Escuela de Bio y Nanotecnologías (EByN)

Universidad de San Martín

Doctorado en Biología Molecular y Biotecnología

Tesis de Doctorado

Métodos computacionales para estudio a gran escala de la respuesta humoral humana frente a patógenos: aplicaciones en la Enfermedad de Chagas.

PhD Thesis

Computational methods for large-scale analysis of the humoral immune response against pathogens in humans: applications in Chagas Disease.

Marzo, 2023

Autor:
Lic. Alejandro Daniel Ricci

Director:
Dr. Fernán Agüero

Contents

Agradecimientos

Bueno, acá estamos. Tras miles de líneas de código, cientos de papers leídos, decenas de congresos y una pandemia, llegó por fin el momento de defender mi tesis y terminar mi doctorado.

Es un momento muy extraño. Cerrar un proyecto que formó gran parte de mi vida y abrir la puerta al futuro incierto que se viene. Da ansiedad e incertidumbre, pero también curiosidad y entusiasmo. Pero bueno, esas son cosas con las que tendré que lidiar mañana. Hoy quería agradecer a mucha gente sin la cual nada de esto hubiera sido posible.

Primero que nada a mi director, Fernán, por darme un lugar en el laboratorio y por ser mi guía en este camino sinuoso que es la bioinformática. Muchas gracias por toda la paciencia, la buena onda y las picadas a lo largo de estos años. Perdón por ser un hereje que sigue usando Windows.

A mis compañeros de laboratorio sin los cuales todo esto habría sido infinitamente más aburrido. A aquellos que estaban para recibirme (Bruno, Carol, Emilio, Ibel, Lanza, Lean, Leo, Lio, Tano), y a los que se fueron agregando a lo largo de los años (Armando, Carla, Emir, Heli, Juli, Mer, Ramiro, Seba). Muchas gracias por todos los almuerzos, encuentros y charlas filosóficas que tuvimos y espero sigamos teniendo en años futuros.

A la institución en la cual pasé estos años de mi vida, el IIB. Gracias por darme un lugar donde podemos trabajar sintiéndonos acompañados. Un lugar donde no siempre funciona todo, pero donde siempre hay personas que tratan que funcione todo, lo cual es incluso más raro. Gracias también al comedor MENSA por hacerme conocer varias comidas ricas que soy demasiado vago para cocinar. Prometo tratar de hacer alguna.

A Alexandra Elbakyan, fundadora de Sci-Hub. Gracias por arriesgar tu libertad para hacer que el mundo de la ciencia esté un poco más cerca de lo que debería ser.

Por último a mi familia. A mis padres Paco y Norma y a mi hermana Valeria, gracias por estar ahí, día tras día, apoyándome en este viaje. Gracias por escuchar mis protestas cuando los experimentos salían mal y por compartir mis festejos cuando salían bien. Gracias por darme sus opiniones cuando tenía dudas, y por apoyarme por más que no las siga. Sepan que valoro muchísimo todo lo que me ayudaron estos años.

So long, and thanks for all the fish

Ale

About this thesis

Resumen

En los últimos años, los microarreglos de péptidos han evolucionado considerablemente, permitiendo realizar ensayos serológicos de alto rendimiento, donde un gran número de péptidos cortos es analizado en paralelo. Esto abrió la puerta a muchas nuevas posibilidades de investigación, dos de las cuales son descriptas en esta tesis. Primero, se utilizaron datos ya existentes para entrenar un algoritmo que predice antigenicidad en proteínas. Segundo, utilizando microarreglos peptídicos de alta densidad, se analizó el proteoma completo de *Trypanosoma cruzi* (agente causal de la enfermedad de Chagas) con el objetivo de descubrir y caracterizar nuevos epítopos B lineales y estudiar la diversidad de las respuestas inmunes de anticuerpos en diferentes poblaciones humanas. El primer capítulo presenta una introducción al marco teórico y de antecedentes de esta tesis.

El segundo capítulo de esta tesis describe el método computacional que llamamos APRANK (Antigenic Protein and Peptide Ranker), que utiliza diversas propiedades moleculares de las proteínas y péptidos de un patógeno para predecir antigenicidad. APRANK fue entrenado con información de antigenicidad de un conjunto filogenéticamente diverso de 15 patógenos humanos (bacterias y protozoos). El rendimiento de APRANK fue analizado y validado en *Onchocerca volvulus* y *Plasmodium falciparum*. APRANK tuvo éxito prediciendo antigenicidad en todas las especies donde se lo probó, facilitando la selección de grupos de proteínas que estén enriquecidos en proteínas antigénicas.

El tercer capítulo describe el uso de microrreglos peptídicos de alta densidad para el descubrimiento y caracterización de antígenos y epítopos en el contexto de la Enfermedad de Chagas. El repertorio de anticuerpos desarrollados como respuesta a una infección representa una valiosa fuente de marcadores diagnósticos. Estos repertorios fueron estudiados utilizando microrreglos peptídicos de alta densidad y muestras de sueros de pacientes infectados con *T. cruzi* (y controles sanos) a lo largo del continente americano. La búsqueda de antígenos se realizó sobre 30.500 proteínas de dos cepas de *T. cruzi* a nivel individual y poblacional. La matriz de seroprevalencia derivada contiene información de la antigenicidad de todos los epítopos encontrados en 71 sueros individuales. Estos datos permiten estudiar el repertorio inmune de la Enfermedad de Chagas con un detalle sin precedentes, al mismo tiempo que proveen una valiosa fuente de biomarcadores serológicos.

Finalmente, una aplicación web desarrollada tambien en esta tesis que permite navegar y explorar los datos generados en este proyecto se presenta brevemente en el cuarto capítulo.

En resumen, estos capítulos muestran métodos (para varios organismos) y una gran colección de datos experimentales (para la enfermedad de Chagas) que en conjunto aumentan nuestra comprensión de qué es lo que el sistema inmune reconoce en proteínas antigénicas. Identificar qué péptidos son detectados por anticuerpos durante una infección puede ser usado para guiar la producción de nuevos reactivos para mejorar los diagnósticos existentes o desarrollar nuevos inmunoensayos para, por ejemplo, monitorear el tratamiento quimioterápico de pacientes y/o detectar en forma temprana la evolución de la patología.

Palabras clave: Trypanosoma cruzi; enfermedad de Chagas; microarreglo peptídico; serología; anticuerpos; diagnóstico; antígenos; epítopos B; bioinformática; predicción

Abstract

The past few years have seen the advent and evolution of peptide microarray platforms, making it possible to perform high-throughput serological screening of short peptides. This opened the door for many new avenues of research, two of which are described in this thesis. First, we leveraged available datasets resulting from these experiments to create an algorithm that predicts antigenic proteins. Second, we have taken advantage of high-density peptide microarrays to analyze whole proteomes and drive the discovery and characterization of novel linear B-cell epitopes in *Trypanosoma cruzi*, the causative agent of Chagas Disease. We also used these platforms to study the diversity of human antibody repertoires across human populations. The first chapter present an introduction to the background and theoretical framework behind this thesis.

The second chapter of this thesis describes a computational method called APRANK (Antigenic Protein and Peptide Ranker) which integrates multiple molecular features to prioritize potentially antigenic proteins and peptides in a given pathogen proteome. APRANK was trained with antigenicity information from a wide phylogenetic selection of 15 human pathogens (bacteria and protozoa). Performance of APRANK was assessed using non-parametric ROC-curves and it was validated using *Onchocerca volvulus* and *Plasmodium falciparum* as independent data sets. APRANK was successful in predicting antigenicity for all pathogen species tested, facilitating the production of antigen-enriched protein subsets.

The third chapter describes the use of high-density peptide microarrays for the large scale discovery and characterization of antigens and epitopes in the context of Chagas Disease, a lifelong infection caused by the protozoan parasite *Trypanosoma cruzi*. During such a chronic infection, the immune system produces a repertoire of specific antibodies against the pathogen that represent a rich source of diagnostic markers. Here, these repertoires were studied using high-density peptide arrays to analyze serum samples from Chagas Disease patients across the Americas. This thesis presents the proteome-wide search for antigens accross 30,500 proteins in two strains of *T. cruzi*, as well as the subsequent fine mapping of the identified linear epitopes at the individual level and across human populations. During the steps of epitope characterization we also obtained a rich seroprevalence matrix for 71 individuals for all discovered epitopes. These datasets enable the study of the Chagas antibody repertoire at an unprecedented depth and granularity, while also providing a rich dataset of serological biomarkers.

Finally, a website application developed in this thesis allows easy access to the data produced in this project. This is described in the fourth chapter.

In summary, these chapters show methods (for disparate organisms) and a large data set (for Chagas disease) that in concert augment our comprehension of what it is that the immune system recognizes in antigenic proteins. Identfying which peptides are detected by antibodies derived from natural infections can be used to guide the production of new reagents to improve diagnostics and develop new immunoassays such as those necessary to monitor chemotherapeutic treatments of Chagas disease patients and/or provide early detection of evolution of pathology.

Keywords: Trypanosoma cruzi; Chagas Disease; peptide microarray; serology; antibodies; diagnostics; antigens; B-cell epitopes; bioinformatics; prediction

Publications and manuscripts

Publications and manuscripts included in this thesis:

2021 **APRANK: computational prioritization of antigenic proteins and peptides from complete pathogen proteomes.** Alejandro D. Ricci, Mauricio Brunner, Diego Ramoa, Santiago J. Carmona, Morten Nielsen and Fernán Agüero. *Frontiers in Immunology*, volume 12. DOI: 10.3389/fimmu.2021.702552

2023 **The Trypanosoma cruzi Antigen and Epitope Atlas: antibody specificities in Chagas Disease patients across the Americas.** Alejandro D. Ricci, Leonel Bracco, Emir Salas-Sarduy, Janine Ramsey, Melissa S. Nolan, M. Katie Lynn, Jaime Altcheh, Griselda E Ballering, Faustino Torrico, Norival Kesper, Juan C. Villar, Iván S Marcipar, Jorge D. Marco and Fernán Agüero. *Nature Communications*, volume 14: 1850. DOI: 10.1038/s41467-023-37522-9

Other related publications and manuscripts:

2021 **Serological Approaches for Trypanosoma cruzi Strain Typing.** Virginia Balouz, Leonel Bracco, Alejandro D. Ricci, Guadalupe Romer, Fernán Agüero and Carlos A. Buscaglia. *Trends in Parasitology*, volume 37, number 3, pages 214-225. DOI: 10.1016/j.pt.2020.12.002

2023 **Deep serological profiling of the *Trypanosoma cruzi* TSSA antigen reveals different epitopes and modes of recognition by Chagas disease patients.** Guadalupe Romer, Leonel Andres Bracco, Alejandro D. Ricci, Virginia Balouz, Luisa Berná, Juan Carlos Villar, Janine M Ramsey, Melissa S Nolan, Faustino Torrico, Norival Kesper, Jaime Altcheh, Carlos Robello, Carlos A. Buscaglia, Fernán Agüero. *Submitted*.

2023 **Biomarkers of pathology in chronic Chagas disease: changes in the antibody repertoire of patients with chagasic cardiomyopathy**. Alejandro D. Ricci, Justo Carbajales, Mario Principato, Leonel Bracco, Juan Mucci, Fernán Agüero. *In preparation.*

Aims and Objectives

Aims

1. Develop a computational method to predict *in silico* if a peptide or protein from a pathogen is likely to be detected by the immune system of the host.
2. Analyze immune responses against *T. cruzi* across diverse human populations in different geographic areas to discover antigens and epitopes.
3. Identify proteins of interest that might improve the diagnosis and treatment of Chagas disease.

Objectives

- Prediction Model
    1. For a pathogen of interest, obtain a list of known linear epitopes detected by the host B-cells. Repeat this process for several human pathogens, curate and parse the data as necessary.
    2. Focusing on variability and relevant biological properties, select a group of existing computational methods to analyze different features of proteins and peptides based on their sequence.
    3. Using these features, train and test prediction models and analyze their performance.
    4. Validate the best model on additional datasets (not used for training).
- Immunomics of *T. cruzi*
    1. Design high-density peptide microarrays containing overlapping peptides from the proteomes of distinct lineages of *T. cruzi*.
    2. Develop software to analyze and visualize the data, obtaining and characterizing an exhaustive list of *T. cruzi* linear epitopes.
    3. Analyze the seroprevalence for each studied linear epitope of *T. cruzi* using serum samples from individuals from different regions throughout the Americas.
    4. Identify *T. cruzi* linear epitopes with high seroprevalence, which are promising candidates to be used in future studies to improve existing diagnostic methods.
    5. Develop an interactive web application to enable fast and easy exploration of the produced data.

1. Introduction

Humanity has been at war against infectious diseases for as long as it has existed, whether we knew it or not. Observation and deduction allowed us to better understand the origins of these diseases (*"don't eat raw meat", "don't bath in that river", "don't forget to sterilize that"*) and how to treat them (*"eat this herb", "eat this fungus", "eat this pill"*). In these past hundreds of years our knowledge of diseases increased exponentially, aided by tools that allowed us to analyze them with an unprecedented level of detail. However, while we now have information on 7,000 to 10,000 different diseases [1], only a small fraction have an actual treatment or cure [2].

The diseases without cure can be divided into two groups. In the first group are the diseases for which humanity simply does not have the knowledge or technology necessary to cure them. This is understandable, and as humans evolve, less and less diseases should fall into this category. The problem we need to solve is the existence of the other group of diseases, the ones that could be treated, but unfortunately are not, mainly due to funding or logistic reasons. Some of these diseases even have promising findings already published, but for many of them, the rate of translation of knowledge into applications lags far behind the rate at which this knowledge is being generated [2].

It would be somewhat understandable if these relegated diseases were all incredibly rare or affected very few people; however, this is not the case. There is a noticeable disparity in the funding of diseases that affect wealthy people and those that do not. The World Health Organization (WHO) and the Public Library of Science (PLOS) have released lists of *"Neglected Tropical Diseases"* [3], which are underfunded diseases that are predominantly located in tropical areas across the globe and mainly affect resource-poor communities, and for which the WHO estimates 1.7 billion people requires treatment every year [4].

What can be done about this? The ideal solution would be a sustained investment effort by public and private sponsors on the research agenda for these diseases; however, we have no control over that. A more realistic approach from our place as scientists is to make the most out of our limited resources, and to intelligently reuse data obtained from the research of other better known diseases.

This thesis will present two avenues to obtain large amounts of valuable information about the pathogen responsible for a given disease, focusing on finding which proteins and peptides from those pathogens are the ones recognized by the adaptive immune system. The first approach leverages large amounts of curated data to develop a bioinformatics tool with the ability to predict antigenicity of proteins and peptides, starting from a pathogen's proteome. The second approach entails analyzing the blood of infected individuals using high-density microarrays containing peptides from the pathogen of interest. The experimental analysis in this PhD thesis will focus on *Trypanosoma cruzi* and Chagas disease, one of the diseases labeled as *"Neglected Tropical Disease"* by the WHO.

1.1 Organization of this Thesis

Chapter 1 of this thesis introduces key concepts used in this work. Basic concepts of immunology and prediction models will be briefly defined, whilst logistic regression models, peptide microarray technology and Chagas disease will be introduced with more detail and with a special emphasis on serological diagnosis.

Chapter 2 of this thesis describes the development of a computational method called APRANK (Antigenic Protein and Peptide Ranker) which integrates multiple molecular features to prioritize potentially antigenic proteins and peptides in a given pathogen proteome. These features include subcellular localization, presence of repetitive motifs, natively disordered regions, secondary structure, transmembrane spans and predicted interaction with the immune system. We trained and tested this method with several pathogenic bacteria and protozoa, we evaluated this integrative method using non-parametric ROC-curves and leave-one-out cross-validation, and made an unbiased validation using an independent data set.

Chapter 3 of this thesis describes the use of high-density peptide microarrays as a platform for antigen discovery and epitope mapping. In this work we have analyzed the whole proteome of two strains of *Trypanosoma cruzi*, in a two-step strategy. The first step focused on a proteome-wide analysis using pooled sera from diverse human populations across the Americas to find antigenic regions. The second step used individual sera to investigate the seroprevalence for each of the regions found in the first step. With this information we identified novel antibody-binding epitopes associated with the chronic phase of Chagas disease, which have the potential to improve both its diagnosis and facilitate the development of other immunoassays. Through this analysis we also obtained a deeper understanding on the adaptive immune response against a pathogen like *T. cruzi*, finding that most regions were antigenic only for a few individuals, while others were detected by many. These data sets enable the study of the Chagas antibody repertoire at an unprecedented depth and granularity, while also providing a rich source of novel serological biomarkers.

Chapter 4 of this thesis describes the development of a website application that allows easy access to the data produced in the analysis of the whole proteome of two strains of *Trypanosoma cruzi* using high-density peptide microarrays and serum samples from across the Americas (see Chapter 3). The application was created using *Shiny* (an R package) and is an interactive interface that enables exploration of raw and parsed data for all proteins, peptides and serums analyzed. The application also contains interactive visual representations of our data, such as antibody-binding plots.

1.2 B-cell antigens and epitopes

An epitope is the part of an antigen that is recognized by the immune system, specifically by antibodies, B-cells, or T-cells. This thesis deals only with peptidic epitopes recognized by antibodies and B-cell receptors and therefore we use the terms "antibody epitope", "B-cell epitope" and "epitope" interchangeably. By extension, we refer to any protein containing one ore more of these epitopes as a "B-cell antigen" or simply an "antigen".

1.2.1 Linear vs. Conformational epitopes

Epitopes of proteins are usually classified as either linear (continuous) or conformational (discontinuous) depending on whether the amino acids included in the epitope are contiguous in the peptide chain or not. This categorization is not always clear because, for example, it is common for conformational epitopes to contain stretches of a few contiguous residues [5]. Another example is that not every residue in a linear epitope is involved in the binding with the antibody, resulting in some amino acid residues in the linear epitope that can be replaced by others without impairing antibody binding. In practice, any continuous peptide fragment of a protein that is able to bind to antibodies is called a linear epitope [6].

Improved methods of solid-phase peptide synthesis have turned synthetic peptides into a convenient tool for studying the epitopes of proteins. However, currently only linear peptides can be successfully synthesised with these methods, meaning that any analysis based on solid-phase peptide synthesis will miss numerous conformational epitopes (which can be identified by X-ray crystallography of antigen–antibody complexes). In spite of this limitation, linear synthetic peptides have been used extensively for studying the immunological properties of proteins and for replacing them as reagents in the diagnosis of infectious and auto-immune diseases as well as potential synthetic vaccines [7–10].

1.2.2 Antigenicity vs. Immunogenicity

The ability of a peptide to react specifically with the functional binding site of a complementary antibody is known as its antigenic reactivity or antigenicity. While our understanding of peptide antigenicity has improved considerably in recent years, this knowledge focuses strictly on the chemical phenomenon of protein–peptide interactions. The situation is quite different for immunogenicity, which is the ability of the protein or peptide to induce an immune response in a competent host. This makes immunogenicity harder to predict than antigenicity, because it depends on many complex interactions with various elements of the host immune system [9]. The methods used in this thesis will focus on finding antigenicity, which is required but not sufficient to provide immunogenicity.

1.2.3 Development of an antibody response

The B-cell receptor (BCR) is an attached form of antibody, which has specificity for particular epitopes. Each B-cell expresses many BCR molecules on its surface, each with the same specificity. When a BCR binds an antigen, it may internalize it into the cell. If the antigen is a protein, the B-cell processes the antigen into smaller peptides, binds some of those peptides to MHC class II molecules, and presents these peptide–MHC complexes on the cell surface. If a helper (CD4+) T-cell has a T-cell receptor (TCR) that binds the peptide–MHC complex, then the T-cell sends a stimulatory

signal to the B-cell. Thus, B-cell stimulation requires binding to an epitope of an antigen, processing the antigen, and finding a helper T-cell that can bind an epitope of the same antigen [11].

The epitopes recognized by the BCR and TCR may differ, but must be linked on the same antigen molecule to provide matches to both the BCR and TCR. Stimulation of T-cells causes B-cells to divide more rapidly, to undergo somatic hypermutation (in a stage called affinity maturation occurring in germinal centers of the lymphoid tissue), and to switch from IgM to IgG production (and other isotypes). Higher affinity selected B-cells will either differentiate into memory B-cells or into plasma cells, which will begin to secrete higher-affinity and isotype-switched antibodies [11].

Although antibody responses to most protein antigens are dependent on helper T-cells, some bacterial polysaccharides, polymeric proteins, and lipopolysaccharides have special properties that enable them to stimulate naive B-cells in the absence of peptide-specific T-cell help. These antigens are known as thymus-independent antigens (TI antigens) because they can stimulate strong antibody responses in athymic individuals. However, B-cell responses to these TI antigens are influenced by the presence of T-cells, perhaps indirectly through cytokines such as IL-5 since they are greatly diminished in animals that have no T-cells at all. These antigens are important components of the humoral immune response to non-protein antigens that do not engage peptide-specific T-cell help. TI antigens fall into two classes that activate B-cells by two different mechanisms: TI-1 antigens and TI-2 antigens [12].

TI-1 antigens possess an intrinsic activity that can directly induce B-cell division. At high concentration, these molecules cause the proliferation and differentiation of most B-cells regardless of their antigen specificity; this is known as polyclonal activation and TI-1 antigens are thus often called B-cell mitogens. An example of a TI-1 antigen is LPS, which binds to LPS-binding protein and CD14, which then associate with the receptor TLR-4 on B-cells. As mentioned, these antigens activate B-cells only at doses at least 100 times greater than those needed to activate dendritic cells; when B-cells are exposed to lower concentrations of TI-1 antigens, only those B-cells whose B-cell receptors also specifically bind the TI-1 molecules become activated. TI-1 antigens have an important role in defense against several extracellular pathogens, as they arise earlier than thymus-dependent responses since they do not require prior priming and clonal expansion of helper T-cells. However, TI-1 antigens are inefficient inducers of isotype switching, affinity maturation, or memory B-cells, all of which require specific T-cell help[12].

TI-2 antigens consist of molecules such as bacterial capsular polysaccharides that have highly repetitive structures. They contain no intrinsic B-cell-stimulating activity and, unlike TI-1 antigens, they can activate only mature B cells. TI-2 antigens are thought to act by extensively cross-linking the B-cell receptors of mature B cells specific for the antigen. Excessive receptor cross-linking, however, renders mature B cells unresponsive or anergic, just as it does immature B cells. It has been shown that while TI-2 responses are still present in mice that lack a thymus, a complete depletion of T-cells in the mice eliminates responses to TI-2 antigens; how T-cells contribute to TI-2 responses is not clear. Both IgM and IgG antibodies are induced by TI-2 antigens and are likely to be an important part of the humoral immune response in many bacterial infections, such as *Haemophilus influenzae* type B (Hib), where the antibodies to the capsular polysaccharide (which is a TI-2 antigen) have an important role in protective immunity to this bacterium [12].

1.2.4 Immunodominance

Immunization of animals with an exogenous protein or peptide (perhaps using carrier conjugates or adjuvants) nearly always gives rise to high titres of antibodies that bind specifically to that molecule. However, when multiple immunogens are simultaneously presented to the immune system, a competition takes place and the immune response is mounted only against a few immunodominant epitopes. This phenomenon is called immunodominance.

During an infection, a pathogen presents a large number of potential epitopes to the host's immune system, but because of immunodominance, the immune response will be focused on a few immunogenic epitopes while the majority of other subdominant epitopes (otherwise immunogenic) will be ignored [13]. Immunodominance occurs in both T-cell and B-cell responses and their mechanisms are poorly understood [13, 14]. Many factors have been suggested to play a role in determining TCL immunodominance, such as MHC binding, cellular processing, the repertoire of TCR specificities and active participation of CD8 T-cells [14, 15].

One proposed mechanism of immunodominance in B-cell activation is related to the affinity of epitope binding to the B-cell receptor (BCR). The diverse, naïve B-cells secrete IgM antibodies that bind to nearly any epitope. Upon the establishment of an active infection, B-cells that bind epitopes with relatively high equilibrium affinity divide rapidly and dominate the early phase of the immune response by outcompeting other B-cells. However, antibodies that bind too strongly clear the matching antigens quickly and prevent feedback stimulation to their B-cells. On the opposite end of the scale where BCRs have low affinity for the epitopes, these B-cells are outcompeted for stimulation by B-cells with BCRs that have higher affinities for their respective epitopes [13]. This defines an affinity window for immunodominant epitopes in the early immune response. Insufficient T-cell stimulation by these B-cells also leads to suppression of these B-cells by the T-cells [13]. The later phases of B-cell competition and maturation of IgG favor antibodies with increased on-rates of association to epitopes rather than increased equilibrium binding affinity [13, 16]. In addition, the temporal sequence in which the host encounters antigenic variants also influences the specificity of the immune response [13]; hence, immunodominance is also modulated by the antigenic history of the host (e.g., vaccinations, previous infections, environmental antigens).

1.2.5 Serodiagnostic antigens

In our work we will focus on discovering antigens with potential for serological diagnosis for infectious diseases. In this context, serodiagnostic antigens are defined as those proteins or peptides with the ability to discriminate infected patients from healthy controls in an appropriate assay.

During active infections, the humoral immune response is directed against immunodominant epitopes of immunodominant antigens expressed by the pathogen. Therefore, serodiagnostic antigens should contain one or more epitopes that are cross-reactive with these immunodominant pathogen epitopes in their native states (as presented to the immune system).

However, identification of these immunodominant antigens is not a trivial matter, because there is still limited knowledge on what intrinsic features differentiate a pathogen protein targeted by the immune system from other proteins expressed by the pathogen. In this thesis we will detail two methods to obtain this information: peptide microarrays, which are used to directly test the antigenicity of the complete proteome of the pathogen, and prediction models, which use computer algorithms to infer antigenicity based on known properties of each protein or peptide.

1.3 Immunoassays

The term "immunoassay" refers to an analytical method that uses antibodies or antibody-related reagents for the determination of sample components. Because antibody binding is highly selective, these reagents can be used directly even on complex biological matrices such as whole or permeabilized cells, blood, plasma, or urine and still result in methods that are highly specific [17]. In this thesis we will focus on immunoassays that detect the presence of antibodies against a given pathogen in patient sera.

Immunoassay technologies are routinely used for serological diagnosis, where immobilized serodiagnostic antigens capture antigen specific antibodies from serum samples. Some of these assays are: ELISA, immunoblotting, complement fixation, and immunofluorescence tests. We will expand on the two methods that are closely related to our work.

1.3.1 Enzyme-linked immunosorbent assay (ELISA)

The first technique that used antibody-mediated detection was described in 1960, where the antibody was linked to a radioactive signal (radioimmunoassay) [18]. However, because of the health risks, methods not involving radioactivity were sought. The discovery that certain enzyme-substrate combinations produced quantifiable colour changes led to a shift in immunodetection. In 1971, two independent research groups in Europe published papers that described the step-by-step process of performing an enzyme-linked immunosorbent assay, or ELISA, as it is most commonly known today [19, 20].

The ELISA method is used to detect and quantify a specific substance, usually an antigen, in a sample. Nowadays, ELISAs are commonly used in at-home pregnancy testing, and in point-of-care testing, such as to diagnose HIV, hepatitis B and malaria in a clinic [21]. There are four main types of ELISAs, each with its own advantages and disadvantages (see Figure 1.1).

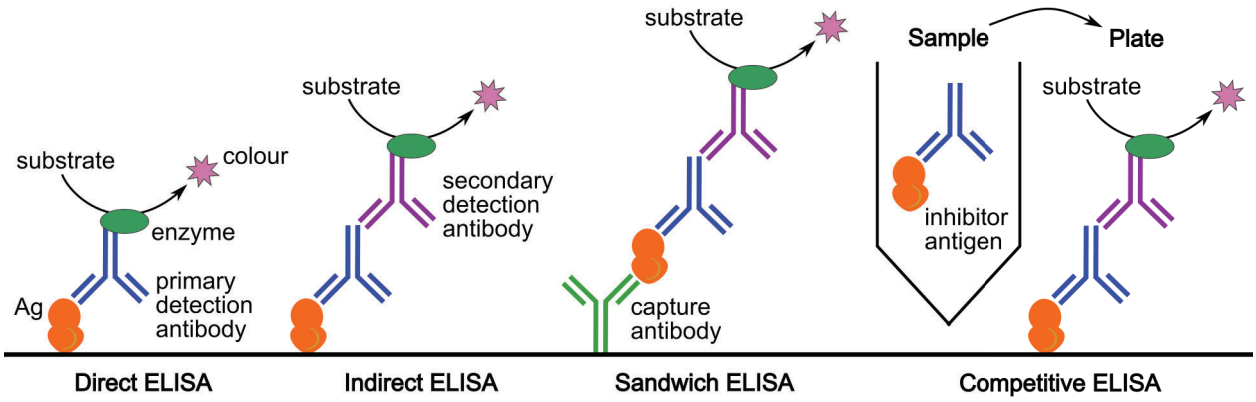


FIGURE 1.1 – ***Types of ELISA.*** *The figure shows the four main types of Enzyme-linked immunosorbent assays, or ELISAs.*

Direct ELISA

The first step to perform a direct ELISA is to immobilize the antigen in a microplate. This is usually done using a carbonate / bicarbonate coating buffer with a pH of at least 9, which facilitates passive absorption of the antigens onto the plate by allowing the antigens to remain soluble and ensuring that they have an overall negative charge that can bind to the positively charged plate [21].

Afterwards, the wells have to be washed with a "washing buffer" (phosphate buffered saline at neutral pH) to remove any unbound assay components, or debris, from the wells. During the washing phase the wells are repeatedly filled with phosphate buffered saline and emptied, which minimizes background noise during detection and increases the specificity of the assay. This washing has to be done here and repeated after each following step [21].

The second step is to add a "blocking buffer", such as bovine serum albumin, to saturate unoccupied binding sites and thus minimize non-specific binding and non-specific protein–protein interactions. A standardized blocking buffer has not been identified as suitable for all assays, but an ideal blocking buffer must have no cross-reactivity with other assay components, minimize denaturation and exhibit low enzyme activity [21].

After another wash, the third step is to add a "primary detection antibody", forming an antigen–antibody complex. This primary detection antibody is directly labelled with an enzyme, but, before adding the enzyme's substrate, the plate is incubated for sufficient time to permit the binding between antigen and antibody and then the wells are washed for the last time to remove excess antibody [21].

In the fourth and final step, the substrate is added, which by reacting with the enzyme attached to the antibodies produces a colour signal indicating the presence of the antigen in the sample. The enzyme–substrate interaction is left in a dark environment for some time and then the reaction is stopped with a specific solution. The enzyme–substrate interaction leads to colour formation that can be detected by a microplate reader and the measurement of the optical density is proportional to the quantity of antigen in the sample [21].

The direct ELISA is the simplest version of the ELISA. It is quicker than other assays and uses only one kind of antibody per test, but it has a lower sensitivity and it requires different labeled antibodies for each antigen of interest [21].

Indirect ELISA

This version is similar to the direct ELISA, but in this case the primary detection antibody has no label. Instead, the fourth step in the indirect ELISA is to add a "secondary detection antibody", which is an antibody that is labelled with an enzyme and has specificity for the primary detection antibody. After washing, the substrate is addded and the produced colour is measured as in the direct ELISA [21].

The two-antibodies method used in the indirect ELISAs allows for signal amplification and makes it easier to use many primary antibodies, resulting in this type of ELISA having high sensitivity and flexibility. However, this method also comes with a chance of cross-reaction occuring with the secondary antibody, which can result in a non-specific signal and can be quantified using calibration curves to compare the assay relative to the analyte being detected [21].

Sandwich ELISA and Competitive ELISA

The two last types of ELISAs are not related to our work, but deserve a mention. In the Sandwich ELISA, the antigen is captured between two antibodies, a "capture antibody", which is immobilized in the microplate, and a "detection antibody", which is added after adding the sample. The detection antibody can have its own label (as in the direct ELISA) or, more likely, the method needs a third labeled antibody (as in the indirect ELISA). This two-antigen *sandwich* makes it possible to obtain high sensitivity and specificity while using samples that had almost no purification (such as in home pregnancy kits). However, this method needs previous knowledge of a primary and secondary antibody that binds to different epitopes on the antigen and work together as matched pairs (meaning they do not compete for antigen binding sites) [21].

The Competitive ELISA is different to the other ELISAs as it uses a competitive binding process. The primary antibody is incubated with an unpurified sample (e.g. in a tube), and binds to any antigen present in the sample. This mix is then added to wells that are pre-coated with the antigen. Any unbound antibodies in the mix will bind to the antigen in the well, so there is an effective competition for the antibodies between the antigens in solution and those in the pre-coated well. After washing and blocking, a secondary antibody conjugated with an enzyme is added that binds to the primary antibody, similar to an indirect ELISA. Competitive ELISA produces an inverse curve, such that a high amount of antigen in the sample yields a lower signal from the substrate, and has the benefits of needing minimal sample purification as well as the ability to measure large range of antigens in a sample. It's specially used for small antigens, where it is not possible to use a sandwich ELISA (because of the need of two independent binding sites in the antigen). However, it has low specificity, so it should not be used for dilute samples [21].

1.3.2 Immunofluorescence

A fluorophore is a a fluorescent chemical compound that can absorb light at a specific wavelength resulting in light emission at a longer or lower energy wavelength. These wavelengths vary greatly between fluorophores; a commonly used fluorophore *Alexa 488* has an excitation peak of 495nm and emits light with an emission peak at 519nm (green spectrum), while another fluorophore *Alexa 594* has an excitation peak of 590nm and an emission peak at 617nm in the red spectrum [22].

It is possible to conjugate a fluorophore to an antibody, resulting in a very useful tool to detect and quantify diverse antigens. This technique is called immunofluorescence (IF) and can be used to visualize proteins in cells (both in suspension and adherent cells), in tissues, as well as in 3D culture-derived spheroids. Combined with the use of a confocal microscope, IF has also the ability to determine the cellular localization of a protein of interest. Similar to the ELISAs, the fluorophore can be attached to the antibody that detects the antigen (direct immunofluorescence) or attached to a secondary antibody that recognizes the unlabeled primary antibody bound to the antigen (indirect immunofluorescence, or IIF). The indirect immunofluorescence technique has the benefit of signal amplification, which is extremely useful for detecting low-abundance targets [22].

Because distinct fluorophores have different excitation and emission wavelengths, multiple antigens can be visualized on the same biological sample by conjugating multiple labelled antibodies to different fluorophores with distinct excitation and emission spectra. This has many uses, specially in experiments that require co-localization of proteins [22].

Fluorescence-linked immunosorbent assay (FLISA)

Immunofluorescence is also used in the fluorescence-linked immunosorbent assay, or FLISA, a variant of an indirect ELISA where the secondary detection antibody is labelled with a fluorophore and the resulting emission is quantified using a fluorometer [21] (see Figure 1.2).

FLISA has many advantages over ELISA, such as not needing a substrate (which also makes the experiment faster by removing one step), requiring less antigen and conjugated antiserum per sample, having a higher signal correlation with concentrations of antigen (specially at low concentrations) and having a higher detection threshold [23–25]. Because FLISA uses fluorophores, it also enables using multiples sets of antibodies at the same time (often up to 2 sets), making it possible to analyze either multiple antigens or multiple types of antibodies (such as IgG and IgM) in the same experiment [24].

The technology behind FLISA is the basis of how most protein and peptide microarrays are used for antigen discovery and epitope mapping or for profiling immune responses [26].

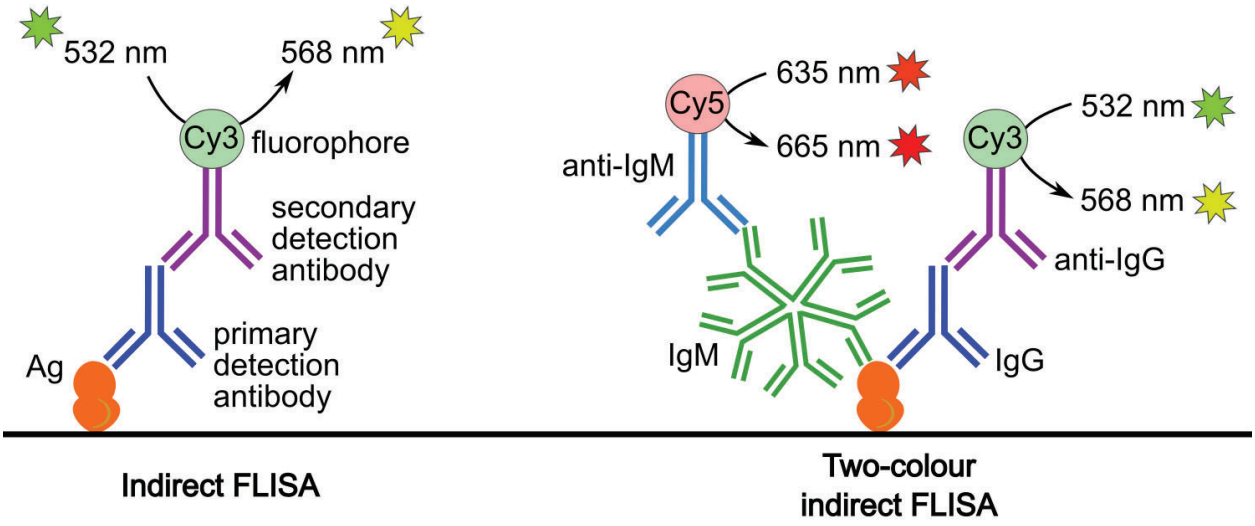


Figure 1.2 – ***Types of FLISA.*** *The figure shows two uses of the Fluorescence-linked immunosorbent assay, or FLISA, a variant of an indirect ELISA where the secondary detection antibody is labelled with a fluorophore. Cy3 and Cy5 are examples of possible fluorophores to be used in this technique.*

1.4 Microarrays as a support platform for immunoassays

Conventional immunoassays such as those described above have a low throughput and they are unable to discern the fine specificities of heterogeneous antibody populations. To surpass these limitations, we set to use microarrays, an emerging technology which enables a highly sensitive high-throughput analysis of antigen specificity.

Briefly, microarrays use a solid surface as a support where several analytes (such as antibodies, DNA, proteins, or peptides) are placed or synthesized in predetermined spots, resulting in a matrix of hundreds, thousands or even millions of analytes. Because there is a known mapping between each spot in the microarray and any given analyte, it is possible to use a single experiment to test interactions for all those analytes in parallel (for example, testing a patient's serum against a protein microarray to find if that patient has antibodies against some of the proteins in the microarray, and if so, which ones).

There were many experiments that led up to the creation of microarrays [27, 28], but one of the most clear predecessor was the "antibody matrix" [29]. Created in 1983, the antibody matrix was a 1 cm by 1 cm solid surface containing a matrix of 100 distinguishable spots (10 columns and 10 rows). Each of these spots contained antibodies of distinct specificities which were capable of serving as minute specific immunoadsorbents for cells bearing on their surface the corresponding antigens. Using this method, scientists had a way to investigate the potential of simultaneous multiple determinations of specific cell surface antigens in one reaction incubation [29].

With time this technology evolved, the dots in the matrix decreased in size, and the process of placing or synthesizing them on the solid surface was automatized by robots [30]. The first time the word "microarray" was used to refer to these type of assay platforms was in 1995, where a microarray of complementary DNA (cDNA) was used to monitor the expression of many genes in parallel in *Arabidopsis thaliana* [31]. These DNA microarrays are also known as "gene chips".

While initially microarrays focused mostly on DNA, many other kinds are being used today. In this thesis we will focus on peptide microarrays, but we will also cover protein microarrays because they are related.

1.4.1 Protein microarrays

Protein microarrays consist on a solid surface where different proteins are immobilized in fixed known positions. This allows the protein microarray, also known as a "protein chip", to be used as an assay system that can obtain the information of many analytes in a single experiment [26]. In our case, we will focus on profiling immune responses by detecting antibody binding to immobilized proteins. To achieve this, the microarrays are incubated with a serum sample (primary antibody) from one or many individuals (e.g. a pool) and then the attached antibodies are measured directly or indirectly by detecting fluorescent or radio-isotope labels.

In the last decades, significant progress has been made after the development of whole-proteome protein arrays. This technology is based on a high-throughput process that includes PCR recombination cloning and expression platforms, which allow the production of thousands of recombinant proteins in parallel. Using this strategy, complete or partial proteome arrays have been produced for many pathogenic organisms [32–37]. This technology has led to the identification of hundreds of new antigens and allowed the first global studies of immune responses [38].

The limitations of protein microarrays are mainly related to the production of recombinant proteins. High-throughput production of proteins for this type of functional proteomics assays may need the combination of different expression platforms and hosts to achieve a significant

degree of the proteome. Optimizing for robust expression of globular, *trans*-membrane, or highly unstructured proteins may require combination of cell-free expression, and/or a combination of different expression hosts, and protocols [33, 39, 40]. However the recent introduction of on-chip expression may solve some of this issues [41, 42]. Still, while these types of microarrays can analyze whole proteomes up to a few thousand proteins (typical of prokaryotic genomes) it can be very laborious or difficult to use them to study larger proteomes such as those of complex eukaryotes.

1.4.2 Peptide microarrays

Peptide microarrays are similar to protein microarrays in that they enable high-throughput biological analysis, but the immobilized analytes are short peptides instead. There are two methods to synthesize the peptides to place on a microarray: Merrifield solid-phase peptide synthesis (SPPS) and *in situ* synthesis.

When using Merrifield solid-phase peptide synthesis, peptides are first synthesized and then placed (*"spotted"*) onto the microarray. Peptides prepared in this way are generally of high quality, having fewer impurities resulting from incomplete synthesis. This benefit of SPPS, however, comes with substantial expense and time associated with the synthesis of hundreds or thousands of distinct sequences [43]. A disadvantage of this method is that most spot formation techniques used for microarray fabrication [44] produced spots of different shapes, which are highly dependent on the spotting technique and other experimental conditions [45]. This can have negative consequences when using microarray data for comparative purposes. This type of disadvantages have been discussed at larged for DNA microarrays (e.g. in studies of differential expression [45]).

This drawback in part motivated the development of *in situ* peptide synthesis, where peptides are directly created on the microarray itself, usually by automated robots. *In situ* peptide synthesis has the benefits that it uses minimal amounts of reagents and eliminates the need (or possibility) for peptide purification. This yields substantial benefits in cost and time needed to prepare the arrays; however, this approach makes it difficult to verify the purity and quality of peptides in the arrays [43]. Even with these drawbacks, *in situ* synthesis is the usual method for most high-throughput analysis based on peptide microarrays.

The first uses of *in situ* synthesis date back to the mid-1980s, where peptides were synthesized on polyethylene or polyacrylic acid solid supports and were used to identify viral antigens for antibody binding using the ELISA assay [46, 47]. Afterwards, the *in situ* method became the basis for three significant approaches to peptide arrays: SPOT, particle-based synthesis, and photolithograpic methods [43].

In situ synthesis: SPOT method

The SPOT method uses fluorenylmethyloxycarbonyl (Fmoc) protected amino acids to synthesize peptides in parallel directly on a membrane support [48]. The process is based on iterative cycles of coupling and washing, where in the first step solutions containing the amino acids and coupling reagents are dispensed onto specific locations of the membrane, and in the second step the entire membrane is washed and treated to remove the terminal amino protecting groups. While initially the amino acid solutions were manually placed onto the membranes via pipettes, more recent versions use automated systems. One can synthesize peptides up to a length of 50 amino acids using this technique, although the optimal range is between 6 and 18 amino acids [49]. This made SPOT synthesis a widely used technique due to their substantially lower costs and the ability to rapidly prepare custom arrays [43]. However, the need to dispense specific reagents in different fields (addressable spots) of the solid support places a limitation on the densities that can be achieved.

In situ synthesis: Particle-based synthesis method

The second *in situ* synthesis method is particle-based synthesis, which was introduced in 2007 [50]. This method works in a similar way to the SPOT method, but it differs in the delivery the amino acids using a 24-ink laser printer to transfer toner particles with the Fmoc-protected amino acids in a solid form. The particles carrying the amino acid molecules are then melted, allowing the coupling reactions. The advantage of this method is that it results in smaller spot sizes, which enables higher density arrays [43].

In situ synthesis: Photolithograpic method

The third *in situ* synthesis method is the photolithographic method, which uses light to direct peptide synthesis on a solid support, typically glass [51]. This method, first implemented in 1991, is based on amino acid reagents protected with a photolabile group (such as the 6-nitroveratryloxycarbonyl group, NVOC), meaning that the next amino acid will be incorporated into a growing peptide only if the peptide is first irradiated to remove the protecting group. The method uses a set of photolithographic masks to select the locations on the solid support (synthesis fields, addressable spots) that will be irradiated, meaning, which growing peptides will add a given amino acid in each synthesis cycle. This method can be automated and can generate very high-density arrays, however, it also needs expensive mask sets and many cycles of synthesis, because only one of the 20 amino acids can be added in each cycle [43].

In recent years, the photolithographic method was improved by introducing a maskless format. In this version, digital light patterns are projected on the support, removing the protection group at specific positions. These light patterns are generated by processing of light using Digital Light Processors, a set of chipsets based on optical micro-electro-mechanical technology. Originally developed in 1987 at Texas Instruments, they are essentially digital micromirrors, which can direct light to precise micron-size locations, hence enabling the addition of amino acids in light-deprotected spots [43]. Another upgrade was the use of more reliable protection groups, such as 3'-nitrophenylpropyloxycarbonyl groups (NPPOC), which works in a similar way than NVOC [52], or tert-Butyloxycarbonyl groups (Boc), which needs the addition of physical barriers and photogenerated acid precursors (PGAs) [53, 54].

While this method requires specialized equipment, it can be used to synthesize millions of peptides on a single slide, opening the door to analyze proteins at a scale that was not possible before [43, 55, 56].

Using peptide microarrays to study antigenicity

One of the most common uses for peptide microarrays is to detect antibody binding to antigens. In these immunoassays, short synthetic peptides, typically consisting of 6 to 18 amino acids, are displayed on the solid support and then used as fixed antigens (probes) for binding of primary antibodies as in indirect ELISA or FLISA assays. The use of chemically synthesized peptides spanning the sequence of pathogen proteins represents one of the most practical ways for large scale identification of serodiagnostic epitopes [57].

While protein microarrays can also be used for this, peptide microarrays have many advantages, the main one being the ability to map the precise location where antibody binding occurs, meaning the epitope [43, 58]. This mapping can be made even more precise by designing peptides that have overlapping sequences and that are tiled spanning the antigen sequence [59].

Besides their capacity to decompose heterogeneous B-cell responses into antibody specificities to short targets, peptide arrays can also differentiate subtle changes in antibody abundance and specificity [43, 58]. Moreover, identified peptides can be directly used as serology reagents [60].

The usual process when studying antigenicity relies on incubation of microarrays with sera from patients or animals, then after several washing steps, reveal the presence of antigen-antibody complexes by adding a secondary antibody with the required specificity, usually anti IgG. This secondary antibody is labeled, usually with a fluorescence label that can be detected by a fluorescence scanner (similar to a FLISA assay). After excitation of microarrays to produce fluorescence, image acquisition and analysis allows extraction of the measurements associated to each peptide (see Figure 1.3). This type of assay is quantitative because the fluorescence signal intensity is correlated to the amount of primary antibody bound to a given peptide spot.

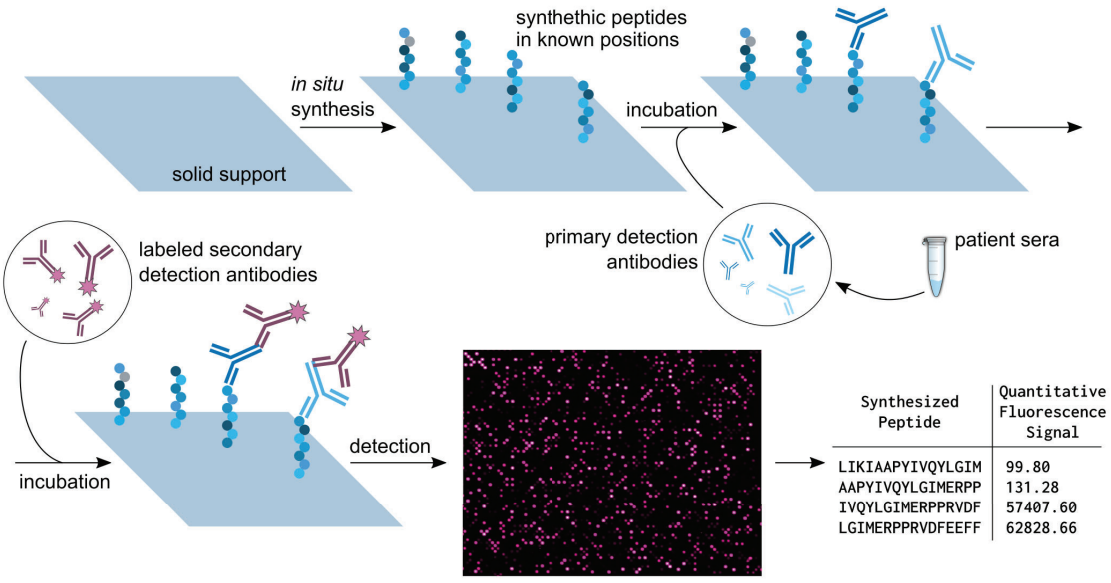


Figure 1.3 – ***Studying antigenicity via peptide microarrays.*** *The figure shows the main steps that are involved in using peptide microarrays to study antigenicity of peptides, usually with the objective of studying antigenicity of pathogenic proteins. The fluorescence signal is measured in arbitrary fluorescence units.*

The sequences of peptides in the microarray are usually derived from existing proteins, however there is a relatively novel "immunosignaturing" method that uses peptide microarrays with random sequences. In this scenario, microarrays are used to analyze how the humoral immune responds to molecular changes usually associated with the development of pathological processes [61]. Binding of antibodies in a sample to random peptides produces a pattern of reactivity, *i*.e. an "*immunosignature*" that is then correlated with a biological or pathological process. This technique has been used to study vaccination efficacy [62], diagnose cancer [61, 63], and study the humoral immune response in healthy humans [64, 65], as well as to predict protein epitopes [66].

There are many examples of peptide microarrays being used to analyze whole-proteome antigenicity. While initially this type of analysis was exclusive of viruses or small bacteria [67, 68], in recent years microarray technology and availability progressed to allow proteome-wide analysis for larger organisms, such as unicellular protozoans and nematodes [37, 69, 70]. In Chapter 3 of this thesis we show a proteome-wide analysis for two strains of *Trypanosoma cruzi*, a unicellular eukaryotic parasite, to identify antibody epitopes associated to Chagas disease, using peptide microarrays displaying close to 3 million peptides each.

1.5 Prediction models

Prediction models are mathematical functions or computer algorithms that allow us to infer future events or outcomes based on current data. There are many kinds of prediction models, and most of them have been used to predict biological features at some point. In this thesis we will focus on linear and logistic regression models, but we will also mention some other commonly used prediction models.

1.5.1 Linear regression models

Linear regression involves specifying the relationship between a single numeric dependent variable (the value to be predicted) and one or more numeric independent variables (the predictors). This kind of regression assumes that the relationship between the independent and dependent variables is linear, and can be fitted onto a linear mathematical function [71]. Linear regression has been used in biology to predict solvent accessibility [72], secondary structural content [73], protein-protein interaction site [74] and folding rate [75], among others.

Simple linear regression

A simple linear regression model defines, via a straight line, the relationship between a dependent variable and a single independent predictor variable. This relationship is shown in equation 1.1, where the intercept, α (alpha), describes where the line crosses the y axis, while the slope, β (beta), describes the change in y given an increase of x [71].

$$y = \alpha + \beta x \tag{1.1}$$

In order to determine the optimal estimates of α and β, an estimation method known as ordinary least squares (OLS) is used. In OLS regression, the slope and intercept are chosen such that they minimize the sum of the squared errors (SSE). The errors, also known as residuals, are the vertical distance between the predicted y value and the actual y value. Because the errors can be over-estimates or under-estimates, they can be positive or negative values. In mathematical terms, the goal of OLS regression can be expressed as the task of minimizing equation 1.2, where e_i is the error, meaning, the difference between the actual value (y_i) and the predicted value ($\hat{y}_i$) for each observation i. The error values are squared to eliminate the negative values and summed across all points in the data (n) [71].

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 \tag{1.2}$$

While the mathetmatic reasoning behind this exceeds the scope of this thesis, it can be shown using calculus that the values of α and β that results in the minimum squared error can be calculated as shown in equations 1.3 and 1.4, where $\bar{y}$ is the mean of the y values, $\bar{x}$ is the mean of the x values, $Cov(x, y)$ is the covariance function for x and y and $Var(x)$ is the variance of x [71].

$$\alpha = \bar{y} - \beta\bar{x} \tag{1.3}$$

$$\beta = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}}{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} = \frac{Cov(x,y)}{Var(x)} \tag{1.4}$$

Multiple linear regression

Most real-world analyses have more than one independent variable. In these scenarios it is necessary to use a different kind of linear regression called "multiple linear regression", which can be understood as an extension of simple linear regression. The goal in both cases is similar: to find values of slope coefficients that minimize the prediction error of a linear equation. The key difference is that for multiple linear regression there are additional terms for the additional independent variables [71].

Multiple regression models are in the form of the following equation 1.5 where, for the observation i, the variable y_i is specified as the sum of an intercept term α plus, for each of m features, the product of the estimated β_j value and the x_{ij} variable (where $1 \leq j \leq m$). An error term $\mathcal{E}_i$ is there as a reminder that the predictions are not perfect. In this equation each feature has a separate effect on the value of y_i, where y_i changes by the amount β_j for each unit increase in feature x_{ij}. The intercept α is the expected value of y_i when the independent variables are all zero [71].

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_m x_{im} + \mathcal{E}_i \tag{1.5}$$

Due to the existence of many independent variables, obtaining the values of α and all the β_j that results in the minimum squared errors is not the same as for simple linear regression. In this case the first step is rewriting the previous equation as in equation 1.6, where β_0 is α and x_0 is always 1.

$$y_i = \beta_0 x_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_m x_{im} + \mathcal{E}_i \tag{1.6}$$

This same formula can now be written using matrices and vectors, as shown in equation 1.7, where Y is a vector containing the prediction for each observation i, β is a vector containing all the β_j values (plus β_0 or α), X is a matrix containing the values of the independent variables for each feature j for each observation i ($m+1$ columns and n rows), and $\mathcal{E}$ is a vector containing the error for each observation i [71].

$$Y = \beta X + \mathcal{E} \tag{1.7}$$

The goal now is to solve for β, the vector of regression coefficients that minimizes the sum of the squared errors between the predicted and actual Y values. Using matrix algebra we can reach equation 1.8, where $\hat{\beta}$ is the best estimate of the vector β, the T indicates the transpose of matrix X, and the negative exponent indicates the matrix inverse [71].

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{1.8}$$

1.5.2 Logistic regression models

Logistic regression, also called binomial logistic regression, is similar to linear regression, only that instead of predicting a numeric value of any range, it predicts the probability of belonging to one of two categories. This makes it ideal for solving classification problems with just two classes [71]. Logistic regression has been used in biology to predict protein solubility [76], protein function [77] and protein subcellular localization [78], among others.

Logistic regression is used to model a binary variable based on one or more other variables, called predictors. The binary variable being modeled is generally referred to as the response variable, or the dependent variable. For a model to fit the data well, it is assumed that the predictors are uncorrelated with one another, that they are significantly related to the response, and that the observations or data elements of a model are also uncorrelated.

A logistic regression is based on the Bernoulli distribution and can be written as shown in equations 1.9 and 1.10, where y_i is the response variable being modeled and μ_i is the probability that y_i has the value of 1 (in general 1 indicates a success, or that the event of interest has occurred). Equation 1.9 is a sigmoid function which calculates μ_i, which can in turn be used to predict the category y_i as shown in equation 1.10 (see also Figure 1.4). In these equations, y_i only has values of 1 or 0, whereas μ_i has values ranging from 0 to 1 [79].

$$\mu_i = \frac{e^{(\beta_0 x_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_m x_{im})}}{1 + e^{(\beta_0 x_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_m x_{im})}} = \frac{e^{(x_i b)}}{1 + e^{(x_i b)}} = \frac{1}{1 + e^{-(x_i b)}} \tag{1.9}$$

$$y_i = \begin{cases} 0, & \mu_i < 0.5 \\ 1, & \mu_i \geq 0.5 \end{cases} \tag{1.10}$$

In equation 1.9, $x_i b$ is the linear predictor of the logistic model for the observation i, and is specified as the sum of an intercept term α (here already as $\beta_0 x_0$) plus, for each of m features, the product of the estimated β_j value and the x_{ij} variable (where $1 \leq j \leq m$). In this equation each feature has a separate effect on the value of $x_i b$, where $x_i b$ changes by the amount β_j for each unit increase in feature x_{ij}. The intercept $\beta_0 x_0$, where β_0 has the value of the actual intercept (α) and x_0 is always 1, is the expected value of $x_i b$ when the independent variables are all zero. As a clarification, μ_i is very unlikely to be exactly 0.5, and it is usually thought to be either above or below the 0.5 threshold [79].

The previous section introduced linear regression models, where there is a linear relationship between the predicted or fitted values of the model and the terms on the right-hand side of equation (see Equation 1.6). This is not the case for the logistic regression, which makes it harder to estimate the parameters. However, it is possible to establish a linear relationship of the predicted value μ_i and the linear predictor $x_i b$ by rewriting the equation using a "link function", which in this case is $ln(\mu_i/(1-\mu_i))$. A logistic regression can then be written as shown in equation 1.11, where the left member of the equation is the link function mentioned above and the rest is the linear predictor from equation 1.9 [79].

$$ln\left(\frac{\mu_i}{1-\mu_i}\right) = x_i b = \beta_0 x_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_m x_{im} \tag{1.11}$$

Notice that $\mu_i/(1-\mu_i)$ is the probability of success divided by the probability of failure, which is the formula for odds of success. The logarithm of the odds has been called by statisticians the "logit function", from which the term logistic regression derives [79].

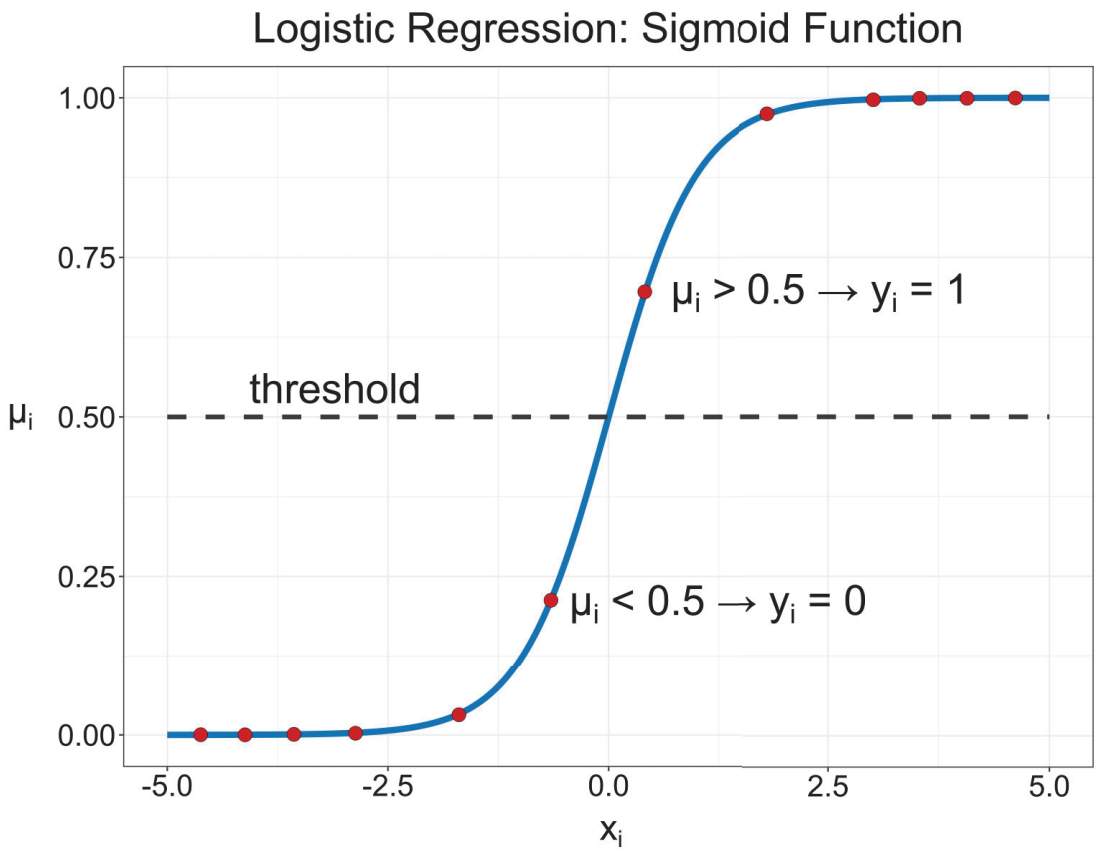


FIGURE 1.4 – *Logistic Regression*. *The figure shows the sigmoid function corresponding to the solved logit function of the logistic regression. In this example there is only one independent variable x_i for each observation i. μ_i refers to the probability of y_i being 1, meaning, the probability of an event occurring, such as a protein being antigenic. Since y_i can only be 1 or 0, we use a threshold (in this case 0.5) to calculate y_i from μ_i.*

The next step is combining data from all n observations to estimate the best β_j possible (the detailed math exceeds the scope of this thesis, but is available in "Practical Guide to Logistic Regression" by Joseph M. Hilbe [79]). The standard method for estimating the parameters of a logistic regression is using "maximum likelihood estimation", or MLE, which uses observed data to calculate the parameters that maximize a likelihood function so that the observed data is most probable [79]. There are many ways to achieve this, but this thesis will focus on the method of "iterative reweighted least squares" (IRLS) [79]. In Chapter 2 of this thesis we utilize IRLS to train a predictor called APRANK, which uses a logistic regression model to infer protein and peptide antigenicity by studying several physical and chemical properties of those proteins and peptides (IRLS is the algorithm used by the *glm* function in the R programming language).

1.5.3 Other prediction models

Decision trees

Decision trees are powerful classifiers that utilize a tree structure to model the relationships among the features and the potential outcomes. This structure mirrors the way a literal tree begins at a wide trunk and splits into narrower and narrower branches, with each branch displaying the possible outcomes of various decisions. A great benefit of decision tree algorithms is that the flowchart-like tree structure is in a human-readable format, providing insight into how and why the model works or fails at a particular task [71]. Decision trees have been used in biology to predict identification of coding regions [80], clinical outcomes [81] and protein-protein interactions [82], among others [83].

Artificial neural networks

Artificial neural networks (ANN) process a set of input signals and return an output signal using a model derived from our understanding of how a biological brain responds to *stimuli* from sensory inputs. Like a brain, it uses a network of interconnected artificial neurons (or nodes) to solve challenging learning problems. However, although they are extremely powerful, their inner workings can be difficult to understand [71]. Artificial neural networks have been used in biology to predict protein secondary structure [84] and MHC antigen presentation [85], among others.

1.5.4 Basics of creating a model

Independently of the prediction model being used, most if not all of them, go through a training phase and an evaluation phase. While there is much to be said about each of these phases, in this thesis we will briefly mention their core concepts.

Training the model

Creating a prediction model usually starts with a large data set. Each observation in this data set has one or more easily measured variables and one harder to measure outcome, which is what the model will try to predict. The goal is then to transform all this data into an abstract form that summarizes how the easily measured variables relate to the hard to measure outcome. This process of fitting a model to a data set is known as "training" [71].

Depending on the prediction model being used, the "best" model can be achieved by using math (such as in Linear regression models) or by using an iterative approach where several parameters are being fitted again and again until some error reaches a local minimum (such as in Artificial neural networks). While trained models do not itself provide new data, they can result in new knowledge, bringing to light important, but previously unseen, patterns and relationships among data [71].

Evaluating the model

Once the model is trained, it is then evaluated on an independent data set in order to judge how well its characterization of the training data generalizes to new, unseen cases. This new data set is usually referred to as "test data set" or "validation data set". It is exceedingly rare for a model to perfectly generalize to every unforeseen case; mistakes when evaluating a model are almost always inevitable. In some prediction models, the information gained in the evaluation phase can then be

used to inform additional training if needed. However, when doing this it is recommended to have yet another truly independent data set to judge the final iteration of the model [71].

A model that performs relatively well during training but relatively poorly during evaluation is said to be overfitted to the training data set because it does not generalize well to the test data set (see Figure 1.5). In practical terms, this means that it has identified a pattern in the data that is not useful for future predictions, meaning, the generalization process has failed. Some possible reasons that can lead to overfitting are: training for too many iterations; measurement errors that lead to having too much noise in the variables; data quality problems, such as corrupted or badly processed data; not having enough data in the training data set; a bias in the training data, meaning, it not being a representative subset of all possible observations; the presence of phenomena that are so complex or so little understood that they impact the data in ways that appear to be random. There are solutions to the problem of overfitting, but most of them are specific to particular machine learning approaches [71].

Because obtaining large amounts of high quality data is not trivial, there are some techniques that allow training and testing with the "same" data. One of the most common is k-fold cross-validation (k-fold CV), which randomly divides the data into k completely separate random partitions called folds. Although k can be set to any number, by far the most common convention is to use 10-fold CV. In this scenario, for each of the 10 folds (each comprising 10 percent of the total data), a machine learning model is built on the remaining 90 percent of data. The fold's 10 percent sample is then used for model evaluation. After the process of training and evaluating the model has occurred 10 times (with 10 different training/testing combinations), the average performance across all folds is reported. An extreme case of k-fold CV is the leave-one-out method, which performs k-fold CV using a fold for each one of the data's examples. This ensures that the greatest amount of data possible is used for training the model [71].

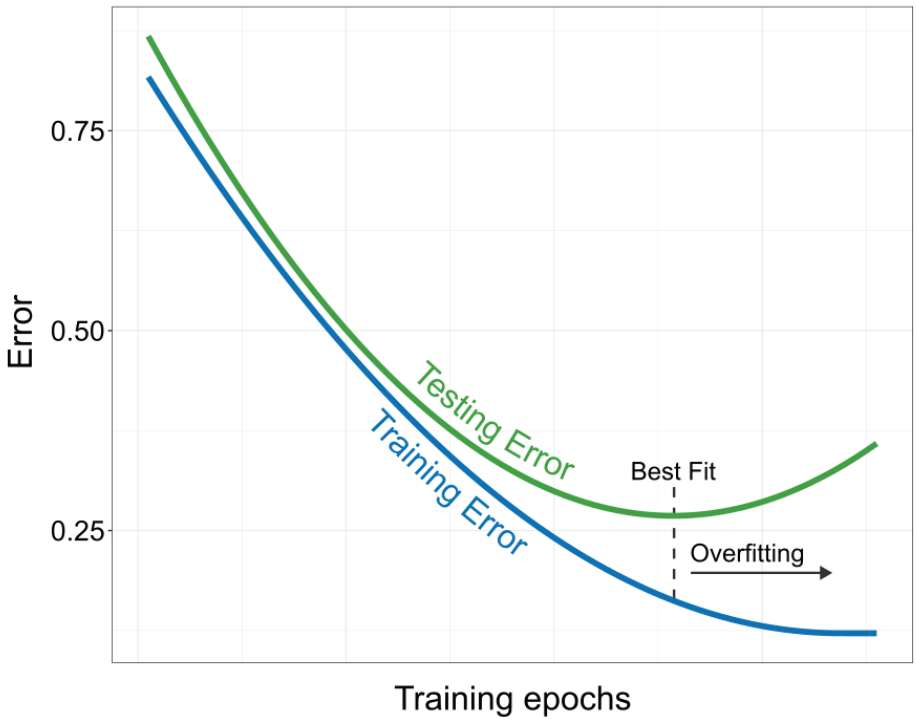


FIGURE 1.5 – ***Example of overfitting***. *The figure shows how the testing and training errors vary across the epochs of the training process. The best fit is achieved when the testing error reaches its minimum value; continuing to train after that point will cause the model to be overfitted to the training data.*

1.6 Chagas disease

Chagas disease, also known as American trypanosomiasis, was discovered in 1909 and is caused by the protozoan parasite *Trypanosoma cruzi*. This parasite is mostly spread to humans and other mammals by the blood-sucking *"kissing bugs"* of the subfamily Triatominae, also known as *"vinchucas"* in Argentina, Bolivia, Chile and Paraguay. Besides vectorial transmission, the disease may also be spread by oral ingestion of the parasite along with tainted foods or drinks, transfusion with infected blood, transplant of infected organs, or vertically from infected mothers to their newborns (congenital infection) [86–89].

This disease is endemic in the Americas and affects between 6 and 8 million people worldwide, with another 65 to 100 million people living in areas with risk for infection [89]. It also affects other animal species, with known infections in up to 150 species of domestic and wild mammals, which act as reservoirs of the parasite [86]. Despite being a disease endemic to the Americas, Chagas disease is also becoming an emerging health problem in many non-endemic areas such as Europe and Japan because of growing population movements [89–91]. While most of the cases of Chagas disease found in non-endemic countries are immigrants from endemic countries, Chagas can still be spread in these non-endemic countries countries through non-vectorial transmission [89].

Due to the difficulties of diagnosing Chagas disease, it is not easy to accurately estimate mortality; however, the most conservative numbers claim that more than 7,000 people die from this disease every year, while others claim that the actual number is closer to 50,000 per year [89, 92, 93]. Notwithstanding this, Chagas disease can have a large impact on a person's life without killing them. This can be measured as "disability-adjusted life years" (DALYs), which refers to the number of years of "healthy" life lost to the disease [94]. In 2013, a study showed that the annual DALYs per individual with chronic Chagas disease was 0.51 (range 0.38–0.60), with a total lifetime DALYs lost per *T. cruzi* infected individual of 3.57 (range 1.18–5.85). In a global scale, the annual burden was estimated to be 806,170 DALYs, with Argentina having the second highest annual burden (after Brazil) with 165,226 DALYs (range 86,008–225,700) [95].

1.6.1 Disease progression

Chagas disease is a heterogeneous condition with a wide variation in clinical course and prognosis. Around 60% to 70% of infected individuals will remain asymptomatic throughout life. Of the other 30% to 40% of infected individuals, some will develop only conduction defects and mild segmental wall motion abnormalities, while others will develop severe symptoms of heart failure, thromboembolic phenomena, and lifethreatening ventricular arrhythmias [88]. Chagas disease evolves through an acute phase and a chronic phase.

Acute phase

Acute Chagas disease occurs with primary infection and typically lasts from 8 to 12 weeks. At this stage, Chagas disease often remains undiagnosed because the majority of patients are asymptomatic or manifest mild and nonspecific symptoms such as fever, malaise, and splenomegaly. In a small percentage of patients, acute infection is marked by inflammation at the site of inoculation, which can be either the pathognomonic "chagoma" (*T. cruzi* skin abscess) or "Romaña sign" (unilateral conjunctivitis and painless swelling of the upper and lower eyelids) [88].

When the acute phase is detected, mild cardiac anomalies such as tachycardia out of proportion to fever are often noted. When more advanced electrocardiographic findings are present, including

right bundle-branch block (RBBB), atrial fibrillation, or ventricular arrhythmias, they signal a worse prognosis. A small proportion of patients present with fulminant acute disease, displaying acute myocarditis, pericardial effusion, meningoencephalitis, or death. These more severe manifestations typically affect immunocompromised individuals and those contracting *T. cruzi* through oral transmission. Regardless of symptoms, the acute phase is marked by microscopically detectable trypomastigotes in the patient's circulating blood [88].

Chronic phase

After 8 to 12 weeks untreated patients enter the chronic phase of *T. cruzi* infection, where parasitemias typically fall below levels detectable by microscopy. Chronically infected patients remain infectious to vectors and can transmit the disease through congenital transmission, blood transfusion, or organ donation [88].

Chronic Chagas disease is subdivided into 4 clinical presentations: indeterminate, digestive, cardiac, or mixed (both digestive and cardiac). After the acute phase of the infection, most patients pass into a chronic indeterminate form defined by positive anti–T cruzi serology, the absence of physical signs or symptoms of disease, a normal ECG, and normal radiographs of the chest, esophagus, and colon [88].

Indeterminate Chagas disease will progress to clinically manifest disease, most commonly dilated cardiomyopathy, at a rate of 1.85% to 7% annually. The progression rate can vary, and the prevalence of the indeterminate form depends on the age of the population. For younger cohorts, over half the patients have the indeterminate form, meaning that their prognosis is good as long as their ECG remains normal (most of them will live up to 10 years before any major complications appears). Among more senior asymptomatic Chagas populations, true indeterminate disease is rare because, given time, most infected individuals will develop characteristic electrocardiographic changes [88].

The gastrointestinal manifestations are less common than Chagas heart disease and are seen mainly in the countries of the Southern Cone (Argentina, Bolivia, Chile, Paraguay, southern Peru, Uruguay, and parts of Brazil). It is hypothesized that geographic specificity results from differences in *T. cruzi* genotypes (TcII, V, and VI in the Southern Cone versus TcI north of the equator), but where regional overlap exists, phenotypic specificity has not been documented. Chagas gastrointestinal disease is the result of enteric nervous system impairment, creating disordered esophageal or colonic motility. Esophageal involvement ranges from mild achalasia to severe megaesophagus, characterized by dysphagia, odynophagia, esophageal reflux, weight loss, aspiration, cough, regurgitation, and increased risk of esophageal carcinoma. The prognosis for patients with digestive forms of the disease is generally good except in those with advanced forms and complications that occasionally cause death [88].

Chagas cardiomyopathy is the most important clinical manifestation of Chagas disease, resulting in the majority of Chagas morbidity and mortality. Most patients that develop Chagas heart disease do so after several decades of the indeterminate form of the disease; however, less than 10% of the patients can progress directly from acute Chagas disease to the chronic cardiac form. Chagas disease distinguishes itself from other cardiomyopathies by having a typical predominant distribution of fibrosis to the posterior and apical regions of the left ventricle and an involvement of the sinus node and electric conduction system. Clinical manifestations of Chagas heart disease result from electric conduction abnormalities, myocardial contractile dysfunction, arrhythmias, or thromboembolism. In most studies, sudden death is the most common overall cause of death (55%–60%), followed by heart failure (25%–30%) and embolic events (10%–15%). The different

stages of Chagas cardiomyopathy as the disease progresses can be seen in Figure 1.6. Predicting which patients will progress to Chagas heart disease is an ongoing challenge and a high-priority area of research [88].

Chagas Disease: Infection With the Parasite *Trypanosoma cruzi*					
Acute Phase	**Chronic Phase**				
Patients infected by *T. cruzi* with findings compatible with acute Chagas disease	**Indeterminate form**	**Chagas cardiomyopathy**			
	A	**B1**	**Chagas dilated cardiomyopathy/heart failure**		
	Patients at risk for developing HF. They have positive serology, neither structural cardiopathy nor HF symptoms. Normal ECG. No digestive changes	Patients with structural cardiopathy, evidenced by electrocardiographic or echocardiographic changes, but with normal global ventricular function and neither current nor previous signs and symptoms of HF	**B2**	**C**	**D**
			Patients with structural cardiopathy characterized by global ventricular dysfunction and neither current nor previous signs and symptoms of HF	Patients with ventricular dysfunction and current or previous symptoms of HF (NYHA FC I, II, III, or IV)	Patients with refractory symptoms of HF at rest despite optimized clinical treatment requiring specialized interventions

Figure 1.6 – ***Progression of Chagas disease into Chagas cardiomyopathy.*** *This table shows the different stages a patient might go through once they are infected with Chagas disease, focusing on the patients that will develop Chagas heart disease. The stages are based on the severity of the disease, according to the "American Heart Association" and "American College of Cardiology" guidelines for the diagnosis and management of heart failure in adults. Arrhythmias and conduction system disease can occur from B1 through D.* ***HF*** *indicates "heart failure", and* ***NYHA FC*** *indicates "New York Heart Association functional class". Extracted from Nunes et al, 2018 [88].*

1.6.2 *Trypanosoma cruzi*

The causative agent of Chagas disease, *Trypanosoma cruzi*, is a unicellular eukaryotic parasite of the order Kinetoplastida, which groups flagellated protists belonging to the phylum Euglenozoa. They are characterized by the presence of an organelle called the kinetoplast (hence these organisms are commonly referred to as "kinetoplastids"). The kinetoplast is an unusual DNA-containing granule located within the single mitochondrion associated with the base of the cell's flagella (the basal body), and it contains the mitochondrial genome of the parasite in the form of a network of concatenated circular DNA molecules (maxicircles and minicircles). Kinetoplastida includes a number of parasites responsible for serious diseases in humans and other animals, as well as various forms found in soil and aquatic environments.

1.6.3 The life cycle of *Trypanosoma cruzi*

The parasite has a complex life cycle illustrated in Figure 1.7. Briefly, an infected triatomine (insect vector) takes a blood meal and releases metacyclic trypomastigotes (an infective, non-replicative parasite form) in its feces near the site of the bite wound. Metacyclic trypomastigotes enter the host through the wound or through intact mucosal membranes, such as the conjunctiva (step 1 in Figure 1.7). Inside the host, the metacyclic trypomastigotes invade cells near the site of inoculation, where they differentiate into intracellular amastigotes (2). Amastigotes multiply by binary fission (3), differentiate into trypomastigotes, and then are released into the circulation as bloodstream trypomastigotes (4). Trypomastigotes infect nucleated cells from a variety of tissues and transform into intracellular amastigotes in new infection sites. Clinical manifestations can result from this infective cycle. The bloodstream trypomastigotes do not replicate. Replication resumes only when the parasites enter another cell or are ingested by an appropriate insect vector. The *"kissing bug"* becomes infected by feeding on human or animal blood that contains circulating parasites (5). The ingested non-replicative trypomastigotes transform into replicating epimastigotes in the vector's midgut (6). The parasites multiply and differentiate in the midgut (7) and differentiate into infective metacyclic trypomastigotes in the hindgut or rectal cell wall (8).

From the point of view of diagnostics, drug discovery and clinical intervention, the mammalian stages (amastigote and trypomastigote) are the relevant stages that take most of the attention of investigators.

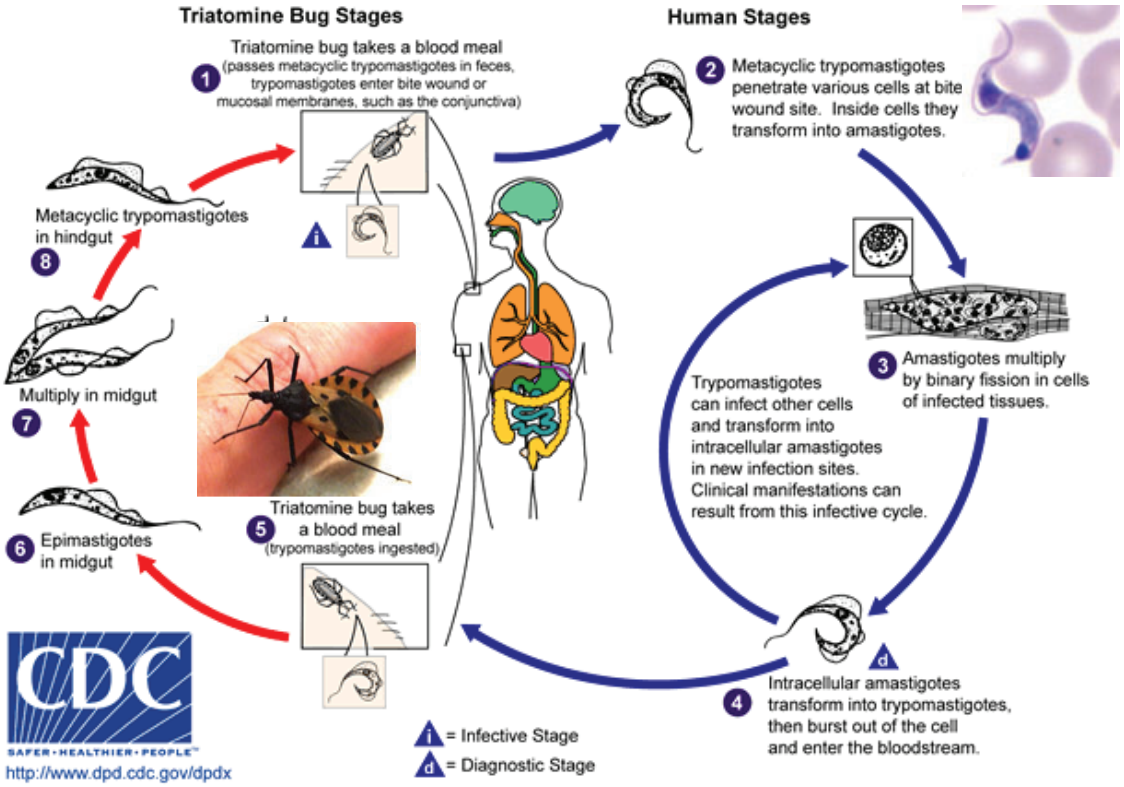


Figure 1.7 – *The life cycle of T. cruzi*

1.6.4 Diagnosis

Many infectious diseases are diagnosed by direct detection of a bacterium, virus, fungus, protozoan, or helminth in a patient with a compatible clinical illness. The methods of detection include cultivation of bacteria and fungi on growth medium, isolation of viruses in cell culture, and identification of the agent microscopically, biochemically, antigenically, or genetically. Visualization of an agent in infected tissue can provide a diagnosis based on specific morphological characteristics or identify the category of organism, for example, gram-positive or gram-negative bacterium or virus. Methods that detect and allow visualization of antigens (immunohistochemistry) or nucleic acid sequences (*in situ* hybridization) provide more specific diagnoses. Detection of specific nucleic acid sequences amplified by polymerase chain reaction is a powerful molecular diagnostic tool [96]. Infectious diseases can also be identified indirectly by detection of a specific immune response, usually antibodies, that develop during the course of illness.

Diagnosing Chagas disease by direct detection is a challenge. The disease is often asymptomatic in its acute phase and evolves into a chronic stage with often mild clinical manifestations [86, 97]. In addition, circulating parasites are difficult to detect during the chronic phase, even by extremely sensitive molecular methods, resulting in the need to rely on indirect serological tests [97–99].

Diagnosis in the acute phase

The natural course of an infection with *T. cruzi* starts with an acute phase, which is estimated to be asymptomatic in more than 95% of the infected individuals. Its manifestations, if present, will arise a few days after the entrance of the parasites. In this phase, parasites multiply exponentially until start of the adaptive immune, which becomes evident 10 to 20 days later. As a consequence, the number of parasites begins to decline and are scarce by the end of the first month. This phase lasts approximately 2 months, and is defined by the relatively easy detection of parasites in the bloodstream, which become more difficult to observe after these first weeks. In this phase parasitological tests, such as parasitological fresh-blood tests, as well as smear and thick drop blood tests are ideal to detect parasites in the infected blood of the patient [97, 100]. If these tests are negative, concentration tests can be carried out. Concentration tests (microhematocrit or Strout test) have a sensitivity of 80-90% and are recommended in the case of patients strongly suspected of having acute Chagas disease and returning negative results for the direct fresh-blood exam [97, 100].

In suspected cases of congenital transmission where the mother is infected with *T. cruzi*, it is important to test for infection of the newborn. This is because current evidence indicates that treatment is highly effective with lower adverse events than those described in adults and cure rates are over 90% in infants treated during the first year of age [101]. Microhaematocrit is the method of choice to identify congenital infection because of its heightened sensitivity and the small amount of blood needed. Microscopic examination of cord blood or peripheral blood of the neonate by this technique is strongly recommended during the first month of life [86, 97]. The infant should then be tested for anti-T cruzi IgG antibodies at 8–12 months of age after the newborn clears maternal antibodies. While this is done to confirm that the infant is indeed infected before starting treatment, this diagnostic method has two major drawbacks: a very high loss-to-treatment risk during pediatric follow-up, and the reduction of drug efficacy if treatment is delayed [97, 102].

Another way of finding parasites in the bloodstream in the acute phase is using molecular amplification of *T. cruzi* DNA, either by conventional polymerase chain reaction (PCR) or by quantitative PCR (qPCR). This method has been shown to be more sensitive and specific than classical parasitological techniques, but there is still no defined standardized technique that would

allow comparable outcomes between laboratories, although there are some being developed [103]. Moreover, this method requires expensive equipment and highly trained personnel which makes it less ideal for point-of-care and rural settings. Therefore, despite its very good performance, molecular detection is not used often beyond regional or national reference laboratories in endemic regions [93, 104].

Diagnosis in the chronic phase

After the acute phase, the infected individual enters the chronic phase, usually in the indeterminate form (no symptoms or signs, no ECG, or X-ray abnormalities). Accurate diagnosis at this stage is important not only for the individual, but also for epidemiological reasons, since they can unknowingly transmit the disease by blood transfusion, organ transplantation and childbirth.

In this phase, circulating parasites are scarce and may be fully absent in the circulation for some periods [97]. However, the appearance of IgG antibodies directed against the antigens of *T. cruzi* makes it possible to diagnose the disease using serological tests, such as ELISA, haemagglutination inhibition assay (HAI) and indirect immunofluorescence (IIF). Specific antibodies remain above detection thresholds for many years, which is advantageous for the serological diagnosis of the infection [93, 104].

The antigens used in these tests can be derived from parasite lysates (which may or may not have been purified), recombinant [105] or purified antigens [106], or other forms of well defined antigens (e.g. synthetic peptides) [60, 105]. While these tests have high levels of sensitivity and specificity on their own, the diagnostic gold standard for reaching a conclusive diagnosis recommended by the World Health Organization (WHO) and the Pan American Health Organization (PAHO) is the use of two independent (e.g. orthogonal) serological tests, and a third test if there are discordant results. Usually these tests are ELISA, HAI and IIF. Two positive results led to a positive diagnosis of infection with *T. cruzi*. However, in the case of ambiguous or discordant results, the tests should be repeated and a third technique should be applied [107].

When diagnosing presence of a parasite using serological tests, it is important to consider the possibility of cross-reactions, meaning, that the serological tests seem to indicate presence of the pathogen being studied, when in fact the individual is infected with another pathogen. This can happen when those two pathogens have shared or cross-reactive antigens, and so they can be detected by individuals infected with either pathogen. For Chagas disease this happens mostly with parasites in the genus *Leishmania*, other kinetoplastid parasites which are responsible for leishmaniasis and are present at the same geographical areas as *T. cruzi*. Cross-reactions can be avoided by excluding from the serological tests any antigens that are shared with similar species present in the same geographic area [97].

Even though parasitemia is low, it is also possible to diagnose infection using indirect parasitological methods such as xenodiagnosis and haemoculture. However, while they are highly specific, they have low sensitivity (20% to 50%), and are labor intensive, making them difficult to apply in clinical settings [100].

1.6.5 Treatment and evaluation of cure

The two main drugs used to treat Chagas disease are benznidazole and nifurtimox. Benznidazole is a 2-nitroimidazole whose main mechanism of action is to generate radical nitro species which can damage the parasite's DNA or cellular machinery. The main side effects observed for benznidazole have been hypersensitivity, bone marrow depletion and peripheral polyneuropathy. These side effects can be controlled with antihistamines, corticosteroids and, in severe cases, suspension of the treatment [108].

As for Nifurtimox, it is a 5-nitrofuran with a similar mechanism of action, involving the production of nitro-anion radicals, which in the presence of oxygen, produce toxic oxygen radicals. The side effects most frequently observed have been anorexia, weight loss, psychological changes, excitability, muscle tremors, somnolence, hallucinations and digestive manifestations such as nausea, vomiting and, occasionally, abdominal pain and diarrhoea. In rare cases, localised convulsions have been observed. These side effects can be controlled with diazepam, cimetidine, metoclopramide, antihistamines and other medications [108].

In the acute phase, treating patients infected with Chagas disease with drugs benznidazole or nifurtimox is capable of preventing a fatal outcome and has a high rate of cure [107, 109]. This also applies for congenital cases and in reactivation due to immunosuppression [89]. The WHO recommends trypanocidal treatment for any patient with acute or congenital *T. cruzi* infection, since they considered that in these scenarios the benefits outweigh the negative aspects [107].

In the chronic phase, however, the consensus on the efficacy of benznidazole and nifurtimox is not clear. While these drugs have shown evidence in some cases of preventing the onset of Chagas disease, delaying the progression of Chagas disease, or reducing the cardiac clinical progression of chronic Chagas disease patients, both drugs have shown low rates of achieving the clinical outcomes in each study while displaying some adverse side effects [89, 93, 107]. Even then, the WHO recommends trypanocidal treatment for children with chronic *T. cruzi* infection and for adults with chronic *T. cruzi* infection and no specific organ damage, since they considered that in these scenarios the benefits outweigh the negative aspects [107].

Another issue when treating Chagas disease is that assessment of cure is not an easy task. Many methods currently being used for evaluating treatment effectiveness, such as serology tests, parasitological tests, detection of lytic antibodies against live trypomastigotes, and PCR based methods [100, 109, 110]. However, all these methods are not sensitive enough or not practical enough to use on daily basis [86, 111]. Hence, there is an urgent need of better quality rapid diagnostic tests to demonstrate within a short time that a cure has been achieved, or that there is a trend towards curing the disease [111].

1.6.6 Serological typification of *Trypanosoma cruzi*

The species *T. cruzi* is genetically diverse and is currently described as comprising six distinct lineages or discrete typing units which can infect humans (DTUs, TcI-TcVI) as well as a seventh DTU found almost exclusively in bats (Tcbat) [112, 113]. These lineages have complex but partially overlapping geographical and ecological distributions and are circumstantially associated with different clinical outcomes [113].

- **TcI** is the lineage that shows the highest genetic heterogeneity and dispersion throughout the Americas, both compatible with a long-term evolution. TcI is ubiquitous in the sylvatic cycle, infecting up to 50 mammalian genera and a major genera of triatomine vectors. Human infection with TcI is prevalent in the north of South America, Central America and Mexico. TcI is associated with Chagas heart disease and there is strong evidence to support that this DTU does not provoke megasyndromes [113].

- **TcII** is as ancient as TcI, but has a more limited geographic distribution. It predominates in the southern and central regions of South America, specially Brazil, and reports in North America are extremely rare. TcII has been isolated mostly from domestic transmission cycles. In those areas where TcII is found, patients show increased events of both Chagas cardiomyopathy and megasyndromes [113].

- **TcIII** is mostly associated with the sylvatic cycle, from northeastern Venezuela to Argentina, with the armadillo (*Dasypus novemcinctus*) as the preferential reservoir. Few human infections have been reported with this lineage in Colombia and in acute cases in Amazonian Brazil [113]. In 2013 there were reports of TcIII infections in domestic dogs in the Argentinean Chaco [113, 114], and in 2015 TcIII was detected in humans for the first time [115].

- **TcIV** is also predominantly associated with the sylvatic cycle, being found in North and South America. Studies of several gene markers indicate that TcIV strains from North and South America are genetically distinct and group separately in phylogenetic analyses. TcIV is the secondary agent of Chagas disease in Venezuela and reported in oral outbreaks in the Brazilian Amazon [113].

- **TcV** and **TcVI** are associated with human Chagas disease in southern countries of South America. However, cases of human infection with TcV and/or TcVI were reported farther north in Ecuador, Colombia, and the Texas in the US [116]. Also, infected vectors and other mammals were reported as north as Mexico and the US [117, 118]. The mammal reservoirs of these DTUs have not been completely defined, although dogs are emerging as potential reservoirs in the Gran Chaco region. Several studies indicate that these two DTUs display minimal inter-lineage diversity and are products of distinct hybridization events, with TcII and TcIII as putative parentals [113].

- **Tcbat** was originally described as a lineage isolated from bats from Central and Southeast Brazil. Similar to other DTUs, Tcbat develops within mammalian cells *in vitro*. However, it lacks virulence and yields low parasitemia in experimentally infected mice. The only evidence of its ability to infect humans is that Tcbat DNA was found in a 5-year-old child in Colombia and in the heart of mummies from Chile [113].

By contributing to the clinical variability of Chagas disease, this large genetic diversity complicates serological diagnosis [119]. Different lineages may express different genes, resulting in proteins with potentially distinct epitopes, with different seroprevalence specially across geographic areas [120].

Trying to study associations between infecting *T. cruzi* lineage and clinical outcome are hampered by the difficulties of isolating *T. cruzi* DNA from the blood of infected patients, which are likely already in the chronic phase of Chagas disease, meaning they have low parasitemia [120]. *T. cruzi* genetic diversity has been partially correlated in *in vitro* systems or animal infection models with clinically relevant phenotypes, such as susceptibility to trypanocidal drugs, tissue distribution, or pathogenesis. In patients, however, these kinds of association remain so far circumstantial and controversial [121].

In patients in the chronic phase of Chagas disease, it is possible to use serological tests to detect antibodies that are produced in response to lineage-specific antigens; however, these kind of antigens are scarce. In 2002, the trypomastigote small surface antigen (TSSA) was described. This protein, encoded by a member of the TcMUCIII mucin gene family, is expressed on the mammalian bloodstream trypomastigote stage of the *T. cruzi* life cycle [122]. TSSA displays small differences across DTUs, which enabled it to work as a marker to differentiate between TcI and the rest of the DTUs, which at the time were all encompassed in TcII [113]. TSSA genetic diversity has allowed serology-based lineage discrimination and many epidemiological studies of Chagas disease [123, 124]. More recent research found that four major TSSA serotypes could be defined: TSSAI (TcI), TSSAII (TcII), TSSAIII (TcIII), and TSSAIV (TcIV). Being hybrids, TcV/TcVI genomes code for both TSSAII and TSSAIII isoforms [121].

1.6.7 Genomics of *Trypanosoma cruzi*

The first whole-genome sequencing of the protozoan pathogen Trypanosoma cruzi was achieved in 2005 using *T. cruzi* strain CL-Brener (TcVI), which was chosen due to being well characterized experimentally at the time. The sequencing revealed that the diploid genome contains a predicted 22,570 proteins encoded by genes, of which 12,570 represent allelic pairs [125]. They also observed that over 50% of the T. cruzi genome is composed of repetitive sequences, which include numerous families of surface proteins (e.g. trans-sialidases, mucins and mucin-associated surface proteins) with hundreds to thousands of members each, as well as substantial numbers of transposable elements, microsatellites and simple tandem repeats. This repetitive nature greatly hampered the assembly of this genome, resulting in a highly fragmented and draft assembly with extensively collapsed high repeat regions [125, 126]. Despite this degree of fragmentation, this draft genome was highly valuable because it led to the identification of several novel species-specific multigene families and gave, for the first time, a draft overview of the genome architecture of T. cruzi. [127]

In the following years some other strains of *T. cruzi* were sequenced, such as Dm28c (TcI) [128], Sylvio X10/1 (TcI) [129], and the subspecies *T. c. marinkellei* [130]. However, they also showed high fragmentation, which for some was even higher than that originally reported for CL-Brener [127].

In 2009, Weatherly et al [131]. tackled the problem of assembly fragmentation by using bacterial artificial chromosome (BAC) library sequencing and previous knowledge of two genomes from closely related species, *Trypanosoma brucei* and *Leishmania*. Using the CL-Brener strain, they managed to scaffold many *T. cruzi* contigs and assembled 41 pairs of chromosomes, a number in agreement with the predicted number of *T. cruzi* chromosomes based upon pulse field gel analysis. These genome had 90% (21,133 of 23,216) of *T. cruzi* known genes annotated in specific positions [131]. While this version of the genome is a huge improvement from the previous one, a large number of gaps were still present in the chromosomes, and many unassigned contigs remained, making it impossible to determine the exact genome content and, in particular, the full repertoires of large gene families [126, 131].

In the recent years new sequencing techniques known as "third-generation sequencing" or "long-read sequencing" have been developed, making it possible to sequence reads of more than 15 kb in average length, some of which are much longer. One of such techniques, called Single Molecular Real-Time (or SMRT), was used in 2018 to sequence the full genome of two clinically and evolutionarily relevant *T. cruzi* strains: TCC, an hybrid strain (TcVI) closely related to CL-Brener, and Dm28c, a non-hybrid strain (TcI). This novel technique enabled an accurate estimation of gene copy numbers, abundance and distribution of repetitive sequences (including satellites and retroelements) [127].

The genes in *T. cruzi* are usually divided in two groups: the multigene families with hundreds of copies (DGF-1, GP63, MASP, mucins, RHS and TS), and those generically defined as "conserved", which include genes encoding proteins with a known function, or "conserved genes", and genes without an assigned function but present in more than one trypanosomatid species, or "hypothetical conserved genes". This analysis using third-generation sequencing revealed that the genome of *T. cruzi* is compartmentalized in two clearly defined regions: a "core compartment" composed of conserved and hypothetical conserved genes, and a "disruptive compartment" composed of the multigene families TS, MASP and mucins. The GP63, DGF-1 and RHS multigene families have a dispersed distribution in the genome, being present in both compartments. The core compartment with its conserved genes has blocks previously described in *T. brucei* and *L. major*. This is not the case for the disruptive compartment, which is mainly composed of species- or genus-specific genes, suggesting that is a recent region of the genome. The disruptive compartment also exhibits higher GC content than the core compartment [127].

1.7 The Leishmaniases

Besides *Trypanosoma cruzi*, in this thesis we will also mention kinetoplastids of the genus *Leishmania*, such as *Leishmania braziliensis* which is the causative agent of the American tegumentary leishmaniasis (ATL). Because of the existing overlaps in the ecoepidemiology of *L. braziliensis* and *T. cruzi* in South America and the reported cross-reactivity of some antigens, it is relevant to introduce here these related pathogens and diseases [132–134].

There are more than 21 species of *Leishmania* that cause various diseases ranging from self-healing tegumentary leishmaniasis (TL) to debilitating and lethal (if untreated) visceral leishmaniasis (VL; also known as kala-azar) [134]. Tegumentary leishmaniasis (TL) affects the skin and mucous membranes and may present as different clinical forms: cutaneous, mucosal, disseminated, and diffuse. Cutaneous leishmaniasis (CL) is the most common presentation of the disease, and mucosal leishmaniasis (ML) occurs in approximately 3–5% of cases of CL [135]. Classic ML occurs secondary to cutaneous lesions. In a minority of cases, however, the ML is primary, without prior or concomitant history of skin lesions.

The World Health Organization (WHO) estimates that there are 900,000 to 1.3 million new cases of leishmaniasis annually throughout the world, with approximately 200,000–400,000 of the visceral form and 700,000 to 1.2 million of the tegumentary form. American tegumentary leishmaniasis (ATL) is caused by the protozoan parasite *Leishmania spp*, and is widespread from the south of the United States to the north of Argentina [136]. At least seven species of *Leishmania* (Trypanosomatidae) are responsible for human ATL in Brazil with the main agents being, by order of prevalence, *Leishmania braziliensis*, *Leishmania amazonensis* and *Leishmania guyanensis* [137].

Leishmania parasites invade mammalian macrophages by receptor-mediated endocytosis, and multiply in the low-pH, amino acid–rich endolysosomes, to which their metabolism and nutrition are adapted [134]. Patients with TL and CL, in whom the pathogen is largely restricted to the skin and is not found in lymphoid tissues, generally develop a curative immune response.

The diagnosis of CL and ML usually employs microscopic detection of organisms in the lesions and skin tests, which consist of intradermal injection of whole or lysed promastigote forms of the pathogen followed by measuring local induration at the injection site 48–72 hours later [134]. However, microscopic detection often shows low sensitivity and requires highly trained personnel; and parasitological skin tests are laborious and suffer from lack of a standardized source of leishmanin antigens [138]. Alternatively, molecular (PCR) [139] and serological methods (ELISA) are used [140]. However they lack specificity, mostly linked to cross-reactivity to *T. cruzi* [140].

2. APRANK: Computational prioritization of epitopes

This chapter describes the development of a computational method called APRANK (Antigenic Protein and Peptide Ranker) which integrates multiple molecular features to prioritize potentially antigenic proteins and peptides in a given pathogen proteome. These features include subcellular localization, presence of repetitive motifs, natively disordered regions, secondary structure, transmembrane spans and predicted interaction with the immune system. We trained and tested this method with several pathogenic bacteria and protozoa, we evaluated this integrative method using non-parametric ROC-curves and leave-one-out cross-validation, and made an unbiased validation using an independent data set.

This chapter is based on the paper ***"APRANK: computational prioritization of antigenic proteins and peptides from complete pathogen proteomes"*** by Ricci AD. *et al.*, published in 2021 [141]. Here, the relevant information from the paper in its final form (accepted manuscript) is transcribed with some minor modifications and updates.

Ricci AD, Brunner M, Ramoa D, Carmona SJ, Nielsen M, Agüero F.
APRANK: Computational Prioritization of Antigenic Proteins
and Peptides From Complete Pathogen Proteomes (2021).
Frontiers in Immunology 12:702552.
doi: 10.3389/fimmu.2021.702552.
PMID: 34335615; PMCID: PMC8320365.

2.1 Introduction

Infectious diseases are one of the first causes of death worldwide, disproportionately affecting poor and young people in developing countries. Several epidemiological and medical strategies exist to deal with these diseases, most of which rely on robust and accurate diagnostic tests. These tests are used to demonstrate infection (presence of the pathogen), to follow up treatments and to monitor the evolution or cure of the disease or the success of field [142].

One of the preferred methods to diagnose infections relies on the detection of pathogen-specific antibodies in the fluids of infected patients (most often serum obtained from blood) [143, 144]. For this reason, there is a big interest in developing reliable methods able to improve the fast and sensitive identification of potential specific antigens.

With the advent of peptide microarray platforms it is now possible to perform high-throughput serological screening of short peptides, which allows for faster discovery of linear antigenic determinants with good potential for diagnostic applications [54]. Taking advantage of complete genome sequences from pathogens, it is theoretically possible to scan every encoded protein with short peptides against sera from infected hosts. However, while this is straightforwardly achieved for viral pathogens and small bacteria, it gets more difficult when dealing with larger bacteria or eukaryotic parasites, since they can reach thousands of proteins with millions of peptides, exceeding the average capacity of standard protein or peptide microarrays [145]. Besides, it is now becoming common to fit in the arrays additional sequence variants obtained from the pathogen population (from diverse strains and clinical isolates). One example are serological strain typing strategies [121], which would stress the capacity of these platforms.

Ultrahigh-density peptide microarrays had been used successfully to map linear epitopes, having an upper theoretical limit of 2 to 3 million unique peptides per array [146]. While these ultrahigh-density peptide microarrays do enable a lot of possibilities, they do not yet have the capacity to analyze whole proteomes of larger pathogens without some preprocessing. It is also worth noting that they are not widely available as lower density arrays and they require substantial processing and downstream work to deal with large proteomes [60, 147, 148].

There are several ways to deal with the problem of not having enough space when accommodating large proteomes in a peptide array, each with their own advantages and disadvantages. Some commonly used methods are: decreasing the overlap between peptides, dividing the proteome among different microarray slides, and using computational methods to prioritize antigens. Here we will focus on the latter method, where we and others have previously shown that a number of protein features can be used to validate and prioritize candidate antigens and epitopes for human pathogens [147, 149–151]. Similar approaches have also been developed into a number of reverse vaccinology programs for bacteria [152].

In previous work from our laboratory, we developed a method that integrates information from a number of calculated molecular and structural features to compute an antigenicity score for proteins and peptides in *Trypanosoma cruzi* [147, 149]. Here, we use machine learning techniques to extend and generalize this concept so that it can be applied to other pathogens. We call this method APRANK (Antigenic Protein and Peptide Ranker) and show how it can be used as a strategy to predict and prioritize diagnostic antigens for several human pathogens.

2.2 Results

2.2.1 Species and Antigenicity

We selected human pathogens from a phylogenetically diverse set of taxa with experimentally validated antigen and/or epitope data to train and test our method. This included gram negative bacteria, gram positive bacteria and eukaryotic protozoans. We did not include viruses in this version of APRANK because they have small proteomes that are already amenable to experimental experimentation (e.g. their full-proteomes fit on stantard low-density protein or peptide microarrays). The species selected to train APRANK and the diseases they cause are shown in Table 2.1.

We obtained the proteomes of these species and split each protein into peptides of 15 residues. Once this was done, we used information from the immune epitope database (IEDB) along with manually extracted information from several papers to tag each protein and peptide as antigenic or non-antigenic. The "non-antigenic" tag in this paper should be understood in the sense of proteins with no prior information on their antigenicity.

It is common practice in the literature to report antigenicity for a single or a few reference proteins or accession numbers, information which is then passed on to databases such as IEDB [153, 154]. Nevertheless, when dealing with complete proteomes, there are usually other paralogs with high sequence similarity to those labeled as antigenic, which are likely to have similar properties and also likely to be antigenic themselves. To improve the learning process of APRANK, and to account for unlabeled proteins, we calculated sequence similarity for all proteins in the 15 analyzed proteomes using blastp from the NCBI BLAST suite [155]. This process gave us clusters of similar proteins, which we then used to spread the antigenicity from proteins labeled as antigenic to similar proteins without that label. The amount of total and antigenic proteins, before and after using BLAST to find similar proteins inside each species, can be seen in Table 2.2.

2.2.2 Protein features and Predictors

To develop a tool that can help identify candidate antigenic proteins and peptides, we used several predictors that focused on different properties of the proteins. On a broad sense, these predictors assess: the antigenicity and/or immunogenicity of proteins [156, 157]; the structural and post-translational features that can be predicted from the protein sequence, some of which may suggest the protein enters the secretory route or is anchored at the membrane [158–160]; the presence of internal tandem repeats in proteins, which have been described to modulate immunogenicity of proteins [161]; and other structural features such as the presence of intrinsically unstructured or exposed regions in proteins, which may effect their presentation in the context of an immune response [160, 162–164]. The detailed list of all predictors can be seen in Table 2.3.

APRANK also uses a couple of custom scripts that measure sequence similarity between each pathogen protein and the human host (CrossReactivity), or itself (SelfSimilarity). The idea behind these measurements was to obtain additional information on highly conserved sequences that may result in e.g. potential lack of immune response (tolerance) if the pathogen sequence is highly similar to a human protein; or cross-reactivity of antigens and epitopes in other proteins from the same pathogen (self-similarity). These predictors provide information on desirable and undesirable properties that then need to be weighted accordingly to achieve good performance at the task of antigen and epitope prediction.

As mentioned before, the predictors were chosen to analyze several properties of the proteins; however, it was expected that some of the predictors would be better than others when predicting antigenicity. To assess this, we analyzed the distribution of the outputs for each of the predictors and used Student's t-test to compare the means between the antigens and the rest of the proteome (see Table 2.4). BepiPred, NetMHCIIpan, NetSurfp and SignalP showed the best results, separating the means of antigens and non-antigens in over 9 of the 15 organisms. Looking at the table, there was an argument to be made for removing some predictors from the model, but we decided to keep them for now and let the model likely assign lower coefficients to them. We talk more about these predictors in this chapter's Discussion.

Pathogen Species	Disease	Group	Taxonomy (Phylum)
Borrelia burgdorferi	Lyme disease	Gram Negative Bacteria	Spirochaetia
Brucella melitensis	Brucellosis		Alpha-proteobacteria
Coxiella burnetii	Q fever		Gamma-proteobacteria
Escherichia coli	Gastroenteritis		Gamma-proteobacteria
Francisella tularensis	Tularemia		Gamma-proteobacteria
Leptospira interrogans	Leptospirosis		Spirochaetia
Porphyromonas gingivalis	Periodontal disease		Bacteroidetes
Mycobacterium leprae	Leprosy	Gram Positive Bacteria	Actinobacteria
Mycobacterium tuberculosis	Tuberculosis		Actinobacteria
Staphylococcus aureus	Bacteremia		Firmicutes
Streptococcus pyogenes	GAS infections		Firmicutes
Leishmania braziliensis	Leishmaniasis	Eukaryotic Protozoa	Euglenozoa
Plasmodium falciparum	Malaria		Apicomplexa
Toxoplasma gondii	Toxoplasmosis		Apicomplexa
Trypanosoma cruzi	Chagas Disease		Euglenozoa

Table 2.1 – ***List of pathogen species used to train APRANK.***

Species	Group	Proteins			Peptides		
		Total	Antigenic		Total	Antigenic	
			Original	After BLAST		Original	After kmer expansion
B. burgdorferi	Gram -	1,390	137	152	386,683	117	863
B. melitensis	Gram -	3,178	13	13	-	-	-
C. burnetii	Gram -	1,853	102	104	-	-	-
E. coli	Gram -	4,778	7	7	1,428,744	9	158
F. tularensis	Gram -	1,556	27	27	-	-	-
L. interrogans	Gram -	3,683	10	10	1,113,309	19	342
P. gingivalis	Gram -	1,881	10	11	626,536	165	1181
M. leprae	Gram +	1,605	7	8	515,942	76	633
M. tuberculosis	Gram +	3,940	81	89	1,268,272	416	4,369
S. aureus	Gram +	2,607	16	16	758,970	55	575
S. pyogenes	Gram +	1,690	13	13	491,619	263	985
L. braziliensis	Eukaryote	8,084	8	12	4,964,396	14	182
P. falciparum	Eukaryote	5,337	106	131	4,009,580	562	9,120
T. gondii	Eukaryote	8,322	15	16	6,535,220	94	457
T. cruzi	Eukaryote	21,170	242	2,480	10,408,841	4,025	7,317

Table 2.2 – ***Amount of antigenic proteins and peptides for each species.*** *This table shows the amount of antigenic proteins and sequences extracted from bibliography and the final amount after processing. For proteins, BLAST was used to also tag as antigenic other proteins of the same species that were similar to the antigenic ones. For peptides, a custom mapping method named "kmer expansion" was used to tag peptides as antigenic based on the antigenic sequences in bibliography (see Methods). We did not have information at peptide level for three of the species.*

Focus	Feature	Predictor	Basis
Stimulation of an immune response	B-cell epitopes	BepiPred 1.0	Antigenicity by HMM
	Binding to MHC Class II molecules	NetMHCIIpan 2.0	ANN trained with peptide and MHC Class II sequence information
Peculiarities in the protein sequence	Glycosylation sites	NetOglyc 3.1d	ANN trained with mucin type GalNAc O-glycosylation sites in mammalian proteins
	GPI-anchored proteins	PredGPI 1.4.3	Discrimination of the anchoring signal by SVM and prediction of the most probable omega-site by HMM
	Signal peptide cleavage sites	SignalP 4.0	Prediction of cleavage sites and a signal peptide/non-signal peptide prediction based on a combination of several ANN
	Tandem repeats	Xstream 1.71	SE algorithm to explicitly locate exact and degenerate tandem repeats TRs of all periods in protein sequences
Three dimensional structure	Disorder	Iupred 1.0	Aminoacids favorable interactions potential
	Parallel coiled coil fold	Paircoil2	Uses pairwise residue probabilities with the Paircoil algorithm and an updated coiled coil database
	Secondary Structure	NetSurfp 1.0	ANN trained with sequence profiles and predicted secondary structure
	Surface access	NetSurfp 1.0	ANN trained to predict the relative surface exposure of the individual amino acid residues
	Transmembrane helices in proteins	TMHMM 2.0c	Membrane protein topology prediction method based on a HMM
Molecular properties	Isoelectric point	Pepstats (EMBOSS 6.6.0.0)	Amino acids pK values
	Molecular Weight	Pepstats (EMBOSS 6.6.0.0)	Amino acids weights
Similarities within itself and with the host	Sequence similarity (pathogen / host)	CrossReactivity	Shared kmers between pathogen and host proteins
	Sequence similarity (pathogen proteins)	SelfSimilarity	Shared kmers between pathogen proteins

Table 2.3 – ***Predictors used to analyze different features of proteins and peptides.*** *CrossReactivity and SelfSimilarity are custom Perl scripts. Acronyms used: ANN (Artificial Neural Network), HMM (Hidden Markov Model), SE (Seed Extension), SVM (Support Vector Machine).*

Predictor	Antigenicity discerning capability			
	All (15)	Gram - (7)	Gram + (4)	Eukaryote (4)
BepiPred	12 (+12 -0)	5 (+5 -0)	4 (+4 -0)	3 (+3 -0)
Isoelectric Point	-7 (+0 -7)	-2 (+0 -2)	-2 (+0 -2)	-3 (+0 -3)
Molecular Weight	0 (+4 -4)	0 (+1 -1)	2 (+2 -0)	-2 (+1 -3)
Iupred	4 (+5 -1)	1 (+1 -0)	2 (+2 -0)	1 (+2 -1)
NetMHCIIpan	-13 (+0 -13)	-5 (+0 -5)	-4 (+0 -4)	-4 (+0 -4)
NetOglyc	3 (+4 -1)	2 (+2 -0)	-1 (+0 -1)	2 (+2 -0)
NetSurfp (RSA)	9 (+9 -0)	5 (+5 -0)	2 (+2 -0)	2 (+2 -0)
NetSurfp (Alpha Helix)	-1 (+2 -3)	0 (+1 -1)	0 (+0 -0)	-1 (+1 -2)
NetSurfp (Beta Strand)	2 (+3 -1)	1 (+1 -0)	0 (+0 -0)	1 (+2 -1)
Paircoil2	2 (+2 -0)	0 (+0 -0)	1 (+1 -0)	1 (+1 -0)
PredGPI	-5 (+2 -7)	-4 (+0 -4)	-3 (+0 -3)	2 (+2 -0)
SignalP	10 (+10 -0)	5 (+5 -0)	3 (+3 -0)	2 (+2 -0)
TMHMM	1 (+3 -2)	1 (+2 -1)	0 (+0 -0)	0 (+1 -1)
Xstream	5 (+6 -1)	0 (+1 -1)	2 (+2 -0)	3 (+3 -0)
Cross Reactivity	2 (+2 -0)	1 (+1 -0)	1 (+1 -0)	0 (+0 -0)
Self Similarity	-3 (+0 -3)	-2 (+0 -2)	0 (+0 -0)	-1 (+0 -1)

Table 2.4 – ***Antigenicity discerning capabilities of predictors.*** *The table shows the net number of species where the mean of the normalized outputs for a given predictor are significantly different for the validated antigens than for the whole proteome (Student's t-test, $p < 0.05$). For each organism, if the mean was significantly greater for the antigenic proteins than for the whole proteome then we added 1, and if the mean was significantly less then we subtracted 1 (this is shown inside the parenthesis). The numbers in the headers of each column correspond to the total amount of analyzed species inside that group.*

2.2.3 Testing APRANK and ROSE on species-specific models

Species-specific models were created to test the method and to compare between balanced and unbalanced training sets, where by "balanced" we meant training sets that had the same amount of antigens and non-antigens. Due to our limited data, the balanced training sets were created using the R package *ROSE*, which works by generating artificial balanced samples from the existing classes according to a smoothed bootstrap approach [165]. As the name implies, the species-specific models worked with only one species at a time, using a fraction of its proteins or peptides to predict antigenicity for the rest. A schematic visualization of this procedure is shown in Figure 2.1.

The first step was running the predictors for all proteins in the selected genome, after which their outputs were parsed, processed and normalized to have them all in a common scale. This new data needed to be divided into training and test sets. Often, training sets represent ∼80% of the data; however, in our case some species had a low number of validated antigens (see Table 2.2), which meant that choosing a 80/20 training/test set split would result in test sets having only a few antigenic proteins. This kind of imbalance tends to compromise the training process, making the model to focus on the prevalent class (non-antigenic) and ignore the rare class (antigenic) [166]. For this reason, when training a model using data from a single species, we chose to split the training and test set 50/50, re-sampling proteins and peptides multiple times. We used the similarity-based clustering of sequences to avoid placing highly similar sequences into both training and test sets, which may had resulted in overfitting.

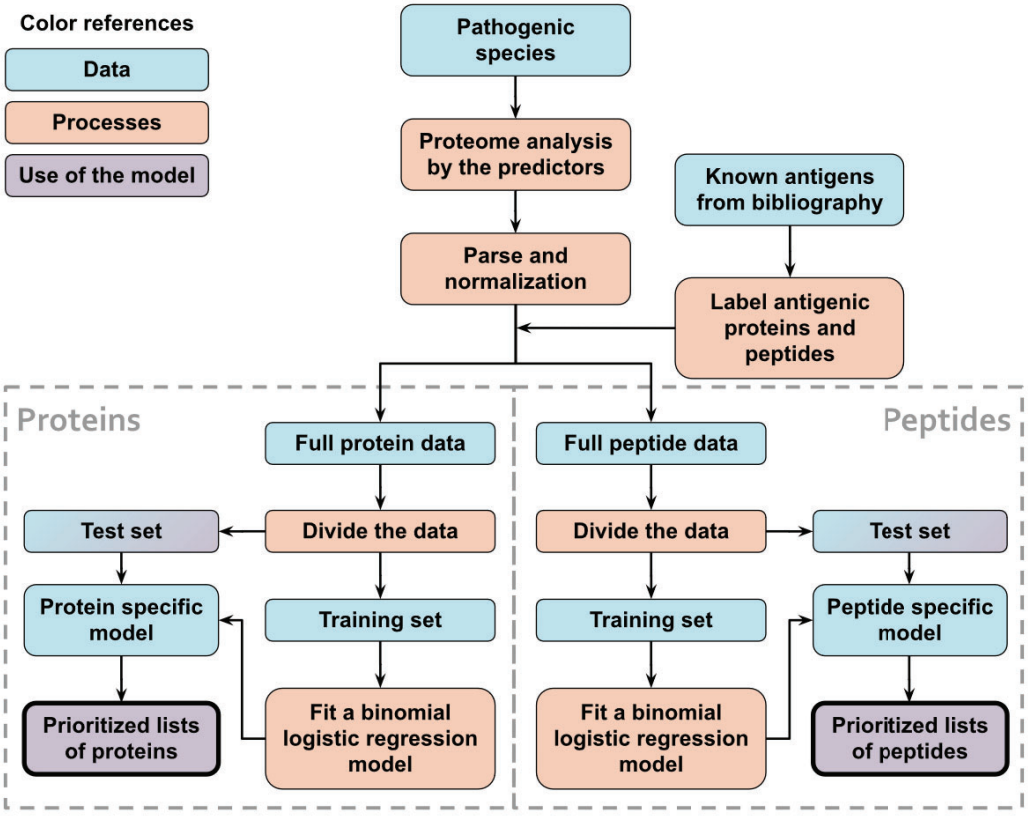


Figure 2.1 – ***Schematic flowchart used to obtain APRANK's species-specific models.*** *With the aim of testing and tuning our method, training and prioritization was performed for both proteins and peptides using data from a single proteome of interest. This process was repeated for all of our 15 species.*

We then used *ROSE* to balance our training sets and used these unbalanced and balanced training sets to fit binomial logistic regression models. We chose this prediction model for two reasons: first, because the coefficients of the model gave us information on how the different predictors affected the model, and second, because it was a model more resilient to the existence of false negatives than other more complex models (and there were false negatives because these were the novel antigens we wanted to find).

For most species this resulted in four models: a balanced protein model, a balanced peptide model, an unbalanced protein model, and an unbalanced peptide model. These models, which we denominated *species-specific models*, were then used to predict the antigenicity of their respective test sets. The performance of APRANK was assessed by measuring the area under the ROC curve (AUC), using known antigens and epitopes in the protein and peptide test sets. This whole process was repeated 50 times, re-sampling which proteins were in the training set and which in the test set. Each iteration resulted in an AUC score, and then the final APRANK AUC score for that species was calculated as the mean of all those AUC scores (see Figure 2.2).

These calculations were done for each of the 15 species, although for 3 of them there was no antigenicity information at the peptide level, and only protein models were calculated. The results are presented in Table 2.5. Our testing showed that APRANK was able to predict antigenicity for proteins and peptides in most cases, with good performance. The only species that did not have a successful prediction were *E. coli* for the protein model, and *M. tuberculosis* and *S. aureus* for the peptide model. In these cases, the final AUC corresponding to the species-specific model was not significantly different than a random prediction. As for the balancing of the data using *ROSE*, it seemed to have mostly positive or neutral effects in the predicting capabilities of our models, which meant we could safely use it in training our pan-species models.

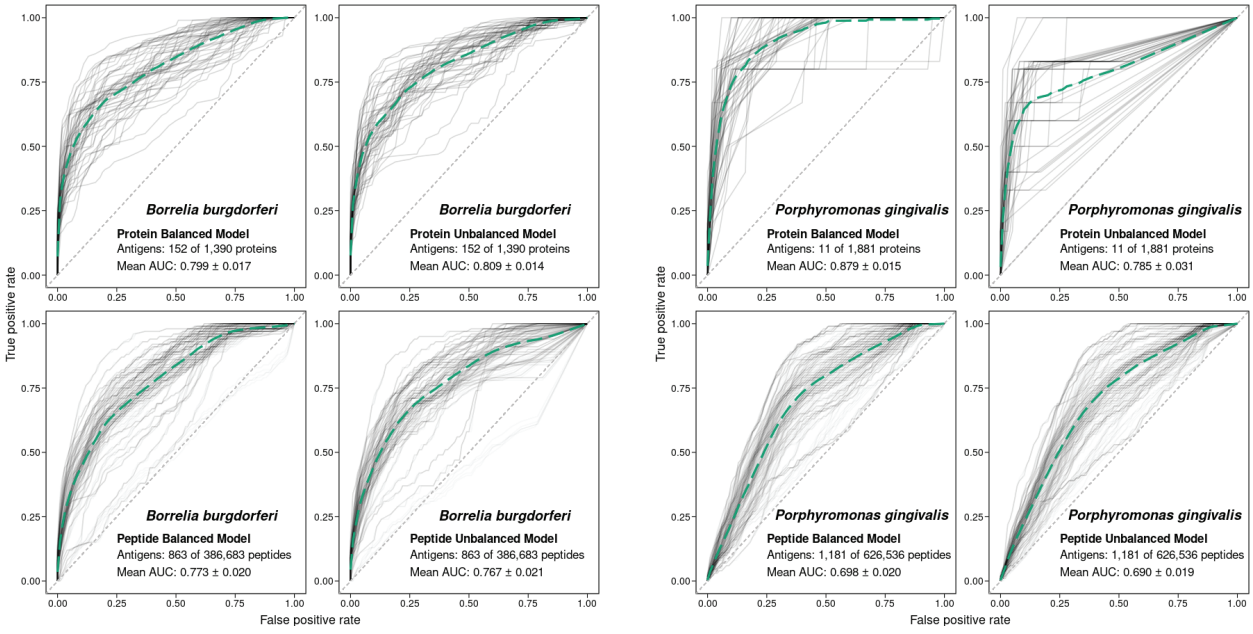


Figure 2.2 – ***Performance of APRANK training using balanced or unbalanced data.*** *Performance of APRANK's species-specific models for B. burgdorferi and P. gingivalis. ROC curves for each iteration of training and testing are shown in light gray, and the average curves are shown in green (dashed lines).*

Species	Proteins			Peptides		
	BTR	Trained with unbalanced data	Trained with balanced data	BTR	Trained with unbalanced data	Trained with balanced data
		Mean AUC	Mean AUC		Mean AUC	Mean AUC
B. burgdorferi	Yes	0.809 ± 0.014	0.799 ± 0.017	Yes	0.767 ± 0.021	0.773 ± 0.020
B. melitensis	Yes	0.710 ± 0.037	0.700 ± 0.033	-	-	-
C. burnetii	Yes	0.611 ± 0.011	0.620 ± 0.010	-	-	-
E. coli	**No**	0.511 ± 0.034	0.515 ± 0.039	Yes	0.584 ± 0.056	0.633 ± 0.047
F. tularensis	Yes	0.783 ± 0.018	**0.807 ± 0.014***	-	-	-
L. interrogans	Yes	0.827 ± 0.033	0.867 ± 0.023	Yes	0.559 ± 0.015	0.565 ± 0.011
P. gingivalis	Yes	0.785 ± 0.031	**0.879 ± 0.015*****	Yes	0.690 ± 0.019	0.698 ± 0.020
M. leprae	Yes	0.633 ± 0.018	0.652 ± 0.018	Yes	0.557 ± 0.029	0.585 ± 0.023
M. tuberculosis	Yes	0.635 ± 0.010	0.647 ± 0.011	**No**	0.508 ± 0.010	0.502 ± 0.010
S. aureus	Yes	0.765 ± 0.032	0.772 ± 0.023	**No**	0.438 ± 0.054	0.420 ± 0.057
S. pyogenes	Yes	0.884 ± 0.039	**0.984 ± 0.003*****	Yes	0.832 ± 0.021	0.844 ± 0.019
L. braziliensis	Yes	**0.719 ± 0.021****	0.673 ± 0.020	Yes	0.778 ± 0.029	**0.867 ± 0.025*****
P. falciparum	Yes	0.821 ± 0.009	0.826 ± 0.007	Yes	0.758 ± 0.016	**0.779 ± 0.012***
T. gondii	Yes	0.656 ± 0.032	**0.744 ± 0.032*****	Yes	**0.646 ± 0.035****	0.584 ± 0.020
T. cruzi	Yes	0.803 ± 0.029	**0.850 ± 0.022***	Yes	0.838 ± 0.019	0.854 ± 0.016

Table 2.5 – ***Prediction results for the specific models.*** *The prediction was considered to be successful if it was significantly Better Than a Random set of scores (BTR). Each specific model was calculated 50 times using different, but overlapping, subsets of data as training and test sets. In bold we show the model with the significantly higher AUC when comparing training with unbalanced or balanced data (Student's t-test, * < 0.05, ** < 0.01, *** < 0.001).*

2.2.4 Development of APRANK as a pan-species ranker of antigens

In the previous section we used protein and peptide data from a given pathogen species to train models that successfully predicted antigenicity for that same organism; however, our end goal was to have models that were able to predict protein and peptide antigenicity for any pathogen. To achieve this, we created models trained with all species, which we called *protein generic models* and *peptide generic models*. A schematic visualization of this procedure is shown in Figure 2.3.

For these models, we used *ROSE* [165] to generate similar sized partitions of balanced data for each of the species, and then we merged these data and fitted two binomial logistic regression models, one for proteins and one for peptides. When using the models to predict the peptide antigenicity scores, we also analyzed the predicting capabilities of what we called the *combined score*, which was a combination of the protein and peptide scores for a given peptide.

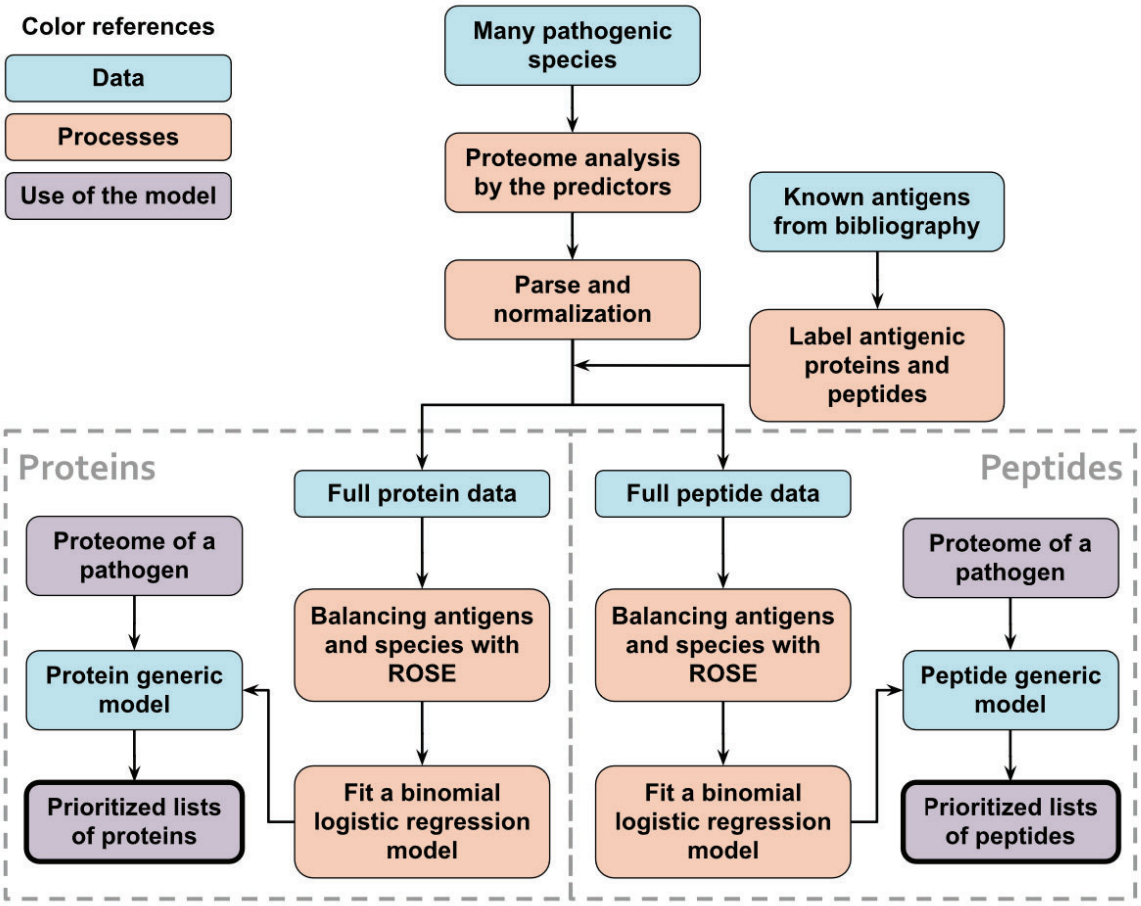


Figure 2.3 – ***Schematic flowchart used to obtain APRANK's generic models.*** *With the aim of creating a set of models that could make predictions for a wide range of species, training and prioritization was performed for both proteins and peptides using combined data from all of our 15 species. When testing the generic models, leave-one-out models were used, where 14 species were used to train the models and the 15th species to test them. This process was repeated for all of our 15 species.*

To validate these models we performed a leave-one-out cross-validation method (LOOCV), hence creating 15 different protein generic models, each time leaving out one species (which was the one being used as test set). For the peptide generic models we followed a similar route, but we ended up with 12 models due to the lack of antigenicity information at peptide level for 3 of the 15 species.

The results for the cross-validation are presented in Table 2.6. The generic protein models were successful in predicting antigenicity for all species, and similar results were obtained also at the peptide level, achieving successful predictions even for *E. coli*, *M. tuberculosis* and *S. aureus*, where the species-specific models performed poorly before. This observation suggests that performance is related to the amount and diversity of recorded antigens.

As for the performance of these generic models, the observed AUC scores obtained similar values to the ones obtained in the species-specific models trained with balanced data, indicating that while these generic models did not have information about the species being tested, the data obtained from all the other 14 species was enough to learn the generic rules that made a protein antigenic. This idea is reinforced by looking at the coefficients for each predictor, which were very robust across all 15 pan-species models, indicating that the different leave-one-out generic models reached a similar conclusion on what makes a protein "antigenic" (see Figure 2.4).

The scores produced by APRANK for each protein and the best peptides for each of these 15 species can be found deposited in Dryad under DOI:10.5061/dryad.zcrjdfnb1.

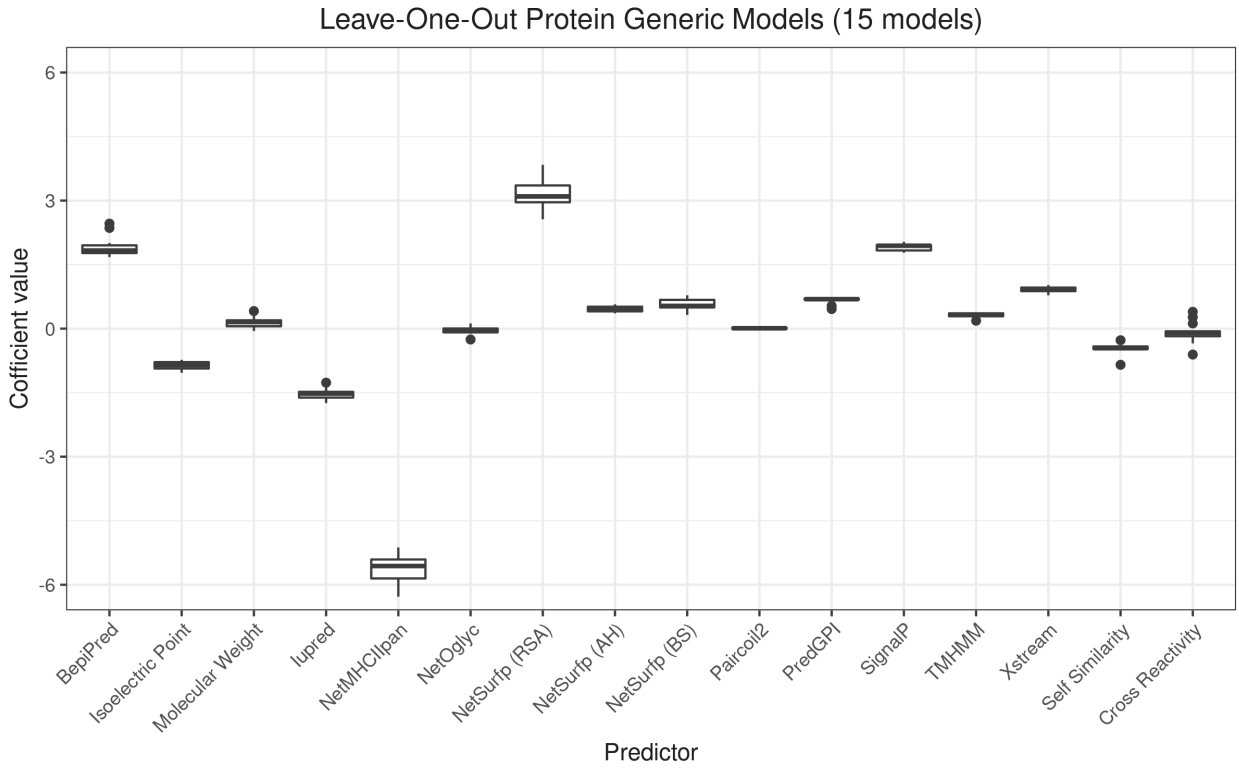


FIGURE 2.4 – *Coefficient values for the leave-one-out generic models.* *Plots were obtained by recording the coefficient of each predictor in the binomial logistic regression models. The different protein models correspond to each of the 15 leave-out-out generic models used to test APRANK. All 15 models converged before reaching the maximum iteration limit when training.*

Species	Proteins		Peptides			
	BTR	LOO model	BTR	LOO model	LOO model + protein scores	Combined score relative AUC gain
B. burgdorferi	Yes	0.786	Yes	0.768	0.950	**23.60%**
B. melitensis	Yes	0.774	-	-	-	-
C. burnetii	Yes	0.620	-	-	-	-
E. coli	Yes	0.754	Yes	0.742	0.780	**5.12%**
F. tularensis	Yes	0.698	-	-	-	-
L. interrogans	Yes	0.947	Yes	0.679	0.948	**39.57%**
P. gingivalis	Yes	0.854	Yes	0.665	0.871	**30.91%**
M. leprae	Yes	0.758	Yes	0.692	0.731	**5.68%**
M. tuberculosis	Yes	0.702	Yes	0.586	0.711	**21.17%**
S. aureus	Yes	0.737	Yes	0.752	0.790	**5.03%**
S. pyogenes	Yes	0.983	Yes	0.838	0.970	**15.81%**
L. braziliensis	Yes	0.709	Yes	0.946	0.878	**-7.20%**
P. falciparum	Yes	0.807	Yes	0.748	0.835	**11.66%**
T. gondii	Yes	0.837	Yes	0.583	0.720	**23.51%**
T. cruzi	Yes	0.867	Yes	0.843	0.857	1.58%

Table 2.6 – ***Prediction results for the leave-one-out generic models.*** *The prediction was considered successful if it was significantly Better Than a Random set of scores (BTR). For peptides, we show both the performance of the model alone, and the performance obtained by combining the protein and peptide scores. In bold we show any difference greater than 5% between the peptide score and the combined score for a given species. LOO Model = Leave-One-Out Model.*

2.2.5 Using APRANK to obtain antigen-enriched sets

Our generic models allowed us to rank proteins and peptides in a given species based on a model trained from other pathogens. Now, we wanted to use these scores to select a subset of proteins or peptides that was enriched in antigens, meaning, that each protein or peptide in that subset had an increased chance of being antigenic when compared to the whole proteome.

For this, we focused on *T. cruzi*, as this was the species with the largest number of recorded antigens within our collection. To obtain fair antigenicity scores for this protein we used the corresponding leave-one-out models created when testing the generic models. We analyzed the distribution of the normalized scores returned by these models, distinguishing between antigenic and non-antigenic proteins and peptides (see Figure 2.5). As was expected, the peak of the scores for the antigens is found to the right of the one for the non-antigens, indicating that the average APRANK score is higher for the antigenic proteins and peptides. It is also worth mentioning that the amount of overlapping seen in these types of plots can be related to the corresponding AUC, where the higher the AUC, the less the overlapping.

Once we had our score distributions, we used them to select an antigen-enriched subset of proteins and peptides. This could be done in one of two ways: either by setting a score threshold or by simply selecting a fixed number of proteins and peptides within the top scores. After analyzing the distribution of score values, we decided to use the first option and selected those proteins and peptides with a normalized APRANK score of at least 0.6. We next calculated what we called *enrichment score* (ES), which was the proportion of antigens in the selected subset relative to the proportion of antigens in the whole proteome (for example, ES = 2 meant you were twice as likely to find an antigen in the subset than in the whole proteome, or in a random subset). In Figure 2.5 we show the enrichment scores for the different normalized scores and the number of proteins and peptides that fall inside or outside those subsets. While the subsets usually contained a small fraction of the whole proteome (close to 10% in most cases), focusing on them would result in a 4 – 6 fold increase in the chances of finding antigens.

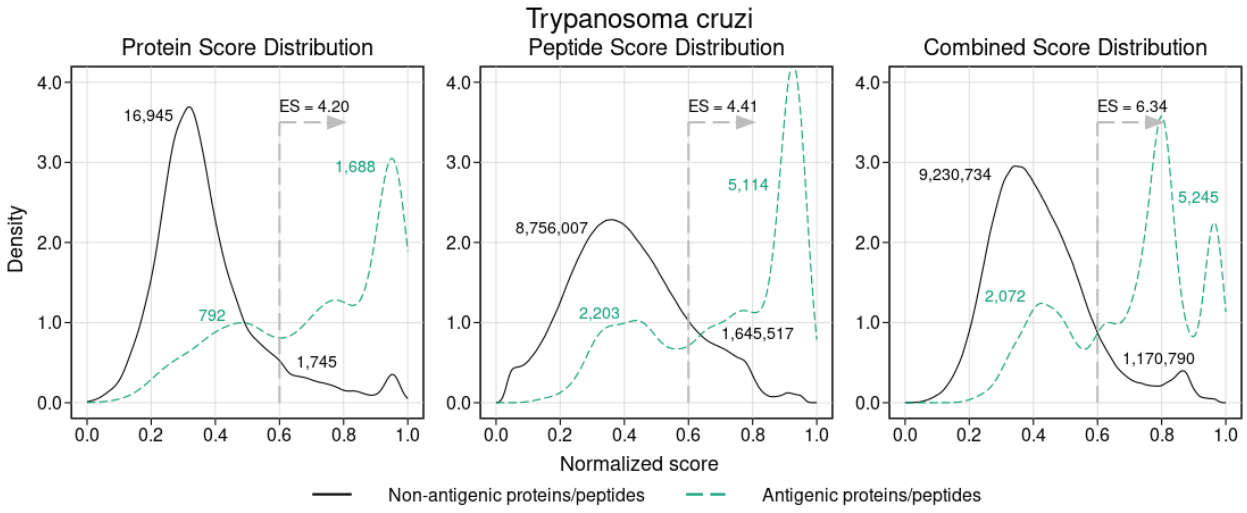


Figure 2.5 – ***Density analysis for the antigenicity scores of T. cruzi.*** *Plots were obtained by analyzing the proteome of T. cruzi with the leave-one-out generic models, and then distinguishing between antigens and non-antigens. The figure shows the enrichment score obtained by keeping only the proteins and peptides with a score greater than 0.6, as well as the amount of antigens and non-antigens that would be inside or outside that subset.*

2.2.6 Assessing the validity of the computational method

Now that we had a working pan-species model, we set to validate the performance achieved by APRANK. For this, we first assessed that the performance was the result of combining information from different predictors, and not from just one or a few of them.

To do this, we selected the predictors that managed to consistently discern antigenic proteins (see Table 2.4) and we calculated the area under the ROC curve (AUC) for both known proteins and known peptides in each case (data not shown). We found that the predictor with best solo predicting capabilities was BepiPred 1.0. We then compared BepiPred's predictions against APRANK's for both the protein and peptide generic models for each species. This is presented in Table 2.7.

We focused on those cases where the AUC changed at least 5% between BepiPred 1.0 and APRANK's generic models. APRANK showed increased predicting capabilities for 11 out of the 15 analyzed proteomes at the level of complete proteins and/or peptides, while showing a decrease in performance only in *M. leprae* at protein level. These results provide validation support to the approach built into APRANK by combining information from many predictors.

As an additional test, we also assessed the performance of APRANK after removing BepiPred 1.0 predictions from our model. This can be seen in Table 2.8. In this simulation we observed that even without BepiPred 1.0 our model reached similar predicting capabilities in most cases, hence suggesting that other predictors and features included in APRANK were able to replace BepiPred when training the model (this is further discussed in this chapter's Discussion).

Finally, to ensure that our model was doing more than simply detecting sequence similarity, we also compared our performance against a "BLAST model", meaning a model that was based solely on how similar a given protein was to a known antigenic protein. The comparison between the performance of this model and APRANK can be seen in Table 2.9. As expected, APRANK achieved a larger AUC for most for the species; however we observed that for *M. leprae* and *L. braziliensis* the "BLAST model" actually resulted in a better prediction. This may be explained because these were species with a small number of validated antigens (test cases) and a with high similarity to other of our selected species. To test this, we repeated this analysis for these two species, but now we removed from the BLAST model the species that were most similar to the one being analyzed (see bottom rows in Table 2.9). The performance under these altered conditions indeed resulted in significantly lower AUCs, matching or falling behind APRANK.

Species	Proteins			Peptides		
	BepiPred score AUC	**APRANK score AUC**	**APRANK relative AUC gain**	**BepiPred score AUC**	**APRANK score AUC**	**APRANK relative AUC gain**
B. burgdorferi	0.729	**0.786**	**7.94%**	0.796	0.768	-3.46%
B. melitensis	0.710	**0.774**	**8.93%**	-	-	-
C. burnetii	0.558	**0.620**	**11.13%**	-	-	-
E. coli	0.587	**0.754**	**28.39%**	0.662	**0.742**	**12.21%**
F. tularensis	0.570	**0.698**	**22.40%**	-	-	-
L. interrogans	0.839	**0.947**	**12.87%**	0.676	0.679	0.42%
P. gingivalis	0.852	0.854	0.25%	0.674	0.665	-1.36%
M. leprae	**0.868**	0.758	**-12.67%**	0.689	0.692	0.51%
M. tuberculosis	0.666	**0.702**	**5.29%**	0.561	0.586	4.58%
S. aureus	0.723	0.737	1.86%	0.767	0.752	-1.93%
S. pyogenes	0.970	0.983	1.33%	0.8	0.838	4.73%
L. braziliensis	0.549	**0.709**	**29.00%**	0.905	0.946	4.48%
P. falciparum	0.793	0.807	1.84%	0.642	**0.748**	**16.42%**
T. gondii	0.579	**0.837**	**44.59%**	0.584	0.583	-0.21%
T. cruzi	0.814	**0.867**	**6.54%**	0.819	0.843	3.03%

Table 2.7 – ***Comparison between APRANK and the predictor with highest solo AUC (BepiPred 1.0).*** *The relative AUC gain shows the increase or decrease of the AUC obtained by our method relative to the one obtained by BepiPred. Differences greater than 5% are shown* ***in bold****.*

Species	Group	Proteins APRANK score			Peptides APRANK score		
		without BepiPred AUC	with BepiPred AUC	Relative AUC gain	without BepiPred AUC	with BepiPred AUC	Relative AUC gain
B. burgdorferi	Gram -	0.777	0.786	1.18%	0.726	**0.768**	**5.78%**
B. melitensis	Gram -	0.749	0.774	3.39%	-	-	-
C. burnetii	Gram -	0.616	0.620	0.61%	-	-	-
E. coli	Gram -	0.751	0.754	0.42%	0.743	0.742	-0.07%
F. tularensis	Gram -	0.714	0.698	-2.15%	-	-	-
L. interrogans	Gram -	0.938	0.947	0.96%	0.646	**0.679**	**5.15%**
P. gingivalis	Gram -	0.847	0.854	0.75%	0.626	**0.665**	**6.19%**
M. leprae	Gram +	0.750	0.758	1.04%	0.657	**0.692**	**5.37%**
M. tuberculosis	Gram +	0.697	0.702	0.66%	0.586	0.586	0.00%
S. aureus	Gram +	0.762	0.737	-3.31%	0.751	0.752	0.19%
S. pyogenes	Gram +	0.983	0.983	0.04%	0.826	0.838	1.47%
L. braziliensis	Eukaryote	0.687	0.709	3.20%	0.928	0.946	1.88%
P. falciparum	Eukaryote	0.801	0.807	0.84%	0.753	0.748	-0.73%
T. gondii	Eukaryote	0.835	0.837	0.27%	0.585	0.583	-0.47%
T. cruzi	Eukaryote	0.869	0.867	-0.29%	0.833	0.843	1.26%

Table 2.8 – *Comparison between APRANK and a version of APRANK without the predictor with highest solo AUC (BepiPred 1.0).* *The relative AUC gain shows the increase or decrease of the AUC obtained by APRANK relative to the version of APRANK without BepiPred. In bold we show differences greater than 5%. Due to the large number of peptides, each individual peptide AUC was calculated as the mean of 5 pseudo-random subsets of 50,000 peptides (see Methods).*

Species	Proteins		
	BLAST AUC	APRANK AUC	Relative AUC gain
B. burgdorferi	0.502	0.786	**56.60%**
B. melitensis	0.637	0.774	**21.49%**
C. burnetii	0.579	0.620	**7.14%**
E. coli	0.677	0.754	**11.44%**
F. tularensis	0.629	0.698	**10.92%**
L. interrogans	0.499	0.947	**89.86%**
P. gingivalis	0.544	0.854	**57.04%**
M. leprae	0.893	0.758	**-15.10%**
M. tuberculosis	0.591	0.702	**18.78%**
S. aureus	0.622	0.737	**18.56%**
S. pyogenes	0.542	0.983	**81.26%**
L. braziliensis	0.951	0.709	**-25.42%**
P. falciparum	0.594	0.807	**35.77%**
T. gondii	0.443	0.837	**88.98%**
T. cruzi	0.501	0.867	**72.95%**
M. Leprae (without M. tuberculosis in BLAST)	0.650	-	**16.63%**
L. braziliensis (without T. cruzi in BLAST)	0.744	-	-4.74%

Table 2.9 – *Comparison between APRANK and a "BLAST model".* *The "BLAST model" worked by assigning to each protein a score related to how similar they were to a recorded antigenic protein. For the two species that resulted in a better prediction when using the "BLAST model", we also tested removing from the BLAST results (and so, from the model) the species that was the most similar to the one being analyzed. In bold we show differences greater than 5%.*

2.2.7 Applying our method on a novel species

To truly validate APRANK, we wanted to test the method on a new species that was not included in our initial training and that had an extensive amount of information on the antigenicity of its proteins and peptides. For this, we searched for publications containing proteome-wide linear epitope screenings using high-density peptide microarrays and selected a recent data set produced by scanning the complete *Onchocerca volvulus* proteome with more than 800,000 short peptides, mostly 15mers (Lagatie et al. [69]). *Onchocerca volvulus* is a nematode and it is the causative agent of Onchocerciasis in humans (also called *"river blindness"*), a disease labeled as *"Neglected Tropical Disease"* by the World Health Organization [167].

We obtained a list of antigens from *O. volvulus* following the same rules applied by the authors to find the peptides they called "immunoreactive" (see Methods in Lagatie et al. [69]), resulting in a set of almost 1,100 antigenic peptides. We tagged a protein as antigenic if it had at least one of these peptides; however, we also kept information on how many "immunoreactive" peptides each protein had for later analysis. Once this was done, we also tagged as antigenic any neighboring peptide that shared at least 8 amino acids with one of these "immunoreactive" peptides.

We next trained APRANK with all our 15 species and then used these models to predict the antigenicity scores for both the proteins and the peptides of *O. volvulus*. An AUC score was calculated for each prediction, comparing the score given by APRANK against the antigenic tag for each protein and peptide. We also calculated the enrichment scores for these scenarios using a score threshold of 0.6 in a similar way that we did for *T. cruzi*. The scores obtained by APRANK for each protein and the best peptides of *O. volvulus* can be found deposited in Dryad under DOI:10.5061/dryad.zcrjdfnb1.

Our method was successful in predicting the antigenicity of proteins and peptides for *O. volvulus*, as shown in Table 2.10. We observed that if we were more strict when tagging a protein as antigenic, meaning requiring more "immunoreactive" peptides, we obtained better performance. When considering as antigenic any protein with 1 "immunoreactive" peptide we had an enrichment score of 2.28, whereas when we increased this requirement to 3 peptides the enrichment score was 5.29 (see Table 2.10, Figure 2.6). Besides validating the performance of APRANK on a new pathogen, this suggests that either our method is better in predicting proteins with many antigenic regions, or that a single reactive peptide from a peptide array screening may provide only weak support for calling of antigens.

For peptides, APRANK obtained an enrichment score of 3.33 – 3.90, also showing an additive effect when combined with the protein score, suggesting that these are effective in predicting antigenicity for *O. volvulus*. Similar to before, we tried being more strict and only considering antigenic peptides in proteins with at least 2 or 3 "immunoreactive" peptides; however this did not seem to affect the predictive performance as much as for whole proteins.

	Total	Score	#MIP	Antigenic	AUC	Antigens with score 0.6	Enrichment score for 0.6
Proteins	12,994	Protein score	1	886	0.677	150	2.28
			2	177	0.713	38	2.89
			3	28	0.828	11	5.29
Peptides	4,872,082	Peptide score	1	1,097 → 14,122	0.800	6,108	3.33
			2	397 → 4,498	0.798	1,995	3.42
			3	104 → 1,182	0.836	598	3.90
		Combined score	1	1,097 → 14,122	0.750	3,376	3.10
			2	397 → 4,498	0.774	1,342	3.88
			3	104 → 1,182	0.871	512	5.63

Table 2.10 – ***Performance of APRANK on Onchocerca volvulus.*** *Proteins and peptides were tagged as antigenic based on the number of Minimum Immunoreactive Peptides (#MIP). For proteins, we considered as antigenic those with at least #MIP immunoreactive peptides. For peptides, we considered as antigenic any immunoreactive peptide found inside proteins with at least #MIP immunoreactive peptides. We show the number of antigenic peptides before and after spreading the antigenicity from the original immunoreactive peptides to their neighboring peptides (before → after). The rule to define an "immunoreactive peptide" was extracted from Lagatie et al. 2017. The enrichment score represents the proportion of antigens in the selected subset relative to the proportion of antigens in the whole proteome.*

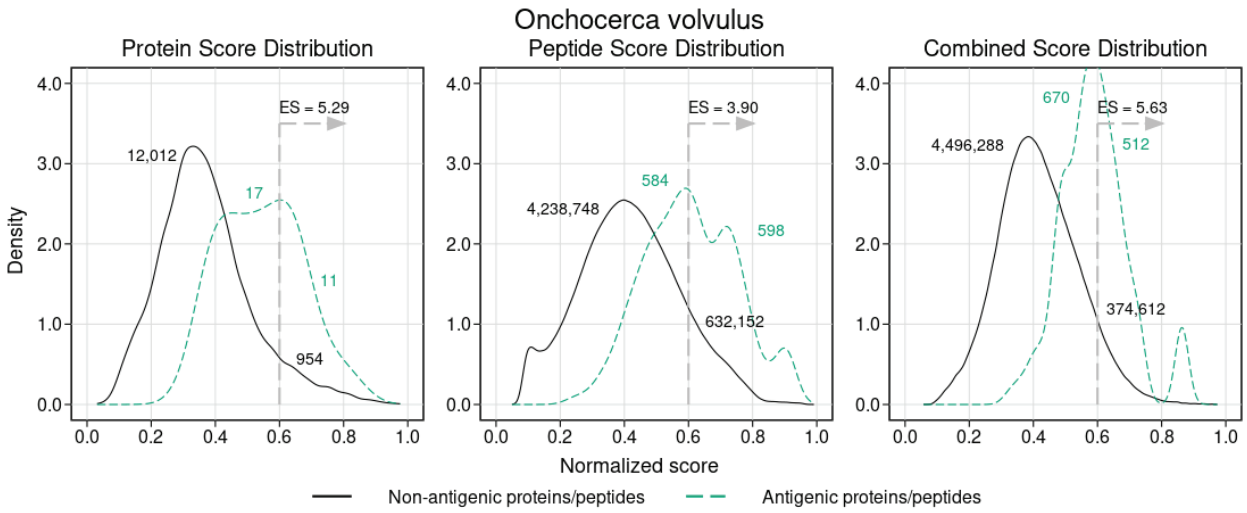


Figure 2.6 – ***Density analysis for the antigenicity scores of Onchocerca volvulus.*** *Plots were obtained by analyzing the proteome of O. volvulus with the final generic models, and then distinguishing between antigens and non-antigens. The figure shows the enrichment score obtained by keeping only the proteins and peptides with a score greater than 0.6, as well as the amount of antigens and non-antigens that would be inside or outside that subset. The plots correspond to the case where a protein was tagged as antigenic if it had at least 3 "immunoreactive" peptides (see Results).*

2.2.8 Applying our method on a novel data set: exploring seroprevalence

As a final step, we also tested performance of APRANK on an additional data set from *Plasmodium falciparum* that was not used as a source of validated antigens in our previous training. In this study, the authors analyzed the proteome of *P. falciparum* using a protein microarray which displayed ∼91% of the proteome, but more importantly, they also analyzed the individual antibody responses of 38 patients in controlled human malaria infections (Obiero et al. [70]). This resulted in a rich set of information on seroprevalence for each analyzed protein.

Whit this information we analyzed if the APRANK scores predicting antigenicity were in any way correlated with the observed seroprevalence. This seroprevalence data encompassed 4,768 unique genes, and was matched against APRANK protein scores for *P. falciparum*. To avoid the possibility of over-fitting, the APRANK scores were those obtained from the leave-one-out generic model trained in 14 species, but leaving out *P. falciparum*. This resulted in 4,343 proteins with information of both seroprevalence (from Obiero et al. [70]) and antigenicity score (from our work).

The results of this analysis are summarized in Figure 2.7. Unlike previous cases where the proteins in the test set were put in binary classes (antigenic vs non-antigenic), here we divided the data in 5 groups, using seroprevalence cutoffs at the 5%, 10%, 20% and 40% levels. The distribution of APRANK scores for these groups showed that proteins with higher seroprevalence also had higher APRANK scores, and hence shift to the right of the plot. This was evident in the separation of the non-antigenic bulk of the proteome (< 5% seroprevalence) from those proteins that are in the 10% - 20% seroprevalence range, and also and importantly in the highly seroprevalente antigens (seroprevalence >= 40%), where the density of the peak shifts further towards higher scores. This was as well supported by the AUC prediction of these two groups, which was 0.660 for the 10% - 20% seroprevalence range and 0.740 for the >= 40% range. While further studies of this kind are necessary to explore the link between antigenicity and seroprevalence, these results further validate APRANK at the task of prioritizing antigenic proteins from complete proteomes.

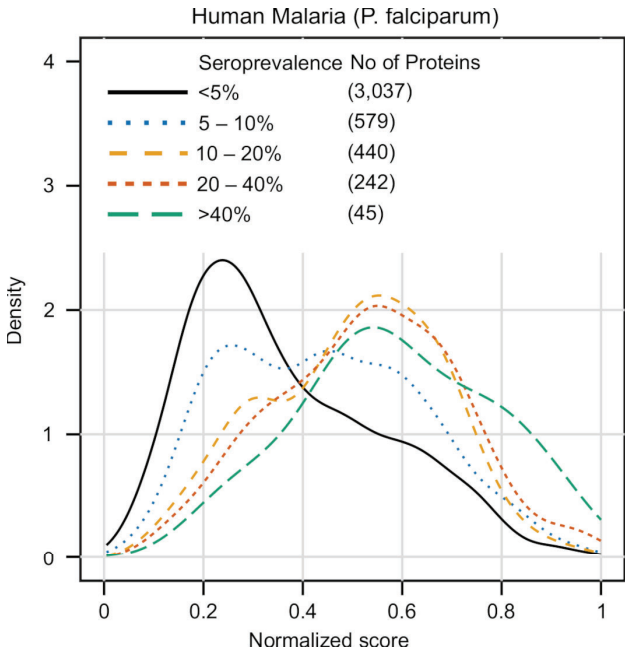


Figure 2.7 – ***Validation of APRANK against antigens with known seroprevalence.*** *Detailed information on the seroprevalence of Plasmodium falciparum proteins in cases of Human Malaria was obtained from Obiero et al. [70] (n = 38). Proteins were clustered in different seroprevalence groups and matched against APRANK antigenicity scores (see Results).*

2.3 Discussion

We present APRANK, a novel method to prioritize and predict the best antigen candidates in a complete pathogen proteome. APRANK relies on a number of protein features that can be calculated for any protein sequence which are then integrated in a pan-species model. We have tested this integrative method using non-parametric ROC-curves and made an unbiased validation using two independent data sets containing recent proteome-wide antigenicity data (*O. volvulus* and *P. falciparum*). In summary, our benchmarks show that by integrating multiple predictors, pooling antigen data from multiple species across a wide phylogenetic selection, and balancing training data sets, APRANK was successful in predicting antigenicity for all pathogen species tested, hence providing a new and improved method to obtain antigen-enriched protein and peptide subsets for a number of downstream applications.

Looking forward

While we are satisfied by APRANK's performance, there are still ways to further improve it. The main issue we had when training our models is the current lack or sparsity of validated epitope and antigen information. Particularly, well validated non-antigenic sets are currently hard to find in the literature, forcing us to count as non-antigenic all proteins and peptides that do not currently have experimental evidence of antigenicity or were not tagged as antigenic in databases (which we know is hardly true). Obtaining validated data about non-antigenic proteins and peptides will improve the training of the models for future versions of APRANK.

We also observed that the performance of APRANK was not considerably affected by removing some individual features. This might indicate that each individual predictor contributes only slightly to the overall performance, or that there might be redundancy between some of the predictors. For example, the features that were used to train BepiPred 1.0 HMMs (propensity scales for secondary structure preference and hydrophilicity of amino acid residues [156]) may overlap with some of the predictors we used for APRANK. Future versions of APRANK will review these overlaps, analyzing the pros and cons of adding novel predictors or removing existing ones.

Regarding the computing performance of APRANK, the majority of its running time is dedicated to run the predictors used internally, most of which run in a reasonable time in a commodity server. However, there are a few bottlenecks, most notably predictions by NetSurfP. This should be improved in a future version in order to offer APRANK e.g. as a web-service. Future work will also explore the possibility to extend APRANK to also use data from other experimental (non-computable) sources, such as evidence of expression derived from proteomic or transcriptomic experiments.

Finally, APRANK is currently focused on finding linear epitopes, and likely missing most of the conformational ones. This is evidently a limitation, but also reflects the current imbalance on experimental validation of linear vs conformational epitopes. There is much more information on linear epitopes and hence the field is ripe to develop applications like APRANK. This also affected the selection of predictors, many of which are also biased to predict/analyze linear features, and the selection of validated antigen and peptide data, which were obtained mostly from peptide microarray data. Introduction of new predictors may increase the amount of conformational information used to rank epitopes, but finding and reporting conformational epitopes would entail large changes to how APRANK currently works. While we believe that the best path forward for APRANK is to focus future work in increasing the accuracy of the prediction for linear epitopes, we do not rule out the possibility of adding the detection of conformational epitopes to this method.

2.4 Materials and methods

All methods are described in detail herein, but are also documented as R and Perl code in APRANK's source code, available at our GitHub Repository, along with the data sets analyzed for this study. Trained models and antigenicity scores were deposited in Dryad under DOI:10.5061/dryad.zcrjdfnb1.

2.4.1 Bioinformatic analysis

FASTA files containing proteins of the species used to train APRANK (see Table 2.1) were downloaded from publicly available database resources (from complete proteomes) and can be found in our GitHub Repository. To comply with requirements of downstream predictors, unusual amino acid characters were replaced by the character "X" and a few proteins with more than 9,999 amino acids were truncated to that size. To obtain information at peptide level, proteins were split into peptides of 15 residues with an overlap of 14 residues between them (meaning an offset of 1 residue between peptides).

The validated FASTA files were analyzed with BepiPred [156], EMBOSS pepstats, Iupred [162], NetMHCIIpan [157], NetOglyc [158], NetSurfp [168], Paircoil2 [163], PredGPI [159], SignalP [160], TMHMM [164], Xstream [161] and two custom perl scripts that analyzed similarity of short peptides against the human genome (NCBI BioProject PRJNA178030). The reasoning of choosing each predictor, what they predict and which version was used can be found in Table 2.3. NetMHCIIpan was run multiple times for different human alleles (DRB1*0101, DRB3*0101, DRB4*0101 and DRB5*0101). The only predictor that needed an extra preprocessing step was PredGPI, which required removing sequences shorter than 41 amino acids and those with an "X" in their sequence. For all purposes, these filtered sequences were assumed to not have a GPI anchor signal. The versions of Ubuntu, R, Perl, packages and modules used to create the computational method, as well as the full console call for each predictor, can be found in Supplementary Tables S2.1 and S2.2.

2.4.2 Compiling a data set of curated antigens

To obtain antigenic proteins and peptides, we extracted information from the immune epitope database (IEDB), as well as information from several papers, most of which relied on data from protein or peptide microarrays combined with sera of infected patients to find new antigens [149, 153, 154, 169–182].

Because different protein identifiers are used across papers, we used either the Uniprot ID mapping tool, the blastp suite of BLAST or a manual mapping to find the corresponding ID or IDs that a given antigen had in our proteomes. The list of all antigenic proteins and peptides used can be found in our GitHub Repository, and their source and the mapping method used in each case can be found in Supplementary Tables S2.4 and S2.5.

For the antigenic peptides, though, mapping the original protein ID to our pathogen proteomes was not enough; we also had to assign the antigenicity to the corresponding peptides within each antigenic protein. However, while our peptides were of fixed length, the curated antigenic sequences varied in size. For this reason, we developed our own mapping method that we called "kmer expansion", which works by marking as antigenic any peptide that shared a kmer of at least 8 amino acids with a curated antigenic sequence for that same protein. The amount of total and antigenic peptides, before and after the "kmer expansion", are listed in Table 2.2.

In the case of *Onchocerca volvulus*, the method we used to derive antigenic proteins and peptides was based on experimental proteome-wide data on antibody-binding to short peptides [69]. We followed the same rules used by these authors to find the peptides they called "immunoreactive". Because these peptides had lengths from 8 to 15 amino acids, we assigned as antigenic any neighboring peptides that shared at least 8 amino acids with them (this is more strict that using our "kmer expansion" strategy because it limits the antigenicity to that section of the protein).

2.4.3 Clustering by sequence similarity

We calculated sequence similarity for all proteins in the 15 analyzed proteomes using blastp from the NCBI BLAST suite [155]. We wanted to filter the BLAST output keeping only the good matches, which meant selecting a similarity threshold. After analyzing different matches, we arrived at a sensible compromise: trying to be as strict as possible without losing much data. For this we kept matches with a percentage of identical amino acids of at least 75% (meaning a *pident* of 0.75), an expected value (*evalue*) less than or equal to 1×10^{-12} and a match length of at least half of the length of the shortest protein in the match.

Using these matches, we created a distance matrix where the *distance* was calculated as the difference between 1 and *pident*. We then applied a single-linkage hierarchical clustering method to group the proteins into a similarity tree and cut this tree using a cutoff of 0.25 (the difference between 1 and the threshold used for *pident*). This process gave us clusters of similar proteins, which we then used to spread the antigenicity from proteins labeled as antigenic to similar proteins without that label. The amount of total and antigenic proteins, before and after using BLAST to find similar proteins inside each species, can be seen in Table 2.2.

When creating the species-specific models, any protein which belonged to a cluster with at least one other antigenic protein was also tagged as antigenic, even across species. As a consequence of this, and to avoid overfitting, proteins in a given cluster were kept together in the training process, meaning they would all be either in the training set or in the test set.

For the generic models, any protein in the training set which belonged to a cluster with at least one other antigenic protein was also tagged as antigenic, even across species (excluding the species being tested). As for the test set, this would also occur, but only inside that same species (the one being tested).

2.4.4 Data normalization

Each predictor used by APRANK varied on how they returned their values. Not only they had different value ranges, but while some of them returned their values per protein, others did so per peptide, kmer, or amino acid. For this reason, we needed to parse and normalize all outputs before feeding their data into our models.

Values returned by each predictor were normalized to fit a numeric range between 0 and 1. Different methods were used to parse and normalize the data for each combination of predictor and model, ranging from linear or sigmoid normalizations to a simple binary indicator of presence or absence of a given feature (such as signal peptide). The methods used to normalize the output for each predictor were the result of analyzing the distribution and spread of these outputs across all of our species for each predictor individually, coupled with biological knowledge of what each predictor was analyzing. Those predictors that returned information exclusively at protein level were not used in the peptide models. The detailed steps on how to parse and normalize the output of each predictor for the protein and the peptide models can be found in Supplementary Table S2.3

and Supplementary Equations S2.1, S2.2 and S2.3. Furthermore, this is also documented in the code available at our GitHub Repository.

2.4.5 Fitting the species-specific models

To fit each protein species-specific model, clusters for that species were divided in training and test sets in a 1:1 ratio due to the low number of recorded antigens for some species. For this same reason, the training set was balanced with *ROSE* [165], generating an artificial training set with a similar number of antigenic and non-antigenic artificial proteins. This process, as well as all other described below, was repeated 50 times by re-sampling the clusters in the training and test sets.

A binomial logistic regression model was fitted for both the balanced and the unbalanced training sets using the generalized linear models in R (function `glm`). Once the balanced and the unbalanced protein models were trained, we used them to predict the scores for the test set. A schematic visualization of this procedure is shown in Figure 2.1. The performance for each model, measured by the area under the ROC curve (AUC), was then calculated using the R package *pROC* [183]. Additionally, two pseudo random set of scores were created by shuffling the scores achieved by both models. These random protein models were used to test if the performance of our models differed significantly from a random prediction.

For the peptide species-specific models, we divided the peptides into training and test sets by simply following the division of the proteins clusters, meaning that if a protein was in the training set for the protein model, its peptides would be in the training set for the peptide model for that iteration. The models were fitted and random scores calculated in a similar manner to the protein models. However, when we attempted to calculate the performance of the peptide models, our test set was too large to calculate performance based on AUC values in a reasonable time. We decided then to sample a subset of 50,000 peptides from the test set in a pseudo-random manner, making sure that the positive peptides were found in the subset and that the fraction of positive vs indeterminate/negative peptides was similar to the one in the test set (but never below 1% unless we ran out of antigens). All AUC values for the different peptide models were calculated using the same subset, and this process was repeated 5 times in each iteration, changing the subset each time.

Once all iterations were finished, we compared the AUC obtained by the balanced and unbalanced versions of the protein and peptide species-specific models using a Student's t-test. Another set of t-tests were used to analyze the difference between each of those models and their relative random model. If the model had a significantly higher AUC than the corresponding random model, we considered the model achieved a successful prediction ($p < 0.05$).

2.4.6 Creating the generic models

The generic (pan-species) models are the actual models used by APRANK. The objective of these models is to generalize predictions of antigenic proteins or peptides for new species (not used for training APRANK). In a broad sense, they have to learn what makes a protein or a peptide antigenic. We achieved this by training the models with a large set of antigenic proteins and peptides from 15 different species, including gram-negative bacteria, gram-positive bacteria and eukaryotic protozoans.

To create the protein generic model, we used *ROSE* [165] to make a balanced training set of 3,000 proteins for each species and then merged all those balanced training sets together. With these data, a logistic regression model was created following the same steps as for the species-specific models. Next, this model was used to predict the scores for the species being analyzed and the performance

of the prediction was calculated the same way as for the species-specific protein models. A schematic visualization of this procedure is shown in Figure 2.3.

We created the peptide generic model in a similar manner, with balanced training sets from each of the species containing 100,000 peptides each. In addition to the regular score calculated by using the model to predict the antigenicity of the test data, we also calculated a combined score, which is simply the mean of the peptide score and the corresponding protein score. The performance of the peptide generic model was calculated the same way as for the species-specific peptide models.

When testing these generic models, we created temporary leave-one-out generic models, where we used 14 of the species to generate the generic protein and peptide models, and then tested the models in the 15th species. We then generated the final protein and peptide generic models using all 15 species and tested them by predicting antigenicity in *Onchocerca volvulus*, a novel species for APRANK, with experimental proteome-wide data [69].

2.4.7 Comparative performance

To discard the possibility that our model was simply detecting sequence similarity, we created a "BLAST model", where we assigned to each protein a score based solely on how similar they were to a known antigenic protein from another organism. The score used was $-log_{10}(evalue)$ and then performance was calculated for each species.

We also wanted to make sure our model was combining information from several predictors. We compared our prediction capabilities against the individual predictor with best AUC, which was BepiPred 1.0. The BepiPred score for each protein and peptide was obtained from the individual amino acid scores following the same steps we used for APRANK, but without normalizing it. The AUCs for the BepiPred peptide scores were calculated the same way as for the peptide species-specific models.

2.5 Supplementary Materials

2.5.1 Supplementary Tables

All Supplementary Tables mentioned in this chapter were deposited as a single Excell file in Figshare under DOI:10.6084/m9.figshare.22047782.v1.
Direct Link: PhD Thesis - Ricci - Chapter 2 - Supplementary Tables.xlsx (Size: 798 KB)

Additionally, Supplementary Tables S2.1, S2.2 and S2.3 can also be found below.

Software	Version
Ubuntu	16.04
R	3.4.3
ROSE (R package)	0.0.3
pROC (R package)	1.12.1
Perl	5.22.1
BioPerl (Perl module)	1.007002

Supplementary Table S2.1 – ***Versions of the software, packages and modules used to create our computational method.***

Predictor	Call	Data extracted
BepiPred 1.0	bepipred $fasta_file -k >$output_file	Score per amino acid
BLAST+ 2.2.31	blastp -query $query_file -db $db_file -outfmt 6 -out $output_file -max_target_seqs 2000	Similarity between proteins (used to assign protein antigenicity)
EMBOSS 6.6.0.0	pepstats -sequence $sequence_file -sprotein1 -aadata Eamino.dat -mwdata Emolwt.dat -termini -nomono -auto -outfile $output_file	Isoelectric Point and Molecular Weight per protein
Iupred 1.0	iupred $sequence_file short >$output_file	Score per amino acid
NetMHCIIpan 2.0	netMHCIIpan -a $allele -f $sequence_file -l $peptide_length >$output_file	%Rank per peptide per MHC II allele used
NetOglyc 3.1d	netOglyc $sequence_file >$output_file	Glycosilation presence per amino acid
NetSurfp 1.0	NetSurfp $sequence_file -a >$output_file	Relative Surface Accessibility, Probability for Alpha-Helix and Probability for Beta-strand per amino acid
Paircoil2	paircoil2 $fasta_file $output_file $error_file	P-score per amino acid
PredGPI 1.4.3	PredGPI.py $filtered_fasta_file >$output_file	Presence and start of GPI per protein
SignalP 4.0	signalp -f long -t $organism_group $fasta_file $output_file	Presence and start of signal peptide per protein and C and S score per amino acid
TMHMM 2.0c	tmhmm $fasta_file >$output_file	Presence of transmembrane helix per protein and amino acid participation in it per amino acid
Xstream 1.71	java -jar $Xstream_path/xstream.jar $sequence_file -d$output_path/	Start, end, period, copy number and consensus error per repeat per protein

Supplementary Table S2.2 – *Third-party software used to retrieve information about the proteins and peptides.* *The call being shown corresponds to those to use under Ubuntu 16.04. Words starting with $ symbolize variables to be replaced by their corresponding values.*

Supplementary Table S2.3 – *Normalization methods used for each predictor in protein and peptide analysis.* *The formulas mentioned are shown in this chapter's Supplementary Equations.*

Predictor's output	Protein	Peptide
BepiPred	Calculate the mean of the BepiPred score for the amino acids inside the protein and normalize it using fixedLinearNormalization with -1.5 and 1.5 as limits	Calculate the mean of the BepiPred score for the amino acids inside the peptide and normalize it using fixedLinearNormalization with -1.5 and 1.5 as limits
Isoelectric Point	Divide the isoelectric point value by 14	-
Molecular Weight	Normalize the molecular weight value using sigmoidNormalization05 with a b of 30,000.	-
Iupred	Calculate the ratio of amino acids inside the protein with an score greater or equal than 0.5	Calculate the ratio of amino acids inside the peptide with an score greater or equal than 0.5
NetMHCIIpan	Calculate the mean of the ranks for the kmers inside the protein, divide it by 100, normalize it using fixedLinearNormalization with 0.05 and 0.5 as limits, and then substract that number from 1	Calculate the mean of the ranks for the kmers inside the peptide, divide it by 100, normalize it using fixedLinearNormalization with 0.05 and 0.5 as limits, and then substract that number from 1
NetOglyc	Calculate the ratio of glycosilated amino acids inside the protein, and normalize it using fixedLinearNormalization with 0 and 0.05 as limits	Check if the peptide has at least 1 glycosilated residue
NetSurfp (RSA)	Calculate the mean of the values for the amino acids inside the protein	Calculate the mean of the values for the amino acids inside the peptide
NetSurfp (Alpha Helix)	Calculate the mean of the values for the amino acids inside the protein	Calculate the mean of the values for the amino acids inside the peptide
NetSurfp (Beta Strand)	Calculate the mean of the values for the amino acids inside the protein	Calculate the mean of the values for the amino acids inside the peptide
Paircoil2	Check if the protein has an amino acid sequence of a given length (50 by default) where all the amino acids has a score above a threshold (0.5 by default)	Calculate the ratio of amino acids inside the peptide with a score above a threshold (0.5 by default)
PredGPI	Check if the protein has a GPI	Check if the peptide is at least in part inside the GPI
SignalP	Check if the protein has a signal peptide	Check if the peptide is at least in part inside the signal peptide
TMHMM	Use the output as it is	Check if the peptide is at least in part inside a transmembrane helix
Xstream	Find the largest copy number in the protein and normalize it using sigmoidNormalization09 with a b of 5	Assign to each amino acid the highest copy number it's involved in, calculate the mean of that value for the amino acids inside the peptide and normalize it using sigmoidNormalization09 with a b of 5
Cross Reactivity	For each kmer in the protein, normalize the amount of times that kmer appears in the host proteome using sigmoidNormalization05 with a b of 1, then calculate the mean of these values	For each kmer in the peptide, normalize the amount of times that kmer appears in the host proteome using sigmoidNormalization05 with a b of 1, then, calculate the mean of these values
Self Similarity	For each kmer in the protein, normalize the amount of other times that kmer appears in the proteome using sigmoidNormalization05 with a b of 1, then calculate the mean of these values	For each kmer in the peptide, normalize the amount of other times that kmer appears in the proteome using sigmoidNormalization05 with a b of 1, then, calculate the mean of these values

Supplementary Table S2.4 – *Source and mapping for the proteins initially tagged as antigenic.* *This table can be found in the Excell file deposited in Figshare.*
Direct Link: PhD Thesis - Ricci - Chapter 2 - Supplementary Tables.xlsx (Size: 798 KB)

Supplementary Table S2.5 – *Source and mapping for the peptides initially tagged as antigenic.* *This table can be found in the Excell file deposited in Figshare.*
Direct Link: PhD Thesis - Ricci - Chapter 2 - Supplementary Tables.xlsx (Size: 798 KB)

2.5.2 Supplementary Equations

$$fixedLinearNormalization(x, m, M) = \begin{cases} 0 & for\ x \leq m \\ \frac{x-m}{M-m} & for\ m < x < M \\ 1 & for\ x \geq M \end{cases} \quad \text{(S2.1)}$$

$$sigmoidNormalization05(x, b) = -1 + \frac{2}{1 + 3^{-\frac{x}{b}}} \quad \text{(S2.2)}$$

$$sigmoidNormalization09(x, b) = -1 + \frac{2}{1 + 20^{-\frac{x}{b}}} \quad \text{(S2.3)}$$

2.5.3 Supplementary Files

Regression models used by APRANK

The trained generic models for proteins and peptides were deposited in Dryad under DOI:10.5061/dryad.zcrjdfnb1. The corresponding files are **.rda**, which are an abbreviation for R Data File. If they are downloaded with some other extension (usually **.gz**), they should be renamed back to **.rda**.
Direct Links:
Protein Model: protein_balanced_generic_model.rda (Size: 18.2 MB)
Peptide Model: peptide_balanced_generic_model.rda (Size: 422 MB)

Antigenicity scores for proteins and peptides

The antigenicity scores returned by APRANK for Proteins and Peptides were submitted separately to Dryad under DOI:10.5061/dryad.zcrjdfnb1. The corresponding file is a compressed **.zip** file containing APRANK predictions for each organism. A README file explains the models used to calculate APRANK's scores in each scenario.
Direct Link: ricci-aprank-antigenicity-scores-supplementary-data.zip (Size: 59.3 MB)

3. High-throughput mapping of antigenic epitopes across the *T. cruzi* proteome

This chapter describes the use of high-density peptide microarrays as a platform for antigen discovery and epitope mapping. In this work we have analyzed the whole proteome of two strains of *Trypanosoma cruzi*, in a two-step strategy. The first step focused on a proteome-wide analysis using pooled sera from diverse human populations across the Americas to find antigenic regions. The second step used individual sera to investigate the seroprevalence for each of the regions found in the first step. With this information we identified novel antibody-binding epitopes associated with the chronic phase of Chagas disease, which have the potential to improve both its diagnosis and facilitate the development of other immunoassays. Through this analysis we also obtained a deeper understanding on the adaptive immune response against a pathogen like *T. cruzi*, finding that most regions were antigenic only for a few individuals, while others were detected by many. These data sets enable the study of the Chagas antibody repertoire at an unprecedented depth and granularity, while also providing a rich source of novel serological biomarkers.

The chapter is based on a paper that has been submitted for publication: ***"A Trypanosoma cruzi Antigen and Epitope Atlas: deep characterization of antibody specificities in Chagas Disease patients across the Americas"*** by Ricci AD, Bracco L *et al.*[184]. Here, the relevant information from the paper in its current form (manuscript deposited as a pre-print in August 2022) is transcribed with some minor modifications and updates.

Ricci AD, Bracco L, Ramsey JM, Nolan MS, Lynn MK, Altcheh J,
Ballering G, Torrico F, Kesper N, Villar JC, Marco JD, Agüero F.
A Trypanosoma cruzi Antigen and Epitope Atlas: deep
characterization of antibody specificities in Chagas Disease
patients across the Americas (2022).
bioRxiv 2022.08.19.504544; doi: https://doi.org/10.1101/2022.08.19.504544
This article is a preprint and has not been certified by peer review.

3.1 Introduction

Although the molecular mechanisms that produce diverse antibody repertoires are precisely understood [185, 186], a comprehensive description of their specificities in different infected individuals has been hindered by the lack of powerful tools.

Synthetic peptides have been used historically to map continuous antibody-binding epitopes or to find key residues for protein binding at small scale [187–189]. The introduction of peptide arrays made it possible to display large numbers of these synthetic peptides on a solid surface at addressable positions [54, 190]. Given the sustained increase in the densities achieved by the in situ synthesis of peptides using maskless photolithography [56, 146, 147, 191–193], it is now possible to display complete proteomes in a single slide [69, 194], opening the door to high-throughput serological screenings.

Chagas Disease, also known as American trypanosomiasis, is a lifelong infection caused by the protozoan parasite Trypanosoma cruzi. Despite being discovered ∼100 years ago, the condition remains a major social and public health problem in Latin America and is regarded as a neglected tropical disease by the World Health Organization [87].

After an initial infection, the parasite evades immune mediated elimination and mounts long-lasting chronic intracellular infections. Due to the low parasitemia observed during the chronic stage of the disease, serological methods are the preferred choice for diagnosis of infection. Although available diagnostic tests give satisfactory results in most cases, there is currently no reference ("gold") standard for diagnosis of infection hence discordant results remain a possible cause of undetected cases [195–197]. Also, there are urgent needs to improve or fill vacant niches with customized serological tools and assays to monitor existing treatments or clinical trials [198–200] and to detect the early onset of Chagas Disease pathology [88], both in active case finding and management and in epidemiological and disease surveillance programmes [201, 202].

In this chapter, we report on the comprehensive survey and characterization of the human antibody responses and specificities against *T. cruzi* using state of the art high-density peptide arrays. This survey investigated the antigenicity of the predicted proteomes of two *T. cruzi* strains in 71 Chagas Disease subjects from diverse human populations across the Americas. As a result, we produced a comprehensive atlas of antigens and their epitopes, providing a unique resource for understanding adaptive immune responses against this parasite and to devise improved serological immunoassays for tackling Chagas Disease.

3.2 Results

3.2.1 Design of a high-density peptide array for antigen discovery

We used high-density peptide arrays to perform a high-resolution antigen discovery screening and epitope mapping for complete *T. cruzi* proteomes. We designed an array that included protein sequences encoded in the genomes of two *T. cruzi* strains: the genome reference CL-Brener strain [125] (19,668 proteins, DTU TcVI, hybrid), and the Sylvio X10 strain [130] (10,832 proteins, DTU TcI, non-hybrid). The selection criteria considered the epidemiological relevance of these representative lineages in the context of Chagas Disease, and the fact that TcVI strains are hybrids of ancestral DTUs TcII and TcIII [113, 203], resulting in most of its genes being represented in the genome by their two ancestral allelic versions [125, 204]. Therefore, by using strains from DTUs TcI and TcVI we maximized the display of relevant peptide variants with only two genomes.

Based on this analysis and the high-density peptide array capacity, we produced a tiling display of 30,500 *T. cruzi* proteins using 16mer peptides with an offset of 4 amino acids (overlap of 12 residues between consecutive peptides). The resulting array design contained 2,441,908 unique peptides, was named **CHAGASTOPE-v1** and was used for the discovery screening (see Figure 3.1). Additional details on the contents of the CHAGASTOPE-v1 array design are available in the Methods, in the Supplementary Tables S3.1, S3.2 and S3.3 and in the Supplementary File S3.8.

3.2.2 Screening reveals distinct antibody repertoires and novel antigens

Using the CHAGASTOPE-v1 array design, we performed a **discovery screening** using pooled serum samples from positive donors and paired negative sample pools from healthy subjects for 6 different geographical regions across the Americas (Figure 3.1). In total, we profiled 12 different pooled serum samples, 6 from infected subjects and 6 from healthy subjects (see Supplementary Table S3.2 for details). To test for cross-reacting epitopes, two additional pools were profiled: a pool from leishmaniasis-positive individuals and a matched pool of leishmaniasis-negative (also Chagas-negative) samples from the same geographic area. All 14 samples were assayed in duplicate, and all technical replicates had high signal correlation (see Supplementary Figure S3.1).

After normalizing the antibody-binding fluorescence signal across experiments, we reconstructed each original protein sequence, used consecutive peptides to perform signal smoothing (to remove outliers, see Methods) and generated visualizations of the antibody-binding signal for each protein and for each assayed sample (see example in Figure 3.2a and the full list in Supplementary File S3.1).

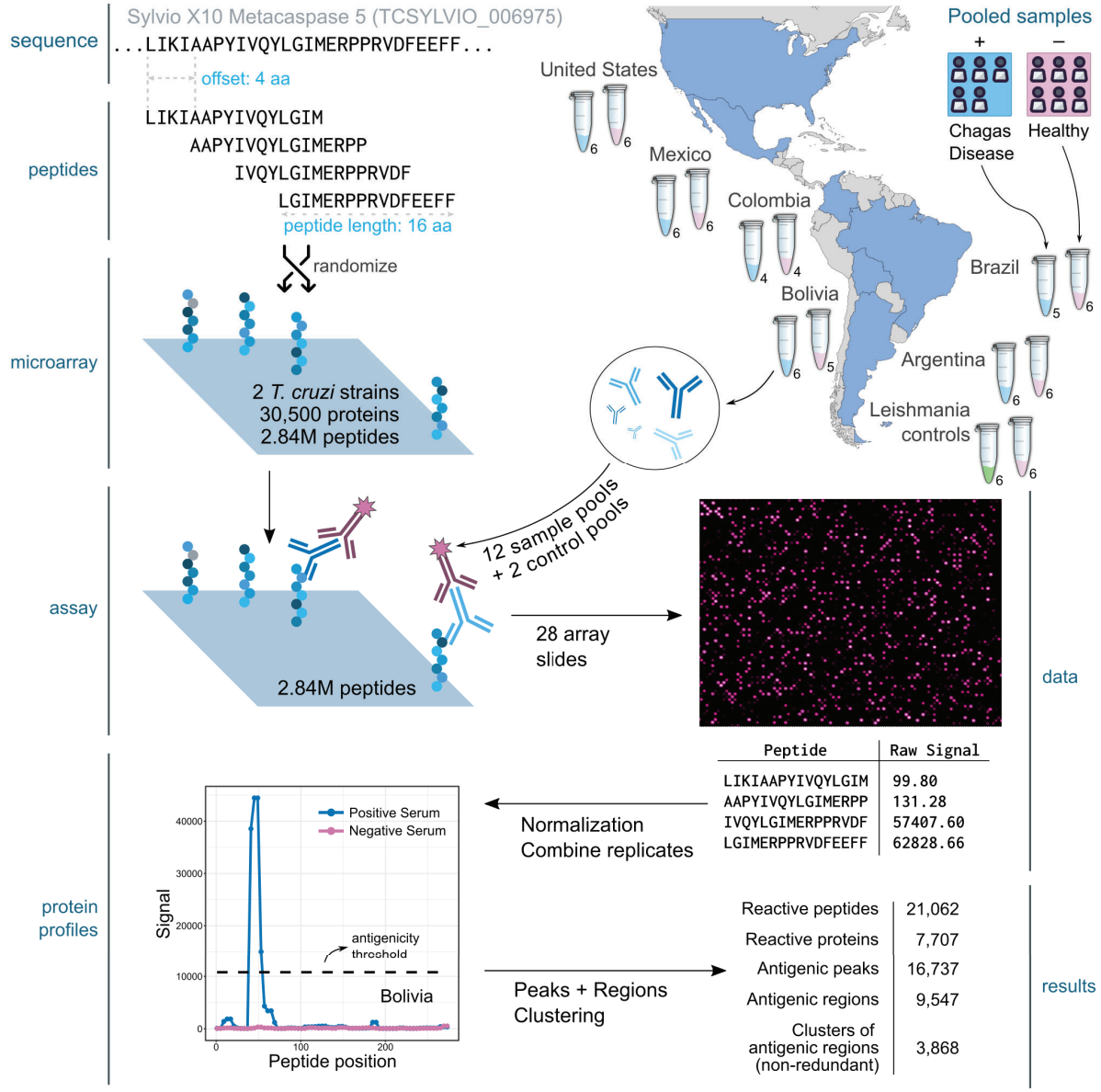


FIGURE 3.1 – *Summary of the Discovery Screening.* *The figure shows a schematic representation of the steps followed to analyze two T. cruzi proteomes (CL-Brener and Sylvio X10) using pooled serum samples across the Americas (one pool from Chagas-infected individuals and one from healthy subjects from Argentina, Brazil, Bolivia, Colombia, Mexico and the United States). The numbers below each tube represent the number of individual sera in each pool. An additional pooled serum sample of Leishmania-infected individuals and its healthy counterpart were used to study cross-reactivity. The protein used for this example is the metacaspase-5 protein from Sylvio X10 (TCSYLVIO_006975) which has 291 residues and was represented in each array using 69 peptides (16mers with an offset of 4 aa), of which only the first 4 peptides are shown.*

To identify the more reactive proteins, we compared signals obtained across experiments with pools of Chagas-negative and Chagas-positive subjects and defined an antigenicity signal threshold. We chose a very conservative threshold of 10,784.80 fluorescence units (the statistical mode plus 4 standard deviations). Any group of two or more consecutive peptides above this threshold was defined as an antibody-binding peak (see Figure 3.2), resulting in 18,199 peaks for the Chagas-positive subjects. We also observed 3,644 reactive peaks in Chagas-negative subjects (see Figure 3.3a and Supplementary Table S3.4). After removing these cross-reactive peaks we obtained 16,737 Chagas-specific antigenic peaks across 7,707 proteins.

Because some peaks were either close or partially overlapping with one another (in the same analyzed sample or across different samples, see Figure 3.2c), we combined neighboring peaks into non-overlapping antigenic regions. This resulted in 9,547 antigenic regions across both proteomes (see Methods). Furthermore, because the analyzed *T. cruzi* genomes have several large gene families, a significant number of reactive regions displayed evident sequence similarity among them. Hence, we grouped these antigenic regions into 3,868 non-redundant clusters based on sequence similarity using protein BLAST (see Figure 3.3b and Methods). The identification of reactive peptides, peaks and regions, and their cognate antigenic proteins is summarized in Table 3.1.

Since one of our objectives was finding novel antigens to diagnose or treat Chagas disease, we compared all 9,547 antigenic regions against a list of known antigens obtained from the Immune Epitope Database (IEDB). We once again used protein BLAST and found that only 16% of our regions showed considerable similarity to a known antigen (see Methods for details). We used this information to tag the regions as "similar to known antigens" or not, and expanded said tag to their corresponding clusters, resulting in 7% of our clusters having regions similar to known antigens (see Figure 3.3b).

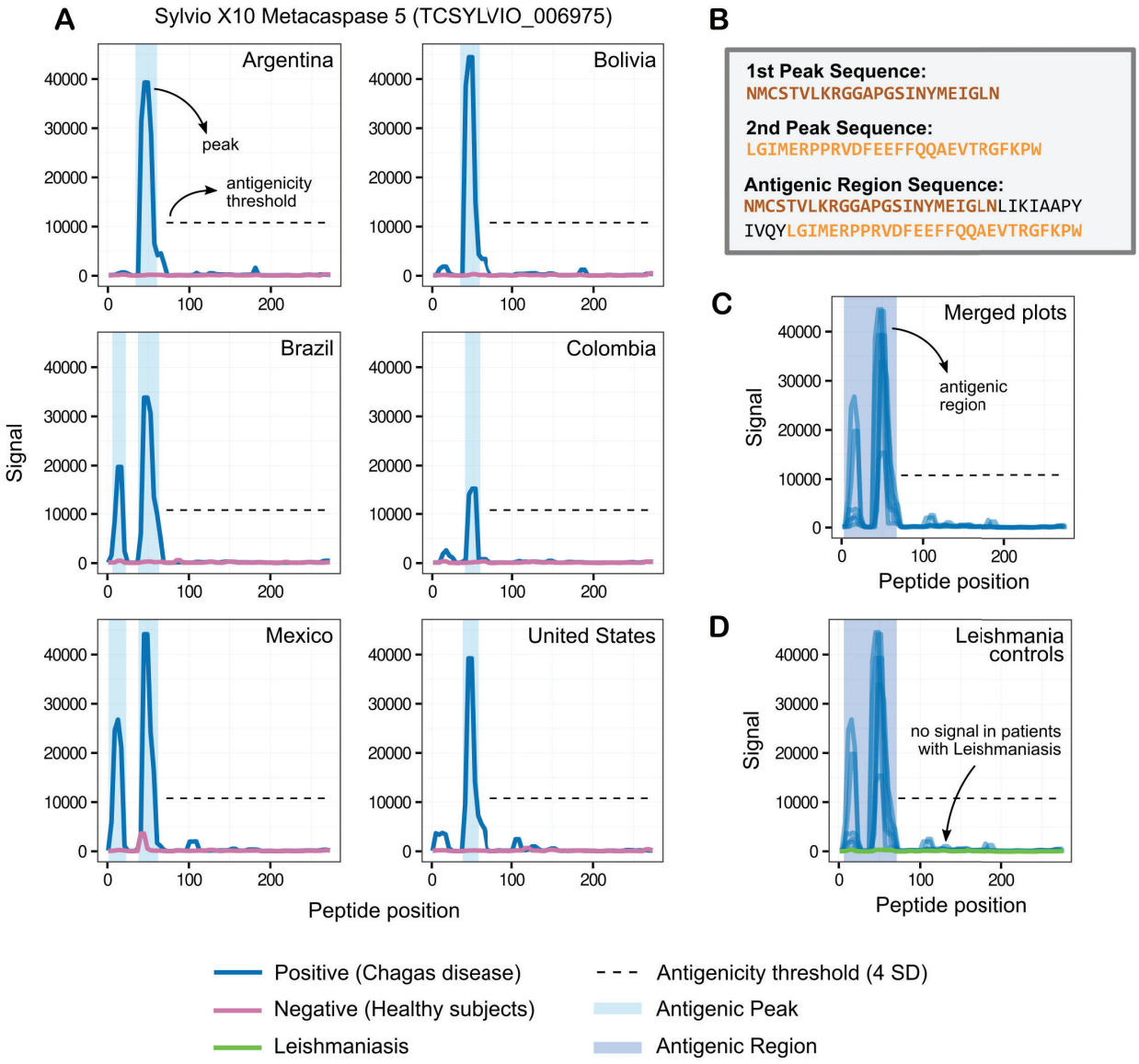


FIGURE 3.2 – ***Antibody-binding profiles, peaks, and regions.*** *The normalized fluorescence signal of each peptide in any given protein was used to produce the antibody-binding profiles for an antigen. The y-axis shows fluorescence units. The x-axis shows peptide positions along the protein sequence. Each subplot was produced using data from 4 high-density peptide arrays (2 replicas for Chagas disease subjects and 2 for matched healthy subjects, see main text). The figure serves to illustrate how we defined peaks (groups of consecutive peptides over the signal threshold) and antigenic regions (groups of neighboring peaks).* ***A)*** *Reactivity subplots for different sera pools for the Sylvio X10 metacaspase-5 protein. The antibody-binding profiles are shown in blue for the Chagas-positive sample pools (infected) and in magenta for the Chagas-negative pools (healthy). Throughout the discovery phase, we used a conservative antigenicity threshold of mode plus four standard deviations, which can be seen as a black dashed line.* ***B)*** *Peptide sequences of the reactive peaks and regions in A.* ***C)*** *Reactivity plot merging data from all sample pools.* ***D)*** *Same as C, and showing the signal for the leishmaniasis-positive serum samples (in green).*

	CL-Brener	Sylvio X10	Pangenome
Total Proteins	19,668	10,832	30,500
Total Peptides (non-redundant)	1,809,351	1,187,499	2,441,908
Reactive Peptides (non-redundant)	16,435	7,709	21,062
Antigenic Peaks	12,720	4,017	16,737
Antigenic Regions	6,710	2,837	9,547
Reactive Proteins	5,328	2,379	7,707
Clusters of Antigenic Regions (non-redundant)	-	-	3,868

Table 3.1 – ***Discovery screening finding summary.*** *This table displays the number of reactive peptides, peaks, regions and proteins found in our first screening. When analyzing the number of peptides each unique sequence was only counted once.*

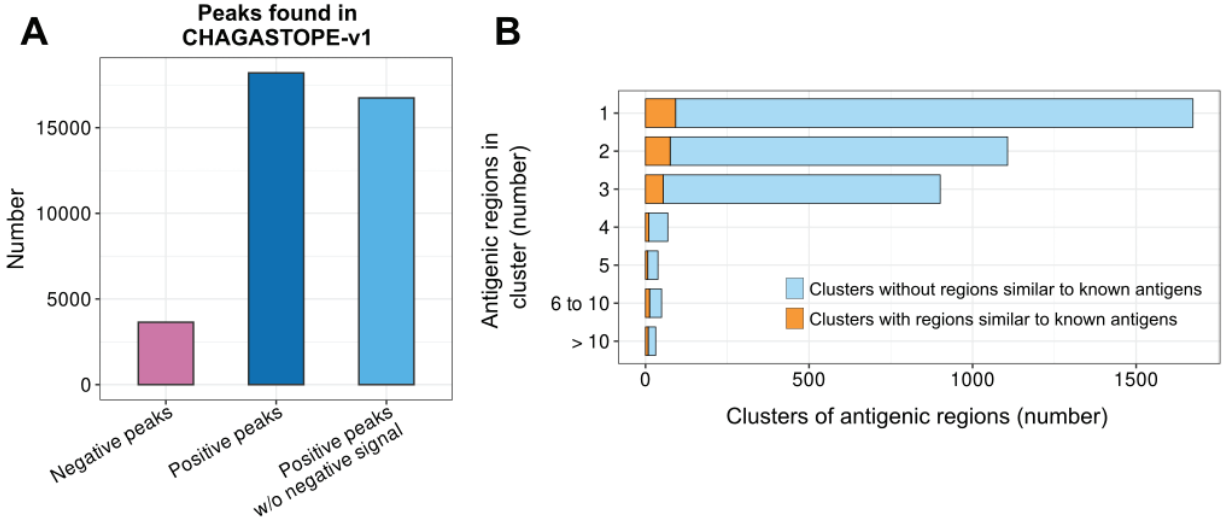


Figure 3.3 – ***Overview of antigenic peaks and regions in CHAGASTOPE-v1.*** ***A)*** *Summary of the observed peaks in Chagas-positive and Chagas-negative subjects (peaks as per the definition in main text and Methods). The rightmost column shows the number of peaks in Chagas-positive subjects that show no significant signal in the negative samples (see Methods for details).* ***B)*** *Visual summary of how the 9,547 antigenic regions are distributed in the 3,868 non-redundant clusters. To assess if a cluster had regions similar to known antigens, we compared the region's sequences against a set of T. cruzi's linear epitopes found in IEDB (see Methods).*

3.2.3 Immune responses in Chagas disease subjects are highly diverse

The complete map of measured antibody-binding reactivities across pooled samples provided a broad view of the diversity of the antibody repertoire developed in response to *T. cruzi* infections.

Before studying the seroprevalence of the clusters of antigenic regions obtained in the previous section, we set to analyze how individual antigenic peptides were shared (or not) between pooled serum samples. To achieve this, we calculated the proportion of non-redundant antigenic peptides each pooled serum shared with each other (see Methods for details). Figure 3.4 shows that no two pooled sera shared more than 30% of their combined antigenic peptides with each other, suggesting that most antigenic peptides analyzed in this first screening were exclusive for each pooled serum. As we have already shown in Figure 3.3, some of the peptides in our microarrays also showed high signal in the negative pooled sera, likely due to cross-reactivity with peptides from another pathogens. The three northernmost geographic regions were highlighted to observe if there were any differences between them and the other three geographic areas, hypothetically related to different *T. cuzi* strains present in each area, but no pattern could be seen at this stage.

We next analyzed how clusters of reactive regions were shared among pooled samples. Once again we observed a large set of non-shared reactive epitopes in each pooled sample (see Figure 3.5). These were not technical artifacts as they were reproducibly identified in the technical replicates and were also supported by the reactivity of overlapping neighboring peptides (pseudo-replicates within each experiment). This set of non-shared antigenic regions was followed by a long tail of shared epitopes with increasing seroprevalence. This observation suggests that the antibody response in Chagas disease is derived from a large and diverse set of antibody-producing B-cell clones, which is supported by what we observed in Figure 3.4.

Particularly important for serology-based applications are the clusters of antigenic regions that resulted antigenic for multiple pooled serum samples of Chagas disease subjects (166 clusters shared by at least 4 of the analyzed pooled samples), including 43 clusters that were reactive in all samples. Table 3.2 shows a selection of 19 clusters that were reactive in all samples, were not reactive against leishmaniasis-positive serum samples and did not match previously known antigenic epitopes in the Immune Epitope Database (see Methods). The seroprevalence for all 3,868 antigenic clusters can be found in Supplementary Table S3.6, while the full list of 9,547 antigenic regions with all their details can be found in Supplementary File S3.2.

We also screened the same array design against a pool of samples from leishmaniasis-positive individuals to identify cross-reacting epitopes (see Supplementary Table S3.5). Overall, there was very low cross-reactivity of *T. cruzi* peptides against this pool. Out of the 3,868 clusters of antigenic regions, only 104 (2.7%) had reactivity against this leishmaniasis sample (using the same threshold used for *T. cruzi* samples). This included 5 of the 43 clusters that had reactivity in all Chagas-positive samples, although for 3 of these only a small percentage of the regions in the cluster were cross-reactive.

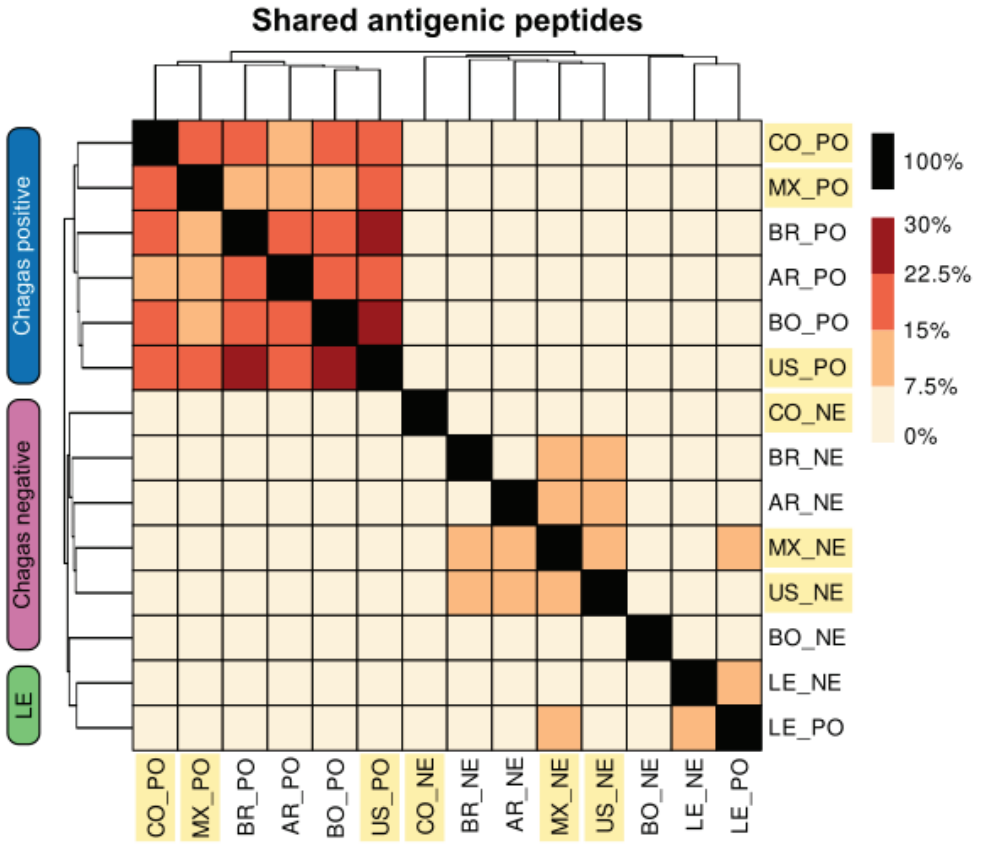


FIGURE 3.4 – ***Overview of antigenic peptides in CHAGASTOPE-v1.*** *Comparative view of the reactivity of pools from Chagas-positive subjects ("_PO") and Chagas-negative ("_NE", healthy) subjects. The heatmap shows the percentage of non-redundant antigenic (positive) peptides that were shared between a pair of serum samples (see Methods). Rows and columns were clustered by similarity, resulting in three distinct clusters as labeled at the left of the plot. Sera from individuals from the three northernmost geographic regions are highlighted. The pools from leishmaniasis-positive ("LE_PO") and negative ("LE_NE") subjects were used to analyze cross-reactivity on a later stage (see Supplementary Table S3.5). See Supplementary Table S3.2 for the codes of patient serum samples (AR = Argentina, BR = Brazil, BO = Bolivia, CO = Colombia, MX = Mexico, US = United States, LE = Leishmaniasis).*

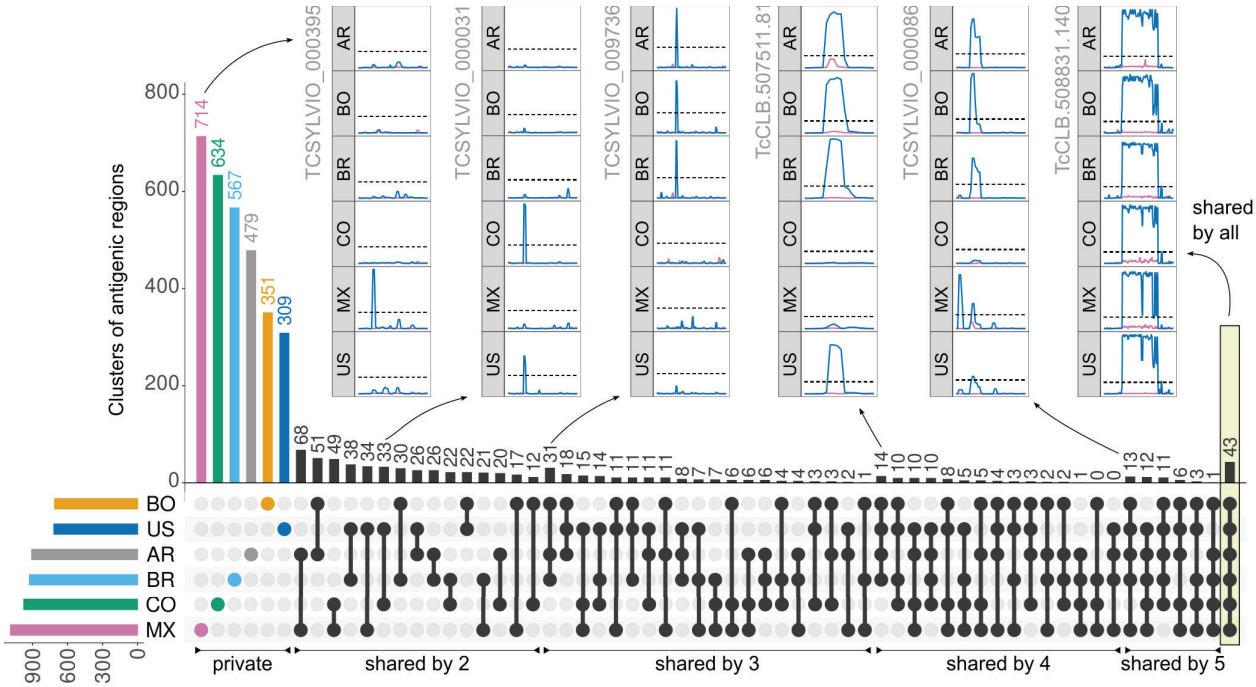


FIGURE 3.5 – ***Diversity of T. cruzi antibody-specific responses in pooled sera*** *Non-redundant clusters of reactive regions in the analyzed T. cruzi proteomes (clustered by sequence similarity) were counted on all intersections of the 6 analyzed pooled samples. A cluster was antigenic in each sera pool when at least one of its regions was antigenic in that pool. The UpSet Plot [205] displays a bar chart with these counts (top), as well as a visual depiction of all the set intersections (bottom, black). The colored bar chart at the bottom left shows the counts of total reactive regions. Pooled samples are AR = Argentina; BR = Brazil; BO = Bolivia; CO = Colombia; MX = Mexico; US = United States. The insets show antibody-binding profiles (as in Figure 3.2) for several selected antigens.*

Cluster		Best Antigenic Region in Cluster					
# Regions	Sero	Locus Identifier	Start - End	Description	Sero	Rep	Best 16mer
3	100%	TCSYLVIO_000971	89 - 144	2-oxoisovalerate dehydrogenase beta subunit, mitochondrial precursor	100%	No	GFAIGMASAGWKPIAE
6	100%	TCSYLVIO_005086	201 - 260	40S ribosomal protein S3A	100%	No	VRTPRFDAQALLNAHG
3	100%	TCSYLVIO_004283	221 - 276	60S ribosomal protein L4	100%	No	KEAMAFLKAIGAVDDV
4	100%	TCSYLVIO_009166	1 - 60	60S ribosomal protein L7	100%	No	SVPAPESAIKRAAFKR
3	100%	TcCLB.507867.80	1229 - 1312	Bait on Hook 2	83%	No	AHMFQVAQAASKNKEG
5	100%	TCSYLVIO_005782	129 - 184	Gim5A protein, glycosomal membrane protein	100%	No	PRGSCKALLPEDAEKK
2	100%	TCSYLVIO_008131	181 - 236	hypothetical protein	100%	No	RITAASADKAERFSSA
1	100%	TCSYLVIO_000132	621 - 661	hypothetical protein	100%	No	LDPAINADKPNRLRDA
2	100%	TCSYLVIO_002563	269 - 324	hypothetical protein	100%	No	CEVVPIRFNEIAAADK
2	100%	TCSYLVIO_003708	1 - 44	hypothetical protein	100%	No	VYEKFEGAVSGDKGYM
3	100%	TCSYLVIO_002923	1249 - 1304	hypothetical protein	100%	No	PVPNFVAATADKPVGT
3	100%	TcCLB.510729.160	85 - 206	hypothetical protein, conserved	100%	No	ANTKGNGVHVVVRKDI
8	100%	TCSYLVIO_006975	1 - 88	metacaspase, cysteine peptidase, Clan CD, family C13	100%	No	ERPPRVDFEEFFQQAE
2	100%	TcCLB.510583.40	29 - 80	mucin TcMUCII	100%	No	ASGAPPVPKPAAKPEV
2	100%	TCSYLVIO_001410	229 - 284	NADH-cytochrome b5 reductase	100%	No	VCGPPPFMEAISGDKD
3	100%	TCSYLVIO_001348	93 - 148	reiske iron-sulfur protein precursor	100%	No	FRKYILKPRLPEELED
2	100%	TCSYLVIO_003288	97 - 152	reticulon domain protein	100%	No	TNRWHLTSDDIHEAVN
169	100%	TCSYLVIO_004530	9 - 244	trans-sialidase	100%	No	GTNSDPDSFSSTNVSG
21	100%	TCSYLVIO_000314	109 - 164	trans-sialidase	33%	No	LYKSGKSGDKKEELIA

Table 3.2 – *Selection of antigens reactive in all Chagas-positive pooled serum samples.* *A list of the 19 clusters of antigenic regions which showed reactivity against all assayed positive pooled serum samples, did not show cross-reactivity against leishmaniasis-positive serum samples, and for which at least 90% of its regions were not similar to already known antigenic epitopes found in IEDB. For each cluster, we show a representative antigenic region with the highest seroprevalence in CHAGASTOPE-v1 for that cluster. "# Regions" = number of regions in said cluster; "Sero" = seroprevalence; "Rep" = repetitive (contains internal tandem repeats). All 3,868 antigenic clusters can be found in Supplementary Table S3.6.*

3.2.4 Identified antigens and epitopes enable a more detailed analysis

The previous screening provided ample information on the diversity and specificities of the antibody repertoire of Chagas disease patients. However, the use of pooled samples limited the analysis of the data. Next, we increased the epitope mapping resolution and assessed the reactivity of epitopes for individual patient samples.

Because 99% of peptides showed no antibody-binding at the screening stage, we removed these from the new design to focus only on the antigenic regions. Working with these smaller regions of the proteome allowed us to increase the epitope mapping resolution, using 16mers with an overlap of 15 residues between consecutive peptides.

This second peptide design was named **CHAGASTOPE-v2** and included peptides from the 9,547 protein regions that were both antigenic (reactive with Chagas-positive samples) and that showed no signal from healthy control subjects. To ensure that entire reactive peaks in each region were included in the design, we included up to 32 additional peptides from the surrounding non-reactive borders of each region (16 from each side). This contributed to the v2 array design with 241,772 unique peptides from this set. Additional details on the contents of the CHAGASTOPE-v2 array design are available in the Methods, in the Supplementary Tables S3.1, S3.2 and S3.3 and in Supplementary File S3.9.

To expand resolution of epitopes down to individual patients, we used sectorized high-density peptide arrays to assay more samples (primary antibodies) in parallel in the same slide (see Supplementary Table S3.1). In these 12-plex arrays, the same set of CHAGASTOPE-v2 peptides were replicated in each sector. A total of 12 CHAGASTOPE-v2 arrays (144 sectors) were used to analyze 71 individual serum samples in duplicate, 33 of which were part of the pools analyzed in the previous step. A set of 38 additional serum samples were analyzed from other Chagas-positive subjects from the same 6 geographic regions (see Supplementary Table S3.2). The replicas showed high Pearson correlation coefficients (>0.8, see Supplementary Figure S3.2). Because these arrays have a much lower content of non-reactive peptides, we recalculated the antigenicity threshold, resulting in a threshold of 5,814.81 relative fluorescence units (statistical mode plus 2.4 standard deviations, see Methods).

3.2.5 Diversity of individual immune responses

In our first screening we were able to find protein regions that were detected by pooled serum samples from across the Americas; however, this design had two major limitations. First, the offset between each peptide was 4 amino acids, which it can sometimes lead to multiple distinct epitopes being visually merged together. Second, the fact that we were working with pooled sera meant that it was not possible to know if an antigenic region was detected by just one individual of the pool or by all of them, which is important information for both private and shared epitopes.

In CHAGASTOPE-v2, the resolution of antibody-binding reactivity at an individual level overcame both of these limitations, providing information on the seroprevalence of each antigenic region and leading us to find different antibody-binding profiles across subjects, three of which can be seen in Figure 3.6.

Figure 3.6a shows a novel antigen with a single antigenic region (reticulon-domain containing protein, TCSYLVIO_003288). The observed seroprevalence at the level of the whole protein is aligned with the seroprevalence observed at the single-epitope level. Figure 3.6b shows a different extreme example, the hypothetical protein encoded by gene TCSYLVIO_005669, another novel antigen. This protein displayed a contiguous antigenic region composed of 7 different antibody-binding peaks (epitopes), each with a unique set of signal and seroprevalence

characteristics. In this instance the global seroprevalence for this antigen was not aligned with the seroprevalence of individual epitopes. Finally, Figure 3.6c shows the repetitive antibody binding profile of the Ag36/MAP antigen [206] for the US samples. While this is a known *T. cruzi* antigen, resolution of how antibodies from different individuals recognize these antigens provide insights to improve diagnostic reagents. As shown in the figure, the antigenic repetitive unit is not recognized in the same way by all individuals; while most individuals display reactivity for all overlapped epitopes in this repeat, some individuals (e.g., US_P3, US_E2) have antibodies with different binding preferences along this antigenic repetitive unit.

Supplementary Files S3.3 and S3.4 contain the antibody-binding profiles for all antigens and all analyzed patient samples tested in CHAGASTOPE-v2 (plots similar to those in Figure 3.6). This is a rich dataset for future studies on the serology and immune responses of Chagas disease patients and serves as a reference for other infectious diseases.

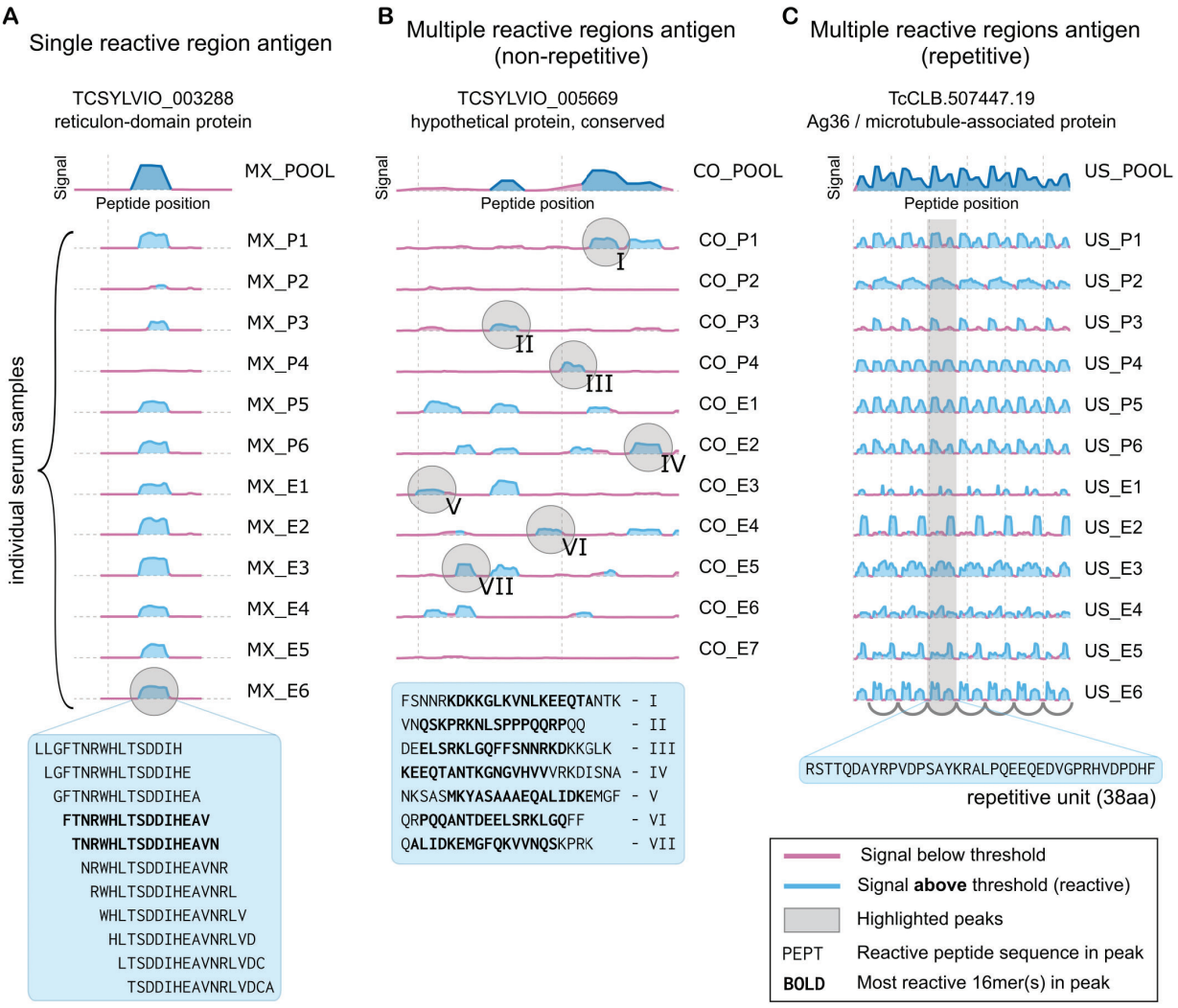


FIGURE 3.6 – *Individual patient resolution and epitope mapping of Chagas Disease antigens.* *Examples of different types of antibody-binding profiles obtained using CHAGASTOPE-v2 arrays. In all cases the reactivity of the pooled samples in the CHAGASTOPE-v1 discovery screening is shown at the top (dark blue). See Supplementary Table S3.2 for the codes of patient serum samples (MX = Mexico, CO = Colombia, US = United States).* ***A)*** *Example antigen with a single reactive region, showcasing peptides with signal above threshold (and most reactive peptides in* ***bold****).* ***B)*** *Non-repetitive antigen displaying multiple reactive peaks (marked using roman numerals), and the corresponding peak sequences below.* ***C)*** *Repetitive antigen displaying heterogeneous recognition of the repetitive unit by different Chagas-positive individuals.*

3.2.6 Individual patient resolution provides insights into seroprevalence

In the discovery screening we observed that most antigenic regions were private, meaning that they were reactive in only one pooled sample. Here, we revisited this analysis using data from the CHAGASTOPE-v2 arrays.

The first thing we did was analyze once again how individual antigenic peptides were shared (or not) between serum samples, but now using the 71 individual serum samples. The proportion of shared peptides was calculated the same way as before (see Methods for details). Figure 3.7 shows that still no two individual sera shared more than 30% of their combined antigenic peptides with each other, revealing that most antigenic peptides were exclusive for each individual serum sample. In total, we observed 88,236 non-redundant antigenic peptides, with each individual displaying reactivity to 5,841 peptides on average, which is in agreement with the previously observed large repertoire of private antigens.

We also observed that the higher proportion of shared peptides were found among individuals from the same or nearby geographic areas. This resulted in a grouping by similarity which suggests that many of the individuals from the three northernmost geographic regions (Colombia, Mexico and the United States) might be infected with a different strain of *T. cruzi* than many of the individuals from the other three geographic areas (Argentina, Brazil and Bolivia). However, at this point this is only a theory since the difference between both groups is not statistically significant.

To see if we could find more serological evidence of different strains of *T. cruzi* affecting individuals from different geographic areas, we focused on peptides found exclusively in CL-Brener or in Sylvio X10. Figure 3.8 summarizes the reactivity of all 71 subjects against these peptides. Subjects displayed reactivity to an average of 3,908 CL-Brener exclusive peptides (min: 2,606; max: 5,113, std: 445) and 941 Sylvio X10 exclusive peptides (min: 557; max: 1,249; std: 151). Comparison between the two strains was performed using Z-scores because of the difference in numbers of peptides in each set (CL-Brener is a larger hybrid genome, see Methods). Most samples from Mexico displayed higher relative reactivity against peptides from the TcI / Sylvio genome (as well as several subjects from Colombia and the United States). Other subjects, particularly from Argentina, Brazil and Bolivia showed higher relative reactivity against TcVI / CL-Brener. While the design of these arrays prevents a more detailed serotyping analysis, these serological reactivity signatures also suggest differences in the infecting *T. cruzi* lineages in these subjects.

Finally, we analyzed the 3,868 clusters of antigenic regions identified at the discovery screening (3,054 private/non-shared, 814 shared) now using 71 individual serum samples. The results of this analysis can be seen in Figure 3.9 and confirmed that the large amount of private antigenic regions was not a consequence of grouping the serum samples into pools, but a biological fact. Close to 85% of the clusters of antigenic regions were detected by less than 20% of the individuals, with most of them being detected by just 1 to 3 individuals. In contrast, we also found 98 promising clusters of antigenic regions that were detected by at least half of the Chagas-positive individuals, with 60% of them not being similar to known antigens found in IEDB (see Methods for details).

Table 3.3 provides additional information on the clusters with higher seroprevalence in CHAGASTOPE-v2, highlighting 18 clusters detected by at least 70% of the subjects that did not show cross-reactivity against leishmaniasis-positive serum samples and did not match previously known antigenic epitopes (see Methods). The list of all 3,868 antigenic clusters and their seroprevalence in CHAGASTOPE-v2 can be found in Supplementary Table S3.7, while the full list of 9,547 antigenic regions with all their details can be found in Supplementary File S3.5.

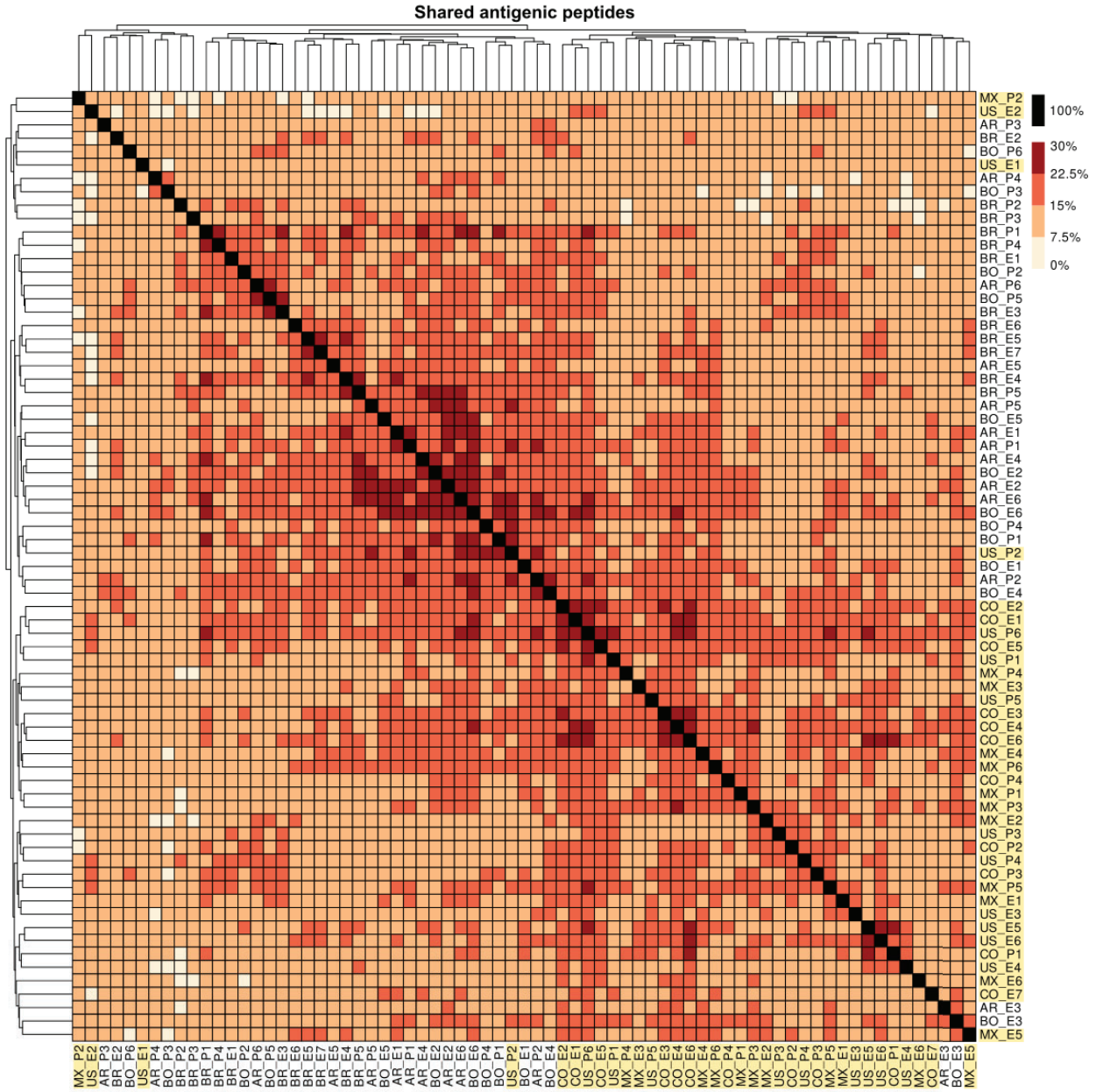


Figure 3.7 – ***Overview of antigenic peptides in CHAGASTOPE-v2*** *Comparative view of the reactivity of individual sera from Chagas-positive subjects. The heatmap shows the percentage of non-redundant antigenic peptides that are shared between a pair of serum samples (see Methods). Rows and columns were clustered by similarity and sera from individuals from the three northernmost geographic regions are highlighted. See Supplementary Table S3.2 for the codes of patient serum samples (AR = Argentina, BR = Brazil, BO = Bolivia, CO = Colombia, MX = Mexico, US = United States).*

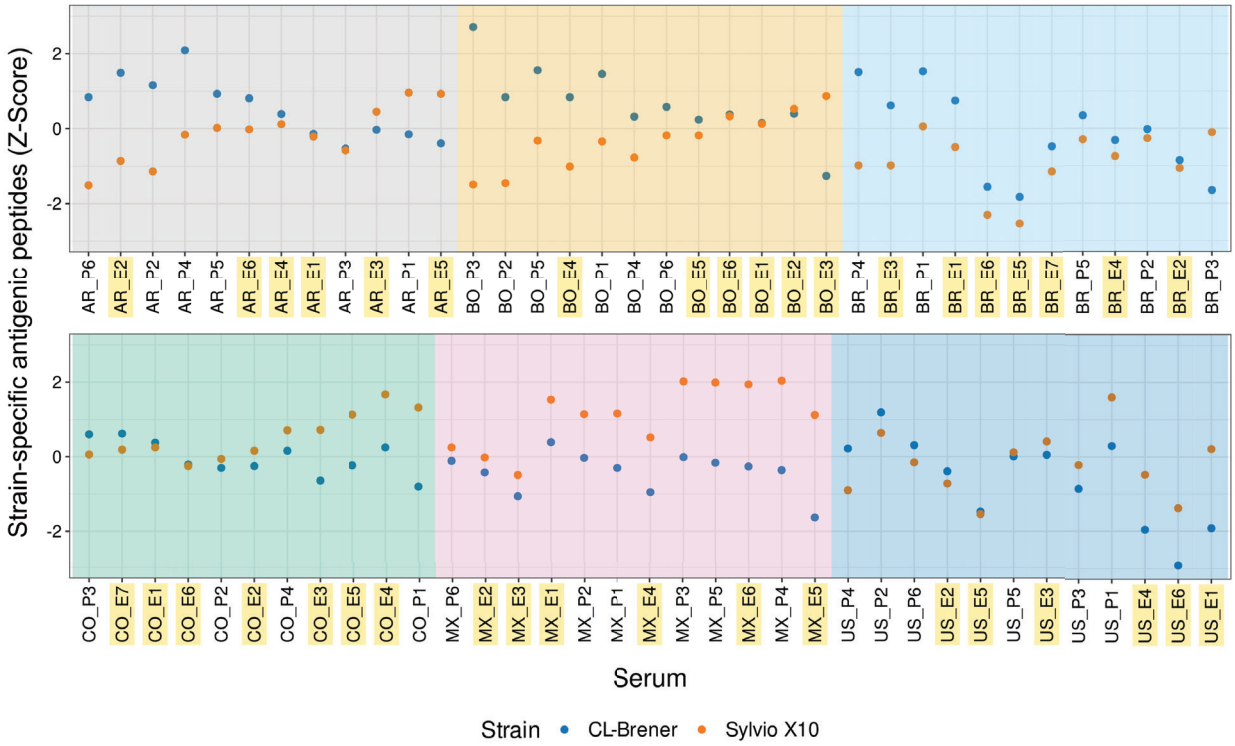


FIGURE 3.8 – *Comparison of strain-specific antigenicity across subjects.* *Counts of each subject's strain-specific reactive peptides (those present in only one strain and with signal above the antigenicity threshold) were standardized using Z-scores. Standardization was necessary because CL-Brener and Sylvio X10 strains have different numbers of encoded proteins (see Methods). Z-scores above 0 and below 0 represent higher and lower relative number of strain-specific reactive peptides, respectively. See Supplementary Table S3.2 for the codes of patient serum samples (AR = Argentina, BR = Brazil, BO = Bolivia, CO = Colombia, MX = Mexico, US = United States). Interpretation of the reactivity observed for the individual sera containing "_E" in their sample name must be done with care because these samples were not used for antigen discovery and selection.*

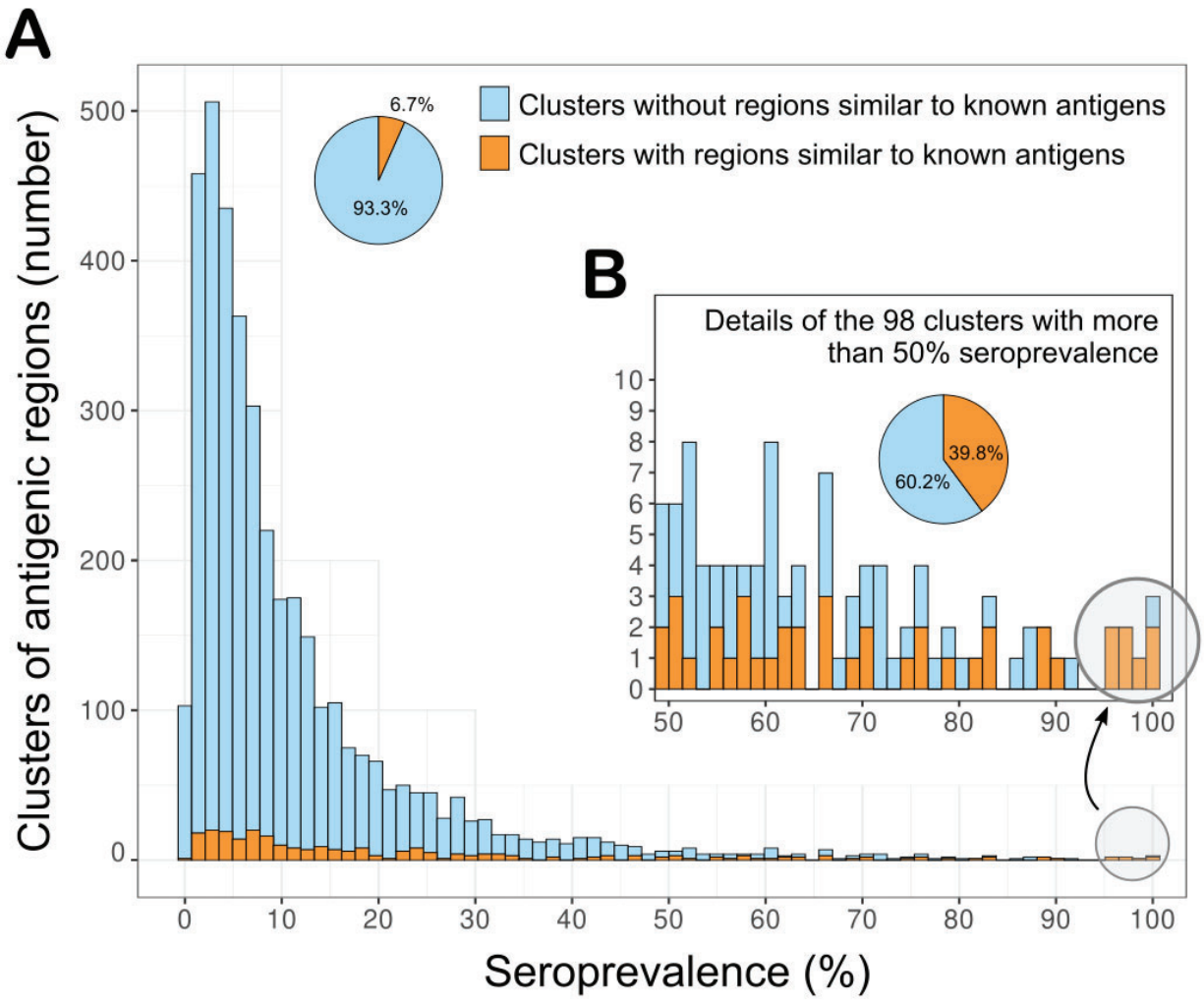


Figure 3.9 – ***Profiling of individual antibody responses reveals the diversity of anti-T. cruzi specific antibodies.*** *Non-redundant clusters of reactive regions (clustered by sequence similarity) were counted on all intersections of the 71 analyzed individual samples. To assess if a cluster had regions similar to known antigens, we compared the region's sequences against a set of T. cruzi's linear epitopes found in IEDB (see Methods).* ***A)*** *Histogram showing the number of clusters that were reactive in each fraction of subjects (shown as seroprevalence).* ***B)*** *Inset zooming in on the 98 clusters reactive in at least 50% of the subjects.*

Cluster		Best Antigenic Region in Cluster					
# Regions	Sero	Locus Identifier	Start - End	Description	Sero	Rep	Best 16mer
169	100%	TCSYLVIO_004530	9 - 244	trans-sialidase	99%	No	GTNSDPDSFSSTNVSG
3	92%	TcCLB.509007.70	1249 - 1308	protein phosphatase 2C	90%	No	AVPNFAAATADKPVGT
48	87%	TcCLB.506289.240	161 - 312	hypothetical protein	55%	No	EGKDERKSGEATAPQV
8	87%	TcCLB.507297.30	1 - 96	metacaspase	86%	No	ERPPRVDVEEFFQQAE
10	83%	TcCLB.506597.33	145 - 193	hypothetical protein	80%	No	LRQIDASPPEPFTAAP
3	80%	TCSYLVIO_004283	221 - 276	60S ribosomal protein L4	80%	No	KEAMAFLKAIGAVDDV
2	79%	TCSYLVIO_009331	481 - 536	structural maintenance of chromosome protein 4	79%	No	LRKIDAATERNGNLVA
1	77%	TcCLB.506941.194	1025 - 1080	hypothetical protein, conserved	77%	No	FRKIDAAVPVPNTSYA
1	76%	TcCLB.507809.119	81 - 140	Autophagy protein Apg6	76%	No	VLRKIEEEFQQLEEQK
2	76%	TCSYLVIO_009815	281 - 336	hypothetical protein	76%	No	FRQIDTEHNDRITAEQ
2	75%	TCSYLVIO_001410	229 - 284	NADH-cytochrome b5 reductase	75%	No	PPFMEAISGDKDFKTS
2	73%	TCSYLVIO_003288	97 - 152	reticulon domain protein	73%	No	LTSDDIHEAVNRLVDC
3	72%	TcCLB.506441.20	641 - 888	cytoskeleton associated protein	70%	Yes	REGRERGYPEEKEDSR
1	72%	TcCLB.506925.460	41 - 92	kinetoplastid kinetochore protein 16	72%	No	LRMIDELAAGVEMWKQ
3	72%	TCSYLVIO_003590	89 - 140	microtubule-associated protein Gb4	72%	No	LREIDDVENHASQSRA
23	72%	TcCLB.510157.10	281 - 332	Mucin-associated surface protein (MASP), subgroup S022	45%	No	AASANKYDTVPQSAGS
3	70%	TcCLB.509139.20	37 - 88	hypothetical protein, conserved	70%	No	SGGAPPTRGGFGAGTS
2	70%	TcCLB.509791.120	65 - 172	kinetoplast-associated protein	70%	No	YDVSKPLDVEKEISKA

Table 3.3 – *Selection of antigens reactive in most Chagas-positive individual serum samples.* *A list of the 18 clusters of antigenic regions which showed reactivity against at least 70% of the positive individual serum samples, did not show cross-reactivity against leishmaniasis-positive serum samples in the previous assay, and for which at least 90% of its regions were not similar to already known antigenic epitopes found in the IEDB. For each cluster, we show a representative antigenic region with the highest seroprevalence in CHAGASTOPE-v2. "# Regions" = number of regions in said cluster; "Sero" = seroprevalence; "Rep" = repetitive (contains internal tandem repeats). All 3,868 antigenic clusters can be found in Supplementary Table S3.7.*

3.3 Discussion

We have produced a detailed characterization of the antibody specificities against linear epitopes in patients with Chagas disease. Previous studies have shown that some pathogens deliberately induce short-lived, polyclonal plasma cells to dilute long-lived, specific antibody responses [207]. In contrast, *Trypanosoma cruzi* induces a massive clonal expansion of B-cells during the acute phase leading to production of parasite-specific and autoreactive antibodies as well as high levels of antibodies with unknown specificity [208, 209]. Our description of a large and diverse number of specificities, composed of mostly non-shared epitopes (low seroprevalence at the population level), support these previous observations, and provide information on the targets of these antibodies.

The power of these massively parallel serological assays lies in the delineation of the responses of individual patients to each identified epitope, hence producing a rich human seroprevalence matrix for these antigens. By grouping similar sequences, our analysis uncovered >3,800 non-redundant clusters of antigenic regions, both public (shared) Chagas antigens as well as private individual anti-*T. cruzi* responses, in 71 individuals from different human populations across the Americas. Of these clusters, 98 were detected by at least 50% of the individuals, making them of great interest for diagnosing and treating Chagas Disease. We compared these antigenic regions against antigens in IEDB and found that 60% of those promising clusters showed no considerable similarity to previously known antigens.

While it was not part of our initial objectives, we also used our data to also look for patterns connecting different strains of *T. cruzi* with the serological profiles of the different individuals. By analyzing the antigenic peptides shared between the different sera, we observed that individuals from the three northernmost geographic regions (Colombia, Mexico and the United States) showed closer patterns of similarity when compared against individuals from the other three geographic regions. Since different geographic areas tend to have different strains of *T. cruzi*, this could suggest that we were seeing peptides that are found exclusively in some of those strains. We also analyzed if each sera recognized more peptides from CL-Brener or from Sylvio X10. There was a tendency of the three southernmost geographic regions (Argentina, Brazil and Bolivia) to show a higher preference for peptides in CL-Brener than the rest, while the three northernmost geographic showed a preference for Sylvio X10 (specially Mexico); this would coincide with the known distribution of both strains in the Americas. While the patterns seemed to be there, neither of these two experiments showed a significant difference between the groups being studied, and further analysis is needed to draw stronger conclusions.

A consequence of the experimental platform we used to detect antibody-binding, is the inherent bias towards linear epitopes (likely missing most conformational epitopes). This is evident in our failure to detect antibody binding to at least one known bona fide antigen: Ag1 [206], also known as FRA or JL7 [105], a cysteine protease (clan CA, family C2, CL-Brener Locus ID: TcCLB.505985.9) which is a component of commercial kits for the diagnosis of *T. cruzi* infection. No reactivity was observed in our short peptide screenings, suggesting that the epitope(s) in this antigen may be conformational.

To identify cross-reactive epitopes against a co-endemic disease caused by a related trypanosomatid parasite, we also assayed a pool of leishmaniasis-positive serum samples against *T. cruzi* peptides from complete proteomes. This is important, as many Chagas disease false positives are suspected to be a consequence of *Leishmania spp.* cross-reactivity [210]. We identified 888 *T. cruzi* peptides that were cross-reactive with these leishmaniasis samples (data in Supplementary Table S3.5). While the number of leishmaniasis samples in the pool may be small (n = 6), the ability to screen at this scale outweighs this potential limitation.

When analyzing if an antigenic region was similar to a previously known antigen, we aimed to be conservative but fair. We used BLASTP to measure similarity, requiring a large percentage of identical matches before tagging an antigenic region as "known", so as to avoid excluding a potential novel epitope from our data. While this increases the likelihood that the regions being tagged as "known" are true positives, it also makes it possible for some of the antigenic regions labeled as "unknown" to be similar to previously known antigens when looking solely at their key epitopes. Since analyzing the key epitopes for all our antigenic regions would require more information that we had (ideally an alanine scan for each of the 3,868 regions), we decided to err on the side of safety and leave those antigens as "unknown". A more detailed manual curation of any antigen is needed before guaranteeing its novelty.

To our knowledge, this Atlas is the largest collection of Chagas disease antigens and epitopes described to date, and the first dataset providing fine resolution of seroprevalence to epitopes in humans. Due to the breadth and diversity of the clinical samples analyzed, this study also provides a large set of experimentally validated negative data (non-antigenic proteins and peptides). This is almost always overlooked, but it represents a highly valuable dataset for training of predictors, which often need to work under the assumption that proteins with no previous information on their antigenicity are non-antigenic [141, 156]. The datasets from the primary discovery screening also provide a large corpus of data on dominant *T. cruzi* peptides reactive to sera from healthy subjects from different human populations.

The observations and the data produced in this study reflect a snapshot in time of the antibody repertoires of each subject. Many questions about these repertoires remain. What is the nature of the private (non-shared) set of antibody specificities? Which epitopes are the targets of short-lived responses? And which are the targets of long-lived responses? The observed low seroprevalence of a large fraction of antigens (non-shared responses) may be explained if this is a fluctuating repertoire. It is thus tempting to speculate that private antigens (as described in this work) may be the target of short-lived or weak antibody responses. Under this scenario, the B-cell clones producing these antibodies may decay after some time, and thus the observed feature of being unique to one or very few individuals in these snapshots may be the telltale of these short-lived immune responses. This agrees with the current view of the complex and focal dynamics of Chagas disease in the host, where waves of parasite bursts from different foci at different moments may direct the immune response to antigens that are uniquely expressed or exposed in different tissue environments [211–213].

The rich set of biomarkers in this Atlas also provide essential information to study the dynamics of the more prevalent (shared) antibody specificities at an unprecedented depth and granularity. In chronic Chagas disease, antibody levels are maintained by the persistence of parasites and antigens [214]. Studying how the antibody repertoires fluctuate upon reduction of parasite loads or elimination, or during development of Chagas disease pathology is a clear path to discover markers for novel immunoassays.

The Human Chagas Antigen and Epitope Atlas is a reference resource that is freely accessible, both as an interactive website and as downloadable data (see Supplementary Materials and Chapter 4 for more details). The resource generated and described herein comprises the collection of antigenic regions of the *Trypanosoma cruzi* pangenome, as revealed by analyzing the anti-*T. cruzi* human antibody repertoires of 71 Chagas Disease patients. These individual antibody repertoires described in detail represent a foundational resource for the community that will serve as a major accelerator for the development of new diagnostics, serology-based immunoassays, vaccines, and to study the dynamics of adaptive immune responses at high resolution.

3.4 Materials and methods

3.4.1 Array Designs

CHAGASTOPE-v1 Design used for antigen and epitope discovery.

Two *T. cruzi* proteomes were used in this design: Sylvio X10 [129] and CL-Brener [125], both retrieved from TriTrypDB Release 5 (2016) [215]. To produce a tiling display of peptides, we first merged sequences from both proteomes and parsed all proteins, removing those shorter than 16 amino acids and duplicates (based on protein ID). This resulted in 30,500 proteins: 10,832 for Sylvio X10 and 19,668 for CL-Brener. Next, we split proteins into peptides of length 16 with an offset of 4 amino acids between consecutive peptides (meaning that there was an overlap of 12 amino acids between those peptides). At this stage we removed duplicate peptide sequences thus each peptide was placed only once in the microarray. This set of 2,441,908 unique peptides is available as Supplementary File S3.8 and as part of the submission to the ArrayExpress Database. Finally, we added other peptides from a number of sources to the CHAGASTOPE-v1 design, such as positive controls that corresponded to previously identified antigenic regions [147], and peptides from other pathogens that are seroprevalent in humans (e.g., cytomegalovirus). These positive controls were repeated 4 times across the array as peptides of length 15 with an offset of 1 amino acid to have a higher resolution for epitope mapping and to match the original conditions of past works. This, along with the trimming of a few peptides in the synthesis (see Array Synthesis below) resulted in a final array design containing 2,842,420 peptides, all of which were present only once in the microarray except for the positive controls. These peptides were assigned randomly to spots in the microarray.

CHAGASTOPE-v2 Design used for characterization of antigenic regions.

Because the aim of this second design was to analyze a smaller subset of peptides in higher detail, we focused on the 9,547 antigenic regions found from the first design. We produced tiling displays of peptides spanning these regions, using peptides of length 16 with an offset of 1 amino acid (maximal resolution for epitope mapping), which resulted in 242,154 unique peptides. These peptides can be found in Supplementary Table S3.9 and as part of the submission to the ArrayExpress Database. Finally, in this array design we also included additional peptide variants not analyzed in this thesis, such as peptides from the *T. cruzi* 231 strain [216] (DTU TcIII), as well as a number of detailed mutagenesis scans (AlaScan) of selected epitopes, among others. While not present in this thesis, the analysis of the mutagenesis scan can be seen in the original paper. The final array design contained 392,299 addressable peptide spots and peptides were assigned randomly to these spots. This design was used to drive synthesis of QX12 (12-plex) arrays, where the same CHAGASTOPE-v2 design was replicated across all 12 sectors of the array (assayed individually).

3.4.2 Array Assays

For the antigen and epitope discovery screening using the CHAGASTOPE-v1 design we produced and assayed 28 1-plex high-density peptide arrays. For the epitope characterization, mapping and seroprevalence study using the CHAGASTOPE-v2 design, we produced and assayed 12 (twelve) 12-plex sectorized high-density peptide arrays. Supplementary Table S3.1 provides an overview of 1-plex and 12-plex arrays used in this work, while Supplementary Table S3.3 provides a list of arrays slides and samples used in each assay.

Array Synthesis

The CHAGASTOPE array designs were synthesized at Roche Sequencing Solutions (Peptide Lab, Madison WI, now Nimble Therapeutics) with a Roche Sequencing Solutions Maskless Array Synthesizer (MAS) by light-directed solid-phase peptide synthesis using an amino-functionalized support (Greiner Bio-One) coupled with a 6-aminohexanoic acid linker and amino acid derivatives carrying a photosensitive 2-(2-nitrophenyl) propyloxycarbonyl (NPPOC) protection group (Orgentis Chemicals). Amino acids (final concentration 20mM) were pre-mixed for 10 min in N,N-Dimethylformamide (DMF, Sigma Aldrich) with N,N,N',N'-Tetramethyl-O-(1H-benzotriazol-1-yl)uronium-hexafluorophosphate (HBTU, Protein Technologies, Inc.; final concentration 20mM) as an activator, 6-Chloro-1-hydroxybenzotriazole (6-Cl-HOBt, Protein Technologies, Inc.; final concentration 20mM) to suppress racemization, and N,N-Diisopropylethylamine (DIPEA, Sigma Aldrich; final concentration 31mM) as base. Activated amino acids were then coupled to the array surface for 3 min. Following each coupling step, the microarray was washed with N-methyl-2-pyrrolidone (NMP, VWR International), and site-specific cleavage of the NPPOC protection group was accomplished by irradiation of an image created by a Digital Micro-Mirror Device (Texas Instruments), projecting 365nm wavelength light. Coupling cycles were repeated to synthesize the full in silico-generated peptide library.

Coupling cycles were limited to avoid extremely long synthesis times, which had the consequence of trimming some peptides in our design by a few amino acids (usually peptides where a single amino acid appeared many times). This occurred in 0.5% of the peptides in the first design and 1.4% of the peptides in the second one, with an average of 1.5 and 1.7 amino acids trimmed in each case respectively. Because this was a rare event, because the trimming removed only one or two amino acids, and because we also smoothed the signal data using a rolling median technique (see below), we assumed this trimming had no substantial impact on analysis of the data.

Sample Binding and Detection

Prior to sample binding, final removal of side-chain protecting groups was performed in 95% trifluoroacetic acid (TFA, Sigma Aldrich), 0.5% Triisopropylsilane (TIPS, TCI Chemicals) for 30 min. Arrays were incubated twice in methanol for 30 s and rinsed four times with reagent-grade water (Ricca Chemical Co.). Arrays were washed for 1 min in TBST (1× TBS, 0.05% Tween-20), washed twice for 1 min in TBS, and exposed to a final wash for 30 s in reagent-grade water.

Serum samples were diluted 1:100 in binding buffer (0.01M Tris-Cl, pH 7.4, 1% alkali-soluble casein, 0.05% Tween-20) and bound to arrays overnight at 4°C. After sample binding, the arrays were washed three times in wash buffer (1× TBS, 0.05% Tween-20), 10 min per wash. Primary sample binding was detected via Alexa Fluor®647-conjugated goat anti-human IgG secondary antibody (Jackson ImmunoResearch #109-605-098). The secondary antibody was diluted 1:10,000 (final concentration 0.1 ng/µl) in the secondary binding buffer (1x TBS, 1% alkali-soluble casein, 0.05% Tween-20). Arrays were incubated with secondary antibody for 3 h at room temperature, then washed three times in wash buffer (10 min per wash), washed for 30 sec in reagent-grade water, and then dried by spinning in a microcentrifuge equipped with an array holder. Fluorescent signal of the secondary antibody was detected by scanning at 635 nm at 2 µm resolution and 15% gain, using an MS200 microarray scanner (Roche NimbleGen). Scanned array images were analyzed with proprietary Roche software to extract fluorescence intensity values for each peptide.

3.4.3 Human Serum Samples

Human serum samples from *T. cruzi*-infected patients and matched negative subjects used in this study were part of the collections of the Laboratorio de Enfermedad de Chagas, Hospital de Niños "Dr. Ricardo Gutierrez" (HNRG, Buenos Aires, Argentina) (AR; n=18); Fundación CEADES (Cochabamba, Bolivia) (BO; n=17); Protozoology Laboratory (LIM 49), Hospital das Clínicas, Faculdade de Medicina, Universidade de São Paulo (São Paulo, Brazil) (BR; n=18); Instituto Nacional de Salud Pública (Tapachula, Mexico) (MX; n=18); Instituto de Cardiología (Bucaramanga, Colombia) (CO; n=15); University of South Carolina (South Carolina, USA) (US; n=18). Human serum samples from patients with American tegumentary leishmaniasis (ATL) and matched negative subjects were from the Instituto de Patología Experimental, Universidad Nacional de Salta (IPE, Salta, Argentina) (LE; n=12). A list of samples and their code identifiers as used in this work are provided in Supplementary Table S3.2.

Chagas Disease patients were in the asymptomatic chronic stage of the disease without cardiac or gastrointestinal compromise (age range: 15 to 96 years old, median: 48). Serum samples were collected from clotted blood obtained by venipuncture and analyzed for *T. cruzi*-specific antibodies with the following commercially or in-house kits: AR: Chagatest ELISA lysate (Laboratorios Wiener, Argentina), Chagatest HAI (Laboratorios Wiener, Argentina), Chagatest ELISA recombinant v4.0 (Wiener Lab, Argentina); BO: recombinant and lysate ELISA; BR: conventional in-house ELISA (confirmed by TESA Blot [217]); CO: ELISA (Chagatek, Organon, Argentina) and HAI (Chagatest, Weiner, Chile); MX: Bio-Rad Chagascreen Plus v4 (recombinant), Test ELISA para Chagas III (Grupo Bios, Chile), Accutrak Chagas Microelisa Test System (Laboratorio Lemos, Argentina), Accutrack Chagatek ELISA recombinante (Laboratorio Lemos, Argentina); US: Chagas Stat-Pak (Chembio, Medford, NY), Hemagen Chagas' Kit ELISA (Hemagen Diagnostics Inc., Columbia, MD), and an in-house TESA-Blot [217]. ATL samples were classified using an in-house ELISA based on crude antigen extracts from promastigotes and amastigotes of 3 different endemic Leishmania species and 2 reference strains [140].

All procedures followed the Declaration of Helsinki Principles. Written informed consent was obtained from all individuals (or from their legal representatives), and all samples were decoded and de-identified before they were provided for research purposes. The procedures were approved by the following ethics committees: Hospital de Niños "Ricardo Gutierrez" (#CEI 14.14); Fundación CEADES (CE-CEADES-4-12-2018; IRB 0990-0279; FWA: 00024189); Comité de Etica en Investigación, Instituto Nacional de Salud Pública, Mexico (CI: 1369, Registro ante CONBIOÉTICA: 17CEI00420160708, Registro ante COFEPRIS: 13 CEI 17 007 36, FWA: 00015605), Comité de Ética en Investigación FOSCAL, Bucaramanga, Colombia (CEI 21/11/14) and Fundación Cardioinfantil (Acta 512/2015), the study sites of the CHICAMOCHA3-equity trial; Baylor College of Medicine (Houston, TX, USA) (#H-35471 and #H-32321), the Gulf Coast Regional Blood Center (Houston) (#13-002), and the South Texas Tissue and Blood Center (San Antonio, TX, USA); and Comisión Provincial de Investigaciones Biomédicas, Ministerio de Salud Pública, Gobierno de la Provincia de Salta, Argentina (Expte 321 - 136934/2018). Samples from LIM 49 were part of an older collection of samples (8-10 years) and qualify as secondary research use of biospecimens for which informed consent was not required. These fall into exemption #4 in the list of exemptions for the requirement of informed consent developed by the Office for Human Research Protections (OHRP), US Department of Health & Human Services, as it did not involve new recruitment of human participants and samples did not include any direct identifier.

3.4.4 Normalization, quality control and removal of outliers (smoothing)

CHAGASTOPE-v1 - Quality control

All experiments were performed in duplicate (same biological sample, duplicate assays on independent array slides). We performed quality control of each pair of replicates for each sample using Bland-Altman (MA) plots and reciprocal signal plots. All replicate array assays showed excellent overall reproducibility (see Supplementary Figure S3.1). As another step of quality control, we analyzed the replicas of the positive controls we placed in the array (these were sections of known antigenic proteins that had their peptides repeated 4 times in the microarray design). For this we used their normalized signals (see below) and we observed excellent overall reproducibility between the replicas, both intra- and inter-array (see Supplementary File S3.10).

CHAGASTOPE-v1 - Quantile normalization

To compare data across experiments, we normalized array data using quantile normalization. Because this method requires similar statistical properties of the underlying distributions, we performed two sets of quantile normalizations, one for the assays using Chagas-positive samples, and one for the assays using negative samples (including those from leishmaniasis-positive serum samples, which produced signal distributions similar to the Chagas-negative samples). We treated replicas as independent samples, resulting in a normalization across 12 array sets for the Chagas-positive samples and a normalization across 16 array sets for the rest. Normalization was performed in R using the function *normalize.quantiles* from the package *preprocessCore*.

CHAGASTOPE-v1 - Smoothing and replicas

To remove outliers, we used two methods combined: a rolling median smoothing procedure and an average via replicas. First, we assigned the normalized signal of each peptide to the corresponding protein sequences. This was done once per serum sample per replica. Next, we used the *rollmedian* function in the R package *zoo* to calculate the rolling median along each protein sequence. We used a window size of 3, meaning that the smoothed signal for each peptide was the median of itself and its two neighboring peptides in the same protein/serum/replica (for peptides at the edges of the protein sequences we added a 0 as the signal of the non-existing neighboring peptide). Next, we combined data from the two replicas to calculate their average and standard deviation, resulting in the final data set that was analyzed and described herein. In this final data set each peptide had 14 associated signal values (6 from Chagas-positive samples, 6 from Chagas-negative samples, 1 from a leishmaniasis-positive sample and 1 from a leishmaniasis-negative sample). These signals can be found in Supplementary File S3.11.

CHAGASTOPE-v2 - Quality control, quantile normalization and smoothing

In the CHAGASTOPE-v2 arrays we followed similar steps as in the first design. **Quality control:** we analyzed each pair of replicates for each sample using Bland-Altman (MA) plots and reciprocal signal plots (see Supplementary Figure S3.2). **Quantile normalization:** In these 12-plex assays, one microarray slide contained 12 sectors, which were assayed separately, hence for all data-analysis purposes 1 array sector was treated as one 1 array data set. Quantile normalization was thus performed for 142 assays (2 replicas for each of 71 individual serum samples). **Smoothing and replicas:** The smoothing and combining of replicas was done exactly as in CHAGASTOPE-v1 and the signals can also be found in Supplementary File S3.11.

3.4.5 Definition of antigenic peaks and regions

CHAGASTOPE-v1 - Antigenicity threshold

To define this threshold, we analyzed the normalized signals (before smoothing) for peptides in the 2 replicas of each of the 6 Chagas-positive and the 6 Chagas-negative pooled samples (totalling 24 samples) and calculated their mode and standard deviation. We then looked at the protein profiles (the smoothed signals) and analyzed the dispersion of the "noise", meaning how high were the signals for the healthy pools. Because this was a discovery screening, we wanted to be very conservative in this choice, making sure to select regions that were truly antigenic. All this, coupled to the amount of space available in our second design, led to us using an antigenicity threshold of 10,784.80 arbitrary fluorescence units (mode plus 4 standard deviations).

CHAGASTOPE-v1 - Peaks

For the discovery screening, we defined as "peak" a group of two or more consecutive peptides with signals greater than the antigenicity threshold. Because we were interested in the discovery of *T. cruzi* antigens and epitopes, we also required each peak to have a maximum signal in a Chagas-positive sample that was at least five times higher than the corresponding maximum signal in the negative samples for those peptides.

CHAGASTOPE-v1 - Regions

Antigenic Regions result from merging of neighboring peaks. For each peak we noted the position in its protein of the first and last peptides. We then expanded this range by moving the start of the peak 16 amino acids to the left and the end of the peak 16 to the right to ensure capturing the whole peak. Then, if two or more of these new "wide peaks" overlapped between each other they were all merged into one. This resulted in 9,547 antigenic regions across both proteomes.

CHAGASTOPE-v1 - Clusters of regions

In our analyzed proteomes there were identical proteins or different proteins sharing significant sequence similarity over a domain or defined sequence region. This was either because they belonged to the same protein family or because the protein was present in both CL-Brener (Esmeraldo and Non-Esmeraldo haplotypes) and Sylvio X10 proteomes. This similarity resulted in several antigenic regions with very similar or identical sequences, which can distort the conclusions drawn from the data. To reduce redundancy, we clustered antigenic regions by sequence similarity using `blastp` (BLAST 2.2.31+ [155]). The "all vs all" comparison across all 9,547 regions was run with the following `blastp` command options and parameters:

```
blastp -db file.FASTA -query file.FASTA -outfmt '6 qseqid sseqid pident
length mismatch gapopen evalue bitscore qseq sseq sstart send'
-word_size 2 -comp_based_stats 0 -max_target_seqs 50000 -matrix BLOSUM80
```

We then kept only matches with a percentage of identical amino acids (*pident*) of at least 80% and a match length of at least 75% of the length of the shortest region in the match. Using these matches, we computed a distance matrix where the distance was calculated as $1 - (pident/100)$ and then applied a single-linkage hierarchical clustering method. The resulting tree was cut at a cutoff of 0.2 ($1 - pidentThreshold$), resulting in 3,868 Distinct Antigenic Regions.

CHAGASTOPE-v2 - Antigenicity threshold

To determine the antigenicity threshold in this experiment, we set to recreate the results obtained in the discovery screening using virtual sample pools. The signal of a virtual pool for a given peptide was the highest signal for that peptide among the individual sera that were part of that pool (and were now being analyzed individually). We then compared the antigenicity of these virtual pools against the antigenicity from our original pools in all clusters of antigenic regions using ROC curves. The goal was to predict the antigenicity in the original pools using the information from the corresponding virtual pools. An original pool was antigenic if it surpassed the 10,784.80 threshold, but for the virtual pools we analyzed possible thresholds of the formula $mode + X * standardDeviation$, where X ranged from 1 to 4 in steps of 0.1 (using the mode and standard deviation from the second design). The best threshold was 5,814.81 (mode plus 2.4 standard deviations) with an AUC of 0.83.

3.4.6 Seroprevalence analysis

CHAGASTOPE-v1 - Shared antigenic peptides

To calculate the proportion of non-redundant antigenic peptides shared between two sera (in this case pooled serum samples), we obtained the list of non-redundant antigenic peptides for each of them. We calculated the proportion as the number of non-redundant peptides those two lists had in common divided by the total number of non-redundant peptides among those two lists (meaning, the intersection of the two sets divided by the union of the two sets).

CHAGASTOPE-v1 - Clusters with regions similar to known antigens

To analyze the novelty of our findings we compared the sequence of all 9,547 antigenic regions found in CHAGASTOPE-v1 against a list of 2,243 known epitopes for *T. cruzi* obtained from the Immune Epitope Database (IEDB) [153]. This list was obtained by selecting in IEDB: **Epitope:** "Linear peptide"; **Assay:** only "B Cell" and "Positive"; **Organism:** "Trypanosoma cruzi (ID:5693)"; **MHC Restriction:** "Any"; **Host:** "Human"; **Disease:** "Any".

We used BLAST to compare this list against our antigenic regions. The command `blastp` was run using the same parameters as before (see above), but we only kept matches with a percentage of identical amino acids (*pident*) of at least 80% and a match length of at least 50% of the length of the shortest sequence in the match. These matches can be seen in Supplementary File S3.12. We tagged a cluster as "having regions similar to known antigens" if at least 10% of its regions matched with a known antigen (which in most cases meant at least 1 of them, see Figure 3.3b).

CHAGASTOPE-v2 - Shared antigenic peptides

Same as for CHAGASTOPE-v1.

CHAGASTOPE-v2 - Comparison between strains

To define if a peptide belonged to CL-Brener or Sylvio X10 (or to both), we looked at each of the peptides from the antigenic regions present in CHAGASTOPE-v2 and mapped these to their cognate proteins. Peptides that were mapped to proteins in both CL-Brener and Sylvio X10 were excluded from this analysis. Next, for each individual sample we counted the number of non-redundant reactive peptides that were exclusive for each strain. When calculating the signal for

each non-redundant peptide we kept the highest signal between the peptides with the same sequence (there was more than one signal per peptide due to the smoothing). Because the CL-Brener proteome is larger (almost double the size of the Sylvio X10 proteome) we standardized the number of reactive non-redundant peptides using two sets of Z-scores, one for each strain. The result of this analysis can be seen in Figure 3.8.

While we performed this analysis using all 71 individual samples from CHAGASTOPE-v2, the interpretation of the reactivity observed for the individual sera containing "_E" in their sample name (those not used on the proteome-wide v1 arrays) has to be done with care, because these 38 serum samples were not used to select peptides for inclusion in the CHAGASTOPE-v2 arrays.

3.4.7 Code availability

Custom software used for data analysis is available at this GitHub Repository: https://github.com/trypanosomatics/The-Chagas-Disease-Antigen-and-Epitope-Atlas.

3.5 Supplementary Materials

3.5.1 Supplementary Figures

All Supplementary Figures mentioned in this chapter were deposited as a single PDF file in Figshare under DOI:10.6084/m9.figshare.22047782.v1.
Direct Link: PhD Thesis - Ricci - Chapter 3 - Supplementary Figures.pdf (Size: 20.42 MB)

Supplementary Figure S3.1 – *Quality control of 1-plex microarray assays.* *Each biological sample (serum pool) was assayed in duplicate (technical replicates). The signal correlation between technical replicates is shown both in reciprocal plots and in MA (Bland-Altman) plots. Each point represents one unique peptide or addressable array spot (n = 2,842,420). The density is shown in a color gradient (see key in figure). In the reciprocal plots, the raw or normalized signal data of one replicate is plotted against the second replicate and two Pearson scores are calculated: the first one using all peptides and the second one using only the ones with the top 1% signals, which are those outside the dashed orange line. In MA plots, each signal is replaced by its log2 and then the average signal of the two replicates (A) is plotted against the difference of the two signals (M).*

Supplementary Figure S3.2 – *Quality Control of 12-plex microarray assays.* *Each biological sample (serum sample from a single individual) was assayed in duplicate (technical replicates) in a separate 12-plex slide. The signal correlation between technical replicates is shown both in reciprocal plots and in MA (Bland-Altman) plots. Each point represents one unique peptide or addressable array spot (n = 392,299). The density is shown in a color gradient (see key in figure). In the reciprocal plots, the raw or normalized signal data of one replicate is plotted against the second replicate and two Pearson scores are calculated: the first one using all peptides and the second one using only the ones with the top 5% signals, which are those outside the dashed orange line. In MA plots, each signal is replaced by its log2 and then the average signal of the two replicates (A) is plotted against the difference of the two signals (M).*

3.5.2 Supplementary Tables

All Supplementary Tables mentioned in this chapter were deposited as a single Excell file in Figshare under DOI:10.6084/m9.figshare.22047782.v1.
Direct Link: PhD Thesis - Ricci - Chapter 3 - Supplementary Tables.xlsx (Size: 1.94 MB)

Supplementary Table S3.1 – *Overview of arrays used in this work.* *Diagram of the microarrays used to analyze each design, CHAGASTOPE-v1 and CHAGASTOPE-v2.*

Supplementary Table S3.2 – *Serum samples and pools.* *Human serum samples used in this study. The Sample Code is an unique identifier for each individual serum. It shows the region plus a letter, which is "N" for negative sera, "P" for positive sera that were part of the pooled serum for that region, or "E" for positive sera that weren't in said pool.*

Supplementary Table S3.3 – *CHAGASTOPE array slides and assays.* *Detailed list of the assayed arrays.*

Supplementary Table S3.4 – *T. cruzi sequences cross-reactive with normal human serum (healthy subjects, Chagas-negative).* *All the proteins where the Chagas-negative pooled serum samples reached an antigenicity signal greater than the antigenicity threshold for CHAGASTOPE-v1. For each protein, the table shows the most reactive peptide for that serum and for the Chagas-positive pooled sera. If there is an overlap of at least 6 amino acids between those peptides, the table shows said overlap and compares the signals reached by each peptide.*

Supplementary Table S3.5 – *T. cruzi sequences cross-reactive with Leishmaniasis samples.* *All the proteins where the leishmaniasis-positive pooled serum samples reached an antigenicity signal greater than the antigenicity threshold for CHAGASTOPE-v1. For each protein, the table shows the most reactive peptide for that serum and for the Chagas-positive pooled sera. If there is an overlap of at least 6 amino acids between those peptides, the table shows said overlap and compares the signals reached by each peptide.*

Supplementary Table S3.6 – *Best antigenic region per cluster in pooled serum samples.* *This table shows a representative antigenic region with the highest seroprevalence in CHAGASTOPE-v1 for each selected cluster. For regions, the seroprevalence ("Sero") is calculated simply as the percentage of sera where the region proved to be antigenic. For clusters, the seroprevalence is calculated as the percentage of sera where any of its regions proved to be antigenic (see Methods). The column "Cross-reactive Regions" shows the number of regions in that cluster that were detected by the leishmaniasis-positive pooled serum. The column "Known Regions" shows the number of regions in that cluster that have a similar sequence to known epitopes found in IEDB (see Methods). The full detailed list of antigenic regions for all clusters can be found in Supplementary File S3.2.*

Supplementary Table S3.7 – *Best antigenic region per cluster in individual serum samples.* *This table shows a representative antigenic region with the highest seroprevalence in CHAGASTOPE-v2 for each selected cluster. For regions, the seroprevalence ("Sero") is calculated simply as the percentage of sera where the region proved to be antigenic. For clusters, the seroprevalence is calculated as the percentage of sera where any of its regions proved to be antigenic (see Methods). The column "Cross-reactive Regions" shows the number of regions in that cluster that were detected by the leishmaniasis-positive pooled serum. The column "Known Regions" shows the number of regions in that cluster that have a similar sequence to known epitopes found in IEDB (see Methods). The full detailed list of antigenic regions for all clusters can be found in Supplementary File S3.5.*

3.5.3 Supplementary Files

All Supplementary Files mentioned in this chapter can be found in the Figshare repository used for the original paper under DOI:10.6084/m9.figshare.19991021.v2 . They were numbered to match those in the original paper for easier linking, which results in some missing numbers. The direct links to each of the Supplementary Files as well as their details are listed below.

Our microarray data was deposited in ArrayExpress. The microarray designs can be found under accession numbers A-MTAB-692 (CHAGASTOPE-v1) and A-MTAB-693 (CHAGASTOPE-v2) and our raw and normalized data can be found under accession numbers E-MTAB-11651 (CHAGASTOPE-v1) and E-MTAB-11655 (CHAGASTOPE-v2). Supplementary Files S3.8 and S3.9 contain the information necessary to map the different peptides found in the microarrays to their position in each protein. The already mapped and parsed signals for each peptide in each protein can be found in Supplementary File S3.11.

Supplementary File S3.1 – ***CHAGASTOPE-v1 profiles for proteins above 3SD.*** *ZIP file containing 3 PDF files with the protein profiles for the proteins in both T. cruzi proteomes analyzed in CHAGASTOPE-v1, showing their antigenicity in the pooled serums. Only proteins that had at least one peptide above an antigenicity threshold of mode + 3 standard deviations are shown.*
Direct Link: Supplementary File S01 - CHAGASTOPE-v1 - Profiles for proteins above 3SD.zip (Size: 2.06 GB)

Supplementary File S3.2 – ***Detailed antigenic region analysis for pooled serums in CHAGASTOPE-v1.*** *Compressed TSV file containing the detailed information for all the antigenic regions found in CHAGASTOPE-v1 and their respective antigenicity in the pooled serum samples.*
Direct Link: Supplementary File S02 - CHAGASTOPE-v1 - Detailed antigenic region analysis for pooled serums.zip (Size: 766 KB)

Supplementary File S3.3 – ***CHAGASTOPE-v2 profiles for proteins above 2.4SD.*** *ZIP file containing a PDF with the squashed protein profiles for the antigenic regions analyzed in CHAGASTOPE-v2, showing their antigenicity in the individual serum samples.*
Direct Link: Supplementary File S03 - CHAGASTOPE-v2 - Profiles for proteins above 2.4SD.zip (Size: 359 MB)

Supplementary File S3.4 – ***CHAGASTOPE-v2 profiles for regions above 2.4SD.*** *ZIP file containing a PDF with the squashed region profiles for the antigenic regions analyzed in CHAGASTOPE-v2, showing their antigenicity in the individual serum samples.*
Direct Link: Supplementary File S04 - CHAGASTOPE-v2 - Profiles for regions above 2.4SD.zip (Size: 264 MB)

Supplementary File S3.5 – ***Detailed antigenic region analysis for individual serums in CHAGASTOPE-v2.*** *Compressed TSV file containing the detailed information for all the antigenic regions found in CHAGASTOPE-v1 and their respective antigenicity in the individual serum samples.*
Direct Link: Supplementary File S05 - CHAGASTOPE-v2 - Detailed antigenic region analysis for individual serums.zip (Size: 906 KB)

Supplementary File S3.8 – ***Mapping of CHAGASTOPE-v1 data to T. cruzi's proteins.*** *Compressed TSV file containing the detailed information to map each peptide in CHAGASTOPE-v1 microarray to their original spots in each protein of T. cruzi in this array.*
Direct Link: Supplementary File S08 - Mapping of CHAGASTOPE-v1 data to T cruzi proteins.zip (Size: 61.8 MB)

Supplementary File S3.9 – ***Mapping of CHAGASTOPE-v2 data to T. cruzi's proteins.*** *Compressed TSV file containing the detailed information to map each peptide in CHAGASTOPE-2 microarray to their original spots in each protein of T. cruzi in this array.*
Direct Link: Supplementary File S09 - Mapping of CHAGASTOPE-v2 data to T cruzi proteins.zip (Size: 7.39 MB)

Supplementary File S3.10 – ***CHAGASTOPE-v1 positive controls.*** *PDF file with the protein profiles for the positive controls analyzed in CHAGASTOPE-v1, showing their antigenicity in the pooled serum samples.*
Direct Link: Supplementary File S10 - CHAGASTOPE-v1 - Positive controls.pdf (Size: 690 KB)

Supplementary File S3.11 – ***CHAGASTOPE-v1 and CHAGASTOPE-v2 smoothed signals.*** *2 ZIP file containing several TSV files (one per serum) with the parsed signal data (normalized and then smoothed) for each peptide in each protein in CHAGASTOPE-v1 and CHAGASTOPE-v2.*
Direct Link to CHAGASTOPE-v1 smoothed signals: Supplementary File S11a - CHAGASTOPE-v1 - Smoothed Signals.zip (Size: 651 MB)
Direct Link to CHAGASTOPE-v2 smoothed signals: Supplementary File S11b - CHAGASTOPE-v2 - Smoothed Signals.zip (Size: 314 MB)

Supplementary File S3.12 – ***Good BLAST matches between antigenic regions and IEDB antigens.*** *Compressed TSV file containing the detailed BLAST information for all good matches between antigenic regions found in CHAGASTOPE-v1 and known antigens found in IEDB.*
Direct Link: Supplementary File S12 - Good BLAST matches between antigenic regions and IEDB antigens.zip (Size: 2.37 MB)

4. Chagastope.org: an interactive web application for array data exploration

This shorter chapter describes the development of a website application that allows easy access to the data produced in the analysis of the whole proteome of two strains of *Trypanosoma cruzi* using high-density peptide microarrays and serum samples from across the Americas (see Chapter 3). The application was created using *Shiny* (an R package) and it is an interactive interface that enables exploration of raw and parsed data for all proteins, peptides and serums analyzed. The application also contains interactive visual representations of our data, such as antibody-binding plots. The current version of this website can be found online at https://chagastope.org/.

4.1 Introduction

In Chapter 3 we used high-density peptide microarrays to analyze the whole proteome of two strains of *Trypanosoma cruzi* and its antigenicity in serum samples from across the Americas. This was done in a two-step study, where first step focused on a proteome-wide analysis using pooled sera from diverse human populations across the Americas to find antigenic regions in *T. cruzi*, and the second step used individual sera to determine the seroprevalence for each of the regions found in the previous step. A short summary of each step is provided below.

In the first step we analyzed the whole proteome of two strains of *Trypanosoma cruzi*, CL-Brener [125] and Sylvio X10 [130], resulting in 30,500 proteins. To fit the proteins in a single high-density peptide microarray we used 16mer peptides with an overlap of 12 residues between consecutive peptides. The resulting array design contained 2,441,908 unique peptides and was named **CHAGASTOPE-v1**. This design was used to screen 14 pooled serum samples, in duplicate. Of those serums, 6 were pools of Chagas-positive individuals from either Argentina, Brazil, Bolivia, Colombia, Mexico or the United States; 6 were pools from healthy individual from those same areas, and 2 were controls for leishmaniasis. From this analysis we extracted antigenic regions, meaning, the regions of the proteins containing the peptides that were detected by the antibodies in the pooled serum samples. We found 9,547 antigenic regions in 7,707 proteins.

The second step focused on analyzing these 9,547 antigenic regions. Working with these smaller regions of the proteome allowed us to increase the epitope mapping resolution, using 16mers with an overlap of 15 residues between consecutive peptides. The resulting array design contained 241,772 unique peptides and was named **CHAGASTOPE-v2**. This design was used to screen 71 individual serum samples, all in duplicate. Some of these individuals were part of the pools used in the previous step, while others were new individuals from those same geographic areas.

This analysis produced many interesting observations already discussed in Chapter 3, as well as large amounts of data which we made available both as TSV files and as an entry in Array Express. While the TSV files are ideal for performing large scale analysis, they might result a bit tedious to work with for anyone interested in just a handful of specific proteins or peptides. For this reason we set to create an interactive website that would allow quick access to our data and enable users to generate visual depictions of antibody-binding to proteins. Due to our experience working with the programming language R, we decided to create this website using *Shiny*, an R package that specializes on building interactive web apps. We called this website **Chagastope Web**.

This chapter functions as a pseudo-manual for Chagastope Web, which can be fund at https://chagastope.org/. We present the contents displayed in the website, provide a tour of the functionality, discuss its current limitations and likely future upgrades, and provide a detail of the underlying framework and tools used in its making.

4.2 Results

Using R and *Shiny* we have developed an interactive website that allows easy exploration of the data we obtained in the analysis of the whole proteome of two strains of *Trypanosoma cruzi* using high-density peptide microarrays (see Chapter 3). Using this website, anyone can access the numeric antibody-binding signals for all peptides in our microarrays, as well as static and dynamic visualizations of said data. The functionality of the site was designed to allow users to get responses to common user questions, such as:

- "What is the signal of peptide `PFGQAAA` in serum samples from Argentina?"
- "What are the peptides reactive in protein TcCLB.508235.230? "
- "How is the shape of the antibody-binding peaks of protein TcCLB.509007.70 in samples AR_P2, AR_P3, AR_P5 and AR_P6 between residues 1250–1280?"

Chagastope Web is divided into different menus and sections, each of them focusing on a different way to explore our data:

- Raw and normalized antibody-binding signals for all peptides
- Smoothed antibody-binding signals for all peptides, grouped by protein
- Smoothed antibody-binding signals for all peptides, grouped by antigenic region
- Static antibody-binding plots for the proteome-wide analysis of *T. cruzi* using pooled samples.
- Static antibody-binding plots of the antigenic proteins using individual serum samples.
- Static antibody-binding plots of the antigenic regions using individual serum samples.
- Single-residue mutagenesis of selected epitopes
- Dynamic antibody-binding plots for all proteins

Below, we explore each of these sections, detailing both what they show and how they work. It is important to note that while this site is ideal for anyone interested in a few proteins or peptides, when doing large scale analysis it is better to use the TSV files provided in the Supplementary Tables and Files of Chapter 3.

4.2.1 Home, Summary and Help

These three sections act as auxiliary sections for the rest of the website, either allowing easier navigation or providing more detailed information.

- As the name suggests, **Home** is the main page of the website and it contains direct links to all other sections of the website which allows for quick navigation (see Figure 4.1).
- **Summary** contains a short summary of what was explained in Chapter 3 of this thesis, focusing only on what is needed to understand the data and plots present in the website.
- **Help** has information about our lab, expands upon the details of the different ways of visualizing our data, and provides links to the TSV files found in Chapter 3 of this thesis.

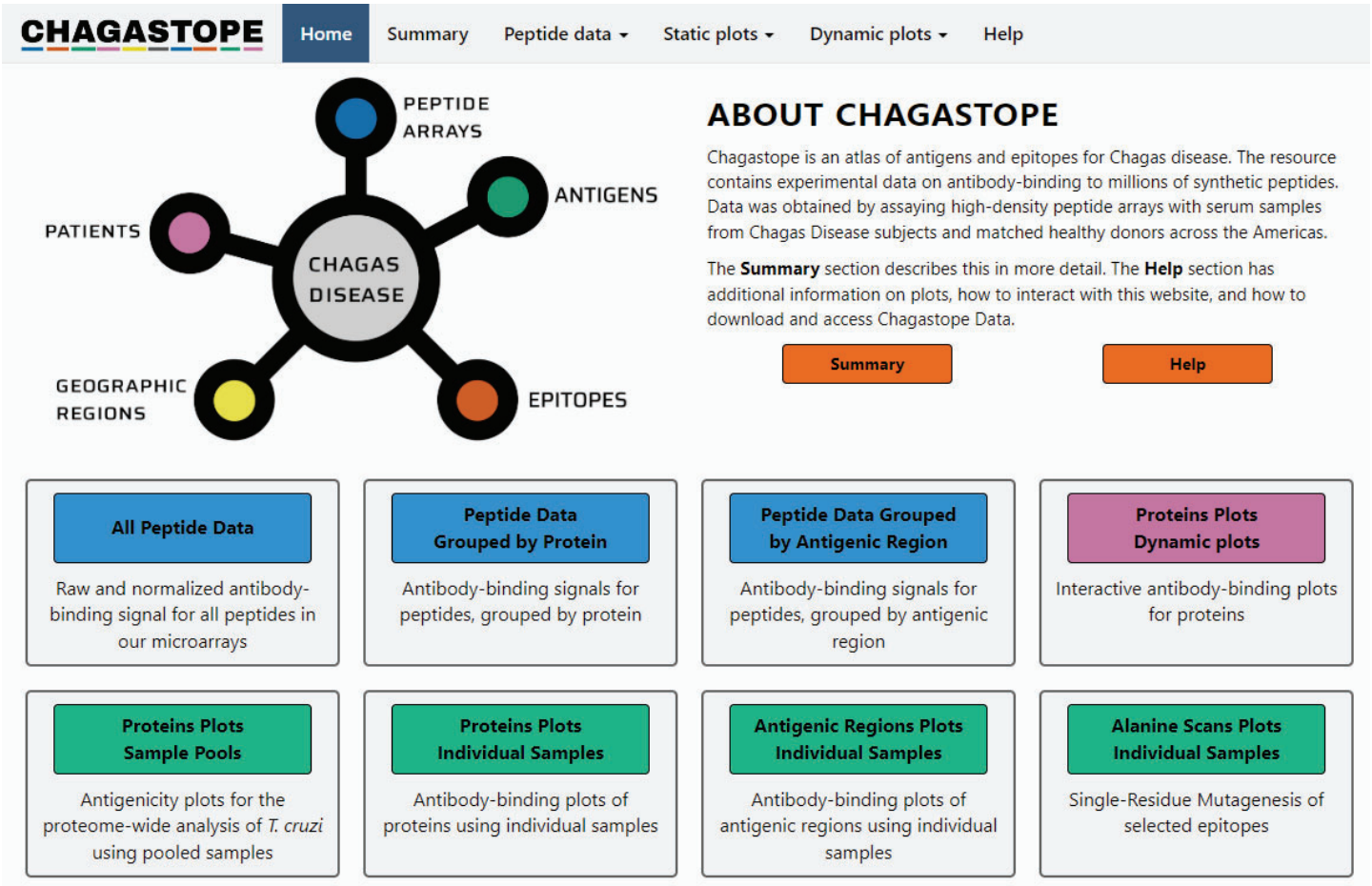


FIGURE 4.1 – *Home page of Chagastope Web*

4.2.2 Antibody-binding signal data

Our experiments produced fluorescence signal values for each peptide in the array and for each assay performed. This data can be searched, explored, and filtered in **Chagastope Web** in tabular format, as well as downloaded.

Raw and normalized antibody-binding signals for all peptides

Peptide data → **All Peptide data** shows the raw and normalized antigenicity signals for each individual peptide. Here antibody-binding to peptides is presented without any additional protein context (e.g. location in a given protein). A screen capture of this section can be seen in Figure 4.2. The table shows:

- **Sequence:** The sequence of the peptide to assay in the microarray (it was not always the peptide tested, see **Truncated** below).
- **Serum:** Code referring to the origin of the serum being tested. The full list of serum codes can be seen in Supplementary Table S3.2 (in Chapter 3).
  - Prefixes: AR (Argentina), BO (Bolivia), BR (Brazil), CO (Colombia), MX (Mexico), US (United States), LE (Leishmaniasis).
  - Suffixes: No suffix (Pooled serum), P# (Individual serum, was part of the pool), E# (Individual serum, was not part of the pool).
  - It is worth noting that selecting a code related to a Pooled Serum (the ones without suffix) will return data from both the Positive and Negative Pooled Sera from that location. It is possible to select just one of them by also using the **Serum Type** filter.
- **Serum Type:** The type of serum being tested (Positive pool, Negative pool or Positive serum).
- **Replica:** Number to differentiate both replicas (the order has no meaning).
- **Raw Signal:** Fluorescence signal value obtained in the microarray for this sequence, serum and replica. It has no post-processing.
- **Normalized Signal:** The **Raw Signal**, but normalized accross all microarrays using quantile normalization (see Methods in Chapter 3 for details).
- **Truncated:** Number of amino acids that were truncated from the end of the peptide before testing due to experimental limitations (it was a rare occurrence, see Methods in Chapter 3).
- **Experiment:** Design used for this test (either CHAGASTOPE-v1 or CHAGASTOPE-v2).

The header of the table allows for a quick sorting and filtering of its contents; however, that filter applies only to the information already in the table, not to all the information in our database (see **Main Filter Options** below). The data being shown in the table can be downloaded by pressing the **Download Data** button. This will download a CSV file which includes all the "pages" of the table, taking in account any filters applied in the header of the table.

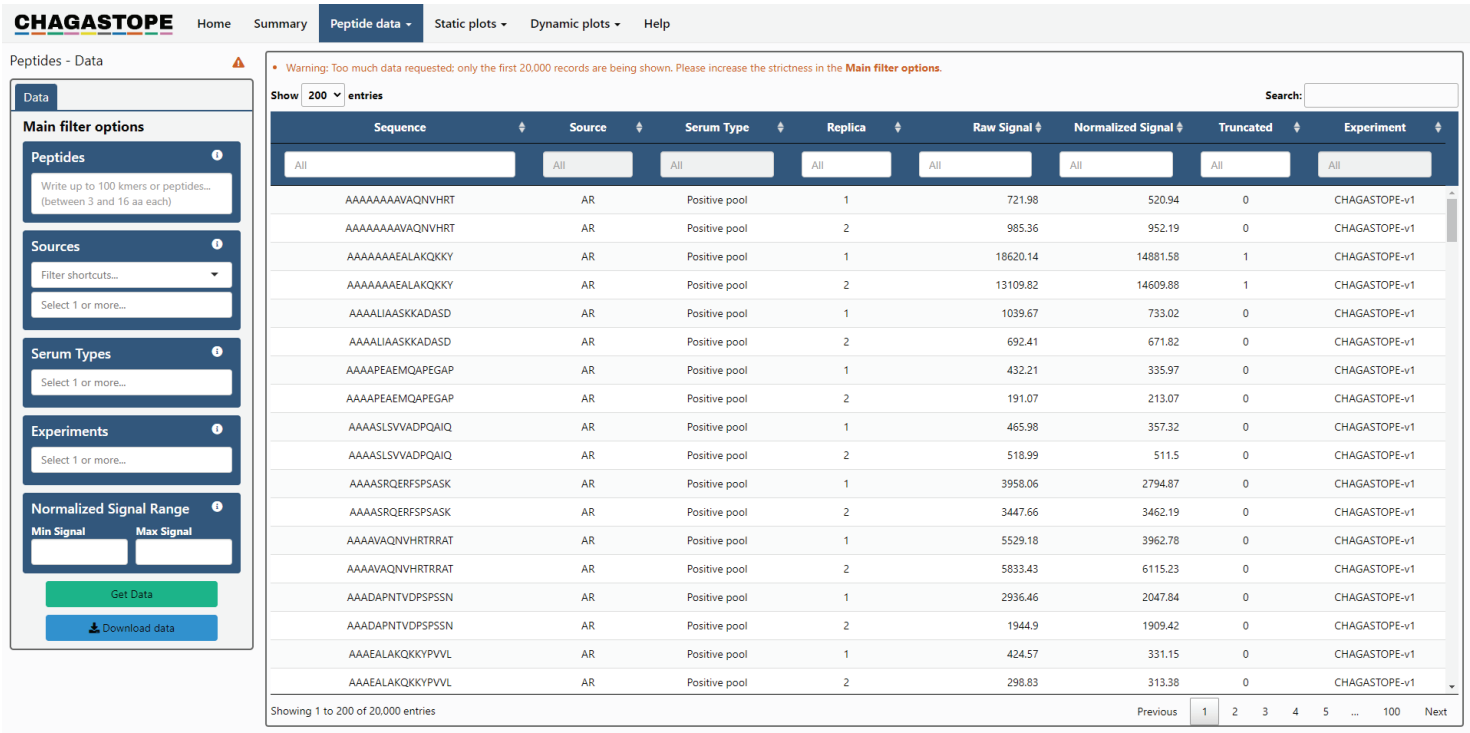


FIGURE 4.2 – *"Peptide data → All peptide data" section of Chagastope Web.* *The table in this figure shows information for each peptide in our microarrays, mainly the raw and normalized antibody-binding signal.*

It is possible to make personalized queries to our database by using the **Main Filter Options** sidebar. This is done by choosing any desired options in the several filters and then pressing the **Get Data** button. The filtering options are:

- **Peptides:** List of kmers to search in our database. All peptides in our database are 16 amino acids long. For 16mers, the database will return any perfect matches. For shorter sequences, the database will return any peptide containing said sequence.
- **Sources:** List of **Serum** codes to retrieve. The select box on the top allows for a faster selection of multiple related serum samples; options starting with "Add" will add that sources to the existing ones, while the other options will replace them.
- **Serum Types:** List of **Serum Types** to retrieve.
- **Experiments:** List of **Experiments** to retrieve.
- **Normalized Signal Range:** Minimum and/or maximum **Normalized Signal** to retrieve.

An example of this process can be seen in Figure 4.3. Hovering over the **(i)** symbols in the sidebar shows a tooltip with a summary of each filter options. It is not necessary to set a value for all filters; selecting no option is equivalent to selecting all.

For the inputs where you have to choose from a pre-existing list (**Sources**, **Serum Types** and **Experiments**) it is possible to copy the selection to the clipboard by clicking in an empty space in that input and pressing **CTRL + C**. In the case of **Sources** you can also do the opposite, meaning you can use **CTRL + V** to paste a list of options that will be then selected automatically. These options have to be pre-existing options in the list and they have to be separated by a comma and no spaces (which is also the format used when copying the data to the clipboard).

It is possible that a **Warning** appears above the table, usually related to too much data being requested. This can be fixed by increasing the strictness of the query. Pressing the Warning symbol hides the text above the table.

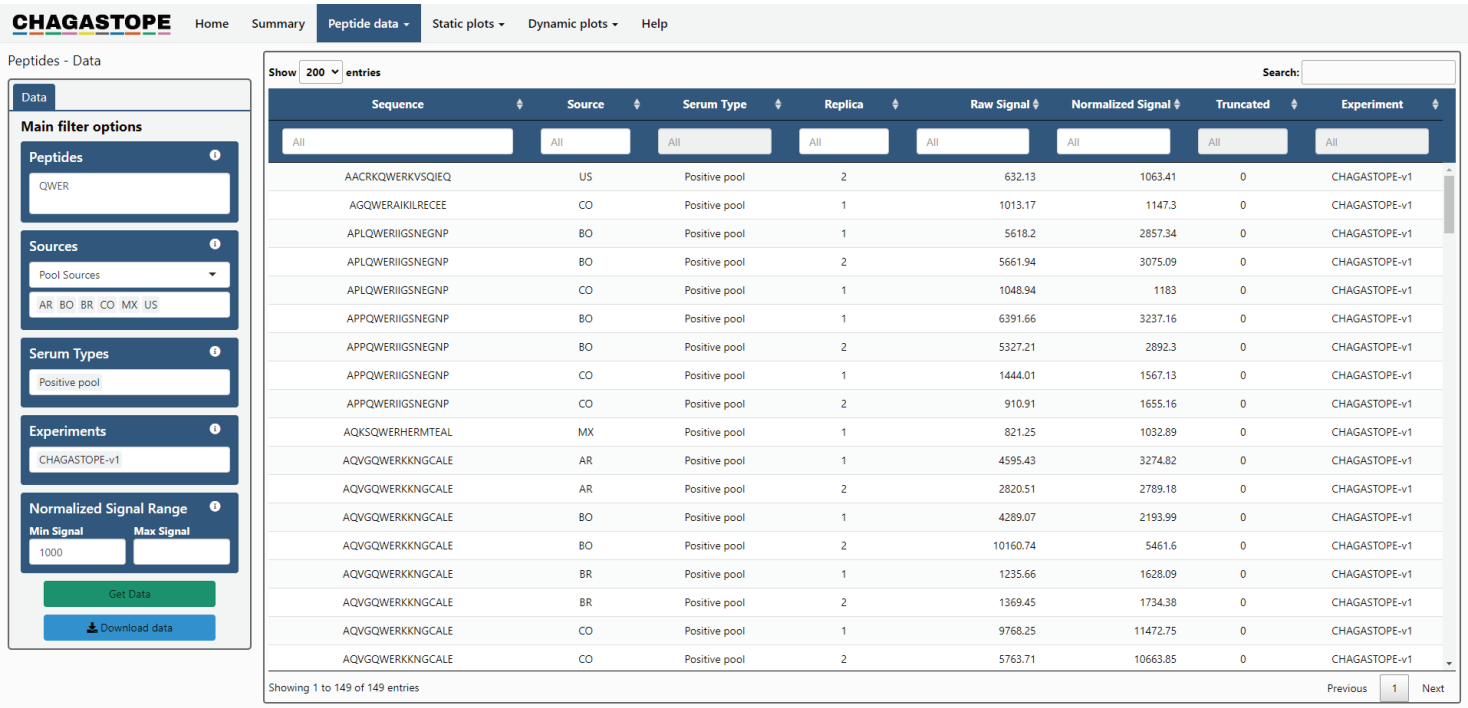


FIGURE 4.3 – ***Applying a filter to the "Peptide data → All peptide data" section of Chagastope Web.*** *An example of how to apply a filter to the data shown on Figure 4.2 by using the "Main filter options" sidebar.*

Smoothed antibody-binding signals for all peptides, grouped by protein

Peptide data → Grouped by protein shows the parsed antigenicity signals for each peptide in each protein analyzed. A screen capture of this section can be seen in Figure 4.4. The table shares many columns with the previous one; the new columns are:

- **Protein:** Locus identifier of the protein containing the peptide.
- **Peptide Start:** Position of the first amino acid of the peptide in the protein (the first peptide has the position 1). Depending on the **Source** and the **Experiment**, the offset between neighboring peptides will be 1 or 4.
- **Smoothed Signal:** The parsed signal of the peptide, obtained by smoothing the **Normalized Signal** of the peptide in the context of the protein and averaging both replicas (see Methods in Chapter 3 for details).
- **SD:** The Standard Deviation of the two replicas used to calculate the **Smoothed Signal**.

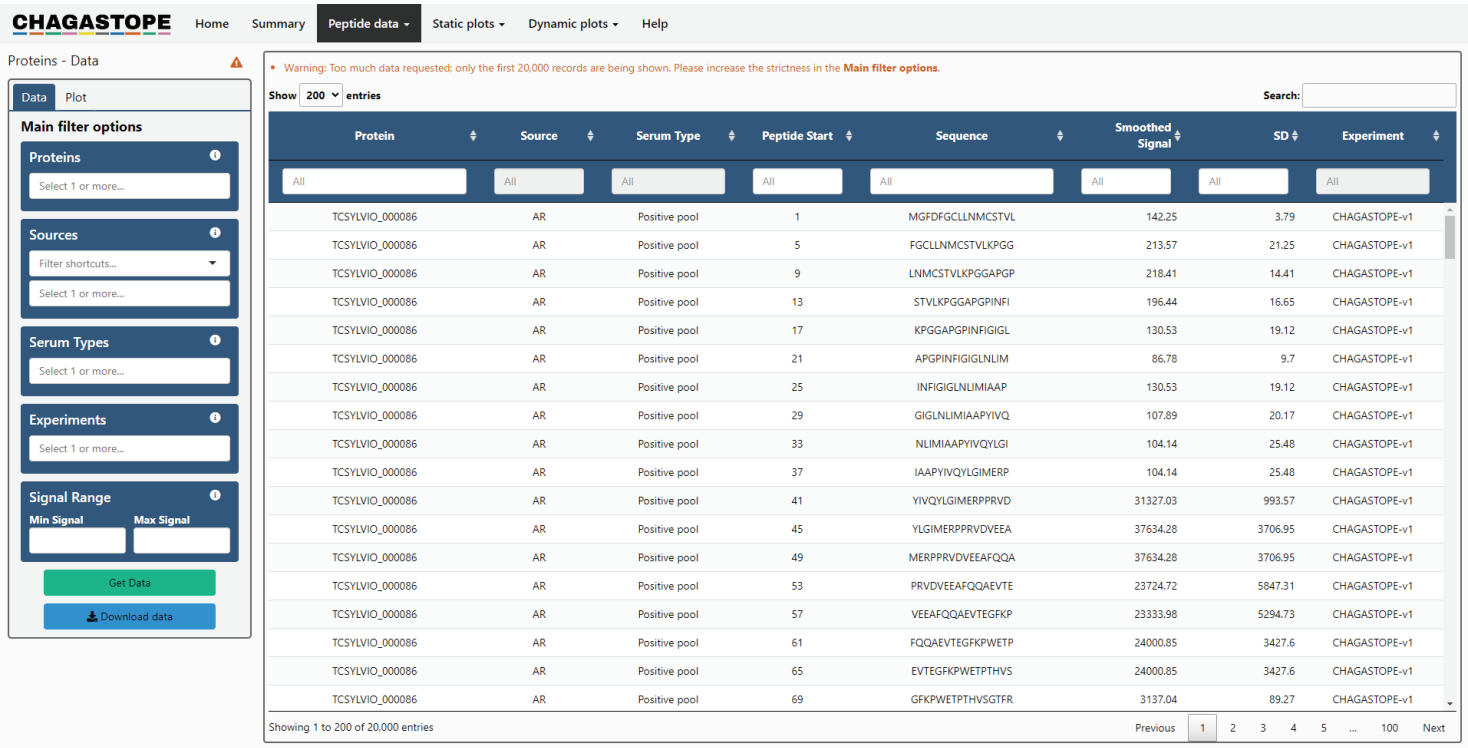


FIGURE 4.4 – *"Peptide data → Grouped by protein" section of Chagastope Web.* *The table in this figure shows information for each peptide in our microarrays grouped by protein, mainly the smoothed antibody-binding signal.*

Same as before, it is possible to make personalized queries to our database by using the **Main Filter Options** sidebar. The new filtering options are:

- **Proteins:** List of the locus identifiers of the proteins containing the peptides. It is possible to select many by pasting a list of locus identifiers (separated by a comma).
- **Signal Range:** Minimum and/or maximum **Smoothed Signal** to retrieve.

Besides the filtering options, the sidebar also has another tab called **Plot**. The buttons in this tab allows for quick access to the Static Plots and Dynamic Plots for whichever Protein is selected in the table. If a button is grayed out that means that there is no plot of that kind for that protein. These plots are discussed in detail below, in Sections 4.2.3 and 4.2.4.

Smoothed antibody-binding signals for all peptides, grouped by antigenic region

Peptide data → Grouped by antigenic region shows the parsed antigenicity signals for each peptide in each antigenic region analyzed. This section is very similar to the one above, with the only difference being that you can search our peptides by antigenic regions instead than by protein. The only new column is:

- **Region:** Identifiers of the antigenic regions containing the peptide, along the corresponding locus identifiers of the protein containing said region. The identifiers of the regions in Chagastope Web are the letter "R" followed by the ID for the region used in Chapter 3. It is possible to select many by pasting a list of region identifiers (separated by a comma).

It is important to note that not all peptides in our study are part of our antigenic regions, so it is possible for a peptide to appear in **Grouped by protein** and not in **Grouped by antigenic region**.

4.2.3 Static antibody-binding plots

We have produced signal plots to visualize the antibody-binding signal for peptides along proteins. These plots allow for a much easier access to our data, as well as show the location of antigenic regions and epitopes in the proteins. Static plots are single-page PDFs containing each of the plots generated in Chapter 3, as well as a few more. When downloading large amounts of plots, it is simpler to do so from the files provided in the Supplementary Files of Chapter 3 or from the links found in the Help section of Chagastope Web.

Antibody-binding plots for the proteome-wide analysis of T. cruzi using pooled samples

Static plots → Proteins - Sample pools (CHAGASTOPE-v1) shows the antibody-binding plots for the most interesting proteins found in our analysis of two strains of *Trypanosoma cruzi* using high-density peptide microarrays and pooled serum samples from all across the Americas. The proteins available are those with at least one peptide with an antigenicity signal above 8,106.10 (mode plus three standard deviations for CHAGASTOPE-v1; this happened in 12,459 out of the 30,500 proteins analyzed). These plots also include the analysis of a pool of Leishmaniasis-positive and Leishmaniasis-negative samples to test for cross-reacting epitopes.

A screen capture of this section can be seen in Figure 4.5. The plots can be loaded by selecting the locus identifier of the corresponding **Protein** in the **Main Filter Options** sidebar and pressing the **Plot Data** button. The plot should be downloadable from the PDF viewer in the browser, but not all browsers behave the same, so this may vary.

It is also possible to arrive at this plot by pressing one of the buttons from the **Plot** tab in the sidebar of the tables. If this is the case, then a button named **Go back to table** will be also visible.

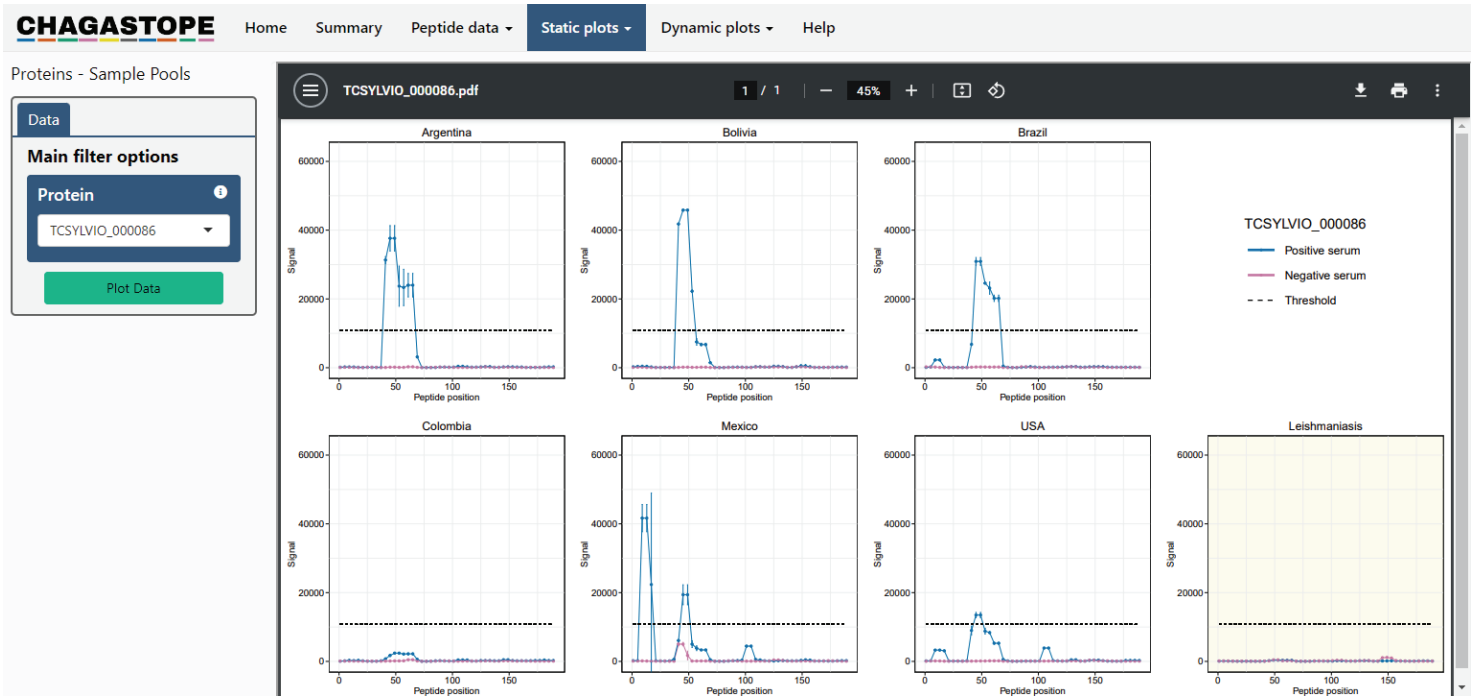


FIGURE 4.5 – *"Static plots → Proteins - Sample pools" section of Chagastope Web.* *The figure shows the antibody-binding plot for the protein selected in the "Main Filter Options" sidebar. These plots correspond to the analysis of the whole protein using pooled serum samples (see CHAGASTOPE-v1 in Chapter 3).*

Antibody-binding plots of antigenic proteins using individual serum samples

Static plots → Proteins - Individual samples (CHAGASTOPE-v2) shows the antibody-binding plots for each of the 7,707 proteins containing the antigenic regions found in CHAGASTOPE-v1, which are now tested using 71 individual serum samples from all across the Americas. The samples are grouped by country of origin, and the first plot in each group corresponds to the profile of that protein in the CHAGASTOPE-v1 arrays.

A screen capture of this section can be seen in Figure 4.6. The plots can be loaded by selecting the locus identifier of the corresponding **Protein** in the **Main Filter Options** sidebar and pressing the **Plot Data** button. The plot should be downloadable from the PDF viewer in the browser, but not all browsers behave the same, so this may vary.

The prefixes in the serum codes mean: AR (Argentina), BO (Bolivia), BR (Brazil), CO (Colombia), MX (Mexico), US (United States). The suffixes in the serum codes mean: No suffix (Pooled serum), P# (Individual serum, was part of the pool), E# (Individual serum, was not part of the pool). In these plots, the line changes colors when the peptide signal surpasses the antigenicity threshold, which is different for CHAGASTOPE-v1 and CHAGASTOPE-v2 (see Methods in Ch apter 3). Empty areas may appear in these plots due to the regions of the proteins that were not analyzed in CHAGASTOPE-v2 arrays (meaning they were not reactive in CHAGASTOPE-v1 arrays). The gray vertical dotted lines appear every 50 amino acids.

It is also possible to arrive at this plot by pressing one of the buttons from the **Plot** tab in the sidebar of the tables. If this is the case, then a button named **Go back to table** will be also visible.

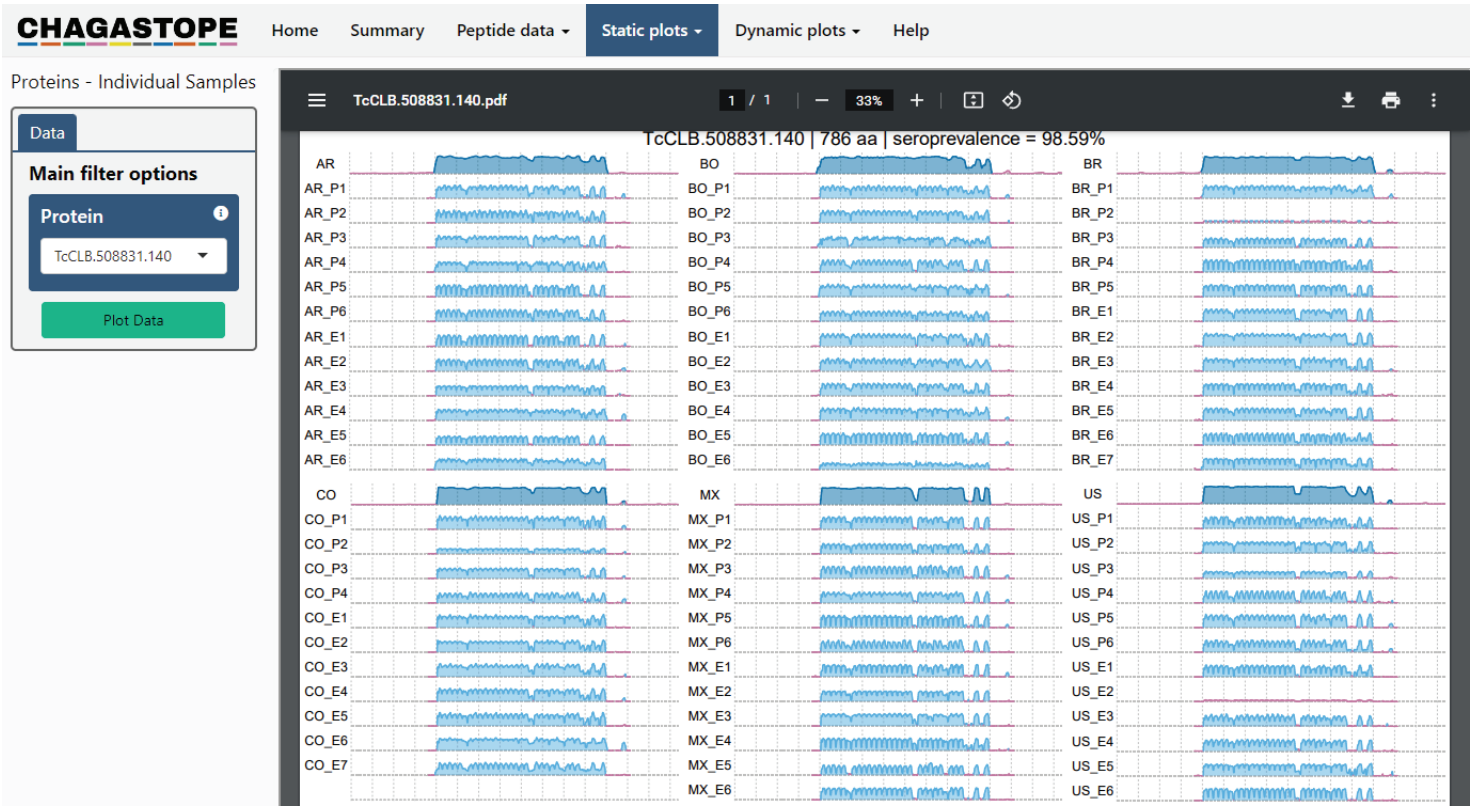


FIGURE 4.6 – ***"Static plots → Proteins - Individual samples" section of Chagastope Web.*** *The figure shows the antibody-binding plot for the protein selected in the "Main Filter Options" sidebar. These plots correspond to the analysis of the antigenic regions using individual serum samples (see CHAGASTOPE-v2 in Chapter 3).*

Antibody-binding plots of antigenic regions using individual serum samples

Static plots → Antigenic regions - Individual samples (CHAGASTOPE-v2) shows the antibody-binding plots for each of the 9,547 antigenic regions found in CHAGASTOPE-v1, which are now tested using 71 individual serum samples from all across the Americas. This section is very similar to the one above, with the only difference being that you can search the plots by antigenic regions instead than by protein. The identifiers of the regions in Chagastope Web are the letter "R" followed by the ID for the region used in Chapter 3.

Single-residue mutagenesis of selected epitopes

Static plots → Alanine Scans - Individual samples (CHAGASTOPE-v2) shows an Alanine Scan of a few selected peptides. We replaced, one at a time, every amino acid of the peptide for Alanine and analyzed how the signal changed for each individual serum. If the amino acid was already Alanine, it was replaced by Glycine. A detailed explanation of the Alanine Scan, which peptides were chosen, and what was concluded from this analysis can be seen in the original paper.

A screen capture of this section can be seen in Figure 4.7. The plots can be loaded by selecting the **Sequence** of the corresponding peptide in the **Main Filter Options** sidebar (found along the locus identifier of its protein) and pressing the **Plot Data** button. The plot should be downloadable from the PDF viewer in the browser, but not all browsers behave the same, so this may vary.

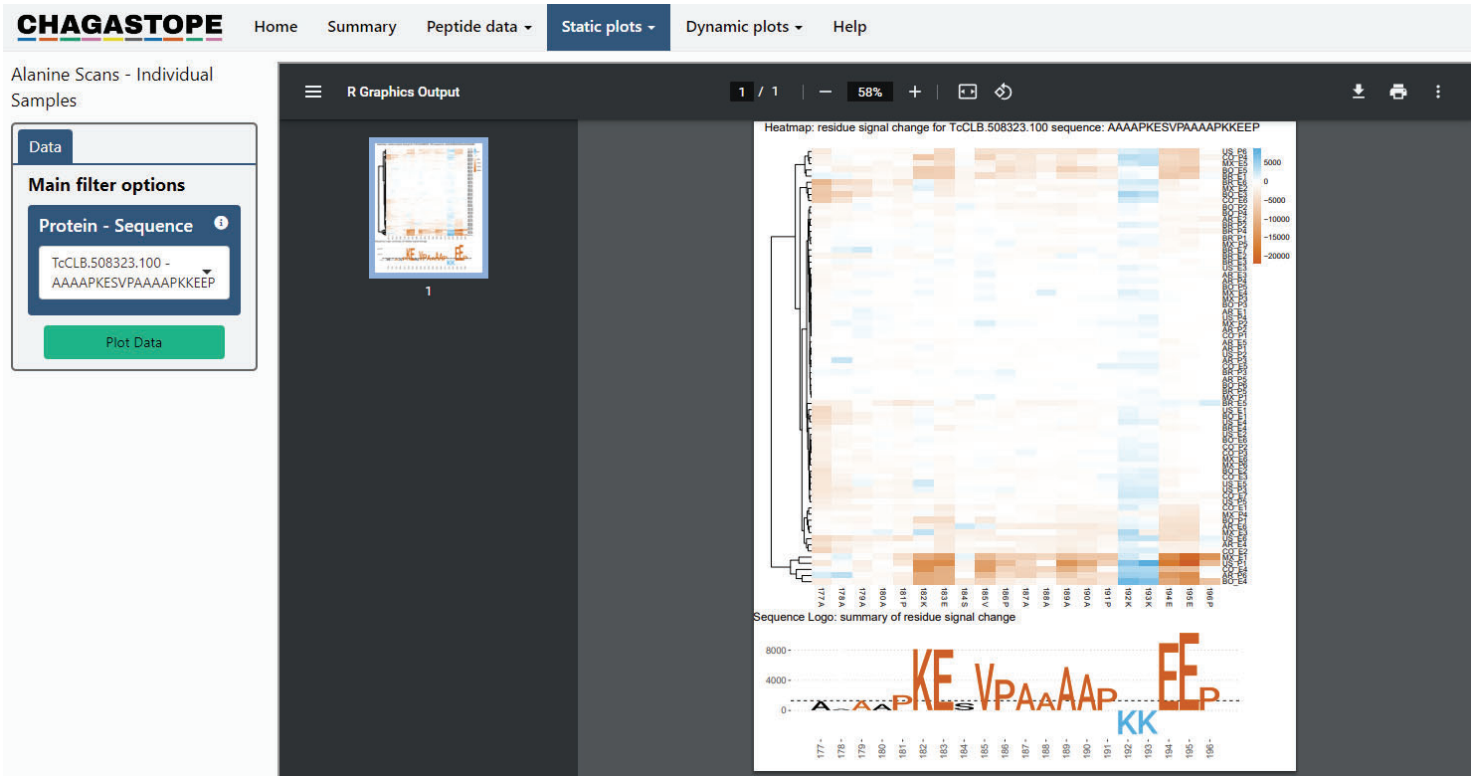


FIGURE 4.7 – ***"Static plots → Alanine Scans - Individual samples" section of Chagastope Web.*** *The figure shows the Alanine Scan plot for the sequence selected in the "Main Filter Options" sidebar. The heatmap shows the variation of the antigenicity signal when replacing a specific amino acid by Alanine (if the amino acid was already Alanine, it was replaced by Glycine). The summary below shows the average variation for all serum samples; amino acids above the cutoff are considered to be the ones that form part of the actual epitope.*

4.2.4 Dynamic antibody-binding plots

Dynamic plots → **Proteins** shows an application that can generate antibody-binding plots in real time from our data. Dynamic Plots allow users to choose which protein and serum samples to plot, as well as changing many visualization options. These plots are also interactive; by being rendered by *Plotly* (an R graphic library), they can be zoomed, re-scaled, panned, and additional information can be displayed by moving the mouse cursor over data points [218]. *Plotly* also allows them to be easily exported in SVG format.

A screen capture of this section can be seen in Figure 4.8. To create a plot is necessary to select one locus identifier in the **Protein** box and between 1 to 18 serum samples in the **Sources by hand** box of the **Main Filter Options** and then pressing the **Plot Data** button. As mentioned before, the resulting plot is made in *Plotly*, which gives some extra funcionality:

- **Hovering above a data point:** shows a tooltip with the peptide's source, source type, position in the protein, smoothed signal and SD, and its sequence.
- **Pressing the items in the plot legend:** Hides or show different parts of the plot (what they are change based on the **Plot Style** on the **Plot options** tab, discussed below). Double-clicking an item on the legend does the opposite, hiding or showing every other item in the legend.
- **Download the plot:** Pressing the Camera icon on the top right toolbar saves the current version of the plot as a non-interactive SVG file.
- **Explore the plot:** The Zoom, Pan, Zoom In and Zoom Out icons on the top right toolbar enable different ways of visually exploring the plot, with the most useful one likely being the Zoom option. This can be used to observe details on a specific area of the protein, but in most cases it will affect only one of the figures. To make the same Zoom for all figures it is necessary to use the **Zoom options** tab (discussed below).
- **Reset the plot:** The Autoscale and Reset Axes icons on the top right toolbar will affect all figures. The Autoscale will change the axis of each figure independently so the data fills all the space possible, while the Reset Axes will change all figures to their original state. It is also possible to reset just one of the figures by double-clicking it.
- **Compare the data:** The final two icons of the top right toolbar change how the tooltip that appears when hovering a data point works. By selecting the Compare data on hover option the plot will show at the same time the tooltip for all data points with the same X value, which in this case means the same position in the protein.

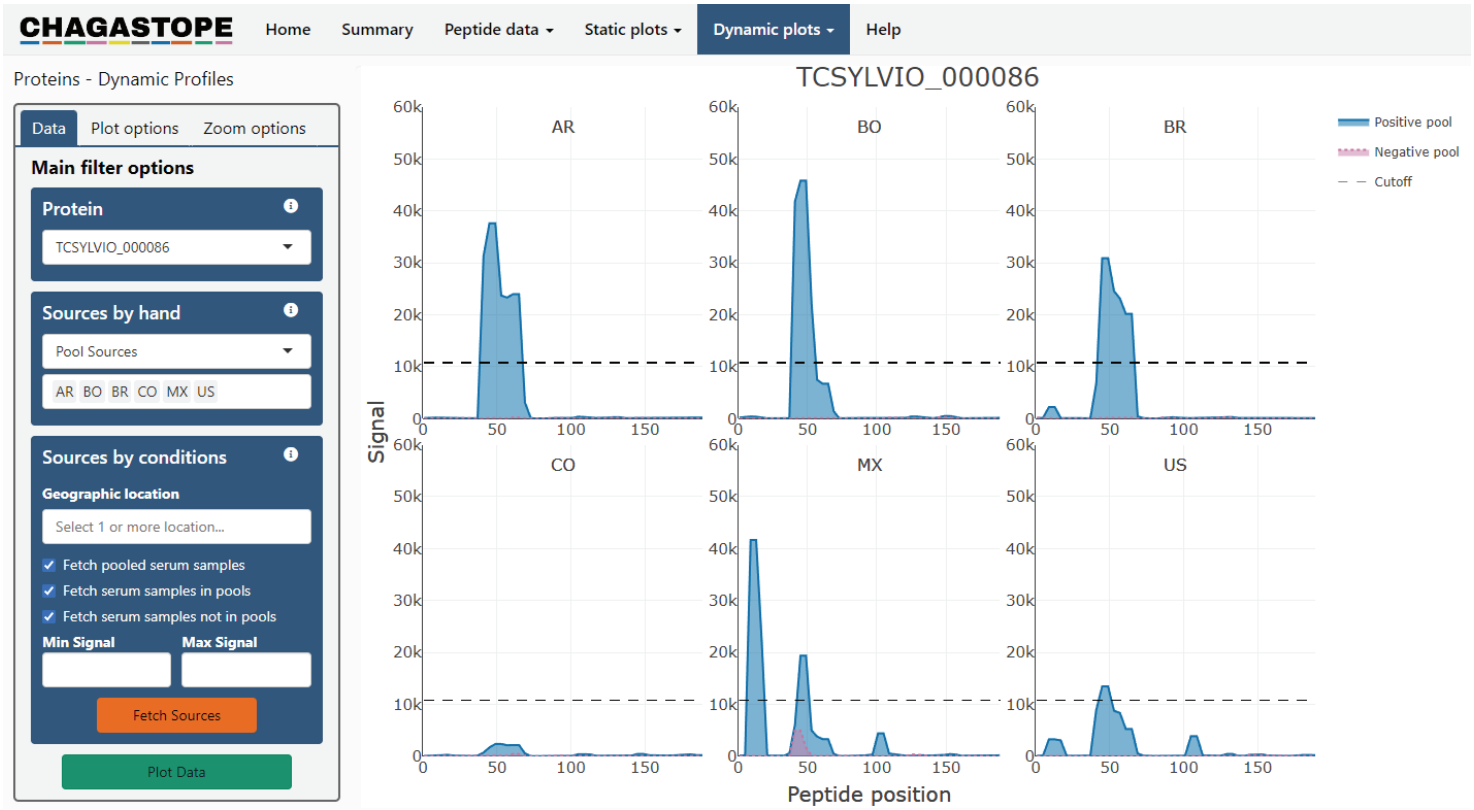


FIGURE 4.8 – *"Dynamic plots → Proteins" section of Chagastope Web.* *The figure shows the dynamic antibody-binding plot for the protein and serum samples selected in the "Main Filter Options" sidebar.*

The **Main Filter Options** sidebar has one more functionality in the **Sources by conditions** box. The **Sources by hand** box is useful for selecting some specific serum samples or serum samples from a specific location, but it is not able to select serum samples that are reactive or not for a given protein. **Sources by conditions** fills this gap, allowing users to select serum samples that fulfill a given condition for the protein selected in the **Protein** box. The possible filters for **Sources by conditions** are:

- **Geographic location:** Refers to where the individual comes from; e.g.: both the pool AR and the individual AR_P1 come from the location AR. While LE is not a location, is also present as an option. Selecting no location means selecting all.
- **Fetch...:** These three check boxes control which serum samples to fetch from the selected locations, such as pools (e.g. AR) or individual serums that are part of those pools (e.g. AR_P1) or not (e.g. AR_E1).
- **Min Signal:** returns the serum samples where at least one peptide has a signal above this number (0 means no filter).
- **Max Signal:** returns the serum samples where all peptides have a signal below this number (0 means no filter).

Once all filters are selected, pressing **Fetch Sources** will populate the input in **Sources by Hand** with sources that fulfill the conditions. Similar to the **Sources** inputs in the tables, you can copy and paste the sources selected for later use.

The other two tabs on the sidebar, **Plot options** and **Zoom options** have customization options for the Dynamic Plot. These can be seen in Figure 4.9, and they are:

- **Show negative signal:** Shows or hides the signal of the pools of healthy serum samples. Useful when selecting a pool in **Sources** (those without suffix).
- **Show standard deviation:** Shows or hides a vertical line for each data point corresponding to its standard deviation.
- **Show legend:** Shows or hides the legend.
- **Show fill:** Shows or hides the fill in the plots, meaning, the color below the lines.
- **Use a fixed scale:** Choose between a fixed scale, or a dynamic scale (when unchecked). A fixed scale will use the number in the text box as the maximum value for the Y axis. The dynamic scale will use the maximum signal for the proteins/sources selected as the maximum value for the Y axis.
- **Plot Style:** Determines how to combine the different sources.
  - **Individual - Horizontal:** It tends towards a three column grid, trying to add as few rows as possible.
  - **Individual - Vertical:** It tends towards a one column grid, trying to add as few columns as possible.
  - **Combined:** It combines all signals in one single figure. In this style, each source is assigned its own entry in the legend, and so it can be filtered in or out using *Plotly*.
- **Threshold:** Shows or hides the line of the antigenicity threshold (it has no functional effect). The select box has the values of the antigenicity thresholds used in our work, but any other value can also be input manually.
- **Zoom Range:** Found in the **Zoom options** tab in the sidebar, it allows for a way to apply the same zoom to all the figures in the plot. When loading a new plot for the first time, the **Min** and **Max** values for that protein will be loaded. By changing these values and pressing the **Zoom** button, all figures will change to show only that section of the plot. Pressing the **Reset Zoom** button will go back to the original values.

Figure 4.10 shows an example of a plot obtained with these options, where we chose the "Combined" **Plot Style**, unchecked the **Show negative signal** and **Show fill** options, set the **Threshold** to an arbitrary value of 15,000 units, used the **Zoom options** to show only the peptides between the position 25 and 75 and selected the *Plotly* option **Compare data on hover** (in the top-right corner of the plot).

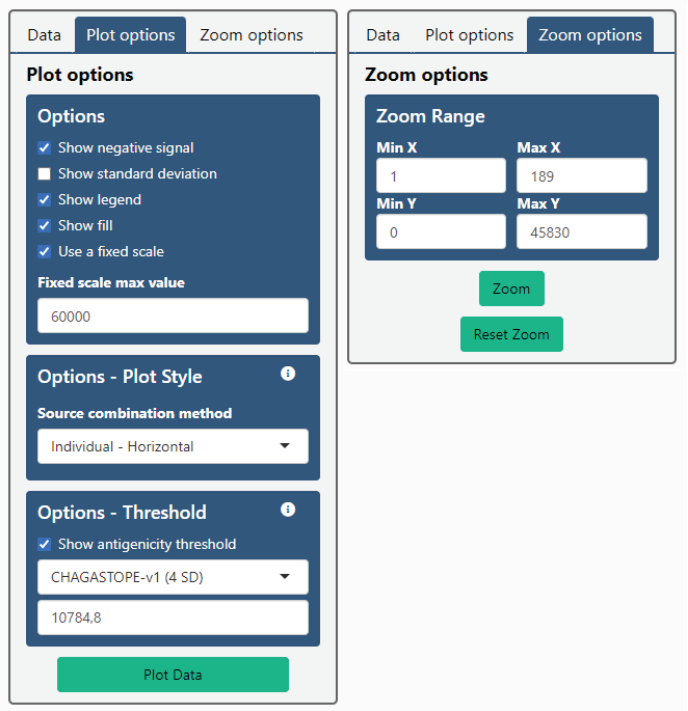


FIGURE 4.9 – *Customization options for the "Dynamic plots → Proteins" section of Chagastope Web.*

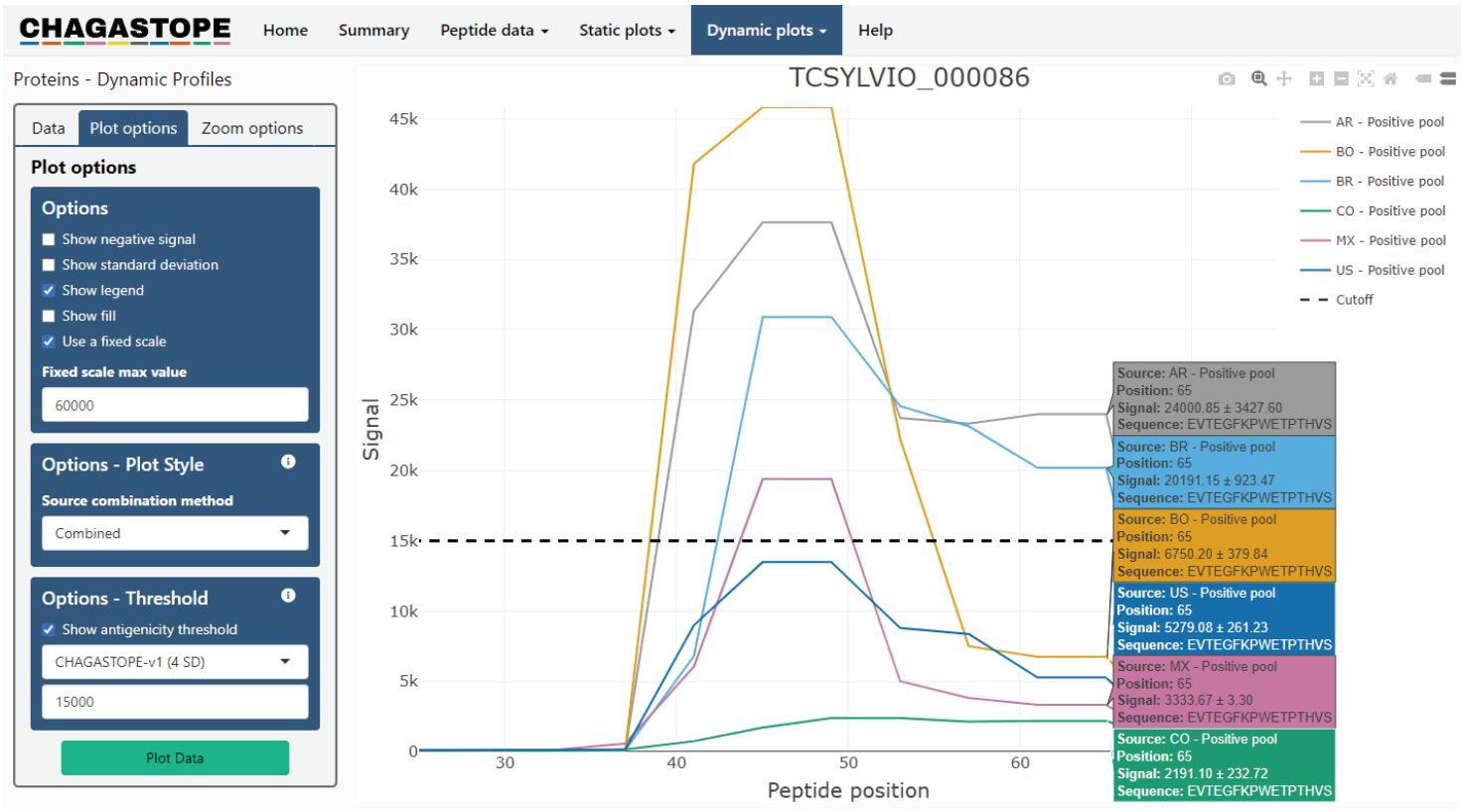


FIGURE 4.10 – *Example of a plot using the customization options for the "Dynamic plots → Proteins" section of Chagastope Web.*

4.3 Discussion

In Chapter 3 we used high-density peptide microarrays to analyze the whole proteome of two strains of *Trypanosoma cruzi* and its antigenicity in serum samples from across the Americas. In this chapter, we have produced an interactive website called **Chagastope Web** that allows for easy exploration of the aforementioned data, allowing users to search for raw, normalized and smoothed signals for all peptides in our microarrays, as well as a diverse variety of antibody-binding plots.

Due to processing and storage reasons, it was necessary to set some limits on what a user was able to request from our database. At this point in time, the tables can return up to 20,000 rows in a query and the dynamic plot can show up to 18 serum samples at the same time. As for the static plots from CHAGASTOPE-v1, we only have those where at least one peptide reached a signal of 8,106.10 (which is 1 standard deviation below the antigenicity threshold used for that design). This happened for 12,459 out of the 30,500 proteins analyzed, meaning that the other 18,041 proteins had a relative low signal. Having said that, all 30,500 proteins can be plotted as a dynamic plot.

We have plans to make some upgrades to Chagastope Web in the future. One of the main things we are trying to improve is making sure that the website works properly on any browser and resolution, including cellphones and tablets. While we have already taken some steps to fix a few related issues, more work and testing is needed to ensure it works everywhere. Another upgrade we are planning is to add a few more plots to the website, such as an interactive Alanine Scan plot where you can choose which serums to use and the summary is created using just those serums. Finally, by hosting Chagastope Web on a better computational cluster we could increase the limits we mentioned above, allowing the users to fetch more data from our database in a single query.

Chagastope Web is freely accessible at https://chagastope.org/. We hope that this interactive website makes the information obtained in Chapter 3 reach a wider audience, allowing anyone who is interested in Chagas disease or in large-scale immune responses against eukaryotic parasites to easily access the antigenicity profile for any protein in the proteome of *T. cruzi*.

4.4 Materials and methods

The **Chagastope Web** website was created to explore the data we obtained in our analysis the proteins of two strains of *Trypanosoma cruzi* and their antigenicity in serum samples from across the Americas (see Chapter 3). This website was created using the *Shiny* package of the R programming language, and can be accessed at https://chagastope.org/. The versions of Ubuntu, MySQL, R, Shiny Server and packages used to create the website can be found in Table 4.1.

Software	Version	Description
Ubuntu	22.04 LTS	Operating system
MySQL	8.0.31	Database manager
R	4.1.2	Programming language
Shiny Server	1.5.18.987	Manages the web server for the application
shiny (R package)	1.7.1	Makes web applications in R
DT (R package)	0.22	Adds interactive tables to Shiny
plotly (R package)	4.10.0	Adds interactive plots to Shiny
htmltools (R package)	0.5.2	Embeds HTML files in Shiny
shinyjs (R package)	2.1.0	Adds javascript to Shiny
shinycssloaders (R package)	1.0.0	Adds loaders to Shiny
shinycustomloader (R package)	0.9.0	Adds loaders to Shiny
data.table (R package)	1.14.2	Enhances internal table management
DBI (R package)	1.1.2	Connects R with MySQL
RMariaDB (R package)	1.2.1	Connects R with MySQL

Table 4.1 – ***Versions of the software and packages used to create our website.***

General Discussion

The knowledge of which pathogen molecules elicit an immune response and are the targets of antibodies during an infection is essential for many diagnostic and clinical applications. In this thesis we followed two strategies to obtain large amounts of valuable information about a given pathogen. The first strategy used existing data to train a prediction model that can then be used to infer antigenicity information for peptides and proteins from complete proteomes. The second strategy was based on the design of high-density peptide microarrays containing peptides from a pathogen of interest and testing them against sera from infected individuals. Each of these methods has its own strengths and limitations.

Prediction models: developing APRANK

Using bioinformatic predictors to find antigenic proteins and peptides has many benefits. These methods make it possible to analyze the antigenicity of hundreds to thousands of proteins/peptides in relatively little time and at a very low cost. This makes them an ideal first step in many studies that relate to antigenicity, specially when studying diseases where samples are difficult to come by. Also, while the design and training of some predictors might need above-average computational resources, the actual predictor can usually be run in a personal computer or it can even be an online service, lowering barriers of accessibility.

The first methods to predict linear B-cell epitopes appeared in the 80s and 90s and relied on "propensity scales". This method worked by providing propensity scores for each amino acid residues based on their physico-chemical properties, and then repetitively averaging these values along the peptide chain [156, 219]. Based on comparative studies from that time [220], the propensity scales that were most successful in predicting antigenicity were the ones related to hydrophilicity [221], secondary structure [222, 223] and accessibility [224], which were then used as the basis for many programs created at the start of the 2000s [225, 226]. While later studies showed that these earlier programs were only marginally better than a random model [227], they were the base for other more accurate models such as Bepipred [156].

The propensity scales are not the only methods used for detecting antigenic proteins. One alternative relies on finding proteins with short tandemly repeated domains, which are known to possess significant antigenicity in some pathogens (e.g. *T. cruzi*) [228]. This is because repeated epitopes in a protein can cross-link B-cell receptors on the surface of a B-cell, leading to T-cell independent activation [12]. Proteins containing perfect or imperfect tandemly repeated sequences can be identified with software such as XSTREAM [161].

Another way to predict antigenicity is to focus on regions of the pathogen's proteins that are potentially exposed to the host immune system, namely alpha-helical coiled-coils and intrinsically unstructured regions, ideally found on the outside of the pathogen [229]. This can be approached in many ways, such as prioritizing proteins that contains a sorting signal directing the protein to the extracellular space (SignalP [160]), contains transmembrane domains (TMHMM [164]), or contains a C-terminal signature sequence for addition of GPI anchor (PredGPI [159]). As for the secondary structure itself, it can be predicted by software such as IUPred [162], Paircoil2 [163], and NetSurfp [168].

Finally, it is also possible to find antigenic proteins by focusing on proteins that are differentially expressed in varied scenarios, for example, proteins from *T. cruzi* that were highly expressed in the parasite stages present in mammalian hosts [230]. While this last method does not directly involve

bioinformatic tools, it shows the possibility of adding an extra filter based on biological knowledge to the results obtained from a prediction model.

All these different methods of predicting linear B-cell epitopes are decent on their own, and it is possible combining them by simply applying one to the output of another. However, they can also be combined using a feature weighting approach, where each filter is not "all or nothing", but an input in a larger prediction model. This last method was used in our lab by Carmona et al. [149] to create a model that identifies candidate diagnostic peptides in the protozoan *Trypanosoma cruzi*, which outperformed alternative prioritizations based on individual properties that existed at the time [149].

While this predictor showed promising results, it was focused solely on *T. cruzi*, which limited its use. Chapter 2 of this thesis describes the extrapolation of this idea into a predictor that is able to infer antigens in any eukaryotic and prokaryotic pathogens. We called this predictor Antigenic Peptide and Protein Ranker, or APRANK for short.

Before discussing our prediction model, it is important to remark that bioinformatic predictors do have limitations, the main one being that they are just generalizations of reality. Their goal is to find combinations of properties that are able to partition proteins and peptides into antigenic and non-antigenic classes; however, this is not a straightforward process. For example, when training and validating a prediction model such as ours, it is necessary to rely on and trust pre-existing data to define which proteins and peptides are antigenic to begin with. Moreover, for negative examples (non-antigenic proteins and peptides) it is usually necessary to rely on lack of evidence as indirect evidence of non-antigenicity. Finally, when training the predictor with large data sets, it is common to assume that different pathogens are detected in similar ways by the immune system, and that all individuals detect the same antigens. Most of these conditions usually are not satisfied, or at least not all of them at the same time. Furthermore, it is also important to have data that is representative of all possible inputs for the predictor, which is not a trivial task. A prediction model will always yield some output, hence it is necessary to conduct proper training, testing and validation to make sure that the trained model returns informative predictions and generalizes well to diverse inputs.

As detailed in Chapter 2, we analyzed 16 properties of each protein/peptide from 15 diverse pathogen species, and used that information to train a binomial logistic regression model called APRANK, which was successful in predicting antigenicity for all pathogen species tested, including an unbiased validation using two independent data sets containing recent proteome-wide antigenicity data (*O. volvulus* and *P. falciparum*). When developing APRANK we set to navigate the limitations mentioned above by using data from a diverse set of pathogens (4 gram positive bacteria, 4 gram negative bacteria and 7 eukaryotic protozoa), using BLAST and kmer expansion to spread antigenicity to similar or identical proteins and peptides, balancing antigenic and non-antigenic datasets using *ROSE* (bootstrap based technique), and validating our models using a leave-one-out cross-validation method. This careful design across all stages led APRANK to successfully predict antigenicity for novel pathogens.

APRANK still has room for improvement. Removing some of the individual predictors showed little to no effect in the overall performance of APRANK, suggesting there might be some redundancy amongst the predictors being used. Also, there are a few bottlenecks regarding the computing performance of APRANK, most notably predictions by NetSurfP (a predictor of protein secondary structure and solvent accessibility for protein residues [168]). Inclusion of NetSurfP calculations incurs a heavy penalty on computing time, which was one of the main reasons that prevented offering APRANK as an online web-service. Also, installing APRANK in a new computer requires some expertise in Linux/Unix operating systems, mostly due to the necessary external dependencies and other third-party software required by APRANK (other predictors). We

cannot distribute some of these other predictors due to licensing issues, which means that we are not able to provide APRANK as a ready-to-use package (or virtual machine). Users interested in installing APRANK have to manually apply for academic licenses (or buy commercial licenses) for some of the third-party software. In summary this prevents APRANK being more widely available, as we would like.

Finally, APRANK is currently focused on finding linear epitopes, and likely missing most of the conformational ones. This is a consequence of the fact that there is much more experimental validation for linear epitopes than for conformational ones, which in turn means that most of the software that predicts and analyzes features is focused on linear features. The recent availability of more reliable structural predictions for proteins based on AlphaFold [231] led to fast improvement of predictors of conformational epitopes, such as DiscoTope [232]. Future versions of APRANK can certainly leverage more conformational and structural features as inputs for its predictions.

Using the method described in this thesis, a complete pathogen proteome can be analyzed in minutes-hours to obtain protein and peptide sets enriched in likely antigenic candidates for a number of downstream applications (see Chapter 2 for details).

High-density peptide microarrays: analyzing T. cruzi proteomes

Biological methods to discover and identify *T. cruzi*'s antigens and their epitopes have evolved greatly in the past decades, as summarized in Figure D1. In the 1960s and 1970s scientists accomplished this task using fractionation of parasite extracts, and obtaining as a result complex mixtures of *T. cruzi*'s antigens of unknown identity [233–236]. Parasite extracts as source of antigens are still being used in commercial or in-house serological assays (parasite lysate ELISAs), and more sophisticated versions still fill important application niches today, such as the TESA blot assay [217]. Developed in the 90s, the TESA blot is an immunoblot containing an extract of trypomastigote excreted-secreted antigens that is used to test the serum of patients for diagnosis of infection. This method is both sensitive and specific in cases of suspected acute or congenital Chagas disease [237]; however, because it is composed of antigens secreted by trypomastigotes (non-replicating parasite cells), it is more cumbersome to produce at large scale. In spite of this, there are today a number of commercial developments using TESA antigens as a base of indirect immunoassays [238, 239].

Over the years, developments on molecular biology enabled new ways of identifying antigens and their epitopes. Molecular cloning of random fragments of parasite DNA into phages that expressed the cloned DNA, followed by screening of these phage libraries using serum from infected patients, led to the identification of several antigenic clones [240]. After phage isolation and sequencing, the identification (sometimes guessing) of their reading frame and translation provided a number of relatively short well-defined antigenic protein sequences. While this allowed identification and cloning of the genes encoding these antigens, the epitopes themselves were only mapped with poor precision, except perhaps for repetitive antigens [206].

Another way of discovering new antigens and epitopes for Chagas disease is through the use of phage display techniques, a method which was developed in 1985, but that became popular many years later. In phage display short peptides are displayed on the capsid of phages. Because of the limited space available on viral capsid proteins to accommodate different inserts lengths, the first versions of this method used random short oligonucleotides instead of parasite-specific DNA fragments [241]. The resulting encoded protein fragments are then not parasite-specific, and any reactive peptides identified from reactive serum samples would only mimic the true (parasite) antigenic sequences. Hence, a further bioinformatic search is required to identify (or guess) the

corresponding parasite gene encoding the antigen. Despite these limitations, this method was used to identify relevant antigens and epitopes for several diseases, including Chagas disease [242, 243]. In more recent years, a combination of phage display with large-scale sequencing and antigen discovery bioinformatic analytical tools, made it possible to include longer parasite-specific DNA inserts, and develop a refined genomic-phage display (gPhage) for *T. cruzi* [244]. While this is a very promising tool, some peptides may end up underrepresented, such as those with structural constrains or those which are toxic to the phage. Also, since it is based on sequencing and mapping the antigenic peptides back to the pathogen's proteome, this method thrives when using pathogens with an already curated genome, and might have trouble when dealing with large proteomes with many similar proteins, such as hybrid strains [244].

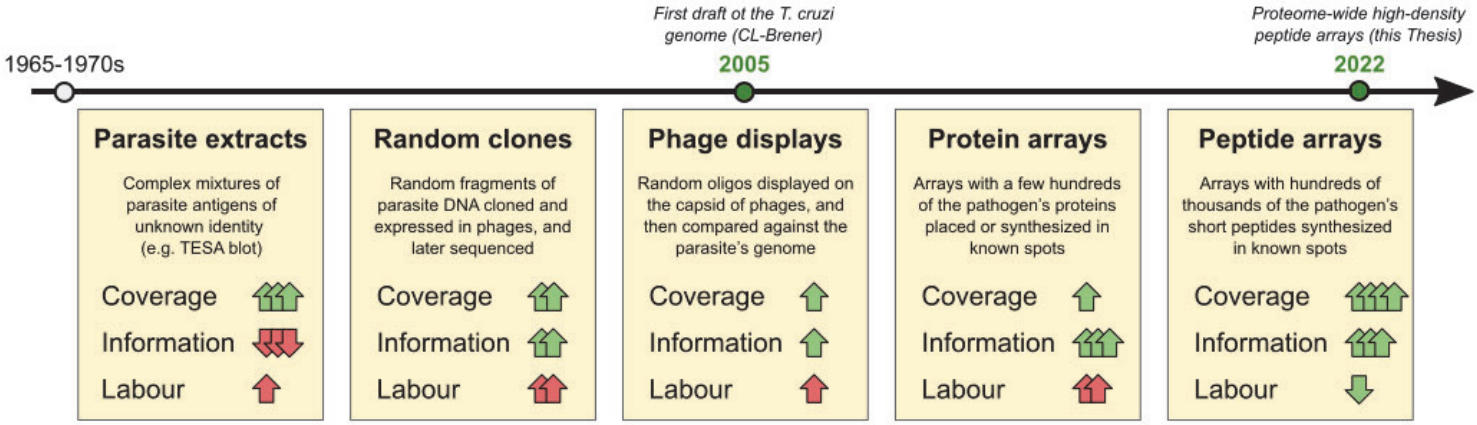


Figure D1 – ***Summary of methods used to discover pathogen's antigens.*** *Timeline showing a selection of methods capable of finding antigens from pathogens such as T. cruzi. The timeline is meant to be a schematic representation and is not at scale.* ***Coverage:*** *estimate of the proteome coverage of the method.* ***Information:*** *estimate of information gained on the identity of antigens and mapping of their epitopes.* ***Labour:*** *estimate of effort required to produce libraries and perform assays. Number and color of arrows shown for Coverage, Information and Labour are rough estimates based on our own analysis of the different methods (Green = good; Red = bad).*

As technology evolved, methods such as protein arrays or bead arrays allowed researchers to study the antigenicity of specific proteins, and to do so for many at the same time. These methods display up to a few hundred proteins at best, and hence require the selection of candidate proteins for inclusion in the array design from those encoded in the proteome. To the best of our knowledge, protein arrays have not been used for antigen discovery or profiling of immune responses in *T. cruzi* (Chagas disease); however, they have been extensively used for other pathogens such as *P. falciparum* (Malaria) [38, 245–247]. As for bead arrays, they have been used to study recombinant *T. cruzi* proteins known to be highly expressed in the parasite stages present in mammalian hosts [230], and those containing tandem repetitive regions [228]. While these methods provided a progress in the field, they requires a degree of effort that prevents scaling it up to complete proteomes.

The advent of peptide microarrays facilitated the study of antigenicity at a large scale [147, 248]. By allowing the simultaneous analysis of thousands to millions of peptides, peptide microarrays enabled faster and simpler ways of studying the complete proteomes of organisms, among many other uses. High-density peptide microarrays can also be divided in several sectors, allowing the analysis of identical "sub-arrays" in the same glass slide using different serum samples. This has the advantage of reducing operator and sample heterogeneity in the assays, as all sub-arrays are processed at the same time under identical conditions. All these advantages opened the door

to precise epitope mapping in the context of individual immune responses (single patient). This was considered when choosing peptide microarrays as a platform for large-scale experimental antigenicity assessment for this thesis. Furthermore, this allowed us to analyze the antigenicity of short peptides covering the whole proteome of two strains of *Trypanosoma cruzi*, the pathogen responsible for Chagas disease.

As mentioned in Chapter 3, high-density peptide microarrays have the disadvantage of usually missing most conformational epitopes, a consequence of analyzing relatively small linear peptides. Conformational epitopes are more likely to be detected using protein microarrays because they contain completely folded proteins or domains, and thus are more likely to expose conformational epitopes [249]. However, as previously discussed, protein microarrays also have limitations in both their coverage and the effort required to produce them. Hence, we consider that peptide microarrays are still the best adapted tool for high-throughput analysis of proteomes in most scenarios.

When analyzing large pathogen proteomes a trade-off has to be made in order to minimize overall costs and maximize proteome and patient (population) coverage. This is what led us to use a two-step screening strategy. In the first step, the proteins were split into overlapping peptides at medium resolution (4 amino acids between the start of each peptide) and they were analyzed using pools of serum samples. This provided us the opportunity to maximize proteome coverage to identify antigenic regions. Next, in a second step we focused on these antigenic regions (discarding non-reactive regions of proteins) and further split these regions into overlapping peptides now at maximum resolution for epitope mapping (1 amino acid between the start of each peptide). Furthermore, through the use of sectorized arrays we analyzed these selected antigenic regions on individual serum samples to assess the seroprevalence for each region and obtained a complete profile of antibody-specificities for each patient. Using this two-step method meant that we analyzed only 7% of the peptides we would have had to analyze if we had used individual serum samples and maximum resolution from the beginning.

By using 40 high-density peptide microarrays spanning two microarray designs (CHAGASTOPE-v1 and v2), we managed to analyze the proteome-wide antigenicity of 14 pools of sera and 71 individual sera, all in duplicate. In total we analyzed over 100 million peptide-serum interactions, which led us to find 3,868 non-redundant clusters of antigenic regions in the proteomes of two strains of *Trypanosoma cruzi*. Of those clusters, 98 were detected by at least 70% of the individuals, with 59 of them not showing similarity to previously known antigens. This list of 59 clusters of antigenic regions are of great interest for future studies on Chagas disease diagnostics. Several of these novel antigens were produced in our laboratory either as synthetic peptides or recombinant proteins, and are being used to develop and optimize new diagnostic reagents (work led by Emir Salas-Sarduy, unpublished). The information obtained on proteome-wide antigenicity was also very valuable as it provided us with a better understanding of how our immune system detects these type of pathogens. Coming full circle with the research presented in Chapter 2, we have now also produced massive information on both antigenic and non-antigenic peptides and proteins for *T. cruzi*. This experimentally derived positive and negative data is essential for improving antigenicity predictors. Finally, but not less important, all this freely available information was used to create searchable and interactive visualizations of the data to facilitate access to this information to all scientists working on Chagas disease (see Chapter 4).

Future Perspectives

"Our work has produced great answers, now someone just needs to figure out which questions they go with."
Randall Munroe, *Alt-text in XKCD 2652*

In this thesis we discussed the benefits and limitations of predictor models and high-throughput experimental screenings using microarrays, while showing a real-life example for each of them and the valuable data that we obtained in each case. These two methods of studying antigenicity do not compete with one another, but actually complement and enhance each other. It is very likely that future versions of APRANK, as well as other predictors, will use the data obtained in Chagastope as part of their training set, allowing them to achieve more accurate predictions. At the same time, predictors such as APRANK can help guide researchers to select which peptides to place in a microarray when dealing with large numbers of proteins or serum panels.

Prediction models: looking forward

The capacity of predictor models grows with computing power and data availability, two things that are ever increasing these days. Once enough data is collected, they will likely be very accurate at predicting the general antigenicity of linear epitopes. This is mostly true already for T-cell linear epitopes [85, 250, 251], but not yet for B-cell linear epitopes, which are the ones being studied in this thesis. While these models are good for predicting antigenicity in general, we will still need experimental assays to study the antigenicity of specific epitopes in any given individual.

Conformational epitopes, however, will likely take a lot more work and time to achieve similar levels of accuracy in predictions. This type of predictions depends on either knowing or being able to successfully predict the correct tertiary (and likely quaternary) structure of the protein under study. This is not a trivial matter, although recently the release of AlphaFold marked a big leap in the ability to predict the tertiary structure of proteins [231], especially since it can be run online for free via Google Colab. That AlphaFold is a game changer can be seen in the impact it had on the development of new and revised DiscoTope algorithms for the prediction of conformational epitopes [232].

Regarding APRANK itself, we have shown in this thesis that using APRANK the chance of finding antigenic peptides increases from ∼3 to ∼6 times versus simply selecting peptides at random; benefit that stems from looking at and combining different protein and peptide properties and features when predicting antigenicity. We hope that this observation is used by others in the future to create even better predictors or improve existing ones.

High-density peptide microarrays: looking forward

In this thesis we have shown that microarray technology is a very useful tool to perform high-throughput analysis of proteins and peptides, which allowed us to obtain a large set of *T. cruzi* peptides that are antigenic in the context of Chagas disease. Combining high-density peptide microarrays with these defined serological markers opens the door for many new avenues of analysis, such as studying how the proteome-wide antigenicity varies in time, likely related to the progress of a disease or the follow-up of a treatment [199, 252–254].

One possible use of these antigenic peptides then is as part of a follow-up in a clinical trial. It is known that immune responses are linked to parasite persistence [214], hence, chemotherapeutic treatments that lead to elimination of the parasite may be accompanied by reduction in antibody titres against defined antigens.

Another possible use of these peptides is studying how and/or if immune responses change during development of the disease. Chronic Chagas disease is usually asymptomatic but can progress to different disease stages, such as Chagas cardiomyopathy, in a fraction of the affected population. However, there is currently no prognostic markers that can predict which patients will develop cardiac or digestive pathologies, or when. Under the hypothesis that there is a controlled balance between the parasite and the host, the presence of antibodies may be used as a surrogate marker of the status of infection. Changes in antibody titres against all or specific antigens may be linked to appearance of pathology, and this is something that can now be investigated at large scale using the markers and epitopes defined in this work [88, 255, 256].

We are following both of these avenues in our research group, and it is likely that many others will follow similar research avenues in the coming years. Other uses of microarray technology are possible, but will depend on the price and availability of microarrays in the future. If they ever become common enough, they are a great tool for diagnosing diseases due to their high sensitivity and large capability, enabling studies such as wide spectrum tests to search for many diseases at once, or more detailed diagnostics that would return more information about the pathogen [257–259].

In our analysis of *Trypanosoma cruzi* using high-density peptide microarrays, we found 59 clusters of antigenic regions that were detected by at least 70% of the individuals and that did not show similarity to previously known antigens. Future studies by our lab or by others will likely analyze these regions in greater detail by trying to find the actual epitopes of each region with methods such as alanine scans, or by testing the regions *in vitro* using assays such as ELISA, among other possibilities.

Our analysis also revealed an interesting pattern of antigenicity, where most of the epitopes were private ones, meaning, they were detected by just a few individuals. This information can be combined with similar studies for other pathogens to determine if this kind of pattern is something ubiquitous, something related to large pathogens or something exclusive of *T. cruzi*. Whichever the answer, knowing this would increase our knowledge of our immune system.

In addition to what was concluded in this thesis and the corresponding papers, there is much information to be analyzed. We have released the full data set of signals for all ∼2.84 million short peptides spanning the complete proteomes of Sylvio X10 and CL Brener, as well as ∼240.000 short peptides spanning the antigenic regions with a higher detail. This information can be downloaded as standalone files or seen online in our website (see Chapter 4 and Supplementary Materials in Chapter 3). We consider this information is very valuable to train new predictors, deduce the actual epitope in a sequence, or learn about differences in antigenicity across different geographic regions.

Acronyms

ANN Artificial neural networks. 19

APRANK Antigenic Protein and Peptide Ranker. 2, 18, 33

ATL American tegumentary leishmaniasis. 31, 86

AUC Area under the ROC Curve. 41, 44, 46, 47, 51, 53, 57, 58, 89

BAC Bacterial artificial chromosome. 29

BCR B-cell receptor. 3–5

Boc tert-Butyloxycarbonyl protecting group. 12

CD Cluster of differentiation. 4, 5

cDNA Complementary DNA. 10

CL Cutaneous leishmaniasis. 31

CV Cross-validation. 20

DALY Disability-adjusted life year. 21

DGF Dispersed gene family. 30

DTU Discrete typing unit. 28, 29, 65, 84

ECG Electrocardiogram. 22, 26

ELISA Enzyme-linked immunosorbent assay. 6–9, 11, 13, 26, 31, 86, 118

FLISA Fluorescence-linked immunosorbent assay. 9, 13

Fmoc Fluorenylmethyloxycarbonyl protecting group. 12

GP Glycoprotein. 30

GPI Glycosyl phosphatidylinositol. 111

HAI Haemagglutination inhibition assay. 26, 86

HIV Human immunodeficiency virus. 6

HMM Hidden Markov Model. 54

IEDB Immune Epitope Database. 35, 55, 67, 77, 82, 89

IF Immunofluorescence. 8

IgG Immunoglobulin G. 4, 5, 9, 13, 25, 26

IgM Immunoglobulin M. 4, 5, 9

IIF Indirect immunofluorescence. 8, 26

IRLS Iterative reweighted least squares. 18

LOOCV Leave-one-out cross-validation. 44

LPS Lipopolysaccharide. 4

MASP Mucin-associated surface protein. 30

MHC Major histocompatibility complex. 3, 5, 19

ML Mucosal leishmaniasis. 31

MLE Maximum likelihood estimation. 18

NCBI National Center for Biotechnology Information. 35, 56

NPPOC 2-(2-nitrophenyl)prop-1-oxycarbonyl protecting group. 12, 85

NVOC 6-nitroveratryloxycarbonyl protecting group. 12

OLS Ordinary least squares. 15

PAHO Pan American Health Organization. 26

PCR Polymerase Chain Reaction. 10, 25, 27, 31

PGA Photogenerated acid. 12

PLOS Public Library of Science. 1

qPCR Quantitative Polymerase Chain Reaction. 25

RBBB Right bundle-branch block. 22

RHS Retrotransposon hot spot. 30

ROC Receiver operating characteristic. 2, 33, 41, 47, 54, 57, 89, 119

SMRT Single Molecular Real-Time. 30

SPPS Solid-phase peptide synthesis. 11

SSE Sum of the squared errors. 15

TCL T-cell line. 5

TCR T-cell receptor. 3–5

TESA Trypomastigote excreted-secreted antigens. 113

TL Tegumentary leishmaniasis. 31

TLR Toll-like receptors. 4

TS Trans-sialidase. 30

TSSA Trypomastigote small surface antigen. 29

VL Visceral leishmaniasis. 31

WHO World Health Organization. 1, 26, 27, 31

Bibliography

[1] Haendel, M. *et al.* How many rare diseases are there? *Nature Reviews Drug Discovery* **19**, 77–78 (2020). URL http://www.nature.com/articles/d41573-019-00180-y.

[2] Tambuyzer, E. *et al.* Therapies for rare diseases: therapeutic modalities, progress and challenges ahead. *Nature Reviews Drug Discovery* **19**, 93–111 (2020). URL https://www.nature.com/articles/s41573-019-0049-9.

[3] Hotez, P. J., Aksoy, S., Brindley, P. J. & Kamhawi, S. What constitutes a neglected tropical disease? *PLOS Neglected Tropical Diseases* **14**, e0008001 (2020). URL https://dx.plos.org/10.1371/journal.pntd.0008001.

[4] Organization, W. H. Neglected tropical diseases: impact of COVID-19 and WHO's response. *Weekly Epidemiological Record* **95**, 461 – 468 (2020). Publisher: World Health Organization.

[5] Kringelum, J. V., Nielsen, M., Padkjær, S. B. & Lund, O. Structural analysis of B-cell epitopes in antibody:protein complexes. *Mol Immunol* **53**, 24–34 (2013). URL http://dx.doi.org/10.1016/j.molimm.2012.06.001. Publisher: Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kemitorvet, Building 208, DK-2800 Lyngby, Denmark.

[6] Van Regenmortel, M. H. V. What is a B-cell epitope? *Methods Mol Biol* **524**, 3–20 (2009). URL http://dx.doi.org/10.1007/978-1-59745-450-6_1. Publisher: Ecole Supérieure de Biotechnologie de Strasbourg, Illkirch Cedex, France.

[7] Gómara, M. J. & Haro, I. Synthetic peptides for the immunodiagnosis of human diseases. *Curr Med Chem* **14**, 531–546 (2007). Publisher: Protein Chemistry, IIQAB-CSIC, Jordi Girona, 18-26 08034 Barcelona, Spain.

[8] Muller, S. Synthetic peptides as tools for diagnosis and therapeutic strategies to treat systemic lupus erythematous. *Autoimmun Rev* **11**, 799–800 (2012). URL http://dx.doi.org/10.1016/j.autrev.2012.02.008. Publisher: CNRS UPR9021, Institut de Biologie Moléculaire et Cellulaire, Strasbourg, France. S.Mueller@ibmc-cnrs.unistrafr.

[9] Van Regenmortel, M. H. Antigenicity and immunogenicity of synthetic peptides. *Biologicals* **29**, 209–213 (2001). URL http://dx.doi.org/10.1006/biol.2001.0308. Publisher: Ecole Supérieure de Biotechnologie de Strasbourg, Boulevard Sébastien Brandt, 67400 Illkirch, France. vanregen@esbs.u-strasbg.fr.

[10] Heiss, K. *et al.* Rapid Response to Pandemic Threats: Immunogenic Epitope Detection of Pandemic Pathogens for Diagnostics and Vaccine Development Using Peptide Microarrays. *Journal of Proteome Research* **19**, 4339–4354 (2020). URL https://pubs.acs.org/doi/10.1021/acs.jproteome.0c00484.

[11] Janeway, C., Janeway Jr, C., Travers, P., Walport, M. & Shlomchik, M. Chapter 3. Antigen recognition by B-cell and T-cell receptors. In *Immunobiology: The Immune System in Health and Disease* (Garland Science, New York, NY, USA, 2001), 5th edn.

[12] Janeway, C., Janeway Jr, C., Travers, P., Walport, M. & Shlomchik, M. Chapter 9. The Humoral Immune Response. In *Immunobiology: The Immune System in Health and Disease* (Garland Science, New York, NY, USA, 2001), 5th edn.

[13] Frank, S. A. *Immunology and evolution of infectious disease* (Princeton University Press, 2002). URL http://www.ncbi.nlm.nih.gov/books/NBK2394/.

[14] Frelinger, J. A. *Immunodominance: the choice of the immune system* (John Wiley & Sons, 2006).

[15] Abbott, R. K. & Crotty, S. Factors in B cell competition and immunodominance. *Immunological Reviews* **296**, 120–131 (2020). URL https://onlinelibrary.wiley.com/doi/10.1111/imr.12861.

[16] Rao, K. V. Selection in a T-dependent primary humoral response: new insights from polypeptide models. *APMIS* **107**, 807–818 (1999). Publisher: Immunology Group, International Centre for Genetic Engineering and Biotechnology, New Delhi, India.

[17] Hage, D. S. Immunoassays. *Analytical Chemistry* **71**, 294–304 (1999). URL https://pubs.acs.org/doi/10.1021/a1999901%2B.

[18] Yalow, R. S. & Berson, S. A. IMMUNOASSAY OF ENDOGENOUS PLASMA INSULIN IN MAN. *Journal of Clinical Investigation* **39**, 1157–1175 (1960). URL http://www.jci.org/articles/view/104130.

[19] Engvall, E. & Perlmann, P. Enzyme-linked immunosorbent assay (ELISA) quantitative assay of immunoglobulin G. *Immunochemistry* **8**, 871–874 (1971). URL https://linkinghub.elsevier.com/retrieve/pii/001927917190454X.

[20] Van Weemen, B. & Schuurs, A. Immunoassay using antigen-enzyme conjugates. *FEBS Letters* **15**, 232–236 (1971). URL http://doi.wiley.com/10.1016/0014-5793%2871%2980319-8.

[21] Shah, K. & Maghsoudlou, P. Enzyme-linked immunosorbent assay (ELISA): the basics. *British Journal of Hospital Medicine* **77**, C98–C101 (2016). URL http://www.magonlinelibrary.com/doi/10.12968/hmed.2016.77.7.C98.

[22] Joshi, S. & Yu, D. Immunofluorescence. In *Basic Science Methods for Clinical Researchers*, 135–150 (Elsevier, 2017). URL https://linkinghub.elsevier.com/retrieve/pii/B9780128030776000084.

[23] Magnusson, K.-E., Bartonek, E., Nordkvist, E., Sundqvist, T. & Asbrink, E. Fluorescence-Linked Immunosorbent Assay (FLISA) for Quantification of Antibodies to Food Antigens. *Immunological Investigations* **16**, 227–240 (1987). URL http://www.tandfonline.com/doi/full/10.3109/08820138709030578.

[24] Swartzman, E. E., Miraglia, S. J., Mellentin-Michelotti, J., Evangelista, L. & Yuan, P. M. A homogeneous and multiplexed immunoassay for high-throughput screening using fluorometric microvolume assay technology. *Analytical Biochemistry* **271**, 143–151 (1999).

[25] Yeo, S.-J. *et al.* Performance of coumarin-derived dendrimer-based fluorescence-linked immunosorbent assay (FLISA) to detect malaria antigen. *Malaria Journal* **13**, 266 (2014). URL https://malariajournal.biomedcentral.com/articles/10.1186/1475-2875-13-266.

[26] Chen, C.-S. & Zhu, H. Protein Microarrays. *BioTechniques* **40**, 423–429 (2006). URL https://www.future-science.com/doi/10.2144/06404TE01.

[27] Southern, E. M. DNA Microarrays: History and Overview. In *DNA Arrays*, vol. 170, 1–15 (Humana Press, New Jersey, 2001). URL http://link.springer.com/10.1385/1-59259-234-1:1.

[28] Kafatos, F. C., Jones, C. & Efstratiadis, A. Determination of nucleic acid sequence homologies and relative concentrations by a dot hybridization procedure. *Nucleic Acids Research* **7**, 1541–1552 (1979). URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/7.6.1541.

[29] Tse-Wen Chang. Binding of cells to matrixes of distinct antibodies coated on solid surface. *Journal of Immunological Methods* **65**, 217–223 (1983). URL https://linkinghub.elsevier.com/retrieve/pii/0022175983903186.

[30] Hoheisel, J. D., Ross, M. T., Zehetner, G. & Lehrach, H. Relational genome analysis using reference libraries and hybridisation fingerprinting. *Journal of Biotechnology* **35**, 121–134 (1994). URL https://linkinghub.elsevier.com/retrieve/pii/0168165694900310.

[31] Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science* **270**, 467–470 (1995). URL https://www.science.org/doi/10.1126/science.270.5235.467.

[32] Davies, D. H. *et al.* Profiling the humoral immune response to infection by using proteome microarrays: high-throughput vaccine and diagnostic antigen discovery. *Proc Natl Acad Sci U S A* **102**, 547–552 (2005). URL http://dx.doi.org/10.1073/pnas.0408782102. Publisher: Center for Virus Research, University of California, Irvine, CA 92697, USA.

[33] Doolan, D. L. Plasmodium immunomics. *Int J Parasitol* **41**, 3–20 (2011). URL http://dx.doi.org/10.1016/j.ijpara.2010.08.002. Publisher: Division of Immunology, Queensland Institute of Medical Research, The Bancroft Centre, 300 Herston Road, P.O. Royal Brisbane Hospital, Brisbane, QLD 4029, Australia. Denise.Doolan@qimr.edu.au.

[34] Liang, L. *et al.* Systems biology approach predicts antibody signature associated with Brucella melitensis infection in humans. *J Proteome Res* **10**, 4813–4824 (2011). URL http://dx.doi.org/10.1021/pr200619r. Publisher: Department of Medicine, Division of Infectious Diseases, University of California, Irvine, California 92697, United States.

[35] Lessa-Aquino, C. *et al.* Proteomic Features Predict Seroreactivity against Leptospiral Antigens in Leptospirosis Patients. *J Proteome Res* **14**, 549–556 (2014). URL http://dx.doi.org/10.1021/pr500718t. Publisher: Fiocruz, Bio-Manguinhos, Brazilian Ministry of Health , Avenida Brasil, 4365 - Manguinhos, Rio de Janeiro, RJ 21040-900, Brazil.

[36] Liu, J., Parrish, J. R., Hines, J., Mansfield, L. & Finley, R. L. A proteome-wide screen of Campylobacter jejuni using protein microarrays identifies novel and conformational antigens. *PLOS ONE* **14**, e0210351 (2019). URL https://dx.plos.org/10.1371/journal.pone.0210351.

[37] Pearson, M. S. *et al.* Immunomics-guided discovery of serum and urine antibodies for diagnosing urogenital schistosomiasis: a biomarker identification study. *The Lancet. Microbe* **2**, e617–e626 (2021).

[38] Bailey, J. A. *et al.* Seroreactivity to a large panel of field-derived Plasmodium falciparum apical membrane antigen 1 and merozoite surface protein 1 variants reflects seasonal and lifetime acquired responses to malaria. *The American Journal of Tropical Medicine and Hygiene* **92**, 9–12 (2015).

[39] Braun, P. & LaBaer, J. High throughput protein production for functional proteomics. *Trends in Biotechnology* **21**, 383–388 (2003). URL https://linkinghub.elsevier.com/retrieve/pii/S0167779903001896.

[40] Rosano, G. L., Morales, E. S. & Ceccarelli, E. A. New tools for recombinant protein production in *Escherichia coli* : A 5-year update. *Protein Science* **28**, 1412–1422 (2019). URL https://onlinelibrary.wiley.com/doi/10.1002/pro.3668.

[41] He, M. *et al.* Printing protein arrays from DNA arrays. *Nature Methods* **5**, 175–177 (2008). URL http://www.nature.com/articles/nmeth.1178.

[42] Hufnagel, K. *et al.* Immunoprofiling of Chlamydia trachomatis using whole-proteome microarrays generated by on-chip in situ expression. *Scientific Reports* **8**, 7503 (2018). URL https://www.nature.com/articles/s41598-018-25918-3.

[43] Szymczak, L. C., Kuo, H.-Y. & Mrksich, M. Peptide Arrays: Development and Application. *Analytical Chemistry* **90**, 266–282 (2018). URL https://pubs.acs.org/doi/10.1021/acs.analchem.7b04380.

[44] Barbulovic-Nad, I. *et al.* Bio-Microarray Fabrication Techniques—A Review. *Critical Reviews in Biotechnology* **26**, 237–259 (2006). URL http://www.tandfonline.com/doi/full/10.1080/07388550600978358.

[45] Kim, H. Y. *et al.* Characterization and simulation of cDNA microarray spots using a novel mathematical model. *BMC Bioinformatics* **8**, 485 (2007). URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-485.

[46] Geysen, H. M., Meloen, R. H. & Barteling, S. J. Use of peptide synthesis to probe viral antigens for epitopes to a resolution of a single amino acid. *Proceedings of the National Academy of Sciences of the United States of America* **81**, 3998–4002 (1984).

[47] Houghten, R. A. General method for the rapid solid-phase synthesis of large numbers of peptides: specificity of antigen-antibody interaction at the level of individual amino acids. *Proceedings of the National Academy of Sciences of the United States of America* **82**, 5131–5135 (1985).

[48] Frank, R. Spot-synthesis: an easy technique for the positionally addressable, parallel chemical synthesis on a membrane support. *Tetrahedron* **48**, 9217–9232 (1992). URL https://linkinghub.elsevier.com/retrieve/pii/S004040200185612X.

[49] Hilpert, K., Winkler, D. F. & Hancock, R. E. Peptide arrays on cellulose support: SPOT synthesis, a time and cost efficient method for synthesis of large numbers of peptides in a parallel and addressable fashion. *Nature Protocols* **2**, 1333–1349 (2007). URL http://www.nature.com/articles/nprot.2007.160.

[50] Beyer, M. *et al.* Combinatorial synthesis of peptide arrays onto a microchip. *Science (New York, N.Y.)* **318**, 1888 (2007).

[51] Fodor, S. P. *et al.* Light-directed, spatially addressable parallel chemical synthesis. *Science* **251**, 767–773 (1991). Publisher: Affymax Research Institute, Palo Alto, CA 94304.

[52] Bhushan, K. R., DeLisi, C. & Laursen, R. A. Synthesis of photolabile 2-(2-nitrophenyl)propyloxycarbonyl protected amino acids. *Tetrahedron Letters* **44**, 8585–8588 (2003). URL https://linkinghub.elsevier.com/retrieve/pii/S0040403903022688.

[53] Gao, X. *et al.* A flexible light-directed DNA chip synthesis gated by deprotection using solution photogenerated acids. *Nucleic Acids Research* **29**, 4744–4750 (2001).

[54] Pellois, J. P. *et al.* Individually addressable parallel peptide synthesis on microchips. *Nature Biotechnology* **20**, 922–926 (2002). URL http://dx.doi.org/10.1038/nbt723. Publisher: Department of Chemistry, University of Houston, Houston, TX 77004-5003, USA.

[55] Komolpis, K., Srivannavit, O. & Gulari, E. Light-directed simultaneous synthesis of oligopeptides on microarray substrate using a photogenerated acid. *Biotechnology Progress* **18**, 641–646 (2002).

[56] Legutki, J. B. *et al.* Scalable high-density peptide arrays for comprehensive health monitoring. *Nature Communications* **5**, 4785 (2014). URL http://dx.doi.org/10.1038/ncomms5785. Publisher: Center for Innovations in Medicine, Biodesign Institute, Arizona State University, Tempe, Arizona 85287, USA.

[57] Panicker, R. C., Huang, X. & Yao, S. Q. Recent advances in peptide-based microarray technologies. *Comb Chem High Throughput Screen* **7**, 547–556 (2004). Publisher: Department of Chemistry, National University of Singapore, 3 Science Drive 3, 117543, Republic of Singapore.

[58] Andresen, H. & Bier, F. F. Peptide microarrays for serum antibody diagnostics. *Methods Mol Biol* **509**, 123–134 (2009). URL http://dx.doi.org/10.1007/978-1-59745-372-1_8. Publisher: Fraunhofer Institut für Biomedizinische Technik, Institutsteil Potsdam, Potsdam, Germany.

[59] Forsström, B. *et al.* Proteome-wide Epitope Mapping of Antibodies Using Ultra-dense Peptide Arrays. *Molecular & Cellular Proteomics* **13**, 1585–1597 (2014). URL http://dx.doi.org/10.1074/mcp.M113.033308. Publisher: hlen@scilifelab.se.

[60] Mucci, J. *et al.* Next-generation ELISA diagnostic assay for Chagas Disease based on the combination of short peptidic epitopes. *PLOS Neglected Tropical Diseases* **11**, e0005972 (2017). URL http://dx.doi.org/10.1371/journal.pntd.0005972.

[61] Chapoval, A. I. *et al.* Immunosignature: Serum Antibody Profiling for Cancer Diagnostics. *Asian Pacific Journal of Cancer Prevention* **16**, 4833–4837 (2015). URL http://koreascience.or.kr/journal/view.jsp?kj=POCPA9&py=2015&vnc=v16n12&sp=4833.

[62] Legutki, J. B. & Johnston, S. A. Immunosignatures can predict vaccine efficacy. *Proceedings of the National Academy of Sciences* **110**, 18614–18619 (2013). URL https://pnas.org/doi/full/10.1073/pnas.1309390110.

[63] Stafford, P., Cichacz, Z., Woodbury, N. W. & Johnston, S. A. Immunosignature system for diagnosis of cancer. *Proceedings of the National Academy of Sciences of the United States of America* **111**, E3072–E3080 (2014). URL http://dx.doi.org/10.1073/pnas.1409432111. Publisher: Center for Innovations in Medicine, The Biodesign Institute, Arizona State University, Tempe, AZ 85287-5901.

[64] Stafford, P., Wrapp, D. & Johnston, S. A. General Assessment of Humoral Activity in Healthy Humans. *Molecular & Cellular Proteomics* **15**, 1610–1621 (2016). URL https://linkinghub.elsevier.com/retrieve/pii/S153594762033601X.

[65] Stafford, P. *et al.* Antibody characterization using immunosignatures. *PLOS ONE* **15**, e0229080 (2020). URL https://dx.plos.org/10.1371/journal.pone.0229080.

[66] Paull, M. L., Johnston, T., Ibsen, K. N., Bozekowski, J. D. & Daugherty, P. S. A general approach for predicting protein epitopes targeted by antibody repertoires using whole proteomes. *PLOS ONE* **14**, e0217668 (2019). URL https://dx.plos.org/10.1371/journal.pone.0217668.

[67] Pérez-Bercoff, L. *et al.* Whole CMV proteome pattern recognition analysis after HSCT identifies unique epitope targets associated with the CMV status. *PLoS One* **9**, e89648 (2014). URL http://dx.doi.org/10.1371/journal.pone.0089648. Publisher: CAST (Center for allogeneic stem cell transplantation), Karolinska Hospital; Division of Therapeutic Immunology (TIM), LabMed Karolinska Institutet, Stockholm, Sweden.

[68] Castro, A. *et al.* Investigation of humoral immune response towards persisting Epstein-Barr virus infections in multiple sclerosis and chronic fatigue syndrome using peptide microarrays (TECH1P. 868). *The Journal of Immunology* **192**, 69–36 (2014). Publisher: Am Assoc Immnol.

[69] Lagatie, O., Van Dorst, B. & Stuyver, L. J. Identification of three immunodominant motifs with atypical isotype profile scattered over the Onchocerca volvulus proteome. *PLOS Neglected Tropical Diseases* **11**, e0005330 (2017). URL https://dx.plos.org/10.1371/journal.pntd.0005330.

[70] Obiero, J. M. *et al.* Antibody Biomarkers Associated with Sterile Protection Induced by Controlled Human Malaria Infection under Chloroquine Prophylaxis. *mSphere* **4**, e00027–19 (2019). URL http://dx.doi.org/10.1128/{mSphereDirect}.00027-19.

[71] Lantz, B. *Machine learning with R: expert techniques for predictive modeling* (Packt Publishing, Birmingham Mumbai, 2019), third edition edn.

[72] Wagner, M., Adamczak, R., Porollo, A. & Meller, J. Linear Regression Models for Solvent Accessibility Prediction in Proteins. *Journal of Computational Biology* **12**, 355–369 (2005). URL http://www.liebertpub.com/doi/10.1089/cmb.2005.12.355.

[73] Lin, Z. & Pan, X.-M. Accurate Prediction of Protein Secondary Structural Content. *Journal of Protein Chemistry* **20**, 217–220 (2001). URL http://link.springer.com/10.1023/A:1010967008838.

[74] Qin, S. & Zhou, H.-X. meta-PPISP: a meta web server for protein-protein interaction site prediction. *Bioinformatics* **23**, 3386–3387 (2007). URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btm434.

[75] Gromiha, M. M., Thangakani, A. M. & Selvaraj, S. FOLD-RATE: prediction of protein folding rates from amino acid sequence. *Nucleic Acids Research* **34**, W70–W74 (2006). URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkl043.

[76] Diaz, A. A. *et al.* Prediction of protein solubility in *Escherichia coli* using logistic regression. *Biotechnology and Bioengineering* **105**, 374–383 (2010). URL https://onlinelibrary.wiley.com/doi/10.1002/bit.22537.

[77] Lee, H., Tu, Z., Deng, M., Sun, F. & Chen, T. Diffusion Kernel-Based Logistic Regression Models for Protein Function Prediction. *OMICS: A Journal of Integrative Biology* **10**, 40–55 (2006). URL http://www.liebertpub.com/doi/10.1089/omi.2006.10.40.

[78] Wan, S., Mak, M.-W. & Kung, S.-Y. mPLR-Loc: An adaptive decision multi-label classifier based on penalized logistic regression for protein subcellular localization prediction. *Analytical Biochemistry* **473**, 14–27 (2015). URL https://linkinghub.elsevier.com/retrieve/pii/S0003269714004965.

[79] Hilbe, J. M. *Practical guide to logistic regression* (2016). URL http://proxy.cm.umoncton.ca/login?url=http://lib.myilibrary.com?id=1004939. OCLC: 982237505.

[80] Salzberg, S. Locating Protein Coding Regions in Human DNA Using a Decision Tree Algorithm. *Journal of Computational Biology* **2**, 473–485 (1995). URL http://www.liebertpub.com/doi/10.1089/cmb.1995.2.473.

[81] Chen, H.-Y. *et al.* A Five-Gene Signature and Clinical Outcome in Non–Small-Cell Lung Cancer. *New England Journal of Medicine* **356**, 11–20 (2007). URL http://www.nejm.org/doi/abs/10.1056/NEJMoa060096.

[82] Zhang, L. V., Wong, S. L., King, O. D. & Roth, F. P. Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinformatics* **5**, 38 (2004). URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-5-38.

[83] Geurts, P., Irrthum, A. & Wehenkel, L. Supervised learning with decision tree-based methods in computational and systems biology. *Molecular BioSystems* **5**, 1593 (2009). URL http://xlink.rsc.org/?DOI=b907946g.

[84] Wardah, W., Khan, M., Sharma, A. & Rashid, M. A. Protein secondary structure prediction using neural networks and deep learning: A review. *Computational Biology and Chemistry* **81**, 1–8 (2019). URL https://linkinghub.elsevier.com/retrieve/pii/S1476927118305012.

[85] Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Research* **48**, W449–W454 (2020). URL https://academic.oup.com/nar/article/48/W1/W449/5837056.

[86] Rassi, A., Rassi, A. & Marin-Neto, J. A. Chagas disease. *Lancet* **375**, 1388–1402 (2010). URL http://dx.doi.org/10.1016/S0140-6736(10)60061-X. Publisher: Division of Cardiology, Anis Rassi Hospital, Goiânia, GO, Brazil. arassijr@terra.com.br.

[87] Pérez-Molina, J. A. & Molina, I. Chagas disease. *The Lancet* **391**, 82–94 (2018). URL https://linkinghub.elsevier.com/retrieve/pii/S0140673617316124.

[88] Nunes, M. C. P. *et al.* Chagas Cardiomyopathy: An Update of Current Clinical Knowledge and Management: A Scientific Statement From the American Heart Association. *Circulation* **138** (2018). URL https://www.ahajournals.org/doi/10.1161/CIR.0000000000000599.

[89] Lidani, K. C. F. *et al.* Chagas Disease: From Discovery to a Worldwide Health Problem. *Frontiers in Public Health* **7**, 166 (2019). URL https://www.frontiersin.org/article/10.3389/fpubh.2019.00166/full.

[90] Schmunis, G. A. Epidemiology of Chagas disease in non-endemic countries: the role of international migration. *Mem Inst Oswaldo Cruz* **102 Suppl 1**, 75–85 (2007). Publisher: Pan American Health Organization/World Health Organization, 525 23rd Street, NW Washington, DC 20037, USA. schmunig@paho.org.

[91] Gascon, J., Bern, C. & Pinazo, M.-J. Chagas disease in Spain, the United States and other non-endemic countries. *Acta Trop* **115**, 22–27 (2010). URL http://dx.doi.org/10.1016/j.actatropica.2009.07.019. Publisher: Centre de Salut Internacional, CRESIB, Hospital Clínic, IDIBAPS, c/Villarroel, 170, 08036, Barcelona, Spain. jgascon@clinic.ub.es.

[92] GBD 2016 Causes of Death Collaborators. Global, regional, and national age-sex specific mortality for 264 causes of death, 1980-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet (London, England)* **390**, 1151–1210 (2017).

[93] Alonso-Padilla, J. *et al.* Strategies to enhance access to diagnosis and treatment for Chagas disease patients in Latin America. *Expert Review of Anti-infective Therapy* **17**, 145–157 (2019). URL https://www.tandfonline.com/doi/full/10.1080/14787210.2019.1577731.

[94] World Health Organization. The global burden of disease : 2004 update 146 (2008). URL https://apps.who.int/iris/handle/10665/43942. Place: Geneva Publisher: World Health Organization.

[95] Lee, B. Y., Bacon, K. M., Bottazzi, M. E. & Hotez, P. J. Global economic burden of Chagas disease: a computational simulation model. *The Lancet. Infectious Diseases* **13**, 342–348 (2013).

[96] Walker, D. Principles of Diagnosis of Infectious Diseases. In *Pathobiology of Human Disease*, 222–225 (Elsevier, 2014). URL https://linkinghub.elsevier.com/retrieve/pii/B9780123864567017135.

[97] Luquetti, A. & Schmuñis, G. Diagnosis of Trypanosoma cruzi infection. In *American Trypanosomiasis Chagas Disease*, 687–730 (Elsevier, 2017). URL https://linkinghub.elsevier.com/retrieve/pii/B9780128010297000307.

[98] Brasil, P. E. A. A., De Castro, L., Hasslocher-Moreno, A. M., Sangenis, L. H. C. & Braga, J. U. ELISA versus PCR for diagnosis of chronic Chagas disease: systematic review and meta-analysis. *BMC Infect Dis* **10**, 337 (2010). URL http://dx.doi.org/10.1186/1471-2334-10-337. Publisher: Instituto de Pesquisa Clínica Evandro Chagas - Fundação Oswaldo Cruz, Rio de Janeiro/RJ, Brazil. pedro.brasil@ipec.fiocruz.br.

[99] Schijman, A. G. *et al.* International study to evaluate PCR methods for detection of Trypanosoma cruzi DNA in blood samples from Chagas disease patients. *PLoS Negl Trop Dis* **5**, e931 (2011). URL http://dx.doi.org/10.1371/journal.pntd.0000931. Publisher: Laboratorio de Biología Molecular de Enfermedad de Chagas, Instituto de Investigaciones en Ingeniería Genética y Biología Molecular (INGEBI-CONICET), Buenos Aires, Argentina. schijman@dna.uba.ar.

[100] Gomes, Y. M., Lorena, V. M. B. & Luquetti, A. O. Diagnosis of Chagas disease: what has been achieved? What remains to be done with regard to diagnosis and follow up studies? *Mem Inst Oswaldo Cruz* **104 Suppl 1**, 115–121 (2009). Publisher: Laboratório de Imunoparasitologia, Departamento de Imunologia, Centro de Pesquisas Aggeu Magalhães-Fiocruz, Recife, PE, Brasil. yara@cpqam.fiocruz.br.

[101] Carlier, Y. *et al.* Congenital Chagas disease: Updated recommendations for prevention, diagnosis, treatment, and follow-up of newborns and siblings, girls, women of childbearing age, and pregnant women. *PLOS Neglected Tropical Diseases* **13**, e0007694 (2019). URL https://dx.plos.org/10.1371/journal.pntd.0007694.

[102] Carlier, Y., Sosa-Estani, S., Luquetti, A. O. & Buekens, P. Congenital Chagas disease: an update. *Memorias Do Instituto Oswaldo Cruz* **110**, 363–368 (2015).

[103] Benatar, A. F. *et al.* Prospective multicenter evaluation of real time PCR Kit prototype for early diagnosis of congenital Chagas disease. *eBioMedicine* **69**, 103450 (2021). URL https://linkinghub.elsevier.com/retrieve/pii/S2352396421002437.

[104] Alonso-Padilla, J., Gallego, M., Schijman, A. G. & Gascon, J. Molecular diagnostics for Chagas disease: up to date and novel methodologies. *Expert Review of Molecular Diagnostics* **17**, 699–710 (2017). URL https://www.tandfonline.com/doi/full/10.1080/14737159.2017.1338566.

[105] da Silveira, J. F., Umezawa, E. S. & Luquetti, A. O. Chagas disease: recombinant Trypanosoma cruzi antigens for serological diagnosis. *Trends in Parasitology* **17**, 286–291 (2001). URL https://linkinghub.elsevier.com/retrieve/pii/S1471492201018979.

[106] Umezawa, E. *et al.* TESA-blot for the diagnosis of Chagas disease in dogs from co-endemic regions for Trypanosoma cruzi, Trypanosoma evansi and Leishmania chagasi. *Acta*

Tropica **111**, 15–20 (2009). URL https://linkinghub.elsevier.com/retrieve/pii/S0001706X09000084.

[107] Pan American Health Organization & World Health Organization. *Guidelines for the diagnosis and treatment of Chagas disease* (2019), pan american health organization edn.

[108] Coura, J. R. Present situation and new strategies for Chagas disease chemotherapy: a proposal. *Memórias do Instituto Oswaldo Cruz* **104**, 549–554 (2009). URL http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0074-02762009000400002&lng=en&tlng=en.

[109] Cançado, J. R. Criteria of Chagas disease cure. *Mem Inst Oswaldo Cruz* **94 Suppl 1**, 331–335 (1999). Publisher: Cadeira de Terapêutica Clínica, Universidade Federal de Minas Gerais, Belo Horizonte, MG, 31270-901, Brasil.

[110] Britto, C. C. Usefulness of PCR-based assays to assess drug efficacy in Chagas disease chemotherapy: value and limitations. *Mem Inst Oswaldo Cruz* **104 Suppl 1**, 122–135 (2009). Publisher: Laboratório de Biologia Molecular e Doenças Endêmicas, Instituto Oswaldo Cruz-Fiocruz, Rio de Janeiro, RJ, Brasil. cbritto@ioc.fiocruz.br.

[111] Médecins Sans Frontières. International meeting: new diagnostic tests are urgently needed to treat patients with Chagas disease. *Rev Soc Bras Med Trop* **41**, 315–319 (2008).

[112] Zingales, B. *et al.* A new consensus for Trypanosoma cruzi intraspecific nomenclature: second revision meeting recommends TcI to TcVI. *Mem Inst Oswaldo Cruz* **104**, 1051–1054 (2009). Publisher: Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, São Paulo, SP, Brasil. zingales@iq.usp.br.

[113] Zingales, B. Trypanosoma cruzi genetic diversity: Something new for something known about Chagas disease manifestations, serodiagnosis and drug sensitivity. *Acta Tropica* **184**, 38–52 (2018). URL https://linkinghub.elsevier.com/retrieve/pii/S0001706X17304266.

[114] Enriquez, G. F. *et al.* Discrete typing units of *Trypanosoma cruzi* identified in rural dogs and cats in the humid Argentinean Chaco. *Parasitology* **140**, 303–308 (2013). URL https://www.cambridge.org/core/product/identifier/S003118201200159X/type/journal_article.

[115] Monje-Rumi, M. M. *et al.* Trypanosoma cruzi diversity in the Gran Chaco: Mixed infections and differential host distribution of TcV and TcVI. *Infection, Genetics and Evolution* **29**, 53–59 (2015). URL https://linkinghub.elsevier.com/retrieve/pii/S1567134814004043.

[116] Garcia, M. N. *et al.* Molecular identification and genotyping of Trypanosoma cruzi DNA in autochthonous Chagas disease patients from Texas, USA. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* **49**, 151–156 (2017).

[117] Murillo-Solano, C. *et al.* Diversity of Trypanosoma cruzi parasites infecting Triatoma dimidiata in Central Veracruz, Mexico, and their One Health ecological interactions. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* **95**, 105050 (2021).

[118] Pronovost, H. *et al.* Deep sequencing reveals multiclonality and new discrete typing units of Trypanosoma cruzi in rodents from the southern United States. *Journal of Microbiology, Immunology, and Infection = Wei Mian Yu Gan Ran Za Zhi* **53**, 622–633 (2020).

[119] Verani, J. R. *et al.* Geographic variation in the sensitivity of recombinant antigen-based rapid tests for chronic Trypanosoma cruzi infection. *Am J Trop Med Hyg* **80**, 410–415 (2009). Publisher: Division of Parasitic Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia 30333, USA. jverani@cdc.gov.

[120] 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). Publisher: Departamento de Morfologia, Instituto de Ciências Biológicas, Belo Horizonte, UFMG, Minas Gerais, Brazil.

[121] Balouz, V. *et al.* Serological Approaches for Trypanosoma cruzi Strain Typing. *Trends in Parasitology* **37**, 214–225 (2021). URL https://www.sciencedirect.com/science/article/pii/S1471492220303445.

[122] Di Noia, J. M., Buscaglia, C. A., Marchi, C. R. D., Almeida, I. C. & Frasch, A. C. C. A Trypanosoma cruzi small surface molecule provides the first immunological evidence that Chagas' disease is due to a single parasite lineage. *J Exp Med* **195**, 401–413 (2002). Publisher: Instituto de Investigaciones Biotecnológicas-Instituto Tecnológico de Chascomús (IIB-INTECH), Universidad Nacional de General San Martín/CONICET, Av. General Paz y Albarellos, San Martin, 1650 Buenos Aires, Argentina.

[123] Burgos, J. M. *et al.* Molecular identification of Trypanosoma cruzi discrete typing units in end-stage chronic Chagas heart disease and reactivation after heart transplantation. *Clin Infect Dis* **51**, 485–495 (2010). URL http://dx.doi.org/10.1086/655680. Publisher: Laboratorio de Biología Molecular de la Enfermedad de Chagas, Instituto de Ingeniería Genética y Biología Molecular, Universidad de Buenos Aires, Ciudad Autónoma de Buenos Aires, Argentina.

[124] Bhattacharyya, T. *et al.* Development of peptide-based lineage-specific serology for chronic Chagas disease: geographical and clinical distribution of epitope recognition. *PLoS Negl Trop Dis* **8**, e2892 (2014). URL http://dx.doi.org/10.1371/journal.pntd.0002892. Publisher: Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, United Kingdom.

[125] El-Sayed, N. M. *et al.* The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease. *Science* **309**, 409–415 (2005). URL http://dx.doi.org/10.1126/science.1112631. Publisher: Department of Parasite Genomics, Institute for Genomic Research, Rockville, MD 20850, USA. nelsayed@tigr.org.

[126] Wang, W. *et al.* Strain-specific genome evolution in Trypanosoma cruzi, the agent of Chagas disease. *PLoS pathogens* **17**, e1009254 (2021).

[127] Berná, L. *et al.* Expanding an expanded genome: long-read sequencing of Trypanosoma cruzi. *Microbial Genomics* **4** (2018).

[128] Grisard, E. C. *et al.* Trypanosoma cruzi Clone Dm28c Draft Genome Sequence. *Genome Announc* **2**, e01114–13 (2014). URL http://dx.doi.org/10.1128/genomeA.01114-13. Publisher: Departamento de Microbiologia, Imunologia e Parasitologia, Universidade Federal de Santa Catarina, Florianópolis, Brazil.

[129] Franzén, O. *et al.* Shotgun Sequencing Analysis of Trypanosoma cruzi I Sylvio X10/1 and Comparison with T. cruzi VI CL Brener. *PLoS Neglected Tropical Diseases* **5**, e984 (2011). URL https://dx.plos.org/10.1371/journal.pntd.0000984. Publisher: Science for Life Laboratory, Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden. oscar.franzen@scilifelab.se.

[130] Franzén, O. *et al.* Comparative genomic analysis of human infective Trypanosoma cruzi lineages with the bat-restricted subspecies T. cruzi marinkellei. *BMC Genomics* **13**, 531 (2012). URL http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-13-531.

[131] Weatherly, D. B., Boehlke, C. & Tarleton, R. L. Chromosome level assembly of the hybrid Trypanosoma cruzi genome. *BMC Genomics* **10**, 255 (2009). URL http://dx.doi.org/10.1186/1471-2164-10-255.

[132] Vexenat, A. d. C., Santana, J. M. & Teixeira, A. R. Cross-reactivity of antibodies in human infections by the kinetoplastid protozoa Trypanosoma cruzi, Leishmania chagasi and Leishmania (Viannia) braziliensis. *Revista do Instituto de Medicina Tropical de São Paulo* **38**, 177–185 (1996). URL http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0036-46651996000300003&lng=en&tlng=en.

[133] Frank, F. M. *et al.* Characterization of human infection by *Leishmania* spp. in the Northwest of Argentina: immune response, double infection with *Trypanosoma cruzi* and species of *Leishmania* involved. *Parasitology* **126**, 31–39 (2003). URL https://www.cambridge.org/core/product/identifier/S0031182002002585/type/journal_article.

[134] Stuart, K. *et al.* Kinetoplastids: related protozoan pathogens, different diseases. *Journal of Clinical Investigation* **118**, 1301–1310 (2008). URL http://www.jci.org/articles/view/33945.

[135] Lessa, M. M. *et al.* Mucosal leishmaniasis: A Retrospective Study of 327 Cases from an Endemic Area of Leishmania (Viannia) braziliensis. *The American Journal of Tropical Medicine and Hygiene* **97**, 761–766 (2017). URL https://ajtmh.org/doi/10.4269/ajtmh.16-0349.

[136] *Manual de Vigilância da Leishmaniose Tegumentar Americana* (Ministéro da Saúde, Secretaria de Vigilância em Saúde, Departamento de Vigilância das Doenças Transmissíveis, Brasilia - DF, 2010). URL https://bvsms.saude.gov.br/bvs/publicacoes/manual_vigilancia_leishmaniose_tegumentar_americana.pdf.

[137] Carvalho, S. H. *et al.* American tegumentary leishmaniasis in Brazil: a critical review of the current therapeutic approach with systemic meglumine antimoniate and short□term possibilities for an alternative treatment. *Tropical Medicine & International Health* **24**, 380–391 (2019). URL https://onlinelibrary.wiley.com/doi/10.1111/tmi.13210.

[138] Carstens-Kass, J., Paulini, K., Lypaczewski, P. & Matlashewski, G. A review of the leishmanin skin test: A neglected test for a neglected disease. *PLOS Neglected Tropical Diseases* **15**, e0009531 (2021). URL https://dx.plos.org/10.1371/journal.pntd.0009531.

[139] González-Marcano, E., Kato, H., Concepción, J. L., Márquez, M. E. & Mondolfi, A. P. Polymerase Chain Reaction Diagnosis of Leishmaniasis: A Species-Specific Approach. In Luthra, R., Singh, R. R. & Patel, K. P. (eds.) *Clinical Applications of PCR*, vol. 1392, 113–124 (Springer New York, New York, NY, 2016). URL http://link.springer.com/10.1007/978-1-4939-3360-0_11. Series Title: Methods in Molecular Biology.

[140] Bracamonte, M. E. *et al.* High performance of an enzyme linked immunosorbent assay for American tegumentary leishmaniasis diagnosis with Leishmania (Viannia) braziliensis amastigotes membrane crude antigens. *PLOS ONE* **15**, e0232829 (2020). URL https://dx.plos.org/10.1371/journal.pone.0232829.

[141] Ricci, A. D. *et al.* APRANK: Computational Prioritization of Antigenic Proteins and Peptides From Complete Pathogen Proteomes. *Frontiers in Immunology* **12** (2021). URL https://www.frontiersin.org/articles/10.3389/fimmu.2021.702552/full.

[142] Peeling, R. W. & Nwaka, S. Drugs and diagnostic innovations to improve global health. *Infectious disease clinics of North America* **25**, 693–705, xi (2011). URL http://dx.doi.org/10.1016/j.idc.2011.06.002. Publisher: Department of Clinical Research, Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK. rosanna.peeling@lshtm.ac.uk.

[143] Washington, J. A. Principles of Diagnosis. In Baron, S. (ed.) *Medical Microbiology* (University of Texas Medical Branch at Galveston, Galveston, TX, 1996), 4th edn. URL https://www.ncbi.nlm.nih.gov/books/NBK8014/. Section: 10.

[144] Vainionpää, R. & Leinikki, P. Diagnostic Techniques: Serological and Molecular Approaches. *Encyclopedia of Virology* 29–37 (2008). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7150332/. Edition: 2008/07/30.

[145] Sutandy, F. X. R., Qian, J., Chen, C.-S. & Zhu, H. Overview of protein microarrays. *Current protocols in protein science* **Chapter 27**, Unit 27.1 (2013).

[146] Buus, S. *et al.* High-resolution Mapping of Linear Antibody Epitopes Using Ultrahigh-density Peptide Microarrays. *Molecular & Cellular Proteomics* **11**, 1790–1800 (2012). URL http://dx.doi.org/10.1074/mcp.M112.020800. Publisher: Laboratory of Experimental Immunology, University of Copenhagen, Copenhagen N, Denmark. sbuus@sund.ku.dk.

[147] Carmona, S. J. *et al.* Towards High-throughput Immunomics for Infectious Diseases: Use of Next-generation Peptide Microarrays for Rapid Discovery and Mapping of Antigenic Determinants. *Molecular & Cellular Proteomics* **14**, 1871–1884 (2015). URL https://linkinghub.elsevier.com/retrieve/pii/S1535947620328541.

[148] Durante, I. M., La Spina, P. E., Carmona, S. J., Agüero, F. & Buscaglia, C. A. High-resolution profiling of linear B-cell epitopes from mucin-associated surface proteins (MASPs) of Trypanosoma cruzi during human infections. *PLoS neglected tropical diseases* **11**, e0005986 (2017).

[149] Carmona, S. J., Sartor, P. A., Leguizamón, M. S., Campetella, O. E. & Agüero, F. Diagnostic Peptide Discovery: Prioritization of Pathogen Diagnostic Markers Using Multiple Features. *PLoS One* **7**, e50748 (2012). URL http://dx.doi.org/10.1371/journal.pone.0050748. Publisher: Instituto de Investigaciones Biotecnológicas, Instituto Tecnológico de Chascomús IIB-INTECH, Universidad Nacional de San Martín, Consejo de Investigaciones Científicas y Técnicas UNSAM-CONICET, Sede San Martín, San Martín, Buenos Aires, Argentina.

[150] Liu, E. W. *et al.* Protein-Specific Features Associated with Variability in Human Antibody Responses to Plasmodium falciparum Malaria Antigens. *The American journal of tropical medicine and hygiene* **98**, 57–66 (2018). URL https://pubmed.ncbi.nlm.nih.gov/29141757. Publisher: The American Society of Tropical Medicine and Hygiene.

[151] Liang, L. & Felgner, P. L. Predicting antigenicity of proteins in a bacterial proteome; a protein microarray and naïve Bayes classification approach. *Chemistry & biodiversity* **9**, 977–990 (2012). URL http://dx.doi.org/10.1002/cbdv.201100360. Publisher: Department of Medicine, Division of Infectious Diseases, University of California, Irvine, CA 92697, USA.

[152] Dalsass, M., Brozzi, A., Medini, D. & Rappuoli, R. Comparison of Open-Source Reverse Vaccinology Programs for Bacterial Vaccine Antigen Discovery. *Frontiers in immunology* **10**, 113 (2019). URL http://dx.doi.org/10.3389/fimmu.2019.00113.

[153] Vita, R. *et al.* The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Research* **47**, D339–D343 (2019).

[154] Martini, S., Nielsen, M., Peters, B. & Sette, A. The Immune Epitope Database and Analysis Resource Program 2003-2018: reflections and outlook. *Immunogenetics* **72**, 57–76 (2020).

[155] Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009). URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-421. Publisher: BioMed Central.

[156] Larsen, J., Lund, O. & Nielsen, M. Improved method for predicting linear B-cell epitopes. *Immunome Research* **2**, 2 (2006). URL http://dx.doi.org/10.1186/1745-7580-2-2.

[157] Nielsen, M., Justesen, S., Lund, O., Lundegaard, C. & Buus, S. NetMHCIIpan-2.0 - Improved silveirapan-specific HLA-DR predictions using a novel concurrent alignment and weight optimization training procedure. *Immunome Research* **6**, 9 (2010). URL http://www.ncbi.nlm.nih.gov/pubmed/21073747.

[158] Julenius, K., Mølgaard, A., Gupta, R. & Brunak, S. Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. *Glycobiology* **15**, 153–164 (2005). URL http://dx.doi.org/10.1093/glycob/cwh151. Publisher: Center for Biological Sequence Analysis, BioCentrum, Building 208, Technical University of Denmark, DK-2800 Lyngby, Denmark. karin.julenius@sbc.su.se.

[159] Pierleoni, A., Martelli, P. L. & Casadio, R. PredGPI: a GPI-anchor predictor. *BMC Bioinformatics* **9**, 392 (2008). URL http://dx.doi.org/10.1186/1471-2105-9-392.

[160] Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods* **8**, 785–786 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/21959131http://www.nature.com/articles/nmeth.1701.

[161] Newman, A. M. & Cooper, J. B. XSTREAM: A practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *BMC Bioinformatics* **8**, 382 (2007). URL http://www.ncbi.nlm.nih.gov/pubmed/17931424.

[162] Dosztányi, Z. Prediction of protein disorder based on IUPred. *Protein science : a publication of the Protein Society* **27**, 331–340 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/29076577. Edition: 2017/11/16 Publisher: John Wiley and Sons Inc.

[163] McDonnell, A. V., Jiang, T., Keating, A. E. & Berger, B. Paircoil2: improved prediction of coiled coils from sequence. *Bioinformatics* **22**, 356–358 (2006). URL http://www.ncbi.nlm.nih.gov/pubmed/16317077.

[164] Krogh, A., Larsson, B., Heijne, G. v. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology* **305**, 567–580 (2001). URL http://dx.doi.org/10.1006/jmbi.2000.4315.

[165] Lunardon, N., Menardi, G. & Torelli, N. ROSE: A Package for Binary Imbalanced Learning. *The R Journal* **6**, 79–89 (2014). URL https://journal.r-project.org/archive/2014/RJ-2014-008/RJ-2014-008.pdf.

[166] Menardi, G. & Torelli, N. Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery* **28**, 92–122 (2014). URL http://link.springer.com/10.1007/s10618-012-0295-5. Publisher: Springer US.

[167] Holmes, P. Neglected tropical diseases in the post-2015 health agenda. *Lancet* **383**, 1803 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24856022.

[168] Klausen, M. S. *et al.* NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins* **87**, 520–527 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/30785653. Edition: 2019/03/09 Place: United States.

[169] Xu, Y., Bruno, J. F. & Luft, B. J. Profiling the humoral immune response to Borrelia burgdorferi infection with protein microarrays. *Microbial Pathogenesis* **45**, 403–407 (2008). URL http://www.ncbi.nlm.nih.gov/pubmed/18976702.

[170] Barbour, A. G. *et al.* A Genome-Wide Proteome Array Reveals a Limited Set of Immunogens in Natural Infections of Humans and White-Footed Mice with Borrelia burgdorferi. *Infection and Immunity* **76**, 3374–3389 (2008). URL http://www.ncbi.nlm.nih.gov/pubmed/18474646.

[171] Richer, J., Johnston, S. A. & Stafford, P. Epitope Identification from Fixed-complexity Random-sequence Peptide Microarrays. *Molecular & Cellular Proteomics* **14**, 136–147 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25368412.

[172] Lawrenz, M. B. *et al.* Human antibody responses to VlsE antigenic variation protein of Borrelia burgdorferi. *Journal of clinical microbiology* **37**, 3997–4004 (1999). URL http://www.ncbi.nlm.nih.gov/pubmed/10565921.

[173] Eyles, J. E. *et al.* ImmunodominantFrancisella tularensis antigens identified using proteome microarray.©Crown Copyright 2007 Dstl. *PROTEOMICS* **7**, 2172–2183 (2007). URL http://www.ncbi.nlm.nih.gov/pubmed/17533643.

[174] Lu, Z. *et al.* Generation and characterization of hybridoma antibodies for immunotherapy of tularemia. *Immunology Letters* **112**, 92–103 (2007). URL http://www.ncbi.nlm.nih.gov/pubmed/17764754.

[175] Kilmury, S. L. N. & Twine, S. M. The francisella tularensis proteome and its recognition by antibodies. *Frontiers in microbiology* **1**, 143 (2010). URL http://www.ncbi.nlm.nih.gov/pubmed/21687770. Publisher: Frontiers Media SA.

[176] Beare, P. A. *et al.* Candidate Antigens for Q Fever Serodiagnosis Revealed by Immunoscreening of a Coxiella burnetii Protein Microarray. *Clinical and Vaccine Immunology* **15**, 1771–1779 (2008). URL http://www.ncbi.nlm.nih.gov/pubmed/18845831.

[177] Wang, X., Xiong, X., Graves, S., Stenos, J. & Wen, B. Protein array of Coxiella burnetii probed with Q fever sera. *Science China Life Sciences* **56**, 453–459 (2013). URL http://www.ncbi.nlm.nih.gov/pubmed/23633077.

[178] Xiong, X., Wang, X., Wen, B., Graves, S. & Stenos, J. Potential serodiagnostic markers for Q fever identified in Coxiella burnetii by immunoproteomic and protein microarray approaches. *BMC Microbiology* **12**, 35 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/22420424.

[179] Vigil, A. *et al.* Profiling the Humoral Immune Response of Acute and Chronic Q Fever by Protein Microarray. *Molecular & Cellular Proteomics* **10**, M110.006304 (2011). URL http://dx.doi.org/10.1074/mcp.M110.006304. Publisher: Department of Medicine, Division of Infectious Diseases, University of California, Irvine, CA 92697, USA. vigila@uci.edu.

[180] Chen, C. *et al.* A systematic approach to evaluate humoral and cellular immune responses to Coxiella burnetii immunoreactive antigens. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases* **15 Suppl 2**, 156–7 (2009). URL http://www.ncbi.nlm.nih.gov/pubmed/19281461. Publisher: NIH Public Access.

[181] Liang, L. *et al.* Large Scale Immune Profiling of Infected Humans and Goats Reveals Differential Recognition of Brucella melitensis Antigens. *PLoS Neglected Tropical Diseases* **4**, e673 (2010). URL http://dx.doi.org/10.1371/journal.pntd.0000673. Publisher: Division of Infectious Diseases, Department of Medicine, University of California Irvine, Irvine, California, USA.

[182] Lessa-Aquino, C. *et al.* Identification of Seroreactive Proteins of Leptospira interrogans Serovar Copenhageni Using a High-Density Protein Microarray Approach. *PLoS Neglected Tropical Diseases* **7**, e2499 (2013). URL http://dx.doi.org/10.1371/journal.pntd.

0002499. Publisher: Bio-Manguinhos, Oswaldo Cruz Foundation, Brazilian Ministry of Health, Rio de Janeiro, Brazil ; Department of Medicine, Division of Infectious Disease, University of California Irvine, Irvine, California, United States of America.

[183] Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics* **12**, 77 (2011). ISBN: 1471-2105 (Electronic) 1471-2105 (Linking).

[184] Ricci, A. D. *et al.* A *Trypanosoma cruzi* Antigen and Epitope Atlas: deep characterization of antibody specificities in Chagas Disease patients across the Americas. Preprint, Immunology (2022). URL http://biorxiv.org/lookup/doi/10.1101/2022.08.19.504544.

[185] Di Noia, J. M. & Neuberger, M. S. Molecular Mechanisms of Antibody Somatic Hypermutation. *Annual Review of Biochemistry* **76**, 1–22 (2007). URL https://www.annualreviews.org/doi/10.1146/annurev.biochem.76.061705.090740.

[186] Neuberger, M. S. Antibody diversification by somatic mutation: from Burnet onwards. *Immunology & Cell Biology* **86**, 124–132 (2008). URL https://onlinelibrary.wiley.com/doi/abs/10.1038/sj.icb.7100160.

[187] Frank, R. & Overwln, H. SPOT Synthesis: Epitope Analysis with Arrays of Synthetic Peptides Prepared on Cellulose Membranes. In *Epitope Mapping Protocols*, vol. 66, 149–170 (Humana Press, New Jersey, 1996). URL http://link.springer.com/10.1385/0-89603-375-9:149.

[188] Geysen, H., Rodda, S. J., Mason, T. J., Tribbick, G. & Schoofs, P. G. Strategies for epitope analysis using peptide synthesis. *Journal of Immunological Methods* **102**, 259–274 (1987). URL https://linkinghub.elsevier.com/retrieve/pii/0022175987900858.

[189] Reineke, U. & Sabat, R. Antibody Epitope Mapping Using SPOT™ Peptide Arrays. In Schutkowski, M. & Reineke, U. (eds.) *Epitope Mapping Protocols*, vol. 524, 145–167 (Humana Press, Totowa, NJ, 2009). URL http://link.springer.com/10.1007/978-1-59745-450-6_11. Series Title: Methods in Molecular Biology™.

[190] Vengesai, A. *et al.* Scoping review of the applications of peptide microarrays on the fight against human infections. *PLOS ONE* **17**, e0248666 (2022). URL https://dx.plos.org/10.1371/journal.pone.0248666.

[191] Hansen, L. B., Buus, S. & Schafer-Nielsen, C. Identification and Mapping of Linear Antibody Epitopes in Human Serum Albumin Using High-Density Peptide Arrays. *PLoS One* **8**, e68902 (2013). URL https://dx.plos.org/10.1371/journal.pone.0068902. Publisher: Laboratory of Experimental Immunology, University of Copenhagen, Copenhagen, Denmark.

[192] Osterbye, T. *et al.* HLA Class II Specificity Assessed by High-Density Peptide Microarray Interactions. *The Journal of Immunology* **205**, 290–299 (2020). URL http://www.jimmunol.org/lookup/doi/10.4049/jimmunol.2000224.

[193] Malovichko, G. & Zhu, X. Single Amino Acid Substitution in the Vicinity of a Receptor-Binding Domain Changes Protein–Peptide Binding Affinity. *ACS Omega* **2**, 5445–5452 (2017). URL https://pubs.acs.org/doi/10.1021/acsomega.7b00963.

[194] Yan, Y. *et al.* Whole Genome–Derived Tiled Peptide Arrays Detect Prediagnostic Autoantibody Signatures in Non–Small-Cell Lung Cancer. *Cancer Research* **79**, 1549–1557 (2019). URL https://aacrjournals.org/cancerres/article/79/7/1549/641540/Whole-Genome-Derived-Tiled-Peptide-Arrays-Detect.

[195] Balouz, V., Agüero, F. & Buscaglia, C. Chagas Disease Diagnostic Applications. In *Advances in Parasitology*, vol. 97, 1–45 (Elsevier, 2017). URL https://linkinghub.elsevier.com/retrieve/pii/S0065308X16301038.

[196] Guzmán-Gómez, D. *et al.* Highly discordant serology against Trypanosoma cruzi in central Veracruz, Mexico: role of the antigen used for diagnostic. *Parasites & Vectors* **8**, 466 (2015). URL https://parasitesandvectors.biomedcentral.com/articles/10.1186/s13071-015-1072-2.

[197] Moure, Z. *et al.* Serodiscordance in chronic Chagas disease diagnosis: a real problem in non-endemic countries. *Clinical Microbiology and Infection* **22**, 788–792 (2016). URL https://linkinghub.elsevier.com/retrieve/pii/S1198743X16301860.

[198] Granjon, E. *et al.* Development of a Novel Multiplex Immunoassay Multi-cruzi for the Serological Confirmation of Chagas Disease. *PLOS Neglected Tropical Diseases* **10**, e0004596 (2016). URL https://dx.plos.org/10.1371/journal.pntd.0004596.

[199] Jurado Medina, L. *et al.* Prediction of parasitological cure in children infected with Trypanosoma cruzi using a novel multiplex serological approach: an observational, retrospective cohort study. *The Lancet Infectious Diseases* **21**, 1141–1150 (2021). URL https://linkinghub.elsevier.com/retrieve/pii/S1473309920307295.

[200] World Health Organization. Research priorities for Chagas disease, human African trypanosomiasis and leishmaniasis. *World Health Organization Technical Report Series* v–xii, 1–100 (2012).

[201] Peeling, R. W. Diagnostics in a digital age: an opportunity to strengthen health systems and improve health outcomes. *International Health* **7**, 384–389 (2015). URL https://academic.oup.com/inthealth/article-lookup/doi/10.1093/inthealth/ihv062.

[202] Peeling, R. W., Boeras, D. I. & Nkengasong, J. Re-imagining the future of diagnosis of Neglected Tropical Diseases. *Computational and Structural Biotechnology Journal* **15**, 271–274 (2017). URL https://linkinghub.elsevier.com/retrieve/pii/S2001037016301015.

[203] Brenière, S. F., Waleckx, E. & Barnabé, C. Over Six Thousand Trypanosoma cruzi Strains Classified into Discrete Typing Units (DTUs): Attempt at an Inventory. *PLOS Neglected Tropical Diseases* **10**, e0004792 (2016). URL https://dx.plos.org/10.1371/journal.pntd.0004792.

[204] Souza, R. T. *et al.* Genome Size, Karyotype Polymorphism and Chromosomal Evolution in Trypanosoma cruzi. *PLoS ONE* **6**, e23042 (2011). URL https://dx.plos.org/10.1371/journal.pone.0023042.

[205] Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: Visualization of Intersecting Sets. *IEEE Transactions on Visualization and Computer Graphics* **20**, 1983–1992 (2014). URL http://ieeexplore.ieee.org/document/6876017/.

[206] Ibañez, C. F. *et al.* Multiple Trypanosoma cruzi antigens containing tandemly repeated amino acid sequence motifs. *Molecular and Biochemical Parasitology* **30**, 27–33 (1988). URL https://linkinghub.elsevier.com/retrieve/pii/0166685188901296. Publisher: Instituto de Investigaciones Bioquimicas Fundacion Campomar, Buenos Aires, Argentina.

[207] Nothelfer, K., Sansonetti, P. J. & Phalipon, A. Pathogen manipulation of B cells: the best defence is a good offence. *Nature Reviews Microbiology* **13**, 173–184 (2015). URL http://www.nature.com/articles/nrmicro3415.

[208] Bermejo, D. A. *et al.* Trypanosoma cruzi infection induces a massive extrafollicular and follicular splenic B-cell response which is a high source of non-parasite-specific antibodies: B-cell response to T. cruzi. *Immunology* **132**, 123–133 (2011). URL https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2567.2010.03347.x.

[209] Minoprio, P. Parasite polyclonal activators: new targets for vaccination approaches? *International Journal for Parasitology* **31**, 588–591 (2001). URL https://linkinghub.elsevier.com/retrieve/pii/S0020751901001710.

[210] Abras, A. *et al.* Serological Diagnosis of Chronic Chagas Disease: Is It Time for a Change? *Journal of Clinical Microbiology* **54**, 1566–1572 (2016). URL https://journals.asm.org/doi/10.1128/JCM.00142-16.

[211] Lewis, M. D. & Kelly, J. M. Putting Infection Dynamics at the Heart of Chagas Disease. *Trends in Parasitology* **32**, 899–911 (2016). URL https://linkinghub.elsevier.com/retrieve/pii/S147149221630126X.

[212] Pérez-Mazliah, D., Ward, A. I. & Lewis, M. D. Host-parasite dynamics in Chagas disease from systemic to hyper-local scales. *Parasite Immunology* **43** (2021). URL https://onlinelibrary.wiley.com/doi/10.1111/pim.12786.

[213] Ward, A. I. *et al.* *In Vivo* Analysis of Trypanosoma cruzi Persistence Foci at Single-Cell Resolution. *mBio* **11**, e01242–20 (2020). URL https://journals.asm.org/doi/10.1128/mBio.01242-20.

[214] Zhang, L. & Tarleton, R. Parasite Persistence Correlates with Disease Severity and Localization in Chronic Chagas' Disease. *The Journal of Infectious Diseases* **180**, 480–486 (1999). URL https://academic.oup.com/jid/article-lookup/doi/10.1086/314889.

[215] Warrenfeltz, S. *et al.* EuPathDB: The Eukaryotic Pathogen Genomics Database Resource. In Kollmar, M. (ed.) *Eukaryotic Genomic Databases*, vol. 1757, 69–113 (Springer New York, New York, NY, 2018). URL http://link.springer.com/10.1007/978-1-4939-7737-6_5. Series Title: Methods in Molecular Biology.

[216] Baptista, R. P. *et al.* Assembly of highly repetitive genomes using short reads: the genome of discrete typing unit III Trypanosoma cruzi strain 231. *Microbial Genomics* **4** (2018). URL https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000156.

[217] Umezawa, E. S. *et al.* Immunoblot assay using excreted-secreted antigens of Trypanosoma cruzi in serodiagnosis of congenital, acute, and chronic Chagas' disease. *Journal of Clinical Microbiology* **34**, 2143–2147 (1996). URL https://journals.asm.org/doi/10.1128/jcm.34.9.2143-2147.1996.

[218] Plotly Technologies Inc. Collaborative data science (2015). URL https://plot.ly. Place: Montreal, QC Publisher: Plotly Technologies Inc.

[219] Hopp, T. P. & Woods, K. R. Prediction of protein antigenic determinants from amino acid sequences. *Proceedings of the National Academy of Sciences* **78**, 3824–3828 (1981). URL https://pnas.org/doi/full/10.1073/pnas.78.6.3824.

[220] Pellequer, J., Westhof, E. & Van Regenmortel, M. Predicting location of continuous epitopes in proteins from their primary structures. In *Methods in Enzymology*, vol. 203, 176–201 (Elsevier, 1991). URL https://linkinghub.elsevier.com/retrieve/pii/007668799103010E.

[221] Parker, J. M. R., Guo, D. & Hodges, R. S. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and x-ray-derived accessible sites. *Biochemistry* **25**, 5425–5432 (1986). URL https://pubs.acs.org/doi/abs/10.1021/bi00367a013.

[222] Chou, P. Y. & Fasman, G. D. Prediction of the Secondary Structure of Proteins from their Amino Acid Sequence. In Meister, A. (ed.) *Advances in Enzymology - and Related Areas of Molecular Biology*, 45–148 (John Wiley & Sons, Inc., Hoboken, NJ, USA, 1979). URL https://onlinelibrary.wiley.com/doi/10.1002/9780470122921.ch2.

[223] Levitt, M. Conformational preferences of amino acids in globular proteins. *Biochemistry* **17**, 4277–4285 (1978). URL https://pubs.acs.org/doi/abs/10.1021/bi00613a026.

[224] Emini, E. A., Hughes, J. V., Perlow, D. S. & Boger, J. Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. *Journal of Virology* **55**, 836–839 (1985). URL https://journals.asm.org/doi/10.1128/jvi.55.3.836-839.1985.

[225] Alix, A. J. Predictive estimation of protein linear epitopes by using the program PEOPLE. *Vaccine* **18**, 311–314 (1999). Publisher: Laboratoire de Spectroscopies et Structures Biomoléculaires, Université de Reims Champagne Ardenne, Faculté des Sciences, BP 1039, 51687 Reims Cedex 2, France. alain.alix@univ-reims.fr.

[226] Odorico, M. & Pellequer, J.-L. BEPITOPE: predicting the location of continuous epitopes and patterns in proteins. *Journal of Molecular Recognition* **16**, 20–22 (2003). URL https://onlinelibrary.wiley.com/doi/10.1002/jmr.602.

[227] Blythe, M. J. & Flower, D. R. Benchmarking B cell epitope prediction: Underperformance of existing methods. *Protein Science* **14**, 246–248 (2004). URL http://doi.wiley.com/10.1110/ps.041059505.

[228] Goto, Y., Carter, D. & Reed, S. G. Immunological dominance of Trypanosoma cruzi tandem repeat proteins. *Infect Immun* **76**, 3967–3974 (2008). URL http://dx.doi.org/10.1128/IAI.00604-08. Publisher: Infectious Disease Research Institute, Seattle, Washington 98104, USA. ygoto@idri.org.

[229] List, C. *et al.* Serodiagnosis of Echinococcus spp. Infection: Explorative Selection of Diagnostic Antigens by Peptide Microarray. *PLoS Neglected Tropical Diseases* **4**, e771 (2010). URL http://www.ncbi.nlm.nih.gov/pubmed/20689813.

[230] Cooley, G. *et al.* High throughput selection of effective serodiagnostics for Trypanosoma cruzi infection. *PLoS Negl Trop Dis* **2**, e316 (2008). URL http://dx.doi.org/10.1371/journal.pntd.0000316. Publisher: Center for Tropical and Emerging Global Diseases and Department of Cellular Biology, University of Georgia, Athens, Georgia, United States of America.

[231] Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021). URL https://www.nature.com/articles/s41586-021-03819-2. Number: 7873 Publisher: Nature Publishing Group.

[232] Høie, M. H. *et al.* DiscoTope-3.0 - Improved B-cell epitope prediction using AlphaFold2 modeling and inverse folding latent representations (2023). URL https://www.biorxiv.org/content/10.1101/2023.02.05.527174v1. Pages: 2023.02.05.527174 Section: New Results.

[233] Teixeira, A. R. & Santos-Buch, C. A. The immunology of experimental Chagas' disease. I. Preparation of Trypanosoma cruzi antigens and humoral antibody response to there antigens. *Journal of immunology (Baltimore, Md. : 1950)* **113**, 859–869 (1974). Place: United States.

[234] Segura, E. L., Campos, J. M., de Ducatenzeiler, A. B. & Cerisola, J. A. Antigens of the subcellular fractions of Trypanosoma cruzi. I. Localization of antigens and proteins in the subcellular fractions. *Medicina* **35**, 451–459 (1975). Place: Argentina.

[235] Segura, E. L. *et al.* Antigens of the Subcellular Fractions of *Trypanosoma cruzi* . II. Flagellar and Membrane Fraction. *The Journal of Protozoology* **24**, 540–543 (1977). URL https://onlinelibrary.wiley.com/doi/10.1111/j.1550-7408.1977.tb01009.x.

[236] Cappa, S. M. G. *et al.* Antigens of Subcellular Fractions of *Trypanosoma cruzi*. III. Humoral immune response and histopathology of immunized mice*. *The Journal of Protozoology* **27**, 467–471 (1980). URL https://onlinelibrary.wiley.com/doi/10.1111/j.1550-7408.1980.tb05399.x.

[237] Noazin, S. *et al.* Trypomastigote Excretory Secretory Antigen Blot Is Associated With Trypanosoma cruzi Load and Detects Congenital T. cruzi Infection in Neonates, Using Anti-Shed Acute Phase Antigen Immunoglobulin M. *The Journal of Infectious Diseases* **219**, 609–618 (2019).

[238] CHAGAS VIRCLIA® - Indirect chemiluminescent immunoassay to test antibodies against Trypanosoma cruzi in human serum/plasma. URL https://www.vircell.com/producto/chagas-virclia/.

[239] Kephera Diagnostics LLC. Point-of-care diagnostic test for T. cruzi (Chagas) infection (2018). URL https://www.sbir.gov/sbirsearch/detail/1565393.

[240] Ibañez, C. F., Affranchino, J. L. & Frasch, A. C. Antigenic determinants of Trypanosoma cruzi defined by cloning of parasite DNA. *Molecular and Biochemical Parasitology* **25**, 175–184 (1987). URL https://linkinghub.elsevier.com/retrieve/pii/0166685187900065.

[241] Kaplan, G. & Gershoni, J. M. A general insert label for peptide display on chimeric filamentous bacteriophages. *Analytical Biochemistry* **420**, 68–72 (2012). URL https://linkinghub.elsevier.com/retrieve/pii/S0003269711005951.

[242] Alvarez, P., Leguizamón, M. S., Buscaglia, C. A., Pitcovsky, T. A. & Campetella, O. Multiple Overlapping Epitopes in the Repetitive Unit of the Shed Acute-Phase Antigen from *Trypanosoma cruzi* Enhance Its Immunogenic Properties. *Infection and Immunity* **69**, 7946–7949 (2001). URL https://journals.asm.org/doi/10.1128/IAI.69.12.7946-7949.2001.

[243] Pitcovsky, T. A. *et al.* Epitope Mapping of *trans* -Sialidase from *Trypanosoma cruzi* Reveals the Presence of Several Cross-Reactive Determinants. *Infection and Immunity* **69**, 1869–1875 (2001). URL https://journals.asm.org/doi/10.1128/IAI.69.3.1869-1875.2001.

[244] Teixeira, A. A. R. *et al.* A refined genome phage display methodology delineates the human antibody response in patients with Chagas disease. *iScience* **24**, 102540 (2021). URL https://linkinghub.elsevier.com/retrieve/pii/S2589004221005083.

[245] Dent, A. E. *et al.* Plasmodium falciparum Protein Microarray Antibody Profiles Correlate With Protection From Symptomatic Malaria in Kenya. *The Journal of Infectious Diseases* **212**, 1429–1438 (2015).

[246] Uplekar, S. *et al.* Characterizing Antibody Responses to Plasmodium vivax and Plasmodium falciparum Antigens in India Using Genome-Scale Protein Microarrays. *PLoS neglected tropical diseases* **11**, e0005323 (2017).

[247] Travassos, M. A. *et al.* Children with cerebral malaria or severe malarial anaemia lack immunity to distinct variant surface antigen subsets. *Scientific Reports* **8**, 6281 (2018).

[248] Reis-Cunha, J. L. *et al.* Genome-Wide Screening and Identification of New Trypanosoma cruzi Antigens with Potential Application for Chronic Chagas Disease Diagnosis. *PLoS ONE* **9**, e106304 (2014). URL https://dx.plos.org/10.1371/journal.pone.0106304.

[249] Ramos-López, P., Irizarry, J., Pino, I. & Blackshaw, S. Antibody Specificity Profiling Using Protein Microarrays. In Rockberg, J. & Nilvebrant, J. (eds.) *Epitope Mapping Protocols*, vol. 1785, 223–229 (Springer New York, New York, NY, 2018). URL http://link.springer.com/10.1007/978-1-4939-7841-0_14. Series Title: Methods in Molecular Biology.

[250] Flower, D. R. Towards in silico prediction of immunogenic epitopes. *Trends in Immunology* **24**, 667–674 (2003). URL https://linkinghub.elsevier.com/retrieve/pii/S1471490603003247.

[251] Jurtz, V. *et al.* NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *The Journal of Immunology* **199**, 3360–3368 (2017). URL http://www.jimmunol.org/lookup/doi/10.4049/jimmunol.1700893.

[252] Zrein, M. *et al.* A novel antibody surrogate biomarker to monitor parasite persistence in Trypanosoma cruzi-infected patients. *PLOS Neglected Tropical Diseases* **12**, e0006226 (2018). URL https://dx.plos.org/10.1371/journal.pntd.0006226.

[253] Altcheh, J. M. *Chagas disease: a clinical approach* (Springer Berlin Heidelberg, New York, NY, 2019), 1st edn.

[254] Murphy, N. *et al.* Assessing antibody decline after chemotherapy of early chronic Chagas disease patients. *Parasites & Vectors* **14**, 543 (2021). URL https://parasitesandvectors.biomedcentral.com/articles/10.1186/s13071-021-05040-6.

[255] Rodríguez-Angulo, H. O. *et al.* Autoantibodies against the immunodominant sCha epitope discriminate the risk of sudden death in chronic Chagas cardiomyopathy. *Annals of the New York Academy of Sciences* **1497**, 27–38 (2021). URL https://onlinelibrary.wiley.com/doi/10.1111/nyas.14586.

[256] Nunes, M. C. P. *et al.* Incidence and Predictors of Progression to Chagas Cardiomyopathy: Long-Term Follow-Up of *Trypanosoma cruzi* –Seropositive Individuals. *Circulation* **144**, 1553–1566 (2021). URL https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.121.055112.

[257] Cao, J. *et al.* The preparation and clinical application of diagnostic DNA microarray for the detection of pathogens in intracranial bacterial and fungal infections. *Experimental and Therapeutic Medicine* (2018). URL http://www.spandidos-publications.com/10.3892/etm.2018.6312.

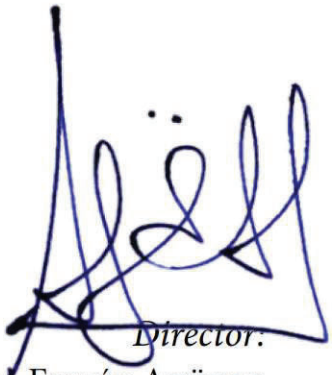
[258] Neagu, M., Bostan, M. & Constantin, C. Protein microarray technology: Assisting personalized medicine in oncology (Review). *World Academy of Sciences Journal* (2019). URL http://www.spandidos-publications.com/10.3892/wasj.2019.15.

[259] Brambilla, D., Chiari, M., Gori, A. & Cretich, M. Towards precision medicine: the role and potential of protein and peptide microarrays. *The Analyst* **144**, 5353–5367 (2019). URL http://xlink.rsc.org/?DOI=C9AN01142K.

Firmas

Dejo constancia que esta versión de la Tésis corresponde a la última versión, incluyendo las correcciones de los jurados.

Autor:
Lic. Alejandro Daniel Ricci



Director:
Dr. Fernán Agüero