messenger**, **Tegwen Marlais**, **Matthew Yeo** for the interesting discussions of protozoan and wormy genomics and to **Michael Miles** for his valuable input in my research and for always welcoming me into his lab whenever I was in London.

To our collaborators in Brazil, **Edmundo C. Grisard**, **Patricia Hermes Stoco**, **Glauber Wagner** in Florianóplois and **Santuza M. R. Teixeira** and **Daniella Bartholomeu** in Belo Horisonte for their hospitality while visiting Brazil and his insightful ideas in kinetoplast biology.

Also, while at CMB, I also had the opportunity to collaborate with several people in very ambitious projects. I am very grateful for their confidence in my work and for giving me the opportunity to play with such fun dataset. I want to thank particularly to:

**András Simon** for being such a great teacher on newt biology and for allowing me to work with newt genomics.

**Jonas Frisén** and **Marta Paterlini** for giving me the opportunity to work with large scale single cell genomic and transcriptomic datasets. To **Jeff Mold** for being such a fun collaborator, for sharing his ideas in immunology.

To **Kirsty Spalding** for the opportunity to work with adipocyte genomics.

Finally, I want to express my gratitude to all the friends I've met during my stay in Sweden. I am very thankful for all the help, advice and fun experiences during all these years.

# REFERENCES


1. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5463–5467 (1977).

2. Human Genome Project's Five-Year Plan (1991-1995). Available at: https://www.genome.gov/10001477/human-genome-projects-fiveyear-plan-19911995/. (Accessed: 28th July 2016)

3. Collins, F. S. Contemplating the end of the beginning. *Genome Res.* **11,** 641–643 (2001).

4. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269,** 496–512 (1995).

5. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274,** 546, 563–7 (1996).

6. C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282,** 2012–2018 (1998).

7. Adams, M. D. *et al.* The genome sequence of Drosophila melanogaster. *Science* **287,** 2185–2195 (2000).

8. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).

9. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

10. El-Sayed, N. M. *et al.* The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease. *Science* **309,** 409–415 (2005).

11. Metzker, M. L. Sequencing technologies—the next generation. *Nat. Rev. Genet.* **11,** 31–46 (2010).

12. Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K. & Mardis, E. R. The next-generation sequencing revolution and its impact on genomics. *Cell* **155,** 27–38 (2013).

13. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376–380 (2005).

14. Mardis, E. R. Next-generation sequencing platforms. *Annu. Rev. Anal. Chem.* **6,** 287–303 (2013).

15. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17,** 333–351 (2016).

16. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).

17. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

18. Ekblom, R. & Wolf, J. B. W. A field guide to whole-genome sequencing, assembly and annotation. *Evol. Appl.* **7,** 1026–1042 (2014).

19. Levene, M. J. *et al.* Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299,** 682–686 (2003).

20. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323,** 133–138 (2009).

21. Loomis, E. W. *et al.* Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. *Genome Res.* **23,** 121–128 (2013).

22. Huddleston, J. *et al.* Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res.* **24,** 688–696 (2014).

23. Au, K. F. *et al.* Characterization of the human ESC transcriptome by hybrid sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **110,** E4821–30 (2013).

24. Quail, M. A. *et al.* A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13,** 341 (2012).

25. Berlin, K. *et al.* Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* 1–11 (2015).

26. Koren, S. *et al.* Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* **30,** 693–700 (2012).

27. Alexeyenko, A. *et al.* Efficient de novo assembly of large and complex genomes by massively parallel sequencing of Fosmid pools. *BMC Genomics* **15,** 439 (2014).

28. Nagarajan, N. & Pop, M. Sequence assembly demystified. *Nat. Rev. Genet.* **14,** 157–167 (2013).

29. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* **13,** 36–46 (2012).

30. Head, S. R. *et al.* Library construction for next-generation sequencing: overviews and challenges. *Biotechniques* **56,** 61–4, 66, 68, passim (2014).

31. Rykalina, V. N. *et al.* Exome sequencing from nanogram amounts of starting DNA: comparing three approaches. *PLoS One* **9,** e101154 (2014).

32. Knierim, E., Lucke, B., Schwarz, J. M., Schuelke, M. & Seelow, D. Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. *PLoS One* **6,** e28240 (2011).

33. Sandberg, R. Entering the era of single-cell transcriptomics in biology and medicine. *Nat. Methods* **11,** 22–24 (2014).

34. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28,** i333–i339 (2012).

35. Miller, J. R., Koren, S. & Sutton, G. Assembly algorithms for next-generation sequencing data. *Genomics* **95,** 315–327 (2010).

36. Myers, E. W. *et al.* A Whole-Genome Assembly of Drosophila. *Science* **287,** 2196–2204 (2000).

37. Simpson, J. T. & Pop, M. The Theory and Practice of Genome Sequence Assembly. *Annu. Rev. Genomics Hum. Genet.* **16,** 153–172 (2015).

38. Zimin, A. *et al.* Sequencing and assembly of the 22-gb loblolly pine genome. *Genetics* **196,** 875–890 (2014).

39. Pevzner, P. A., Tang, H. & Waterman, M. S. An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 9748–9753 (2001).

40. Simpson, J. T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19,** 1117–1123 (2009).

41. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19,** 455–477 (2012).

42. Nystedt, B. *et al.* The Norway spruce genome sequence and conifer genome evolution. *Nature* **497,** 579–584 (2013).

43. Sahlin, K., Vezzi, F., Nystedt, B., Lundeberg, J. & Arvestad, L. BESST--efficient scaffolding of large fragmented assemblies. *BMC Bioinformatics* **15,** 281 (2014).

44. Neale, D. B. *et al.* Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biol.* **15,** R59 (2014).

45. Boetzer, M. & Pirovano, W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence

information. *BMC Bioinformatics* **15,** 211 (2014).

46. Howe, K. & Wood, J. M. Using optical mapping data for the improvement of vertebrate genome assemblies. *Gigascience* **4,** 10 (2015).

47. Salzberg, S. L. *et al.* GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res.* **22,** 557–567 (2012).

48. Bradnam, K. R. *et al.* Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* **2,** 10 (2013).

49. Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 1513–1518 (2011).

50. Lien, S. *et al.* The Atlantic salmon genome provides insights into rediploidization. *Nature* **533,** 200–205 (2016).

51. Ainsworth, C. DNA is life's blueprint? No, there's far more to it than that. *New Scientist*

52. Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342,** 1235587 (2013).

53. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526,** 75–81 (2015).

54. Sudmant, P. H. *et al.* Global diversity, population stratification, and selection of human copy-number variation. *Science* **349,** aab3761 (2015).

55. Yeaman, S. Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proc. Natl. Acad. Sci. U. S. A.* **110,** E1743–51 (2013).

56. Huang, W. *et al.* Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. *Genome Res.* **24,** 1193–1208 (2014).

57. Chen, L., Zhou, W., Zhang, L. & Zhang, F. Genome architecture and its roles in human copy number variation. *Genomics Inform.* **12,** 136–144 (2014).

58. Fan, S. & Meyer, A. Evolution of genomic structural variation and genomic architecture in the adaptive radiations of African cichlid fishes. *Front. Genet.* **5,** 163 (2014).

59. Wendel, J. F., Jackson, S. A., Meyers, B. C. & Wing, R. A. Evolution of plant genome architecture. *Genome Biol.* **17,** 37 (2016).

60. Simossis, V., Kleinjung, J. & Heringa, J. An overview of multiple sequence

alignment. *Curr. Protoc. Bioinformatics* **Chapter 3,** Unit 3.7 (2003).

61. Iantorno, S., Gori, K., Goldman, N., Gil, M. & Dessimoz, C. Who watches the watchmen? An appraisal of benchmarks for multiple sequence alignment. *Methods Mol. Biol.* **1079,** 59–73 (2014).

62. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

63. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).

64. Reinert, K., Langmead, B., Weese, D. & Evers, D. J. Alignment of Next-Generation Sequencing Reads. *Annu. Rev. Genomics Hum. Genet.* **16,** 133–151 (2015).

65. Montgomery, S. B. *et al.* The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res.* **23,** 749–761 (2013).

66. Lunter, G. & Goodson, M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* **21,** 936–939 (2011).

67. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

68. Hts-specs by samtools. Available at: https://samtools.github.io/hts-specs. (Accessed: 27th June 2016)

69. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43,** 11.10.1–33 (2013).

70. Picard Tools - By Broad Institute. Available at: https://broadinstitute.github.io/picard/. (Accessed: 28th June 2016)

71. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31,** 2032–2034 (2015).

72. Genome Sequencing and Analysis Group @ Broad Institute. GATK | GATK Best Practices. Available at: https://www.broadinstitute.org/gatk/guide/best-practices.php. (Accessed: 28th June 2016)

73. Pabinger, S. *et al.* A survey of tools for variant analysis of next-generation genome sequencing data. *Brief. Bioinform.* **15,** 256–278 (2014).

74. Hwang, S., Kim, E., Lee, I. & Marcotte, E. M. Systematic comparison of variant calling pipelines

using gold standard personal exome variants. *Sci. Rep.* **5,** 17875 (2015).

75. Olson, N. D. *et al.* Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Front. Genet.* **6,** 235 (2015).

76. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

77. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).

78. Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32,** 246–251 (2014).

79. Reumers, J. *et al.* Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. *Nat. Biotechnol.* **30,** 61–68 (2012).

80. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).

81. DePristo, M. a. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–498 (2011).

82. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv [q-bio.GN]* (2012).

83. Chiang, C. *et al.* SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat. Methods* **12,** 966–968 (2015).

84. Li, H. FermiKit: assembly-based variant calling for Illumina resequencing data. 1–2 (2015).

85. Li, H. Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. 1–7 (2012).

86. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. & Mcvean, G. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Publishing Group* **44,** 226–232 (2012).

87. Maslov, D. A., Podlipaev, S. A. & Lukes, J. Phylogeny of the kinetoplastida: taxonomic problems and insights into the evolution of parasitism. *Mem. Inst. Oswaldo Cruz* **96,** 397–402 (2001).

88. WHO | Chagas disease. (2014).

89. Chagas, C. Nova tripanozomiaze humana: estudos sobre a morfolojia e o ciclo evolutivo do Schizotrypanum cruzi n. gen., n. sp., ajente etiolojico de

nova entidade morbida do homem. *Mem. Inst. Oswaldo Cruz* **1,** 159–218 (1909).

90. WHO | Chagas disease (American trypanosomiasis). (2016).

91. Pecoul, B. *et al.* The BENEFIT Trial: Where Do We Go from Here? *PLoS Negl. Trop. Dis.* **10,** 2–5 (2016).

92. De Souza, W. Basic cell biology of Trypanosoma cruzi. *Curr. Pharm. Des.* **8,** 269–285 (2002).

93. Goldenberg, S. & Avila, A. R. *Aspects of Trypanosoma cruzi stage differentiation*. **75,** 285–305 (Elsevier Ltd., 2011).

94. Andrade, L. O. & Andrews, N. W. The Trypanosoma cruzi–host-cell interplay: location, invasion, retention. *Nat. Rev. Microbiol.* **3,** 819–823 (2005).

95. Tyler, K. M. & Engman, D. M. The life cycle of Trypanosoma cruzi revisited. *Int. J. Parasitol.* **31,** 472–481 (2001).

96. Carrea, A. & Diambra, L. Systems Biology Approach to Model the Life Cycle of Trypanosoma cruzi. *PLoS One* **11,** e0146947 (2016).

97. Flávia Nardy, A., Freire-de-Lima, C. G. & Morrot, A. Immune Evasion Strategies of Trypanosoma cruzi. *J Immunol Res* **2015,** 178947 (2015).

98. Geiger, A. *et al.* Escaping Deleterious Immune Response in Their Hosts: Lessons from Trypanosomatids. *Front. Immunol.* **7,** 212 (2016).

99. Sacks, D. & Sher, A. Evasion of innate immunity by parasitic protozoa. *Nat. Immunol.* **3,** 1041–1047 (2002).

100. Lopes, M. F., Zamboni, D. S., Lujan, H. D. & Rodrigues, M. M. Immunity to protozoan parasites. *J. Parasitol. Res.* **2012,** 250793 (2012).

101. Watanabe Costa, R., da Silveira, J. F. & Bahia, D. Interactions between Trypanosoma cruzi Secreted Proteins and Host Cell Signaling Pathways. *Front. Microbiol.* **7,** 388 (2016).

102. Padilla, A. M., Simpson, L. J. & Tarleton, R. L. Insufficient TLR activation contributes to the slow development of CD8+ T cell responses in Trypanosoma cruzi infection. *J. Immunol.* **183,** 1245–1252 (2009).

103. Andrade, L. O. & Andrews, N. W. Lysosomal fusion is essential for the retention of Trypanosoma cruzi inside host cells. *J. Exp. Med.* **200,** 1135–1143 (2004).

104. Fernandes, M. C. & Andrews, N. W. Host cell invasion by Trypanosoma cruzi: a unique strategy that promotes persistence. *FEMS Microbiol. Rev.* **36,**

734–747 (2012).

105. Cardoso, M. S., Reis-Cunha, J. L. & Bartholomeu, D. C. Evasion of the Immune Response by Trypanosoma cruzi during Acute Infection. *Front. Immunol.* **6,** 659 (2015).

106. Díaz, M. L., Solari, A. & González, C. I. Differential expression of Trypanosoma cruzi I associated with clinical forms of Chagas disease: overexpression of oxidative stress proteins in acute patient isolate. *J. Proteomics* **74,** 1673–1682 (2011).

107. Gupta, S., Wen, J.-J. & Garg, N. J. Oxidative Stress in Chagas Disease. *Interdiscip. Perspect. Infect. Dis.* **2009,** (2009).

108. Romano, P. S., Arboit, M. A., Vázquez, C. L. & Colombo, M. I. The autophagic pathway is a key component in the lysosomal dependent entry of Trypanosoma cruzi into the host cell. *Autophagy* **5,** 6–18 (2009).

109. Rubin-de-Celis, S. S. C., Uemura, H., Yoshida, N. & Schenkman, S. Expression of trypomastigote trans-sialidase in metacyclic forms of Trypanosoma cruzi increases parasite escape from its parasitophorous vacuole. *Cell. Microbiol.* **8,** 1888–1898 (2006).

110. Caradonna, K. L. & Burleigh, B. a. *Mechanisms of host cell invasion by Trypanosoma cruzi*. **76,** 33–61 (Elsevier Ltd., 2011).

111. Yoshida, N. & Cortez, M. Trypanosoma cruzi: parasite and host cell signaling during the invasion process. *Subcell. Biochem.* **47,** 82–91 (2008).

112. Osorio, L., Ríos, I., Gutiérrez, B. & González, J. Virulence factors of Trypanosoma cruzi: who is who? *Microbes Infect.* **14,** 1390–1402 (2012).

113. Hoft, D. F., Lynch, R. G. & Kirchhoff, L. V. Kinetic analysis of antigen-specific immune responses in resistant and susceptible mice during infection with Trypanosoma cruzi. *J. Immunol.* **151,** 7038–7047 (1993).

114. Schenkman, S. *et al.* Mucin-like glycoproteins linked to the membrane by glycosylphosphatidylinositol anchor are the major acceptors of sialic acid in a reaction catalyzed by trans-sialidase in metacyclic forms of Trypanosoma cruzi. *Mol. Biochem. Parasitol.* **59,** 293–303 (1993).

115. Pereira, M. E., Zhang, K., Gong, Y., Herrera, E. M. & Ming, M. Invasive phenotype of Trypanosoma cruzi restricted to a population expressing trans-sialidase. *Infect. Immun.* **64,** 3884–3892 (1996).

116. Di Noia, J. M., D'Orso, I., Aslund, L., Sánchez, D. O. & Frasch, A. C. The Trypanosoma cruzi mucin family is transcribed from hundreds of genes having hypervariable regions. *J. Biol. Chem.* **273,** 10843–10850 (1998).

117. Previato, J. O., Andrade, A. F., Pessolani, M. C. & Mendonça-Previato, L. Incorporation of sialic acid into Trypanosoma cruzi macromolecules. A proposal for a new metabolic route. *Mol. Biochem. Parasitol.* **16,** 85–96 (1985).

118. Nardy, A. F. F. R., Freire-de-Lima, C. G., Pérez, A. R. & Morrot, A. Role of Trypanosoma cruzi Trans-sialidase on the Escape from Host Immune Surveillance. *Front. Microbiol.* **7,** 348 (2016).

119. Arner, E. *et al.* Database of Trypanosoma cruzi repeated genes: 20 000 additional gene variants. *BMC Genomics* **8,** 1–15 (2007).

120. Freitas, L. M. *et al.* Genomic analyses, gene expression and antigenic profile of the trans-sialidase superfamily of Trypanosoma cruzi reveal an undetected level of complexity. *PLoS One* **6,** e25914 (2011).

121. Tribulatti, M. V., Mucci, J., Van Rooijen, N., Leguizamón, M. S. & Campetella, O. The trans-sialidase from Trypanosoma cruzi induces thrombocytopenia during acute Chagas' disease by reducing the platelet sialic acid contents. *Infect. Immun.* **73,** 201–207 (2005).

122. Kipnis, T. L., David, J. R., Alper, C. A., Sher, A. & da Silva, W. D. Enzymatic treatment transforms trypomastigotes of Trypanosoma cruzi into activators of alternative complement pathway and potentiates their uptake by macrophages. *Proc. Natl. Acad. Sci. U. S. A.* **78,** 602–605 (1981).

123. De Pablos, L. M. & Osuna, A. Multigene families in Trypanosoma cruzi and their role in infectivity. *Infect. Immun.* **80,** 2258–2264 (2012).

124. Freire-de-Lima, L. *et al.* Trypanosoma cruzi subverts host cell sialylation and may compromise antigen-specific CD8+ T cell responses. *J. Biol. Chem.* **285,** 13388–13396 (2010).

125. Singh, I. *Textbook of Human Histology: With Colour Atlas & Practical Guide*. (Jaypee Brothers Publishers, 2010).

126. Buscaglia, C. A., Campo, V. A., Frasch, A. C. C. & Di Noia, J. M. Trypanosoma cruzi surface mucins: host-dependent coat diversity. *Nat. Rev. Microbiol.* **4,** 229–236 (2006).

127. Eugenia Giorgi, M. & de Lederkremer,

R. M. Trans-sialidase and mucins of Trypanosoma cruzi: an important interplay for the parasite. *Carbohydr. Res.* **346,** 1389–1393 (2011).

128. Cánepa, G. E., Mesías, A. C., Yu, H., Chen, X. & Buscaglia, C. A. Structural features affecting trafficking, processing, and secretion of Trypanosoma cruzi mucins. *J. Biol. Chem.* **287,** 26365–26376 (2012).

129. Gravina, H. D., Antonelli, L., Gazzinelli, R. T. & Ropert, C. Differential use of TLR2 and TLR9 in the regulation of immune responses during the infection with Trypanosoma cruzi. *PLoS One* **8,** e63100 (2013).

130. Bartholomeu, D. C. *et al.* Genomic organization and expression profile of the mucin-associated surface protein (masp) family of the human pathogen Trypanosoma cruzi. *Nucleic Acids Res.* **37,** 3407–3417 (2009).

131. Serna, C. *et al.* A synthetic peptide from Trypanosoma cruzi mucin-like associated surface protein as candidate for a vaccine against Chagas disease. *Vaccine* **32,** 3525–3532 (2014).

132. De Pablos, L. M. *et al.* Differential expression and characterization of a member of the mucin-associated surface protein family secreted by Trypanosoma cruzi. *Infect. Immun.* **79,** 3993–4001 (2011).

133. dos Santos, S. L. *et al.* The MASP family of Trypanosoma cruzi: changes in gene expression and antigenic profile during the acute phase of experimental infection. *PLoS Negl. Trop. Dis.* **6,** e1779 (2012).

134. Kayama, H. & Takeda, K. The innate immune response to Trypanosoma cruzi infection. *Microbes Infect.* **12,** 511–517 (2010).

135. Bafica, A. *et al.* Cutting edge: TLR9 and TLR2 signaling together account for MyD88-dependent control of parasitemia in Trypanosoma cruzi infection. *J. Immunol.* **177,** 3515–3519 (2006).

136. Gazzinelli, R. T. & Denkers, E. Y. Protozoan encounters with Toll-like receptor signalling pathways: implications for host parasitism. *Nat. Rev. Immunol.* **6,** 895–906 (2006).

137. Tarleton, R. L. CD8+. *Semin. Immunopathol.* **37,** 233–238 (2015).

138. Padilla, A. M., Bustamante, J. M. & Tarleton, R. L. CD8+ T cells in Trypanosoma cruzi infection. *Curr. Opin. Immunol.* **21,** 385–390 (2009).

139. Martin, D. L. *et al.* CD8+ T-Cell responses to Trypanosoma cruzi are highly focused on strain-variant trans-sialidase epitopes. *PLoS Pathog.* **2,** e77 (2006).

140. Lanham, S. M., Grendon, J. M., Miles, M. A., Povoa, M. M. & De Souza, A. A. A. A comparison of electrophoretic methods for isoenzyme characterization of trypanosomatids. I: Standard stocks of Trypanosoma cruzi zymodemes from northeast Brazil. *Trans. R. Soc. Trop. Med. Hyg.* **75,** 742–750 (1981).

141. Morel, C. *et al.* Strains and clones of Trypanosoma cruzi can be characterized by pattern of restriction endonuclease products of kinetoplast DNA minicircles. *Proc. Natl. Acad. Sci. U. S. A.* **77,** 6810–6814 (1980).

142. Tibayrenc, M. & Ayala, F. J. Towards a population genetics of microorganisms: The clonal theory of parasitic protozoa. *Parasitol. Today* **7,** 228–232 (1991).

143. Tibayrenc, M. & Ayala, F. J. How clonal are Trypanosoma and Leishmania? *Trends Parasitol.* **29,** 264–269 (2013).

144. Telleria, J. *et al.* Trypanosoma cruzi: sequence analysis of the variable region of kinetoplast minicircles. *Exp. Parasitol.* **114,** 279–288 (2006).

145. Tibayrenc, M. Genetic subdivisions within Trypanosoma cruzi(Discrete Typing Units) and their relevance for molecular epidemiology and experimental evolution. *Kinetoplastid Biol. Dis.* **2,** 1–6 (2003).

146. Zingales, B. *et al.* The revised Trypanosoma cruzi subspecific nomenclature: rationale, epidemiological relevance and research applications. *Infect. Genet. Evol.* **12,** 240–253 (2012).

147. Lewis, M. D. *et al.* Genotyping of Trypanosoma cruzi: Systematic selection of assays allowing rapid and accurate discrimination of all known lineages. *Am. J. Trop. Med. Hyg.* **81,** 1041–1049 (2009).

148. Vago, A. R. *et al.* Genetic characterization of Trypanosoma cruzi directly from tissues of patients with chronic Chagas disease: differential distribution of genetic types into diverse organs. *Am. J. Pathol.* **156,** 1805–1809 (2000).

149. Machado, C. A. & Ayala, F. J. Nucleotide sequences provide evidence of genetic exchange among distantly related lineages of Trypanosoma cruzi. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 7396–7401 (2001).

150. Gaunt, M. W., Yeo, M., Frame, I. A. & Stothard, J. R. Mechanism of genetic exchange in American trypanosomes. **421,** (2003).

151. Westenberger, S. J., Barnabé, C., Campbell, D. A. & Sturm, N. R. Two

hybridization events define the population structure of Trypanosoma cruzi. *Genetics* **171,** 527–543 (2005).

152. De Freitas, J. M. *et al.* Ancestral genomes, sex, and the population structure of Trypanosoma cruzi. *PLoS Pathog.* **2,** 0226–0235 (2006).

153. Ramírez, J. D. *et al.* Contemporary cryptic sexuality in Trypanosoma cruzi. *Mol. Ecol.* **21,** 4216–4226 (2012).

154. Messenger, L. A. *et al.* Ecological host fitting of Trypanosoma cruzi TcI in Bolivia: mosaic population structure, hybridization and a role for humans in Andean parasite dispersal. *Mol. Ecol.* **24,** 2406–2422 (2015).

155. Messenger, L. A. & Miles, M. A. Evidence and importance of genetic exchange among field populations of Trypanosoma cruzi. *Acta Trop.* **151,** 150–155 (2015).

156. Tibayrenc, M., Michel, T. & Ayala, F. J. in *Advances in Parasitology* 253–268 (2014).

157. Tibayrenc, M. & Ayala, F. J. Cryptosporidium,Giardia, Cryptococcus, Pneumocystis genetic variability: cryptic biological species or clonal near-clades? *PLoS Pathog.* **10,** e1003908 (2014).

158. Tibayrenc, M. & Ayala, F. J. The population genetics of Trypanosoma cruzi revisited in the light of the predominant clonal evolution model. *Acta Trop.* **151,** 156–165 (2015).

159. Ramírez, J. D. & Llewellyn, M. S. Reproductive clonality in protozoan pathogens - truth or artefact? *Mol. Ecol.* 4195–4202 (2014).

160. Llewellyn, M. S. *et al.* Genome-scale multilocus microsatellite typing of Trypanosoma cruzi discrete typing unit I reveals phylogeographic structure and specific genotypes linked to human infection. *PLoS Pathog.* **5,** e1000410 (2009).

161. El-Sayed, N. M. *et al.* Comparative genomics of trypanosomatid parasitic protozoa. *Science* **309,** 404–409 (2005).

162. Clayton, J. The promise of T. cruzi genomics. *Nature* **465,** S16–7 (2010).

163. Andersson, B. The Trypanosoma cruzi genome; conserved core genes and extremely variable surface molecule families. *Res. Microbiol.* **162,** 619–625 (2011).

164. Pinto, C. M., Kalko, E. K. V., Cottontail, I., Wellinghausen, N. & Cottontail, V. M. TcBat a bat-exclusive lineage of Trypanosoma cruzi in the Panama Canal Zone, with comments on its classification and the

use of the 18S rRNA gene for lineage identification. *Infect. Genet. Evol.* **12,** 1328–1332 (2012).

165. Garcia, E. S., Castro, D. P., Figueiredo, M. B., Genta, F. A. & Azambuja, P. Trypanosoma rangeli: a new perspective for studying the modulation of immune reactions of Rhodnius prolixus. *Parasit. Vectors* **2,** 33 (2009).

166. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–829 (2008).

167. Miller, J. R. *et al.* Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* **24,** 2818–2824 (2008).

168. Vicedomini, R., Vezzi, F., Scalabrin, S., Arvestad, L. & Policriti, A. GAM-NGS: genomic assemblies merger for next generation sequencing. *BMC Bioinformatics* **14 Suppl 7,** S6 (2013).

169. Franzén, O. *et al.* Shotgun sequencing analysis of Trypanosoma cruzi i Sylvio X10/1 and comparison with T. cruzi VI CL Brener. *PLoS Negl. Trop. Dis.* **5,** 1–9 (2011).

170. Natali, L. *et al.* The repetitive component of the sunflower genome as shown by different procedures for assembling next generation sequencing reads. *BMC Genomics* **14,** 686 (2013).

171. Nowak, R. M. Assembly of repetitive regions using next-generation sequencing data. *arXiv [q-bio.GN]* (2014).

172. Albalat, R. & Cañestro, C. Evolution by gene loss. *Nat. Rev. Genet.* (2016). doi:10.1038/nrg.2016.39

173. Guhl, F. & Vallejo, G. A. Trypanosoma (Herpetosoma) rangeli Tejera, 1920: an updated review. *Mem. Inst. Oswaldo Cruz* **98,** 435–442 (2003).

174. Mendonça, A. G., Alves, R. J. & Pereira-Leal, J. B. Loss of genetic redundancy in reductive genome evolution. *PLoS Comput. Biol.* **7,** e1001082 (2011).

175. Lespinet, O., Wolf, Y. I., Koonin, E. V. & Aravind, L. The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res.* **12,** 1048–1059 (2002).

176. de Jonge, R. *et al.* Extensive chromosomal reshuffling drives evolution of virulence in an asexual pathogen. *Genome Res.* **23,** 1271–1282 (2013).

177. Dong, S., Raffaele, S. & Kamoun, S. The two-speed genomes of filamentous pathogens: Waltz with plants. *Curr. Opin. Genet. Dev.* **35,**

57–65 (2015).

178. Faino, L. *et al.* Transposons passively and actively contribute to evolution of the two-speed genome of a fungal pathogen. *Genome Res.* (2016). doi:10.1101/gr.204974.116

179. Esnault, C., Maestre, J. & Heidmann, T. Human LINE retrotransposons generate processed pseudogenes. *Nat. Genet.* **24,** 363–367 (2000).

180. Ewing, A. D. *et al.* Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. *Genome Biol.* **14,** R22 (2013).

181. Brandström, M., Bagshaw, A. T., Gemmell, N. J. & Ellegren, H. The relationship between microsatellite polymorphism and recombination hot spots in the human genome. *Mol. Biol. Evol.* **25,** 2579–2587 (2008).

182. Miles, A. *et al.* Genome variation and meiotic recombination in Plasmodium falciparum: insights from deep sequencing of genetic crosses. *bioRxiv* 024182 (2015). doi:10.1101/024182

183. Bhérer, C. & Auton, A. in *eLS* (John Wiley & Sons, Ltd, 2001).

184. Leushkin, E. V. & Bazykin, G. A. Short indels are subject to insertion-biased gene conversion. *Evolution* **67,** 2604–2613 (2013).

185. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12,** 363–376 (2011).

186. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* **14,** 125–138 (2013).

187. Erwin, J. A., Marchetto, M. C. & Gage, F. H. Mobile DNA elements in the generation of diversity and complexity in the brain. *Nat. Rev. Neurosci.* **15,** 497–506 (2014).

188. Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nat. Rev. Genet.* **10,** 551–564 (2009).

189. Sfeir, A. & Symington, L. S. Microhomology-Mediated End Joining: A Back-up Survival Mechanism or Dedicated Pathway? *Trends Biochem. Sci.* **40,** 701–714 (2015).

190. Glover, L., Alsford, S. & Horn, D. DNA break site at fragile subtelomeres determines probability and mechanism of antigenic variation in African trypanosomes. *PLoS Pathog.* **9,** e1003260 (2013).

191. Peng, D., Kurup, S. P., Yao, P. Y.,

Minning, T. A. & Tarleton, R. L. CRISPR-Cas9-mediated single-gene and gene family disruption in Trypanosoma cruzi. *MBio* **6,** e02097–14 (2015).

192. Robberecht, C., Voet, T., Zamani Esteki, M., Nowakowska, B. A. & Vermeesch, J. R. Nonallelic homologous recombination between retrotransposable elements is a driver of de novo unbalanced translocations. *Genome Res.* **23,** 411–418 (2013).

193. Veltman, J. A. & Brunner, H. G. De novo mutations in human genetic disease. *Nat. Rev. Genet.* **13,** 565–575 (2012).

194. Peacock, L., Bailey, M., Carrington, M. & Gibson, W. Meiosis and haploid gametes in the pathogen Trypanosoma brucei. *Curr. Biol.* **24,** 181–186 (2014).

195. Conrad, D. F. & Hurles, M. E. The population genetics of structural variation. *Nat. Genet.* **39,** S30–6 (2007).

196. Ghezraoui, H. *et al.* Chromosomal translocations in human cells are generated by canonical nonhomologous end-joining. *Mol. Cell* **55,** 829–842 (2014).

197. Janoušek, V., Laukaitis, C. M., Yanchukov, A. & Karn, R. The roles of LINEs, LTRs and SINEs in lineage-specific gene family expansions in the human and mouse genomes. *bioRxiv* 042309 (2016). doi:10.1101/042309

198. Crellen, T. *et al.* Whole genome resequencing of the human parasite Schistosoma mansoni reveals population history and effects of selection. *Sci. Rep.* **6,** 20954 (2016).

199. Assefa, S. *et al.* Population genomic structure and adaptation in the zoonotic malaria parasite Plasmodium knowlesi. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 13027–13032 (2015).

200. VanLiere, J. M. & Rosenberg, N. A. Mathematical properties of the r2 measure of linkage disequilibrium. *Theor. Popul. Biol.* **74,** 130–137 (2008).

201. Raffaele, S. & Kamoun, S. Genome evolution in filamentous plant pathogens: why bigger can be better. *Nat. Rev. Microbiol.* **10,** 417–430 (2012).

202. Hartl, D. L. *et al.* The paradoxical population genetics of Plasmodium falciparum. *Trends Parasitol.* **18,** 266–272 (2002).

203. Weedall, G. D. & Conway, D. J. Detecting signatures of balancing selection to identify targets of anti-parasite immunity. *Trends Parasitol.*

**26,** 363–369 (2010).

204. Charlesworth, D. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet.* **2,** e64 (2006).

205. King, K. C., Stelkens, R. B., Webster, J. P., Smith, D. F. & Brockhurst, M. A. Hybridization in Parasites: Consequences for Adaptive Evolution, Pathogenesis, and Public Health in a Changing World. *PLoS Pathog.* **11,** e1005098 (2015).

206. Tibayrenc, M., Ward, P., Moya, A. & Ayala, F. J. Natural populations of Trypanosoma cruzi, the agent of Chagas disease, have a complex multiclonal structure. *Proc. Natl. Acad. Sci. U. S. A.* **83,** 115–119 (1986).

207. Tibayrenc, M., Kjellberg, F. & Ayala, F. J. A clonal theory of parasitic protozoa: the population structures of Entamoeba, Giardia, Leishmania, Naegleria, Plasmodium, Trichomonas, and Trypanosoma and their medical and taxonomical consequences. *Proc. Natl. Acad. Sci. U. S. A.* **87,** 2414–2418 (1990).

208. Roberts, R. J., Carneiro, M. O. & Schatz, M. C. The advantages of SMRT sequencing. *Genome Biol.* **14,** 1–4 (2013).

209. Sakai, H. *et al.* The power of single molecule real-time sequencing technology in the de novo assembly of a eukaryotic genome. *Sci. Rep.* **5,** 16780 (2015).

210. Ghedin, E. *et al.* Gene synteny and evolution of genome architecture in trypanosomatids. *Mol. Biochem. Parasitol.* **134,** 183–191 (2004).

211. Souza, R. T. *et al.* Genome size, karyotype polymorphism and chromosomal evolution in Trypanosoma cruzi. *PLoS One* **6,** e23042 (2011).

212. Weckselblatt, B., Hermetz, K. E. & Rudd, M. K. Unbalanced translocations arise from diverse mutational mechanisms including chromothripsis. *Genome Res.* **25,** 937–947 (2015).

213. Lewis, M. D. *et al.* Flow cytometric analysis and microsatellite genotyping reveal extensive DNA content variation in Trypanosoma cruzi populations and expose contrasts between natural and experimental hybrids. *Int. J. Parasitol.* **39,** 1305–1317 (2009).

214. Lewis, M. D. *et al.* Recent , Independent and Anthropogenic Origins of Trypanosoma cruzi Hybrids. **5,** (2011).

215. Minning, T. a., Weatherly, D. B., Flibotte, S. & Tarleton, R. L. Widespread, focal copy number variations (CNV) and whole chromosome aneuploidies in Trypanosoma cruzi strains revealed by array comparative genomic hybridization. *BMC Genomics* **12,** 139 (2011).

216. Myler, P. J. Molecular variation in trypanosomes. *Acta Trop.* **53,** 205–225 (1993).

217. Branche, C., Ochaya, S., Aslund, L. & Andersson, B. Comparative karyotyping as a tool for genome structure analysis of Trypanosoma cruzi. *Mol. Biochem. Parasitol.* **147,** 30–38 (2006).

218. Ibarra-Laclette, E. *et al.* Architecture and evolution of a minute plant genome. *Nature* **498,** 94–98 (2013).

219. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).

220. Bishop, A. J., Louis, E. J. & Borts, R. H. Minisatellite variants generated in yeast meiosis involve DNA removal during gene conversion. *Genetics* **156,** 7–20 (2000).

221. Gendrel, C. G., Boulet, A. & Dutreix, M. (CA/GT)(n) microsatellites affect homologous recombination during yeast meiosis. *Genes Dev.* **14,** 1261–1268 (2000).

222. Hunter, N. Meiotic Recombination: The Essence of Heredity. *Cold Spring Harb. Perspect. Biol.* **7,** (2015).

223. Weedall, G. D. & Hall, N. Sexual reproduction and genetic exchange in parasitic protists. *Parasitology* **142 Suppl 1,** S120–7 (2015).

224. Müller, J. & Hemphill, A. Drug target identification in protozoan parasites. *Expert Opin. Drug Discov.* **11,** 815–824 (2016).

225. Kodama, Y., Shumway, M., Leinonen, R. & International Nucleotide Sequence Database Collaboration. The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.* **40,** D54–6 (2012).

226. Ledford, H. AstraZeneca launches project to sequence 2 million genomes. *Nature* **532,** 427 (2016).

227. Wildenhain, J. *et al.* Prediction of Synergism from Chemical-Genetic Interactions by Machine Learning. *Cell Syst* **1,** 383–395 (2015).

228. Lötsch, J. & Ultsch, A. Process Pharmacology: A Pharmacological Data Science Approach to Drug Development and Therapy. *CPT Pharmacometrics Syst Pharmacol* **5,** 192–200 (2016).