UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS BIOLÓGICAS E DA SAÚDE
PROGRAMA DE PÓS-GRADUAÇÃO EM
GENÉTICA EVOLUTIVA E BIOLOGIA MOLECULAR

JHONNE PEDRO PEDOTT SANTANA

LOCALIZAÇÃO DE REGIÕES POTENCIAIS PARA INTEGRAÇÃO DO kDNA DE *Trypanosoma cruzi* NO GENOMA HUMANO

Dissertação apresentada ao Programa de Pós-graduação em Genética Evolutiva e Biologia Molecular da Universidade Federal de São Carlos, como requisito parcial à obtenção do título de Mestre.

SÃO CARLOS
2016

JHONNE PEDRO PEDOTT SANTANA

LOCALIZAÇÃO DE REGIÕES POTENCIAIS PARA INTEGRAÇÃO DO kDNA DE *Trypanosoma cruzi* NO GENOMA HUMANO

Dissertação apresentada ao Programa de Pós-graduação em Genética Evolutiva e Biologia Molecular da Universidade Federal de São Carlos, como requisito parcial à obtenção do título de Mestre.

Orientador: Prof. Dr. Eduardo Leonardecz
Coorientador: Prof. Dr. Ricardo Carneiro Borra

SÃO CARLOS
2016

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária UFSCar
Processamento Técnico
com os dados fornecidos pelo(a) autor(a)

Santana, Jhonne Pedro Pedott
S232l Localização de regiões potenciais para integração do kDNA de Trypanosoma cruzi no genoma humano / Jhonne Pedro Pedott Santana. -- São Carlos : UFSCar, 2016.
72 p.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2016.

1. Transferência horizontal de genes. 2. Análise bioinformática. 3. Genoma humano. 4. Doença de Chagas. 5. Trypanosoma cruzi. I. Título.

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Biológicas e da Saúde
Programa de Pós-Graduação em Genética Evolutiva e Biologia Molecular

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado do candidato Jhonne Pedro Pedott Santana, realizada em 23/03/2016:

Prof. Dr. Eduardo Leonardecz Neto
UFSCar

Prof. Dr. Ricardo Carneiro Borra
UFSCar

Profa. Dra. Bianca Baccili Zanotto
EMBRAPA

"Todas as grandezas desse mundo não valem um bom amigo."

Voltaire

AGRADECIMENTOS

Agradeço de antemão a todos que de alguma maneira passaram pela minha vida e contribuíram para a construção de quem sou hoje.

À minha noiva, amor da minha vida, Naiane, pelo companheirismo, paciência, carinho e apoio emocional prestado durante todos esses 6 anos de relacionamento.

Aos meus pais, João e Luci, pelo amor incondicional, ensinamentos e apoio para que eu pudesse estudar e chegar cada vez mais longe.

Aos meus irmãos, Cacio, Laura e Mariana, por todos os momentos de carinho e lições de vida.

Ao meu irmão Rubian, pelos momentos de alegria durante toda minha vida.

À minha segunda família, Antonia, Charles e Rafaela, pelo amor e incentivo a continuar buscando meus sonhos.

Aos queridos Joelson e Andrieli, por todos os momentos de descontração e palavras de ânimo e confiança.

À minha amiga Naiana, pela amizade construída, apoio e suporte durante todo o mestrado.

Aos amigos que considero extremamente especiais: Camila, João Pedro e Tatiele. Vocês tornam meus dias melhores, possibilitando momentos únicos e inesquecíveis de felicidade e sincera amizade.

Aos meus amigos de laboratório, Leonardo e Pablo, sem os quais nada disso seria possível.

Em especial, ao meu orientador, prof. Dr. Eduardo Leonardecz, por toda a paciência, confiança, ensinamentos e orientação prestada na realização deste trabalho.

Ao meu coorientador, prof. Dr. Ricardo Borra, pela ajuda e colaboração para a conclusão deste projeto.

Aos professores Dr. Francis Nunes e Dra. Polyana Tizioto, por terem aceitado fazer parte de minha banca de qualificação e a Dra. Bianca Baccili Zanotto Vigna pela disponibilidade em participar da banca de defesa, por todos os conselhos e ensinamentos.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

BLAST	*Basic Local Alignment Search Tool*
BLASTN	*Nucleotide Basic Local Alignment Search Tool*
BLAT	*BLAST-like Alignment Tool*
Motivos CA's	Assinaturas de Adenina e Citosina
CLUSTAL W	*Cluster Analysis Tool*
CRAN	*Comprehensive R Archive Network*
DCA	Doença de Chagas Aguda
DCC	Doença de Chagas Crônica
DN	Doenças Negligenciadas
DNA	Ácido Desoxirribonucleico
EUA	Estados Unidos da América
GOLD	*Genomes OnLine Database*
GRCh38	*Genome Reference Consortium Human Build 38*
gRNA	RNA guia
HGP	Projeto Genoma Humano
HGT	Transferência Horizontal de Genes
IUPAC	União Internacional de Química Pura e Aplicada
kDNA	DNA do Cinetoplasto
LINE	Longos Elementos Nucleares Intercalados

mRNA	RNA Mensageiro
NCBI	Centro Nacional de Informação Biotecnológica dos EUA
pb	Pares de Bases
RNA	Ácido Ribonucleico
SINE	Curtos Elementos Nucleares Intercalados
tRNA	RNA Transportador
UCSC	*University of California Santa Cruz*
UF	Unidade Federativa

SUMÁRIO

RESUMO

LOCALIZAÇÃO DE REGIÕES POTENCIAIS PARA INTEGRAÇÃO DO kDNA DE *Trypanosoma cruzi* NO GENOMA HUMANO

Com o sequenciamento do genoma humano e tantas outras espécies, abre-se agora uma nova janela de oportunidades analíticas. Podemos pensar em fazer buscas orientadas dentro dessa massa enorme de dados publicados em bancos de dados biológicos. Tendo isso em foco, buscamos estruturar uma forma automatizada de busca dentro do genoma humano, pela qual pudéssemos inferir sobre os sítios mais prováveis de integração de DNA exógeno. Para isso utilizamos como modelo os trabalhos que indicam que a doença de Chagas é produzida pela introgressão do kDNA de *Trypanosoma cruzi* no genoma hospedeiro, por meio de herança genética horizontal. Já foi demonstrado experimentalmente que micro-homologias ricas em adenina e citosina medeiam as integrações de minicírculos de kDNA do *T. cruzi*, no genoma de vertebrados. Deste modo, o presente trabalho propõe uma maneira eficiente, fácil e rápida para a busca e localização de múltiplas assinaturas dos sinalizadores que propiciam a introgressão do kDNA exógeno no genoma humano, através da criação de um conjunto de scripts para análises *in silico*, adaptados a grandes arquivos de sequências. Foram desenvolvidos três scripts, baseados na linguagem R: para permutação de elementos (ácidos nucleicos ou aminoácidos); para busca, agrupamento e plotagem das correspondências em genoma; e para contagem total de correspondências e contagem por janela cromossômica. Todas as assinaturas compostas por adenina e citosina (motivos CA's) foram devidamente identificadas no genoma humano, porém não foi identificado nenhum ponto mais suscetível à integração do kDNA de *T. cruzi*. Com os dados obtidos, um mapa genético foi criado, listando as correspondências em cada banda citogenética, porém não foi possível identificar qual cromossomo possui maior propensão à mutações, já que quanto maior o cromossomo, maior é a quantidade de correspondências presentes.

Descritores: Transferência Horizontal de Genes, Análise Bioinformática, Genoma Humano, Doença de Chagas, *Trypanosoma cruzi*, kDNA.

ABSTRACT

LOCALIZATION OF POTENTIAL REGIONS FOR INTEGRATION OF *Trypanosoma cruzi* KDNA IN THE HUMAN GENOME

Knowledge about horizontal gene transfer has been proposed even before the determination of the molecular structure of DNA. It has been experimentally shown that micro-homologies rich in adenine and cytosine mediates the integration of *Trypanosoma cruzi's* kDNA minicircle, in the vertebrate genome. After human genome sequencing, the genome characterization of different organisms has been one of the main driving forces of science, providing a quantity of biological data for modern biomedical research, unprecedented in the history of science. However, even though traditional DNA mapping algorithms are highly accurate, they operate at a much lower rate than that needed for the next generation sequencers to accumulate new data. This great asymmetry between data generation and analysis capability requires the rapid evolution of mapping and reading algorithms so that this large volume of information can be worked through targeted searches. Thus, this work proposes an efficient, fast and easy way to search and locate multiple signatures of indicators that allow exogenous kDNA integration in the human genome, by creating a set of scripts for *in silico* analysis adapted to large files sequences. Three scripts based in R language were developed: to permute the elements (nucleic acids or amino acids codes); for search, grouping and plotting matches in genome; and for counting total matches and chromosomal window. All adenine and cytosine signatures were properly identified in the human genome, but no point more susceptible to *T. cruzi* kDNA integration was identified. With the obtained data, a genetic map was created, listing all matchings in each cytogenetic band, but it was not possible to identify which chromosome was more prone to mutations, since the bigger the chromosome is, the higher the quantity of matches are.

Key words: Horizontal Gene Transfer, Bioinformatics Analysis, Human Genome, Chagas Disease, *Trypanosoma cruzi*, kDNA.

1. INTRODUÇÃO

1.1. FERRAMENTAS BIOINFORMÁTICAS: BREVE HISTÓRICO

Nos primeiros projetos de sequenciamento genômico, como o trabalho que sequenciou a primeira molécula de ácido ribonucleico (RNA), utilizando RNA transportador (tRNA) de alanina em *Escherichia coli* (HOLLEY *et al.*, 1965), e o primeiro sequenciamento de ácido desoxirribonucleico (DNA) moderno, através do método Sanger baseado em didesoxirribonucleotídeos (SANGER *et al.*, 1977), foram necessários meses de processamento, mesmo fazendo uso de modelos reduzidos e algoritmos bem mais simples que os atuais.

Descrito em 1977, o método Sanger permaneceu como único método de sequenciamento de ácidos nucléicos utilizado durante décadas e serviu de base para a era genômica na biologia, permitindo diversos projetos de sequenciamento de bactérias (HUTCHISON, 2007). Mesmo com todo o sucesso obtido, a eficiência, custo e tempo de processamento impediam o avanço das pesquisas nesta área, o que abriu portas para o desenvolvimento de outras técnicas de sequenciamento de DNA. No início dos anos 2000, com o aperfeiçoamento e automatização das tecnologias dos equipamentos de sequenciamento foi possível gerar de duas a três vezes mais dados do que os obtidos pela tecnologia Sanger, por meio de máquinas que geram sequências 24h por dia e a um custo inferior (Figura 1) (MARDIS, 2008).

Desde a conclusão do primeiro sequenciamento de genoma de um organismo de vida livre, o *Haemophilus influenzae*, em 1995 (FLEISCHMANN *et al.*, 1995), a caracterização dos genomas de diferentes organismos foi uma das principais forças motrizes da ciência Genômica.

O sequenciamento do genoma humano, formalmente iniciado em Outubro de 1990 pelo Projeto Genoma Humano (HGP) e anunciado em 2001 pelo International Human Genome Sequencing Consortium (LANDER *et al.*, 2001), abriu portas para uma era de abrangentes descrições sobre a saúde humana, provendo uma quantidade de dados biológicos para a investigação biomédica moderna sem precedentes na história da ciência.

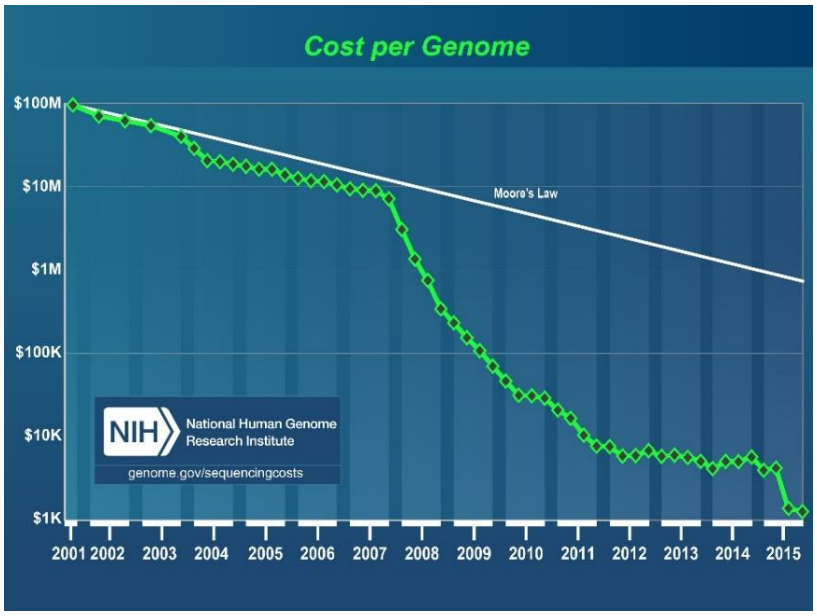


Figura 1: *Evolução dos custos para o sequenciamento de um genoma* (WETTERSTRAND, 2016). Os dados evidenciam a redução dos preços (2001 a 2015) para sequenciar um genoma do tamanho do genoma humano, além de comparar com dados hipotéticos da Lei de Moore (duplicação do poder computacional a cada dois anos na indústria tecnológica). A súbita queda no custo a partir de janeiro de 2008 representa a transição para tecnologias de sequenciamento de "segunda geração" (*next-generation*).

Com a conclusão do HGP e, consequentemente, com o advento do estudo da genômica (VENTER *et al.*, 2001), a disponibilidade de listas de genomas completamente sequenciados cresceu rapidamente (Figura 2), transformando a biologia em diversos sentidos, no qual destaca-se a emergência da bioinformática.

O domínio da bioinformática nos permitiu organizar, manipular, gerenciar, integrar e analisar dados complexos e variados *in silico*, em um universo crescente de informações, proporcionando grandes avanços na compreensão dos detalhes moleculares da biologia tradicional (LESK, 2014).

Desta maneira, com os crescentes investimentos na genômica e a integração de diversas áreas do conhecimento, foi inevitável o surgimento de novas áreas, tais como, a proteômica, responsável por analisar e identificar as proteínas expressas (ZHOU *et al.*, 2002), e a transcriptômica, que estuda a transcrição do DNA para o RNA (STUART *et al.*, 2003).

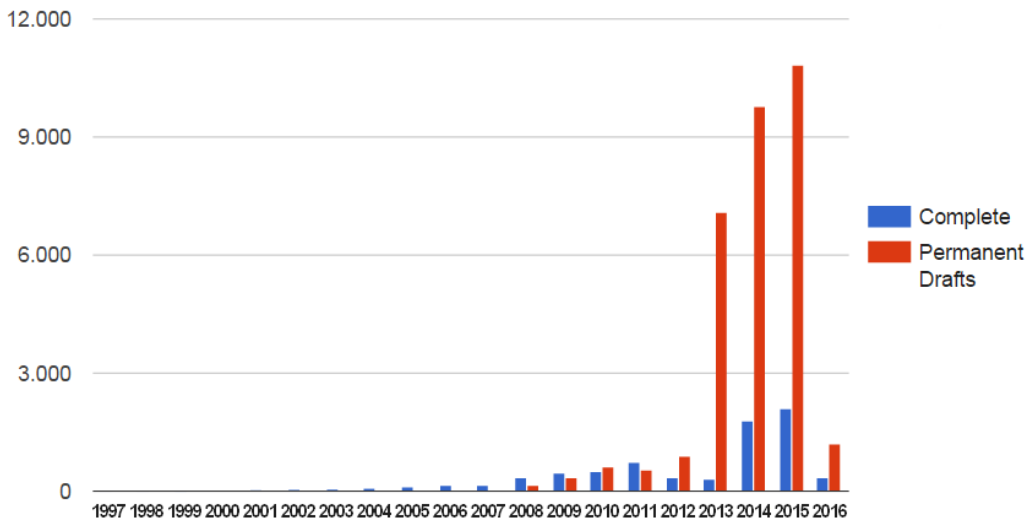


Figura 2: *Projetos de genoma completos e disponíveis no GOLD* (GENOMES ONLINE DATABASE, 2016). Os dados não cumulativos, desde 1997, demonstram o aumento significativo do número de projetos nos últimos anos.

Além das citadas anteriormente, uma variedade de subdisciplinas das chamadas “Ciências Ômicas” tem surgido, cada uma com seu próprio conjunto de instrumentos, técnicas, softwares e base de dados, com o objetivo de caracterizar o maior número possível de biomoléculas de um mesmo grupo, como DNA, RNA, proteínas ou metabólitos (ESPINDOLA *et al.*, 2010). Os dados gerados por todas as áreas das “Ômicas”, através de processos de sequenciamento em larga escala, podem ser exploradas com o uso de poderosas ferramentas de busca, interligadas a fontes de informação digital disponíveis para acesso público por meio de banco de dados.

Com crescimento acentuado nos últimos anos, os bancos de dados de Biologia Molecular oferecem informações detalhadas de moléculas ou funções específicas. O Centro Nacional de Informação Biotecnológica dos Estados Unidos (NCBI) é considerado o banco de dados central sobre informações genômicas. Existem outros bancos de dados similares em países europeus e no Japão, porém todos trocam dados a cada 24 horas com o NCBI, de forma a permitir uma maior integração entre os mesmos. Podemos destacar o *GenBank* como o principal banco de dados do NCBI, que disponibiliza publicamente cerca de 185 milhões de sequências de nucleotídeos para mais de 340 mil espécies formalmente descritas (CLARK *et al.*, 2016).

A existência de bancos de dados “secundários” para organização de informações é tão necessária quanto a preservação dos dados originais no *GenBank*, como no caso do *UniGene*, responsável por reunir sequências parciais do transcriptoma de um organismo em *clusters* (PONTIUS *et al.*, 2003), ou do *RefSeq*, responsável por agrupar a mais representativa sequência de um transcrito, denominada sequência de referência (PRUITT *et al.*, 2012).

Desde 1988, o NCBI conduz a pesquisa e desenvolvimento sobre representação, integração e recuperação de dados de biologia molecular e da literatura biomédica (Figura 3) (NATIONAL LIBRARY OF MEDICINE, 2016).

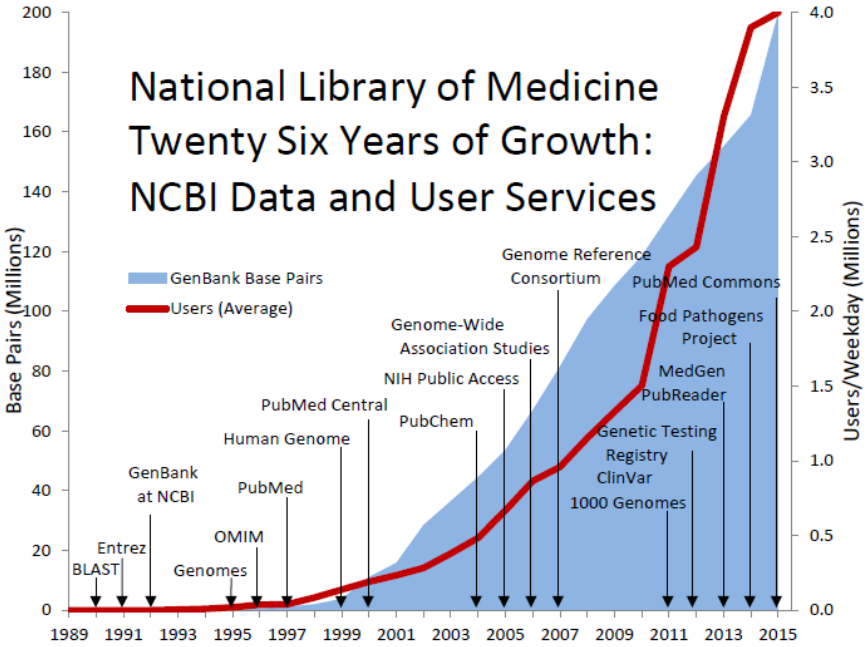


Figura 3: *Quantidade de dados e média de usuários no GenBank* (NATIONAL LIBRARY OF MEDICINE, 2016). O gráfico representa o crescimento da quantidade de informação disponível no *GenBank*, bem como a média de usuários ativos na plataforma, nos últimos 26 anos.

Atualmente, os genes têm sido descobertos aos milhares e, consequentemente, a quantidade de dados a serem analisados no futuro é cada vez maior. Utilizando programas de predição gênica, estima-se que existam aproximadamente 20,000 genes que codificam proteínas e mais de 18,000 genes de RNA distribuídos somente no genoma humano (PENNISI, 2012).

O Bioconductor é um ambiente de engenharia de software flexível, que permite desenvolver as ferramentas necessárias para análise destes dados. Fornece estruturas e métodos que permitem a análise genômica de dados de alto rendimento no contexto do ambiente de programação estatística oferecido pelo software R, suportando diversos tipos de dados de sequenciamento; associando recursos de anotação; e permitindo a rápida criação de fluxos de trabalho que combinam vários tipos de dados e ferramentas para a inferência estatística, análise de redes e visualização em todas as fases de um projeto de geração de dados para publicação. (HUBER *et al.*, 2015).

No entanto, mesmo que algoritmos tradicionais de mapeamento de DNA, como o BLAST (*Basic Local Alignment Search Tool*) (ALTSCHUL *et al.*, 1997), BLAT (*BLAST-like Alignment Tool*) (KENT, 2002) ou CLUSTAL W (*Cluster Analysis Tool*) (THOMPSON *et al.*, 1994), sejam altamente precisos, operam em um ritmo muito inferior que os sequenciadores de nova geração acumulam novos dados. Esta grande assimetria entre geração de dados e capacidade de análise força a rápida evolução dos algoritmos de mapeamento e leitura, já que interpretar os dados é essencial para consumar plenamente o valor biológico, clínico e científico deste imenso e contínuo fluxo de dados ômicos (NOCQ *et al.*, 2013; ZHU *et al.*, 2015).

Deste modo, é de suma importância que haja a integração das atividades de bancos de dados e algoritmos computacionais, bem como o desenvolvimento de novas ferramentas para análise de dados massivos. Buscas orientadas destes dados, que ajudem a elucidar desde características básicas até mecanismos complexos de infecção/patogenia ainda desconhecidos de doenças que acometem a população em geral, principalmente daquelas tidas como negligenciadas, são necessárias para que esse grande volume de informação seja trabalhado.

1.2. DOENÇA DE CHAGAS

1.2.1. ASPECTOS GERAIS E DISTRIBUIÇÃO

As doenças negligenciadas (DN) compõem um grupo de doenças infecciosas altamente prevalentes que afetam a qualidade de vida e geram impactos

socioeconômicos negativos para a população dos países subdesenvolvidos, despertando pouco atrativo financeiro por parte da grande indústria farmacêutica (KEALEY e SMITH, 2010).

Uma das DN mais conhecidas é a Doença de Chagas ou Tripanossomíase Americana, considerada a maior endemia infecciosa do Mundo Ocidental, que se estende por todo o continente americano, indo do sul dos Estados Unidos (EUA) até o sul da Argentina, regiões em que se encontram os triatomíneos, hospedeiros invertebrados e vetores do *Trypanosoma cruzi* (CAVALIER-SMITH, 2003), agente etiológico da doença (WORLD HEALTH ORGANIZATION, 2002; 2012).

Esta antropozoonose foi inicialmente descrita em 1909 pelo médico e pesquisador brasileiro Carlos Ribeiro Justiniano das Chagas, que além de relatar o agente etiológico (*T. cruzi*), identificou o vetor de transmissão, a interação parasito-hospedeiro, os elementos epidemiológicos e a progressão clínica da doença (CHAGAS, 1909).

Já foi considerada exclusiva às localidades rurais com condições habitacionais precárias e população de baixa renda, porém, em 2007, após mortes por transfusões ou enxertos de órgãos em alguns pacientes na Europa, pesquisas da OMS encontraram diversos casos da doença entre imigrantes latino-americanos. A OMS decidiu então lançar uma campanha em países não endêmicos, cujo objetivo foi avaliar a prevalência e possível transmissão da infecção por *T. cruzi* em países onde a transmissão vetorial nunca foi relatada. Os resultados obtidos com essa iniciativa convenceu muitos países, incluindo Austrália, Canadá, Japão, EUA e vários países europeus, a desenvolverem sistemas de vigilância para o controle da doença (WORLD HEALTH ORGANIZATION, 2015).

Estima-se que exista cerca de 50 mil novos casos de doença de Chagas por ano, com prevalência de, aproximadamente, 8 milhões de pessoas infectadas em todo o mundo (Figura 4) (WORLD HEALTH ORGANIZATION, 2015).

Hoje, um sistema de vigilância da doença de Chagas mundial focado principalmente na transmissão da infecção através de transfusão de sangue e transplante de órgãos, além da transmissão congênita, está estabelecido. Recentemente foram encontrados vetores capazes de transmitir a infecção, aumentando o risco da instalação da transmissão vetorial em alguns países,

especialmente na Ásia. Estima-se que existam de 68 a 123 mil casos em nove países europeus: Bélgica, França, Alemanha, Itália, Países Baixos, Portugal, Espanha, Suíça e Reino Unido (BASILE *et al.*, 2011; WORLD HEALTH ORGANIZATION, 2012).

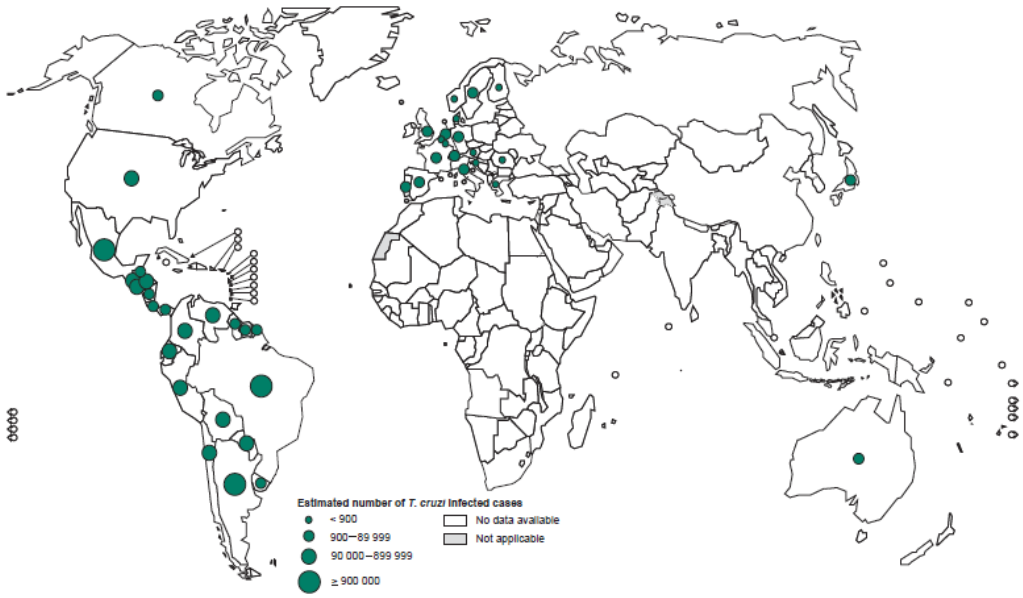


Figura 4: *Distribuição mundial dos casos de doença de Chagas* (WORLD HEALTH ORGANIZATION, 2015). Todos os casos reportados à OMS, de 2010 a 2013, evidenciando que a grande maioria dos casos ocorrem na América Latina.

A condição habitacional da população das áreas rurais, que residem em casas com paredes de barro (pau-a-pique) e casebres de palha, favorece a vida de várias espécies de triatomíneos, insetos hematófagos pertencentes à ordem *Hemíptera*, subordem *Heteroptera*, família *Reduviidae* e subfamília *Triatominae*. No Brasil são conhecidos popularmente como “barbeiro”, “chupão”, “fincão”, “furão”, “bicudo”, “bicho-de-parede”, dentre outros nomes, dependendo da região do país (LENT, 1999).

No território brasileiro, a endemia existe onde o *Triatoma infestans* Klug, 1834 é a principal espécie transmissora do agente etiológico para o homem, em especial na região Norte, com o estado do Pará liderando o número de casos (Tabela 1). Entretanto, outras espécies de barbeiro (*Panstrongylus megistus* Burmeister, 1835, *Triatoma sórdida* Stål, 1859*, Triatoma Brasiliensis* Neiva, 1911 *e*

Rhodnius sp. Stål, 1859) com hábitos silvestres ou domiciliados, também transmitem o *T. cruzi* para o homem (REY, 2008; WORLD HEALTH ORGANIZATION, 2012).

Tabela 1: *Casos confirmados de doença de Chagas aguda no Brasil* (MINISTÉRIO DA SAÚDE - BRASIL, 2015). A distribuição dos casos por Região, Unidade Federativa (UF) de residência e forma de transmissão, em território brasileiro, de 2000 a 2013, nos permite inferir que a condição habitacional e presença do vetor está intimamente relacionada com a quantidade de casos.

Região/Unidade da Federação	Forma de transmissão					Total	%
	Oral	Vetorial	Vertical	Ignorada*	Outras**		
Norte	1.023	70	1	329	7	1.430	91,1
Rondônia	0	2	0	0	0	2	0,1
Acre	5	2	0	0	0	7	0,4
Amazonas	56	14	0	7	1	78	5,0
Roraima	0	0	0	1	0	1	0,1
Pará	812	49	1	306	5	1.173	74,7
Amapá	131	1	0	13	1	146	9,3
Tocantins	19	2	0	2	0	23	1,5
Nordeste	33	14	1	23	2	73	4,6
Maranhão	11	7	0	5	1	24	1,5
Piauí	0	3	0	1	0	4	0,3
Ceará	8	1	0	0	0	9	0,6
Rio Grande do Norte	1	0	1	0	0	2	0,1
Paraíba	0	0	0	1	0	1	0,1
Pernambuco	0	2	0	15	0	17	1,1
Sergipe	0	1	0	1	0	2	0,1
Bahia	13	0	0	0	1	14	0,9
Sudeste	0	2	1	8	1	12	0,8
Minas Gerais	0	0	0	6	0	6	0,4
Espírito Santo	0	0	0	0	1	1	0,1
Rio de Janeiro	0	1	0	0	0	1	0,1
São Paulo	0	1	1	2	0	4	0,3
Sul	25	0	3	0	0	28	1,8
Santa Catarina	24	0	0	0	0	24	1,5
Rio Grande do Sul	1	0	3	0	0	4	0,3
Centro-Oeste	0	14	0	12	1	27	1,7
Mato Grosso	0	4	0	0	0	4	0,3
Goiás	0	10	0	12	1	23	1,5
Brasil	1.081	100	6	372	11	1.570	100,0

A doença de Chagas crônica é tida por muitos pesquisadores como incurável (WORLD HEALTH ORGANIZATION, 2002). Mesmo após mais de 100 anos de sua descoberta, não existe vacina contra o *T. cruzi* e o tratamento com droga é insatisfatório (LAURIA-PIRES *et al.*, 2000). Assim, postulamos que tratamento e prevenção pressupõem reconhecimento prévio dos mecanismos de produção das lesões da doença de Chagas.

Na América Latina, o controle dos vetores é o método mais útil para prevenir a doença de Chagas. Também é vital o rastreio do sangue para prevenir a infecção

através de transfusão e transplante de órgãos, além do diagnóstico da infecção em mulheres grávidas e recém-nascidos em todo o mundo (WORLD HEALTH ORGANIZATION, 2015).

Estudos indicam que a doença de Chagas é desencadeada não apenas pela presença do *T. cruzi*, mas pelas modificações que a introgressão de seu DNA do Cinetoplasto (kDNA) produz no genoma do hospedeiro, e atribuímos a essas mutações (e por extensão, os *loci* em que ocorrem), como os causadores das reações de autoimunidade que caracterizam a doença (SIMÕES-BARBOSA *et al.*, 2006).

1.2.2. AGENTE ETIOLÓGICO

Como já dito anteriormente, o agente etiológico da doença de Chagas é o *Trypanosoma cruzi*, organismo eucarionte unicelular. Taxonomicamente, a espécie pertence ao reino *Excavata*, filo *Euglenozoa*, classe *Kinetoplastida*, ordem *Trypanosomatida*, família *Trypanosomatidae* e gênero *Trypanosoma*. (GRUBY, 1843; CHAGAS, 1909; BAKER, 1963; HONIGBERG, 1963; WHITTAKER e MARGULIS, 1978; CAVALIER-SMITH, 1981; 2003; MOREIRA *et al.*, 2004).

O *T. cruzi* pode se apresentar em diferentes formas celulares, dotadas de características morfológicas e biológicas distintas, em seu ciclo biológico, pleomorfismo este, que se torna evidente principalmente quando há a transição entre hospedeiro vertebrado (mamíferos) e invertebrado (triatomíneos). Devido a diversos fatores, o parasita pode manifestar-se sob as formas amastigota (arredondada, imóvel e sem flagelo livre), epimastigota (fusiforme, presença de membrana ambulante e com flagelo livre) e tripomastigota (alongada, longa membrana ambulante e com flagelo livre (Figura 5) (NEVES *et al.*, 2011).

Os principais mecanismos de transmissão da doença de Chagas são: vetorial, principal via de transmissão para o homem, ocorrendo por contaminação com as fezes do vetor; oral, mais importante via para os animais silvestres, ocorrendo a partir da ingestão de vetores e reservatórios infectados; transfusional, que é ocasional e ocorre pela infecção de receptores de transfusões sanguíneas; e transplacentário, via de transmissão congênita. O controle dos vetores é o método mais útil para prevenir a doença de Chagas, na América Latina. Também é vital o

rastreio do sangue para prevenir a infecção através de transfusão e transplante de órgãos, além do diagnóstico da infecção em mulheres grávidas e recém-nascidos (COURA, 2006; 2015; WORLD HEALTH ORGANIZATION, 2015).

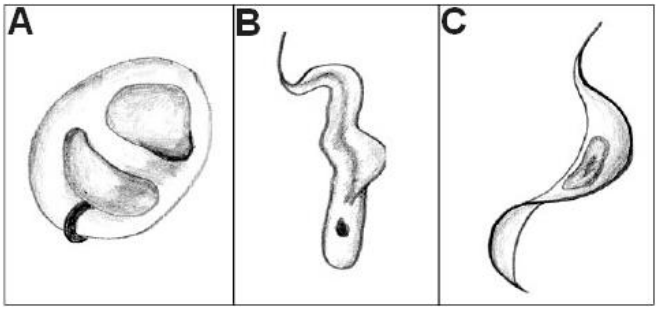


Figura 5: *Formas celulares do Trypanosoma cruzi* (TOSO *et al.*, 2011). **A:** Forma Amastigota; **B:** Forma Epimastigota; e **C:** Forma Tripomastigota.

O *T. cruzi* apresenta ciclo biológico heteroxênico, já que possui um hospedeiro vertebrado (várias espécies de mamíferos) e um invertebrado (alguns insetos hematófagos). Quando o triatomíneo infectado pratica o repasto sanguíneo em mamíferos, ocorre a eliminação de tripomastigotas sobre a pele ou mucosas do hospedeiro, próximo do local da picada, pelas fezes do inseto. Os tripomastigotas penetram no hospedeiro através da lesão tecidual pela picada/coceira ou em mucosas intactas. Dentro do hospedeiro, os tripomastigotas infectam células de uma grande variedade de tecidos, onde diferenciam-se em amastigotas intracelulares, que se multiplicam binariamente no citoplasma. Diferenciam-se novamente em tripomastigotas e são liberados em espaços intercelulares ou circulação sanguínea. Os tripomastigotas metacíclicos podem invadir novas células, num ciclo infectante contínuo, ou ser ingerido pelo triatomíneo durante o repasto em algum hospedeiro vertebrado. O triatomíneo infecta-se dos parasitas sanguíneos circulantes, que se diferenciam em epimastigotas no intestino médio do vetor e multiplicam-se por divisão binária. Os epimastigotas migram para o intestino posterior e diferenciam-se novamente na forma infectante dos hospedeiros vertebrados, o tripomastigota metacíclico (Figura 6) (BRENER, 1985; TYLER e ENGMAN, 2001).

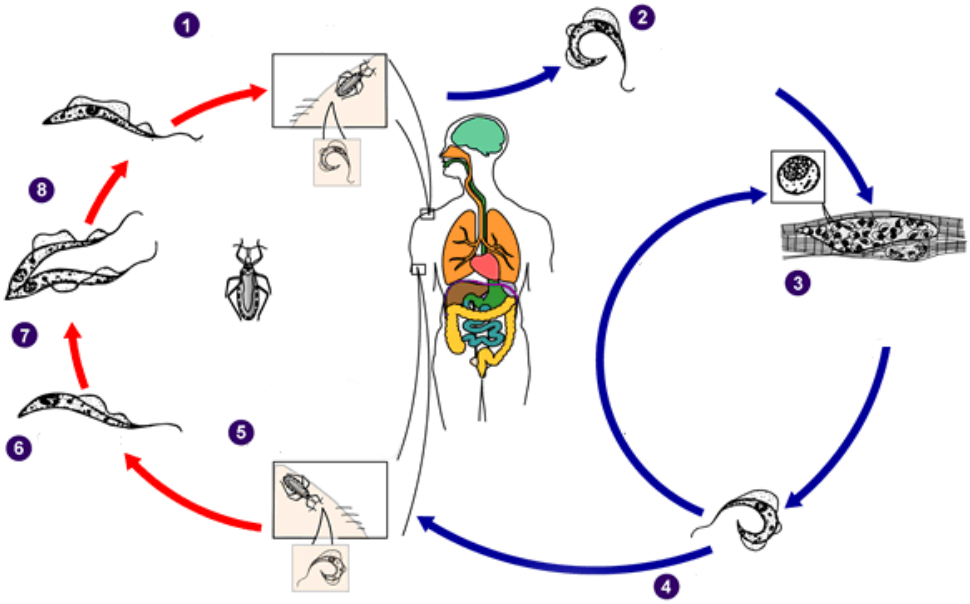


Figura 6: *Ciclo biológico do Trypanosoma cruzi* (CLAYTON, 2010). **1:** Infecção do homem pelo triatomíneo infectado; **2:** Diferenciação dos tripomastigotas em amastigotas; **3:** Multiplicação dos amastigotas; **4:** Liberação dos tripomastigotas no sangue; **5:** Infecção do triatomíneo por parasitas circulantes; **6:** Diferenciação dos tripomastigotas em epimastigotas; **7:** Multiplicação dos epimastigotas; **8:** Diferenciação dos epimastigotas em tripomastigotas.

1.2.3. PATOGENIA

A doença de Chagas pode ser dividida em duas fases: uma curta fase aguda e uma longa fase crônica (DUTRA *et al.*, 2005).

Cerca de 95% dos casos da doença de Chagas aguda (DCA) é assintomática, mas também podem haver manifestações clínicas leves, como febre, diarreia, fadiga ou taquicardia (TEIXEIRA, 2007).

No Brasil, a letalidade média anual por DCA ao longo dos últimos anos (2005 a 2013) foi de 4,21% (Figura 7). Também foram registrados 33 óbitos, sendo 55% deles em pacientes do sexo masculino com mediana de idade de 38 anos e 39% deles apresentando algum tipo de sinal de gravidade como, por exemplo, insuficiência cardíaca congestiva (MINISTÉRIO DA SAÚDE - BRASIL, 2015).

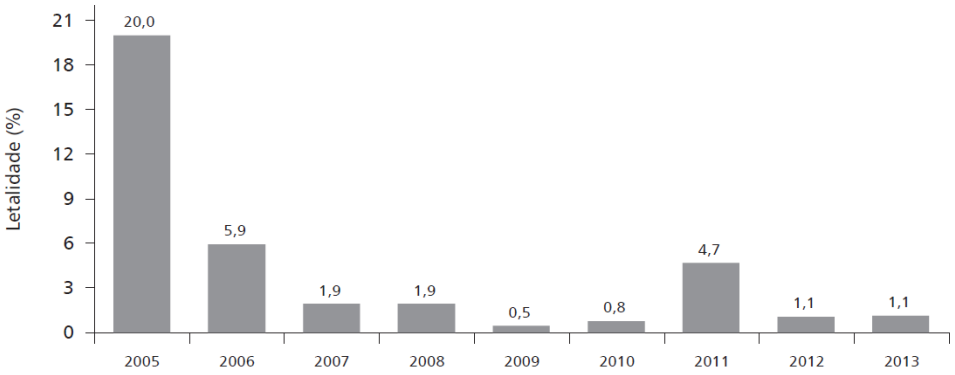


Figura 7: *Letalidade anual de doença de Chagas aguda no Brasil* (MINISTÉRIO DA SAÚDE - BRASIL, 2015). Os dados observados de 2005 a 2013 revelaram elevada letalidade (20,0%) em 2005, que coincidiu com o surto de Chagas aguda por transmissão oral em Santa Catarina. Nos anos subsequentes houve significativa redução da letalidade.

Usualmente, o paciente infectado pelo *T. cruzi* não fica doente quando a parasitemia é alta, podendo desenvolver as lesões graves da doença de Chagas crônica (DCC) até mesmo décadas mais tarde, quando o parasito é dificilmente encontrado. Estudos da doença em coelhos mostraram aspectos semelhantes àqueles que se descrevem para a doença humana, em que a patogênese da doença tem sido associada com infiltrados de células mononucleares e rejeição de células cardíacas do hospedeiro. A manifestação da doença pode não depender, exclusivamente, da ação direta do parasito sobre células alvo do hospedeiro e, por isso, muitos autores afirmam que a resposta imune geneticamente regulada desempenha papel importante na patogênese da doença de Chagas (TEIXEIRA *et al.*, 1978; LEON *et al.*, 2001; GIRONES e FRESNO, 2003; LEON e ENGMAN, 2003; TEIXEIRA, 2007).

Um problema muito importante na pesquisa sobre doença de Chagas é a determinação da patogenia das lesões associadas com as manifestações clínicas. A documentação que já existente sobre a transferência horizontal de kDNA de *T. cruzi* para o genoma de mamíferos e aves, provê informação fundamental para investigar a origem da autoimunidade (TEIXEIRA *et al.*, 1991; SIMÕES-BARBOSA *et al.*, 2006; TEIXEIRA *et al.*, 2011).

Até 30% dos infectados com DCC desenvolvem alterações cardíacas e até 10% desenvolvem alterações digestivas e/ou neurológicas, que podem exigir tratamento específico (WORLD HEALTH ORGANIZATION, 2015).

1.2.4. ORGANIZAÇÃO GÊNICA DO *Trypanosoma cruzi*

Os membros da ordem Kinetoplastida apresentam mitocôndria única que contém uma região rica em DNA (kDNA), o cinetoplasto, constituída por moléculas dupla-fita circulares, e dois tipos de anéis, os minicírculos e maxicírculos, concatenados em uma única rede. O kDNA lembra uma rede de pescar no bojo de uma canoa: a rede é formada de minicírculos e os maxicírculos formam a corda de puxar a rede (Figura 8). Antes da replicação, os minicírculos são decatenados e liberados pela Topoisomerase II ATP-dependente no saco flagelar (bojo da canoa) na matriz mitocondrial. A rede é duplicada e a progênie de minicírculos são concatenados no fim da fase S do ciclo celular. Várias proteínas ligantes de DNA interferem na iniciação da replicação do kDNA (LIU *et al.*, 2005).

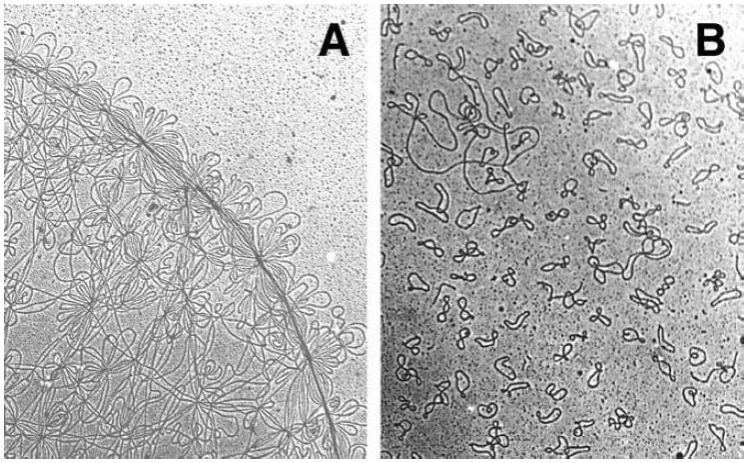


Figura 8: *Micrografia eletrônica da rede de kDNA em Crithidia fasciculata* (SHAPIRO *et al.*, 1999). **A:** Rede concatenada de kDNA com a presença dos minicírculos e dos maxicírculos. **B:** kDNA decatenado pela Topoisomerase II, com a presença de vários minicírculos livres e um maxicírculo.

Como demonstrado em trabalho com *Crithidia fasciculata*, o kDNA possui sequências com ângulos de torção que geram grande curvaturas (SCIPIONI *et al.*,

2004), o que dificulta sobremaneira o obtenção de sequências fidedignas para essas regiões.

Os maxicírculos são moléculas de DNA funcionalmente análogas às mitocôndrias dos eucariotos e codificam os produtos gênicos típicos desta organela (proteínas do complexo respiratório). Estão distribuídos em dezenas de cópias possuindo entre 20 e 40 mil pares de bases (Figura 9) (WESTENBERGER *et al.*, 2006). A expressão dos genes codificadores das proteínas de maxicírculos é bastante complexa e os RNAs transcritos são dependentes de processamento (edição de RNA) (LANDWEBER e GILBERT, 1994), mecanismo este, responsável por formar RNAs mensageiros (mRNAs) mitocondriais com códons de iniciação e de terminação corretos e com fases abertas de leitura (LUKES *et al.*, 2002; SIMPSON *et al.*, 2002).

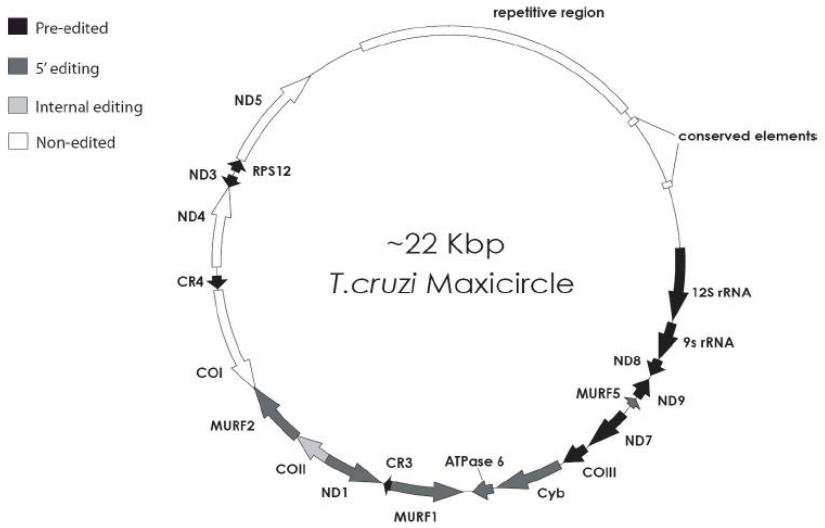


Figura 9: *Esquematização do maxicírculo de Trypanosoma cruzi* (WESTENBERGER *et al.*, 2006). O maxicírculo possui, aproximadamente, 22 mil pares de bases e todos os genes anotados são mostrados como setas, indicando a direção de codificação.

Os minicírculos são moléculas pequenas de 1,4 mil pares de bases (Figura 10), com sequência heterogênea e presentes como milhares de cópias concatenadas, formando a rede de kDNA. Codificam as moléculas de RNAs guias (gRNAs) que auxiliam na editoração dos mRNAs dos maxicírculos mitocondriais (SIMPSON, 1987). Apresentam uma peculiar organização molecular, com 4 regiões

conservadas (de 120 a 160 pares de bases – pb) intercaladas por 4 regiões variáveis (de 240 a 320 pb) (GUIMARO *et al.*, 2014).

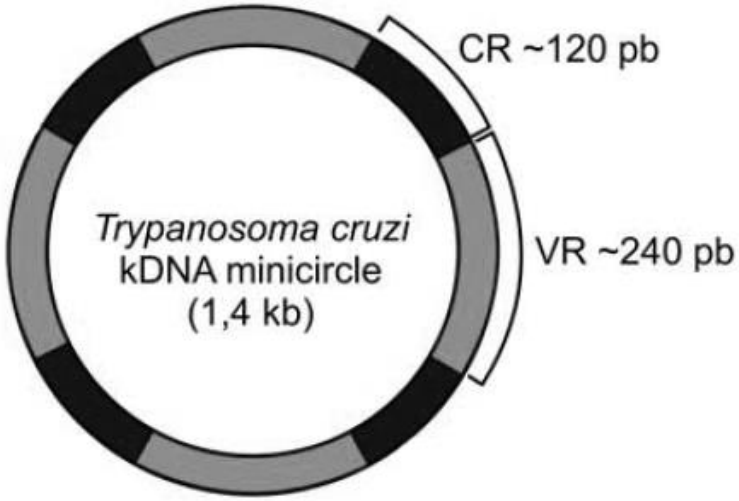


Figura 10: *Esquematização do minicírculo de Trypanosoma cruzi* (GUIMARO *et al.*, 2014). O minicírculo possui 1,4 mil pares de bases, com 4 regiões conservadas (CR) de 120 pb e 4 regiões variáveis (VR) de 240 pb.

1.3. INTRODUÇÃO À TRANSFERÊNCIA HORIZONTAL DE GENES

Numa perspectiva histórica, a transferência de DNA entre organismos não relacionados, situados em diferentes Reinos, é similar aos eventos sugeridos por Margulis e Sagan para explicar a origem de células eucariontes (MARGULIS e SAGAN, 2003). A transferência horizontal de genes (HGT) refere-se à troca de material genético entre células/genomas de espécies distantes filogeneticamente, cujo conceito não é novo, com estudos datando mais de 50 anos (FREEMAN, 1951; BROWN, 2003). Pode ser encarada como um mecanismo alternativo de aquisição de genes em que organismos transferem material genético para outros organismos geneticamente incompatíveis, em que um indivíduo pode dispor de uma reserva genética distinta de seu DNA, o que pode ser benéfico ou não

Entretanto, alguns acreditam que há exagero, possivelmente resultante de métodos inadequados de identificação de HGT (KURLAND *et al.*, 2003). Outros sugerem que a HGT seja uma força evolutiva (KLEIN e TAKAHATA, 2002). O

avanço científico do sequenciamento do genoma mostrou que a HGT está associada com crescimento e remodelagem de DNA (LANDER *et al.*, 2001). HGT pode ser estudada na célula *in vitro*, mas o mecanismo de integração do DNA exógeno é pouco reconhecido (TEIXEIRA *et al.*, 1994; SIMÕES-BARBOSA *et al.*, 2006).

Estudos *in vitro* mostraram que os elementos LINE-1 (longos elementos nucleares intercalados) são sítios eletivos de HGT (KUSAKABE *et al.*, 2001). A maioria das cópias de LINEs é truncada (PAVLICEK *et al.*, 2002) e apresentam nas regiões repetidas ricas em adenina os sítios de remodelamento de éxons (OSTERTAG e KAZAZIAN, 2001). Os LINEs carregam esses motivos diretos e invertidos de elementos repetitivos não-autônomos SINEs (curtos elementos nucleares intercalados), apresentando micro-homologias intrínsecas que se associam com as junções de Holliday, e polimorfismo local (ZINGLER *et al.*, 2005; BABUSHOK *et al.*, 2006; SIMÕES-BARBOSA *et al.*, 2006). O genoma de vertebrados contém LINEs e SINEs que invadiram o genoma de espécies há 150 milhões de anos (KAZAZIAN e MORAN, 1998). A HGT do DNA do parasito para o genoma do hospedeiro é um evento constante, documentado em modelos *in vitro* e *in vivo* (TEIXEIRA *et al.*, 1994; SIMÕES-BARBOSA *et al.*, 2006; HECHT *et al.*, 2010).

1.4. INTEGRAÇÃO DO kDNA DE *T. cruzi* NO GENOMA DO HOSPEDEIRO

Trabalhos prévios indicam que o kDNA do *T. cruzi* integra-se no genoma do hospedeiro vertebrado, como consequência da infecção (TEIXEIRA *et al.*, 1991; TEIXEIRA *et al.*, 1994). As sequências de minicírculos do DNA mitocondrial do parasito foram encontradas integradas em retrotransposons LINE-1 de mamíferos (Figura 11) (SIMÕES-BARBOSA *et al.*, 2006).

As integrações do kDNA ocorreram durante a infecção natural em coelhos, já em aves, a infecção tem que ser induzida intra ovo. A utilização desses modelos da doença pode esclarecer sobre a origem da sua patogênese. Por exemplo, o modelo da ave permite avaliar a importância da autoimunidade na patogênese da doença, já que as infecções pelo *T. cruzi* se instalam na primeira semana de desenvolvimento do embrião, e as aves nascem refratárias ao protozoário, devido à

sua temperatura corpórea. Diante desta observação em aves, foi verificado que as mutações no genoma de pacientes chagásicos são transferidas para os descendentes nas gerações subsequentes (SIMÕES-BARBOSA *et al.*, 2006; HECHT *et al.*, 2010).

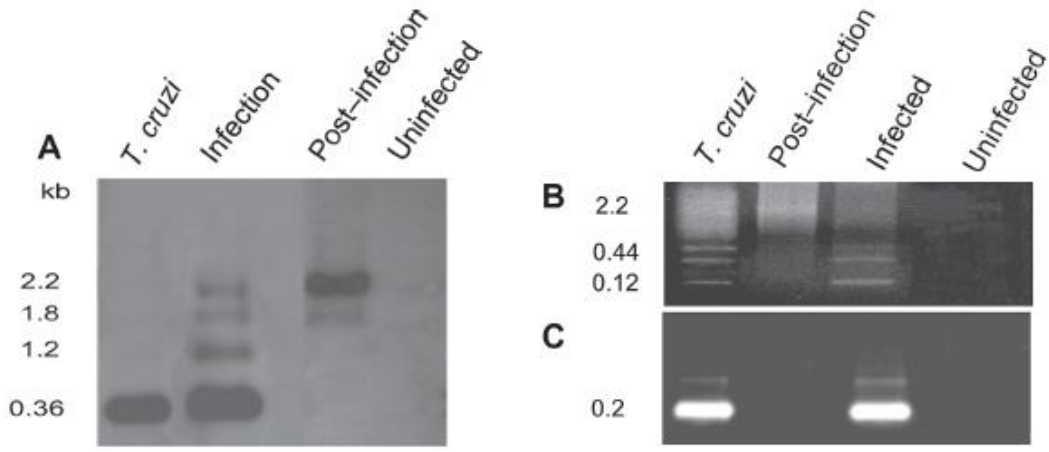


Figura 11: *Integração de minicírculos de T. cruzi no genoma de macrófagos* (SIMÕES-BARBOSA *et al.*, 2006). **A:** Hibridação Southern de macrófagos; **B:** Coloração com brometo de etídio dos minicírculos (amplificados por PCR); **C:** Coloração com brometo de etídio da cepa de *T. cruzi*.

Um aspecto interessante da relação parasita-hospedeiro nas infecções pelo *T. cruzi* parece ser a sua longa permanência no interior da célula, geralmente ao longo de toda a vida do organismo hospedador. Neste ambiente, o simbionte sofre ciclos de multiplicação, que podem ocorrer sincronicamente com a replicação da célula hospedeira. Nesta ocasião surgem as condições essenciais para a transferência de DNA entre as duas espécies filogeneticamente muito distantes. Regiões ricas em micro-homologias com mais de 6 repetições dos nucleotídeos adenina e citosina (motivos CA's) medeiam a integração dos minicírculos de kDNA exógeno no DNA hospedeiro (Figura 12) (HECHT *et al.*, 2010).

O kDNA integrado em LINE-1 do macrófago humano pode ser mobilizado do cromossomo 4 para o cromossomo 5, via transcrição reversa do elemento transponível-kDNA, rompendo a fase aberta de leitura e silenciando o gene p15. Esta é uma demonstração de patologia molecular decorrente da mutação induzida (SIMÕES-BARBOSA *et al.*, 2006). Os retrotransposons LINE-1 são conhecidos

genitores de inserções mutagênicas no lócus da β-globina e em outros genes RP2, uma vez que eles possuem a maquinaria endógena – transposases: polimerase I e transcriptase reversa - para mobilização de sequências de DNA dentro do genoma, gerando rearranjo de éxons (LANDER *et al.*, 2001). Promotores situados a montante no genoma iniciam a transcrição de LINE-1 geralmente confinado em células (TRELOGAN e MARTIN, 1995), mas a retrotransposição de um LINE-1 somático foi correlacionado com hemofilia (KAZAZIAN e MORAN, 1998). Por último, foram encontradas mutações em LINE-1 em pacientes com cardiomiopatia dilatada (YOSHIDA *et al.*, 1998).

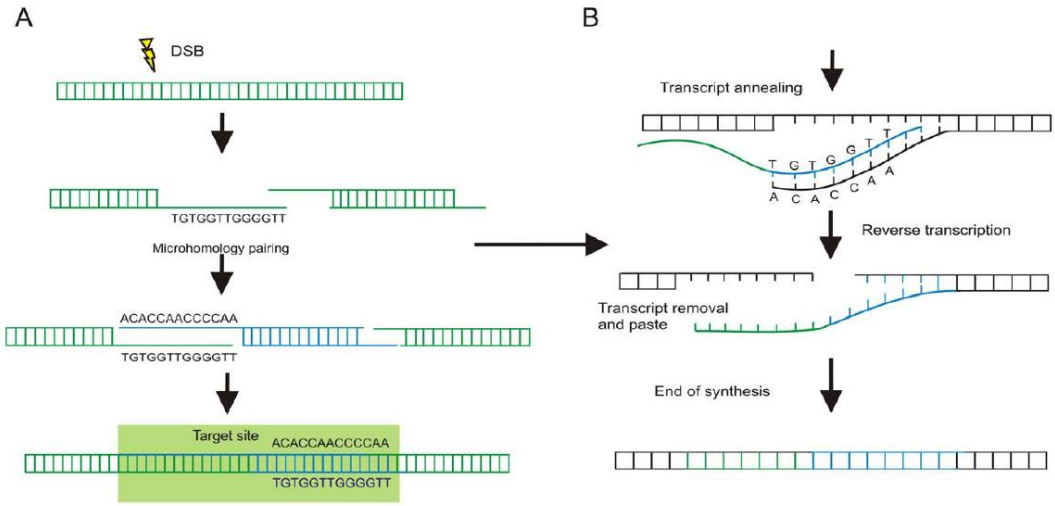


Figura 12: *Integração e replicação dos minicírculos no genoma humano* (HECHT *et al.*, 2010). **A:** Infecção induzida no DNA hospedeiro (verde), quebra da dupla fita (DSB) e integração da sequência do minicírculo de kDNA (azul) mediada por regiões de micro-homologias; **B:** Replicação da sequência kDNA-LINE por transcrição reversa, utilizando a maquinaria de recombinação autônoma do LINE-1.

Um grande volume de informação já foi produzido, os quais comprovam a transferência de sequências de minicírculos de kDNA do *T. cruzi* para os elementos LINE-1 nos cromossomos da célula hospedeira, bem como que micro-homologias ricas em motivos CA's medeiam essas integrações (TEIXEIRA *et al.*, 1991; TEIXEIRA *et al.*, 1994; SIMÕES-BARBOSA *et al.*, 2006; HECHT *et al.*, 2010).

2. OBJETIVOS

2.1. OBJETIVO GERAL

O objetivo deste trabalho é propor uma maneira eficiente, fácil e rápida para a busca e localização de múltiplas assinaturas dos sinalizadores que propiciam a introgressão do kDNA exógeno de *T. cruzi* no genoma humano, através de um conjunto de scripts, baseados na linguagem R e adaptados a grandes arquivos de sequências.

2.2. OBJETIVOS ESPECÍFICOS

Este trabalho objetiva:

1. buscar sinalizadores que indicam a introgressão do kDNA exógeno de *T. cruzi* no genoma humano;
2. montar um mapa genético dos pontos prováveis de integração do kDNA;
3. identificar se há cromossomos com maior “fragilidade” ou maior propensão a serem mutados por meio de herança horizontal.

3. MATERIAL E MÉTODOS

3.1. SOFTWARE R

O ambiente do software R foi utilizado para as análises computacionais, provendo uma ampla variedade de técnicas estatísticas (modelagens, classificações, clustering e testes estatísticos clássicos) e gráficas para a manipulação dos dados. Devido ao status de software em desenvolvimento ativo, a última versão utilizada e completamente compatível com as análises e pacotes utilizados é a versão "R 3.2.3", de 10/12/2015, disponível em: https://www.R-project.org/ (R CORE TEAM, 2015).

Disponível para diversas plataformas, incluindo os sistemas Unix e Windows, e contando, atualmente, com cerca de 8000 pacotes adicionais disponíveis através de repositórios, como CRAN (*Comprehensive R Archive Network*) e Bioconductor, o principal motivo da escolha da linguagem "R" foi sua alta compatibilidade e disponibilidade de pacotes de funções e dados biológicos, sobremaneira com bancos de dados de genomas.

3.2. PACOTES ADICIONAIS

3.2.1. PACOTES BÁSICOS

Os pacotes *"R base"*, responsável pelas funções básicas de linguagem, entrada/saída de arquivos, aritmética e suporte de programação; *"R utils"*, contendo uma coleção de funções utilitárias; e *"R grDevices"*, responsável pelas funções gráficas e controle dos dispositivos gráficos utilizados, são intrínsecos ao próprio software e foram utilizados na montagem dos scripts (R CORE TEAM, 2015).

3.2.2. PACOTE *"doParallel"*

A fim de distribuir e executar tarefas em paralelo e, consequentemente, diminuir o tempo de rotina, foi utilizado o pacote *"doParallel"* (versão 1.0.10, de 14/10/2015), disponível em: https://CRAN.R-project.org/package=doParallel, na construção de todos os scripts. Atua como uma interface entre os pacotes *"parallel"* e *"foreach"*, incorporados ao software R desde a versão 2.14.0, e é compatível com

sistemas Unix e Windows. É caracterizado pela distribuição das tarefas, em especial as que envolvam laços de repetição, entre cada núcleo de uma máquina que possua processadores multinúcleos, criando unidades lógicas gerenciadas como um objeto "cluster" (REVOLUTION ANALYTICS e WESTON, 2015).

3.2.3. PACOTE *"gtools"*

O pacote *"gtools"* (versão 3.5.0, de 26/05/2015), disponível em: https://CRAN.R-project.org/package=gtools, é caracterizado por possuir diversas funções estatísticas, responsáveis por auxiliar a programação em "R". Foi utilizado para gerar as máscaras combinatórias, para posterior busca e localização no genoma (WARNES *et al.*, 2015).

3.2.4. PACOTE *"BiocInstaller"*

O pacote *"BiocInstaller*" (versão 1.20.1, de 12/02/2016), disponível em: https://bioconductor.org/biocLite.R, foi utilizado para instalar o Bioconductor no ambiente do software R. O Bioconductor é um projeto de código aberto e gratuito, iniciado em 2001, para a análise e compreensão de dados genômicos gerados por experimentos *in silico* na biologia molecular, através do fornecimento de ferramentas de alto rendimento, baseadas na linguagem R (TENENBAUM e BIOCORE TEAM, 2016).

3.2.5. PACOTE *"BSgenome"*

Fornecido pelo Bioconductor no software R e disponível em: https://www.bioconductor.org/packages/BSgenome/, o pacote *"BSgenome"* (versão 1.38.0, de 12/02/2016) é um *container* para armazenamento de sequências completas de genomas de diversos organismos, estas por sua vez, criadas e disponibilizadas por terceiros para a comunidade Bioconductor (PAGES, 2016).

3.2.6. PACOTE *"BSgenome.Hsapiens.NCBI.GRCh38"*

Por meio do pacote *"BSgenome"* foi obtido o pacote de dados genômicos *"BSgenome.Hsapiens.NCBI.GRCh38"* (versão 1.3.1000, de 16/12/2015). Baseado

em *containers "Biostrings"* para cadeias de caracteres, este pacote de dados compreende a sequência completa do genoma humano (*Homo sapiens*), como fornecida pelo *Genome Reference Consortium Human Build 38* (GRCh38), em 17/12/2013 (THE BIOCONDUCTOR DEV TEAM, 2015).

3.2.7. PACOTE *"Biostrings"*

O pacote *"Biostrings"* (versão 2.38.4, de 12/02/2016), disponível em: https://www.bioconductor.org/packages/Biostrings/, compreende *containers* eficientes de cadeias de caracteres, algoritmos de busca de correspondências e diversas outras utilidades para manipulação de grandes sequências biológicas (como sequências de DNA ou RNA) ou conjuntos de sequências de maneira rápida e fácil. Foi utilizado para a varredura do genoma em busca de correspondências com as sequências alvo (PAGES *et al.*, 2016).

3.3. CONSTRUÇÃO DOS SCRIPTS

3.3.1. OBTENÇÃO DAS SEQUÊNCIAS

O primeiro passo para varrer o genoma humano em busca das regiões alvos (ricas em motivos CA's) foi a criação de arquivos com todas as combinações possíveis destes dois nucleotídeos, através de um processo matemático conhecido por permutação. Deste modo, um script foi construído para permutar estes elementos, em sequências de tamanho pré-estabelecido, utilizando a função *"permutations"*, do pacote *"gtools"*. Para isso, definimos os nucleotídeos a serem permutados ("A" e "C") e o tamanho de cada sequência (6 nucleotídeos).

Por padrão, este pacote separa cada elemento da cadeia de caracteres permutada em colunas, assim, a função *"gsub"* foi utilizada para agrupar os caracteres, sem espaçamento entre eles, em uma única coluna.

3.3.2. VARREDURA DO GENOMA

Após a criação do arquivo com todas sequências CA's possíveis, foi necessária a criação de um novo script, este, por sua vez, responsável por varrer o genoma humano em busca de correspondências.

O primeiro passo do processo de varredura foi carregar a biblioteca de genoma *"BSgenome.Hsapiens.NCBI.GRCh38"* e definir qual cromossomo será analisado. Depois o arquivo de entrada contendo as sequências CA's, geradas anteriormente, foi carregado, como uma matriz, e alocado como um conjunto de objetos *"DNAString"* em um *container "DNAStringSet"*, através do pacote *"Biostrings"*. Para manipularmos essas sequências de DNA, utilizamos *containters* da classe *"PDict"*, que armazena dicionários pré-processados de caracteres de DNA (*"DNAStringSet"*).

Os *containers "PDict"* podem ser rapidamente requeridos pela função *"matchPDict"*, que utiliza esse dicionário pré-processado com sequências de tamanho constante (tamanho 6) e busca as correspondências, de maneira eficiente, retornando um objeto "*MIndex"* (um *container* básico que armazena cada correspondência encontrada, informando a composição, tamanho, posição inicial e posição final).

A função *"matchPDict"* percorre o genoma base-a-base, assim, caso porções da correspondência encontrada anteriormente constituam uma nova correspondência, também é exportada (Figura 13).

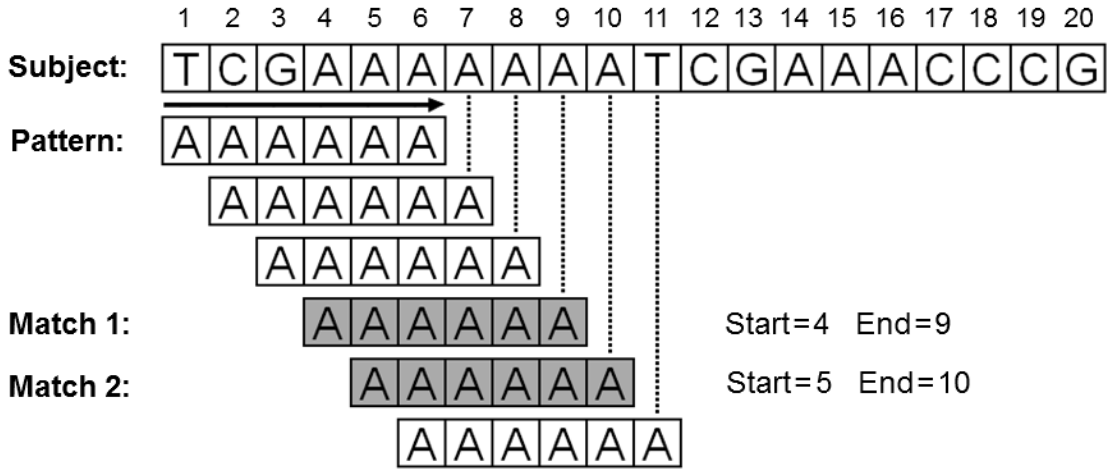


Figura 13: Função *"matchPDict"* para busca de correspondências, utilizando um dicionário com tamanho constante, em banco de dados de genoma.

3.3.3. AGRUPAMENTO DE REGIÕES JUSTAPOSTAS

Como as sequências buscadas são de tamanho constante, foi necessário agrupar as correspondências justapostas encontradas anteriormente, para que o número de correspondências coincida com a quantidade real de nucleotídeos no genoma.

Deste modo, foi construído um script que reordena as correspondências por posição inicial e verifica se a mesma encontra-se dentro da correspondência anterior. Caso pertença, a sequência do grupo anterior é conservada e o último caractere da correspondência abaixo é anexado à antiga sequência, utilizando a função *"substr"*, permitindo assim que hajam sequências com tamanho maior que 6 caracteres.

3.3.4. PLOTAGEM DO MAPA GENÉTICO

A plotagem das correspondências de cada cromossomo foi realizada a partir de um script que analisa um arquivo contendo as posições iniciais e finais das correspondências e plota cada uma em arquivos ".png", através da função *"plot"* do pacote gráfico *"grDevices"*, no software R. O pacote de genoma foi utilizado para obter o tamanho do cromossomo que será analisado.

O gráfico final foi particionado em 50 imagens (este valor pode ser alterado livremente), mostrando a quantidade e localização das sequências naquela região do cromossomo.

3.3.5. CONTAGEM DE CORRESPONDÊNCIAS

O último script desenvolvido foi o de contagem de correspondências, responsável pela contagem total de cada correspondência, em determinado cromossomo, e pela contagem das sequências por janelas, em bandas cromossômicas.

Para a contagem total, utilizamos as funções *"table"*, responsável por somar todas as sequências idênticas de uma matriz, e *"order"*, para reordenar as sequências somadas, por frequência.

A contagem por bandas cromossômicas também foi realizada, utilizando como entrada os valores da posição inicial e final de cada banda citogenética do arquivo *"cytoBand.txt.gz"*, fornecido pelo banco de dados de anotação do genoma (ftp://hgdownload.cse.ucsc.edu/goldenPath/hg38/database/), da *University of California Santa Cruz* (UCSC) Genome Bioinformatics, em 10/08/2014.

1	0	4300000	p13
2	4300000	9400000	p12
3	9400000	13700000	p11.2
4	13700000	15000000	p11.1
5	15000000	17400000	q11.1
6	17400000	21700000	q11.21
7	21700000	23100000	q11.22
8	23100000	25500000	q11.23
9	25500000	29200000	q12.1
10	29200000	31800000	q12.2

Figura 14: *Exemplo do arquivo de entrada "cytoband.txt"*. Através deste arquivo obtivemos a posição inicial (coluna 1), posição final (coluna 2) e nome das bandas cromossômicas (coluna 3), para cada cromossomo humano.

4. RESULTADOS E DISCUSSÃO

4.1. SCRIPT - PERMUTAÇÃO DE NUCLEOTÍDEOS

O script de permutação de nucleotídeos foi desenvolvido, inicialmente, para permutar apenas dois elementos em sequências de tamanho predeterminado, sem a utilização de nenhum pacote adicional ao software R, em que a definição dos caracteres a serem permutados era manual (Figura 15).

```
String_composition = readline("Define elements->")
Nucleo = strsplit(String_composition, "")[[1]]
word=""
n=-1
while (n<1 | n>99){
  String_length = readline("Define length->")
  n = as.integer(String_length)
  if (n>1000) break
}

if(n<1000){
  inic=2^(n)
  fim=2^(n+1)-1
  w=1
  Vector=matrix(nrow=2^n,ncol=1)
  Namefile=paste("OUTPUT_Permutacao_",n,".csv",sep="")
for(x in inic:fim){
  cod=intToBits(x)
  for(z in 1:n){
    word=paste(word,Nucleo[as.integer(cod[z])+1],sep="")
  }

  print(word)
  Vector[w]=word
  w=w+1
  word=""
}

write.table(Vector, file =Namefile,
            append = FALSE, quote = FALSE, dec=",",
            sep = "\t",row.names=FALSE, col.names=FALSE)
}
```

Figura 15: Script funcional para permutação de 2 nucleotídeos sem a utilização de pacotes adicionais ao Software R.

Porém, para que o alcance do script fosse ampliado, utilizamos o pacote *"gtools"* (permuta as sequências com qualquer número de caracteres) e o pacote *"doParallel"* (paraleliza a rotina), para o desenvolvimento de um novo script para permutação (Anexo 1).

Script: String Permutation

No input file
Output: Matrix with all possible string permutations

1 **Select: [1] Permute nucleic acids codes; or [2] Permute amino acids codes**
2 **makeCluster ("doParallel")**
3 **Set characters to be permuted**
4 **Set the length of string**
5 **Permute elements with "gtools"**
6 n^r (n = elements; r = string)
7 Allow repeats
8 Merge all columns
9 Remove spaces between characters
10 **Export Vector (Output)**

Figura 16: Resumo do script responsável por permutar sequências de caracteres no Software R.

Esse novo script tem por características, a possibilidade de escolha entre permutar códigos de ácidos nucleicos ou códigos de aminoácidos, utilizando nomenclatura adaptada da União Internacional de Química Pura e Aplicada (IUPAC) (CORNISH-BOWDEN, 1985), e a possibilidade de escolha do tamanho da sequência a ser gerada.

Uma matriz contendo todas as permutações possíveis dos elementos informados, em sequências de tamanho predeterminado dispostas em linhas, foi gerada e posteriormente exportada para o arquivo de saída (Figura 17).

1	AAAAAA	17	ACAAAA	33	CAAAAA	49	CCAAAA
2	AAAAAC	18	ACAAAC	34	CAAAAC	50	CCAAAC
3	AAAACA	19	ACAACA	35	CAAACA	51	CCAACA
4	AAAACC	20	ACAACC	36	CAAACC	52	CCAACC
5	AAACAA	21	ACACAA	37	CAACAA	53	CCACAA
6	AAACAC	22	ACACAC	38	CAACAC	54	CCACAC
7	AAACCA	23	ACACCA	39	CAACCA	55	CCACCA
8	AAACCC	24	ACACCC	40	CAACCC	56	CCACCC
9	AACAAA	25	ACCAAA	41	CACAAA	57	CCCAAA
10	AACAAC	26	ACCAAC	42	CACAAC	58	CCCAAC
11	AACACA	27	ACCACA	43	CACACA	59	CCCACA
12	AACACC	28	ACCACC	44	CACACC	60	CCCACC
13	AACCAA	29	ACCCAA	45	CACCAA	61	CCCCAA
14	AACCAC	30	ACCCAC	46	CACCAC	62	CCCCAC
15	AACCCA	31	ACCCCA	47	CACCCA	63	CCCCCA
16	AACCCC	32	ACCCCC	48	CACCCC	64	CCCCCC

Figura 17: Arquivo de saída do script de permutação, ilustrando as 64 permutações possíveis com 2 nucleotídeos em sequências de tamanho 6.

4.2. SCRIPT - BUSCA, AGRUPAMENTO E PLOTAGEM DE SEQUÊNCIAS

A junção das tarefas de busca, agrupamento e plotagem, em um único script (Anexo 2), nos permitiu exportar os arquivos de saída e os carregar para a próxima análise em uma mesma rotina, tornando o trabalho com estas ferramentas mais funcional. O script conta com a possibilidade de escolha entre: busca, agrupamento e plotagem; apenas buscar; apenas agrupar; ou apenas plotar (Figura 18).

A primeira parte do script consiste em buscar dentro de banco de dados de genomas disponíveis no Bioconductor para o software R (neste caso, no genoma

humano) todas as correspondências de determinada(s) sequência(s), carregadas previamente como arquivo de entrada.

Script: Matches Search, Grouping and Plotting

Input: Matrix with genome sequences
Output: Files with matchings (grouped or not) and .png images

```
Select: [1] Search, group and plot; [2] Only search; [3] Only group; or [4] Only plot
makeCluster ("doParallel")
Searching strings in genome
    Require genome library (e.g. "BSgenome.Hsapiens.NCBI.GRCh38")
    Set which chromosome will be searched
    Load input strings
    Store strings as "DNAStringSet" container for genome libraries
    Convert strings in "PDict" object
    Search matches in genome using "matchPDict" function
    Export Vector (Output 1)
Grouping juxtaposed sequences
    Load input matches (e.g. Output 1)
    Reorder string rows by match start postition
    for i in 2:end row do
        Read each match start and end position
        if start position [i] <= end position [i-1] do
            Conserve old sequence and add last characther of string [i]
            Consensus region size = end position - start position + 1
    Export Vector (Output 2)
Plotting matches
    Load input sequences (e.g. Output 1 or Output 2)
    Require genome library and set which chromosome will be plotted
    Open new graphic device
    Set chromosome length and how many bp each image will display
    for i in seq from 0, to chromosome lenght, by window size do
        Create a new plot frame
        for r in 1:end row do
            if match start position [r] is inside plot window = plot frame
        Paste graphic content to graphic device and export in .png (Output 3)
    Shut down graphic device
```

Figura 18: Resumo do script responsável por buscar, agrupar e plotar as correspondências CA's no Software R.

A ideia inicial era realizar a busca localmente, com o genoma baixado previamente da base de dados G*enBank* (ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/). Porém, em cada rotina, os dados genômicos do cromossomo a ser analisado tinham de ser carregados em arquivos de entrada, aumentando ainda mais o tempo de trabalho. Isso fez com que optássemos por utilizar o pacote de genoma humano

(*"BSgenome.Hsapiens.NCBI.GRCh38"*), disponibilizado pela NCBI para o ambiente do software R.

É importante ressaltar que o *container "DNAStringSet"* (do pacote *"Biostrings"*), dedicado a armazenar longas sequências de caracteres biológicos, permite apenas o trabalho com sequências baseadas no alfabeto de DNA, de acordo com a nomenclatura da IUPAC, já que as letras são codificadas para otimização em algoritmos de busca. Portanto, no script anterior, de permutação, deve-se seguir as orientações para a utilização da opção com os códigos de ácidos nucleicos.

O pré-processamento dos dados através da função *"matchPDict"* foi utilizado em detrimento ao *"matchPattern"* ou *"vmatchPattern"*, pois a rotina é geralmente mais eficiente nos laços de repetição, o que diminuiu, de maneira considerável, o esforço computacional e, consequentemente, o tempo de rotina do script de busca (PAGES *et al.*, 2016).

O processo de busca gerou um arquivo de saída que nos informa a posição inicial, posição final, tamanho e composição das correspondências (Figura 19).

		p_start	p_end	width	group_name
1	1	10510357	10510362	6	AAAAAA
2	1	10510358	10510363	6	AAAAAA
3	1	10510359	10510364	6	AAAAAA
4	1	10510360	10510365	6	AAAAAA
5	1	10510519	10510524	6	AAAAAA
6	1	10510520	10510525	6	AAAAAA
7	1	10510521	10510526	6	AAAAAA
8	1	10510522	10510527	6	AAAAAA
9	1	10510523	10510528	6	AAAAAA

Figura 19: *Exemplo do primeiro arquivo de saída (Output 1) do script de busca, agrupamento e plotagem*. O *Output 1* informa o grupo (coluna 1), posição inicial (coluna 2), posição final (coluna 3), tamanho (coluna 4) e composição (coluna 5), de todas as correspondências de tamanho constante.

Por padrão, a busca utilizando o *"Biostrings"* não agrupa as correspondências justapostas. Dessa maneira, a segunda parte desse script é responsável, justamente, por agrupar essas correspondências armazenadas no *Output 1*.

O processo de agrupamento se inicia com a criação de um *data frame* a partir do *Output 1*, que foi ordenado de forma crescente pelas posições iniciais das correspondências.

Foi definido que, caso a posição inicial de dada sequência fosse menor que a posição final da sequência prévia, ambas deveriam encontrar-se dentro de uma mesma sequência de número maior. Essas linhas foram então agrupadas, sempre mantendo o último caractere da sequência a ser anexada.

As novas sequências foram reagrupadas em uma matriz e exportadas para um novo arquivo de saída, similar ao primeiro, mas com sequências de todos os tamanhos possíveis, de 6 até 447 nucleotídeos, no presente trabalho (Figura 20).

	seq	p_start	p_end	width	group_name
2	1	10510321	10510326	6	CCAACA
3	2	10510332	10510338	7	AAACCCC
4	3	10510357	10510365	9	AAAAAAAAA
5	4	10510469	10510478	10	CACACCACCC
6	5	10510519	10510528	10	AAAAAAAAAA
7	6	10510634	10510641	8	AAAACAAA
8	7	10510656	10510661	6	CAAAAA
9	8	10510730	10510735	6	AAACAA
10	9	10510748	10510753	6	ACAAAA

Figura 20: *Exemplo do segundo arquivo de saída (Output 2) do script de busca, agrupamento e plotagem*. O *Output 2* informa a contagem (coluna 1), posição inicial (coluna 2), posição final (coluna 3), tamanho (coluna 4) e composição (coluna 5), de todas as correspondências justapostas agrupadas.

A última parte do script nos fornece a opção de plotar as correspondências de determinado cromossomo, em arquivos de saída de imagens “.png”, utilizando o pacote gráfico do software R (Figura 21).

A análise dos dados de plotagem se mostrou ineficaz para elucidar os locais mais prováveis de ocorrer integração do kDNA de *T. cruzi* no genoma humano, visto que existem um grande número de correspondências em cada cromossomo (levando a pintar praticamente toda a área da janela de cada imagem .png), com distribuição homogênea.

Dessa maneira, não foi possível, por meio apenas desta técnica, tirar conclusões a respeito dos pontos de mutação utilizando esse script de plotagem. Faz-se necessário ainda um grande avanço nas buscas para esse tipo de inferência, precisaríamos sobrepor esse mapa obtido, sobre um mapa dos *loci* codantes no genoma e ainda sobre um mapa de localização potencial de elementos transponíveis.

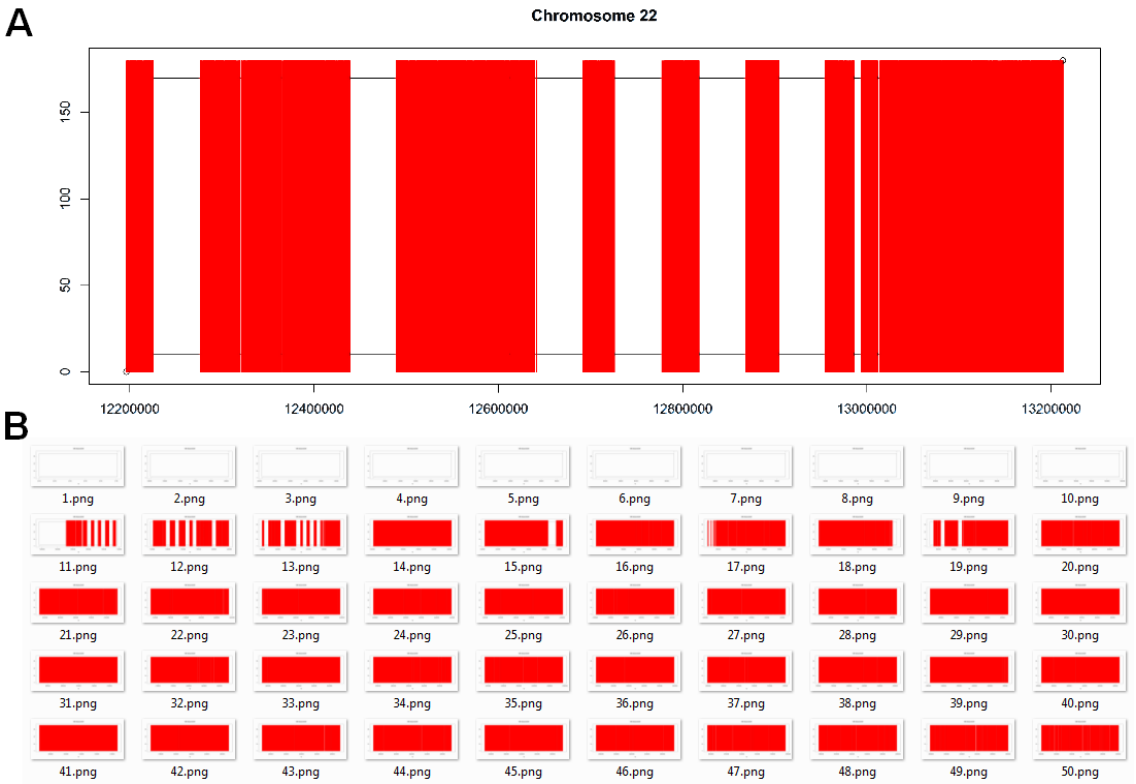


Figura 21: *Exemplo dos arquivos de saída (Output 3) do script de busca, agrupamento e plotagem.* **A:** Cada correspondência está representada por uma linha vermelha dentro de uma janela cromossômica, representando, neste exemplo, uma porção do Cromossomo 22; **B:** As 50 imagens que compõem todas as correspondências no Cromossomo 22.

4.3. SCRIPT - CONTAGEM DE CORRESPONDÊNCIAS

O script de contagem de correspondências (Anexo 3) foi construído para permitir analisar a frequência de cada correspondência por cromossomo e o número de correspondências em cada janela cromossômica (Figura 22).

Script: Matches Counting

Input: Matrix with sequences
Output: Files with matches counting

```
1  Select: [1] Count Total Matches; or [2] Count Windows Matches
2  makeCluster ("doParallel")
3  Set which chromosome will be counted
4  Counting total matches by chromosome
5      Load input sequences
6      Sum all identical strings ("table")
7      Reorder rows by string frequency
8      Export Vector (Output 1)
9  Counting matches by chromosomal set windows
10     Load input sequences
11     Load input windows size
12     for i in 1:end row do
13         if match start position [i] is inside window = count match
14         if match start position [i] > window size = break
15     Export Vector (Output 2)
```

Figura 22: Resumo do script responsável pela contagem das correspondências CA's no Software R.

A primeira parte do script de contagem é responsável por somar todas as correspondências idênticas de cada cromossomo e exportar os valores da frequência das sequências para um arquivo de saída (Figura 23).

1	group	freq	11	CCAACA	5471	21	CCACCC	3767
2	CCCAAA	10323	12	CCCCCA	5366	22	AAAAAAA	3667
3	AAAAAA	9003	13	AAAACA	5079	23	AACAAA	3549
4	AAACAC	7905	14	ACAAAA	4842	24	AACACA	3531
5	CACCCA	6837	15	AACCCA	4745	25	CACCAC	3493
6	CCACCA	6648	16	CAAAAA	4693	26	CCAAAA	3310
7	AAAAAC	6532	17	CCCACA	4464	27	CCCCAC	3277
8	CCCACC	5656	18	CACACC	4009	28	ACCCCA	3273
9	CACACA	5547	19	CACAAA	3914	29	ACCACA	3207
10	ACAAAAA	5532	20	AAACAA	3796	30	AAACCC	2990

Figura 23: *Exemplo do arquivo de saída (Output 1) do script de contagem*. O *Output 1* informa a composição (coluna 1) e a frequência (coluna 2) de cada uma das diferentes correspondências que aparecem no Cromossomo 22.

A análise desses dados nos permitiu encontrar as correspondências mais frequentes em cada cromossomo (Tabela 2). Observamos que as correspondências mais comuns são as menores, com 6 ou 7 nucleotídeos, deste modo incluímos as sequências maiores que 20 nucleotídeos mais frequentes, já que regiões maiores representam uma maior probabilidade para o possível pareamento e integração.

Tabela 2: Correspondências mais encontradas em cada cromossomo e sua respectiva frequência.

Chr	Sequência	Freq	Sequência (≥20)	Freq
1	AAAAAA	74449	CAAAAAAAAAAAAAAAAAAA	795
2	AAAAAA	84472	CAAAAAAAAAAAAAAAAAAA	675
3	AAAAAA	72179	CAAAAAAAAAAAAAAAAAAA	514
4	AAAAAA	75045	CAAAAAAAAAAAAAAAAAAA	438
5	AAAAAA	65534	CAAAAAAAAAAAAAAAAAAA	469
6	AAAAAA	62469	CAAAAAAAAAAAAAAAAAAA	469
7	AAAAAA	54856	CAAAAAAAAAAAAAAAAAAA	505
8	AAAAAA	50839	CAAAAAAAAAAAAAAAAAAA	382
9	AAAAAA	40147	CAAAAAAAAAAAAAAAAAAA	409
10	AAAAAA	42846	CAAAAAAAAAAAAAAAAAAA	421
11	AAAAAA	42374	CAAAAAAAAAAAAAAAAAAA	380
12	AAAAAA	45336	CAAAAAAAAAAAAAAAAAAA	450
13	AAAAAA	38097	CAAAAAAAAAAAAAAAAAAA	208
14	AAAAAA	30375	CAAAAAAAAAAAAAAAAAAA	287
15	AAAAAA	26472	CAAAAAAAAAAAAAAAAAAA	299
16	AAAAAA	21757	CAAAAAAAAAAAAAAAAAAA	320
17	CCCAAA	23963	CAAAAAAAAAAAAAAAAAAA	452
18	AAAAAA	27019	CAAAAAAAAAAAAAAAAAAA	173
19	CCCAAA	20673	CAAAAAAAAAAAAAAAAAAA	413
20	AAAAAA	17104	CAAAAAAAAAAAAAAAAAAA	220
21	AAAAAA	13533	CAAAAAAAAAAAAAAAAAAA	94
22	CCCAAA	10323	CAAAAAAAAAAAAAAAAAAA	197
X	AAAAAA	54653	CAAAAAAAAAAAAAAAAAAA	379
Y	AAAAAA	8925	CAAAAAAAAAAAAAAAAAAA	61

Também destacamos as sequências mais longas encontradas no genoma humano (Anexo 4), contendo mais de 100 nucleotídeos cada, dentre as quais podemos destacar a maior correspondência, com 447 nucleotídeos (Figura 24).

```
ACACACACAC ACACCCCCCA CACACACACA CCCCCCACAC ACACCACACA
CACACCCCAC ACACACAACC ACACCCCACA CACACAACCA CACACACACC
ACACACACAC CCCACACACA CCACACACAC ACCACACACC CCACACACAC
ACCCCACACA CACCACACAC CACACACACA CCCCACACAC AACCACACAC
CACACACCAC ACACACACCA CACACCACAC CACACACACA CCACACCACA
CACACACCAC ACACACACAC CACACCACAC ACACCACACA CACACCACAC
AACACCCCCC ACACACACAC CACACACACA CACCACACAC ACCACACACA
CACCACACAC CCCACACACA CACCACACAC ACACACCACA CACACACACA
CCCCACACAC ACACACACCC CCCCCCCACA CACACACACA CACACCA
```

Figura 24: *Maior correspondência encontrada no genoma humano*. Esta sequência compreende 447 nucleotídeos e está localizada no Cromossomo 2.

Fornecido pelo NCBI, em 06/08/2014, e disponível em: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF_000209065.1_ASM20906v1/GCF_000209065.1_ASM20906v1_genomic.fna.gz, o genoma do *T. cruzi* foi utilizado para comparar todas essas sequências de interesse em busca de correspondências. Deste modo, destacamos as 10 sequências mais encontradas no genoma humano, bem como 10 sequências com mais de 20 nucleotídeos mais frequentes e todas as sequências com mais de 100 nucleotídeos e comparamos ao genoma do *T. cruzi* (Tabela 3).

Tabela 3: Correspondências mais encontradas no genoma humano e sua frequência no genoma de *T. cruzi*.

Sequência	**Freq (*Homo sapiens*)**	**Freq (*T. cruzi*)**
AAAAAA	991734	126201
CCCAAA	600131	25839
AAAACA	572210	74563
CAAAAA	526624	97027
ACAAAA	472389	86905
AAAAAAA	426999	72729
AAAAAC	419187	66207
AACAAA	397977	80328
AAACAC	372329	44339
CCAAAA	368615	37002

Através da plataforma do software BLAST, disponibilizado pelo NCBI, buscamos essas sequências de interesse, com mais de 20 nucleotídeos, utilizando a ferramenta BLASTN (*Nucleotide BLAST*) (versão 2.3.1), para buscas curtas e comparações entre espécies (ALTSCHUL *et al.*, 1997) (Tabela 4).

As correspondências no genoma de *T. cruzi* mostraram que a frequência das correspondências mais encontradas não seguem o mesmo padrão de frequência do genoma humano, porém, assim como no genoma humano, seguem o padrão de tamanho, ou seja, conforme a quantidade de nucleotídeos aumenta, menor a chance de se encontrar tal correspondência.

Nenhuma das sequências com mais de 100 nucleotídeos retornou similaridades através da ferramenta BLASTN.

Tabela 4: Correspondências com mais de 20 nucleotídeos mais encontradas no genoma humano e sua frequência no genoma de *T. cruzi*.

Sequência	Freq (*Homo sapiens*)	Freq (*T. cruzi*)
CAAAAAAAAAAAAAAAAAAA	9009	4959
CAAAAAAAAAAAAAAAAAAAA	7820	4789
CAAAAAAAAAAAAAAAAAAAAA	7083	4914
CAAAAAAAAAAAAAAAAAAAAAA	6470	4781
CAAAAAAAAAAAAAAAAAAAAAAA	5738	4933
CAAAAAAAAAAAAAAAAAAAAAAAA	4904	4612
CAAAAAAAAAAAAAAAAAAAAAAAAA	4007	4781
CAAAAAAAAAAAAAAAAAAAAAAAAAA	3100	3836
AAAAAAAAAAAAAAAAAAA	2650	5287
CAAAAAAAAAAAAAAAAAAAAAAAAAAA	2432	3888

A segunda parte do script de contagem é responsável por particionar o cromossomo em janelas cromossômicas contar as correspondências, utilizando como entrada os valores da posição inicial e final de cada banda citogenética do arquivo *“cytoBand.txt.gz”*.

O arquivo de saída exportado é similar ao arquivo de entrada contendo os valores das bandas e nos informa a quantidade total de correspondências naquela região (Figura 25).

A

1	0	4300000	p13	
2	4300000	9400000	p12	
3	9400000	13700000	p11.2	
4	13700000	15000000	p11.1	
5	15000000	17400000	q11.1	
6	17400000	21700000	q11.21	
7	21700000	23100000	q11.22	
8	23100000	25500000	q11.23	
9	25500000	29200000	q12.1	
10	29200000	31800000	q12.2	

B

1	1	0	4300000	0
2	2	4300000	9400000	0
3	3	9400000	13700000	24524
4	4	13700000	1.5e+07	9579
5	5	1.5e+07	17400000	28019
6	6	17400000	21700000	47063
7	7	21700000	23100000	16452
8	8	23100000	25500000	25691
9	9	25500000	29200000	41208
10	10	29200000	31800000	28905

Figura 25: *Exemplo do arquivo de entrada e segundo arquivo de saída (Output 2) do script de contagem.* **A:** Arquivo de entrada informando a posição inicial (coluna 1), posição final (coluna 2) e nome (coluna 3), das bandas cromossômicas do Cromossomo 22; **B:** Arquivo de saída informando o número de banda (coluna 1), posição inicial (coluna 2), posição final (coluna 3) e contagem (coluna 4), das correspondências em cada janela no mesmo cromossomo.

A análise dos dados obtidos pelo script de contagem por janela, nos permitiu criar um mapa genético (Anexo 5) com todas as assinaturas em cada banda cromossômica (Figura 26), nos permitindo identificar que a distribuição dos motivos CA's se dá de modo homogêneo, não indicando porções de maior ou menor propensão para introgressões de kDNA.

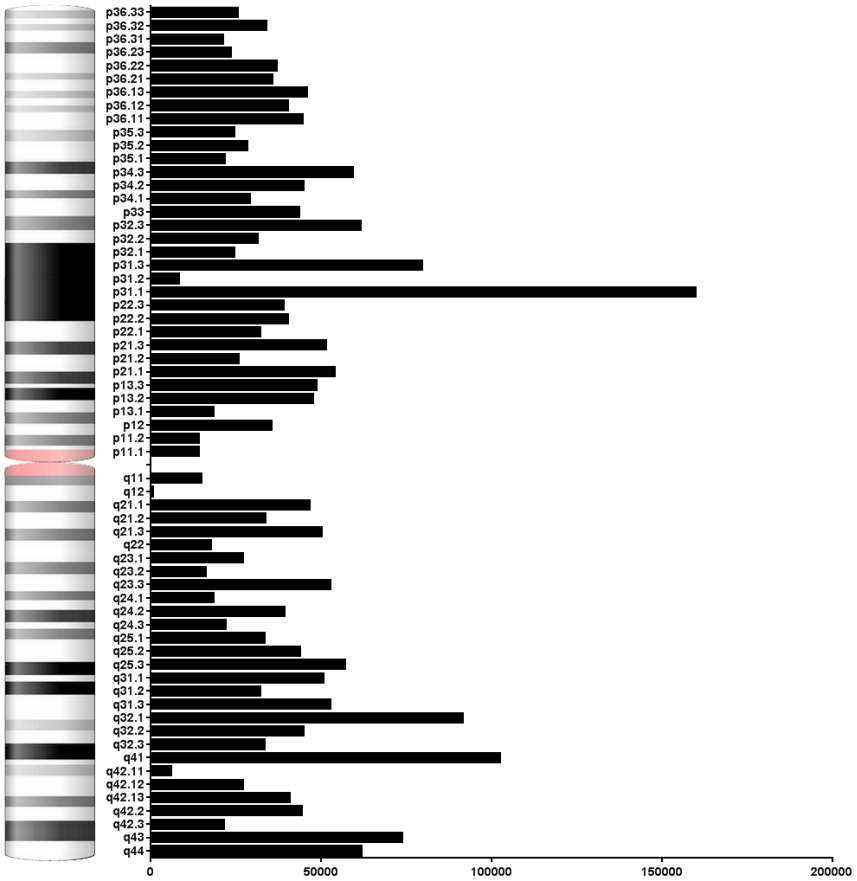


Figura 26: Exemplo do gráfico de frequência dos motivos CA's em cada banda citogenética do Cromossomo 1.

A quantidade total de correspondências está diretamente relacionada ao tamanho da banda, deste modo, não se pode inferir quais bandas são mais suscetíveis à integração do kDNA de *T. cruzi*. Porém percebemos que ao aproximar-se do centrômero a taxa de motivos CA's é baixa, quando não é nula, indicando pontos pouco prováveis de integração, o que também pode ser explicado pelas limitações tecnológicas do próprio sequenciamento, já que cerca de 1% do genoma

não está sequenciado devido à grande quantidade de repetições (NATIONAL HUMAN GENOME RESEARCH INSTITUTE, 2010).

Prioritariamente, a intenção era associar este mapa de correspondências com um mapa que nos indicasse os locais mais prováveis de elementos transponíveis, em especial retrotransposons do tipo LINE-1, visto que o kDNA exógeno possui uma predisposição para integrar-se nestes pontos (SIMÕES-BARBOSA *et al.*, 2006), todavia ainda não existe na literatura um mapa dos elementos transponíveis no genoma humano.

Para inferir sobre pontos mais suscetíveis à integração do kDNA de *T. cruzi,* será necessário a sobreposição deste mapa com um mapa dos *loci* codantes já descritos e um mapa de localização potencial de elementos transponíveis.

5. CONCLUSÃO

Foram desenvolvidos 3 scripts baseados na linguagem R: permutação de elementos (ácidos nucleicos ou aminoácidos); busca, agrupamento e plotagem das correspondências em genoma; contagem total de correspondências e contagem por janela cromossômica.

Os scripts foram eficientes, com rápida capacidade de resposta, mesmo em máquinas não muito robustas.

Os scripts criados podem sem facilmente ajustados para buscas orientadas, utilizando outros tipos de sequências base, com qualquer combinação de nucleotídeos desejada.

Todas as assinaturas dos motivos CA’s foram devidamente identificadas no genoma humano.

Um mapa genético foi desenvolvido, listando as correspondências em cada banda citogenética, em todos os cromossomos.

REFERÊNCIAS

ALTSCHUL, S. F.; MADDEN, T. L.; SCHAFFER, A. A.; ZHANG, J.; ZHANG, Z.; MILLER, W.; LIPMAN, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Res,** v. 25, n. 17, p. 3389-402, Sep 1 1997.

BABUSHOK, D. V.; OSTERTAG, E. M.; COURTNEY, C. E.; CHOI, J. M.; KAZAZIAN, H. H., JR. L1 integration in a transgenic mouse model. **Genome Res,** v. 16, n. 2, p. 240-50, Feb 2006.

BAKER, J. R. Speculations on the evolution of the family Trypanosomatidae Doflein, 1901. **Exp Parasitol,** v. 13, p. 219-33, Apr 1963.

BASILE, L.; JANSA, J. M.; CARLIER, Y.; SALAMANCA, D. D.; ANGHEBEN, A.; BARTOLONI, A. *et al.* Chagas disease in European countries: the challenge of a surveillance system. **Euro Surveill,** v. 16, n. 37, 2011.

BRENER, Z. Relações parasita-hospedeiro na Doença de Chagas: mecanismos de infecção e doença. **Ann Soc Belge Med Trop,** v. 65, n. 1, p. 9-13, 1985.

BROWN, J. R. Ancient horizontal gene transfer. **Nat Rev Genet,** v. 4, n. 2, p. 121-32, Feb 2003.

CAVALIER-SMITH, T. Eukaryote kingdoms: seven or nine? **Biosystems,** v. 14, n. 3-4, p. 461-81, 1981.

CAVALIER-SMITH, T. The excavate protozoan phyla Metamonada Grasse emend. (Anaeromonadea, Parabasalia, Carpediemonas, Eopharyngia) and Loukozoa emend. (Jakobea, Malawimonas): their evolutionary affinities and new higher taxa. **Int J Syst Evol Microbiol,** v. 53, n. Pt 6, p. 1741-58, Nov 2003.

CHAGAS, C. Nova tripanozomiaze humana: estudos sobre a morfolojia e o ciclo evolutivo do *Schizotrypanum cruzi* n. gen., n. sp., ajente etiolojico de nova entidade morbida do homem. **Memórias do Instituto Oswaldo Cruz,** v. 1, p. 159-218, 1909.

CLARK, K.; KARSCH-MIZRACHI, I.; LIPMAN, D. J.; OSTELL, J.; SAYERS, E. W. GenBank. **Nucleic Acids Res,** v. 44, n. D1, p. D67-72, Jan 4 2016.

CLAYTON, J. Chagas disease 101. **Nature,** v. 465, n. 7301, p. S4-5, Jun 24 2010.

CORNISH-BOWDEN, A. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. **Nucleic Acids Res,** v. 13, n. 9, p. 3021-30, May 10 1985.

COURA, J. R. Transmissão da infecção chagásica por via oral na história natural da doença de Chagas. **Rev Soc Bras Med Trop,** v. 39, n. IV, p. 113-117, 2006.

COURA, J. R. The main sceneries of Chagas disease transmission. The vectors, blood and oral transmissions--a comprehensive review. **Mem Inst Oswaldo Cruz,** v. 110, n. 3, p. 277-82, May 2015.

DUTRA, W. O.; ROCHA, M. O.; TEIXEIRA, M. M. The clinical immunology of human Chagas disease. **Trends Parasitol,** v. 21, n. 12, p. 581-7, Dec 2005.

ESPINDOLA, F. S.; CALÁBRIA, L. K.; REZENDE, A. A. A. D.; PEREIRA, B. B.; SANTANA, F. A.; AMARAL, I. M. R. *et al.* Recursos de bioinformática aplicados às ciências ômicas como genômica, transcriptômica, proteômica, interatômica e metabolômica. **Biosci. j. (Online),** v. 26, n. 3, p. 463-477, 06 2010.

FLEISCHMANN, R. D.; ADAMS, M. D.; WHITE, O.; CLAYTON, R. A.; KIRKNESS, E. F.; KERLAVAGE, A. R. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. **Science,** v. 269, n. 5223, p. 496-512, Jul 28 1995.

FREEMAN, V. J. Studies on the virulence of bacteriophage-infected strains of *Corynebacterium diphtheriae*. **J Bacteriol,** v. 61, n. 6, p. 675-88, Jun 1951.

GENOMES ONLINE DATABASE. Complete and Permanent Draft Genome Totals in GOLD. 2016. Disponível em: < https://gold.jgi.doe.gov/statistics >. Acesso em: 25/02/2016.

GIRONES, N.; FRESNO, M. Etiology of Chagas disease myocarditis: autoimmunity, parasite persistence, or both? **Trends Parasitol,** v. 19, n. 1, p. 19-22, Jan 2003.

GRUBY, D. Recherches et observations sur une nouvelle espèce d'hématozoaire *Trypanosoma sanguinis*. **Comptes Rendus Academie des Sciences**, n. 17, p. 1134-1136, 1843.

GUIMARO, M. C.; ALVES, R. M.; ROSE, E.; SOUSA, A. O.; DE CASSIA ROSA, A.; HECHT, M. M. *et al.* Inhibition of autoimmune Chagas-like heart disease by bone marrow transplantation. **PLoS Negl Trop Dis,** v. 8, n. 12, p. e3384, Dec 2014.

HECHT, M. M.; NITZ, N.; ARAUJO, P. F.; SOUSA, A. O.; ROSA ADE, C.; GOMES, D. A. *et al.* Inheritance of DNA transferred from American trypanosomes to human hosts. **PLoS One,** v. 5, n. 2, p. e9181, 2010.

HOLLEY, R. W.; APGAR, J.; EVERETT, G. A.; MADISON, J. T.; MARQUISEE, M.; MERRILL, S. H. *et al.* Structure of a Ribonucleic Acid. **Science,** v. 147, n. 3664, p. 1462-5, Mar 19 1965.

HONIGBERG, B. M. Evolutionary and systematic relationships in the flagellate order Trichomonadida Kirby. **J Protozool,** v. 10, p. 20-63, Feb 1963.

HUBER, W.; CAREY, V. J.; GENTLEMAN, R.; ANDERS, S.; CARLSON, M.; CARVALHO, B. S. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. **Nat Methods,** v. 12, n. 2, p. 115-21, Feb 2015.

HUTCHISON, C. A., 3RD. DNA sequencing: bench to bedside and beyond. **Nucleic Acids Res,** v. 35, n. 18, p. 6227-37, 2007.

KAZAZIAN, H. H., JR.; MORAN, J. V. The impact of L1 retrotransposons on the human genome. **Nat Genet,** v. 19, n. 1, p. 19-24, May 1998.

KEALEY, A.; SMITH, R. Neglected tropical diseases: infection, modeling, and control. **J Health Care Poor Underserved,** v. 21, n. 1, p. 53-69, Feb 2010.

KENT, W. J. BLAT--the BLAST-like alignment tool. **Genome Res,** v. 12, n. 4, p. 656-64, Apr 2002.

KLEIN, J.; TAKAHATA, N. **Where Do We Come From? The Molecular Evidence for Human Descent**. New York: Springer, 2002.

KURLAND, C. G.; CANBACK, B.; BERG, O. G. Horizontal gene transfer: a critical view. **Proc Natl Acad Sci U S A,** v. 100, n. 17, p. 9658-62, Aug 19 2003.

KUSAKABE, T.; SUGIMOTO, Y.; MAEDA, T.; NAKAJIMA, Y.; MIYANO, M.; NISHIKAWA, J. *et al.* Linearization and integration of DNA into cells preferentially occurs at intrinsically curved regions from human LINE-1 repetitive element. **Gene,** v. 274, n. 1-2, p. 271-81, Aug 22 2001.

LANDER, E. S.; LINTON, L. M.; BIRREN, B.; NUSBAUM, C.; ZODY, M. C.; BALDWIN, J. *et al.* Initial sequencing and analysis of the human genome. **Nature,** v. 409, n. 6822, p. 860-921, Feb 15 2001.

LANDWEBER, L. F.; GILBERT, W. Phylogenetic analysis of RNA editing: a primitive genetic phenomenon. **Proc Natl Acad Sci U S A,** v. 91, n. 3, p. 918-21, Feb 1 1994.

LAURIA-PIRES, L.; BRAGA, M. S.; VEXENAT, A. C.; NITZ, N.; SIMÕES-BARBOSA, A.; TINOCO, D. L.; TEIXEIRA, A. R. Progressive chronic Chagas heart disease ten years after treatment with anti-*Trypanosoma cruzi* nitroderivatives. **Am J Trop Med Hyg,** v. 63, n. 3-4, p. 111-8, Sep-Oct 2000.

LENT, H. Evolução dos conhecimentos sobre vetores da doença de Chagas 90 anos após sua descoberta. **Memórias do Instituto Oswaldo Cruz,** v. 94, p. 89-92, 1999.

LEON, J. S.; ENGMAN, D. M. The significance of autoimmunity in the pathogenesis of Chagas heart disease. **Front Biosci,** v. 8, p. e315-22, May 1 2003.

LEON, J. S.; GODSEL, L. M.; WANG, K.; ENGMAN, D. M. Cardiac myosin autoimmunity in acute Chagas' heart disease. **Infect Immun,** v. 69, n. 9, p. 5643-9, Sep 2001.

LESK, A. M. **Introduction to Bioinformatics**. Oxford University Press, 2014.

LIU, B.; LIU, Y.; MOTYKA, S. A.; AGBO, E. E.; ENGLUND, P. T. Fellowship of the rings: the replication of kinetoplast DNA. **Trends Parasitol,** v. 21, n. 8, p. 363-9, Aug 2005.

LUKES, J.; GUILBRIDE, D. L.; VOTYPKA, J.; ZIKOVA, A.; BENNE, R.; ENGLUND, P. T. Kinetoplast DNA network: evolution of an improbable structure. **Eukaryot Cell,** v. 1, n. 4, p. 495-502, Aug 2002.

MARDIS, E. R. The impact of next-generation sequencing technology on genetics. **Trends Genet,** v. 24, n. 3, p. 133-41, Mar 2008.

MARGULIS, L.; SAGAN, D. **Acquiring genomes. A theory of the origins of species**. International Ed. New York: Basic Books, 2003.

MINISTÉRIO DA SAÚDE - BRASIL, S. D. V. E. S. Doença de Chagas aguda no Brasil: série histórica de 2000 a 2013. **Boletim Epidemiológico,** v. 46, n. 21, p. 9, 2015.

MOREIRA, D.; LOPEZ-GARCIA, P.; VICKERMAN, K. An updated view of kinetoplastid phylogeny using environmental sequences and a closer outgroup: proposal for a new classification of the class Kinetoplastea. **Int J Syst Evol Microbiol,** v. 54, n. Pt 5, p. 1861-75, Sep 2004.

NATIONAL HUMAN GENOME RESEARCH INSTITUTE. The Human Genome Project Completion: Frequently Asked Questions. 2010. Disponível em: < https://www.genome.gov/11006943 >. Acesso em: 06/02/2016.

NATIONAL LIBRARY OF MEDICINE. NLM Congressional Justification FY 2017. 2016. Disponível em: < https://www.nlm.nih.gov/about/appropriations.html >. Acesso em: 28/02/2016.

NEVES, D. P.; MELO, A. L.; LINARDI, P. M. **Parasitologia humana**. 12ª ed. Rio de Janeiro: Atheneu, 2011. 546p.

NOCQ, J.; CELTON, M.; GENDRON, P.; LEMIEUX, S.; WILHELM, B. T. Harnessing virtual machines to simplify next-generation DNA sequencing analysis. **Bioinformatics,** v. 29, n. 17, p. 2075-83, Sep 1 2013.

OSTERTAG, E. M.; KAZAZIAN, H. H., JR. Biology of mammalian L1 retrotransposons. **Annu Rev Genet,** v. 35, p. 501-38, 2001.

PAGES, H. **BSgenome: Infrastructure for Biostrings-based genome data packages and support for efficient SNP representation**: R package version 1.38.0. 2016.

PAGES, H.; ABOYOUN, P.; GENTLEMAN, R.; DEBROY, S. **Biostrings: String objects representing biological sequences, and matching algorithms**: R package version 2.38.3. 2016.

PAVLICEK, A.; PACES, J.; ELLEDER, D.; HEJNAR, J. Processed pseudogenes of human endogenous retroviruses generated by LINEs: their integration, stability, and distribution. **Genome Res,** v. 12, n. 3, p. 391-9, Mar 2002.

PENNISI, E. Genomics. ENCODE project writes eulogy for junk DNA. **Science,** v. 337, n. 6099, p. 1159, 1161, Sep 7 2012.

PONTIUS, J. U.; WAGNER, L.; SCHULER, G. D. UniGene: A Unified View of the Transcriptome. In: (Ed.). **The NCBI Handbook**. Bethesda (MD): National Center for Biotechnology Information, 2003.

PRUITT, K. D.; TATUSOVA, T.; BROWN, G. R.; MAGLOTT, D. R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. **Nucleic Acids Res,** v. 40, n. Database issue, p. D130-5, Jan 2012.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**: R Foundation for Statistical Computing, Vienna, Austria. 2015.

REVOLUTION ANALYTICS; WESTON, S. **doParallel: Foreach Parallel Adaptor for the 'parallel' Package**: R package version 1.0.10. 2015.

REY, L. **Parasitologia**. 4ª ed. Rio de Janeiro: Guanabara Koogan, 2008. 930p.

SANGER, F.; AIR, G. M.; BARRELL, B. G.; BROWN, N. L.; COULSON, A. R.; FIDDES, C. A. *et al.* Nucleotide sequence of bacteriophage phi X174 DNA. **Nature,** v. 265, n. 5596, p. 687-95, Feb 24 1977.

SCIPIONI, A.; PISANO, S.; ANSELMI, C.; SAVINO, M.; DE SANTIS, P. Dual role of sequence-dependent DNA curvature in nucleosome stability: the critical test of highly bent Crithidia fasciculata DNA tract. **Biophys Chem,** v. 107, n. 1, p. 7-17, Jan 1 2004.

SHAPIRO, T. A.; KLEIN, V. A.; ENGLUND, P. T. Isolation of kinetoplast DNA. **Methods Mol Biol,** v. 94, p. 61-7, 1999.

SIMÕES-BARBOSA, A.; ARGANARAZ, E. R.; BARROS, A. M.; ROSA ADE, C.; ALVES, N. P.; LOUVANDINI, P. *et al.* Hitchhiking *Trypanosoma cruzi* minicircle DNA affects gene expression in human host cells via LINE-1 retrotransposon. **Mem Inst Oswaldo Cruz,** v. 101, n. 8, p. 833-43, Dec 2006.

SIMPSON, A. G.; LUKES, J.; ROGER, A. J. The evolutionary history of kinetoplastids and their kinetoplasts. **Mol Biol Evol,** v. 19, n. 12, p. 2071-83, Dec 2002.

SIMPSON, L. The mitochondrial genome of kinetoplastid protozoa: genomic organization, transcription, replication, and evolution. **Annu Rev Microbiol,** v. 41, p. 363-82, 1987.

STUART, J. M.; SEGAL, E.; KOLLER, D.; KIM, S. K. A gene-coexpression network for global discovery of conserved genetic modules. **Science,** v. 302, n. 5643, p. 249-55, Oct 10 2003.

TEIXEIRA, A. R. **Doenças de Chagas e Evolução**. 1ª ed. Brasília: FINATEC/UNB, 2007.

TEIXEIRA, A. R.; ARGANARAZ, E. R.; FREITAS, L. H., JR.; LACAVA, Z. G.; SANTANA, J. M.; LUNA, H. Possible integration of *Trypanosoma cruzi* kDNA minicircles into the host cell genome by infection. **Mutat Res,** v. 305, n. 2, p. 197-209, Mar 1 1994.

TEIXEIRA, A. R.; GOMES, C.; NITZ, N.; SOUSA, A. O.; ALVES, R. M.; GUIMARO, M. C. *et al. Trypanosoma cruzi* in the chicken model: Chagas-like heart disease in the absence of parasitism. **PLoS Negl Trop Dis,** v. 5, n. 3, p. e1000, 2011.

TEIXEIRA, A. R.; LACAVA, Z.; SANTANA, J. M.; LUNA, H. Inserção de DNA de *Trypanosoma cruzi* no genoma de célula hospedeira de mamífero por meio de infecção. **Revista da Sociedade Brasileira de Medicina Tropical,** v. 24, p. 55-58, 1991.

TEIXEIRA, A. R.; TEIXEIRA, G.; MACEDO, V.; PRATA, A. Acquired cell-mediated immunodepression in acute Chagas' disease. **J Clin Invest,** v. 62, n. 6, p. 1132-41, Dec 1978.

TENENBAUM, D.; BIOCORE TEAM. **BiocInstaller: Install/Update Bioconductor and CRAN Packages**: R package version 1.20.1. 2016.

THE BIOCONDUCTOR DEV TEAM. **BSgenome.Hsapiens.NCBI.GRCh38: Full genome sequences for Homo sapiens (GRCh38)**: R package version 1.3.1000. 2015.

THOMPSON, J. D.; HIGGINS, D. G.; GIBSON, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. **Nucleic Acids Res,** v. 22, n. 22, p. 4673-80, Nov 11 1994.

TOSO, A. M.; VIAL, F. U.; GALANTI, N. Transmisión de la enfermedad de Chagas por vía oral. **Revista médica de Chile,** v. 139, p. 258-266, 2011.

TRELOGAN, S. A.; MARTIN, S. L. Tightly regulated, developmentally specific expression of the first open reading frame from LINE-1 during mouse embryogenesis. **Proc Natl Acad Sci U S A,** v. 92, n. 5, p. 1520-4, Feb 28 1995.

TYLER, K. M.; ENGMAN, D. M. The life cycle of *Trypanosoma cruzi* revisited. **Int J Parasitol,** v. 31, n. 5-6, p. 472-81, May 1 2001.

VENTER, J. C.; ADAMS, M. D.; MYERS, E. W.; LI, P. W.; MURAL, R. J.; SUTTON, G. G. *et al.* The sequence of the human genome. **Science,** v. 291, n. 5507, p. 1304-51, Feb 16 2001.

WARNES, G. R.; BOLKER, B.; LUMLEY, T. **gtools: Various R Programming Tools**: R package version 3.5.0. 2015.

WESTENBERGER, S. J.; CERQUEIRA, G. C.; EL-SAYED, N. M.; ZINGALES, B.; CAMPBELL, D. A.; STURM, N. R. *Trypanosoma cruzi* mitochondrial maxicircles display species- and strain-specific variation and a conserved element in the non-coding region. **BMC Genomics,** v. 7, p. 60, 2006.

WETTERSTRAND, K. A. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). 2016. Disponível em: < www.genome.gov/sequencingcosts >. Acesso em: 20/01/2016.

WHITTAKER, R. H.; MARGULIS, L. Protist classification and the kingdoms of organisms. **Biosystems,** v. 10, n. 1-2, p. 3-18, Apr 1978.

WORLD HEALTH ORGANIZATION. Control of Chagas disease: Second report of the WHO Expert Committee. **World Health Organ Tech Rep Ser**, n. 905, p. 1-109, 2002.

WORLD HEALTH ORGANIZATION. Research priorities for Chagas disease, human African trypanosomiasis and leishmaniasis. **World Health Organ Tech Rep Ser**, n. 975, p. v-xii, 1-100, 2012.

WORLD HEALTH ORGANIZATION. Investing to overcome the global impact of neglected tropical diseases. **Third WHO report on neglected tropical diseases**, p. 191, 2015.

YOSHIDA, K.; NAKAMURA, A.; YAZAKI, M.; IKEDA, S.; TAKEDA, S. Insertional mutation by transposable element, L1, in the DMD gene results in X-linked dilated cardiomyopathy. **Hum Mol Genet,** v. 7, n. 7, p. 1129-32, Jul 1998.

ZHOU, H.; RANISH, J. A.; WATTS, J. D.; AEBERSOLD, R. Quantitative proteome analysis by solid-phase isotope tagging and mass spectrometry. **Nat Biotechnol,** v. 20, n. 5, p. 512-5, May 2002.

ZHU, J.; SHI, Z.; WANG, J.; ZHANG, B. Empowering biologists with multi-omics data: colorectal cancer as a paradigm. **Bioinformatics,** v. 31, n. 9, p. 1436-43, May 1 2015.

ZINGLER, N.; WILLHOEFT, U.; BROSE, H. P.; SCHODER, V.; JAHNS, T.; HANSCHMANN, K. M. *et al.* Analysis of 5' junctions of human LINE-1 and Alu retrotransposons suggests an alternative model for 5'-end attachment requiring microhomology-mediated end-joining. **Genome Res,** v. 15, n. 6, p. 780-9, Jun 2005.

ANEXO 1

Script 01 – Permutação

```
# Escolha da função a ser realizada.
Options=readline("Options:\n [1] Permute Nucleic Acids Codes\n [2] Permute Amino Acids
Codes\n [3] Exit\n->")

Selection=paste("t", Options, sep="")
switch(Selection, t1={Construct=1},  #Permutar caracteres de Ácidos nucleicos
                  t2={Construct=2},  #Permutar caracteres de Aminoácidos
                  t3={Construct=3},  #Sair
                  stop("INVALID CHOICE, PLEASE RESTART!"))

# Paralelização da rotina.
library(doParallel)
cores = detectCores(logical = TRUE)
cl = makeCluster(cores)
registerDoParallel(cl)

if(Construct==1){

        String_composition = readline("Define the elements to be permuted (Only letters!):\n
        [G] Guanine                                    [K] Ketone
        [A] Adenine                                    [S] Strong interaction
        [T] Thymine                                    [W] Weak interaction
        [C] Cytosine                                   [H] not-G
        [U] Uracil                                     [B] not-A
        [R] puRine                                     [V] not-T (not-U)
        [Y] pYrimidine                                 [D] not-C
        [M] aMino                                      [N] aNy\n->")

}

if(Construct==2){

        String_composition = readline("Define the elements to be permuted (Only letters!):\n
        [G] Glycine                                    [R] Arginine
        [A] Alanine                                    [W] Tryptophan
        [V] Valine                                     [S] Serine
        [L] Leucine                                    [T] Threonine
        [I] Isoleucine                                 [D] Aspartic acid
        [P] Proline                                    [E] Glutamic acid
        [F] Phenylalanine                              [N] Asparagine
        [Y] Tyrosine                                   [Q] Glutamine
        [C] Cysteine                                   [B] Aspartic acid or asparagine
        [M] Methionine                                 [Z] Glutamic acid or glutamine
        [H] Histidine                                  [O] Terminator
        [K] Lysine                                     [X] Unknown\n->")

}
```

```
if(Construct==1 | Construct==2){

# Seleção do tamanho da sequência.
  String_length = readline("Define the length of the string->")
  r = as.integer(String_length)

# Aloca a matriz Vector e define o nome do arquivo de saída.
  Nucleo = strsplit(String_composition, "")[[1]]
  n = nchar(String_composition)

  library(gtools)
  String_permut = permutations(n = n, r = r, repeats.allowed = TRUE,  Nucleo)
  Vector = gsub(" ","", matrix(do.call(paste, as.data.frame(String_permut))))

  Namefile = paste("OUTPUT_Permutation_", String_composition, "_", r, ".csv", sep = "")

# Exporta os dados da permutação para o arquivo de saída.
  write.table(Vector, file = Namefile,
                append = FALSE, quote = FALSE, dec = ",",
                sep = "\t", row.names = FALSE, col.names = FALSE)

}

if(Construct==3){

  print("Successfully Exit")#Stop

}
```

ANEXO 2

Script 02 – Busca, Agrupamento e Plotagem

```
# Escolha da função a ser realizada.
Options=readline("Options:\n [1] Search, Group and Plot\n [2] Only Search\n [3] Only
Group\n [4] Only Plot\n [5] Exit\n->")

Selection=paste("t", Options, sep="")
switch(Selection, t1={Construct=1},  #Busca, Agrupamento e Plotagem de Sequências
                  t2={Construct=2},  #Somente Busca no Genoma
                  t3={Construct=3},  #Somente Agrupamento de Regiões Justapostas
                  t4={Construct=4},  #Somente Plotagem do mapa de Correspondências
                  t5={Construct=5},  #Sair
                  stop("INVALID CHOICE, PLEASE RESTART!"))

# Paralelização da rotina.
library(doParallel)
cores = detectCores(logical = TRUE)
cl = makeCluster(cores)
registerDoParallel(cl)

if(Construct==1 | Construct==2){

        # Informar a biblioteca de genoma a ser buscada e qual cromossomo será analisado.
        require (BSgenome.Hsapiens.NCBI.GRCh38)
        Genome = BSgenome.Hsapiens.NCBI.GRCh38
        Chr_search = readline("Define the chromosome to be searched->")

        # Indicar o caminho do arquivo de entrada com as strings a serem buscadas.
        INPUT_String = read.table("OUTPUT_Permutation_AC_6.csv", quote="\"")
        String_matrix = as.matrix (INPUT_String)
        String_objDNA = DNAStringSet(String_matrix, use.names=TRUE)

        # Tranforma a string em objeto PDict.
        String_objPDict = PDict(String_objDNA,
                                max.mismatch=NA, tb.start=NA, tb.end=NA,
                                tb.width=NA, algorithm="ACtree2",
                                skip.invalid.patterns=FALSE)

        # Busca as correspondências das strings na biblioteca do genoma.
        Matches = matchPDict(String_objPDict,
                             Genome[[Chr_search]], max.mismatch=0, min.mismatch=0,
                             with.indels=FALSE, fixed=TRUE,
                             algorithm="auto", verbose=FALSE)

        # Transforma a matriz em data frame.
        String_matches = as.data.frame(Matches)
        Endtable = nrow(String_matches)
        String_matches[1:Endtable,]$group_name=toupper(c(String_matrix[String_matches[1
        :Endtable,]$group]))
```

```
	# Reordena as colunas.
	group = as.numeric(String_matches[,1])
	group_name = as.character(String_matches[,2])
	p_start = as.numeric(String_matches[,3])
	p_end = as.numeric(String_matches[,4])
	width = as.numeric(String_matches[,5])
	Matches_sort = data.frame(cbind(group, p_start, p_end, width, group_name))

	# Informar o nome do arquivo de saída para a exportação das correspondências.
	Namefile = paste("OUTPUT_Search_Chr", Chr_search, ".txt", sep = "")
	write.table(Matches_sort, file = Namefile,
			append = FALSE, quote = FALSE,
			sep = "\t", row.names = FALSE, col.names = TRUE)

}

if(Construct==1 | Construct==3) {

	# Indicar o caminho do arquivo de entrada com as sequências a serem agrupadas.
	INPUT_Matches = read.table("OUTPUT_Search_Chr22.txt", quote="\"")
	Matches_matrix = as.matrix(INPUT_Matches)

	# Define qual cromossmo será agrupado.
	if(Construct==3) {
		Chr_search = readline("Define the chromosome to be grouped->")
	}

	# Recorte das colunas a serem analisadas. Importante: verificar se o arquivo de
entrada possui cabeçalho, caso não possua, retirar o [-1].
	group_name = as.character(Matches_matrix[-1,5])
	p_start = as.numeric(Matches_matrix[-1,2])
	p_end = as.numeric(Matches_matrix[-1,3])

	# Reordenamento pela posição inicial da correspondência.
	EndReg = data.frame(cbind(p_start, p_end, group_name))
	Order_sort = order(p_start, decreasing=FALSE)
	EndReg_sort = EndReg[Order_sort,]

	# Cria uma matriz molde para exportar os dados a seguir.
	Endmatrix = nrow(Matches_matrix)-1
	Endmatrix_100 = 100/Endmatrix
	MatrixReg = matrix(nrow=Endmatrix, ncol=5, byrow = TRUE)
	Reg=1

	# Determina quais colunas correspondem ao início e fim da correspondência.
	Start_consensus = as.numeric(as.character(EndReg_sort$p_start[1]))
	End_consensus = as.numeric(as.character(EndReg_sort$p_end[1]))
	Grouped_strings = as.character(EndReg_sort$group_name[1])
	Grouping_start = TRUE
```

```
# Processo de agrupamento das sequências justapostas.
        for (i in 2:Endmatrix){
                p_start_sort=as.numeric(as.character(EndReg_sort$p_start[i]))
                p_end_sort=as.numeric(as.character(EndReg_sort$p_end[i-1]))

                if (p_start_sort<=p_end_sort){

                        End_consensus=as.numeric(as.character
                        (EndReg_sort$p_end[i]))

                                if(Grouping_start==TRUE){Grouped_strings =
                                as.character(EndReg_sort$group_name[i-1])}

                        Grouping_start=FALSE
                        Grouped_strings_end=substr(as.character(
                        EndReg_sort$group_name[i]), 6, 6)
                        Grouped_strings = paste(Grouped_strings,
                        Grouped_strings_end, sep = "")

                } else {
                         MatrixReg[Reg,1]=Reg
                         MatrixReg[Reg,2]=Start_consensus
                         MatrixReg[Reg,3]=End_consensus
                         MatrixReg[Reg,4]=End_consensus-Start_consensus+1
                         MatrixReg[Reg,5]=Grouped_strings

                         #Informa na tela a região e porcentagem da tarefa.
                         Reg=Reg+1
                         cat(c("Regiao:",Reg,"\n"))
                         cat(c("Tarefa executada:",round(
                         i*Endmatrix_100,digits=2),"%\n"))

                         Start_consensus=as.numeric(as.character(
                         EndReg_sort$p_start[i]))
                         End_consensus=as.numeric(as.character(
                         EndReg_sort$p_end[i]))
                         Grouped_strings=as.character(EndReg_sort$group_name[i])
                         Grouping_start=TRUE

                }

        }

# Cria uma matriz contendo os dados obtidos no agrupamento.
Grouped_matches = MatrixReg[c(-(Reg:(Endmatrix))),]
colnames(Grouped_matches) = c("seq", "p_start", "p_end", "width", "group_name")

# Arquivo de saída para a exportação das sequências agrupadas.
Namefile = paste("OUTPUT_Grouping_Chr", Chr_search, ".txt", sep = "")
write.table(Grouped_matches, file = Namefile,
            append = FALSE, quote = FALSE,
            sep = "\t",row.names=FALSE, col.names=TRUE)
}
```

```
if(Construct==1 | Construct==4) {

	# Indicar o caminho dos arquivos de entrada com as sequências.
	INPUT_Sequences = read.table("OUTPUT_Grouping_Chr22.txt", quote="\"")
	Sequences_matrix = as.matrix(INPUT_Sequences)

	# Define qual cromossmo será plotado.
	if(Construct==4) {
		require (BSgenome.Hsapiens.NCBI.GRCh38)
		Genome = BSgenome.Hsapiens.NCBI.GRCh38
		Chr_search = readline("Define the chromosome to be plotted->")
	}

	# Recorte das colunas a serem analisadas. Importante: verificar se o arquivo de
entrada possui cabeçalho, caso não possua, retirar o [-1].
	p_start = as.numeric(Sequences_matrix[-1,2])
	p_end = as.numeric(Sequences_matrix[-1,3])

	# Cria uma matriz molde para exportar os dados a seguir.
	Endmatrix = nrow(Sequences_matrix)-1

	# Cria um novo dispositivo gráfico.
	library(graphics)
	dev.new(width=10, height=5)

	# Define o tamanho da janela em cada arquivo png (Por padrão Cromossomo/50).
	Chr_size = nchar(Genome[[Chr_search]])
	Chr_window = Chr_size/50

	# Processo de criação dos objetos gráficos.
	arq = 1

	for (i in seq(0, Chr_size, by = Chr_window)){
		plot.new()
		bordai=i
		bordaf=i+Chr_window
		plot(c(bordai,bordaf), c(0,180), main=paste(
		"Chromosome ",Chr_search,sep=""),xlab="bp",ylab="")
		rect(bordai,10,bordaf,170, border="black")
		cat(c("Janela:",i/Chr_window,"\n"))

		for(r in 1:Endmatrix){
			if((p_start[r]>bordai) && (p_start[r]<=bordaf)){
				rect(p_start[r],0,p_end[r],180, border="red",
				col=p_end[r]-p_start[r]+1)
			}
			if(p_start[r]>bordaf) {
				cat(c("pre-break","\n"))
				break
			}

		}
```

```
                # Copia o conteúdo gráfico para o dispositivo criado e exporta para png.
                dev.print(png,paste("Window_plot ",arq,".png",sep=""),
                width=13.8,height=6, units="in", res=300)
                arq = arq+1

        }

        # Desliga o dispositivo gráfico.
        dev.off()

}

if(Construct==5) {

print("Successfully Exit")#Stop

}
```

ANEXO 3

Script 03 – Contagem

```
# Escolha da função a ser realizada.
Options=readline("Options:\n [1] Count Total Matches by Chromosome\n [2] Count Matches
by Chromosomal Set Windows\n [3] Exit\n->")

Selection=paste("t", Options, sep="")
switch(Selection, t1={Construct=1},  #Contagem total de correspondências
                  t2={Construct=2},  #Contagem em janelas cromossomais
                  t3={Construct=3},  #Sair
                  stop("INVALID CHOICE, PLEASE RESTART!"))

# Paralelização da rotina.
library(doParallel)
cores = detectCores(logical = TRUE)
cl = makeCluster(cores)
registerDoParallel(cl)

# Define de qual cromossmo será feita a contagem.
Chr_search = readline("Define the chromosome to be grouped->")

if(Construct==1) {

        # Indicar o caminho do arquivo de entrada com as sequências a serem contadas.
        INPUT_Sequences = read.table("OUTPUT_Grouping_Chr22.txt", quote="\"")
        Sequences_matrix = as.matrix(INPUT_Sequences)

        # Recorte da coluna a ser analisada e posterior contagem. Importante: verificar se o
arquivo de entrada possui cabeçalho, caso não possua, retirar o [-1].
        Sequences_name = as.character(Sequences_matrix[-1,5])
        Sequences_frame = data.frame(cbind(Sequences_name))
        Search_sequences = table(Sequences_frame)
        Search_matrix = as.data.frame(Search_sequences)

        # Reordenamento dos dados pela frequência de cada sequência.
        Sequences_freq = as.numeric(Search_matrix[,2])
        Sequences_group = as.character(Search_matrix[,1])
        Sequences_sort = data.frame(cbind(Sequences_group, Sequences_freq))
        Sort_order = order(Sequences_freq, decreasing=TRUE)

        # Cria uma matriz contendo os dados obtidos na contagem.
        Sequences_count = Sequences_sort[Sort_order,]
        colnames(Sequences_count) = c("group_name", "freq")

        # Informar o nome do arquivo de saída para a exportação das sequências contadas.
        Namefile = paste("OUTPUT_Total_Count_Chr", Chr_search, ".txt", sep = "")
        write.table(Sequences_count, file = Namefile,
                    append = FALSE, quote = FALSE,
                    sep = "\t", row.names = FALSE, col.names = TRUE)
 }
```

```
if(Construct==2){

	# Indicar o caminho dos arquivos com as sequências e tamanho das janelas.
	INPUT_Sequences = read.table("OUTPUT_Grouping_Chr22.txt", quote="\"")
	Sequences_matrix = as.matrix(INPUT_Sequences)

	INPUT_Window = read.table("INPUT_cytoBand_Chr22.txt", quote="\"")
	Window_matrix = as.matrix(INPUT_Window[,-3])

	# Recorte das colunas a serem analisadas. Importante: verificar se o arquivo de
entrada possui cabeçalho, caso não possua, retirar o [-1].
	p_start = as.numeric(Sequences_matrix[-1,2])
	p_end = as.numeric(Sequences_matrix[-1,3])

	# Cria uma matriz molde para exportar os dados a seguir.
	Endmatrix = nrow(Sequences_matrix)-1
	Endmatrix_100 = 100/Endmatrix
	Endwindow = nrow(Window_matrix)

	# Processo de contagem das sequências.
	Window_count = matrix(nrow=Endwindow+1, ncol=4, byrow = TRUE)
	m=1
	for(j in 1:Endwindow){
		n=1

		for(i in 1:Endmatrix){
			if((p_start[i]>Window_matrix[j,1]) && (p_start[i]<=Window_matrix[j,2])){
				start_in = p_start[i]
				end_in = p_end[i]
				n=n+1

				#Informa na tela a região e porcentagem da tarefa.
				cat(c("Sequencia:",n,"\n"))
				cat(c("Tarefa	executada:",round(
				i*Endmatrix_100,digits=2),"%\n"))
			}

			if(p_start[i]>Window_matrix[j,2]) {
				cat(c("pre-break","\n"))
				break
			}
		}

		Window_count[m,1]=j
		Window_count[m,2]=Window_matrix[j,1]
		Window_count[m,3]=Window_matrix[j,2]
		Window_count[m,4]=n-1
		m=m+1
		cat(c("RegiaoCromossomal:",m,"\n"))
	}

	Total_count = sum(Window_count [,4], na.rm = TRUE)
	Window_count [Endwindow+1,4]= Total_count
```

```
        # Informar o arquivo de saída para a exportação da contagem das sequências.
        Namefile = paste("OUTPUT_Window_Count_Chr", Chr_search, ".txt", sep = "")
        write.table(Window_count, file = Namefile,
                    append = FALSE, quote = FALSE,
                    sep = "\t", row.names = FALSE, col.names = FALSE)

}

## SAIR ##
if(Construct==3) {

print("Successfully Exit")#Stop

}
```

ANEXO 4

Longas Correspondências

Cromossomo 2 – 447 nucleotídeos;

ACACACACACACACCCCCCACACACACACACCCCCCACACACACCACACACACACCCCACACACACAA
CCACACCCCACACACACAACCACACACACACCACACACACACCCCACACACACCACACACACACCACA
CACCCCACACACACACCCCACACACACCACACACCACACACACACCCCACACACAACCACACACCACA
CACCACACACACACCACACACCACACCACACACACACCACACCACACACACACCACACACACACACCA
CACCACACACACCACACACACACCACACAACACCCCCCCACACACACACCACACACACACACCACACAC
ACCACACACACACCACACACCCCACACACACACCACACACACACACCACACACACACACACCCCCACAC
ACACACACACCCCCCCCCCCACACACACACACACACACCA

Cromossomo 21 – 268 nucleotídeos;

CACCACCACCACCACCACCACCACCACCACCACCACCACCACCACCACCACCACCACCACCACCACCA
CCACCACCACCACCACCACAACCACCACCACCACCACCACCACCACCACCACCACCACCACCACCACC
ACCACCACCACCACCACCACCACCACCACCACCACCACCACCACCACCACCACCACCACCACCACCAC
CACCACCACCACCAACACCACCACCACCACCACCACCACCACCACAACAACAACACCACAC

Cromossomo 2 – 173 nucleotídeos;

CCCACACACCACACCCACACACCACACCCACACACACCCACACCCACACCCACACACACCACACCCAC
ACACCACACCCACACCCACACACCACACCCACACCACACCCACACACCACACACCACACCCACACCCA
CACACACCACACCCACACACCACACCCACACACACCC

Cromossomo 8 – 172 nucleotídeos;

CCCAACACCACCCCCACACCACCCCCAACACCACCCCCACACCACCCCAACACCACCCCCAACACCAC
CCCCACACCACCCCAACACCACCCCCAACACCACCCCCAACACCACCCCCAACACCACCCCCACACCA
CCCCCACACCACCCCCACCACCACCCCAACACCACC

Cromossomo 1 – 170 nucleotídeos;

ACACACACCACACACACACACACACACACACACACACACACACACACACCACACACACACACACACAC
ACACACACACACACACCACACACACACACACACACACACACACACACACACCACACACACACACCACA
CACACACACACACACACACACACACACACACC

Cromossomo 5 – 168 nucleotídeos;

ACACACACACACCACACACACACCAACACACACCACACACACCACCAACACACACACCCCACACACCA
ACACACACACCCCACACACCAACACACACACCCCACACACCACACACACCAACACACAACCACACACA
CCACCAACACAACCACCAACACACACCACACA

Cromossomo 10 – 166 nucleotídeos;

ACACCACACACCCCACACACACCACACACACCACACACCACACCACACACCCCACACACACCACACAC
CCCACACACACCCCCACACACACCCCCACACACACCACACACACACACCACACACCACACACCCCACACA
CACACCACACACACCACACACACACCCCAC

Cromossomo X – 166 nucleotídeos;

CCCCCACACACACACCCCACACACCACACACACACACACCACACACCACACACACACCCCACACACCC
CACACCCCACACACACCCCCCACACACCACACCACACACACACACACCCACACCCCACACACACACACC
ACACCACACACACCACACCACACACACACA

Cromossomo 2 – 165 nucleotídeos;

CACACACACCACCACACACACCACACACACCACACACACACCACACACACCACCACACACACCACACA
CACCACACACACCACACACACCACCACACACCACACACACCACCACACACACCACACACACAACACAC
ACCACCACACACCACACACACCACACACA

Cromossomo 3 – 158 nucleotídeos;

ACAACACACAACCACACACCACACACACAACCACACAACCAACCACACACCACACACACCCACACACA
CCAACACAACACACAACCACACACCACACACACAACCACACAACCACACACCACACACAACCACACAC
ACAACCACACACACCACACACA

Cromossomo 3 – 154 nucleotídeos;

ACACACACCACACACAACCACACAACCCCACACACAACCACACACAACCACACAACCACACACCACAC
ACCACACACAACCACACAACACACAAACACACAACACACAACCACACACAACACACACCACACACAAC
CACACACCACACACAACA

Cromossomo 10 – 150 nucleotídeos;

ACACACCACACACCCCACACACCCCACACACACCACACACACCCACCACACACACACCCCACACACAC
ACCACACACCAACCACACACCACACACACCACACCACACACACACACCACACACACCACACACCACAC
ACACCACACACACA

Cromossomo 11 – 146 nucleotídeos;

ACACACCCACACACACCCCACACACACACACCCACACACACCCACACACACACCACACACACACAACACA
CACCCCCCACACACAAACACCACACACACACCACACACAAACACCACACACACACACCACACACAAAC
ACACACCACA

Cromossomo 7 – 146 nucleotídeos;

CACACACCCCACACCCCACACACACCCCACACACACACCACACACACCCCCCCCACACACACACCCCAC
ACACACCCCACACACCACACACACACACCCACCACACACACCCCACACCCCCACCACACACACCCCAC
ACACACCACC

Cromossomo 3 – 142 nucleotídeos;

ACCACACAACCACACCACACACCACACACACAACCACACAACCACACACCACACACCACACACAACCA
CACACACCACAACCACACACACAACACACACAACCACACACCACACAACCACACCCACCACAACCACA
CACACC

Cromossomo 9 – 141 nucleotídeos;

AACACACACACACACACACACACACAAACACACACACACACACACAAACACACACACACACACACAAACA
CACACACAAACACACACACACACACAAACACACACACACACACACACACACACACAAACACACACACA
CACAC

Cromossomo 1 – 140 nucleotídeos;

CACACACACAAACACACACACACACAAACACACACACACACACACACACACACACACACACAAACACACA
CACACACAAACACACACAAACACACACAAACACACACAAACACACACAAACACACACAAACACACACA
AACA

Cromossomo 5 – 140 nucleotídeos;

CACACACACACCACACACACACACCACACACACACACAAACCACACACCACACACACACCACACACAC
ACCACACACACCACACACCCCACACACACCAACCACACACACCCCCCACACACACACCACACACACCA
CACA

Cromossomo 16 – 139 nucleotídeos;

CCCACACCCCCACACCCACACACACACACCCCCACACACACCCCCACACCCACACACCCACACACCAC
ACACACACACCCCCACACCCACCCCACACACCCACACACCCACACACCACACACACACCCACACACAC
CCA

Cromossomo 4 – 138 nucleotídeos;

ACACACCCACACCCACACACACACCCACAAACACACACACACACCCCCACACCCACACACACACCCAC
AAACACACACACCCACACACACACACAAACACACACACACACACACACACCCACACACACACCCACAC
AC

Cromossomo X – 137 nucleotídeos;

CACACACCACACACACCACACACCACACACACACCACACCACACACCACACACACCACACACACACAC
CACACACCACACACACACCACACACCACACACACCACACCAAACACACACACCACACACACACAACAC
A

Cromossomo 5 – 122 nucleotídeos;

ACACACACACACCCACACACACACACACCCCCACACCCACCCACACCCACCCACACCCACACACCCAC
ACACACACACACCCCCACACACACCCACACCCACACACCCACACACACCCCACAC

Cromossomo 21 – 121 nucleotídeos;

CCACACACACCACACCACAAACCACACACCACACACCCACACACCACACACACCACACCACAAACCAC
ACACCACACACCCACACACCACACACACCACACCACAAACCACACACCACACA

Cromossomo X – 120 nucleotídeos;

CACACACACCACACACACACACCACACCACACACACACCACACACCACACACCACACACACACCAC
ACACCACACACACACCACCACACACACACACCCCACACACCACACACACACA

Cromossomo 17 – 119 nucleotídeos;

CCCCCAACACACACACACACACACACACACACACACACACCCCAACACACACACACACACCCCAACACAC
ACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACA

Cromossomo 10 – 119 nucleotídeos;

CCCCCCACACCCACACACACACCCACACACCCCCCCACCCCCCCCCACACACACACCCCCCCCACACACC
CCCCCACACACACCCCCCCCACACACACCCCCCCCCCCCCCCCCACACACACA

Cromossomo 14 – 118 nucleotídeos;

ACACACACACACACACACACACACACACACACAACACACACACACACAACACACACACACACAACACACA
CACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACAAA

Cromossomo X – 116 nucleotídeos;

ACACCACACACACCACACACCACACCACACACACACCACACACCACACACACACCACACCACACACCA
CACACACACACCCCACACACCACACACACCCCACACACACCCCCCAAAC

Cromossomo 2 – 115 nucleotídeos;

ACCACACACCCCCACACACCCCACACACCCCCACACACCCCACACACCCCACACACACCACACACCCC
ACACACACACCACACACCCCACACACCCCACACACCCCACACACACA

Cromossomo 2 – 115 nucleotídeos;

CACACCCAACACACACACCCAACACACACCCCCAACACACACACACCCAACACACACACACCCAACAC
ACACCCCCAACACACACACCCCAACACACACACACACCCAACACACA

Cromossomo 3 – 111 nucleotídeos;

ACCCACACACCACACCACACACCCACACACCCCACACCCACACACCCACACACCACACACCCCACACA
ACCACACACACCACACACACACCCCACACACACCCCACACCCACA

Cromossomo 5 – 111 nucleotídeos;

CAAACACCACACCACACACACCCCACACACACACCCCCACCACACCACACACACACCACACACCACAC
CACACACACACCACACACACACCCCACACACACCACACACACAC

Cromossomo 2 – 110 nucleotídeos;

AACACACACCCAACACACACACCCCCAAACACACCCCCAACACACACACAAACACACCCAACACACACA
CACCCAACACACACCCCCCAACACACACACCCCCACCACACAAC

Cromossomo 7 – 110 nucleotídeos;

CAAAAAAAAAAAAAAACAACAAAAAACCAAAACACACACACACACACAAAACACACAAAACCACACACA
CACACACACACACACACACACACACACACACACAACCAAAAAAC

Cromossomo 16 – 109 nucleotídeos;

ACACACACCCCACACACACCCCCACACACACCCCACACACACCCCACACACCCCACACCCCCACACCC
CCCCACACACCCCCCACACACCACACACACACCACACACAC

Cromossomo 16 – 109 nucleotídeos;

CACACACACACCACACACCCCCACACACACCCCACACACACACCACACACACACCACACACACCCCAC
ACCCCCCACACACCCCCCACACACACCCCCACACCCCCACACAC

Cromossomo 1 – 108 nucleotídeos;

ACACACACACACACACACACACACACACACACACACCACACACACACACACACACACACACACACACA
CACACACACACACACACACACACACACACACACACACACACA

Cromossomo 9 – 107 nucleotídeos;

CCCCCACCCCCCCAACACACACACACACACACACACACACACACACACACACACCAACACACACACACACA
CACCCCAACACACACACACACACACACACACACACACCA

Cromossomo 11 – 105 nucleotídeos;

ACACACCCCACACACACCACCACACCACACACACACACCACCACACACACACCCCACACCACACACAC
ACACCACACACACCACACACACACCACACACACCACA

Cromossomo X – 105 nucleotídeos;

CAAAACAAACAACAACAACAAAAAACACACAACCACCCCCACAAAACAAAACAAAACAAAACAACAAC
AACAACAAAAAAAAAAAAAAAAAAAAAAAACCAAA

Cromossomo 7 – 104 nucleotídeos;

ACACACACACACACACACACACACCACACCACACACACCCCACACACACACACCCCACACACAC
CACACACACACACCACACACACCCCACACACACACA

Cromossomo 3 – 103 nucleotídeos;

ACACACCACACAACCACACACACCACACACAACCACAAACACCACACACCACACACAACACACACAAC
CACACACACAACACAACACACAACCACACACAACA

Cromossomo 17 – 103 nucleotídeos;

CAAAAAAACAAAAAACAAAACACACACACACACACACACACACACACACACACACACACCAAAACCAAAA
CCAAAACCAAAACCAAAAAACAAACAAAAAACACC

Cromossomo 19 – 103 nucleotídeos;

CAAAACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACAC
ACACACACACACACACACACACACACACACACACA

Cromossomo 3 – 102 nucleotídeos;

CAACACACACACACCCACACCCACACACACCCACACCCACACACACCCACACCCACACACACCCACAC
CCACACCCACACACACCCACACCCACACACACAC

Cromossomo 21 – 101 nucleotídeos;

ACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACAC
ACACACACACACACACACACACACACACACCCC

Cromossomo 12 – 101 nucleotídeos;

CAAAAACAAAACAAAACAAAACCAAAAACCAAAAACCAACCAACCAAACAAAAAACCAAACACACACA
CACACACACACACACACACACACACACACAAAA

Cromossomo 3 – 100 nucleotídeos;

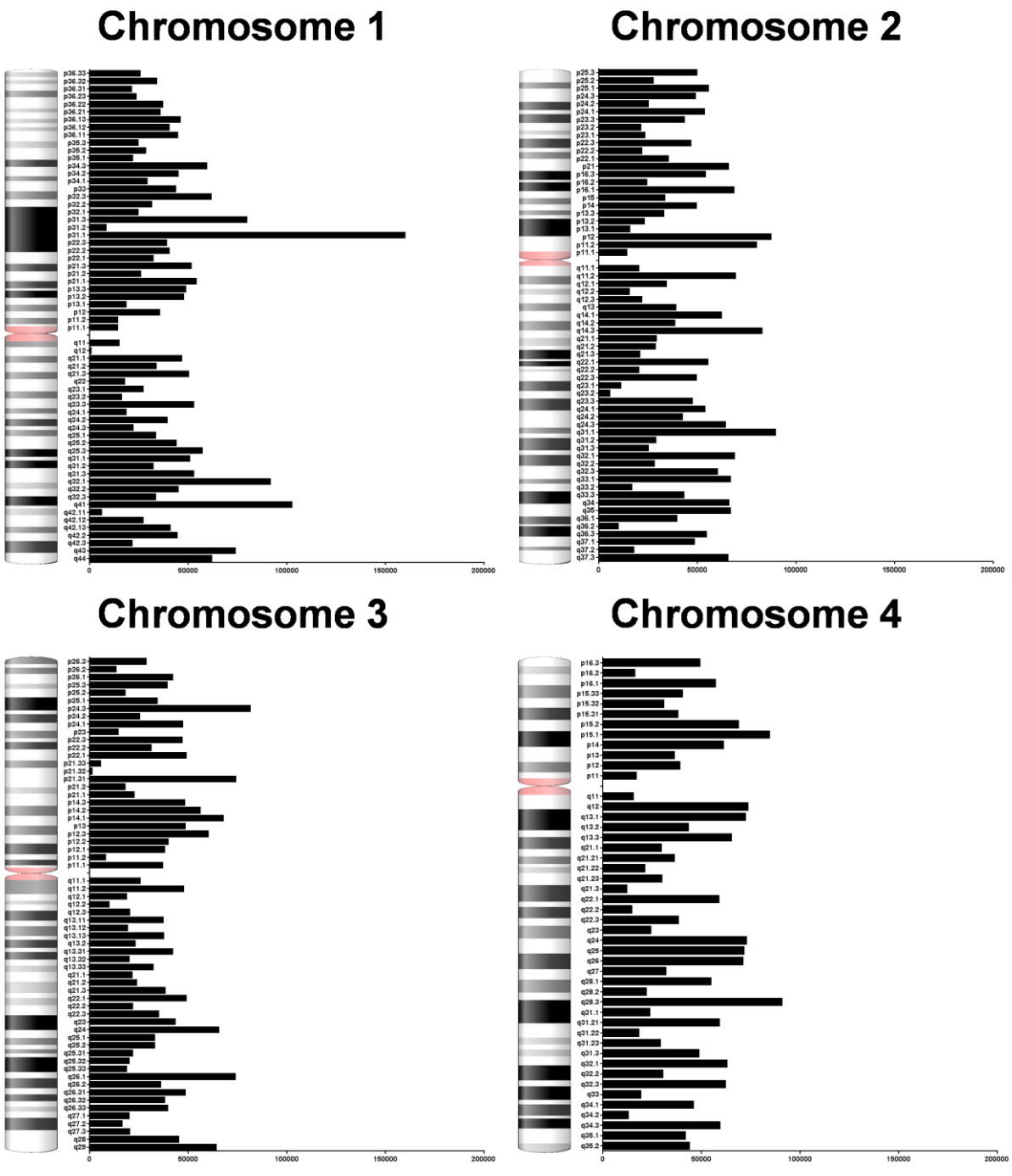
AAAAAAAAAAAAAAAACAAAAAAAACAAAACACACACACACACACACACACACACACACACACACACA
CACAACAAACCAACACCCAAAACCACAAAACA

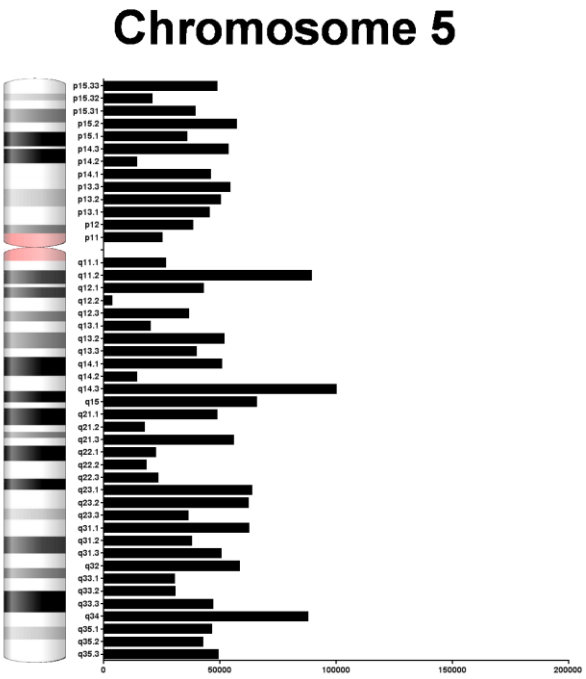
Cromossomo 18 – 100 nucleotídeos;

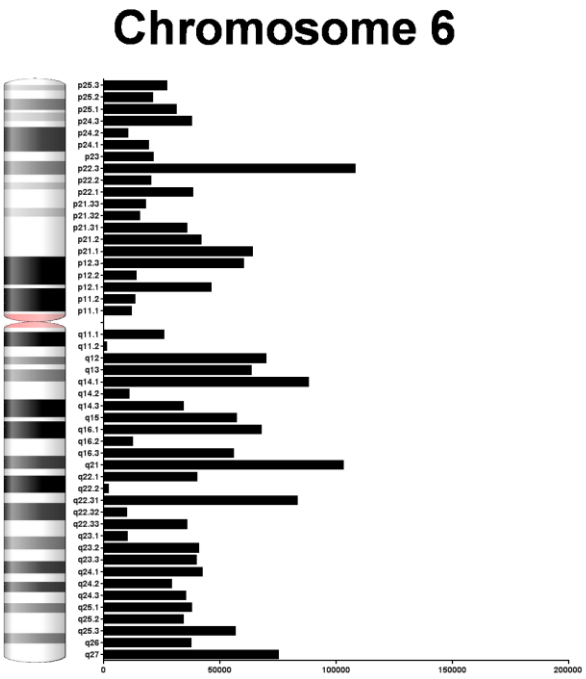
CAAAAAAAAAAAAAAAAAAAAAACCAACAAAAAAACACCAAAAAAACAAACAAACAAAAAACCCCCC
AAAAACCAAAAAACAAACAAACAAAAAAACAA

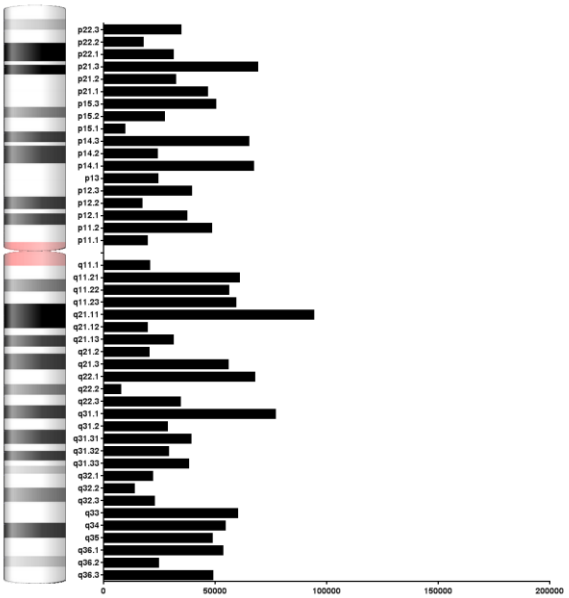
ANEXO 5

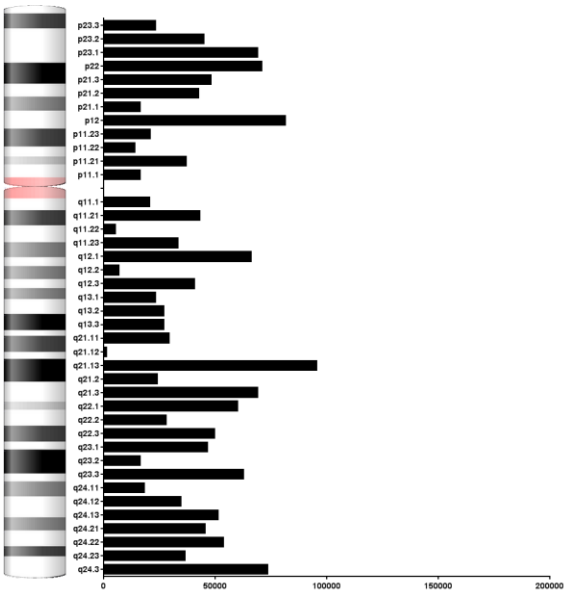
Mapa Genético das Correspondências CA's



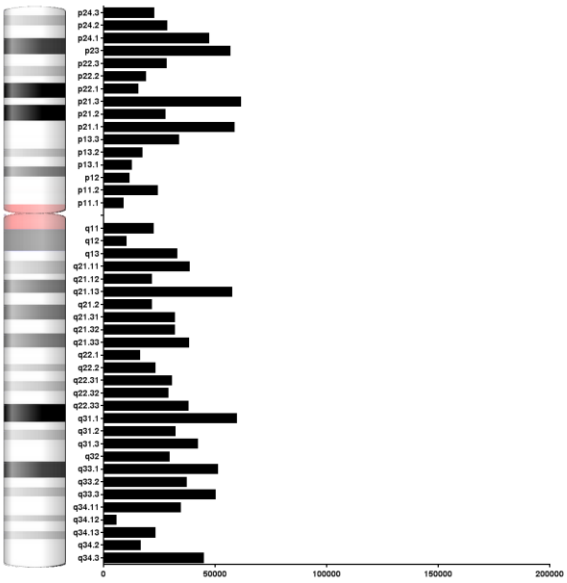

Chromosome 5
p15.33
p15.32
p15.31
p15.2
p15.1
p14.3
p14.2
p14.1
p13.3
p13.2
p13.1
p12
p11
q11.1
q11.2
q12.1
q12.2
q12.3
q13.1
q13.2
q13.3
q14.1
q14.2
q14.3
q15
q21.1
q21.2
q21.3
q22.1
q22.2
q22.3
q23.1
q23.2
q23.3
q31.1
q31.2
q31.3
q32
q33.1
q33.2
q33.3
q34
q35.1
q35.2
q35.3
0
50000
100000
150000
200000


Chromosome 6
p25.3
p25.2
p25.1
p24.3
p24.2
p24.1
p23
p22.3
p22.2
p22.1
p21.33
p21.32
p21.31
p21.2
p21.1
p12.3
p12.2
p12.1
p11.2
p11.1
q11.1
q11.2
q12
q13
q14.1
q14.2
q14.3
q15
q16.1
q16.2
q16.3
q21
q22.1
q22.2
q22.31
q22.32
q22.33
q23.1
q23.2
q23.3
q24.1
q24.2
q24.3
q25.1
q25.2
q25.3
q26
q27
0
50000
100000
150000
200000

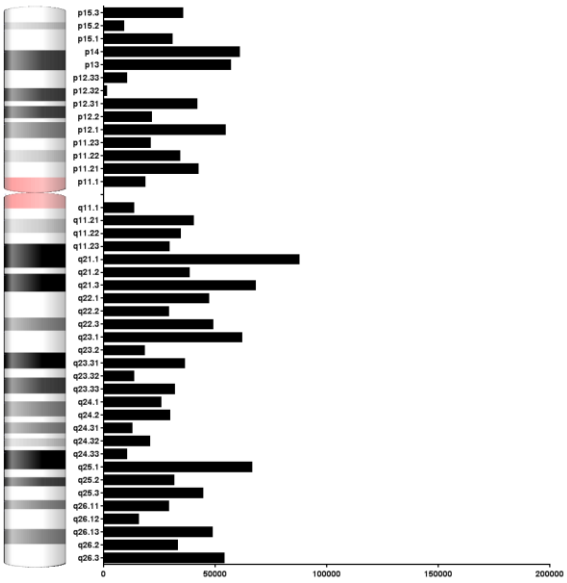

Chromosome 7
p22.3
p22.2
p22.1
p21.3
p21.2
p21.1
p15.3
p15.2
p15.1
p14.3
p14.2
p14.1
p13
p12.3
p12.2
p12.1
p11.2
p11.1
q11.1
q11.21
q11.22
q11.23
q21.11
q21.12
q21.13
q21.2
q21.3
q22.1
q22.2
q22.3
q31.1
q31.2
q31.31
q31.32
q31.33
q32.1
q32.2
q32.3
q33
q34
q35
q36.1
q36.2
q36.3
0
50000
100000
150000
200000


Chromosome 8
p23.3
p23.2
p23.1
p22
p21.3
p21.2
p21.1
p12
p11.23
p11.22
p11.21
p11.1
q11.1
q11.21
q11.22
q11.23
q12.1
q12.2
q12.3
q13.1
q13.2
q13.3
q21.11
q21.12
q21.2
q21.3
q22.1
q22.2
q22.3
q23.1
q23.2
q23.3
q24.11
q24.12
q24.13
q24.21
q24.22
q24.23
q24.3
0
50000
100000
150000
200000

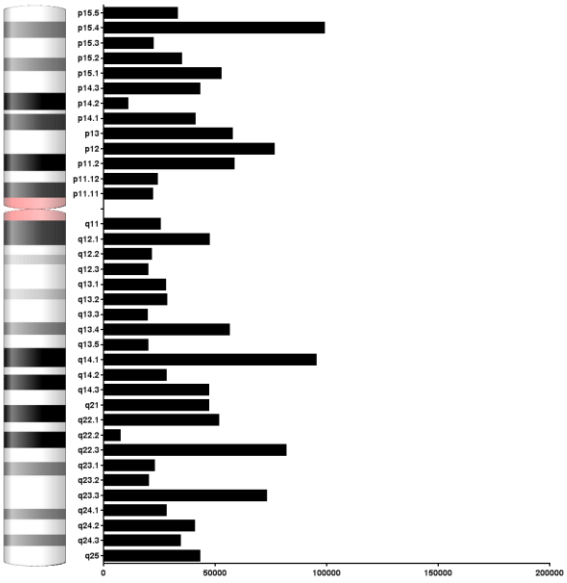
Chromosome 9



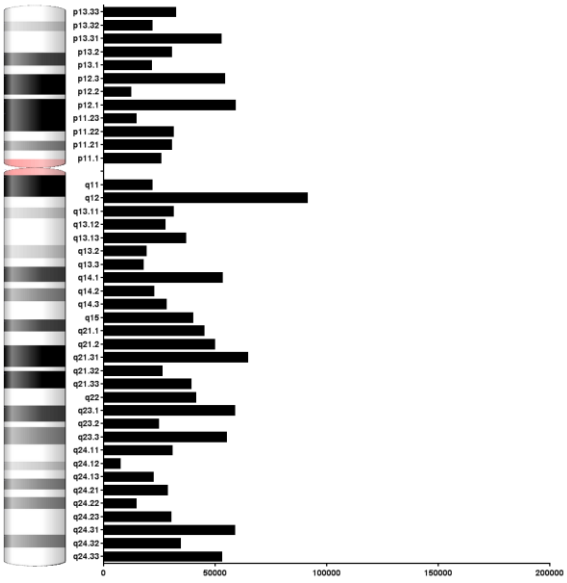
Chromosome 10



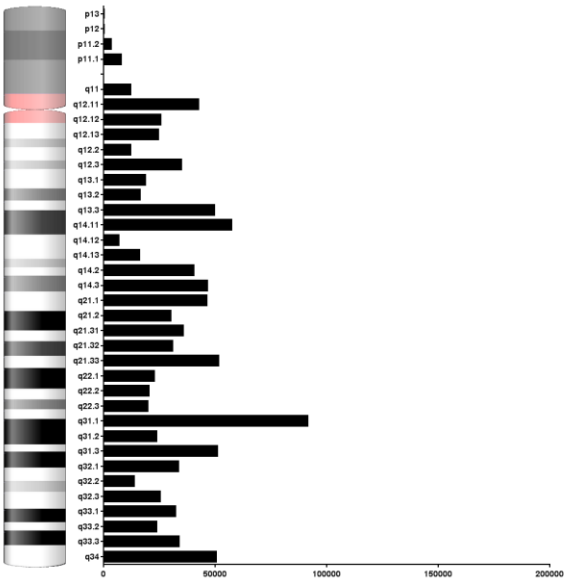
Chromosome 11



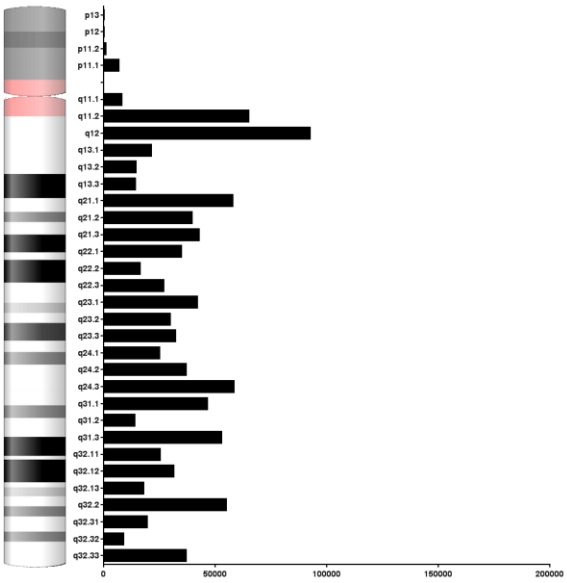
Chromosome 12



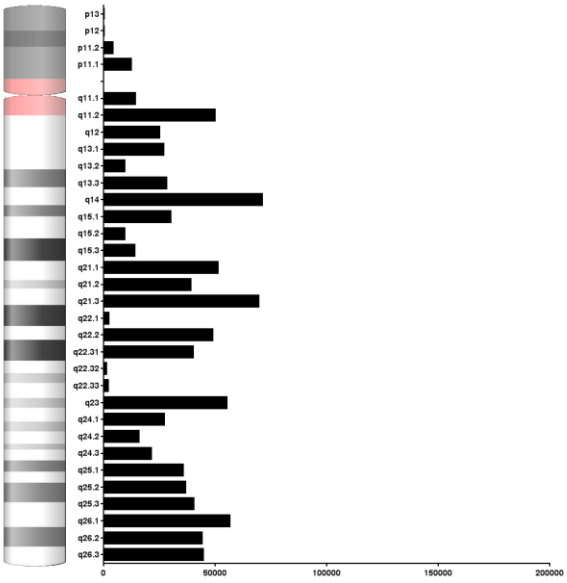
Chromosome 13



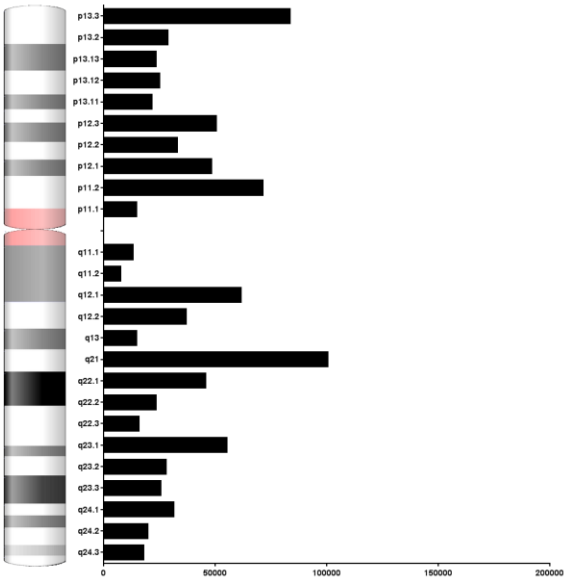
Chromosome 14



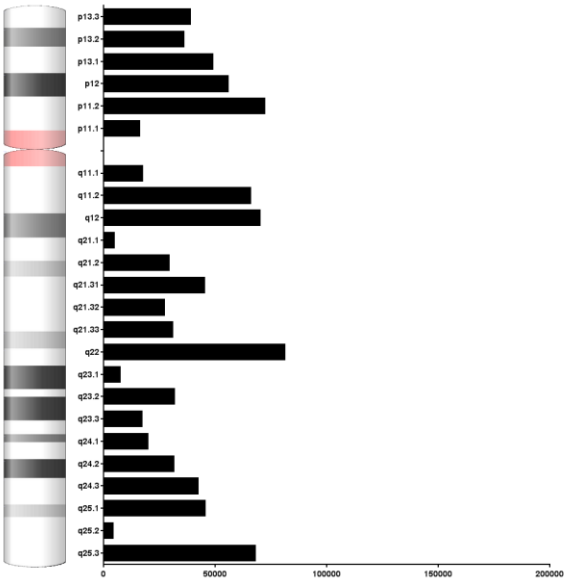
Chromosome 15



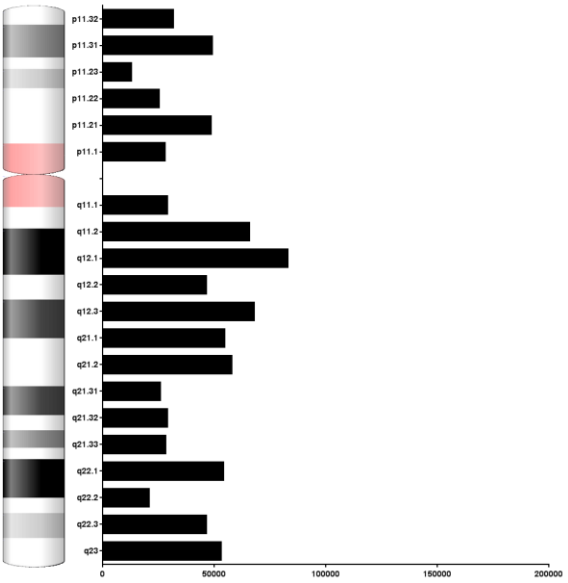
Chromosome 16



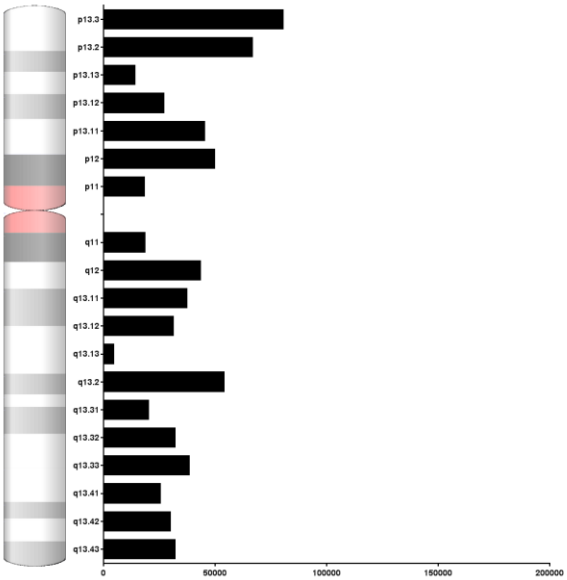
Chromosome 17



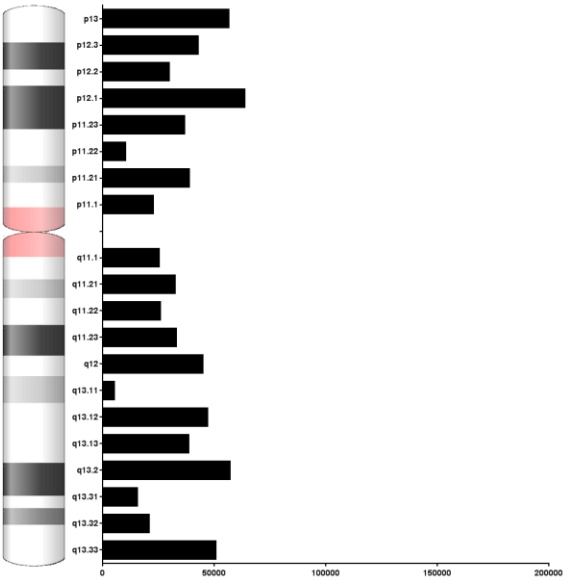
Chromosome 18

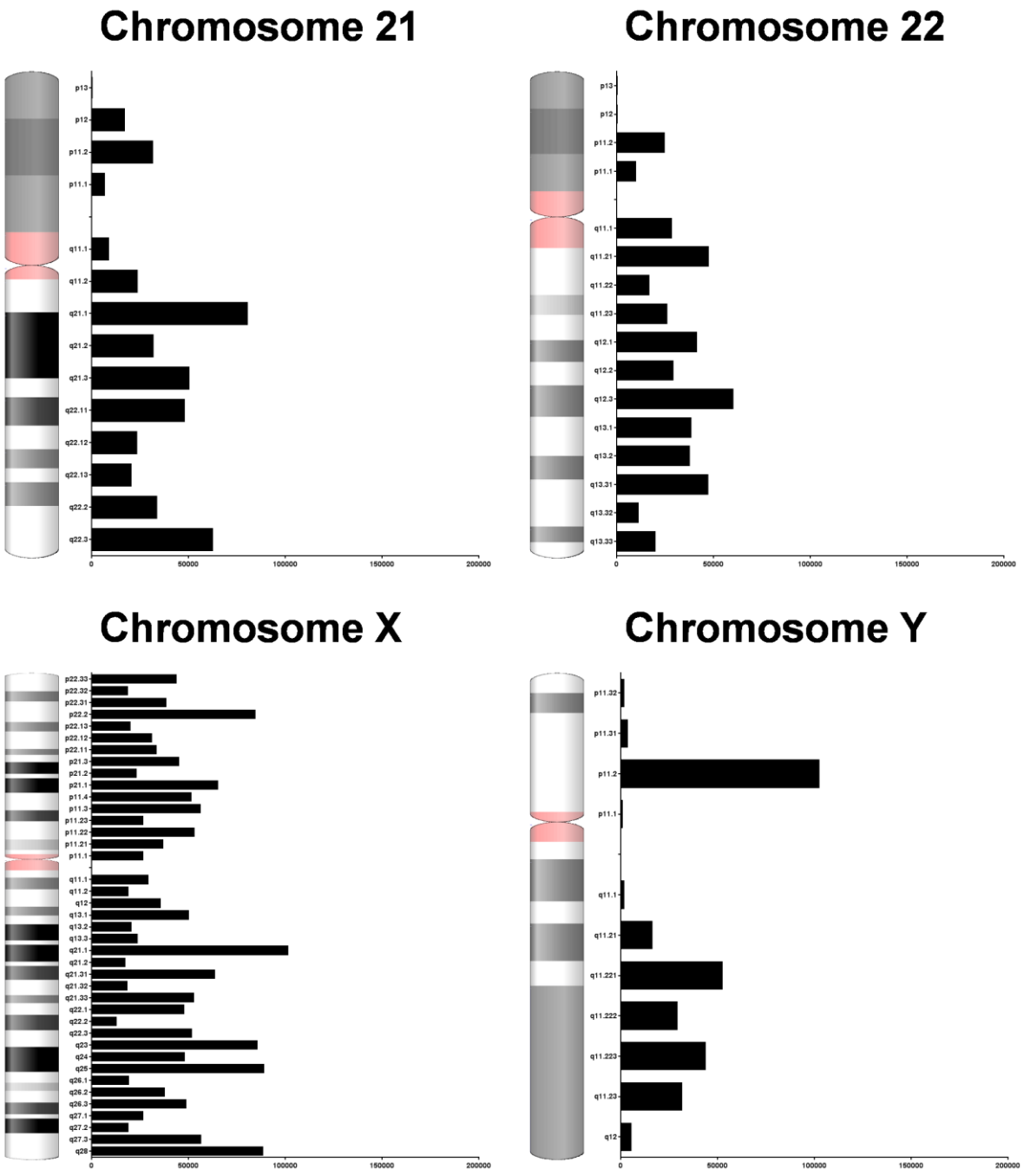


Chromosome 19



Chromosome 20




Chromosome 21
p13
p12
p11.2
p11.1
q11.1
q11.2
q21.1
q21.2
q21.3
q22.11
q22.12
q22.13
q22.2
q22.3
0
50000
100000
150000
200000
Chromosome 22
p13
p12
p11.2
p11.1
q11.1
q11.21
q11.22
q11.23
q12.1
q12.2
q12.3
q13.1
q13.2
q13.31
q13.32
q13.33
0
50000
100000
150000
200000
Chromosome X
p22.33
p22.32
p22.31
p22.2
p22.13
p22.12
p22.11
p21.3
p21.2
p21.1
p11.4
p11.3
p11.23
p11.22
p11.21
p11.1
q11.1
q11.2
q12
q13.1
q13.2
q13.3
q21.1
q21.2
q21.31
q21.32
q21.33
q22.1
q22.2
q22.3
q23
q24
q25
q26.1
q26.2
q26.3
q27.1
q27.2
q27.3
q28
0
50000
100000
150000
200000
Chromosome Y
p11.32
p11.31
p11.2
p11.1
q11.1
q11.21
q11.221
q11.222
q11.223
q11.23
q12
0
50000
100000
150000
200000