Juliana Gonçalves

Licenciada em Biologia

Características do genoma humano associadas à integração do HIV – análise bioinformática

Dissertação para obtenção do Grau de Mestre em
Genética Molecular e Biomedicina

Orientador: Prof. Doutora Maria Aldina Brás, Professora Auxiliar, Faculdade de Ciências Médicas da Universidade Nova de Lisboa

Co-orientador: Prof. Doutora Inês Jorge Sequeira, Professora Auxiliar, Faculdade de Ciências e Tecnologias da Universidade Nova de Lisboa

Co-orientador: Doutora Elsa Moreira, Investigadora, Faculdade de Ciências e Tecnologias da Universidade Nova de Lisboa

Júri:

Presidente: Prof. Doutora Paula Maria Theriaga Mendes Bernardo Gonçalves

Arguente: Prof. Doutor Nuno Filipe da Rocha Guerreiro de Oliveira

Vogal: Prof. Doutora Maria Aldina Lopes Brás

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Janeiro de 2014

Juliana Gonçalves

Licenciada em Biologia

Características do genoma humano associadas à integração do HIV – análise bioinformática

Dissertação para obtenção do Grau de Mestre em
Genética Molecular e Biomedicina

Orientador: Prof. Doutora Maria Aldina Brás, Professora Auxiliar, Faculdade de Ciências Médicas da Universidade Nova de Lisboa

Co-orientador: Prof. Doutora Inês Jorge Sequeira, Professora Auxiliar, Faculdade de Ciências e Tecnologias da Universidade Nova de Lisboa

Co-orientador: Doutora Elsa Moreira, Investigadora, Faculdade de Ciências e Tecnologias da Universidade Nova de Lisboa

Júri:

Presidente: Prof. Doutora Paula Maria Theriaga Mendes Bernardo Gonçalves

Arguente: Prof. Doutor Nuno Filipe da Rocha Guerreiro de Oliveira

Vogal: Prof. Doutora Maria Aldina Lopes Brás

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Janeiro de 2014

Características do genoma humano associadas à integração do HIV – análise bioinformática

Os resultados discutidos nesta tese originaram:

Publicações em revistas internacionais:

Sequeira IJ, **Gonçalves J**, Moreira E, Mexia JT, Rueff J e Brás A. *Genetic and Statistical Study of HIV Integration in the Human Genome.* Numerical Analysis and Applied Mathematics ICNAAM 2013 – AIP Conference Proceedings.

Comunicações orais em encontros internacionais:

Sequeira IJ, **Gonçalves J**, Moreira E, Mexia JT, Rueff J e Brás A. *Genetic and Statistical Study of HIV Integration in the Human Genome.* 11th International conference of numerical analysis and applied mathematics. Rodes, Grécia, 2013.

Comunicações orais em encontros nacionais:

Gonçalves J, Brás A, Sequeira IJ e Moreira E. *Human Genomics Features Associated with HIV Integration – a Bioinformatic Analysis.* Jornadas Intercalares das Dissertações Anuais dos Mestrados. Faculdade de Ciências e Tecnologia – Universidade Nova de Lisboa, Lisboa, Portugal, 2013.

Sequeira IJ, Moreira E, Mexia JT, **Gonçalves J**, Brás A e Rueff J. *Analysis of HIV Integration Sites in Human Chromosomes by Genetic and Statistical Methods.* International conference and advanced school planet earth, dynamics, games and science. Fundação Calouste Gulbenkian, Lisboa, Portugal, 2013.

Agradecimentos

Com a realização desta tese fecha-se mais um ciclo importante e por isso, não posso deixar de expressar o meu agradecimento a todos os que me ajudaram e apoiaram.

Agradeço ao Professor Doutor José Rueff, director do Departamento de Genética da Faculdade de Ciências Médicas da Universidade Nova de Lisboa, por toda a ajuda oferecida e por todo o incentivo.

Agradeço à Professora Doutora Aldina Brás por me ter recebido e orientado durante a realização desta tese. Agradeço também toda a amizade e carinho demonstrado durante este tempo e todo o apoio. Não posso deixar de agradecer todas as sugestões apresentadas durante a escrita da tese.

Agradeço também à Professora Doutora Inês Sequeira por todo o tempo dedicado, pela simpatia e por toda a ajuda prestada não só a nível da realização da tese como também ao nível da escrita.

Agradeço igualmente à Doutora Elsa Moreira por toda a simpatia, auxílio e disponibilidade demonstrada durante a realização e escrita desta tese.

A todas as pessoas que integram o grupo do Departamento de Genética da Faculdade de Ciências Médicas da Universidade Nova de Lisboa, quero agradecer a amizade, a motivação e o espírito de entreajuda sempre demonstrados. Agradeço também todos os conselhos e sugestões apresentados durante a realização desta tese.

À minha mãe Laurinda e ao meu pai José, a quem devo tudo o que sou, quero agradecer por todo o apoio, ajuda e carinho que sempre demostraram. Por nunca deixarem de acreditar em mim e nunca me deixarem desistir.

Ao meu irmão Helder agradeço por toda a amizade e pelo auxílio, pelas longas viagens em que partilhámos longas conversas e por todas as gargalhadas que trocámos juntos.

Agradeço também o apoio demostrado por toda a minha família e por todas as alegrias que me deram.

Ao meu namorado, Ricardo agradeço toda a ajuda prestada e por todo o ânimo que me deu, mesmo quando os dias corriam menos bem. Agradeço todo o tempo dedicado e toda a paciência. Não posso também deixar de agradecer todas as alegrias e momentos que

vivemos juntos ao longo destes anos.

Agradeço a todos os meus amigos que ficaram da família que formei em Évora durante a Licenciatura por toda a amizade e carinho ao longo destes anos. Agradeço especialmente à Vânia de Araújo pelo apoio dado especialmente durante este ano, uma amiga com quem sempre pude contar e que sei que estará sempre lá para tudo. Agradeço também à Margarida Figueira e ao Tiago Neves pelos dias bem passados, por todos os conselhos e pela amizade verdadeira que juntos criámos.

Às amigas do grupo de mestrado agradeço as tardes e horas de almoço de risada e alegria. Agradeço especialmente à Sara por todo o suporte a amizade ao longo deste ano. Quero deixar um agradecimento muito especial às minhas colegas de casa Graça e Clara pelo ano que vivemos juntas, por todos os planos concretizados e por estarem sempre presentes quando precisei.

Resumo

O vírus da imunodeficiência humana (HIV), para completar o seu ciclo de vida necessita de integrar o seu genoma no genoma humano, sendo que esta integração é feita em locais específicos. Ainda não estão completamente clarificadas as preferências de integração do HIV, por isso o nosso objectivo é estudar estas preferências utilizando as posições de sítios de integração de HIV-1 DNA e HIV-2 DNA isolados de células mononucleadas do sangue periférico (PBMCs) e de HIV-1 DNA isolado de células T Jurkat. As posições dos sítios de integração foram obtidas por recurso a bases de dados. Analisámos se os sítios de integração do HIV coincidiam com as bandas *Giemsa* claras que, sendo muito activas transcripcionalmente poderiam favorecer a integração. Analisámos também as preferências de integração do HIV relativamente aos sítios frágeis (FSs), alvos preferenciais para a integração de outros vírus. Para este estudo utilizámos dois testes não paramétricos, o dos sinais e de Wilcoxon e uma análise de variância (ANOVA). Os resultados mostraram que o HIV-1 DNA integra com maior intensidade nas bandas *Giemsa* claras, enquanto que o HIV-2 não apresenta preferências. Já os FSs não constituem alvos preferenciais para a integração deste vírus, integrando o HIV-1 DNA isolado de PBMCs com mais intensidade nas regiões não frágeis. É importante conhecer as preferências de integração do HIV, uma vez que os vectores retrovirais são utilizados em terapia génica, podendo assim contribuir-se para diminuir os riscos associados a esta terapia.

Palavras-chave: Vírus da Imunodeficiência Humana, bandas *Giemsa*, sítios frágeis cromossómicos, testes não paramétricos, análise de variância

Abstract

To complete its life cycle the Human Immunodeficiency Virus (HIV) integrates its genome in the human genome. This virus chooses specific sites and characteristics for the integration. Currently, it is not completely clarified the preferences of HIV integrations, therefore our aim is to study these preferences of the virus using the positions of integration sites of HIV-1 DNA and HIV-2 DNA isolated from peripheral mononuclear blood cells and HIV-1 DNA isolated from Jurkat T cells. The positions of integrations sites were obtained from databases. We analysed if HIV integrations sites coincide with Giemsa light bands that could be preferential targets of the virus since they are transcriptionally actives, as well as with fragile sites (FSs) that are preferential targets for the integration of other virus. In this thesis were used the signal and Wilcoxon tests, two nonparametric methods and the analysis of variance (ANOVA). The results show that HIV-1 DNA integrates with major intensity in Giemsa light bands while HIV-2 shows no preferences. Relatively to FSs, they are not preferential targets for the integration of this virus. HIV-1 DNA isolated from PBMCs integrates with major intensity in non fragile regions. It is important to know the integration preferences of HIV because retroviral vectors are used in gene therapy, so our data can be beneficial for contributing to decrease the risks associated to this therapy.

Keywords: Human Immunodeficiency Virus, Giemsa bands, chromosomal fragile sites, nonparametric tests, variance analysis

Índice Geral

Índice de Figuras

Índice de Tabelas

Lista de Abreviaturas e Siglas

AAV	Vírus Adeno-associado (do inglês *Adeno-Associated Virus*)
AIDS	Síndrome da Imunodeficiência Adquirida (do inglês *Acquired Immunodeficiency Syndrome*)
ANOVA	Análise de Variância (do inglês *Analysis of Variance*)
apc	Afidicolina (do inglês *Aphidicolin*)
A	Adenina
ATR	do inglês *Ataxia-Telangiectasia and Rad3-Related*
ATM	do inglês *Ataxia-Telangiectasia Mutaded*
BER	Reparação por Excisão de Pares de Bases (do inglês *Base Excision Repair*)
BLAST	do inglês *Basic Local Alignment SearchTool*
BLAT	do inglês *BLAST-like alignment tool*
bp	Pares de Bases (do inglês b*ase pair*)
BrdU	Bromodeoxiuridina (do inglês *Bromodeoxyuridine*)
C	Citosina
CFS	Sítio Frágil Comum (do inglês *Common Fragile Site*)
DNA	Ácido Desoxirribonucleico (do inglês *Desoxyrribonucleic Acid*)
DSB	Quebra da dupla cadeia de DNA (do inglês *Double-Strand Break*)
DSBR	Reparação da Quebra da Dupla Cadeia (do inglês *Double-Strand Break Repair*)
EBV	Vírus Epstein-Barr (do inglês *Epstein-Barr Virus*)
ERFS	Sítios Frágeis de Replicação Precoce (do inglês Early Replicated Fragile Sites)
FHIT	do inglês *Fragile Histidine Triad*
FISH	Hibridação *in situ* por Fluorescência (do inglês *Fluorescence in situ Hybridization*)
FoSTeS	do inglês *Fork Stalling and Template Switching*
FR	Região Frágil (do inglês *Fragile Region*)
FS	Sítio Frágil (do inglês *Fragile Site*)
G	Guanina
HIV	Vírus da Imunodeficiência Humana (do inglês *Human Immunodeficiency Virus*)
HIV-1	Vírus da Imunodeficiência Humana tipo 1 (do inglês *Human Immunodeficiency Virus type 1*)
HIV-2	Vírus da Imunodeficiência Humana tipo 2 (do inglês *Human Immunodeficiency Virus type 2*)
HPV	Vírus do Papiloma Humano (do inglês *Human Papiloma Virus*)
HR	Recombinação Homóloga (do inglês *Homologous Recombination*)
IL-2	Interleucina 2
IN	Integrase (do inglês *Integrase*)
Kb	Kilo bases (do inglês Kilobase)
LEDGF	do inglês *Lens Epithelium-Derived Growth Factor*
LINE	do inglês *Long Interspersed Repetitive Sequences*

LTR	Repetições Terminais Longas (do inglês *Long Terminal Repeats*)
MMBIR	do inglês *Microhomologies-mediated break-induced replication*
MMEJ	do inglês *Microhomology-mediated end-joning*
MMR	Reparação por *mismatch* (do inglês *Mismatch Repair*)
NAHR	do inglês *Non Allelic Homologous Recombination*
NCBI	do inglês National Center for Biotechnology Information
NER	Reparação por Excisão de Nucleótidos (do inglês *Nucleotide Excision Repair*)
NHEJ	do inglês *Non-Homologous End Joining*
NOR	Região dos Organizadores Nucleolares (do inglês Nucleolar Organizer Region)
NFR	Região Não Frágil (do inglês *Non Fragile Region*)
NFS	Sítio Não Frágil (do inglês *Non Fragile Site*)
PIC	Complexo de Pré Integração (do inglês *Pre-Integration Complexos*)
PBMC	Células Mononucleadas de Sangue Periférico (do inglês *Peripheral Blood Mononuclear Cells*)
RFS	Sítio Frágil Raro (do inglês *Rare Fragile Site*)
RNA	Ácido Rinonucleico (do inglês *Ribonucleic Acid*)
SCE	Troca de Crómatides Irmãs (do inglês *Sister Chromatid Exchange*)
SINE	do inglês *Short Interspersed Repetitive DNA Sequences*
SIV	Vírus da Imunodeficiência dos Símios (do inglês *Simian Immunodeficiency Virus*)
SSBR	Reparação da Quebra em Cadeia Simples (do inglês *Single-Stranded Break Repair*)
T	Timina
Vpr	Proteína Viral R (do inglês *Viral Protein R*)
Vpx	Proteína Viral X (do inglês *Viral Protein X*)
WWOX	do inglês *WW Domain-containing Oxireductase*
3D	3 Dimensões
5-azaC	5-Azacitidina (do inglês *5-azacytidine*)

I. Introdução

I.1 Genoma Humano

O genoma define a natureza de cada indivíduo, contendo longas sequências de ácidos nucleicos e toda a maquinaria necessária para construir um organismo, podendo ser funcionalmente dividido em genes (Lewin, 2004). No genoma, podemos encontrar sequências de DNA não-repetitivo e sequências de DNA repetitivo, sendo que dentro do DNA repetitivo podemos encontrar o DNA moderadamente repetitivo. Neste último DNA podemos observar a existência de transposões que sendo pequenas sequências, possuem a capacidade de se movimentarem no genoma ou até mesmo de fazerem novas cópias de si próprias (Lewin, 2004). Estes transposões estão também envolvidos nos rearranjos que ocorrem no genoma.

O desenvolvimento de técnicas de citogenética permitiu fazer novos estudos no que diz respeito ao genoma humano, nomeadamente o desenvolvimento da hibridização *in situ* por fluorescência (FISH) que permitiu visualizar os domínios da cromatina, os genes individualmente e a posição e organização de cada cromossoma (Bickmore, 2013).

A evolução das técnicas 3D permitiu também complementar os estudos realizados *in vivo*. Por exemplo, para se compreender melhor as funções do genoma *in vivo* é necessário ter em consideração os resultados obtidos da estrutura 3D dos cromossomas no núcleo, para assim estudar melhor a estrutura da cromatina e a expressão de genes (Bickmore e van Steensel, 2013).

Com o avanço das descobertas acerca do genoma humano, foram estudados os territórios cromossómicos que constituem assim o parâmetro básico da arquitectura do núcleo. Experiências realizadas mostram que micro-irradiações de uma pequena parte do núcleo apenas danificam um pequeno subconjunto do complemento do cromossoma mitótico, o que prova, indirectamente a existência de territórios cromossómicos (Cremer e Cremer, 2010). A visualização dos territórios cromossómicos só foi possível por volta de 1980 com o surgimento de técnicas de hibridização *in situ*. A distribuição radial dos territórios cromossómicos não é feita ao acaso, existindo estudos baseados em FISH que mostram que por exemplo no cromossoma 19 os territórios cromossómicos são encontrados no interior do núcleo de linfócitos, enquanto que os territórios do cromossoma 18 pobres em genes estão localizados na periferia do núcleo (Croft *et al.*, 1999; Cremer *et al.*, 2003). Isto permite concluir que a densidade de genes pode estar correlacionada com a distribuição radial no núcleo. Foram também realizados estudos em fibroblastos humanos com 3D FISH que permitiram distinguir territórios cromossómicos para os 22 pares de autossomas e para os 2 cromossomas sexuais (Bolzer *et al.*, 2005), aparecendo estes como estruturas com múltiplas formas de domínios da cromatina de ordem elevada (Küpper *et al.*, 2007). Um *locus*, tem a capacidade para sair e localizar-se fora do seu território cromossómico no núcleo, aumentando assim as probabilidades de interagir com sequências de outros cromossomas (Bickmore e van Steensel, 2013).

Na cromatina, podemos distinguir entre domínios activos e domínios inactivos, sendo que os domínios inactivos são mais limitados na sua habilidade para interagir a longas distâncias genómicas quando comparados com os domínios activos (Bickmore, 2013). As regiões dos domínios inactivos

possuem até uma tendência para se associarem umas com as outras (Simonis *et al.*, 2006). As regiões inactivas são normalmente encontradas dentro do seu território cromossómico (Bickmore, 2013), sendo as interacções entre domínios inactivos encontradas em porções restritas de cada braço do cromossoma.

O Projecto do Genoma Humano permitiu iniciar uma nova era de estudos em torno da genética, levando a grandes avanços na descoberta de genes específicos de doenças, na constituição de cada cromossoma, entre muitas outras. A sequenciação do genoma humano foi também o ponto de partida para a sequenciação e descoberta de outros organismos, nomeadamente os vírus, sendo possível comparar os diversos genomas e inferir sobre a evolução de certos genes.

O genoma humano possui uma grande maquinaria para que ocorra a replicação do DNA, a formação de novas células e para que todos os processos sejam regulados. Todas as características do genoma fazem com que este seja um hospedeiro preferencial para certos vírus, que assim o conseguem infectar e reproduzir-se.

I.1.1 Cromossomas Humanos

O genoma humano é constituído por vinte e dois pares de autossomas e dois cromossomas sexuais, sendo que a organização dentro de cada cromossoma não é feita ao acaso, existindo algumas regiões que podem possuir um elevado conteúdo em genes e outras um conteúdo mais baixo (Nussbaum *et al.*, 2007). Existem também regiões que são codificantes e outras que simplesmente não codificam para nenhum gene ou função específica. Na sequência do referido anteriormente, com a técnica de FISH foi possível provar que os cromossomas não estão organizados e distribuídos de forma aleatória, apresentando preferências de posicionamento relativamente à periferia do núcleo ou ao seu interior (Bickmore, 2013). Dentro dos cromossomas existe também uma organização, estando as regiões pobres em genes orientadas na direcção da periferia do núcleo quando comparadas com as regiões ricas em genes do mesmo cromossoma (Küpper *et al.*, 2007). A organização radial dos cromossomas pode ter consequências na formação de anomalias estruturais do genoma humano, sendo que a organização do núcleo não é rígida, de modo que os cromossomas nem sempre se apresentam no seu local preferido (Bickmore, 2013).

Cada cromossoma é constituído por um conjunto de genes estando cada um localizado no seu *locus* (Lewin, 2004).

Podemos recorrer a numerosas técnicas para visualizar, identificar e estudar a estrutura dos cromossomas humanos. Nesta tese faremos especial atenção às bandas *Giemsa* e aos sítios frágeis (FSs) cromossómicos, sendo que cada uma destas estruturas possui uma sequência típica.

I.1.1.1 Do DNA ao Cromossoma

A unidade básica fundamental da estrutura da cromatina é o nucleossoma que é constituído pelo DNA e pelas histonas H2A, H2B, H3 e H4. As histonas apresentam uma grande densidade de aminoácidos com carga positiva, o que lhes permite ligarem-se firmemente à dupla hélice de DNA que possui carga negativa (Speicher, 2010).

Existem dois tipos de variantes de histonas que diferem entre si na sua sequência primária, na sua expressão durante o ciclo celular e na sua distribuição no genoma. O primeiro tipo, as histonas canónicas apresentam um grande pico de expressão durante a fase S, enquanto que o outro tipo, as variantes de substituição possuem expressão independentemente da fase S (Szenker *et al.*, 2011).

Para além do domínio funcional das histonas que lhe permite o contacto histona-histona e histona-DNA, estas possuem o domínio que contém sítios para modificações pós-translacionais, tais como acetilação, metilação, fosforilação e ubiquitinação (Speicher, 2010). As modificações que ocorrem nas histonas afectam a estrutura do nucleossoma ou até mesmo criam locais para a ligação de proteínas não histónicas, podendo levar a modificações nas propriedades da cromatina (Lewin, 2004). A metilação das histonas está associada com a inactividade. Já a acetilação das histonas permite que a cromatina esteja acessível para as proteínas que se ligam ao DNA (Tse *et al.*, 1998), tornando assim a cromatina mais eficientemente transcrita (Nightingale *et al.*, 1998).

O DNA é organizado em cada cromossoma de um modo altamente eficaz, para permitir a transcrição individual de genes, a regulação génica, a replicação, o emparelhamento dos cromossomas homólogos durante a meiose, o *crossing-over* meiótico e outras funções (Spurbeck *et al.*, 2004). O termo cromossoma deriva da expressão *colored body*, a qual significa corpo com cor (Dolan, 2011), tendo sido identificados pela primeira vez no século XIX. Para além de DNA e histonas, são ainda constituídos por proteínas não histónicas (Dolan, 2011).

Cada cromossoma apresenta-se como uma repetição de nucleossomas e segmentos curtos de DNA que se ligam a cada nucleossoma (Speicher, 2010).

O nucleossoma e as fibras de cromatina estão num contínuo estado de fluxo. É actualmente conhecido que a organização da cromatina está envolvida no controlo da transcrição e na regulação da segregação cromossómica, replicação e recombinação (Wolffe, 2001; Holmquist e Ashley, 2006). O nucleossoma constitui o primeiro nível de organização, possuindo um octâmero de histonas e sendo um componente invariável da eucromatina e da heterocromatina na interfase (Lewin, 2004).

A estrutura do DNA possui toda a informação química que permite a transmissão da informação genética de uma geração para outra.

I.1.1.2 Organização do material genético nos cromossomas humanos

Durante a divisão celular, o cromossoma tem a aparência em forma de x que se tornou um símbolo nas ciências da vida (Uhlmann, 2013). Durante a interfase a cromatina está descondensada, começando a condensar à medida que avança a divisão celular, sendo completamente visível a

estrutura e morfologia do cromossoma na metafase (Spurbeck *et al.*, 2004). Nessa fase do ciclo celular, podemos visualizar em cada cromossoma um centrómero, dois telómeros e braço curto e longo. Podem-se distinguir assim três tipos de cromossomas, os metacêntricos em que o centrómero se localiza no centro do cromossoma, os acrocêntricos cujos centrómeros se apresentam quase na extremidade terminal do cromossoma e os submetacêntricos nos quais o centrómero se encontra mais próximo de uma extremidade do cromossoma (Spurbeck *et al.*, 2004). Os acrocêntricos possuem no seu braço curto as regiões dos organizadores nucleolares (NORs), sendo estas constituídas por heterocromatina (Sullivan *et al.*, 2001).

Os cromossomas possuem proteínas que são responsáveis por manter a sua estrutura, sendo elas de dois grupos distintos, as condensinas e as coesinas. As primeiras estão envolvidas na condensação durante a mitose, controlando a estrutura global do cromossoma, enquanto que as coesinas estão presentes nas ligações entre cromátides irmãs (Lewin, 2004).

I.1.1.2.1 Centrómeros

Os centrómeros são entidades individuais durante a interfase, sendo identificados durante a metafase (Boyarchuk *et al.*, 2011). As suas sequências funcionam através da interacção com uma proteína de estrutura hierárquica denominada de cinetócoro (Cheeseman e Desai, 2008), sendo que o centrómero representa o *locus* onde as fibras do cinetócoro se ligam às cromátides (Cleveland *et al.*, 2003). É assim possível que durante a divisão celular, ocorra a correcta segregação dos cromossomas pela qual os centrómeros são responsáveis (Allshire e Karpen, 2008). As sequências dos centrómeros apresentam variações entre indivíduos (Warburton *et al.*, 1991), o que leva a que cada genoma humano possua uma sequência personalizada em termos de composição e organização.

Os centrómeros possuem uma grande quantidade de satélites alfa, que pertencem a uma das famílias de satélites ricos em Adenina (A) e Timina (T) (Manuelidis, 1978). Estes satélites alfa interagem com as proteínas internas do cinetócoro, tendo sido demonstrado que apresentam competência para estabelecerem de novo a identidade dos centrómeros em ensaios com cromossomas artificiais (Schueler *et al.*, 2001). Os satélites alfa são regiões bastante instáveis apresentando tendência a rearranjos (Alkan *et al.*, 2011). Tal como as sequências dos centrómeros, também os satélites alfa variam entre indivíduos não relacionados, o que fornece uma variação no DNA para se investigar padrões de polimorfismos dos centrómeros na população (Hayden, 2012).

Podemos distinguir dois domínios da cromatina dento dos centrómeros, a cromatina central e a heterocromatina adjacente pericêntrica. O primeiro domínio é responsável pela formação do cinetócoro, sendo que se caracteriza pela presença do CenH3, uma variante específica da histona H3 (Boyarchuk *et al.*, 2011). O segundo domínio tem como principal característica o facto de não apresentar a variante da H3, o CenH3. O CenH3, é importante para assegurar o correcto funcionamento dos centrómeros e preservar a sua identidade, já que tem como funções formar partículas como as do nucleossoma e organizar a cromatina centromérica (Boyarchuk *et al.*, 2011).

Embora seja determinante não é o único factor que contribui para a identidade dos centrómeros, sendo também de considerar a incorporação de outras variantes de histonas. A própria organização 3D do centrómero contribui igualmente para a identidade do mesmo (Jansen *et al.*, 2007), nomeadamente as condensinas que são importantes para a deposição e retenção do CenH3 nos centrómeros e que formam complexos proteicos que contribuem para a organização da cromatina (Yong-Gonzalez *et al.*, 2007; Samoshkin *et al.*, 2009).

O estudo destas regiões genómicas é muito importante, uma vez que algumas doenças têm sido associados com regiões que coincidem com lacunas centroméricas, nomeadamente a esclerose múltipla (Reich *et al.*, 2005) e o cancro (Stacey *et al.*, 2008). No entanto, o estudo dos centrómeros apresenta ainda muitos desafios, na medida em que é necessário uma montagem através de milhões de pares de bases (bp) de sequências altamente repetitivas e pouco identificadas (Hayden, 2012).

I.1.1.2.2 Telómeros

Os telómeros foram descobertos em 1930 primeiro por Herman Muller e depois por Barbara McClintock. São as regiões localizadas no final dos cromossomas (de Lange, 2005) que conferem protecção de eventos que promovem a instabilidade genética como fusões de um telómero a outro e degradação das regiões terminais (Blackburn, 2005). Estas estruturas são bastante dinâmicas e o número de repetições teloméricas pode variar de célula para célula (Zakian, 2012) e até mesmo entre alelos no mesmo telómero.

Os telómeros dos mamíferos são constituídos por DNA de dupla cadeia com pequenas repetições *tandem* TTAGGG (Palm e de Lange, 2008) seguidas do terminal 3´ rico em G de cadeia simples (Lu *et al.*, 2013) ligado a um complexo com seis polipéptidos diferentes denominado de *shelterin* ou telossoma (de Lange, 2005). Este complexo é formado pelas proteínas teloméricas 1 e 2 (TRF1 e TRF2), pelas proteínas que interagem com a proteína nuclear 2 (TIN2), pela proteína protectora de telómeros 1 (POT1), pelo repressor activador da proteína 1 (RAP1) e pela TPP1 (Liu *et al.*, 2004; Palm e de Lange, 2008). Este complexo conduz não só à síntese de DNA telomérico como também afecta a estrutura terminal do telómero (Schluth-Bolard *et al.*, 2010). O DNA do telómero é distinto do resto do genoma e foi provado que inibe alguns sinais de respostas a danos no DNA, nomeadamente da recombinação homóloga (HR) e da *non-homologous end joining* (NHEJ) (de Lange, 2009). Esta inibição, pode ser vista como um mecanismo para prevenir a fusão do final dos telómeros (Smogorzewska *et al.*, 2002).

Os telómeros possuem um problema no final da replicação devido a limitações na polimerase de DNA em completar a replicação das regiões terminais de moléculas lineares (Chan e Blackburn, 2004), sendo que este problema leva a que ocorra um encurtamento dos telómeros em cada ciclo de replicação (Richter e von Zglinicki, 2007; Armanios *et al.*, 2009). A telomerase, enzima presente nos telómeros previne que ocorra este encurtamento da estrutura. Existem também outros factores que podem levar ao encurtamento dos telómeros, como o *stress* oxidativo (von Zglinicki, 2002) e também

o avançar da idade (Armanios *et al.*, 2009), sendo que a perda destes pode trazer consequências graves nomeadamente dificuldades para a célula se dividir com sucesso.

I.1.1.2.3 Eucromatina *versus* Heterocromatina

A cromatina pode ser dividida em eucromatina e heterocromatina, apresentando cada um dos tipos de material características distintas.

A eucromatina representa a maior componente do genoma ocupando a maior parte do núcleo e, apresenta-se menos densamente empacotada quando comparada com a heterocromatina (Lewin, 2004). As regiões de eucromatina, apresentam diferentes estados de condensação durante a interfase e na mitose. São regiões que possuem genes activos, por isso são necessárias para a expressão génica.

A heterocromatina é normalmente encontrada nos centrómeros, sendo composta por satélites de DNA e descrita como permanentemente enrolada e inerte (Lewin, 2004). Esta cromatina pode ser visualizada durante todo o ciclo celular, uma vez que é possível fazer-se coloração, tendo sido identificada como possuindo poucos ou nenhuns genes activos (Speicher, 2010). Encontra-se normalmente mais densamente empacotada com fibras, conseguindo passar o ciclo celular com poucas alterações no seu grau de condensação, sendo que possui proteínas que se ligam à cromatina e que impedem os factores de transcrição de activarem promotores nestas regiões (Lewin, 2004). A heterocromatina pode ainda ser subdividida em constitutiva ou facultativa, sendo a primeira observada mais perto dos centrómeros e idêntica em todas as células.

I.1.1.2.4 Técnicas de Bandeamento

No passado, e anteriormente à descoberta das técnicas de bandeamento dos cromossomas, estes apenas podiam ser distinguidos pelo seu tamanho e a posição relativa do centrómero (Lewin, 2004). Agora, é possível analisar e identificar cada cromossoma recorrendo a técnicas de bandeamento. Com a atribuição destes padrões, tornou-se possível identificar translocações, deleções ou inserções nos vários cromossomas, permitindo também o estudo mais detalhado de doenças. Para realizar este bandeamento existem diversas técnicas que podem ser utlizadas, dependendo as bandas geradas do tipo de tratamento e consequente resposta do cromossoma (Lewin, 2004). O bandeamento refere-se aos padrões gerados pelas bandas alternando entre escuro e claro visível por microscopia óptica ao longo do cromossoma, sendo único para cada cromossoma (Dolan, 2011).

A primeira técnica utilizada foi designada por bandeamento-Q, devido ao facto de se utilizar o corante fluorescente quinacrina (Dolan, 2011). Esta técnica foi descrita pela primeira vez por Carpersson *et al.* (1970) (Carpersson *et al.*, 1970) tendo sido as bandas designadas de bandas-Q que

surgiam num padrão alternado de brilho e sem brilho. Esta técnica de bandeamento é bastante útil na medida que permite estabelecer a natureza específica de anomalias cromossómicas, quer numéricas quer estruturais (Spurbeck *et al.*, 2004).

A técnica mais utilizada é a das bandas-G, na qual os cromossomas são tratados com tripsina e depois corados com *Giemsa* (Seabright, 1971). Para além de permitir também visualizar anomalias estruturais e numéricas, esta técnica fornece também informações acerca de anomalias constitucionais e doenças adquiridas (Spurbeck *et al.*, 2004), podendo ser utilizada em preparações de células de vários tecidos como sangue e tumores.

Outra das técnicas utilizadas é denominada de bandeamento-R corando esta a eucromatina, tal como o bandeamento-G. Estas duas técnicas permitem criar um padrão alternado de bandas escuras e claras. Pelo bandeamento-R surgem coradas as bandas escuras que aparecem a claro quando utilizado o bandeamento-G e vice versa (Holmquist, 1992). A técnica do bandeamento-R permitiu grandes avanços no que diz respeito ao final dos cromossomas que no bandeamento-G apareciam menos nítidos. Com esta técnica, algumas regiões terminais aparecem escuras e por isso torna-se mais fácil serem analisadas (Dolan, 2011). Nesta técnica, utiliza-se tampão fosfato quente para incubação, seguida da coloração *Giemsa*.

Temos também a técnica de bandeamento-C que permite visualizar a heterocromatina centromérica, tendo sido descrita pela primeira vez por Pardue e Gall (1970) (Pardue e Gall, 1970). Esta técnica permite detectar rearranjos estruturais que estão envolvidos nas regiões pericêntricas dos cromossomas 1, 9 e 16 e na zona distal do braço longo do cromossoma Y (Arrighi e Hsu, 1971; Speicher, 2010). O método mais utilizado para produzir estas bandas consiste no tratamento dos cromossomas com ácido hidroclorídrico, seguido de uma incubação com hidróxido de bário e uma lavagem com soluções salinas, sendo no fim as preparações coradas com *Giemsa* (Spurbeck *et al.*, 2004). Esta técnica é muito útil para detecção de múltiplos centrómeros e para distinguir entre células dadoras e receptoras nos transplantes de medula.

Existe ainda outra técnica de bandeamento destinada especificamente às NORs, denominada bandeamento Ag-NOR. O termo Ag refere-se ao símbolo químico da prata, indicando assim que esta técnica é realizada por impregnação de prata (Speicher, 2010). Esta técnica permite corar o DNA ribossomal que está activo dos braços curtos dos acrocêntricos.

I.2. Bandas *Giemsa*

Como afirmado acima, a técnica de coloração diferencial mais utilizada para identificar os cromossomas individuais é a coloração *Giemsa* (Nussbaum *et al.*, 2007). A utilização da enzima proteolítica tripsina permite digerir as proteínas dos cromossomas e é seguida da coloração *Giemsa* para se conseguir obter um padrão alternado de bandas escuras e claras (Nussbaum *et al.*, 2007; Speicher, 2010). Estas bandas permitem estudar não só a estrutura do cromossoma como também a sua morfologia. As bandas escuras, também designadas de bandas-G (*Giemsa*) apresentam

características e padrões muito distintos das bandas claras denominadas de bandas-R (*Reverse*) (Carvalho *et al.*, 2001; Niimura e Gojobori, 2002).

A nomenclatura destas bandas foi definida na Conferência de Paris em 1971 atribuindo-se um número aos autossomas de 1 a 22 sendo os cromossomas sexuais designados de X e Y. Cada braço do cromossoma, curto (p) e longo (q), foi dividido em regiões numeradas e dentro de cada uma delas as diferentes bandas foram designadas por um número.

A distribuição das bandas nos vários cromossomas pode também ser obtida através de coloração de cromossomas *in silico*, utilizando para isso programas de computador e sequências de DNA do projecto do genoma humano (Niimura e Gojobori, 2002). Esta distribuição de bandas é feita de acordo com a percentagem de Guanina (G) e Citosina (C), podendo assim observar-se claramente quais as regiões mais ricas ou pobres em G/C. Nesta tese foram usadas as bandas obtidas *in silico* por Niimura e Gojobori (2002), as quais se encontram representadas na figura I.1. O trabalho destes autores é também citado e desenvolvido no clássico livro de referência "Vogel and Motulsky´s Human Genetics" (Speicher, 2010).

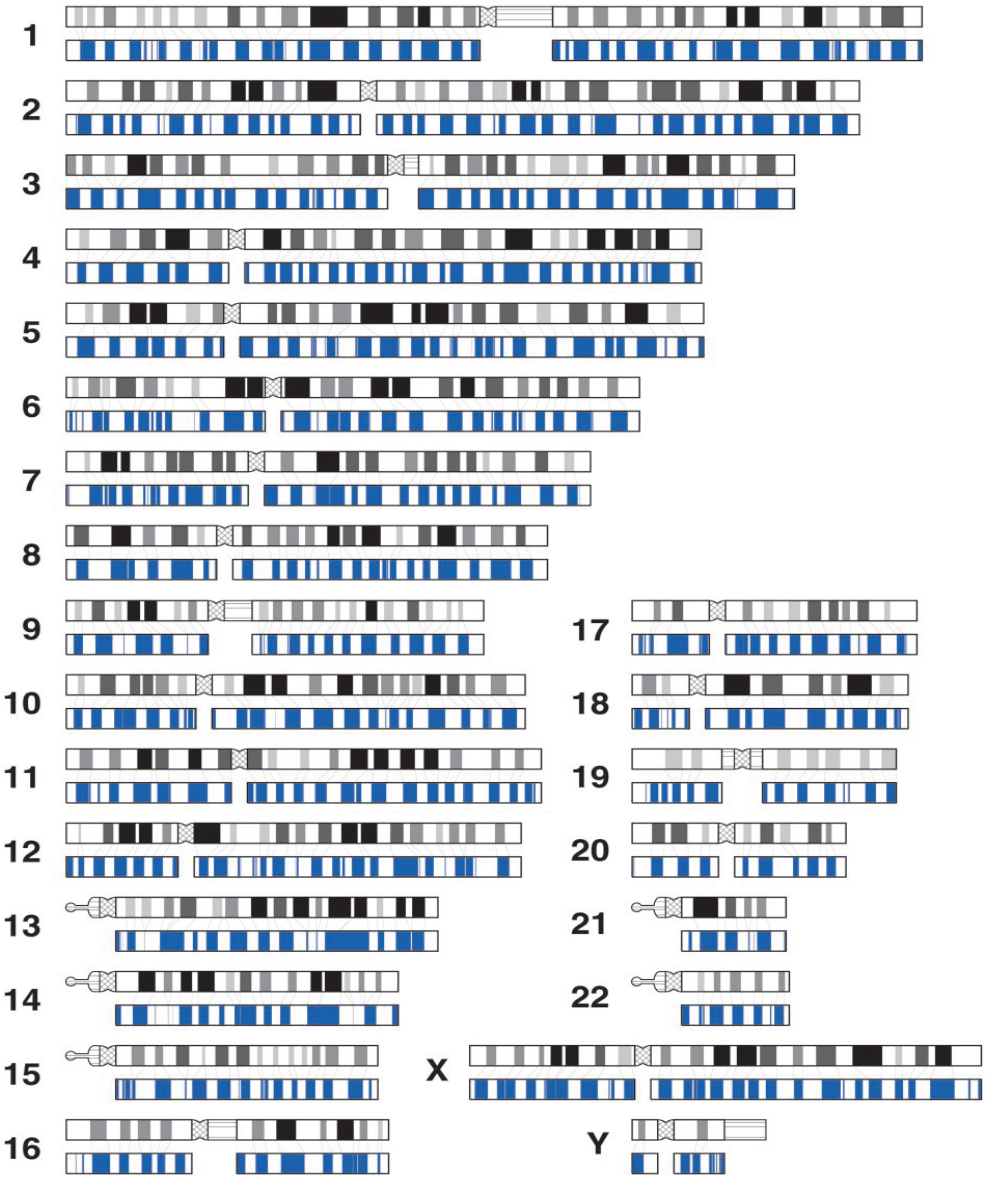


Figura I.1 – (página anterior) Representação esquemática das bandas *Giemsa* e *in* silico. A cinzento encontram-se as bandas *Giemsa*, enquanto que a azul se encontram as bandas *in silico* para todos os cromossomas. Os braços p e q estão representados à esquerda e à direita, respectivamente. Para os cromossomas 1, 3, 9, 16, 19 e Y a região da heterocromatina encontra-se representada com linhas na horizontal. A região do centrómero encontra-se representada pelas linhas diagonais em cruz. Retirado de Niimura e Gojobori, 2002.

Esta coloração *in silico* permitiu grandes avanços, resolvendo o problema de que várias experiências revelavam que as correspondências entre bandas *Giemsa* escuras e regiões pobres em G/C por um lado, e bandas *Giemsa* claras e regiões ricas em G/C por outro, eram bastante ténues. Assim, esta técnica mostra claramente que as bandas *Giemsa* escuras são regiões pobres em G/C comparativamente com as regiões flanqueadoras (Speicher, 2010).

Apesar do extraordinário avanço dos métodos citogenéticos nos últimos anos, o bandeamento-G permanece um tópico actual (Dolan, 2011; Francke, 2013), quer pela necessidade do seu uso para complementar técnicas recentemente desenvolvidas como por exemplo a FISH, a CGH e a sequenciação completa do genoma, quer porque o material genético contido nas bandas *Giemsa* claras é significativamente diferente do material contido nas bandas *Giemsa* escuras, reflectindo diferentes estruturas e funções, como detalhado abaixo.

I.2.1 Bandas *Giemsa* Escuras ou R claras

As bandas *Giemsa* escuras apresentam um conteúdo muito pobre em G/C (Federico *et al.*, 2000), sendo ricas em sequências de A/T. Estas bandas apresentam ainda uma baixa densidade de genes (Watanable e Maekawa, 2013) contendo preferencialmente genes específicos de tecidos e replicando mais tardiamente quando comparadas com as bandas *Giemsa* claras (Carvalho *et al.*, 2001). São ainda transcricionalmente menos activas. As repetições mais presentes neste tipo de bandas são as LINEs (*long interspersed repetitive sequences*) (Bickmore e van Steensel, 2013). O DNA das bandas *Giemsa* escuras encontra-se localizado na periferia do núcleo (Cremer e Cremer, 2001).

Recentemente foi mostrada evidência de que o bandeamento-G está relacionado com a posição do DNA que foi replicado na fase S tardia do ciclo celular (Hoshi e Ushiki, 2011). Domínios de replicação tardia são geralmente pobres em genes, ricos em A/T e muitas vezes localizados na periferia do núcleo (Hiratani *et al.*, 2009). Estes domínios de replicação podem assemelhar-se aos domínios constitutivos associados à lâmina nuclear que são universalmente caracterizados por longos fragmentos de DNA com alto teor A/T podendo contribuir para a arquitectura basal do cromossoma (Meuleman *et al.*, 2013).

Tabela I.1 – Resumo das características que distinguem as bandas-G que replicam tardiamente das bandas-R que replicam precocemente. Adaptado de Holmquist e Ashley, 2006.

Coram de escuro depois da tripsina – *Giemsa*
Replicam tardiamente
Pobres em genes
Sem genes *housekeeping*
Ricas em A/T
Pobres em SINEs, Pobres em Alu
Ricas em LINEs, Ricas em L1
Ilhas CpG são raras
Recombinação meiótica infrequente

I.2.2 Bandas *Giemsa* Claras ou R escuras

Ao contrário das bandas *Giemsa* escuras, as bandas *Giemsa* claras caracterizam-se por possuírem uma elevada densidade de genes nomeadamente genes *housekeeping* (Cremer e Cremer, 2001), apresentando sequências ricas em G/C e contendo um grande número de ilhas CpG. Possuem ainda um grande conteúdo em repetições Alu e SINEs (*short interspersed repetitive DNA sequences*) (Speicher, 2010), sendo transcricionalmente mais activas. A proximidade espacial dos genes influi directamente a co-regulação da transcrição (Stefano *et al.*, 2013). Este elevado conteúdo de repetições Alu nas bandas ricas em G/C pode ser explicada pelo facto de as repetições Alu possuírem como alvo para inserção as regiões ricas em G/C e ainda pelo facto das repetições Alu serem seleccionadas positivamente nas regiões ricas em G/C (International Human Genome Sequencing Consortium, 2001). O facto de este tipo de bandas possuir um conteúdo elevado de G/C e como tal uma cromatina mais aberta faz com que estas tenham um papel importante no início da integração de provírus (Costantini *et al.*, 2012).

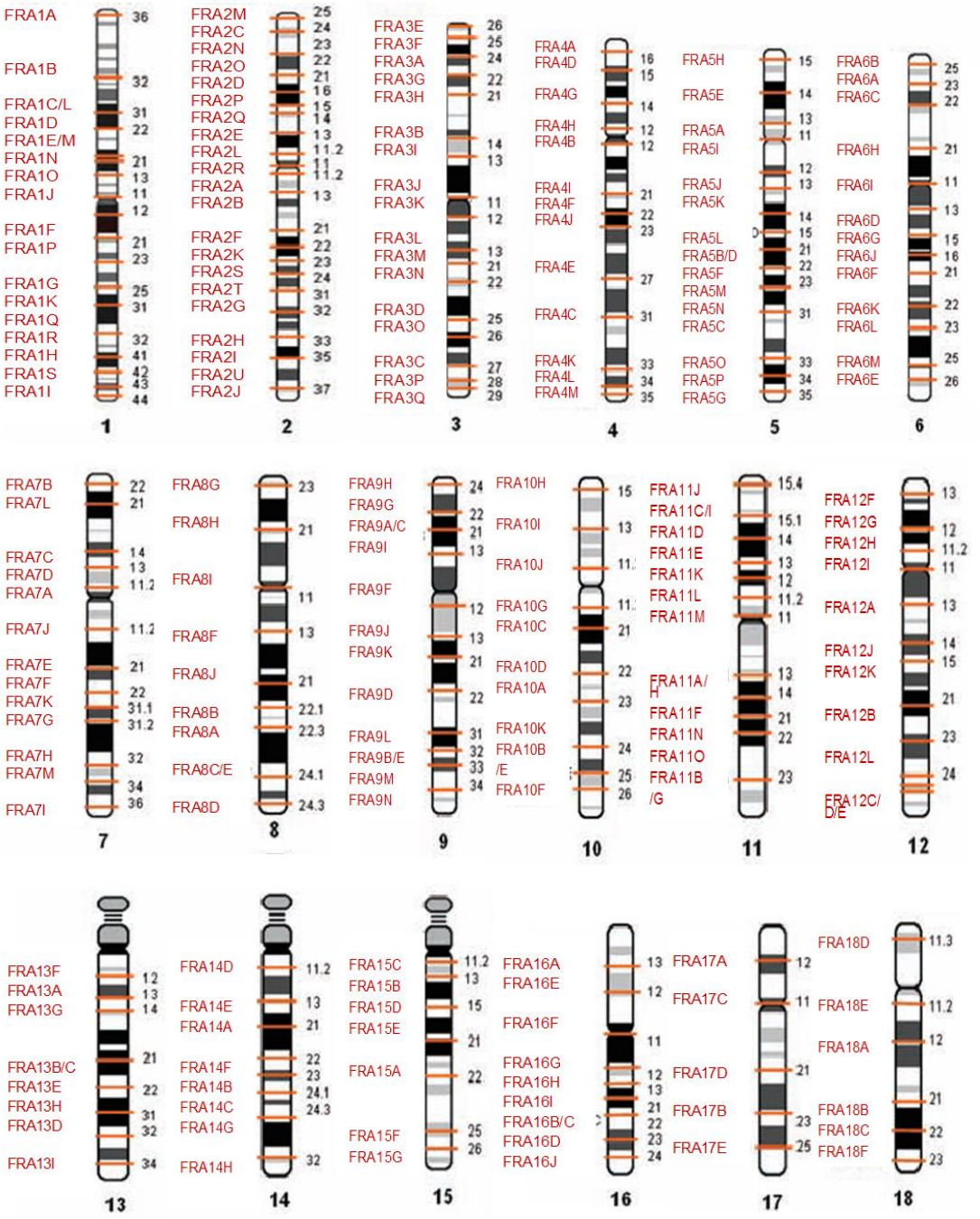
Em termos de replicação, estas bandas replicam mais precocemente do que as bandas *Giemsa* escuras (Cross e Bird, 1995), embora condensem mais tarde, o que pode estar relacionado com o facto de as bandas R possuírem o DNA da cromatina menos condensado (Yokota *et al.*, 1997). O DNA das bandas *Giemsa* claras encontra-se no interior do núcleo (Küpper *et al.*, 2007).

I.3 Sítios Frágeis

Os sítios frágeis cromossómicos foram descritos pela primeira vez em 1965 por Anatole Dekaban (Dekaban, 1965) embora a designação de FS só tenha sido atribuída cinco anos depois por Magenis (Magenis *et al.*, 1970). São *loci* hereditários dos cromossomas humanos que são

susceptíveis à ocorrência de lacunas, quebras ou rearranjos quando sujeitos a condições de *stress* (Büttel *et al.*, 2004; Mrasek *et al.*, 2010) ou quando tratados com agentes químicos específicos (Durkin e Glover, 2007).

Os FSs apresentam uma nomenclatura que é específica para cada cromossoma e que se inicia sempre com a abreviatura FRA seguida do número do respectivo cromossoma e uma letra maiúscula, iniciando-se de A para Z, de acordo com o surgimento da descrição do FS de pter para qter (Mrasek *et al.*, 2010). A nomenclatura e respectiva posição dos FSs utilizadas nesta dissertação baseiam-se em Mrasek *et al.* (2010) (Mrasek *et al.*, 2010) e Lukusa e Fryns (2008) (Lukusa e Fryns, 2008) e apresentam-se esquematizadas na figura I.2.



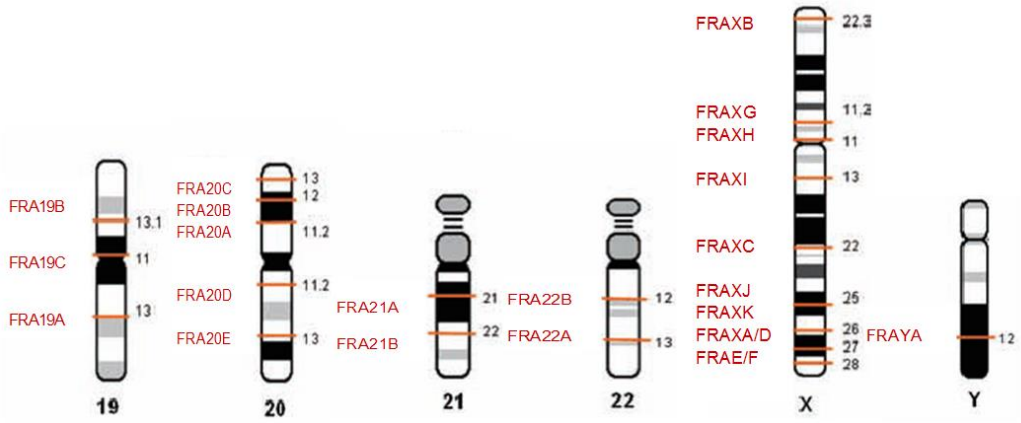


Figura I.2 – Esquema de todos os FSs conhecidos com a respectiva posição em cada cromossoma. Adaptado de Mrasek *et al*., 2010 com a adição de FSs referidos por Lukusa e Fryns, 2008.

Estudos indicam que os FSs podem ter participado em processos evolutivos, uma vez que podem ser conservados entre espécies (Ruiz-Herrera *et al.*, 2002).

Os FSs estão muitas vezes envolvidos em processos que ocorrem no genoma humano, nomeadamente a troca de cromátides irmãs (SCE) (Glover e Stein, 1987; Hirsch, 1991; Durkin e Glover, 2007), deleções e translocações (Glover e Stein, 1988), amplificações de genes intra-cromossomais (Coquelle *et al.*, 1997) e ainda na integração de plasmídeos. Foi também demostrado que os FSs constituem locais preferenciais para a integração de alguns vírus como o vírus do papiloma humano (HPV) 16 e 18 (Matovina *et al.*, 2009), o vírus da hepatite B (Feitelson e Lee, 2007) e o vírus Epstein-Barr (EBV) (Luo *et al.*, 2004).

Os FSs encontram-se distribuídos pela população em diferentes frequências, podendo assim ser divididos em sítios frágeis comuns (CFS) e sítios frágeis raros (RFS) (Durkin e Glover, 2007; Lukusa e Fryns, 2008).

No entanto, existem características que são comuns aos dois tipos de FSs, nomeadamente o facto de ambos possuírem indutores que de uma maneira ou de outra interferem na replicação do DNA, contribuindo para a fragilidade na região. Tanto os CFSs como os RFSs têm a capacidade para formar estruturas secundárias como *harpin*, *stem-loop* e estruturas cruciformes, representadas na figura I.3 que podem interferir com a elongação na replicação (Schwartz *et al.*, 2006). Outra característica comum é o facto de ambos possuírem uma tendência para a organização da cromatina inadequada, uma vez que tanto as repetições CGG como as repetições A/T desfavorecem a montagem dos nucleossomas. Os dois tipos de FS possuem DNA muito flexível. Todas as características comuns aos CFSs e RFSs interferem na elongação aquando da replicação e na estrutura da cromatina. A organização específica da cromatina dos FSs pode promover a formação das estruturas secundárias, retardar a forquilha de replicação ou até mesmo causar a falha na condensação da cromatina (Wang, 2006).

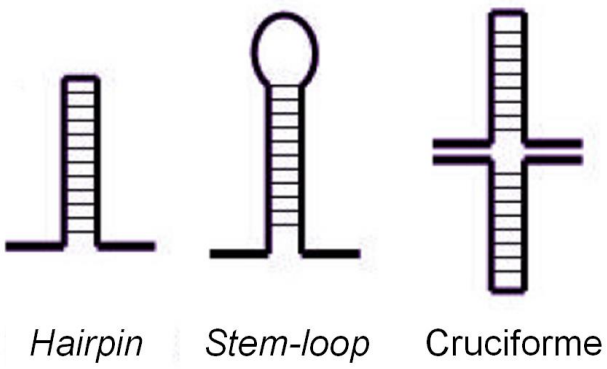


Figura I.3 - Representação das estruturas secundárias formadas pelos CFSs e RFSs. Adaptado de (Freudenreich, 2007).

I.3.1 Sítios Frágeis Comuns

Os CFSs fazem parte da constituição normal dos cromossomas, estando presentes em todos os indivíduos da população (Büttel *et al.*, 2004; Durkin e Glover, 2007; Freudenreich, 2007; Mrasek *et al.*, 2010), podendo ser subdivididos em grupos consoante o agente indutor, nomeadamente a afidicolina (apc), a bromodeoxiuridina (BrdU) e a 5-azacitidina (5-azaC).

A maioria dos CFSs (77) são induzidos por apc, um inibidor da polimerase α e δ. Este agente indutor deve ser utilizado em pequenas concentrações para evitar que ocorra paragem do ciclo celular, existindo apenas inibição parcial da replicação (Lukusa e Fryns, 2008). Depois do tratamento com apc, os CFSs dão origem a uma elevada frequência de deleções e translocações (Glover e Stein, 1988).

Dos 88 CFSs, 7 são induzidos por BrdU, sendo que o tempo óptimo para a exposição ao agente é entre 6 e 12 horas (Sutherland *et al.*, 1985). Este composto é um análogo da base timidina que, durante a síntese do DNA consegue substituir esta base e ser incorporado no DNA sintetizado de novo durante a fase S do ciclo celular (Crane *et al.*, 2011). É também bastante tóxico para as células.

Os 4 CFSs restantes são induzidos por 5-azaC que é um inibidor da metil-transferase do DNA (Büttel *et al.*, 2004). É também um análogo da citosina que consegue ser incorporado na sequência de DNA durante a replicação em substituição da citosina.

Além de todos estes agentes químicos indutores existem alguns factores ambientais que podem contribuir para a instabilidade nos CFSs, tais como o fumo do tabaco (Ban *et al.*, 1965) e a exposição à radiação (Pyatenko *et al.*, 2013).

FRA3B e FRA16D são os dois CFSs mais expressos e aqueles que se encontram melhor caracterizados molecularmente (Durkin e Glover, 2007). Estes dois CFSs são muito importantes uma vez que ambos se encontram em genes supressores de tumores, o FRA3B no gene *FHIT* (*fragile histidine triad*) e o FRA16D no gene *WWOX* (*WW domain-containing oxidoreductase*) (Durkin e Glover, 2007). O primeiro gene encontra-se no braço curto do cromossoma 3 na banda p14.2 (Shi *et*

al., 2000), sendo alvo de rearranjos cromossómicos. Este gene encontra-se muitas vezes inactivado em alguns tumores humanos e está envolvido em deleções que ocorrem na sua região (Pekarsky *et al.*, 2002). O WWOX localiza-se na posição 16q23.3-24.1 e é muito expresso em células epiteliais secretoras de órgãos endócrinos e reprodutores (Yang e Zhang, 2008). A inactivação deste gene pode contribuir para o desenvolvimento de cancro (Abdeen *et al.*, 2011).

Os CFSs são regiões de grande instabilidade, o que pode estar associado a sequências específicas ou à dinâmica de replicação existente nessas regiões (Büttel *et al.*, 2004). Podem existir vários factores que contribuem para essa instabilidade, nomeadamente o facto dos CFSs serem regiões de replicação tardia (Büttel *et al.*, 2004), serem regiões ricas em A/T e possuírem sequências de alta flexibilidade (Zlotorynski *et al.*, 2003) e ainda o facto de estarem envolvidos nas SCE (Ma *et al.*, 2012). Também contribui para a instabilidade dos CFSs o facto de estes estarem localizados em regiões na interface das bandas-G e R uma vez que estas regiões são instáveis e correspondem às regiões de transição no tempo de replicação, como se pode observar na figura I.4 (Watanable *et al.*, 2004). Os CFSs são ainda considerados como sítios preferenciais para a variação estrutural em células estaminais (Hussein *et al.*, 2011). Estando também associados à amplificação de genes durante o desenvolvimento de cancros (Ma *et al.*, 2012); pode assim dizer-se que os CFSs possuem uma relevância médica (Fungtammasan *et al.*, 2012). Recentemente foram ainda publicadas evidências para a predominância de rearranjos cromossómicos constitucionais nos FSs (Sequeira *et al.*, 2013).

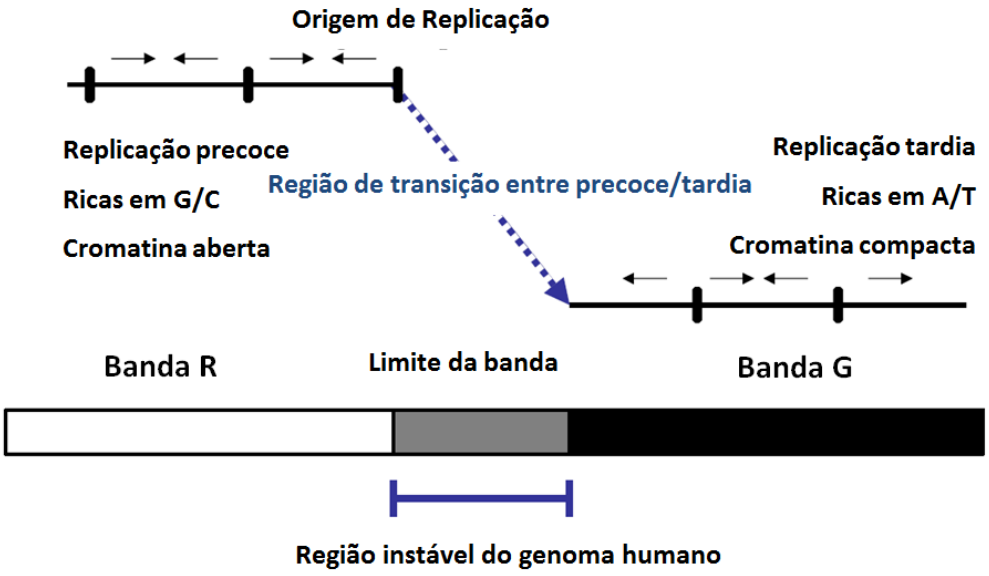


Figura I.4 - Representação da região na interfase das bandas-G e R. Adaptado de (Watanable *et al.*, 2004).

Estas regiões localizadas na interfase das bandas-G e R são regiões do genoma genética e epigeneticamente instáveis, sendo que existe nestas zonas uma correspondência entre as transições no tempo de replicação e a compactação da cromatina (Watanable e Maekawa, 2013). Estas regiões possuem ainda uma tendência para formar estruturas *non-B-DNA*, o que aumenta o risco para a

travagem da forquilha de replicação. Assim a falha no completar da replicação leva à instabilidade genética, o que pode também resultar em doenças humanas.

As células humanas estão sob constantes ameaças, como por exemplo a exposição a luz ultravioleta, a agentes químicos e a produtos do metabolismo celular normal. Para conseguirem sobreviver, as células humanas desenvolveram vários mecanismos e proteínas que controlam a progressão do ciclo celular para que a integridade genómica seja mantida. As células desenvolveram mecanismos de *checkpoint* que podem atrasar a divisão celular caso ocorra algum dano no DNA, para que este seja reparado (Harrison e Haber, 2006). Existem duas cinases, *ataxia-telangiectasia and Rad3-related* (ATR) e *ataxia-telangiectasia mutated* (ATM) que são *checkpoints* para danos no DNA e que funcionam em vias sobrepostas (Ma *et al.*, 2012). ATR tem um papel essencial quando as células são expostas a tratamentos com luz ultravioleta e quando a progressão da forquilha de replicação é bloqueada por agentes, como por exemplo a apc que induz os CFSs (Abraham, 2001). O défice desta cinase leva ao aumento da instabilidade nos FSs após stress replicativo, como o induzido pela apc, o que leva à fragmentação dos cromossomas (Casper *et al.*, 2002). No entanto, foram observados CFSs em células com ausência de ATR sem a indução de apc (Ma *et al.*, 2012). ATM está associada com a quebra da dupla cadeia do DNA (DSB) possuindo também um papel na estabilidade dos FSs na ausência da ATR (Ozeri-Galai *et al.*, 2008; Ma *et al.*, 2012), embora a ausência de ATM sozinha não provoque um aumento de expressão dos CFSs (Ma *et al.*, 2012). Estudos demostraram que em células com ausência de ATR e ATM ocorre um aumento do número de constrições e lacunas comparativamente a células com ausência apenas de ATR (Ozeri-Galai *et al.*, 2008). Assim, apesar da ATR possuir um efeito mais significativo na estabilidade dos FSs, pode afirmar-se que existe um efeito sinérgico de ATR e ATM na regulação da fragilidade dos FSs.

Existem outras proteínas que são também *checkpoints* do ciclo celular e que têm um papel na manutenção da estabilidade dos FSs como por exemplo a BRCA1, a SMC1 e a FANCD2 (Arlt *et al.*, 2006).

A formação de DSBs pode estar relacionada com a indução dos CFSs por apc, pois pode levar ao colapso da forquilha de replicação. Por outro lado as DSBs estão envolvidas na indução de lacunas nos FSs sob *stress* replicativo (Schwartz *et al.*, 2005). Este tipo de dano pode ser considerado o mais perigoso para a integridade do genoma, uma vez que a falha na sua reparação pode levar à morte celular ou a rearranjos cromossómicos (Khanna e Jackson, 2001). Para reparar as DSBs existem duas vias principais: a NHEJ e a HR, as quais também estão envolvidas na manutenção da estabilidade dos FSs (Schwartz *et al.*, 2005; Ozeri-Galai *et al.*, 2008). A NHEJ é um mecanismo de recombinação não homóloga que pode ocorrer em qualquer fase do ciclo celular, tendo especial importância na fase G1 e é dos mecanismos primários a ser utilizado para resolver DSBs, não requerendo nenhuma arquitectura genómica específica (Vissers e Stankiewicz, 2012), enquanto que a HR requer a presença de cromátides ou cromossomas homólogos.

Dos mecanismos de reparação mediados por homologias nas junções que se relacionam com os rearranjos que ocorrem nos CFSs, o mecanismo mediado pelas microhomologias é o mais frequentemente encontrado, como mostra a figura I.5. Os processos de rearranjos mais associados às microhomologias são a NHEJ, a *microhomology-mediated end joining* (MMEJ), as

microhomologies-mediated break-induced replication (MMBIR) e as *fork stalling and template switching* (FoSTeS) (Mitsui *et al.*, 2010). Por outro lado, os mecanismos de recombinação homóloga não alélica (NAHR) associados às junções com homologia, representado a verde na figura I.5 são os menos frequentes.

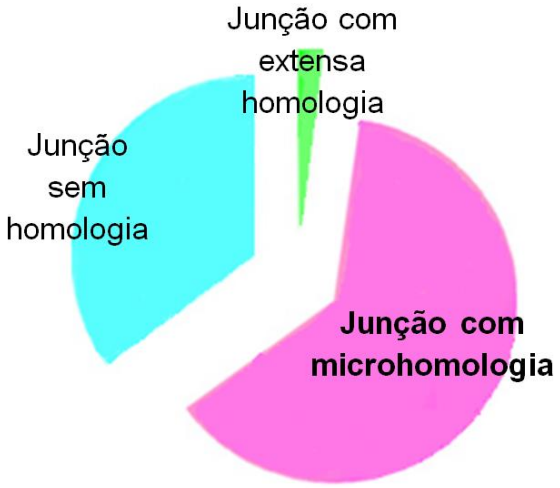


Figura I.5 – Representação esquemática da predominância do mecanismo de reparação mediado por microhomologias envolvido nos processos de rearranjos que ocorrem nos CFSs. Adaptado de Mitsui *et al*., 2010.

Continuam a surgir novos estudos em volta dos CFSs, estando os respectivos genes a surgir como potenciais biomarcadores com valor na previsão de resposta a quimioterapia a qual causa danos no DNA em tecidos saudáveis e tumores (Ma *et al.*, 2012).

I.3.2 Sítios Frágeis Raros

Os RFSs estão presentes numa pequena porção da população, sendo a sua percentagem inferior a 5% (Schwartz *et al.*, 2006; Mitsui e Tsuji, 2011), podendo por isso ser utilizados em estudos genéticos familiares (Sutherland, 1988). Tal como os CFSs, também os RFSs são classificados consoante o modo como são induzidos. Assim, podemos distingui-los em dois grandes grupos, os sensíveis ao folato e os não sensíveis ao folato (Lukusa e Fryns, 2008).

Os RFSs sensíveis ao folato constituem a maioria dos RFSs e a sua indução pode ser realizada de diferentes maneiras, adicionando ou retirando componentes ao meio de cultura. Podem ser induzidos em meios de cultura que apresentem um défice em ácido fólico e T ou em meios enriquecidos em fluorodeoxiuridina ou ainda em meios enriquecidos em metotrexato (Lukusa e Fryns, 2008), sendo que todos resultam na síntese perturbada de DNA. Estes RFSs são ainda caracterizados por possuírem expansões da repetição microssatélite CGG (Schwartz *et al.*, 2006) que têm a capacidade para formar estruturas secundárias estáveis, o que depende do comprimento e da pureza da sequência (Schwartz *et al.*, 2006). A formação destas estruturas secundárias pode perturbar a elongação na replicação do DNA e consequentemente estar envolvida na estabilidade da

região. As repetições CGG apresentam ainda uma baixa eficiência na montagem dos nucleossomas, o que pode levar à observação de constrições nestes FSs.

Estas expansões dos microssatélites podem levar a atrasos mentais associados aos RFSs FRAXA e FRAXE. O FRAXA encontra-se localizado em Xq27.3 e está associado ao síndrome do X frágil, que é um dos atrasos mentais hereditários mais graves, enquanto que o FRAXE está localizado em Xq28 estando associado a um atraso mental não específico ligado ao X (Knight *et al.*, 1993). Além das expansões de microssatélites, há evidências *in vivo* da ocorrência de quebras cromossómicas nos FSs, o que foi documentado pela observação de deleções em alguns doentes com síndrome de X frágil e com síndrome de Jacobsen (Gedeon *et al.*, 1992; Wöhrle *et al.*, 1992; Jones *et al.*, 1995). Estes dados corroboram as observações *in vitro* de que a seguir à indução da expressão do FS, uma proporção de células têm duplicações e deleções do material genético distal ao FS (Sutherland e Baker, 2000).

Os RFSs não sensíveis ao folato podem ainda ser subdivididos em dois grupos, os induzidos por distamicina A e compostos relacionados, berenil, netropsina e Hoechst33258 e os induzidos por BrdU (Lukusa e Fryns, 2008). A distamicina A e compostos relacionados impedem a replicação do DNA ligando-se a este com uma elevada afinidade para o pequeno sulco de sequências ricas em A/T (Abu-Daya *et al.*, 1995).

Tanto os RFSs sensíveis ao folato como os não sensíveis ao folato apresentam expansões de repetições, o que mostra que existem polimorfismos no número de cópias na população (Schwartz *et al.*, 2006).

I.3.3 Sítios Frágeis de Replicação Precoce

Os sítios frágeis de replicação precoce (ERFSs), foram descobertos recentemente, tendo sido classificados como uma nova classe de FSs, estando presentes nas células dos mamíferos e contribuindo para rearranjos recorrentes durante a formação de linfomas (Barlow *et al.*, 2013). Apesar deste novo tipo de FS apresentar algumas características distintas dos CFSs, estando algumas delas esquematizadas na figura I.6, destacando-se as diferenças no conteúdo e o diferente estado da cromatina (Mortusewicz *et al.*, 2013), apresentam também algumas semelhanças. Tanto a fragilidade dos CFSs como a dos ERFSs é aumentada por inibição da ATR, por stress oncogénico e também por deficiência em HR (Barlow *et al.*, 2013).

Os ERFSs possuem algumas características semelhantes às características das bandas *Giemsa* claras, como o facto de possuírem um conteúdo rico em G/C, serem regiões de replicação precoce ou ainda o facto de possuírem uma configuração de cromatina aberta. Sabe-se igualmente que os ERFSs estão enriquecidos em regiões que são transcripcionalmente activas, podendo ser a fonte de muitos rearranjos encontrados em cancro (Barlow *et al.*, 2013).

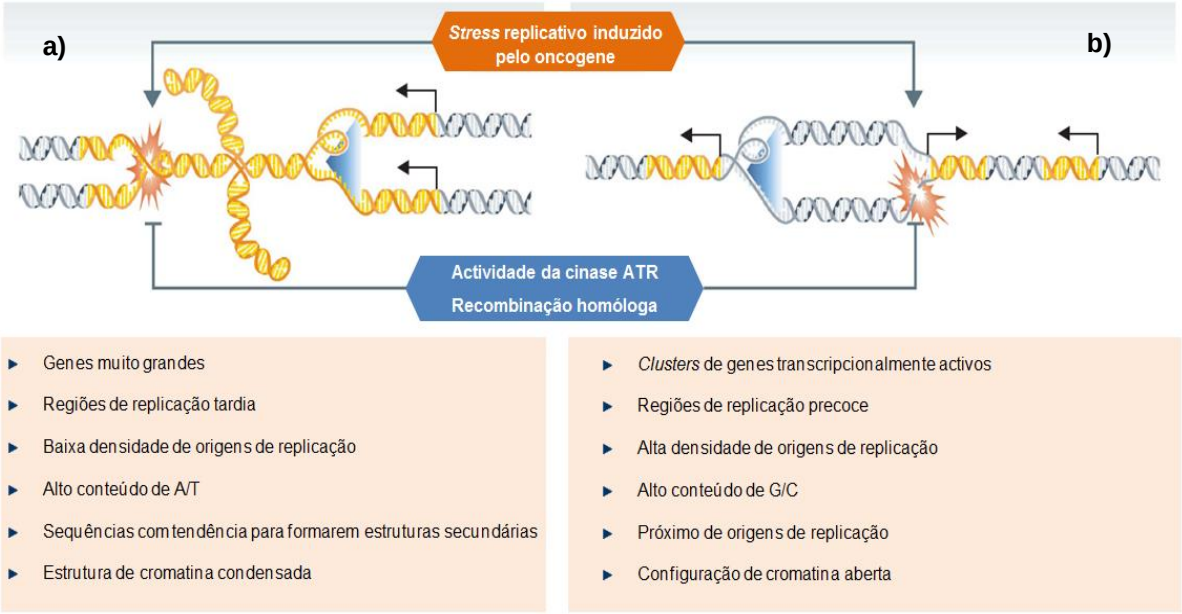


Figura I.6 – Representação esquemática dos ERFSs comparativamente aos CFSs, onde está presente um sumário das principais diferenças entre os dois tipos de FSs. a) Corresponde aos CFSs. b) Corresponde aos ERFSs. Adaptado de Mortusewicz *et al*., 2013.

Barlow *et al*. (2013) (Barlow *et al.*, 2013) realizaram estudos em ratinhos, tendo já conseguido obter as localizações de alguns *loci* humanos associados com os ERFSs em células B de linfomas, sendo que os *loci* são enriquecidos de elementos repetitivos como transposões.

I.4 Vírus da Imunodeficiência Humana

O Vírus da Imunodeficiência Humana (HIV) pertence à família dos retrovírus e ao género lentivírus, os quais são caracterizados por um longo período de incubação. O HIV desenvolveu-se a partir do vírus da imunodeficiência dos símios (SIV), encontrado em primatas da África, tendo este cruzado a barreira das espécies e passado dos macacos para os humanos (Worobey *et al.*, 2008).

O primeiro vírus foi isolado em 1983 por Barré-Sinoussi *et al*. a partir do sangue de um doente (Barré-Sinoussi *et al.*, 1983). Cerca de três anos depois, foi isolado o segundo vírus causador do síndrome da imunodeficiência adquirida (AIDS) por uma equipa que integrou a investigadora portuguesa Odete Santos Ferreira (Clavel *et al.*, 1986a). Assim, dentro do HIV podem distinguir-se dois tipos diferentes de vírus, o vírus da imunodeficiência humana tipo 1 (HIV-1) e o vírus da imunodeficiência humana tipo 2 (HIV-2).

Os retrovírus são vírus cujo genoma consiste em duas cópias de cadeia simples de RNA e que possuem um envelope (Smith e Daniel, 2006), sendo que para se replicarem necessitam que ocorra uma integração estável do seu material genético no genoma do hospedeiro (Turlure *et al.*, 2004). Este vírus consegue infectar células que não estão em divisão (Craigie e Bushman, 2012),

sendo que tem de assegurar que a célula sobrevive o tempo necessário para maximizar a produção de outros vírus (Lilley *et al.*, 2007).

I.4.1 Ciclo de Vida do Vírus

Para que ocorra integração o vírus necessita ligar-se a receptores que existem na membrana da célula hospedeira, nomeadamente os CD4 dos linfócitos T (Vilanova e Ferreira, 2007). São ainda células alvo os macrófagos, os monócitos, as células dendríticas, entre outras. Assim, ocorre a fusão entre a membrana do vírus e a membrana celular, permitindo que ocorra libertação do núcleo do vírus no citoplasma da célula hospedeira (Ciuffi e Bushman, 2006), como se encontra representado na figura I.7. O retrovírus utiliza a maquinaria de transcrição do hospedeiro para completar o seu ciclo de vida.

Uma vez no citoplasma da célula hospedeira, forma-se um complexo de pré-integração (PIC) dentro do qual o genoma RNA do vírus é convertido em DNA (DNA provírus) pela enzima transcriptase reversa (Herschhorn e Hizi, 2010), para que possa depois ser transportado para o interior do núcleo através de um poro nuclear, como demostrado na figura I.7.

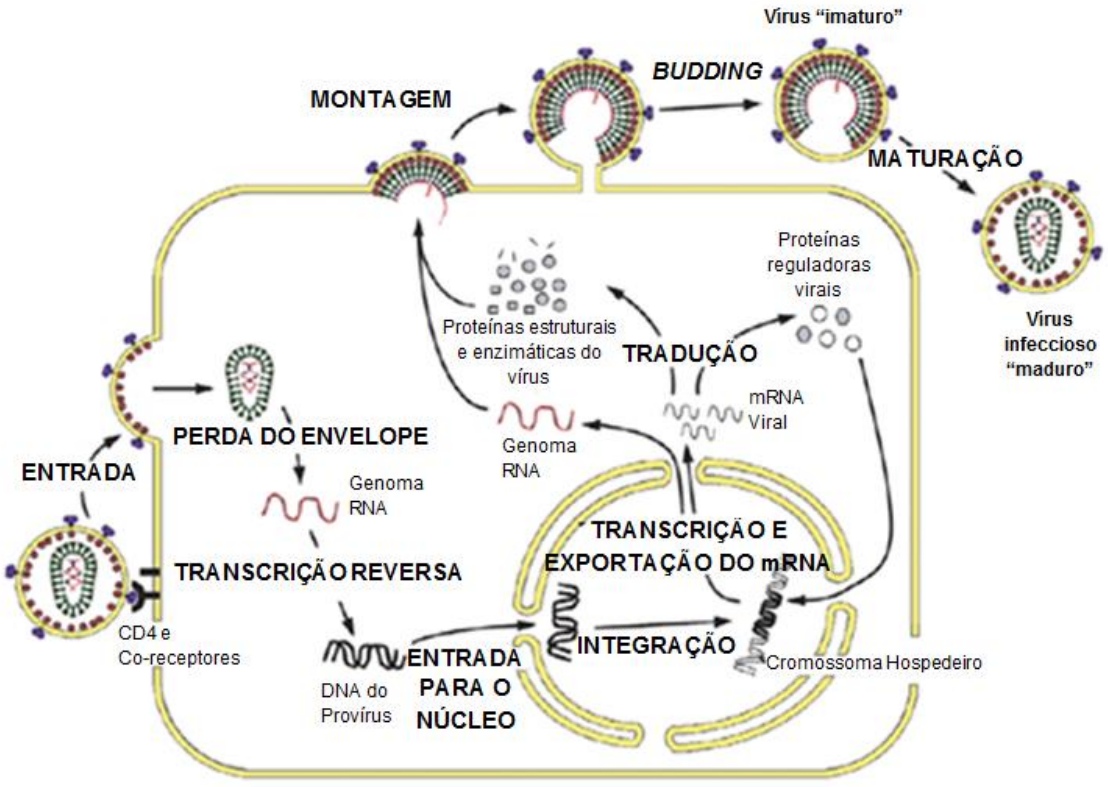


Figura I.7 – Esquema ilustrativo do ciclo de replicação do vírus. Adaptado de (Ganser-Pornillos *et al.*, 2008).

Após a entrada no núcleo da célula hospedeira, ocorre a integração do genoma do vírus, que decorre em três passos, os quais aparecem representados na figura I.8. Primeiramente ocorre o

processing, no qual a enzima integrase (IN) remove dois nucleótidos do extremo 3´do DNA viral (Craigie e Bushman, 2012). Segue-se o *joining* onde a IN catalisa uma reacção na qual o extremo 3´do DNA viral é unido ao DNA da célula hospedeira (Daniel e Smith, 2008). O último passo da integração é designado por *postintegration repair* onde é feito um corte na extremidade 5´do DNA viral, havendo ligação das sequências da célula hospedeira a esse extremo 5´ e depois uma reconstituição da estrutura da cromatina no sítio de integração, sendo necessária a actuação das proteínas de reparação do DNA da célula hospedeira (Skalka e Katz, 2005).

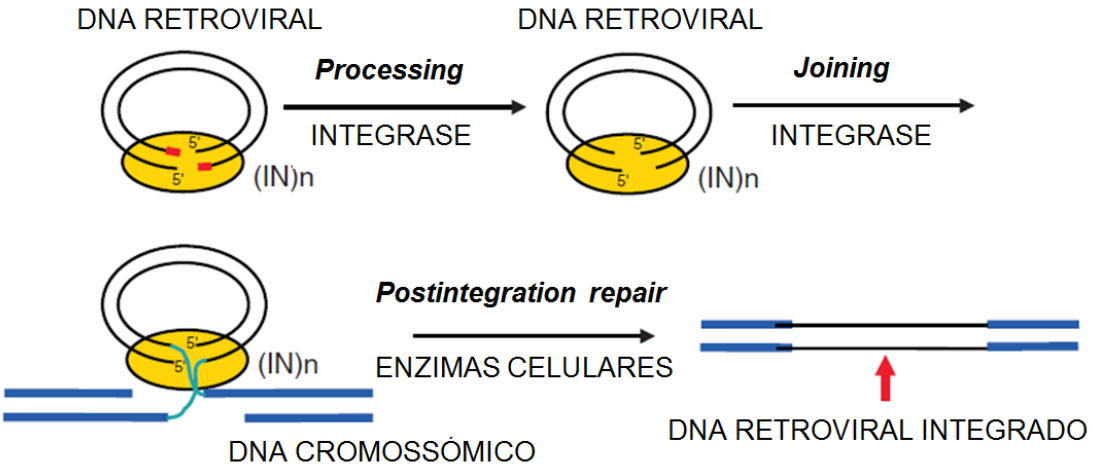


Figura I.8 – Esquema da integração do retrovírus no DNA cromossomal. Adaptado de (Daniel e Smith, 2008).

Após a integração, ocorre a transcrição no núcleo da célula hospedeira seguida da tradução no citoplasma, ficando disponíveis todos os componentes necessários para a montagem de um novo vírus. Ocorre o empacotamento de todos os componentes do vírus e este sai da célula hospedeira por *Budding* arrastando consigo um pedaço da membrana da célula que irá ser incorporado no seu envelope (Turlure *et al.*, 2004). O vírus vai depois infectar novas células, iniciando um novo ciclo viral.

Quando integra o seu genoma, o retrovírus insere promotores virais perto dos genes do hospedeiro, o que pode afectar a expressão e consequentemente a função desses mesmos genes na célula hospedeira (Ciuffi e Bushman, 2006). Quando ocorre a integração do vírus, a célula assume-a como um evento de DSBs, activando assim os mecanismos de reparação do DNA para reparar a quebra provocada pelo vírus (Smith e Daniel, 2006). Se as células infectadas não possuírem os componentes dos mecanismos de reparação de DSBs, o passo de *postintegration repair* falha, podendo levar à morte da célula. Durante a integração, o vírus utiliza proteínas do hospedeiro que podem funcionar como nuclease, polimerase e ligase, estar relacionadas com o rearranjo da estrutura da cromatina e com a indução dos *checkpoints* do ciclo celular (Smith e Daniel, 2006).

Foi realizado um estudo com siRNAs que mostrou que a via de reparação de DNA por excisão de pares de bases (BER) é a mais envolvida na replicação do HIV, sendo que a via de NHEJ, foi das que apresentou a frequência mais baixa, como se pode observar na figura I.9. Curiosamente, a via de NHEJ é frequentemente relacionada com a replicação do HIV (Li *et al.*, 2001; Daniel *et al.*, 2004).

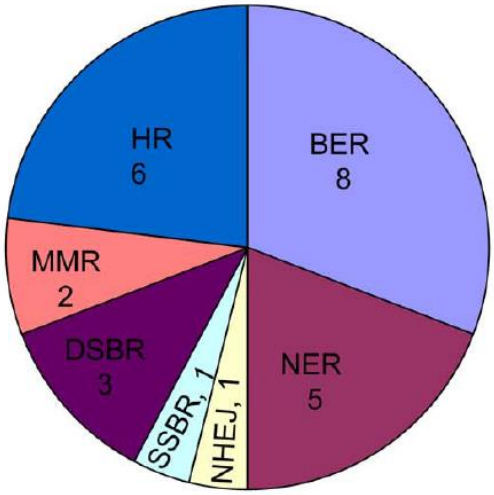


Figura I.9 – Representação das vias de reparação do DNA conseguidas através do estudo de siRNAs. HR refere-se a recombinação homóloga, BER a reparação por excisão de pares de bases, NER a reparação por excisão de nucleótidos, DSBR a reparação da quebra de dupla cadeia, MMR a reparação por *mismatch*, SSBR a reparação da quebra em cadeia simples e NHEJ a recombinação *nonhomologous end-joining*. Adaptado de (Espeseth *et al.*, 2011).

I.4.2 HIV-1 e HIV-2

O genoma do HIV-1 consiste de duas cópias de RNA, cada uma com cerca de 9.2 Kilobases (Kb) de tamanho (Layne *et al.*, 1992). Este genoma codifica 16 proteínas, sendo que os componentes essenciais do vírus incluem glicoproteínas do envelope e da cápside e as enzimas transcriptase reversa, IN e protease (Engelman e Cherepanov, 2012).

A transcriptase reversa é a enzima que converte o genoma RNA do vírus em DNA e a IN permite a integração do vírus no genoma humano, tal como descrito na seção I.4.1. A protease do HIV faz parte da família das proteases aspárticas (Fun *et al.*, 2012) e intervém na fase final do ciclo de vida do vírus, permitindo que as partículas virais imaturas passem a ser vírus infectantes por vias proteolíticas (Petit *et al.*, 1994).

O HIV-1 codifica ainda uma proteína acessória R do vírus (Vpr) de 14 KDa que possui um papel importante na infecção sendo, por exemplo, algumas das suas funções conhecidas, a modulação da transcrição do genoma do vírus (Sawaya *et al.*, 2000), a facilitação da transcrição reversa, a indução de defeitos na mitose e a disrupção do ciclo celular (Chang *et al.*, 2004). A Vpr pode levar à formação de quebras do genoma do hospedeiro e consequentemente à integração do vírus. A Vpr do HIV-1 possui ainda a capacidade para promover a apoptose durante a infecção do HIV-1, enquanto que a Vpr do HIV-2 não induz apoptose em linhas celulares (Romani e Engelbrecht, 2009). Por causar DSBs, esta proteína viral faz aumentar a frequência de HR (Nakai-Murakami *et al.*, 2007). Agentes químicos, como a apc podem também gerar DSBs, o que faz aumentar a taxa de integração do vírus no genoma hospedeiro (Groschel e Bushman, 2005).

O HIV-1 possui um *lens epithelium-derived growth factor* (LEDGF)/p75 que possui a habilidade de se ligar fortemente à IN (Turlure *et al.*, 2004) e de a proteger contra a proteólise que poderá ocorrer *in vivo*.

O genoma do HIV-2 possui 9.5 Kb de tamanho (Clavel *et al.*, 1986b), existindo diferenças substanciais nas sequências genómicas dos dois vírus, sendo que as proteínas codificadas são também diferentes (Clavel *et al.*, 1986a). O HIV-2 possui não só a Vpr como também outra proteína viral designada de proteína viral X (Vpx). A Vpx tem como principais funções reduzir as sobreposições funcionais no vírus e afastar alvos não funcionais que possam ser prejudiciais para o vírus (Casey *et al.*, 2010). A Vpr do HIV-2, tal com a do HIV-1 também está envolvida na paragem do ciclo celular na fase G2, embora não esteja relacionada com a passagem do PIC para o interior do núcleo (Fletcher *et al.*, 1996).

O HIV-2 apresenta uma virulência e uma capacidade de infecção menor do que o HIV-1.

I.4.3 Vectores derivados do HIV

Os vectores derivados do HIV são um dos veículos mais utilizados em terapia génica para o transporte de genes, uma vez que conseguem estabilizar e inserir com precisão novas sequências de DNA em vários tipos celulares (Ciuffi e Bushman, 2006) e conseguem ter um efeito terapêutico de longa duração (Edelstein *et al.*, 2007). É importante conhecer bem os alvos e preferências de integração dos retrovírus, uma vez que têm sido associados efeitos adversos à integração de vectores de retrovírus próximos de proto-oncogenes (Hacein-Bey-Abina *et al.*, 2010).

Há possibilidade de interacção entre os vectores usados na terapia génica e o lentivírus *wild-type* obtido por infecção natural (Dirac *et al.*, 2002).

O método mais comum para se obter um vector a partir de retrovírus potencialmente patogénicos é delectando os genes virais e fazendo uma substituição destes por unidade de expressão. Assim aos vectores irá faltar a maioria das proteínas codificantes, mas tendo sempre o sinal de empacotamento e as sequências terminais repetidas (LTRs) que são necessárias para a integração do DNA (Nagel *et al.*, 2012).

I.5 Interacção do HIV com o Genoma Humano

A integração do vírus no genoma humano causa alterações ao nível molecular e epigenético, o que afecta a expressão de genes, podendo levar à activação de oncogenes (Ciuffi e Bushman, 2006). O HIV, após a sua entrada na célula, interfere ainda com a síntese e função de proteínas celulares normais, podendo mesmo levar à lise das células infectadas (Vilanova e Ferreira, 2007). No momento da entrada, o hospedeiro desencadeia uma resposta imune contra o vírus para bloquear a infecção e até mesmo eliminar as células infectadas para impedir a propagação do vírus.

Dentro do genoma humano, existem determinadas características que levam o vírus a escolher o sítio de integração que mais favorece a sua replicação e infecção. Existem estudos que relacionam os locais alvos de integração com os locais mais ricos em genes que podem ser importantes para a expressão eficiente do genoma do vírus (Schröder *et al.*, 2002), sendo que dentro dos genes a integração é feita preferencialmente nos intrões devido ao seu tamanho. A transcrição está também relacionada com a integração do vírus, sendo que o vírus integra mais em locais activos transcripcionalmente, já que estes locais aumentam a acessibilidade à cromatina (Schröder *et al.*, 2002). Zonas em que a cromatina se encontra mais aberta, ou seja, menos condensada também são alvos preferenciais para a integração do vírus.

A integração estável do vírus pode levar à modulação da estrutura da cromatina no genoma humano levando à ocorrência de alterações na organização da estrutura da cromatina.

Os provírus retrovirais estão muitas vezes relacionados com os transposões eucarióticos, sendo estes elementos designados de retrotransposões. Os retrotransposões não possuem actividade de transposões, mas no entanto apresentam sequências que são reconhecidas como substrato para a transposição pelos elementos activos (Lewin, 2004).

I.6 Objectivos

I.6.1 Objectivo Principal

O presente trabalho teve como objectivo principal estudar as características do genoma humano associadas à integração do HIV, incidindo especificamente nas bandas *Giemsa* e nos FSs. As bandas *Giemsa* claras são ricas em genes e zonas transcricionalmente muito activas, o que suscita a questão se serão sítios preferenciais de integração do HIV, visto ser um vírus com preferência de integração em zonas com estas características. Em relação aos FSs, existem alguns vírus como o HPV16, o vírus Adeno-associado (AAV) e o EBV que têm uma preferência de integração nestas regiões do genoma, facto que não está claro para o HIV.

Este trabalho teve por base análises bioinformáticas e estatísticas para se inferir sobre as preferências de integração do vírus.

I.6.2 Objectivos Específicos

Com este estudo, pretendeu-se elucidar as preferências de integração do HIV-1 DNA e HIV-2 DNA isolados de células mononucleadas do sangue periférico (PBMCs) nas regiões referidas na secção I.6.1.

Outro dos objectivos deste trabalho foi estudar um vasto conjunto de sítios de integração de HIV-1 DNA isolado de células T Jurkat e aferir sobre as suas preferências de integração também nas regiões referidas na secção I.6.1.

II. Materiais e Métodos

II.1 Obtenção dos Dados

Para a realização deste trabalho recorreu-se a bases de dados, tendo sido utilizados dados sobre sítios de integração de HIV-1 DNA e HIV-2 DNA isolados a partir de células mononucleadas do sangue periférico (PBMCs) e de HIV-1 DNA isolados a partir de células T Jurkat. Os dados foram originalmente obtidos conforme descrito nas secções II.1.1, II.1.2 e II.1.3.

Nesta tese, referimo-nos frequentemente a HIV-1 DNA e HIV-2 DNA pretendendo abranger os respectivos sítios de integração.

II.1.1 HIV-1 DNA Isolado de PBMCs

Os sítios de integração do HIV-1 DNA foram obtidos por Mitchell *et al.* (2004) (Mitchell *et al.*, 2004) a partir de PBMCs. Inicialmente os autores separaram PBMCs de sangue humano usando um gradiente de ficol. Depois, PBMCs pré-estimuladas com fitohemaglutinina e Interleucina 2 (IL-2) foram infectadas com vectores baseados em HIV-1. As células foram analisadas por citometria de fluxo para assim se determinar a dimensão da infecção. O DNA das células infectadas foi depois isolado, clivado com enzimas de restrição que não cortavam dentro do genoma do vector e ligado a *linkers* de DNA, para isolarem os sítios de integração. De seguida, estes sítios foram amplificados usando dois *primers* diferentes, um que se ligava à extremidade final do DNA viral e outro que se ligava ao *linker*. Os produtos de amplificação foram clonados por PCR mediado por ligação e sequenciados. Finalmente, os sítios de integração foram mapeados na sequência do genoma humano e as características locais dos sítios de integração foram quantificadas. Todas as novas sequências de sítios de integração foram depositadas no *National Center for Biotechnology Information* (NCBI).

II.1.2 HIV-2 DNA Isolado de PBMCs

Para o HIV-2 DNA foram utilizados sítios de integração também isolados a partir de PBMCs obtidos por MacNeil *et al.* (2006) (MacNeil *et al.*, 2006). Inicialmente, os autores infectaram PBMCs com um isolado primário de HIV-2. A estirpe utilizada de HIV-2 foi previamente isolada de uma mulher (Kanki *et al.*, 1992). Foram utilizadas PBMCs de um dador normal, para a infecção, sendo que as células foram separadas e estimuladas durante 72 horas com fitohemaglutinina em meio completo com IL-2. Depois, foi removida a fitohemaglutinina e as células foram infectadas com a estirpe de HIV-2. Passadas 24 horas após infecção, e para remover sobrenadantes residuais de vírus, as células foram lavadas e resuspendidas em meio fresco com IL-2. Para sequenciarem os sítios de integração do HIV-2, criaram uma biblioteca por *linker-mediated nested PCR*, contendo esta

fragmentos de DNA que continham a extremidade final 5´da LTR do HIV integrado e, a montante DNA do genoma humano correspondente à integração no local da junção. Por fim, utilizaram o programa *BLAST-like alignment tool* (BLAT) para mapear os sítios de integração, tendo colocado os dados obtidos no NCBI.

II.1.3 HIV-1 DNA Isolado de células T Jurkat

Estes sítios de integração obtidos por Wang *et al.* (2007) (Wang *et al.*, 2007) foram isolados a partir de células T Jurkat. Para isolarem as amostras, os autores começaram por incubar as células T Jurkat com vectores baseados em HIV, usando seguidamente PCR mediado por ligação para prepararem fragmentos de DNA da junção hospedeiro-vírus. Para clivar o genoma humano foi utilizada uma mistura de duas enzimas de restrição, uma que usava MseI que reconhece um local de quatro bases e outra que usava um *pool* de enzimas, nomeadamente AvrII, SpeI e NheI, que reconhecem sequências de seis bases. O DNA celular digerido foi ligado a *linkers*, sendo amplificada a junção entre o DNA celular e viral. Os autores utilizaram a técnica de massively parallel pyrosequencing, da qual depois de um controlo de qualidade resultaram os sítios únicos no genoma humano utilizados nesta tese, tendo ao autores fornecido para o estudo a posição do sítio de integração. Os dados estão disponíveis numa base de dados *online* da *UCSC Genome Bioinformatics*.

II.1.4 Obtenção das posições de integração do HIV-1 DNA e HIV-2 DNA isolados de PBMCs

Para a realização deste trabalho, e para a obtenção das sequências dos sítios de integração do HIV-1 DNA e HIV-2 DNA isolados de PBMCs, baseámo-nos em Soto *et al.* (2011) (Soto *et al.*, 2011). Estes autores, utilizaram um conjunto de sequências de sítios de integração para o HIV-1 DNA obtidas por Mitchell *et al.* (2004) (Mitchell *et al.*, 2004) e para o HIV-2 DNA obtidas por MacNeil *et al.* (2006) (MacNeil *et al.*, 2006). A esse conjunto de sequências, Soto *et al.* (2011) aplicaram determinados critérios de selecção, obtendo as sequências de integração utilizadas nesta tese. Para serem seleccionadas e consideradas sítios de integração autênticos, as sequências tinham de corresponder aos seguintes critérios, conter o terminal 3´ LTR do HIV-1 DNA ou HIV-2 DNA; possuir correspondência com o DNA genómico dentro de 5 bp do fim da LTR viral; possuir pelo menos 95% de homologia com as sequências do genoma humano em toda a região sequenciada; possuir correspondência com um único *locus* genético humano com pelo menos 95% de homologia em toda a região sequenciada e possuir um tamanho mínimo de 50 bp. Para a descriminação destes critérios, os autores tiveram em consideração os critérios de MacNeil *et al* (2006) (MacNeil *et al.*, 2006).

Assim, colocámos os números de acesso fornecidos por Soto *et al.* (2001) (Soto *et al.*, 2011)

no *Basic Local Alignment Tool* (BLAST) de forma a alinhar e confirmar as sequências obtendo as respectivas posições iniciais e finais dos sítios de integração do HIV. Estas posições foram colocadas em tabelas e ordenadas pelo cromossoma correspondente. Na figura II.1 encontra-se um exemplo de um alinhamento obtido no BLAST com o número de acesso CL529704 correspondente ao HIV-1 DNA isolado de PBMCs.

Figura II.1 – Exemplo de um alinhamento obtido no BLAST para o HIV-1 DNA isolado de PBMCs. Do alinhamento, podemos retirar o cromossoma a que corresponde (7) e a posição inicial (1137380) e final (1137878) da sequência do sítio de integração.

II.2 Bandas *Giemsa*

Este estudo foi realizado com as bandas *Giemsa* obtidas *in silico* por Niimura e Gojobori (2002) (Niimura e Gojobori, 2002), tendo os autores disponibilizado as posições finais e iniciais de cada banda Giemsa escura por cromossoma. Para conseguirem obter as sequências das bandas *Giemsa in silico*, os autores obtiveram as sequências do DNA do genoma humano da base de dados *UCSC Genome Bioinformatics* e definiram as posições relativas de cada limite entre as bandas *Giemsa* vizinhas relativamente à porção da eucromatina total de cada braço cromossómico a partir do

estudo realizado por Francke (1994) (Francke, 1994). De seguida, basearam-se no método computadorizado descrito por Needleman e Wunsch (1970) (Needleman e Wunsch, 1970) para obterem uma pontuação de similaridade entre as bandas *Giemsa* e bandas *in silico* para conseguirem assim determinar o alinhamento óptimo das bandas *Giemsa*. Após esse alinhamento, os autores aplicaram testes estatísticos que lhes permitiram avaliar o significado estatístico das semelhanças entre as bandas *Giemsa* e as bandas *in silico* e concluir que os padrões das bandas *Giemsa* foram reconstruídos com sucesso pelo tratamento *in silico* (Niimura e Gojobori, 2002). Utilizaram ainda vários programas computacionais para detectarem regiões onde o conteúdo em G/C era mais baixo do que as regiões flanqueadoras, conseguindo obter um diagrama com as várias bandas representado na figura I.1. A tabela com as posições iniciais e finais das bandas *in silico* está apresentada no anexo.

II.2.1 HIV-1 DNA e HIV-2 DNA isolado de PBMCs

O tratamento dos dados foi realizado de igual forma para HIV-1 DNA e HIV-2 DNA, embora a análise tenha sido feita em separado para cada HIV. Para o HIV-1 DNA possuíamos 140 sítios de integração, enquanto que para o HIV-2 DNA possuíamos 132.

Tendo a posição inicial e final de cada sítio de integração de HIV-1 DNA e HIV-2 DNA para cada cromossoma e a posição inicial e final de cada banda *Giemsa* escura, elaboramos tabelas para verificar se existia ou não co-localização dos sítios de integração do HIV com as bandas escuras por cromossoma. Deste modo, as integrações virais foram classificadas em dois grupos: *sim* se existisse co-localização com as bandas *Giemsa* escuras e *não* se não existisse. Existiam também alguns sítios de integração que co-localizavam com os centrómeros ou braços curtos dos acrocêntricos, tendo sido estes retirados do total das integrações e analisados de uma forma diferente, calculando apenas a frequência. Assim, dos 140 sítios de integração iniciais para o HIV-1 DNA ficámos com 130, tendo dados para 21 cromossomas, uma vez que no cromossoma 15 e 21 as integrações ocorreram todas nos centrómeros. Para o HIV-2 DNA, dos 132 sítios de integração iniciais ficámos com 120, tendo ficado com dados para os 23 cromossomas, perdendo-se apenas integrações dentro de alguns cromossomas.

Com o objectivo de saber em que tipo de banda o vírus integrava com maior frequência, calculámos duas medidas de integração diferentes, que designámos por *rácio em extensão* e *intensidade em número*. Foi necessário calcular estas medidas porque as bandas possuíam comprimentos diferentes e, comparar simplesmente o número de integrações virais em cada banda induziria a resultados enganadores. O *rácio em extensão* permite ter em consideração o comprimento do sítio de integração viral e o tamanho total da banda, sendo que o resultado é dado pelas seguintes fórmulas:

$$r_{banda\ escura} = \frac{l_{sim}}{l_{banda\ escura}} \quad ; \quad r_{banda\ clara} = \frac{l_{não}}{l_{banda\ clara}},$$

onde l_{sim} é o comprimento do sítio de integração viral nas bandas escuras, $l_{banda\ escura}$ é o tamanho das bandas escuras, $l_{não}$ é o comprimento do sítio de integração viral nas bandas claras e $l_{banda\ clara}$ é o comprimento das bandas claras.

A outra medida de integração utilizada, a *intensidade em número* é definida como a frequência de integração que ocorre em cada banda ponderada pelo comprimento da mesma. Esta intensidade é calculada segundo as fórmulas seguintes:

$$i_{banda\ escura} = \frac{n_{sim}}{l_{banda\ escura}} \quad ; \quad i_{banda\ clara} = \frac{n_{não}}{l_{banda\ clara}},$$

onde n_{sim} representa o número de integrações virais nas bandas escuras e $n_{não}$ representa o número de integrações virais nas bandas claras.

Depois de calcularmos o *rácio em extensão* e a *intensidade em número*, obtivemos para cada cromossoma um par (x, y), em que x representa a medida de integração nas bandas escuras e y a mediada de integração nas bandas claras. Com estes pares obtivemos uma representação gráfica, onde cada ponto corresponde a um cromossoma. Com vista a inferir sobre a preferência de integração de cada vírus, definimos a mesma escala em ambos os eixos do gráfico e representámos uma linha diagonal correspondente à recta de equação $y = x$. O que permitiu visualizar o número de cromossomas em que $y > x$ e em que $y < x$, ou seja, se o vírus tem preferência de integração nas bandas escuras ou nas bandas claras.

Para comprovar estatisticamente o que podemos inferir dos gráficos, aplicámos dois testes não paramétricos diferentes, o teste dos sinais e o teste de Wilcoxon (Siegel, 1975). Estes dois testes permitem comparar duas amostras dependentes, como é o nosso caso, pois os pares (x, y) foram calculados a partir da mesma amostra de integrações virais, pelo que são obviamente dependentes. Testámos a hipótese nula H0:$x = y$, isto é, testámos se é igualmente provável que o vírus integre tanto nas bandas escuras como nas bandas claras. Por outras palavras, testamos se relativamente a qualquer cromossoma é igualmente provável ter-se $x - y > 0$ ou $x - y < 0$ (Pestana e Velosa, 2002). O teste dos sinais focaliza-se no sentido da diferença entre x e y para cada cromossoma. Se H0 for verdadeira pode-se esperar que cerca de metade das diferenças tenha sinal "+" e a outra metade tenha sinal "-". Portanto sob a validade da hipótese nula, o número de sinais "+" na diferença entre as observações tem distribuição amostral *Binomial* $(N, \frac{1}{2})$, onde N é o número de pares (x, y) cuja diferença é positiva ou negativa; sendo N inferior a 25. Como por cromossoma temos um par (x, y) e todas as diferenças são não nulas, temos $N = 21$ quando testamos a integração do HIV-1 DNA e $N = 23$ para o HIV-2 DNA. A hipótese alternativa, H1, que pode ser $x \neq y$, ou $x > y$, ou ainda $x < y$, leva a considerar uma região de rejeição bilateral ou unilateral, à direita ou à esquerda, respectivamente. Tanto para o HIV-1 DNA como para o HIV-2 DNA conseguimos predizer o sentido da diferença de integração e formulámos as seguintes hipóteses: H0: o vírus integra com igual intensidade nas bandas escuras e nas bandas claras e H1: o vírus integra com maior intensidade nas bandas claras do que nas bandas escuras. Sendo K o número de sinais que ocorrem com menor frequência, o $p - value$ é dado, para um teste unilateral, pela probabilidade de se observar K ou

menos diferenças com o sinal menos frequente, se H0 for verdadeira. Considerámos em ambos os testes um nível de significância $\alpha = 1\%$. Quanto menor for o $p - value$, menor é a consistência entre os dados e H0. Assim, se $p - value < \alpha$ devemos rejeitar H0 ao nível de significância α.

Aplicámos depois um segundo teste, para testar as mesmas hipóteses, o teste de Wilcoxon, por ser um teste mais potente do que o teste dos sinais (Siegel, 1975). No caso do resultado de ambos ser contraditório, é tomado em conta o resultado dado pelo teste de Wilcoxon, uma vez que este teste é mais poderoso do que o teste dos sinais. Para obter a estatística do teste de Wilcoxon, T, atribui-se um *rank* ao valor absoluto da diferença, afectado do respectivo sinal. Assim, o valor da estatística do teste calculada $Tobs$ será a menor soma dos *ranks* do mesmo sinal. Utilizando N como o número de cromossomas e o valor de $Tobs$ consultámos a tabela específica para este teste onde obtivemos o valor de T crítico. Se o valor de $Tobs$ fosse inferior ao valor de T crítico, então rejeitaríamos H0.

Numa segunda fase do trabalho realizámos uma abordagem diferente na qual o HIV-1 DNA e o HIV-2 DNA foram analisados conjuntamente, sendo que o objectivo era saber qual a influência do tipo de vírus e do tipo de bandas na intensidade de integração. Para tal utilizámos uma análise de variância (ANOVA) a dois factores, considerando como factores o vírus e a banda e as intensidades de integração como nossa variável resposta. Ao factor vírus atribuímos dois níveis, HIV-1 e HIV-2 e ao factor banda os níveis atribuídos foram banda escura e banda clara, resultando assim em quatro tratamentos diferentes em terminologia ANOVA: HIV-1 x banda escura, HIV-1 x banda clara, HIV-2 x banda escura e HIV-2 x banda clara. Para que cada um dos cromossomas fosse considerado como uma observação replicada dos tratamentos foi, necessário obter outra medida denominada *intensidade em proporção*. A nova medida permitiu-nos tomar em consideração o facto do número de sítios de integração não ser igual por cromossoma e como tal poder originar resultados enganosos. Nesta intensidade considera-se a proporção em vez do número de sítios de integração virais por cromossoma ,sendo dada pelas seguintes fórmulas:

$$i_{banda\ escura} = \frac{\frac{n_{sim}}{n_{sim}+n_{não}}}{l_{banda\ escura}} \quad ; \quad i_{banda\ clara} = \frac{\frac{n_{não}}{n_{sim}+n_{não}}}{l_{banda\ clara}}.$$

Para cada um dos tratamentos referidos e cromossoma obteve-se uma intensidade, sendo que o valor das intensidades foi multiplicado por 10^8 afim de obtermos os resultados numa escala maior sem qualquer influência nos resultados dos testes estatísticos. Com os valores das intensidades atribuídos a cada tratamento, calculámos para cada factor a soma dos quadrados, os quadrados médios, os graus de liberdade e a razão de variância, segundo as fórmulas da tabela II.1. A ANOVA permite-nos ainda observar se existe alguma interacção entre o vírus e as bandas e dá-nos também o erro ou resíduo, isto é, a fonte de variância que resta depois de determinada a variância originada pelos factores vírus e bandas e pela interacção destes.

Tabela II.1 – Quadro resumo dos cálculos efectuados para a ANOVA. Neste quadro, *r* refere-se ao número de níveis do vírus que será sempre 2, o HIV-1 DNA e o HIV-2 DNA; *k* refere-se ao número de níveis das bandas que também será sempre 2, bandas claras e bandas escuras; *n* é o número total de observações em todos os tratamentos; *SQ* refere-se à soma dos quadrados; *QM* refere-se aos quadrados médios; *L* diz respeito ao vírus; *C* diz respeito às bandas; *E* ao erro e *T* ao total.

ORIGEM DA VARIÂNCIA	GRAUS DE LIBERDADE	SOMAS DE QUADRADOS	QUADRADOS MÉDIOS	RAZÃO DE VARIÂNCIA
Entre linhas	r-1	$SQL = \frac{\sum_{i=1}^{r} Y_{i,\circ,\circ}^2}{kc} - \frac{Y_{\circ,\circ,\circ}^2}{n}$	$QML = \frac{SQL}{r-1}$	$F_{0L} = \frac{QML}{QME}$
Entre colunas	k-1	$SQC = \frac{\sum_{j=1}^{k} Y_{\circ,j,\circ}^2}{rc} - \frac{Y_{\circ,\circ,\circ}^2}{n}$	$QMC = \frac{SQC}{k-1}$	$F_{0C} = \frac{QMC}{QME}$
Interacção	rk-r-k+1	$SQ(LC) = \frac{\sum_{i=1}^{r}\sum_{j=1}^{k} Y_{i,j,\circ}^2}{c} - \frac{\sum_{i=1}^{r} Y_{i,\circ,\circ}^2}{kc} - \frac{\sum_{j=1}^{k} Y_{\circ,j,\circ}^2}{rc} + \frac{Y_{\circ,\circ,\circ}^2}{n}$	$QM(LC) = \frac{SQ(LC)}{\text{rk-r-k+1}}$	$F_{0LC} = \frac{QM(LC)}{QME}$
Erro ou Resíduo	n-rk	$SQE = SQT - SQL - SQC - SQ(LC)$	$QME = \frac{SQE}{n-rk}$	
Total	n-1	$SQT = \sum_{i=1}^{r}\sum_{j=1}^{k}\sum_{l=1}^{c} Y_{i,j,l}^2 - nY_{\bullet\bullet\bullet}^2$		

Esta ANOVA com dois factores permite-nos testar três hipóteses nulas, que se podem traduzir da seguinte forma:

1) H01: O factor vírus não tem influência nas intensidades de integração, ou seja, não há diferenças entre as intensidades de integração do HIV-1 DNA e HIV-2 DNA;
2) H02: O factor banda não tem influência nas intensidades de integração, ou seja, não há diferenças entre as intensidades de integração nas bandas claras e nas bandas escuras;
3) H03: Não há interacção entre vírus e banda a influenciar as intensidades de integração.

Para um nível de significância de 5% consultámos a tabela de quantis da distribuição F. Se o valor obtido na razão de variância, a nossa estatística de teste, for superior ao valor do quantil de F, rejeitamos a hipótese nula e o factor associado a essa hipótese será significativo a 5%, senão este não é significativo, concluindo-se que o factor não tem influência nas intensidades de integração. No caso da interacção entre factores, se a hipótese for rejeitada conclui-se que um dos níveis do factor vírus combinado com um dos níveis do factor banda estão a influenciar as intensidade de integração.

Para completar a ANOVA e sabermos que tipo de vírus ou banda estaria a influenciar mais a integração, calculámos as médias das intensidades de cada tratamento.

II.2.2 HIV-1 DNA isolado de células T Jurkat

Para o HIV-1 DNA isolado de células T Jurkat, possuíamos 44236 sítios de integração do vírus, sendo que para estes apenas tínhamos informação acerca da posição exacta de integração, não possuíamos a extensão da região de integração como para o HIV-1 DNA e HIV-2 DNA isolado de PBMCs. Elaborámos igualmente tabelas de co-localização dos sítios de integração com as bandas *Giemsa* escuras, sendo também aqui retirados do total os sítios de integração que co-localizavam com os centrómeros e braços curtos dos acrocêntricos. Assim, para este HIV-1 DNA trabalhámos com um total de 42912 sítios de integração.

Após a elaboração da tabela procedemos ao cálculo da *intensidade em número* utilizando a metodologia já descrita na secção II.2.1. Não foi possível neste caso calcularmos o *rácio em extensão* devido ao facto de não possuirmos a extensão dos sítios de integração. De seguida, com os pares (x, y) obtidos do cálculo da intensidade elaborámos o gráfico como descrito anteriormente. Realizámos também os testes estatísticos dos sinais e de Wilcoxon já descritos na secção II.2.1. Para os dois testes utilizámos as seguintes hipóteses, H0: as integrações virais ocorrem com igual intensidade nas bandas escuras e claras; H1: as integrações virais ocorrem com maior intensidade nas bandas claras e um $N = 24$ (22 autossomas, cromossomas X e Y).

Para o HIV-1 DNA isolado de células T Jurkat não foi possível realizar os cálculos da ANOVA visto apenas possuirmos dados sobre um tipo de HIV, não sendo assim possível distribuir os dados em quatro tratamentos como realizado para o HIV isolado de PBMCs.

II.3 Sítios Frágeis

II.3.1 Obtenção das regiões frágeis e das regiões não frágeis

O genoma humano foi dividido em regiões frágeis (FRs) e regiões não frágeis (NFRs) de acordo com a sua posição. Para tal, utilizámos uma lista de FS obtida de Mrasek *et al.* (2010) (Mrasek *et al.*, 2010) que continha os FSs e as respectivas posições em cada cromossoma, sendo essa lista completada por FSs obtidos de Lukusa e Fryns (2008) (Lukusa e Fryns, 2008). De seguida, para determinar a sequência genómica dos FSs recorremos à base de dados *NCBI MapViewer* (37.2).

Para dividir o genoma humano, duas bandas consecutivas associadas com FSs foram agrupadas para formar uma FR (Laganà *et al.*, 2010), sendo que a região entre duas FRs foi considerada uma NFR. Obtivemos assim uma tabela (tabela II.2) com a posição inicial e final de cada FR e NFR por cromossoma.

O cromossoma Y não foi considerado neste estudo pelo facto de não possuir FS bem definidos, existindo apenas um provável FS (Holden *et al.*, 1986).

Tabela II.2 – Divisão do genoma em FRs e NFRs. Para cada região é indicada a posição inicial e final da sequência, sempre em bp.

Regiões Frágeis				Regiões Não Frágeis	
Sítio Frágil	**Localização**	**Posição inicial (bp)**	**Posição final (bp)**	**Posição inicial (bp)**	**Posição final (bp)**
FRA1A	1p36	0	27.800.000	27.800.000	34.400.000
FRA1	1p34	34.400.000	46.500.000	46.500.000	51.000.000
FRA1B/L/D/M/E	1p32/1p21.2	51.000.000	102.000.000	102.000.000	107.000.000
FRA1N	1p13	107.000.000	117.600.000	117.600.000	120.700.000
FRA1O/J/F	1p11/q21	120.700.000	153.300.000	153.300.000	154.800.000
FRA1P	1q23	154.800.000	163.800.000	163.800.000	171.200.000
FRA1G	1q25.1	171.200.000	174.300.000	174.300.000	184.000.000
FRA1K/Q/R/H/HS/I	1q31-1qter	184.000.000	247.200.000	-	-
FRA2M	2p25	0	12.800.000	12.800.000	17.000.000
FRA2C	2p24.2	17.000.000	19.100.000	19.100.000	23.900.000
FRA2N/O	2p21-23	23.900.000	47.600.000	47.600.000	52.700.000
FRA2D	2p16.2	52.700.000	54.800.000	54.800.000	61.100.000
FRA2P/Q/E	2p13-15	61.100.000	75.400.000	75.400.000	83.700.000
FRA2L/R/A	2p11-2q11	83.700.000	102.100.000	102.100.000	108.600.000
FRA2B	2q13	108.600.000	113.800.000	113.800.000	118.600.000
FRA2	2q14.2-14.3	118.600.000	129.600.000	129.600.000	134.800.000
FRA2F	2q21.3	134.800.000	136.600.000	136.600.000	144.700.000
FRA2K/S/T/G/H	2q22.3-32.1	144.700.000	189.100.000	189.100.000	197.100.000
FRA2I	2q33	197.100.000	209.100.000	209.100.000	215.100.000
FRA2U	2q35	215.100.000	221.300.000	221.300.000	237.000.000
FRA2J	2q37.3-qter	237.000.000	243.000.000	-	-
FRA3E/F	3p25-26	0	14.700.000	14.700.000	23.800.000
FRA3A	3p24.2	23.800.000	26.400.000	26.400.000	32.100.000
FRA3G/H	3p22-p21	32.100.000	54.400.000	54.400.000	58.500.000
FRA3B	3p14.2	58.500.000	63.700.000	63.700.000	71.800.000
FRA3I	3p13	71.800.000	74.200.000	74.200.000	91.700.000
FRA3J/K	3cen-q12	91.700.000	104.400.000	104.400.000	115.000.000
FRA3L/M/N	3q13.3-23	115.000.000	144.400.000	144.400.000	150.000.000
FRA3D/O/C/P/Q	3q25-29	150.000.000	200.000.000	-	-
				0	5.200.000
FRA4A/D/G	4p16.1-14	5.200.000	40.900.000	40.900.000	45.600.000
FRA4H	4p12	45.600.000	48.700.000	48.700.000	52.400.000
FRA4 B	4q12	52.400.000	59.200.000	59.200.000	76.500.000
FRA4I/F/J	4q21-23	76.500.000	102.500.000	102.500.000	120.600.000
FRA4E	4q27	120.600.000	124.000.000	124.000.000	139.500.000
FRA4C	4q31.1	139.500.000	141.700.000	141.700.000	155.100.000
FRA4C/K/L/M	4q32-q35	155.100.000	191.300.000	-	-

FRA5H/E/A	5p15-13	0	42.000.000	42.000.000	45.800.000
FRA5I/J/K/L/D/F/M/N/C	5p11-q31.1	45.800.000	135.400.000	135.400.000	147.200.000
FRA5O/P/G	5q33-35	147.200.000	180.900.000	-	-
				0	4.100.000
FRA6B	6p25.1	4.100.000	7.000.000	7.000.000	13.500.000
FRA6A	6p23	13.500.000	15.500.000	15.500.000	23.500.000
FRA6C	6p22.2	23.500.000	26.100.000	26.100.000	40.600.000
FRA6H	6p21.1	40.600.000	45.200.000	45.200.000	57.200.000
FRA6I	6p11-q11	57.200.000	63.500.000	63.500.000	70.000.000
FRA6D	6q13	70.000.000	75.900.000	75.900.000	87.500.000
FRA6G	6q15	87.500.000	92.100.000	92.100.000	99.900.000
FRA6J/F/K/L	6q16.3-23	99.900.000	139.100.000	139.100.000	149.100.000
FRA6M/E	6q25-26	149.100.000	164.400.000	164.400.000	170.900.000
FRA7B/L	7p22-21	0	19.500.000	19.500.000	35.600.000
FRA7C	7p14.2	35.600.000	37.500.000	37.500.000	43.300.000
FRA7D	7p13	43.300.000	46.600.000	46.600.000	53.900.000
FRA7A	7p11.2	53.900.000	57.400.000	57.400.000	71.800.000
FRA7J	7q11.23	71.800.000	77.400.000	77.400.000	90.900.000
FRA7E	7q21.2	90.900.000	92.600.000	92.600.000	97.900.000
FRA7F/K/G	7q22-31.2	97.900.000	117.200.000	117.200.000	130.100.000
FRA7H	7q32.3	130.100.000	132.400.000	132.400.000	137.300.000
FRA7M	7q34	137.300.000	142.800.000	142.800.000	147.500.000
FRA7I	7q36	147.500.000	158.800.000	-	-
FRA8G	8p23	0	12.700.000	12.700.000	19.100.000
FRA8H	8p21	19.100.000	29.700.000	29.700.000	38.500.000
FRA8I	8p11-q11	38.500.000	55.600.000	55.600.000	66.100.000
FRA8F	8q13	66.100.000	74.000.000	74.000.000	87.200.000
FRA8J/B	8q21.3-22.1	87.200.000	99.100.000	99.100.000	101.600.000
FRA8A	8q22.3	101.600.000	106.100.000	106.100.000	117.700.000
FRA8C	8q24.1	117.700.000	127.300.000	127.300.000	140.000.000
FRA8D	8q24.3	140.000.000	146.300.000	-	-
FRA9H	9p24	0	9.000.000	9.000.000	14.100.000
FRA9G/A/I	9p22-13	14.100.000	40.200.000	40.200.000	60.300.000
FRA9F/J/K/D	9q12-22.1	60.300.000	91.000.000	91.000.000	102.000.000
FRA9L/B/M/N	9q31-34	102.000.000	140.000.000	-	-
FRA10H	10p15	0	6.700.000	6.700.000	12.300.000
FRA10I/J	10p13-11.2	12.300.000	38.800.000	38.800.000	42.100.000
FRA10G/C/D	10q11.2-22.1	42.100.000	74.600.000	74.600.000	89.600.000
FRA10A	10q23.3	89.600.000	98.000.000	98.000.000	99.400.000
FRA10K	10q24.2	99.400.000	102.000.000	102.000.000	111.800.000
FRA10E	10q25.2	111.800.000	114.900.000	114.900.000	119.100.000
FRA10F	10q26.1	119.100.000	127.400.000	127.400.000	135.400.000
				0	2.800.000

FRA11J	11p15.4-15.3	2.800.000	12.600.000	12.600.000	16.100.000
FRA11C	11p15.1	16.100.000	21.600.000	21.600.000	26.000.000
FRA11D	11p14.2	26.000.000	27.200.000	27.200.000	31.000.000
FRA11E/K/L/M	11p13-q11	31.000.000	56.400.000	56.400.000	69.200.000
FRA11H	11q13.3	69.200.000	70.700.000	70.700.000	85.300.000
FRA11F	11q14.2	85.300.000	87.900.000	87.900.000	92.300.000
FRA11N/O	11q21-22	92.300.000	110.000.000	110.000.000	115.400.000
FRA11G	11q23.3	115.400.000	120.700.000	120.700.000	134.500.000
FRA12F/G/H/I	12p13-q11	0	37.000.000	37.000.000	44.600.000
FRA12A	12q13.1	44.600.000	53.100.000	53.100.000	56.300.000
FRA12J/K	12q14-15	56.300.000	69.800.000	69.800.000	78.700.000
FRA12B	12q21.3	78.700.000	91.200.000	91.200.000	94.800.000
FRA12L	12q23.1	94.800.000	100.000.000	100.000.000	107.500.000
FRA12E	12q24	107.500.000	132.300.000	-	-
				0	32.900.000
FRA13A	13q13.2	32.900.000	34.700.000	34.700.000	39.500.000
FRA13G/B/C/E/H/D/I	13q14-34	39.500.000	114.000.000	-	-
				0	19.100.000
FRA14D	14q11.2	19.100.000	23.600.000	23.600.000	31.800.000
FRA14E	14q13	31.800.000	36.900.000	36.900.000	41.000.000
FRA14A	14q21.2	41.000.000	43.200.000	43.200.000	48.300.000
FRA14F/B/C	14q22-24.1	48.300.000	69.300.000	69.300.000	72.900.000
FRA14G	14q24.3	72.900.000	78.400.000	78.400.000	88.900.000
FRA14H	14q32	88.900.000	106.400.000	-	-
				0	18.400.000
FRA15C	15q11.2	18.400.000	23.300.000	23.300.000	25.700.000
FRA15B	15q13	25.700.000	31.400.000	31.400.000	37.900.000
FRA15D/E/A	15q15	37.900.000	65.300.000	65.300.000	70.400.000
FRA15/F/G	15q24-26	70.400.000	100.300.000	-	-
				0	14.700.000
FRA16A	16p13.11	14.700.000	16.700.000	16.700.000	22.000.000
FRA16E/F/G/H/I/C	16p12.1-q22.1	22.000.000	69.000.000	69.000.000	78.200.000
FRA16D	16q23.2	78.200.000	80.500.000	80.500.000	82.700.000
FRA16J	16q24	82.700.000	88.800.000	-	-
				0	11.200.000
FRA17A/C	17p12-q11	11.200.000	28.800.000	28.800.000	35.400.000
FRA17D	17q21	35.400.000	47.600.000	47.600.000	54.900.000
FRA17B	17q23.1	54.900.000	55.600.000	55.600.000	59.900.000
FRA17E	17q24-25	59.900.000	78.800.000	-	-
FRA18D	18p11.3	0	7.200.000	7.200.000	17.300.000
FRA18E	18q11.2	17.300.000	23.300.000	23.300.000	31.000.000
FRA18A	18q12.2	31.000.000	35.500.000	35.500.000	52.000.000

FRA18B/C/F	18q21.3-23	52.000.000	76.100.000	-	-
FRA19B	19p13	0	19.800.000	19.800.000	26.700.000
FRA19C	19p11/q11	26.700.000	30.200.000	30.200.000	37.100.000
FRA19A	19q13	37.100.000	63.800.000	-	-
FRA20C	20p13	0	5.000.000	5.000.000	9.000.000
FRA20B	20p12.2	9.000.000	11.900.000	11.900.000	17.800.000
FRA20A	20p11.23	17.800.000	21.200.000	21.200.000	28.400.000
FRA20D	20q11.2	28.400.000	37.100.000	37.100.000	41.100.000
FRA20E	20q13.1	41.100.000	49.200.000	49.200.000	54.400.000
FRA20	20q13.3	54.400.000	62.400.000	-	-
				0	13.200.000
FRA21/A/B	21q11.2-22.1	13.200.000	38.600.000	38.600.000	46.900.000
				0	27.900.000
FRA22B	22q12.2	27.900.000	30.500.000	30.500.000	35.900.000
FRA22A	22q13	35.900.000	49.700.000		
				0	6.000.000
FRAXB	Xp22.31	6.000.000	9.500.000	9.500.000	37.500.000
FRAXG/H	Xp11-q11	37.500.000	65.100.000	65.100.000	67.700.000
FRAXI	Xq13	67.700.000	76.000.000	76.000.000	98.200.000
FRAXC	Xq22.1	98.200.000	102.500.000	102.500.000	120.700.000
FRAXJ/K	Xq25-26	120.700.000	137.800.000	137.800.000	140.100.000
FRAXD/A/E	Xq27.2-28	140.100.000	154.900.000	-	-

II.3.2 HIV-1 DNA e HIV-2 DNA isolado de PBMCs

Para o HIV isolado de PBMCs, possuíamos 140 sítios de integração para o HIV-1 DNA e 132 para o HIV-2 DNA. Os dados dos HIV foram tratados separadamente, mas aplicando o mesmo procedimento.

Começámos por elaborar tabelas de co-localização utilizando os sítios de integração do vírus e a posições das FRs e NFRs (tabela II.2), atribuindo um *sim* se existisse co-localização com uma FR ou um *não* se a co-localização fosse com uma NFR. De seguida, calculámos o *rácio em extensão* segundo as fórmulas que se seguem:

$$r_{FR} = \frac{l_{sim}}{l_{FR}} \quad ; \quad r_{NFR} = \frac{l_{\text{não}}}{l_{NFR}}.$$

Nestas fórmulas, l_{sim} representa o comprimento da integração viral na FR; l_{FR} representa o tamanho da FR; $l_{não}$ representa o comprimento da integração viral na NFR e l_{NFR} representa o tamanho da NFR. Calculámos também a *intensidade em número* dada pelas seguintes fórmulas:

$$i_{FR} = \frac{n_{sim}}{l_{FR}} \quad ; \quad i_{NFR} = \frac{n_{não}}{l_{NFR}}.$$

Aqui, n_{sim} representa o número de integrações virais na FR e $n_{não}$ representa o número de integrações virais na NFR.

De seguida, realizámos o teste dos sinais e o teste de Wilcoxon, como descritos na secção II.2.1. Para os FSs, possuíamos 23 cromossomas para cada um dos tipos de HIV. Para ambos os testes e no que diz respeito ao HIV-1 DNA formulámos as seguintes hipóteses, H0: as integrações virais ocorrem com igual intensidade nas FRs e nas NFRs; H1: as integrações virais ocorrem com maior intensidade nas NFRs. Para o HIV-2 DNA as hipóteses formuladas foram, H0: as integrações virais ocorrem com igual intensidade nas FRs e nas NFRs; H1: as integrações virais ocorrem com maior intensidade nas FRs. Utilizámos para os dois testes um $N = 23$.

Para a realização da ANOVA calculámos a *intensidade em proporção* segundo as fórmulas:

$$i_{FR} = \frac{\frac{n_{sim}}{n_{sim}+n_{não}}}{l_{FR}} \quad ; \quad i_{NFR} = \frac{\frac{n_{não}}{n_{sim}+n_{não}}}{l_{NFR}}.$$

A ANOVA foi realizada segundo o procedimento descrito na secção II.2.1 utilizando dois factores, o vírus (HIV-1 DNA e HIV-2 DNA) e a região (FR e NFR) resultando os quatro tratamentos diferentes, HIV-1 x FR, HIV-1 x NFR, HIV-2 x FR e HIV-2 x NFR.

Para os cálculos na ANOVA, formulámos as seguintes hipóteses:

1) H01: O factor vírus não tem influência nas intensidades de integração, ou seja, não há diferenças entre as intensidades de integração do HIV-1 DNA e HIV-2 DNA;
2) H02: O factor região não tem influência nas intensidades de integração, ou seja, não há diferenças entre as intensidades de integração nas FRs e nas NFRs;
3) H03: Não há interacção entre vírus e região a influenciar as intensidades de integração.

Neste caso, possuíamos 23 cromossomas para cada tratamento, ou seja o mesmo número de observações por amostra verificando-se assim uma situação de equilíbrio nos dados na qual a ANOVA é robusta (Ito, 1980).

No final, e para completar a análise calculámos as médias com as intensidades de cada tratamento.

II.3.3 HIV-1 DNA isolado de células T Jurkat

Para a análise do HIV-1 DNA isolado de células T Jurkat, utilizámos 44150 sítios de integração viral, deixando de fora 86 sítios de integração correspondentes ao cromossoma Y, uma vez que este não possui FS bem definidos (Holden *et al.*, 1986).

Elaborámos também tabelas de co-localização dos sítios de integração viral com as FRs ou

com as NFRs e calculámos a *intensidade em número*. Com a obtenção dos pares (x, y) criámos o gráfico sempre com a linha na diagonal para conseguirmos prever se as integrações ocorreriam com maior intensidade nas FRs ou nas NFRs. Realizámos os testes dos sinais e de Wilcoxon descritos na secção II.2.1 tendo por base as seguintes hipóteses, H0: as integrações virais ocorrem com igual intensidade nas FRs e nas NFRs; H1: as integrações virais ocorrem com maior intensidade nas FRs e utilizando um $N = 23$.

Pelas mesmas razões descritas na secção II.2.2 também aqui não foi possível calcular o *rácio em extensão* nem realizar o teste da ANOVA.

III. Resultados

III.1 Bandas *Giemsa*

III.1.1 HIV-1 DNA e HIV-2 DNA isolados de PBMCs

III.1.1.1 Testes não paramétricos

Para o HIV-1 DNA isolado de PBMCs, após os cálculos do *rácio em extensão* e da *intensidade em número*, elaborámos os gráficos com os pares (x, y) que se encontram representados na figura III.1 e III.2, respectivamente.

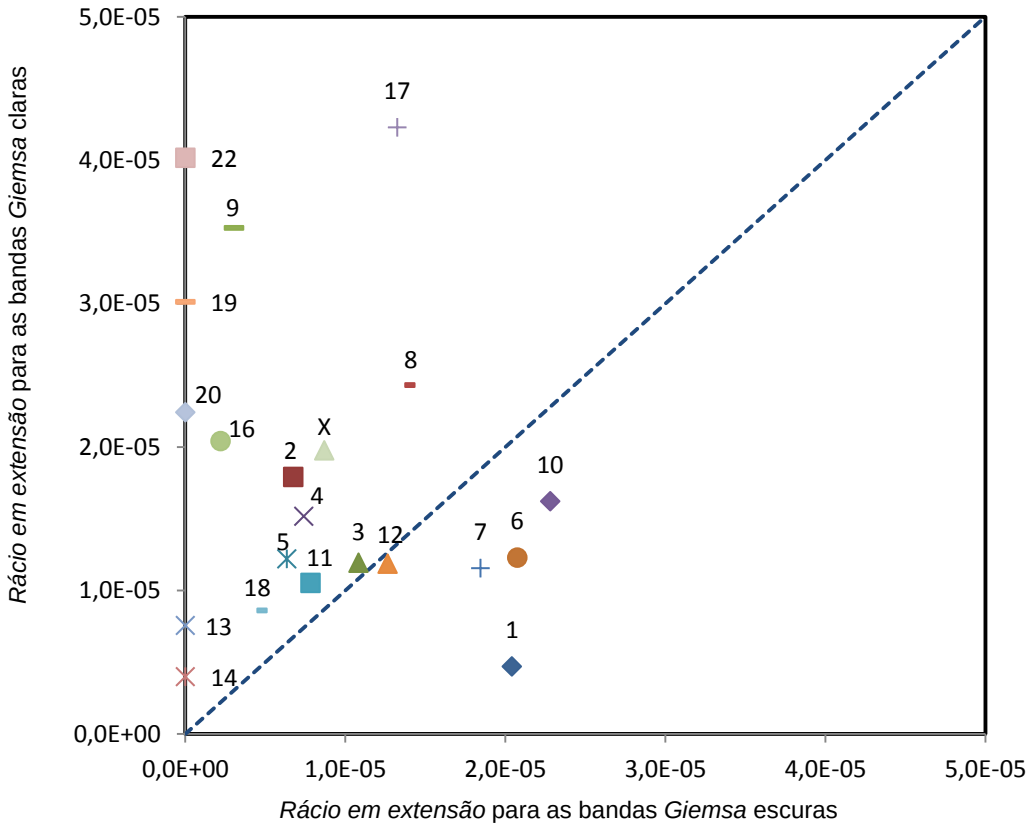


Figura III.1 - Representação do *rácio em extensão* para a integração do HIV-1 DNA nas bandas *Giemsa* escuras *versus* bandas *Giemsa* claras. Cada ponto representa um cromossoma cujas coordenadas são os pares (x, y) obtidos no cálculo do rácio. A linha a tracejado representa a recta $y = x$ que nos permite visualizar a possível preferência de integração do vírus.

O gráfico representado na figura III.1 sugere que existirá uma preferência para o vírus se integrar nas bandas *Giemsa* claras, uma vez que existe um maior número de pontos acima da recta a tracejado. No teste dos sinais, obtivemos um $p - value$ de 0,013 que é superior ao nível de significância de 1%. Assim, não rejeitamos H0 e concluímos que as integrações ocorrem com igual intensidade nos dois tipos de bandas. No entanto, para o teste de Wilcoxon obtivemos um valor de $Tobs$ inferior ao valor

de T crítico apresentado na tabela III.1. Assim rejeitámos H0 sendo o resultado que as integrações de HIV-1 DNA ocorrem com maior intensidade nas bandas *Giemsa* claras. Como o resultado dos dois testes é contraditório, tomámos como certo o resultado dado pelo teste de Wilcoxon, visto ser um teste mais poderoso. O resultado sugerido pelo gráfico é então confirmado pelo teste estatístico.

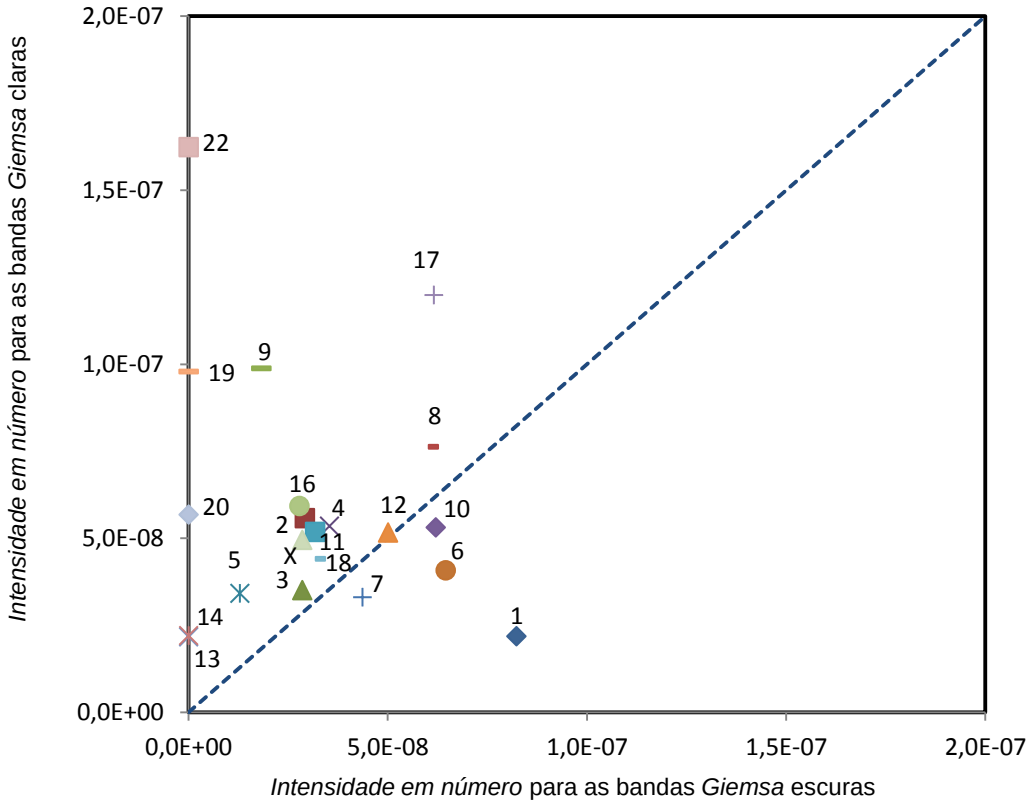


Figura III.2 - Representação da *intensidade em número* para a integração do HIV-1 DNA nas bandas *Giemsa* escuras *versus* bandas *Giemsa* claras. Cada ponto representa um cromossoma cujas coordenadas são os pares (x, y) obtidos no cálculo da intensidade. A linha diagonal representa a recta $y = x$ que nos permite visualizar a possível preferência de integração do vírus.

Analisando a figura III.2 verifica-se que o vírus terá preferência para se integrar nas bandas *Giemsa* claras. O teste dos sinais está de acordo com o visualizado no gráfico, pois o $p - value$ obtido de 0,004 é inferior ao nível de significância. O teste de Wilcoxon está igualmente em concordância, sendo o valor de $Tobs$ inferior ao valor de T crítico, como se pode verificar na tabela III.1. Assim, para o HIV-1 DNA os resultados indicam que o vírus integra com maior intensidade nas bandas *Giemsa* claras do que nas bandas *Giemsa* escuras.

Tanto para o *rácio em extensão* como para a *intensidade em número* o resultado obtido foi o mesmo, ou seja, o vírus tem preferência pelas bandas claras, estando as duas medidas calculadas em concordância.

Quanto ao HIV-2 DNA, os gráficos obtidos encontram-se na figura III.3 para o *rácio em extensão* e na III.4 para a *intensidade em número*.

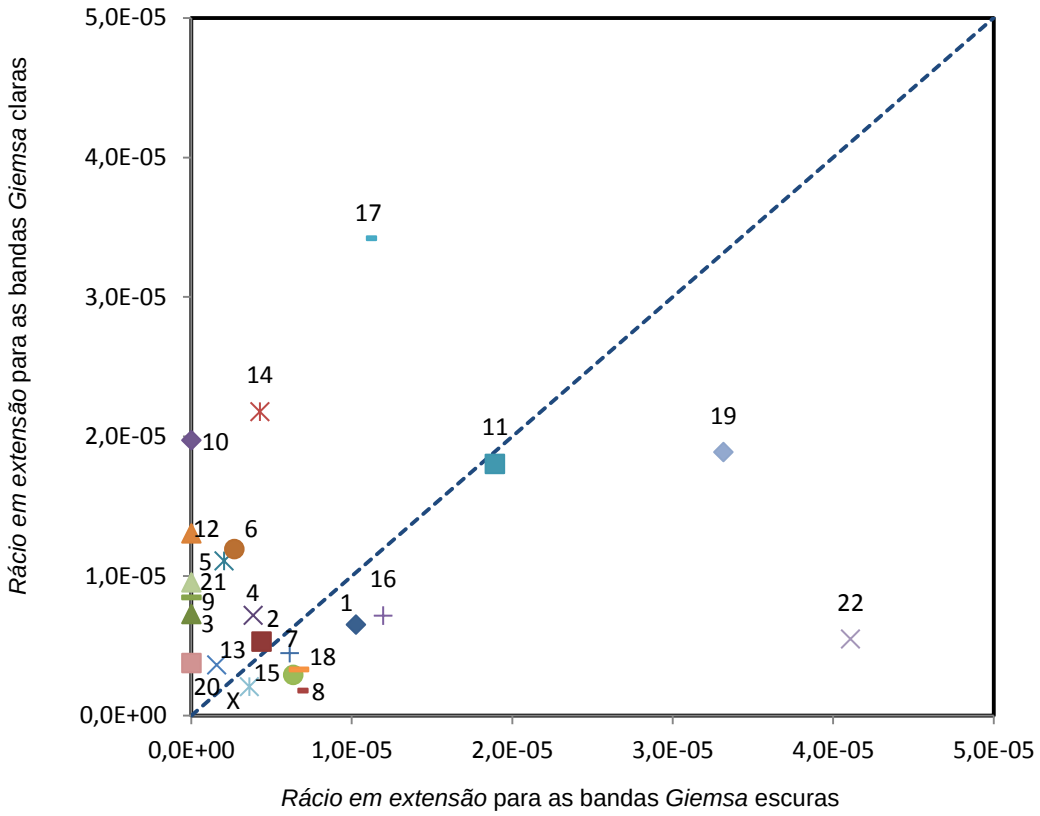


Figura III.3 - Representação do *rácio em extensão* para a integração do HIV-2 DNA nas bandas *Giemsa* escuras *versus* bandas *Giemsa* claras. Cada ponto representa um cromossoma cujas coordenadas são os pares (x, y) obtidos no cálculo do rácio. A linha a tracejado representa a recta $y = x$ que nos permite visualizar a possível preferência de integração do vírus.

Ao visualizarmos a figura III.3 não conseguimos perceber se existirá uma preferência de integração por parte do HIV-2 DNA, existindo apenas uma ligeira inclinação para as bandas *Giemsa* claras. No teste dos sinais, obtivemos um $p - value$ de 0,339 que é claramente superior ao nível de significância utilizado de 1%, o que nos leva a não rejeitar a H0. Os valores obtidos para o teste de Wilcoxon levam-nos igualmente a não rejeitar a H0. Assim, para o HIV-2 DNA, ambos os testes dizem que o vírus integra com igual intensidade nas bandas claras e escuras, não apresentando uma preferência por nenhum tipo de banda.

Na figura III.4 observa-se uma distribuição dos cromossomas bastante próxima da diagonal, ou seja, não se observa uma distinta preferência de integração. Todos os valores dos testes dos sinais e de Wilcoxon nos levam a não rejeitar a H0, logo para a *intensidade em número* o vírus integra com igual intensidade nos dois tipos de bandas.

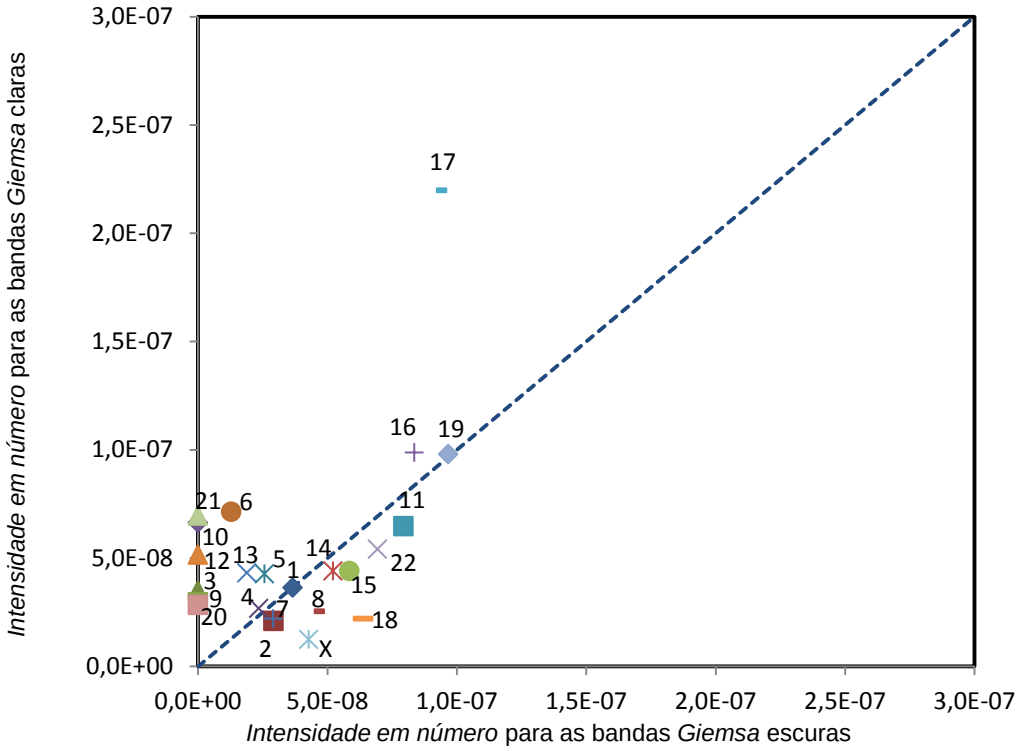


Figura III.4 - Representação da *intensidade em número* para a integração do HIV-2 DNA nas bandas *Giemsa* escuras *versus* bandas *Giemsa* claras. Cada ponto representa um cromossoma cujas coordenadas são os pares (x, y) obtidos no cálculo da intensidade. A linha diagonal representa a recta $y = x$ que nos permite visualizar a possível preferência de integração do vírus.

Os resultados encontrados para o HIV-2 DNA isolado de PBMCs foram coerentes nos dois tipos de medida utilizados, sendo a conclusão final de que o HIV-2 DNA não apresenta uma preferência para um tipo de banda, integrando nos dois com igual intensidade.

Assim, utilizando estes dois testes de estatística não paramétrica verifica-se uma distinção entre o HIV-1 DNA e o HIV-2 DNA, apresentando o primeiro uma preferência para integrar nas bandas *Giemsa* claras e o segundo não apresentando uma preferência específica de integração.

Tabela III.1 – Resultados dos testes dos sinais e de Wilcoxon para HIV-1 DNA e HIV-2 DNA. Para o teste dos sinais é apresentado o valor do $p - value$ para o *rácio em extensão* e a *intensidade em número* para cada um dos HIV. Para o teste de Wilcoxon estão representados os valores do T obs e do T crítico para cada vírus e para cada medida.

	$p - value$ **: teste dos sinais**		$Tobs$**: teste de Wilcoxon**		T**crítico: teste de Wilcoxon**
	Rácio em extensão	*Intensidade em número*	*Rácio em extensão*	*Intensidade em número*	*Rácio em extensão e intensidade em número*
HIV-1 DNA	0,013	0,004	42	38	49
HIV-2 DNA	0,339	0,339	97	86	62

III.1.1.2 Teste da ANOVA

Na segunda parte do trabalho utilizámos o teste da ANOVA que aplicámos às intensidades calculadas para os quatro tratamentos em estudo, sendo eles HIV-1 x banda escura, HIV-1 x banda clara, HIV-2 x banda escura e HIV-2 x banda clara. Os resultados obtidos encontram-se esquematizados na tabela III.2.

Tabela III.2 – Resumo dos resultados obtidos com a aplicação do teste da ANOVA.

Fonte	**Graus de liberdade**	**Somas dos quadrados**	**Quadrados médios**	**Razão de variância** (F)	$p - value$
Vírus	1	0,64	0,64	0,62	0,4333
Bandas	1	11,61	11,61	11,2	0,0012
Vírus x Bandas	1	0,45	0,45	0,43	0,5138
Erro	84	87,04	1,04		
Total	87	99,74			

Para o teste da ANOVA, o nível de significância que utilizámos foi de 5%. Para interpretarmos os resultados, olhamos para o valor de $p - value$ que nos dá a probabilidade de rejeitar H0. Se esse valor for inferior ao nível de significância, então o valor de F é significativo. Neste caso, o único valor de $p - value$ inferior a 5% é o que corresponde ao valor de F de 11,2 fazendo das bandas o único factor significativo. A interacção vírus x bandas e o vírus em si não são estatisticamente significativos. Isto significa que as bandas influenciam significativamente a integração viral.

Para completar a ANOVA calculámos as médias por tratamento encontrando-se estas apresentadas na tabela III.3.

Tabela III.3 - Quadro resumo com as médias calculadas por tratamento.

	Bandas		
Vírus	Escuras	Claras	Total
HIV-1 DNA	0,444898	1,321348	0,883123
HIV-2 DNA	0,75943	1,348836	1,054133
Total	0,602164	1,335092	

A interpretação para este resultado é que a intensidade de integração nas bandas claras é significativamente superior à das bandas escuras, sendo que em termos da ANOVA não foi considerada significativa a diferença de intensidades que existe entre o HIV-1 DNA e o HIV-2 DNA.

III.1.2 HIV-1 DNA isolado de células T Jurkat

Para o HIV-1 isolado de células T Jurkat, obteve-se o gráfico representado na figura III.5 após os cálculos da *intensidade em número*.

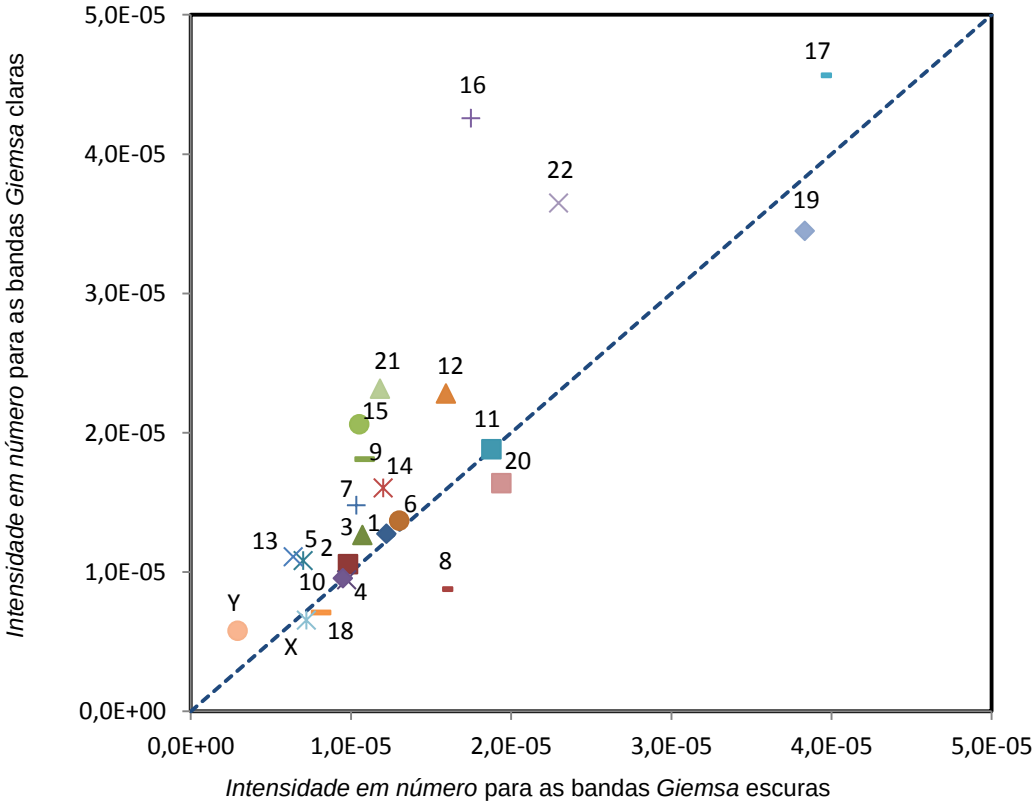


Figura III.5 - Representação gráfica do resultado para o cálculo da *intensidade em número* para a integração do HIV-1 isolado de células T Jurkat nas bandas *Giemsa* claras e escuras. Cada ponto representa um cromossoma cujas coordenadas são os pares (x, y) resultantes do cálculo da intensidade. Na diagonal encontra-se representada a recta que permite visualizar as preferências de integração do vírus.

Pela observação da figura, visualiza-se uma preferência para o vírus se integrar mais nas bandas *Giemsa* claras. No teste dos sinais, obtivemos um $p - value$ de 0,011 superior ao nível de significância de 1% levando à não rejeição da hipótese de que as integrações ocorrem com igual intensidade nos dois tipos de bandas. No entanto, para o teste de Wilcoxon, o valor de T crítico foi de 69 e o de $Tobs$ foi de 59. Sendo o $Tobs$ inferior ao T crítico, rejeitamos H0 concluindo-se então que as integrações de HIV-1 DNA ocorrem com maior intensidade nas bandas *Giemsa* claras. Os dois testes apontam resultados diferentes, pelo que tomamos como certo o resultado do teste de Wilcoxon.

III.2 Sítios Frágeis

III.2.1 HIV-1 DNA e HIV-2 DNA isolado de PBMCs

III.2.1.1 Testes não paramétricos

Para o HIV-1 DNA isolado de PBMCs obtivemos o gráfico representado na figura III.6 com os pares (x, y) resultantes do cálculo do *rácio em extensão*.

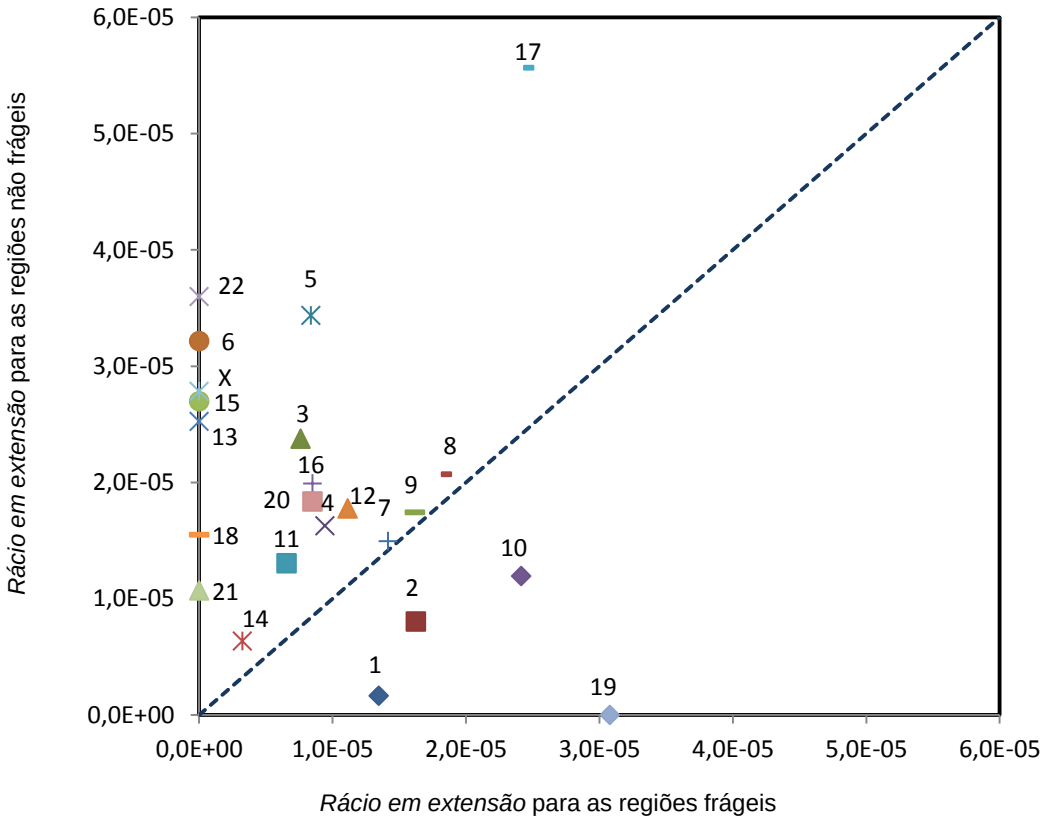


Figura III.6 - Representação gráfica da integração do HIV-1 DNA isolado de PBMCs nas FRs *versus* NFRs. Cada cromossoma está representado por um ponto cujas coordenadas (x, y) resultam do cálculo do *rácio em extensão*. A linha diagonal representa a recta $y = x$ que nos permite visualizar a possível preferência de integração do vírus.

Na figura acima representada, observa-se uma tendência para o vírus se integrar mais nas NFRs, facto esse que é confirmado no teste dos sinais onde obtivemos um $p - value$ de 0,001 inferior ao nível de significância, o que nos leva a rejeitar H0 em favor de H1. Conclui-se assim que o vírus integra com maior intensidade nas NFRs. No teste de Wilcoxon rejeitámos igualmente a H0 visto termos obtido um valor de $Tobs$ inferior ao valor de T crítico, como se pode observar na tabela III.4.

Para os dois testes aplicados obtivemos o mesmo resultado de que o HIV-1 DNA isolado de PBMCs integra com maior intensidade nas NFRs. Este resultado confirma também o que inicialmente

tínhamos observado na figura III.6, sendo também que novamente os dois testes estão em concordância.

Para os cálculos da *intensidade em número* obteve-se o gráfico representado na figura III.7, observando-se que existe uma preferência para a integração nas NFRs.

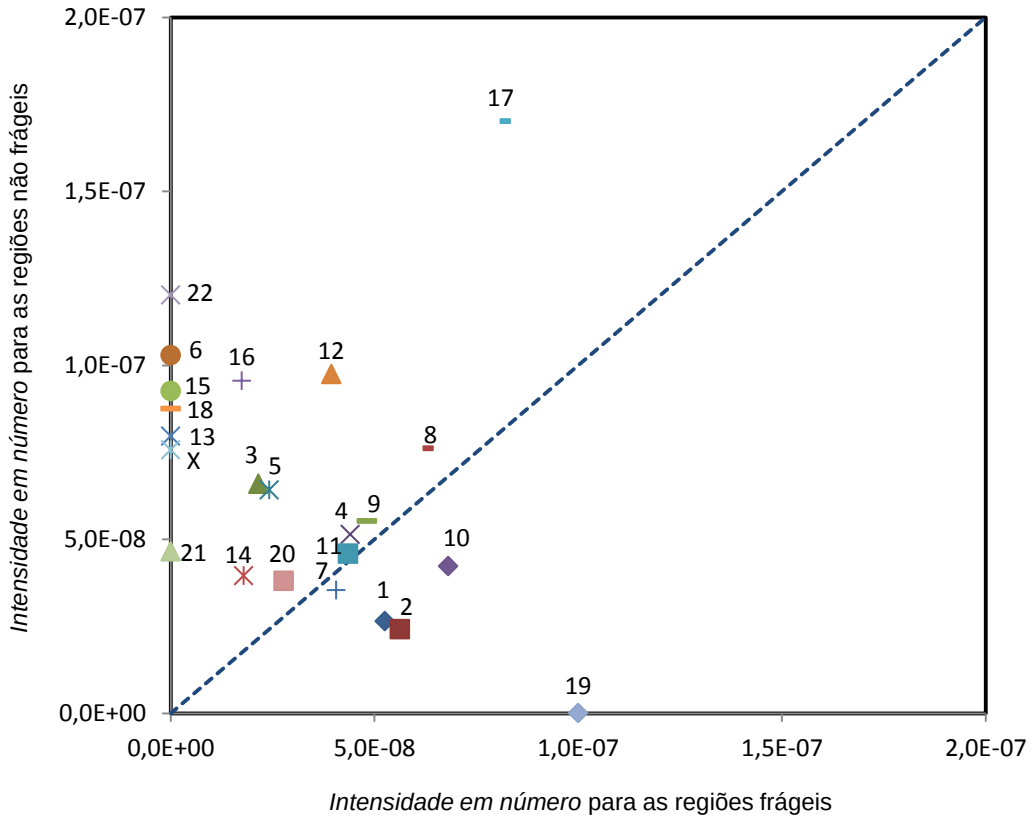


Figura III.7 – Gráfico para a integração do HIV-1 DNA isolado de PBMCs nas FRs *versus* NFRs. Cada ponto representa um cromossoma cujas coordenadas são os pares (x, y) obtidos no cálculo da intensidade. A linha diagonal representa a recta $y = x$ que nos permite visualizar a possível preferência de integração do vírus.

Para a *intensidade em número*, o resultado do teste dos sinais indica que devemos rejeitar a H0, já que obtivemos um $p - value$ inferior ao nível de significância, como se pode observar na tabela III.4. No que diz respeito ao teste de Wilcoxon e tendo nós obtido um valor de $Tobs$ inferior ao T crítico, igualmente rejeitamos a H0. Assim, verifica-se que existem diferenças nas integrações virais nas FRs e nas NFRs, sendo que o vírus integra com maior intensidade nas NFRs.

Para o HIV-1 DNA a conclusão geral é de que o vírus integra preferencialmente nas NFRs estando as duas medidas calculadas em concordância.

Em relação ao HIV-2 DNA isolado de PBMCs, os resultados obtidos encontram-se nas figuras III.8 e III.9 para o *rácio em extensão* e para a *intensidade em número*, respectivamente.

Ao analisarmos a figura III.8 verificamos que existem mais algumas integrações nas FRs do que nas NFRs apesar de muitos dos cromossomas ficarem representados perto ou mesmo sobre a

linha diagonal. Para confirmar a falta de tendência do gráfico, realizámos então o teste dos sinais segundo o qual o vírus não tem preferência de integração, integrando com igual intensidade nas FRs e nas NFRs, já que o valor de $p - value$ nos leva a não rejeitar H0. O mesmo se verifica para o teste de Wilcoxon que se apresenta com um $Tobs$ superior ao T crítico. Assim, os resultados para esta medida indicam que o HIV-2 DNA não apresenta preferência de integração.

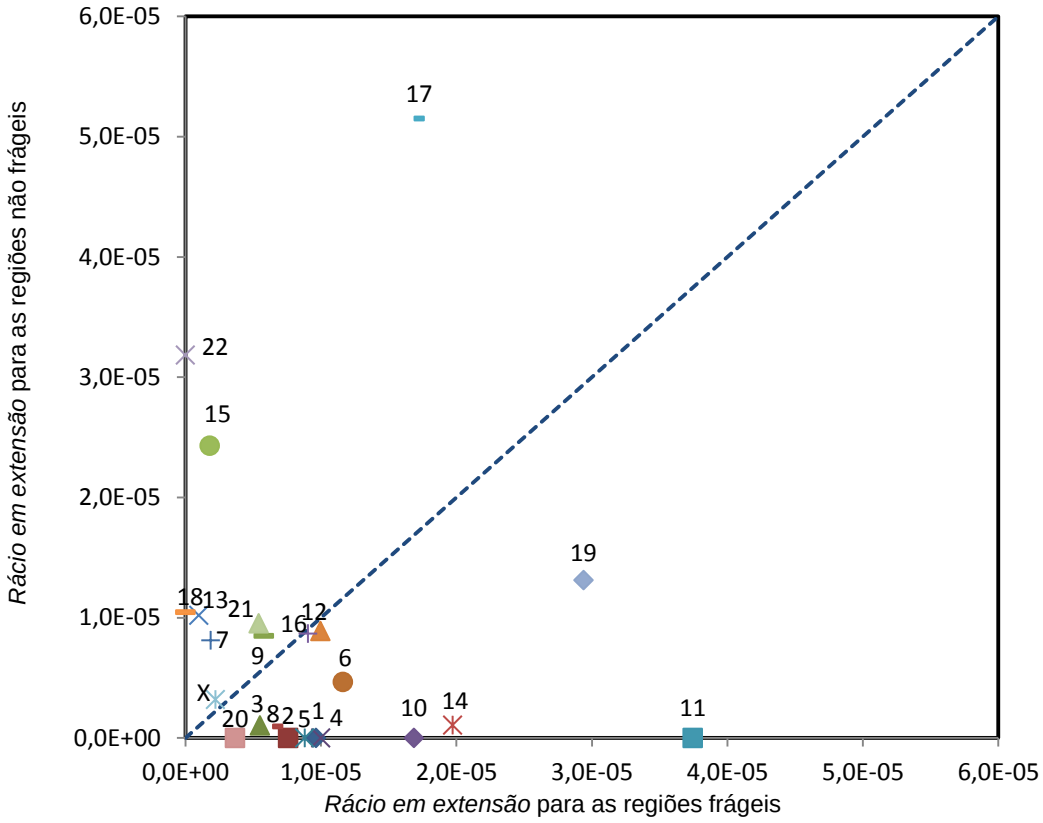


Figura III.8 - Representação gráfica da integração do HIV-2 DNA isolado de PBMCs nas FRs *versus* NFRs. Cada cromossoma encontra-se representado por um ponto tendo como coordenadas (x, y) os pares calculados para o *rácio em extensão*. A azul encontra-se representada a recta $y = x$ que nos permite prever a possível preferência de integração do vírus.

No gráfico da figura III.9 encontram-se mais integrações nas FRs, mas e como já havia sido verificado para o *rácio em extensão* do HIV-2 DNA, existem alguns cromossomas muito próximos da linha a tracejado. Para o teste dos sinais obteve-se um $p\ value$ de 0,339 que é claramente superior ao nível de significância de 1%, sendo que como tal não rejeitamos a H0 de que o vírus não possui uma preferência de integração. Com o teste de Wilcoxon, chegámos à mesma conclusão, uma vez que o $Tobs$ é superior ao T crítico, como se pode verificar na tabela III.4.

Para o HIV-2 DNA isolado de PBMCs, tanto para o *rácio em extensão* como para a *intensidade em número*, os resultados indicam que este tipo de vírus integra com igual intensidade na FRs e nas NFRs.

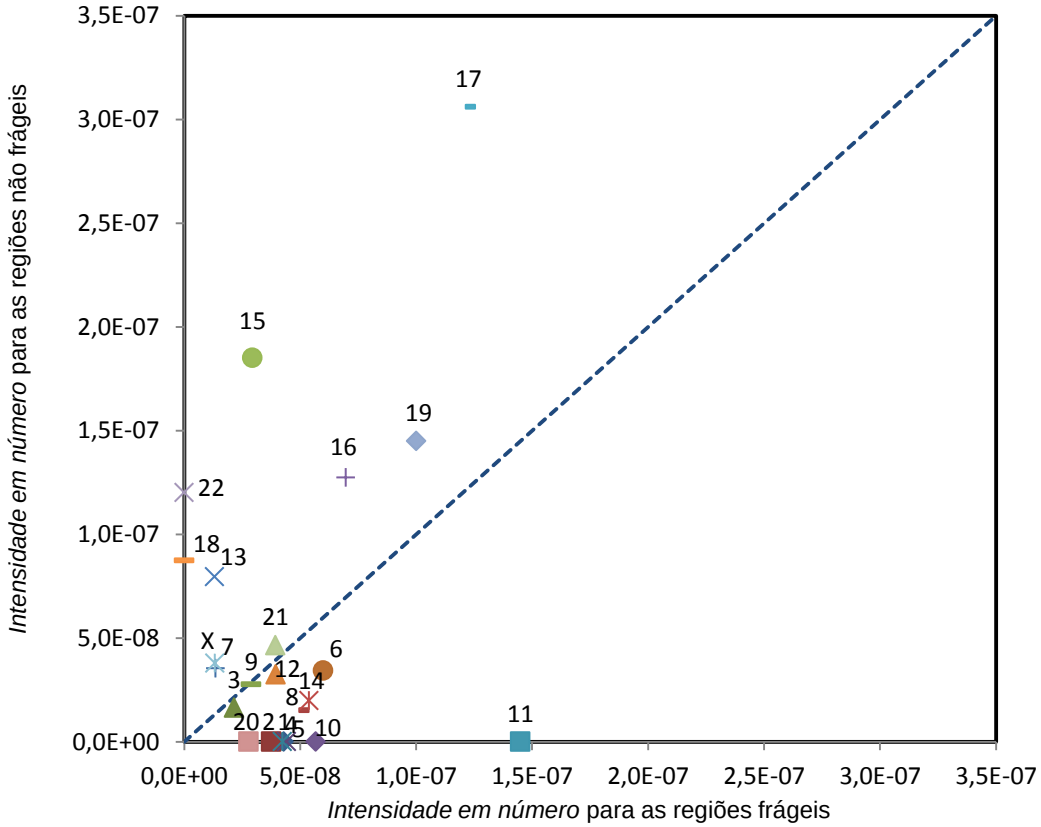


Figura III.9 – Representação do resultado para a integração do HIV-2 DNA isolado de PBMCs nas FRs *versus* NFRs. Cada ponto representa um cromossoma cujas coordenadas são os pares (x, y) obtidos no cálculo da intensidade. A linha diagonal representa a recta $y = x$ que nos permite visualizar a possível preferência de integração do vírus.

Tabela III.4 – Resultados dos testes dos sinais e de Wilcoxon para HIV-1 DNA e HIV-2 DNA. Para o teste dos sinais é apresentado o valor do $p - value$ para o *rácio em extensão* e a *intensidade em número* para cada um dos HIV. Para o teste de Wilcoxon estão representados os valores do $Tobs$ e do T crítico para cada vírus e para cada medida.

	$p - value$: teste dos sinais		**$Tobs$: teste de Wilcoxon**		**Tcrítico: teste de Wilcoxon**
	Rácio em extensão	*Intensidade em número*	*Rácio em extensão*	*Intensidade em número*	*Rácio em extensão e intensidade em número*
HIV-1 DNA	0,001	0,005	53	50	62
HIV-2 DNA	0,202	0,339	113	127	62

III.2.1.2 Teste da ANOVA

Após os cálculos para a ANOVA obtivemos os resultados que se encontram na tabela III.5. Tal como para as bandas *Giemsa*, também nos FSs o nível de significância utilizado foi de 5%.

Tabela III.5 – Quadro resumo dos resultados da ANOVA.

Fonte	**Graus de liberdade**	**Somas dos quadrados**	**Quadrados médios**	**Razão de variância** (F)	$p - value$
Vírus	1	0,11	0,11	0,14	0,710
Região	1	7,75	7,75	9,98	0,002
Vírus x Região	1	3,71	3,71	4,78	0,031
Erro	88	68,32	0,78		
Total	91	79,89			

Com base nestes resultados, podemos concluir que os valores significativos de F são para o factor região e para a interacção entre o vírus e a região, sendo que o factor vírus não é significativo. Assim, podemos dizer que o factor região influencia significativamente a integração do vírus. Particularmente, quando analisamos a tabela III.6 contendo as médias calculadas por tratamento, verificamos que a intensidade de integração nas NFRs é bastante superior à das FRs (1,2546 contra 0,6741).

Tabela III.6 – Resultados para o cálculo da média por tratamento.

	Região		
Vírus	FRs	NFRs	Total
HIV-1 DNA	0,5075	1,4898	0,9986
HIV-2 DNA	0,8408	1,0194	0,9301
Total	0,6741	1,2546	

Quanto ao facto de a interacção vírus x região ser significativa, isto quer dizer que um nível particular do factor vírus combinado com um nível particular do factor região influenciam a intensidade de integração. Quando olhamos para a tabela III.6 observamos que o HIV-1 DNA nas NFRs possui uma intensidade média de integração elevada quando comparada com os outros três tratamentos (HIV-1 DNA x FRs, HIV-2 DNA x NFRs, HIV-2 DNA x FRs).

Assim, como conclusão do teste da ANOVA podemos dizer que as integrações ocorrem significativamente com maior intensidade nas NFRs e que em particular o HIV-1 DNA é o que contribui mais para esse facto.

Comparando os resultados do teste dos sinais e de Wilcoxon que diziam que o HIV-1 integrava com maior intensidade nas NFRs, os resultados do teste da ANOVA estão em concordância.

III.2.2 HIV-1 DNA isolado de células T Jurkat

Para o HIV-1 DNA isolado de células T Jurkat, o gráfico elaborado após o cálculo da *intensidade em número* encontra-se na figura III.10.

Na figura III.10 visualiza-se uma possível preferência de integração nas FRs, o que não é confirmado pelos testes. No teste dos sinais obtivemos um $p - value$ de 0,202 que é superior ao nível de significância, o que significa que não rejeitamos H0. Para o teste de Wilcoxon, obtivemos um valor de $Tobs$ de 82 e um valor de T crítico de 62. Sendo o $Tobs$ superior, não rejeitamos igualmente H0. Concluímos então que o HIV-1 DNA isolado de células T Jurkat não apresenta uma preferência de integração, sendo que integra com igual intensidade nas duas regiões.

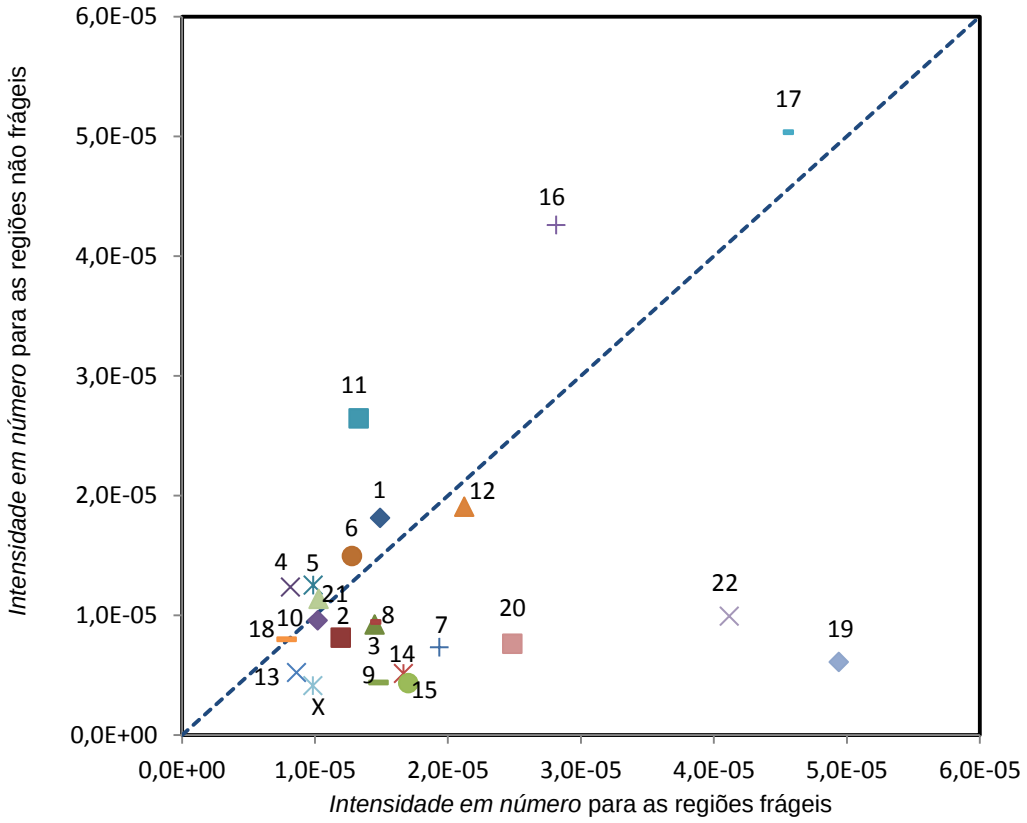


Figura III.10 – Gráfico para a integração do HIV-1 DNA isolado de células T Jurkat nas FRs *versus* NFRs. Cada ponto representa um cromossoma cujas coordenadas são os pares (x, y) obtidos no cálculo da intensidade. A linha diagonal representa a recta $y = x$ que nos permite visualizar a possível preferência de integração do vírus.

IV. Discussão

O HIV é um retrovírus que se distingue de muitos outros vírus não só por possuir uma enzima transcriptase reversa para produzir a dupla cópia do DNA a partir do genoma RNA viral, como também na forma como se integra no genoma do hospedeiro (Craigie e Bushman, 2012). Os vectores de retrovírus possuem um potencial papel na terapia génica não só devido à sua capacidade para introduzirem material genético nas células alvo (Nagel *et al.*, 2012) como também pelo facto de se integrarem de uma forma estável no genoma hospedeiro (Kay *et al.*, 2001). Assim, a compreensão das preferências de integração deste vírus é bastante importante devido à aplicação da inserção retroviral em terapia génica, área onde efeitos adversos têm sido associados com a integração de vectores retrovirais perto de proto-oncogenes (Howe *et al.*, 2008; Hacein-Bey-Abina *et al.*, 2010). O conhecimento das preferências de integração pode também auxiliar na escolha de genes apropriados introduzidos por vectores, o que por sua vez poderá minimizar a possível toxicidade na integração (Schröder *et al.*, 2002). Assim, e para compreendermos mais sobre estas preferências, fomos estudar as integrações do HIV nas bandas *Giemsa* e nos FSs.

No genoma humano existem parâmetros genéticos que se relacionam, o que leva a que não seja fácil identificar quais os determinantes primários da escolha do alvo de integração do HIV (Craigie e Bushman, 2012), já que os sítios de integração do HIV não estão distribuídos ao acaso no genoma humano, encontrando-se em regiões enriquecidas em genes activos (Schröder *et al.*, 2002).

Estudos anteriores revelaram que o HIV favorece a integração em unidades transcripcionalmente activas (Berry *et al.*, 2006; Wang *et al.*, 2007), sendo que estas unidades estão muitas vezes associadas a regiões com elevada densidade de genes, regiões com elevados conteúdos G/C, regiões com alta densidade de ilhas CpG e frequências elevadas de repetições *Alu* (Craigie e Bushman, 2012). O facto destas regiões serem também alvos preferenciais de integração do HIV, está de acordo com os resultados obtidos por nós, de que tanto o HIV-1 DNA isolado de PBMCs como o HIV-1 DNA isolado de células T Jurkat integram preferencialmente nas bandas *Giemsa* claras, uma vez que este tipo de bandas é também rico em ilhas CpG, possuindo um elevado conteúdo G/C (Niimura e Gojobori, 2002). Estes resultados estão também de acordo com os resultados obtidos por Elleder *et al.* (2002) (Elleder *et al.*, 2002) segundo os quais o HIV integra preferencialmente nas bandas R, ou seja, *Giemsa* claras e em regiões de cromatina aberta. A enzima IN do HIV-1 interage com componentes do complexo remodelador da cromatina (Kalpana *et al.*, 1994), pelo que os mecanismos de integração envolvem uma grande disponibilidade de cromatina aberta (Elleder *et al.*, 2002). A preferência do vírus para regiões transcripcionalmente activas pode dever-se ao facto de estas regiões permitirem uma eficiente continuação do ciclo de replicação do vírus, já que permitem uma maior transcrição do provírus (Elleder *et al.*, 2002), aumentando assim as chances para a expressão dos genes virais (Wang *et al.*, 2007). As células T infectadas pelo HIV têm tipicamente uma semi-vida de apenas um ou dois dias antes que as células sejam mortas pelo sistema imune ou pela toxicidade da infecção (Perelson *et al.*, 1996), o que faz com que o HIV tenha um tempo limitado para a produção de descendência (Craigie e Bushman, 2012). Integrando-se em regiões transcripcionalmente activas, o vírus assegura a transcrição dos seus genes, uma vez que foi demostrado que a integração em unidades de transcrição é geralmente favorável para a transcrição

eficiente (Jordan *et al.*, 2001; Lewinski *et al.*, 2005). O achado de que o HIV-1 integra preferencialmente nas bandas *Giemsa* claras pode ainda ser justificado pelo facto de estas serem ricas em genes (Holmquist e Ashley, 2006), uma vez que o ambiente intranuclear dos genes activos pode conduzir à integração do vírus, sendo que esta preferência pode ter se desenvolvido para facilitar a eficiente expressão dos genes de HIV após a infecção (Schröder *et al.*, 2002). Outra característica das bandas *Giemsa* claras por oposição às bandas *Giemsa* escuras é relativa aos seus elevados níveis de acetilação das histonas H3 e H4 (Sadoni *et al.*, 1999; Holmquist e Ashley, 2006) e foi demostrado que a frequência de integração do HIV foi associada com as modificações epigenéticas, incluindo a acetilação da H3 e H4 (Wang *et al.*, 2007), dando ainda consistência ao nosso achado de que o HIV integra mais nas bandas *Giemsa* claras.

Wang *et al.* (2007) (Wang *et al.*, 2007) realizaram estudos com HIV e verificaram que a integração deste vírus era desfavorecida nos centrómeros, o que está de acordo com os resultados encontrados por nós em que obtivemos poucas integrações nos centrómeros (dados não mostrados), e se pode relacionar com a não transcrição da heterocromatina constitutiva (Lewin, 2004). Pelo contrário, o mesmo grupo de autores (Wang *et al.*, 2007) encontrou que as regiões subteloméricas favoreciam a integração do HIV, de acordo com o facto destas regiões serem relativamente mais transcritas (Riethman *et al.*, 2004).

No que diz respeito ao HIV-2 DNA isolado de PBMCs, chegámos à conclusão de que este vírus integra com igual intensidade nos dois tipos de bandas, não apresentando uma tendência específica como no caso do HIV-1 DNA isolado de PBMCs e de células T Jurkat. O mesmo se verifica quando analisamos os resultados para a integração nos FSs, sendo que mais uma vez, o HIV-2 DNA isolado de PBMCs integra com igual intensidade nas FRs e nas NFRs. Estas diferenças encontradas entre o HIV-1 DNA e o HIV-2 DNA podem ser devidas quer a diferenças na constituição dos dois vírus como a diferenças a nível clínico. A nível proteico, estes dois tipos de HIV possuem uma grande diferença na constituição. Enquanto que o HIV-1 necessita apenas de uma proteína, a Vpr, para promover a infecção em células que não estão em divisão e despoletar a paragem do ciclo celular na fase G2 em células em divisão, o HIV-2 necessita de duas proteínas para desempenhar as mesmas funções, a Vpx e a Vpr (Casey *et al.*, 2010). A nível clínico a grande diferença entre os dois vírus é que a progressão para a imunodeficiência ocorre mais lentamente no HIV-2 do que no HIV-1, apresentado o HIV-2 uma transmissibilidade mais baixa (Nyamweya *et al.*, 2013).

Ainda em relação às bandas *Giemsa*, o resultado do teste da ANOVA indicou que eram as bandas a influenciar a integração, ou seja, é o facto de ser banda *Giemsa* escura ou banda *Giemsa* clara que leva a que o vírus se integre e, em particular este integra significativamente mais nas bandas claras do que nas escuras. Este resultado está também de acordo com o facto de o HIV seleccionar o sítio de integração de acordo com as suas características, e por isso poder ser o factor banda a influenciar a preferência, já que cada tipo de banda possui características diferentes que podem ser mais ou menos vantajosas para a integração do vírus. Na figura IV.1 encontra-se representado um esquema ilustrativo das preferências de integração do HIV-1 DNA isolado de PBMCs.

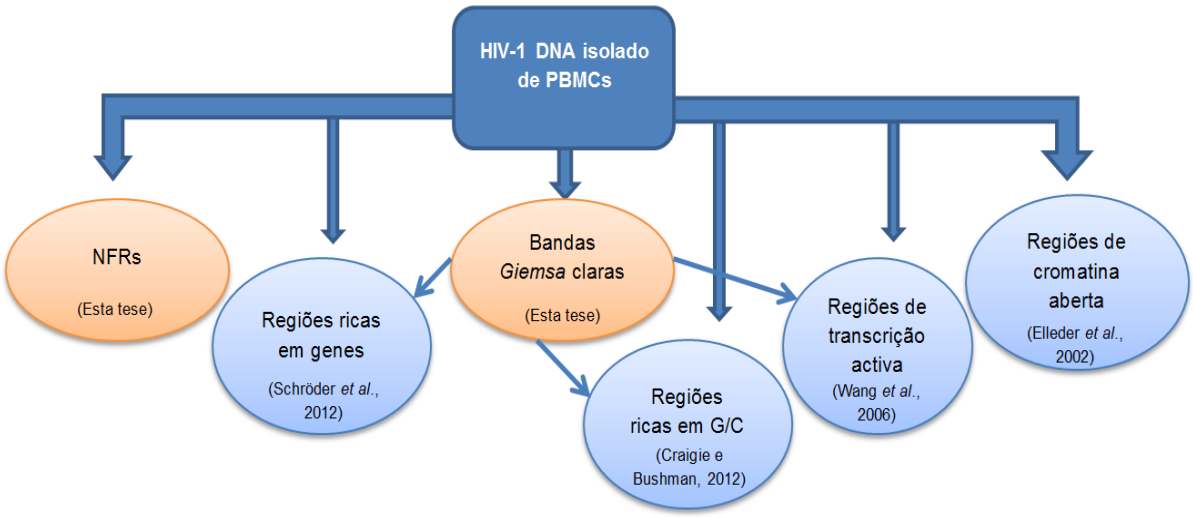


Figura IV.1 - Esquema ilustrativo das preferências de integração do HIV-1 DNA isolado de PBMCs.

Para os FSs existem estudos que revelam que estes são alvos preferenciais para a integração de alguns vírus, nomeadamente do HPV16 e HPV18 (Wilke *et al.*, 1996; Matovina *et al.*, 2009), do vírus Epstein-Barr (Luo *et al.*, 2004) e do vírus da hepatite B (Feitelson e Lee, 2007), o que não se verifica para o HIV, já que os resultados indicaram que o HIV-1 DNA isolado de PBMCs integra com maior intensidade nas NFRs e que o HIV-2 DNA isolado de PBMCs e o HIV-1 DNA isolado de células T Jurkat não apresenta uma preferência de integração, podendo assim concluir-se que o HIV não integra preferencialmente nas FRs. Assim verifica-se que o HIV apresenta um padrão de integração diferente de outros vírus, no que diz respeito aos FSs. Uma possível explicação para esta diferença pode residir nas diferentes fases do ciclo celular em que o vírus entra nas células humanas. Por exemplo, enquanto que a progressão no ciclo celular através de mitose é crítica para a infecção do HPV (Pyeon *et al.*, 2009), o HIV pode infectar células que não estão em divisão (Craigie e Bushman, 2012). Logo, se a cromatina está em diferentes fases do ciclo celular é lógico que ocorram diferenças na selecção dos sítios de integração.

O facto de o HIV não ter tendência para integrar nos FSs pode estar relacionado com a tendência destes sítios para formarem estruturas secundárias como *harpin* e estruturas cruciformes que interferem na replicação (Schwartz *et al.*, 2006), o que pode dificultar a integração do genoma do vírus e sua replicação. A constituição em bases dos FSs pode também ser uma das razões pela não preferência do vírus para as FRs, uma vez que estas possuem um elevado conteúdo A/T (Zlotorynski *et al.*, 2003; Mortusewicz *et al.*, 2013). O HIV para conseguir completar o seu ciclo de vida, necessita de integrar o seu genoma no genoma hospedeiro para aí conseguir replicar-se e criar novos vírus para infectar outras células (Turlure *et al.*, 2004), pelo que sendo os FSs sítios de grande instabilidade genómica (Durkin e Glover, 2007) poderão não ser os alvos de integração mais favoráveis ao vírus. Outra das razões que poderá justificar a não integração nas FRs poderá ser a menor actividade transcripcional destas regiões em relação às bandas *Giemsa* claras que são alvos preferenciais para a integração. Os FSs são locais sujeitos aos acontecimentos de quebras ou lacunas de forma espontânea ou induzida (Büttel *et al.*, 2004), pelo que para o HIV podem não ser

sítios de preferência de integração, já que o vírus poderá não conseguir integrar-se facilmente ou até não conseguir concluir a sua replicação de modo a gerar novos vírus.

O facto de o HIV-1 DNA isolado de PBMCs integrar preferencialmente nas NFRs, está de acordo com estudos realizados por Bester *et al.* (2006) (Bester *et al.*, 2006) segundo os quais a análise de correlação entre FSs e integração de HIV em várias linhas celulares revelou não existir integração preferencial nos FSs. Por outro lado, Kim *et al.* (2008) (Kim *et al.*, 2008) realizaram um estudo no qual obtiveram o resultado de que a percentagem de integração do HIV-1 nos CFSs era mais elevada do que a do controlo aleatório utlizado por eles; sendo que estes autores realizaram o estudo em células do cancro da próstata não fazendo uma comparação entre FRs e NFRs.

Para os FSs, o teste da ANOVA confirmou os resultados dos testes dos sinais e de Wilcoxon, na medida em que verificou que o HIV-1 DNA isolado de PBMCs era o que mais contribuía para que as integrações ocorressem nas NFRs. Verificámos também nesta análise de variância que era o factor região que estava a influenciar significativamente a integração, ou seja, que provavelmente serão as características da região em si que levam o vírus a integrar mais ou menos nas NFRs ou nas FRs.

Em relação aos gráficos apresentados, é curioso observar que tanto nos gráficos relativos às bandas *Giemsa*, como nos relativos aos FSs existem cromossomas que se destacam do aglomerado maior, como é o caso do cromossoma 17. Não podemos dissociar deste resultado os factos de Soto *et al*. (2001) (Soto *et al.*, 2011) terem encontrado que este era o cromossoma com maior número de integrações virais e de este cromossoma ser rico em genes, em SINEs e em conteúdo CpG (Zody *et al.*, 2006).

Os resultados encontrados foram diferentes entre as PBMCs e as células T Jurkat, o que pode estar relacionado primeiramente com o tamanho da amostra utilizado, já que o número de sítios de integração para as PBMCs era bastante inferior ao número de sítios de integração para as células T Jurkat. Esta diferença de resultados pode também estar associada ao facto de as metodologias para a obtenção dos sítios de integração diferirem nos dois tipos de células. Nomeadamente, para a obtenção dos sítios de integração isolados de células T Jurkat por Wang *et al.* (2007) (Wang *et al.*, 2007) os autores utilizaram a técnica de *massively parallel pyrosequencing*, a qual lhes permitiu gerar a colecção maior de sítios de integração e a qual não tinha sido utilizada nos estudos de Mitchell *et al.* (2004) (Mitchell *et al.*, 2004) e de MacNeil *et al.* (2006) (MacNeil *et al.*, 2006). Esta técnica de pirosequenciação, aplicada ao HIV permite também investigações mais profundas na distribuição dos sítios de integração dos vectores de HIV (Bushman *et al.*, 2008). As diferenças encontradas podem ser ainda devidas ao facto de a integração ser específica do tipo celular, já que células neoplásicas e células não totalmente diferenciadas podem ter um comportamento diferente do de outras células no que diz respeito à integração viral. Biasco *et al.* (2001) (Biasco *et al.*, 2011) descobriram que os vectores retrovirais possuíam preferências de integração que eram específicas do tipo celular e se relacionavam com o estado da cromatina das células alvo no momento da transdução. Nagel *et al.* (2012) (Nagel *et al.*, 2012) verificaram que os sítios de integração de sequências retrovirais em células HeLa, que são neoplásicas tal como as células T Jurkat, estavam significativamente mais

internas no núcleo que as suas regiões homólogas, enquanto que outros tipos celulares não apresentavam esta diferença.

IV.1 Principais conclusões

Em conclusão, nesta tese conseguimos aplicar métodos estatísticos e genéticos para estudarmos os sítios de integração do HIV nos FSs e nas bandas *Giemsa*. Concluímos que os dois tipos de HIV possuem preferências de integração diferentes, sendo que no que diz respeito às bandas *Giemsa*, o HIV-1 DNA isolado de PBMCs e isolado de células T Jurkat integra com maior intensidade nas bandas *Giemsa* claras, enquanto que o HIV-2 DNA isolado de PBMCs não apresenta uma preferência clara por um tipo de bandas específico, integrando com igual intensidade nas bandas escuras e claras. No que diz respeito aos cálculos da ANOVA, concluímos que o factor banda influencia significativamente a integração do vírus, mostrando este em particular preferência pelas bandas claras.

Em relação aos FSs, o HIV-2 DNA isolado de PBMCs mais uma vez não apresenta uma preferência sendo que integra com igual intensidade nas FRs e nas NFRs. Aqui, também o HIV-1 DNA isolado de células T Jurkat integra com igual intensidade nas duas regiões. Já o HIV-1 DNA isolado de PBMCs integra com maior intensidade nas NFRs, facto que é confirmado não só pelos testes dos sinais e de Wilcoxon, como também pela ANOVA.

Este trabalho é baseado em análises estatísticas que permitem complementar os estudos laboratoriais. Contudo, existem factores que actuam *in vivo* que afectam os processos biológicos, incluindo mecanismos fisiológicos e que podem influenciar a selecção dos sítios de integração do vírus. Alguns destes factores podem ser o tipo celular presente, a fase do ciclo celular, o estado transcripcional da célula e a localização do DNA cromossómico.

IV.2 Perspectivas Futuras

Em relação ás bandas *Giemsa* existem já estudos publicados, não com o HIV-2, mas com o HIV-1. Assim, com esta nossa contribuição, o assunto fica mais clarificado.

No que diz respeito aos FSs, existem estudos que os relacionam com as integrações de outros vírus, sendo que para o HIV ainda não são conhecidas muitas informações. Assim, para estudos futuros, poderá utilizar-se a divisão do genoma em FRs e NFRs elaborada por nós para estudar melhor as integrações deste retrovírus. Nomeadamente, poder-se-ão obter outros sítios de integração do HIV por metodologias diferentes e verificar se os resultados serão os mesmos ou não e compará-los, afim de se ir conhecendo melhor quais as preferências para a integração deste vírus que tantos problemas causa na nossa população.

Por outro lado, e como continuação desta tese poder-se-ia utilizar as mesmas sequências de integração do vírus e verificar se este possuía uma tendência para se integrar nos ERFSs. Com a obtenção das sequências destes ERFS poder-se-iam fazer estudos semelhantes ao apresentado nesta tese para verificar se o HIV apresentaria uma tendência para se integrar nos ERFSs ou não e por fim, poder-se-iam comparar resultados entre esta nova classe e os CFSs. Este estudo seria pertinente, já que este novo tipo de FSs apresenta características em comum com as bandas *Giemsa* claras e portanto se levantar a hipótese de estes serem alvos preferenciais de integração do HIV.

V. Bibliografia

Abdeen, S. K., Salah, Z., Maly, B., Smith, Y., Tufail, R., Abu-Odeh, M., Zanesi, N., Croce, C. M., Nawaz, Z. e Aqeilan, R. I. 2011. *WWOX* inactivation enhances mammary tumorigenisis. Oncogene 30:3900-3906.

Abraham, R. T. 2001. Cell Cycle checkpoint signaling through the ATM and ATR kinases. Genes & Development 15:2177-2196.

Abu-Daya, A., Brown, P. M. e Fox, K. R. 1995. DNA sequence preferences of several AT-selective minor groove binding ligands. Nucleic Acids Research 23:3385-3392.

Alkan, C., Cardone, M. F., Catacchio, C. R., Antonacci, F., O´Brien, S. J., Ryder, O. A., Purgato, S., Zoli, M., Valle, G. D., Eichler, E. E. e Ventura, M. 2011. Genome-wide characterization of centromeric satellites from multiple mammalian genomes. Genome Research 21:137-145.

Allshire, R. C. e Karpen, G. H. 2008. Epigenetics regulation of centromeric chromatin: old dogs, new tricks? Nature Reviews in Genetics 9:923-937.

Arlt, M. F., Durkin, S. G., Ragland, R. L. e Glover, T. W. 2006. Common fragile sites as targets for chromosome rearrangements. DNA Repair 5:1126-1135.

Armanios, M., Alder, J. K., Parry, E. M., Karim, B., Strong, M. A. e Greider, C. W. 2009. Short telomeres are sufficient to cause the degenerative defects associated with aging. The American Journal of Human Genetics 85:823-832.

Arrighi, F. E. e Hsu, T. C. 1971. Localization of heterochromatin in human chromosomes. Cytogenetic and Genome Research 10:81-86.

Ban, S., Cologne, J. B. e Neiishi, K. 1965. Effect of radiation and cigarette smoking on expression of FUdR-inducible common fragile sites in huma peripheral lymphocytes. Mutation Research 334:197-203.

Barlow, J. H., Faryabi, R. B., Callén, E., Wong, N., Malhowski, A., Chen, H. T., Gutierrez-Cruz, G., Sun, H.-W., Mckinnon, P., Wright, G., Casellas, R., Robbiani, D. F., Staudt, L., Fernandez-Capetillo, O. e Nussenzweig, A. 2013. Identification of early replicating fragile sites that contribute to genome instability. Cell 152:620-632.

Barré-Sinoussi, F., Chremann, J. C., Rey, F., Nugeyre, M. T., Chamaret, S., Gruest, J., Dauguet, C., Axler-Blin, C., Vézinet-Brun, F., Rouzioux, C., Rozenbaum, W. e Montagnier, L. 1983. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). Science 220:868-871.

Berry, C., Hannenhalli, S., Leipzig, J. e Bushman, F. D. 2006. Selection of target sites for mobile DNA integration in the human genome. PLOS Computational Biology 2:1450-1462.

Bester, A. C., Schwartz, M., Schmidt, M., Garrique, A., Hacein-Bay-Abina, S., Cvazzana-Calvo, M., Ben-Porat, N., von Kalle, C., Fischer, A. e Kerem, B. 2006. Fragile sites are preferential targets for integrations of MLV vectors in gene theraphy. Gene Theraphy 13:1057-1059.

Biasco, L., Ambrosi, A., Pellin, D., Bartholomae, C., Brigida, I., Grazia, M., Serio, C. D., von Kalle, C., Schmidt, M. e Aiuti, A. 2011. Integration profile of retroviral vector in gene theraphy treated

patients is cell-specific according to gene expression and chromatin conformation of target cell. EMBO Molecular Medicine 3:89-101.

Bickmore, W. A. 2013. The Spatial Organization of the Human Genome. Annual Review of Genomics and Human Genetics 14:67-84.

Bickmore, Wendy A. e van Steensel, B. 2013. Genome Architecture: Domain Organization of Interphase Chromosomes. Cell 152:1270-1284.

Blackburn, E. H. 2005. Telomeres and telomerase: their mechanisms of action and the effects of altering their functions. Federation of European Biochemical Societies 579:859-862.

Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Müller, S., Elis, R., Cremer, C., Speicher, M. R. e Cremer, T. 2005. Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. PLOS Biology 3:826-842.

Boyarchuk, E., de Oca, R. M. e Almouzni, G. 2011. Cell cycle dynamics of histones variants at the centromere, a model for chromosomal landmarks. Current Opinion in Cell Biology 23:266-276.

Bushman, F. D., Hoffmann, C., Ronen, K., Malani, N., Minkah, N., Rose, H. M., Tebas, P. e Wang, G. P. 2008. Massively parallel pyrosequencing in HIV research. AIDS 22:1411-1415.

Büttel, I., Fechter, A. e Schwab, M. 2004. Common fragile sites and cancer: Targeted cloning by insertional mutagenis. Annual New York Academy of Sciences 1028:14-27.

Carpersson, T., Zech, L. e Johansson, C. 1970. Differential binding of alkylating fluorochromes in human chromosomes. Experimental Cell Research 60:315-319.

Carvalho, C., Pereira, H. M., Ferreira, J., Pina, C., Mendonça, D., Rosa, A. C. e Carmo-Fonseca, M. 2001. Chromosomal G-dark bands determine the spatial organization of centromeric heterochromatin in the nucleus. Molecular Biology of the Cell 12:3563-3572.

Casey, L., wen, X. e de Noronha, C. M. C. 2010. The functions of the HIV1 protein Vpr and its action through the DCAF1, DDB1, Cullin4 ubiquitin ligase. Cytokine 51:1-9.

Casper, A. M., Nghiem, P., Arlt, M. F. e Glover, T. W. 2002. ATR regulates fragile site stability. Cell 111:779-789.

Chan, S. R. W. L. e Blackburn, E. H. 2004. Telomeres and telomerase. Philosophical Transactions of The Royal Society 359:109-121.

Chang, F., Re, F., Sebastian, S., Sazer, S. e Luban, J. 2004. HIV-1 Vpr induces defects in mitosis, cytokinesis, nuclear structure, and centrsomes. Molecular Biology of the Cell 15:1793-1801.

Cheeseman, I. M. e Desai, A. 2008. Molecular architecture of the kinetochore-microtubule interface. Nature Reviews Molecular Cell Biology 9:33-46.

Ciuffi, A. e Bushman, F. D. 2006. Retroviral DNA integration: HIV and the role of LEDGF/p75. Trends in Genetics 22:388-395.

Clavel, F., Guétard, D., Brun-Vézinet, F., Chamaret, S., Rey, M.-A., Santos-Ferreira, M. O., Laurent, A. G., Dauguet, C., Katlama, C., Rouzioux, C., Klatzmann, D., Champalimaud, J. L. e Montagnier, L. 1986a. Isolation of a new human retrovirus from West African patients with AIDS. Science 233:343-346.

Clavel, F., Guyader, M., Guétard, D., Sallé, M., Montagnier, L. e Alizon, M. 1986b. Molecular cloning of the human immune deficiency virus type 2. Nature 324:691-695.

Cleveland, D. W., Mao, Y. e Sullivan, K. F. 2003. Centromeres and kinetochores: from epigenetics to mitotic checkpoint signaling. Cell 112:407-421.

Coquelle, A., Pipiras, E., Toledo, F., Buttin, G. e Debatisse, M. 1997. Expression of fragile sites triggers intrachromosomal mammalian gene amplification and sets boundaries to early amplicons. Cell 89:215-225.

Costantini, M., Auletta, F. e Bernardi, G. 2012. The distributions of a "new" and "old" Alu sequences in the human genome: the solution of a "mistery". Molecular Biology and Evolution 29:421-427.

Craigie, R. e Bushman, F. D. 2012. HIV DNA integration. Cold Spring Harbor Perspectives in Medicine 2:1-18.

Crane, J., Mittar, D., Soni, D. e Mclntyre, C. 2011. Cell cycle analysis using the BD BrdU FITC assay on the BD FACSVERSE™ System. BD Biosciences 1-12.

Cremer, M., Küpper, K., Wagler, B., Wizelman, L., von Hase, J., Weiland, Y., Kreja, L., Diebold, J., Speicher, M. R. e Cremer, T. 2003. Inheritance of gene density - related higher order chromatin arrangements in normal and tumor cell nuclei. The Journal of Cell Biology 162:809-820.

Cremer, T. e Cremer, C. 2001. Cromosome territories, nuclear architecture and gene regulation in mammalian cells. Nature 2:292-301.

Cremer, T. e Cremer, M. 2010. Chromosome Territories. Cold Spring Harbor Perspectives in Biology 2:1-22.

Croft, J. A., Bridger, J. M., Boyle, S., Perry, P., Teagure, P. e Bickmore, W. A. 1999. Differences in the localization and morphology of chromosomes in the human nucleus. The Journal of Cell Biology 145:1119-1131.

Cross, S. H. e Bird, A. P. 1995. CpG islands and genes. Current Opinion in Genetics and Development 5:309-314.

Daniel, R., Greger, J. G., Katz, R. A., Taganov, K. D., Wu, X., Kappes, J. C. e Skalka, A. M. 2004. Evidence that stable retroviral transduction and cell survival following DNA integration depend on components of the nonhomologous end joining repair pathway. Journal of Virology 78:8573-8581.

Daniel, R. e Smith, J. A. 2008. Integration site slection by retroviral vectors: Molecular mechanism and clinical consequences. Human Gene Therapy 19:557-568.

de Lange, T. 2005. Shelterin: the protein complex that shapes and safeguards human telomeres. Genes & Development 19:2100-2110.

de Lange, T. 2009. How telomeres solve the end-protection problem. Science 326:948-952.

Dekaban, A. 1965. Persisting clone of cells with an abnormal chromosome in a woman previously irradiated. Journal of Nuclear Medicine 6:740-746.

Dirac, A. M. G., Huthoff, H., Kjems, J. e Berkhout, B. 2002. Requirements for RNA heterodimerization of the human immunodeficiency virus type 1 (HIV-1) and HIV-2 genomes. Journal of General Virology 83:2533-2542.

Dolan, M. 2011. The role of Giemsa stain in cytogenetics. Biotechnic & Histochemistry 86:94-97.

Durkin, S. G. e Glover, T. W. 2007. Chromosome fragile sites. Annual Review of Genetics 41:169-192.

Edelstein, M. L., Abedi, M., R. e Wixon, J. 2007. Gene therapy clinical trials wordwild to 2007 - an update. The Journal of Gene Medicine 9:833-842.

Elleder, D., Pavlícek, A., Paces, J. e Hejnar, J. 2002. Preferential integration of human immunodeficiency virus type 1 into genes, cytogenetic R bands and GC-rich DNA regions: insight from the human genome sequence. Federation of European Biochemical Societies 517:285-286.

Engelman, A. e Cherepanov, P. 2012. The structural biology of HIV-1: mechanistic and therapeutic insights. Nature Reviews Microbiology 10:279-290.

Espeseth, A. S., Fishel, R., Hazuda, D., Huang, Q., Xu, M., Yoder, K. e Zhou, H. 2011. siRNA screening of a targeted library of DNA repair factors in HIV infection reveals a role for base excision repair in HIV integration. PLos ONE 6:1-8.

Federico, C., Andreozzi, L., Saccone, S. e Bernardi, G. 2000. Gene density in the Giemsa bands of human chromosomes. Chromosome Research 8:737-746.

Feitelson, M. A. e Lee, J. 2007. Hepatitis B virus integration, fragile sites, and hepatocarcinogenesis. Cancer Letters 252:157-170.

Fletcher, T. M., Brichacek, B., Sharova, N., Newman, M. A., Stivahtis, G., Sharp, P. M., Emerman, M., Hahn, B. H. e Stevenson, M. 1996. Nuclear import and cell cycle arrest functions of the HIV-1 Vpr protein ar encoded by two separate genes in HIV-2/SIV_{SM}. 15:6155-6165.

Francke, U. 1994. Dgitized and differentially shaded human chromosome ideograms for genomic applications. Cytogenetics and Cell Genetics 65:206-218.

Francke, U. 2013. 2012 Willian Allan Award: Adventures in cytogenetics. The American Journal of Human Genetics 92:325-337.

Freudenreich, C. H. 2007. Chromosome fragility: molecular mechanisms and cellular consequences. Frontiers in Bioscience 12:4911-4924.

Fun, A., Wensing, A. M. J., Verheyen, J. e Nijhuis, M. 2012. Human Immunodeficiency Virus gag and protease: partners in resistance Retrovirology 9:63-77.

Fungtammasan, A., Walsh, E., Chiaromonte, F., Eckert, K. A. e Makova, K. D. 2012. A genome-wide analysis of common fragile sites: What features determine chromosomal instability in the human genome? Genome Research 22:993-1005.

Ganser-Pornillos, B. K., Yeager, M. e Sundquist, W. I. 2008. The structural biology of HIV assembly. Current Opinion in Structural Biology 18:203-217.

Gedeon, A. K., Baker, E., Robinson, H., Partington, M. W., Gross, B., Manca, A., Korn, B., Poustka, A., Yu, S., Sutherland, G. R. e Mulley, J. C. 1992. Fragile X syndrome without CCG amplification has an FMR1 deletion. Nature Genetics 1:341-344.

Glover, T. W. e Stein, C. K. 1987. Induction os sister chromatid exchanges at common fragile sites. American Society of Human Genetics 41:882-890.

Glover, T. W. e Stein, C. K. 1988. Chromosome breakage and recombination at fragile sites. American Society of Human Genetics 43:265-273.

Groschel, B. e Bushman, F. 2005. Cell cycle arrest in G_2/M promotes early steps of infection by human immunodeficiency virus. Journal of Virology 79:5695-5704.

Hacein-Bey-Abina, S., Hauer, J., Lim, A., Picard, C., Wang, G. P., Berry, C. C., Martinache, C., Rieux-Laucat, F., Latour, S., Belohradsky, B. H., Leiva, L., Sorensen, R., Debré, M., Casanova, J. L., Blanche, S., Durandy, A., Bushman, F. D., Fischer, A. e Cavazzana-Calvo, M. 2010. Efficacy of gene therapy for X-linked severe combined immunodeficiency. The New England Journal of Medicine 363:355-364.

Harrison, J. C. e Haber, J. E. 2006. Surviving the breakup: the DNA damage checkpoint. The Annual Review of Genetics 40:209-235.

Hayden, K. E. 2012. Human centromere genomics: now it´s personal. Chromosome Research 20:621-633.

Herschhorn, A. e Hizi, A. 2010. Retroviral reverse transcriptases. Cellular and Molecular Life Sciences 67:2717-2747.

Hiratani, I., Takebayashi, S.-i., Lu, J. e Gilbert, D. M. 2009. Replication timing and transcriptional control: beyond cause and effect-part II. Current Opinion in Genetics & Development 19:142-149.

Hirsch, B. 1991. Sister chromatid exchanges are preferentially induced at expressed and nonexpressed common fragile sites. Human genetics 87:302-306.

Holden, J. J., Ridgway, P. e Smith, A. 1986. A possible fragile-site at Yq12: case report and possible relevance to *de novo* structural rearrangements of the Y-chromosome. American Journal of Medical Genetics 23:545-555.

Holmquist, G. P. 1992. Chromosome bands, their chromatin flavors, and theis functional features. The American Journal of Human Genetics 51:17-37.

Holmquist, G. P. e Ashley, T. 2006. Chromosome organization and chromatin modification: influence on genome function and evolution. Cytogenetic and Genome Research 114:96-125.

Hoshi, O. e Ushiki, T. 2011. Replication banding patterns in human chromosomes detected using 5-ethynyl-2´-deoxyuridine incorporation. Acta of Histochemica et Cytochemica 44:233-237.

Howe, S. J., Mansour, M. R., Schwarzwaelder, K., Bartholomae, C., Hubank, M., Kempski, H., Brugman, M. H., Pike-Overzet, K., Chatters, S. J., de Ridder, D., Gilmour, K. C., Adams, S., Thornhill, S. I., Parsley, K. L., Staal, J. T., Gale, R. E., Linch, D. C., Bayford, J., Brown, L., Quaye, M., Kinnon, C., Ancliff, P., Webb, D. K., Schmidt, M., von Kalle, C., Gaspar, H. B. e Thrasher, A. J. 2008. Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients. The Journal of Clinical Investigation 118:3143-3150.

Hussein, S. M., Batada, N. N., Vuoristo, S., Ching, R. W., Autio, R., Narva, E., Ng, S., Sourour, M., Hamalainen, R., Olsson, C., Lundin, K., Mikkola, M., Trokovic, R., Peitz, M., Brustle, O., Bazett-Jones, D. P., Alitalo, K., Lahesmaa, R., Nagy, A. e Otonkoski, T. 2011. Copy number variation and selection during reprogramming to pluripotency. Nature 471:58-64.

International Human Genome Sequencing Consortium 2001. Initial sequencing and analysis of the human genome. Nature 409:860-921.

Ito, P. K. 1980. Robusteness of ANOVA and MANOVA test procedures. *In* Handbook of Statistics Analysis of Variance (Krishnaiah, P. K. ed) pp 199-236, North-Holland Publishing Company.

Jansen, L. E. T., Black, B. E., Foltz, D. R. e Cleveland, D. W. 2007. Propagation of centromeric chromatin requires exit from mitosis. The Journal of Cell Biology 176:795-805.

Jones, C., Penny, L., Mattina, T., Yu, S., Baker, E., Voullaire, L., Langdon, W. Y., Sutherland, G. R., Richards, R. I. e Tunnacliffe, A. 1995. Association of a chromosome deletion syndrome with a fragile site within the proto-oncogene CBL2. Nature 376:145-149.

Jordan, A., Defechereux, P. e Verdin, E. 2001. The site of HIV-1 integration in the human genome determines basal transcriptional activity and response to Tat trasactivaction. The EMBO Journal 20:1726-1738.

Kalpana, G. V., Marmon, S., Wang, W., Crabtree, G. R. e Goff, S. P. 1994. Binding and stimulation of HIV-1 integrase by a human homolog of yeast transcritpion factor SNF5 Science 266:2002-2006.

Kanki, P., M'Boup, S., Marlink, R., Travers, K., Hsieh, C. C., Gueye, A., Boye, C., Sankale, J. L., Donnelly, C., Leisenring, W. e et al. 1992. Prevalence and risk determinants of human immunodeficiency virus type 2 (HIV-2) and human immunodeficiency virus type 1 (HIV-1) in west African female prostitutes. Am J Epidemiol 136:895-907.

Kay, M. A., Glorioso, J. C. e Naldini, L. 2001. Viral vectors for gene theraphy: the art of turning infectious agents into vehicles of therapeutics. Nature Medicine 7:33-40.

Khanna, K. K. e Jackson, S. P. 2001. DNA double-strand breaks: signaling, repair and the cancer connection. Nature Genetics 27:247-254.

Kim, S., Kim, N., Dong, B., Boren, D., Lee, S. A., Gupta, J. D., Gaughan, C., Klein, E. A., Lee, C., Silverman, R. H. e Chow, S. A. 2008. Integration site preference of xenotropic murine leukemia virus-related, a new human retrovirus associated with prostate cancer. Journal of Virology 82:9964-9977.

Knight, S. J. L., Flannery, A. V., Hirst, M. C., Campbell, L., Christodoulou, Z., Phelps, S. R., Pointon, J., Middleton-Price, H. R., Barnicoat, A., Pembrey, M. E., Holland, J., Oostra, B. A., Bobrow, M. e Davies, K. E. 1993. Trinucleotide repeat amplification and hypermethylation of a CpG island in FRAXE mental retardation. Cell 74:127-134.

Küpper, K., Kölbl, A., Biener, D., Dittrich, S., von Hase, J., Thormeyer, T., Fiegler, H., Carter, N. P., Speicher, M. R., Cremer, T. e Cremer, M. 2007. Radial chromatin positioning is shaped by local gene density, not by gene expression. Chromosoma 116:285-306.

Laganà, A., Russo, F., Sismeiro, C., Giugno, R., Pulvirenti, A. e Ferro, A. 2010. Variability in the incidence of miRNAs and genes in fragile sites and the role of repeats and CpG islands in the distribution of genetic material. PLos ONE 5:1-8.

Layne, S. P., Merges, M. J., Dembo, M., Spouge, J. L., Conley, S. R., Moore, J. P., Raina, J. L., Renz, H., Gelderblom, H. R. e Narat, P. L. 1992. Factors underlying spontaneous inactivation and susceptibility to neutralization of human immunodeficiency virus. Virology 189:695-714.

Lewin, B. 2004. Genes VIII. Pearson Prentice Hall, New Jersey, United States of America.

Lewinski, M. K., Bisgrove, D., Shinn, P., Chen, H., Hoffman, C., Hannenhali, S., Verdin, E., Berry, C. C., Ecker, J. R. e Bushman, F. D. 2005. Genome-wide analysis of chromosomal features repressing human immunodeficiency virus transcription. Journal of Virology 79:6610-6619.

Li, L., Olvera, J. M., Yoder, K. E., Mitchell, R. S., Butler, S. L., Lieber, M., Martin, S. L. e Bushman, F. D. 2001. Role of the non-homologous DNA end joining pathway in the early steps of retroviral infection. The EMBO Journal 20:3272-3281.

Lilley, C. E., Schwartz, R. A. e Weitzman, M. D. 2007. Using or abusing: viruses and the cellular DNA damage response. Trends in Microbiology 15:119-126.

Liu, D., O'Connor, M. S., Qin, J. e Songyang, Z. 2004. Telosome, a mammalian telomere-associated complex formed by multiple telomeric proteins. The Journal of Biological Chemistry 279:51338-51342.

Lu, W., Zhang, Y., Liu, D., Songyang, Z. e Wan, M. 2013. Telomeres - structure, function, and regulation. Experimental Cell Research 319:133-141.

Lukusa, T. e Fryns, J. P. 2008. Human chromosome fragility. Biochimica et Biophysica Acta 1779:3-16.

Luo, W.-J., Takakuwa, T., Ham, M. F., Wada, N., Liu, A., Fujita, S., Sakane-Ishikawa, E. e Aozasa, K. 2004. Epstein-Barr virus is integrated between REL and BCL-11A in American Burkitt lymphoma cell line (NAB-2). Laboratory Investigation 84:1193-1199.

Ma, K., Qiu, L., Mrasek, K., Zhang, J., Liehr, T., Quintana, L. G. e Li, Z. 2012. Common fragile sites: genomic hotspots of DNA damage and carcinogenesis. International Journal of Molecular Sciences 13:11974-11999.

MacNeil, A., Sankalé, J.-L., Meloni, S. T., Sarr, A. D., Mboup, S. e Kanki, P. 2006. Genomic sites of Human Immunodeficiency Virus type 2 (HIV-2) integration: similarities to HIV-1 in vitro and possible differences in vivo. Journal of Virology 80:7316-7321.

Magenis, R. E., Hecht, F. e Lovrien, E. W. 1970. Heritable fragile site on chromosome 16: probable localization of haptoglobin locus in man. Science 170:85-87.

Manuelidis, L. 1978. Chromosomal localization of complex and simple repeat human DNAs. Chromosoma 66:23-32.

Matovina, M., Sabol, I., Grubisic, G., Gasperov, N. M. e Grce, M. 2009. Identification of human papillomavirus type 16 integration sites in high-grade precancerous cervical lesions. Gynecologic Oncology 113:120-127.

Meuleman, W., Peric-Hupkes, D., Kind, J., Beaudry, J.-B., Pagie, L., Kellis, M., Reinders, M., Wessels, L. e van Steensel, B. 2013. Constitutive nuclear lamina-genome interactions are highly conserved and associations with A/T-rich sequence. Genome Research 23:270-280.

Mitchell, R. S., Beitzel, B. F., Schroder, A. R. W., Shinn, P., Chen, H., Berry, C. C., Ecker, J. R. e Bushman, F. D. 2004. Retroviral DNA integration: ASLV, HIV and MLV show distinct target site preferences. PLOS Biology 2:1127-1137.

Mitsui, J., Takahashi, Y., Goto, J., Tomiyama, H., Ishikawa, S., Yoshino, H., Minami, N., Smith, D. I., Lesage, S., Aburatani, H., Nishino, I., Brice, A., Hattori, N. e Tsuji, S. 2010. Mechanisms of genomic instabilities underlying two common fragile-site-associated loci, PARK2 and DMD, in germ cell and cancer cell lines. The American Journal of Human Genetics 87:75-89.

Mitsui, J. e Tsuji, S. 2011. Common chromosomal fragile sites: breakages and rearrangements in somatic and germline cells. Atlas of Genetics and Cytogenetics in Oncology and Haematology 15:1089-1096.

Mortusewicz, O., Patrick, H. e Helleday, T. 2013. Early replication fragile sites: where replication-transcription collisions cause genetic instability. The EMBO Journal 32:493-495.

Mrasek, K., Schoder, C., Teichmann, A.-C., Behr, K., Franze, B., Wilhelm, K., Blaurock, N., Claussen, U., Liehr, T. e Weise, A. 2010. Global screening and extended nomenclature for 230 aphidicolin-inducible fragile sites, including 61 yet unreported ones. International Journal of Oncology 36:929-940.

Nagel, J., Grob, B., Meggendorfer, M., Preiss, C., Grez, M., Brack-Werner, R. e Dietzel, S. 2012. Stably integrated and expressed retroviral sequences can influence nuclear location and chromatin condensation of the integration locus. Chromosoma 121:353-367.

Nakai-Murakami, C., Shimur, M., Kinomoto, M., Takizawa, Y., Tokunaga, K., Taguchi, T., Hoshino, S., Miyagawa, K., Sata, T., Kurumizaka, H., Yuo, A. e Ishizaka, Y. 2007. HIV-1 Vpr induces ATM-dependent cellular signal with enhanced homologous recombination. Oncogene 26:477-486.

Needleman, S. B. e Wunsch, C. D. 1970. A general method applicable to the search for similarities in the amino acid sequenced of two proteins. Journal of Molecular Biology 48:443-453.

Nightingale, K. P., Wellinger, R. E., Sogo, J. M. e Becker, P. B. 1998. Histone acetylation facilitates RNA polymerase II transcription of the *Drosophila hsp26* gene in chromatin. The EMBO Journal 17:2865-2876.

Niimura, Y. e Gojobori, T. 2002. *In silico* chromosome staining: Reconstruction of Giemsa bands from the whole human genome sequence. PNAS 99:797-802.

Nussbaum, R. L., McInnes, R. R. e Willard, H. F. 2007. Thompson & Thompson Genetics in Medicine. Saunders Elsevier, Philadelphia.

Nyamweya, S., Hegedus, A., Jaye, A., Rowland-Jones, S., Flanagan, K. L. e Macallan, D. C. 2013. Comparing HIV-1 and HIV-2 infections: Lessons for viral immunopathogenesis. Reviews in Medical Virology 23:221-240.

Ozeri-Galai, E., Schwartz, M., Rahat, A. e Kerem, B. 2008. Interplay between ATM and ATR in the regulation of common fragile site stability. Oncogene 27:2109-2117.

Palm, W. e de Lange, T. 2008. How shelterin protects mammalian telomeres. Annual Review of Genetics 42:301-334.

Pardue, M. L. e Gall, J. G. 1970. Chromosomal localization of mouse satellite DNA. Science 168:1356-1358.

Pekarsky, Y., Zanesi, N., Palamarchuk, A., Huebner, K. e Croce, C. M. 2002. *FHIT*: from gene discovery to cancer treatment and prevention. The Lancet Oncology 3:748-754.

Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M. e Ho, D. D. 1996. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science 271:1582-1586.

Pestana, D. D. e Velosa, S. F. 2002. Introdução à Probabilidade e à Estatística. Fundação Calouste Gulbenkian, ISNB:972-31-0954-9.

Petit, S. C., Moody, M. D., Wehbie, R. S., Kaplan, A. H., Nantermet, P. V., Klein, C. A. e Swanstrom, R. 1994. The p2 domain of human immunodeficiency virus type 1 gag regulates sequential proteolytic processing and is required to produce fully infections virions. Journal of Virology 68:8017-8027.

Pyatenko, V. S., Eidelman, Y. A., Khvostunov, I. K. e Andreev, S. G. 2013. Radiation- induced chromosomal instability under constrained growth of irradiated cells. Biochemistry, Biophysics and Molecular Biology 451:190-193.

Pyeon, D., Pearce, S. M., Lank, S. M., Ahlquist, P. e Lambert, P. F. 2009. Establishment of human papillomavirus infection requires cell cycle progression. PLoS Pathogens 5:1-9.

Reich, D., Patterson, N., De Janger, P. L., McDonald, G., Waliszewska, A., Tandon, A., Lincol, R. R., DeLoa, C., Fruhan, S. A., Cabre, P., Bera, O., Semana, G., Kelly, A. M., Francis, D. A., Ardlie, K., Omar, K., Cree, B. A., Hausser, S. L., Oksenberg, J. R. e Hafler, D. A. 2005. A whole-genome adixture scan finds a candidate locus for multiple sclerosis susceptibility. Nature Genetics 37:1113-1118.

Richter, T. e von Zglinicki, T. 2007. A continuous correlation between oxidative stress and telomere shortening in fibroblasts. Experimental Gerontology 42:1039-1042.

Riethman, H., Ambrosini, A., Castaneda, C., Finklestein, J., Hu, X.-L., Mudunuri, U., Paul, S. e Wei, J. 2004. Mapping and initial analysis of human subtelomeric sequence assemblies. Genome Research 14:18-28.

Romani, B. e Engelbrecht, S. 2009. Human immunodeficiency virus type 1 Vpr: functions and molecular interactions. Journal of General Virology 90:1795-1805.

Ruiz-Herrera, A., Ponsà, M., García, F., Egozcue, J. e Gracía, M. 2002. Fragile sites in human and *Macaca fascicularis* chromosomes are breakpoints in chromosome evolution. Chromosome Research 10:33-44.

Sadoni, N., Langer, S., Fauth, C., Bernardi, G., Cremer, T., Turner, B. M. e Zink, D. 1999. Nuclear organization of mammalian genomes: polar chromosome territories build up functionally distinct higher order compartments. The Journal of Cell Biology 146:1211-1226.

Samoshkin, A., Arnaoutov, A., Jansen, L. E. T., Ouspenski, I., Dye, L., Karpova, T., McNally, J., Dasso, M., Cleveland, D. W. e Strunnikov, A. 2009. Human condensin function is essential for centromeric chromatin assembly and proper sister kinetochore orientation. PLos ONE 4:6831-6846.

Sawaya, B. E., Khalili, K., Gordon, J., Taubes, R. e Amini, S. 2000. Cooperative interaction between HIV-1 regulatory proteins Tat and Vpr modulates transcription of the viral genome. The Journal of Biological Chemistry 275:35209-35214.

Schluth-Bolard, C., Ottaviani, A., Bah, A., Boussouar, A., Gilson, E. e Magdinier, F. 2010. Dynamics and plasticity of chromosome ends: consequences in human pathologies. Atlas of Genetics and Cytogenetics in Oncology and Haematology 14:501-524.

Schröder, A. R. W., Shinn, P., Huaming, C., Berry, C., Ecker, J. R. e Bushman, F. 2002. HIV-1 integration in the human genome favors active genes and local hotspots. Cell 110:521-529.

Schueler, M. G., Higgins, A. W., Rudd, M. K., Gustashaw, K. e Willard, H. F. 2001. Genomic and genetic definition of a functional human centromere. Science 294:109-115.

Schwartz, M., Zlotorynski, E., Golderg, M., Ozeri, E., Rahat, A., le Sage, C., Chen, B. P. C., Chen, D. J., Agami, R. e Kerem, B. 2005. Homologous recombination and nonhomologous end-joining repair pathways regulate fragile site stability. Genes & Development 19:2715-2726.

Schwartz, M., Zlotorynski, E. e Kerem, B. 2006. The molecular basis of common and rare fragile sites. Cancer Letters 232:13-26.

Seabright, M. 1971. A rapid banding technique for human chromosomes. The Lancet 2:971-972.

Sequeira, I. J., Mexia, J. T., Santiago, J., Mamede, R., Silva, E., Santos, J., Faria, D., Rueff, J. e Brás, A. 2013. Predominance of constitutional chromosomal rearrangements in human chromosomal fragile sites. Open Journal of Genetics 3:8-13.

Shi, Y., Zou, M., Farid, N. R. e Paterson, M. C. 2000. Association of *FHIT* (fragile histidine triad), a candidate tumour suppressor gene, with the ubiquitin-conjugating enzyme hUBC9. Biochemical Journal 353:443-448.

Siegel, S. 1975. O caso de duas amostras relacionadas. *In* Estatística Não-Paramétrica para as Ciências do Comportamento (McGraw-Hill ed pp 75-93, Brasil.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B. e de Laat, W. 2006. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nature Genetics 38:1348-1354.

Skalka, A. e Katz, R. 2005. Retroviral DNA integration and the DNA damage response. Cell Death and Differentiation 12:971-978.

Smith, J. A. e Daniel, R. 2006. Following the path of the virus: The exploitation of host DNA repair mechanisms by retroviruses. ACS Chemical Biology 1:217-226.

Smogorzewska, A., Karlseder, J., Holtgreve-Grez, H., Jauch, A. e de Lange, T. 2002. DNA ligase IV-dependent NHEJ of deprotected mammalian telomeres in G1 and G2. Current Biology 12:1635-1644.

Soto, M. J., Peña, À. e Vallejo, F. G. 2011. A genomic and bioinformatics analysis of the integration of HIV in peripheral blood mononuclear cells. AIDS Research and Human Retroviruses 27:547-555.

Speicher, M. R. 2010. Chromosomes. *In* Vogel and Motulsky´s Human Genetics (Speicher, M. R., Antonarakis, S. E., Motulsky, A. G. eds), 4th ed., pp 55-86, Springer, London.

Spurbeck, J. L., Adams, S. A., Stupca, P. J. e Dewald, G. W. 2004. Primer on medical genomics; Part XI: visualizing human chromosomes. Medical Genomics 79:58-75.

Stacey, S. N., Manolescu, A., Sulem, P., Thorlacius, S., Gudjonsson, S. A., Jonsson, G. F., Jakobsdottir, M., Berghtorsson, J. T., Gudmundsson, J., Aben, K. K., Strobbe, L. J., Swinkels, D. W., van Engelenburg, K. C., Henderson, B. E., Kolonel, L. N., Le Marchand, L., millastre, E., Andress, R., Saez, B., Lambea, J., Godino, J., Polo, E., Tres, A., Picelli, S., Rantala, J., Margolin, S., Jonsson, T., Sigurdsson, H., Jonsdottir, T., Hrafnkelsson, J., Johannsson, J., Sveinsson, T., Myrdal, G., Grimsson, H. N., Sveinsdottir, S. G., Alexiusdottir, K., Saemundsdottir, J., Sigudsson, A., Kostic, J., Gudmundsson, L., Kristjansson, K., Masson, G., Fackenthal, J. D., Adebamowo, C., Ogundiran, T., Olopade, O. I., Haiman, C. A., Lindblom, A., Mayordomo, J. I., Kiemeney, L. A., Gulcher, J. R., Rafnar, T., Thorsteinsdottir, U., Johannsson, O. T., Kong, A. e Stefansson, K. 2008. Common variants on chromsome 5_p12 confer susceptibility to estrogen receptor-positive brast cancer. Nature Genetics 40:703-706.

Stefano, M. D., Rosa, A., Belcastro, V., di Bernardo, D. e Micheletti, C. 2013. Colocalization of coregulated genes: a steered molecular dynamics study of human chromosome19. PLOS Computational Biology 9:1-13.

Sullivan, G. J., Bridger, J. M., Cuthbert, A. P., Newbold, R. F., Bickmore, W. A. e McStay, B. 2001. Human acrocentric chromosomes with transcriptionally silent nucleolar organizer regions associate with nucleoli. The EMBO Journal 20:2867-2877.

Sutherland, G. R. 1988. The role of nucleotides in human fragile site expression. Mutation Research 200:207-213.

Sutherland, G. R. e Baker, E. 2000. The clinical significance of fragile sites on human chromosomes. Clinical Genetics 58:157-161.

Sutherland, G. R., Parslow, M. I. e Baker, E. 1985. New classes of common fragile sites induced by 5-azacytidine and bromodeoxyuridine. Human genetics 69:233-237.

Szenker, E., Ray-Gallet, D. e Almouzni, G. 2011. The double face of the histone variant H3.3. Cell Research 21:421-434.

Tse, C., Sera, T., Wolffe, A. P. e Hansen, J. C. 1998. Disruption of higher-order folding by core histone acetylation dramatically enhances transcription of nucleosomal arrays by RNA polymerase III. Molecular and Cellular Biology 18:4629-4638.

Turlure, F., Devroe, E., Silver, P. A. e Engelman, A. 2004. Human cell proteins and Human Immunodeficiency Virus DNA integration. Frontiers in Bioscience 9:3187-3208.

Uhlmann, F. 2013. Open questions: chromosome condensation-why does chromosomes look like a chromosome? BMC Biology 11:9-10.

Vilanova, M. e Ferreira, P. 2007. Imunologia da Infecção. *In* Fundamentos de Imunologia (Arosa, F. A., Cardoso, E. M., Pacheco, F. C. eds), pp 193-195, LIDEL, Lisboa.

Vissers, L. E. L. M. e Stankiewicz, P. 2012. Microdeletion and microduplication syndromes. *In* Genomic Structural Variants Methods and Protocols (Feuk, L. ed pp 50-51, Springer, London.

von Zglinicki, T. 2002. Oxidative stress shortens telomeres. Trends in Biochemical Sciences 27:339-344.

Wang, G. P., Ciuffi, A., Leipzig, J., Berry, C. C. e Bushman, F. D. 2007. HIV integration site selection: Analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. Genome Research 17:1186-1194.

Wang, Y.-H. 2006. Chromatin structure of human chromosomal fragile sites. Cancer Letters 232:70-78.

Warburton, P. E., Greig, G. M., Haa, T. e Willard, H. F. 1991. PCR amplification of chromosome-specific alpha satellite DNA: definition of centromeric STS markers and polymorphic analysis. Genomics 11:324-333.

Watanable, Y., Ikemura, T. e Sugimura, H. 2004. Amplicons on human chromosome 11q are located in the early/late-switch regions of replication timing. Genomics 84:796-805.

Watanable, Y. e Maekawa, M. 2013. R/G-band boundaries: Genomic instability and human disease. Clinica Chimica Acta 419:108-112.

Wilke, C. M., Hall, B. K., Hoge, A., Paradee, W., Smith, D. I. e Glover, T. W. 1996. FRA3B extends over a broad region and contains a spontaneous HPV16 integration site: direct evidence for the coincidence of viral integration sites and fragile sites. Human Molecular Genetics 5:187-195.

Wöhrle, D., Kotzot, D., Hrst, M. C., Manca, A., Korn, B., Schmidt, A., Barbi, G., Rott, H.-D., Poustka, A., Davies, K. E. e Steinbach, P. 1992. A microdeletion of less than 250Kb, including the proximal part of the FMR-1 gene and the fragile -X site, in a male with the clinical phenotype of fragile-X syndrome. American Journal of Human Genetics 51:299-306.

Wolffe, A. P. 2001. Chromatin remodeling: why it is important in cancer. Oncogene 20:2988-2990.

Worobey, M., Gemmel, M., Teuwen, D. E., Haselkorn, T., Kunstman, K., Bunce, M., Muyembe, J.-J., Kabongo, J.-M., Kalengayi, R. M., Marck, E. V., Gilbert, M. T. P. e Wolinsky, S. M. 2008. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature 455:661-665.

Yang, J. e Zhang, W. 2008. WWOX tumor suppressor gene. Histology and Histopathology 23:877-882.

Yokota, H., Singer, M. J., van den Engh, G. J. e Trask, B. J. 1997. Regional differences in the compaction of chromatin in human G_0/G_1 interphase nuclei. Chromosome Research 5:157-166.

Yong-Gonzalez, V., Wang, B.-D., Butylin, P., Ouspenski, I. e Strunnikov, A. 2007. Condensin function at centromere chromatin facilitates proper kinetochore tension and ensures correct mitotic segregation of sister chromatids. Genes to Cells 12:1075-1090.

Zakian, V. A. 2012. Telomeres: the beginnings and ends of eukaryotic chromosomes. Experimental Cell Research 318:1456-1460.

Zlotorynski, E., Rahat, A., Skaug, J., Ben-Porat, N., Ozeri, E., Hershberg, R., Levi, A., Scherer, S. W., Margalit, H. e Kerem, B. 2003. Molecular basis for expression of common and rare fragile sites. Molecular and Celular Biology 23:7143-7151.

Zody, M. C., Garber, M., Adams, D. J., Sharpe, T., Harrow, J., Lupski, J. R., Nicholson, C., Searle, S. M., Wilming, L., Young, S. K., Abouelleil, A., Allen, N. R., Bi, W., Bloom, T., Borowsky, M. L., Bugalter, B. E., Butler, J., Chang, J. L., Chen, C. K., Cook, A., Corum, B., Cuomo, C. a., de Jong, P. J., Decaprio, D., Dewar, K., FitzGerald, M., Gilbert, J., Gibson, R., Gnerre, S., Goldstein, S., Grafham, D. V., Grocock, R., Hafez, N., Hapogian, D. S., Hart, E., Norman, C. H., Humphray, S., Jaffe, D. B., Jones, M., Kamal, M., Khodiyar, V. K., LaButti, K., Laird, G., Lehoczky, J., Liu, X., Lokyitsang, T., Loveland, J., Lui, A., Macdonald, P., Major, J. E., Matthews, L., Mauceli, E., McCarroll, S. A., Mihalev, A. H., Mudge, J., Nguyen, C., Nicol, R., O´Leary, S. B., Osoegawa, K., Schwartz, D. C., Shaw-Smith, C., Stankiewicz, P., Steward, C., Swarbreck, D., Venkataraman, V., Whittaker, C. A., Yang, X., Zimmer, A. R., Bradley, A., Hubbard, T., Birren, B. W., Rogers, J., Lander, E. S. e Nusbaum, C. 2006. DNA sequence of human chromosome 17 and analysis of rearrangement in the human lineage. Nature 440:1045-1049.

Anexos

Tabela - Representação das posições iniciais e finais de cada banda *Giemsa* escura cedida pelos autores (Niimura e Gojobori, 2002). N.D. representa as posições não detectadas pelo método *in silico*.

Cromossoma 1

Posição Inicial	Posição Final	Banda-G correspondente
4465001	5195000	p36.32
5235001	9265000	p36.23
12445001	15495000	p36.21
20355001	24095000	p36.12
29715001	31705000	p35.3
36605001	40395000	p35.1
46045001	48745000	p34.2
54475001	58835000	p33
63045001	66415000	p32.2
67055001	69905000	p31.3
78455001	84955000	p31.1
97875001	101445000	p22.2
107175001	111255000	p21.3
114315001	119995000	p21.1
124005001	127535000	p13.2
130325001	132845000	p12
133214276	163214275	Centrómero + heterocromatina
169959276	174779275	q21.2
178759276	181399275	q22
183779276	186489275	q23.2
190509276	196309275	q24.1
199009276	201989275	q24.3
207219276	209619275	q25.2
215959276	221159275	q31.1
223799276	229539275	q31.3
236709276	242769275	q32.2
247249276	252289275	q41
255049276	256069275	q42.12
265029276	268799275	q42.2
272299276	276949275	q43

Cromossoma 2

Posição Inicial	Posição Final	Banda-G correspondente
3645001	8005000	p25.2
17185001	18925000	p24.3
21125001	24215000	p24.1
31535001	31585000	p23.2

33165001	37925000	p22.3
40995001	43765000	p22.1
50705001	55275000	p16.3
58695001	62335000	p16.1
65145001	66995000	p14
69025001	71945000	p13.2
78815001	82395000	p12.0
94725253	97725252	Centrómero
106220253	109130252	q12.1
110880253	113240252	q12.3
118180253	122370252	q14.1
126620253	131290252	q14.3
137160253	139350252	q21.2
143860253	148900252	q22.1
148920253	148940252	q22.3
150910253	154560252	q23.2
159330253	163250252	q24.1
168060253	174920252	q24.3
186790253	189880252	q31.2
191860253	196570252	q32.1
200040253	203640252	q32.3
206640253	209130252	q33.2
218890253	225530252	q34
230630253	233870252	q36.1
235370253	240470252	q36.3
244560253	249410252	q37.2

Cromossoma 3

Posição Inicial	Posição Final	Banda-G correspondente
1	4355000	p26.3
7885001	11865000	p26.1
15765001	15795000	p25.2
16285001	18455000	p24.3
23105001	30325000	p24.1
35355001	38545000	p22.3
41375001	44315000	p22.1
51065001	54475000	p21.32
67485001	70915000	p21.2
75175001	78175000	p14.3
86205001	86835000	p14.1
89315001	93145000	p12.3
96575001	100285000	p12.1
103476284	106476283	Centrómero

106476284	111401283	q12.2
115751284	120251283	q13.11
123421284	126251283	q13.13
129451284	133711283	q13.31
135961284	138721283	q13.33
141281284	143601283	q21.2
148021284	152211283	q22.1
154971284	158071283	q22.3
163761284	169151283	q24
173131284	177041283	q25.2
179891284	181431283	q25.32
184711284	193391283	q26.1
197931284	201611283	q26.31
204381284	208121283	q26.33
209101284	209331283	q27.2
215031284	222241283	q28

Cromossoma 4

Posição Inicial	Posição Final	Banda-G correspondente
3635001	6375000	p16.2
10465001	14995000	p15.33
19355001	24875000	p15.31
35445001	39385000	p15.1
45875001	50185000	p13
52289776	55289775	Centrómero
61554776	65204775	q13.1
66614776	70534775	q13.3
76074776	78924775	q21.21
82394776	85504775	q21.23
94034776	98164775	q22.1
100544776	103214775	q22.3
109134776	114044775	q24
121424776	125614775	q26
131804776	135294775	q28.1
138704776	146674775	q28.3
150294776	153964775	q31.21
157554776	160584775	q31.23
169854776	173154775	q32.1
175694776	177094775	q32.3
179644776	182014775	q34.1
186274776	192174775	q34.3
196924776	200214775	q35.2

Cromossoma 5

Posição Inicial	Posição Final	Banda-G correspondente
3525001	9115000	p15.32
13885001	17245000	p15.2
22315001	26375000	p14.3
27545001	31605000	p14.1
35125001	38575000	p13.2
44275001	47215000	p12
50695404	53695403	Centrómero
62670404	66430403	q12.1
69170404	72950403	q12.3
77400404	77850403	q13.2
81700404	84450403	q14.1
98860404	104620403	q14.3
108970404	113200403	q21.1
114950404	117970403	q21.3
122480404	123760403	q22.2
129430404	131920403	q23.1
141620404	145890403	q23.3
152130404	155700403	q31.2
161180404	167900403	q32
171380404	174110403	q33.2
181620404	187960403	q34
194660404	197700403	q35.2

Cromossoma 6

Posição Inicial	Posição Final	Banda-G correspondente
995001	1975000	p25.2
8335001	11645000	p24.3
12025001	13755000	p24.1
18725001	21695000	p22.3
23205001	26665000	p22.1
31805001	34025000	p21.32
43015001	46125000	p21.2
50755001	57225000	p12.3
60135001	63195000	p12.1
63993123	66993122	Centrómero
71128123	76458122	q12
81808123	85558122	q14.1
85918123	89348122	q14.3
99408123	104028122	q16.1
108288123	112698122	q16.3
120578123	125358122	q22.1
128638123	132118122	q22.31

135628123	138418122	q22.33
143438123	144398122	q23.2
148898123	151938122	q24.1
154608123	158068122	q24.3
162148123	166838122	q25.2
173118123	178578122	q26

Cromossoma 7

Posição Inicial	Posição Final	Banda-G correspondente
7415001	11685000	p22.2
13625001	15875000	p21.3
17975001	21275000	p21.1
27115001	28725000	p15.2
32895001	36175000	p14.3
38375001	43885000	p14.1
47985001	51475000	p12.3
53395001	56385000	p12.1
58586186	61586185	Centrómero
69981186	73701185	q11.22
78141186	82631185	q21.11
89681186	91631185	q21.13
94331186	99431185	q21.3
105181186	108061185	q22.2
111211186	114651185	q31.1
122201186	124981185	q31.31
127651186	131271185	q31.33
135251186	136461185	q32.2
139021186	144251185	q33
150401186	154721185	q35
159141186	162471185	q36.2

Cromossoma 8

Posição Inicial	Posição Final	Banda-G correspondente
3185001	7085000	p23.2
14325001	19345000	p22
26235001	29565000	p21.2
33785001	40755000	p12
44155001	45895000	p11.22
48399986	51399985	Centrómero
54424986	57924985	q11.22
64134986	66554985	q12.1
68924986	72234985	q12.3
74704986	76294985	q13.2
82394986	86574985	q21.11

89964986	92904985	q21.13
95304986	98704985	q21.3
102024986	104074985	q22.2
112634986	114994985	q23.1
117594986	123604985	q23.3
127804986	130764985	q24.12
134504986	141054985	q24.21
143904986	148054985	q24.23

Cromossoma 9

Posição Inicial	Posição Final	Banda-G correspondente
2365001	5325000	p24.2
9395001	14815000	p23
18515001	18695000	p22.2
22325001	27825000	p21.3
30385001	34355000	p21.1
38915001	39125000	p13.2
42935001	45670000	p12
45666974	67666973	Centrómero + heterocromatina
70211974	70761973	q21.11
72601974	75731973	q21.13
79671974	83091973	q21.31
85691974	87631973	q21.33
91251974	93721973	q22.2
96811974	99341973	q22.32
102811974	107611973	q31.1
112391974	115171973	q31.3
118391974	122941973	q33.1
123071974	123081973	q33.3
125731974	129851973	q34.12
134011974	137421973	q34.2

Cromossoma 10

Posição Inicial	Posição Final	Banda-G correspondente
2385001	5145000	p15.2
8015001	11375000	p14
18605001	22455000	p12.33
25835001	29925000	p12.31
33385001	36245000	p12.1
37925001	38715000	p11.22
41725291	44725290	Centrómero
52600291	56080290	q11.22
57460291	61600290	q21.1
68230291	73230290	q21.3

77450291	83690290	q22.2
87330291	92280290	q23.1
95180291	98170290	q23.31
100600291	100820290	q23.33
101890291	104880290	q24.2
107340291	109380290	q24.32
114400291	119830290	q25.1
125060291	127870290	q25.3
131990291	133830290	q26.12
138170291	142750290	q26.2

Cromossoma 11

Posição Inicial	Posição Final	Banda-G correspondente
3085001	7205000	p15.4
13785001	16865000	p15.2
21585001	27245000	p14.3
29305001	32915000	p14.1
39235001	45485000	p12
50965001	53110000	p11.12
53107611	56107610	Centrómero
58892611	62442610	q12.1
65312611	67792610	q12.3
71042611	73682610	q13.2
78042611	81642610	q13.4
86262611	90452610	q14.1
96052611	100812610	q14.3
106292611	110732610	q22.1
114122611	117602610	q22.3
119732611	122732610	q23.2
133902611	137302610	q24.1
140602611	143172610	q24.3

Cromossoma 12

Posição Inicial	Posição Final	Banda-G correspondente
3875001	5715000	p13.32
8775001	12335000	p13.2
15455001	19465000	p12.3
21615001	24995000	p12.1
27765001	31055000	p11.22
35960233	38960232	Centrómero
40515233	43495232	q12
45325233	49685232	q13.12
60295233	62225232	q13.2
64175233	68125232	q14.1

72015233	72525232	q14.3
76695233	80195232	q21.1
89745233	93815232	q21.31
95675233	98285232	q21.33
103145233	111115232	q23.1
113395233	116085232	q23.3
119565233	119595232	q24.12
123045233	126125232	q24.21
126415233	128785232	q24.23
136665233	140455232	q24.32

Cromossoma 13

Posição Inicial	Posição Final	Banda-G correspondente
1	16000000	Centrómero + braço p
19685001	22015000	q12.12
25285001	25335000	q12.2
30235001	32235000	q13.1
32265001	38005000	q13.3
45165001	48775000	q14.12
52825001	58285000	q14.2
61935001	65545000	q21.1
68645001	73215000	q21.31
77385001	77945000	q21.33
80575001	80835000	q22.2
83375001	87925000	q31.1
87945001	97585000	q31.3
100425001	102425000	q32.2
106225001	110915000	q33.1
111395001	115305000	q33.3

Cromossoma 14

Posição Inicial	Posição Final	Banda-G correspondente
1	16000000	Centrómero + braço p
21465001	26175000	q12
32845001	32875000	q13.2
36445001	40515000	q21.1
42715001	45895000	q21.3
54255001	55935000	q22.2
56135001	61225000	q23.1
63695001	66205000	q23.3
68985001	72015000	q24.2
77675001	81395000	q31.1
81405001	87955000	q31.3
92495001	92555000	q32.12

96105001	96185000	q32.2
96195001	99845000	q32.32

Cromossoma 15

Posição Inicial	Posição Final	Banda-G correspondente
1	17000000	Centrómero + braço p
22245001	23285000	q12
29685001	34725000	q13.2
39395001	39705000	q14
41635001	46245000	q15.2
49705001	53955000	q21.1
57295001	60255000	q21.3
63735001	64265000	q22.2
67985001	70795000	q22.32
73695001	75445000	q23
N.D.	N.D.	q24.2
78175001	81245000	q25.1
83765001	86845000	q25.3
91855001	96565000	q26.2

Cromossoma 16

Posição Inicial	Posição Final	Banda-G correspondente
8335001	14055000	p13.2
17025001	20105000	p13.12
24275001	27455000	p12.3
29345001	31405000	p12.1
40403277	55403276	Centrómero + heterocromatina
61498277	65378276	q12.2
69648277	77188276	q21
83178277	86278276	q22.2
89358277	92758276	q23.1
94368277	97508276	q23.3
100748277	101548276	q24.2

Cromossoma 17

Posição Inicial	Posição Final	Banda-G correspondente
5765001	6745000	p13.2
11285001	18095000	p12
24844219	27844218	Centrómero
35809219	39019218	q12
41899219	44559218	q21.2
48569219	51949218	q21.32
55679219	60759218	q22
N.D.	N.D.	q23.2

64529219	68069218	q24.1
73589219	77929218	q24.3
83059219	85529218	q25.2

Cromossoma 18

Posição Inicial	Posição Final	Banda-G correspondente
5435001	8615000	p11.31
12775001	13915000	p11.22
18411927	21411926	Centrómero
28626927	32396926	q12.1
40136927	46896926	q12.3
54256927	60056926	q21.2
63756927	66066926	q21.32
68816927	73626926	q22.1
76126927	79696926	q22.3

Cromossoma 19

Posição Inicial	Posição Final	Banda-G correspondente
12985001	15605000	p13.2
17775001	19915000	p13.12
28890319	31890318	Centrómero
35845319	38645318	q13.12
43105319	46325318	q13.2
51775319	55365318	q13.32
65005319	68375318	q13.41
71815319	74765318	q13.43

Cromossoma 20

Posição Inicial	Posição Final	Banda-G correspondente
6125001	9905000	p12.3
14685001	18235000	p12.1
21905001	25125000	p11.22
27795595	30795594	Centrómero
33410595	35210594	q11.22
38510595	43040594	q12
49640595	51440594	q13.12
54250595	57770594	q13.2
60430595	63390594	q13.32

Cromossoma 21

Posição Inicial	Posição Final	Banda-G correspondente
1	11281116	Centrómero + braço p
17196117	22876116	q21.1
25306117	29286116	q21.3

32866117	33176116	q22.12
35626117	40046116	q22.2

Cromossoma 22

Posição Inicial	Posição Final	Banda-G correspondente
1	13000000	Centrómero + braço p
18695001	19425000	q11.22
22065001	26455000	q12.1
29335001	32905000	q12.3
36225001	38945000	q13.2
42885001	45885000	q13.32

Cromossoma X

Posição Inicial	Posição Final	Banda-G correspondente
1645001	5795000	p22.32
11565001	14655000	p22.2
18145001	21265000	p22.12
24635001	27535000	p21.3
27625001	31625000	p21.1
35705001	41185000	p11.3
49725001	53130000	p11.22
53125381	56125380	Centrómero
58640381	62880380	q12
71020381	73600380	q13.2
75270381	79430380	q21.1
82420381	86190380	q21.31
93160381	96640380	q21.33
102270381	105080380	q22.2
111210381	117250380	q23
126330381	130530380	q25
134790381	137370380	q26.2
139120381	141390380	q27.1
146140381	153950380	q27.3

Cromossoma Y

Posição Inicial	Posição Final	Banda-G correspondente
1	3695000	p11.31
8245413	11245412	Centrómero
17910413	19710412	q11.22